Федеральное государственное бюджетное учреждение науки
Институт физиологии им. И. П. Павлова Российской академии наук

На правах рукописи

МАЛАХОВА ЕКАТЕРИНА ЮРЬЕВНА

ИССЛЕДОВАНИЕ И ИНТЕРПРЕТАЦИЯ ПРИ ПОМОЩИ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ МЕХАНИЗМОВ ОПИСАНИЯ ЗРИТЕЛЬНЫХ ОБРАЗОВ В НИЖНЕВИСОЧНОЙ КОРЕ ГОЛОВНОГО МОЗГА ПРИМАТОВ

Специальность 1.5.5 – Физиология человека и животных

Диссертация на соискание ученой степени
кандидата биологических наук

Научный руководитель:
доктор медицинских наук,
профессор
Шелепин Юрий Евгеньевич

Санкт-Петербург
2025

Оглавление

Введение

Актуальность темы исследования

Зрительная система позволяет живым организмам получать информацию о состоянии окружающей среды, осуществлять целенаправленную деятельность и эффективно взаимодействовать с другими организмами и окружающей средой. Понимание принципов зрительного восприятия является одной из ключевых задач нейробиологии.

Распознавание образов напрямую связано с выделением статистических свойств воспринимаемой (наблюдаемой) зрительной сцены. Особенности окружающей среды также оказывают существенное влияние и на формирование самой зрительной системы. Изучению свойств натуральных сцен и их взаимосвязи с характеристиками нейронов в зрительной коре посвящено большое количество научных работ в отечественной и зарубежной литературе.

Сравнительно новым подходом к изучению статистических свойств воспринимаемых изображений является использование искусственных нейронных сетей. Построенные на принципах, присущих зрительной системе приматов, сверточные нейронные сети вывели задачу распознавания образов на новый уровень, зарекомендовав себя не только как универсальный инструмент для работы с изображениями, но и как наиболее точная модель зрительной системы.

Получение новых сведений о принципах обработки информации в зрительной коре головного мозга с использованием нейросетевых моделей и их связи с реальными физиологическими механизмами восприятия представляет интерес как с научной точки зрения, так и для практического применения. Это может способствовать более глубокому пониманию процессов восприятия и принятия решений в биологических системах, а также улучшению стабильности и адаптивности компьютерных систем обработки изображений, что особенно важно в таких областях, как клиническая диагностика, транспортные средства с автоматическим управлением, системы обеспечения безопасности и т. д.

На сегодняшний день хорошо изучены функции нейронов первичной зрительной коры (поле 17 по Бродману BA17), известна организация их рецептивных полей (Hubel, Wiesel, 1962; 1965; Шелепин, 1981; Шевелев, 1984), функциональная и структурная организация проводящих каналов зрительной системы, в том числе наличие специализированных путей распространения информации (Глезер, 1966; Кок, 1967; Mishkin, Ungerleider, 1982; Праздникова, Глезер, Данилова, 1985).

Проводятся исследования процессов и на высших уровнях зрительной системы – заднетеменной и нижневисочной коры (Кок, 1967; Глезер и др., 1975; Яковлев, 1982; Праздникова, Данилова, Мешкенайте, 1989). Внесен большой вклад в изучение механизмов

инвариантности зрительного восприятия как основополагающего принципа распознавания объектов (Глезер, 1966; Шелепин, 1973; Шелепин, Чихман, Фореман, 2008; Яковлев, 1982). Выявлено, что отдельные клетки нижневисочной коры отвечают преимущественно на сложные стимулы, такие как изображения рук и лиц (Gross, Rocha-Miranda, Bender, 1972). В ряде работ было продемонстрировано, что для распознавания объектов в коре формируются представления, чувствительные к углу обзора, но инвариантные к ряду других признаков (Logothetis, Sheinberg, 1996); инвариантность может достигаться за счет непрерывного кодирования признаков близлежащими нейрональными колонками (Tanaka, 1996); кодирование объектов может осуществляться при помощи различных комбинаций активных и неактивных колонок для отдельных признаков (Tsunoda et al., 2001) и др.

Во многом предваряя исследования в области нейрофизиологии, осуществляются и модельные исследования зрительного сигнала как такового. Установлены принципы осуществления временной и пространственно-частотной фильтрации сигнала на начальных этапах обработки (Глезер, Цуккерман, 1961) и последующей согласованной фильтрации зрительной информации на более высоких уровнях (Глезер, 1985; Шелепин, 1981; Glezer, 1995). Предложены модели согласованной фильтрации и распознавания объектов в условиях шума и вариаций входных паттернов (Красильников 1958, 1986; Grossberg, 1976).

Следует отметить, что хотя указанные основополагающие работы и значительно продвинули понимание роли и взаимодействия различных областей мозга в распознавании объектов, тем не менее, точные принципы и механизмы этого процесса до сих пор не до конца понятны и остаются предметом активных исследований. Для изучения работы зрительной системы применяются методы *in silico* моделирования и машинного обучения, позволяющие охватить различные аспекты сенсорного восприятия, исследовать механизмы представления информации и характеристики отдельных нейронов.

В то же время открытия в области нейрофизиологии находят отражение в архитектуре программных решений автоматического распознавания образов, например, в таких широко применяемых моделях, как сверточные нейронные сети (Fukushima, 1980; LeCun et al., 1989). Помимо заложенных структурных аналогий со зрительной системой, в сверточных нейронных сетях наблюдается и ряд функциональных сходств (Cadieu et al., 2014, Schrimpf, 2018). Указанное обстоятельство хотя и не говорит об идентичности протекающих внутри процессов, но позволяет рассматривать их в качестве модели для изучения принципов кодирования зрительной информации.

Актуальным направлением в нейрофизиологии является и разработка методов повышения качества процесса исследования при помощи искусственного интеллекта. Помимо технологических инноваций, направленных на повышение точности работы оборудования и

анализ больших массивов данных, можно выделить проекты, нацеленные на усовершенствование самого экспериментального поиска. Например, методы управления ходом эксперимента, такие как поиск оптимальных настроек экспериментального оборудования при помощи байесовской оптимизации либо же использование алгоритмов активного поиска для максимизации получаемой информации за счет адаптивного подбора стимулов на основе регистрируемых психофизических ответов. В диссертационном исследовании рассматривается возможность направленного создания стимульного материала во время проведения экспериментальной сессии при помощи генеративных методов искусственного интеллекта.

Понимание нейрофизиологических механизмов распознавания образов является значимым также и для области интерпретируемого машинного обучения. Задачей этого отдельного направления на пересечении нейро и компьютерных наук является исследование уже обученных моделей, сопоставление их с восприятием и механизмами принятия решений человеком. Подобные исследования важны для оценки стабильности работы искусственных систем и понимания их возможных ограничений и отличий.

Цель и задачи исследования

В данной работе исследуется процесс формирования и распознавания зрительного образа в коре головного мозга. Значительное внимание уделяется особенностям кодирования зрительных образов и категорий на различных этапах обработки сигнала.

Целью работы является изучение и интерпретация механизмов описания зрительных образов нейронными сетями высших отделов зрительной системы приматов, а именно нижневисочной коры, обеспечивающих распознавание объектов и оценку их значения для наблюдателя, в модельных исследованиях с применением искусственных нейронных сетей, выполняющих аналогичные задачи распознавания.

Поставленная цель достигается решением следующих *задач*:

1. Планирование исследований, предварительная обработка и статистический анализ данных нейрональной активности нижневисочной коры для задач моделирования процесса распознавания зрительных образов.
2. Моделирование ответа нейронов нижневисочной коры на предъявляемые зрительные стимулы при помощи сверточных нейронных сетей; сопоставление модели с нейрофизиологическими данными.
3. Разработка и реализация методов и алгоритмов исследования кодирования зрительной информации и интерпретации работы моделей зрительной системы на различных этапах распознавания, применимых как для изучения свойств искусственных нейронных сетей, так и для обработки данных нейрофизиологических исследований.

4. Исследование взаимосвязи статистических свойств изображений и функций нейронов на различных уровнях обработки зрительного сигнала.

Научная новизна

1. Предложен новый подход к исследованию функций нейронов высших отделов зрительной системы посредством применения генеративных нейронных сетей, а также разработан программный комплекс, позволяющий создание стимулов непосредственно во время проведения нейрофизиологических экспериментов на основе регистрируемой нейрональной активности.

2. Впервые показана возможность контролируемой активации нейрональных колонок при помощи создания абстрактных экспериментальных стимулов во время проведения нейрофизиологического эксперимента. Проанализированы характеристики подхода и их влияние на нейрональный ответ.

3. Проведен многосторонний анализ применения искусственных нейронных сетей в модельных исследованиях высших отделов коры. Получены новые данные о пространстве описания информации как отдельными нейронами, так и популяциями, включая влияние характеристик сигнала и задачи на формирование этого пространства.

Основные положения, выносимые на защиту

1. Разработанный *метод адаптивной генерации стимулов, использующий генеративно-состязательные сети в сочетании с моделями нейронных ответов нижневисочной коры,* позволяет создавать изображения, вызывающие целенаправленную активацию нейронных ансамблей, и открывает возможность исследования функциональной специализации нейронов без ограничения фиксированными наборами экспериментальных стимулов.
2. Показаны *оппонентные отношения между пространством описания сигнала и пространством постановки задачи*, определяющие структуру и динамику преобразования информации на разных этапах обработки зрительных стимулов в нейронных сетях. В частности, описание зрительной сцены на низкоуровневых этапах (начальные сверточные слои) отражает статистику входного сигнала, тогда как высокоуровневые (как сверточные, так и полносвязные слои) – адаптируются под цели и задачи деятельности наблюдателя.
3. *Преобразование зрительной информации* носит фазовый характер: описание зрительных образов, изначально подчиняющееся структуре входного сигнала, подвергается резкой перестройке и переходит в продолжительную стадию, сопровождающуюся трансформацией представлений, максимальной сложностью описания и распределенным

кодированием, после чего происходит процесс интенсивной консолидации признаков и переструктурирование описания в соответствии со стоящей перед испытуемыми задачей.

Теоретическая и практическая значимость

Теоретическая значимость работы заключается в углублении понимания нейрофизиологических механизмов кодирования зрительной информации в высших отделах зрительной системы приматов. Методологически значимым является предложенный подход к изучению функций нейронов высших областей зрительной системы, основанный на визуализации высокоуровневых репрезентаций с использованием генеративно-состязательных нейронных сетей.

Практическая значимость заключается в разработке и реализации предложенных методов в виде программных решений, применимых при проведении нейрофизиологических исследований зрительной системы. На основе данных микроэлектродной регистрации нейронов нижневисочной коры макак выполнено моделирование зрительного восприятия, включая распознавание образов и предсказание нейронных откликов с учетом семантики и контекста зрительных репрезентаций. Показана схожесть откликов искусственных нейронных сетей с характеристиками зрительного восприятия как на уровне принятия решений, так и на уровне активности кортикальных колонок нижневисочной коры. Рассмотрены особенности работы с искусственными нейронными сетями, как моделью зрительной системы приматов, выделены сходства, отличия и ограничения моделей.

Собран и систематизирован обширный материал, касающийся механизмов распознавания зрительных образов, что способствует дальнейшему изучению кодирования и интерпретации нейронной активности. Разработанные подходы позволяют углубить понимание взаимосвязи между сигналами, задачами и процессами формирования репрезентаций, а также найти новые пути для интеграции нейрофизиологических данных с методами искусственного интеллекта.

Личный вклад автора

Материалы, вошедшие в данную работу, обсуждались и публиковались автором совместно с научным руководителем. Автор непосредственно участвовала в формулировании цели и задач исследования, разработке дизайна экспериментов и выборе методов исследования; проводила обработку психофизиологических и нейрональных данных с применением методов статистического анализа, машинного обучения и искусственных нейронных сетей. Автором были разработаны и реализованы методы модельных исследований (пространство описания, прототипы категорий), а также метод создания стимульного материла при помощи генеративно-состязательных и диффузионных сетей. Экспериментальные нейрофизиологические данные

были получены методом внеклеточной микроэлектродной записи в серии экспериментов в лаборатории Манабу Танифуджи в RIKEN Brain Science Institute (Япония). Автор участвовала в части экспериментов, разработала программный комплекс для обработки нейрональной активности и создания стимульного материла, а также проводила последующий анализ полученных данных. Сотрудниками лаборатории Танифуджи также был предложен метод анализа функции нейронов при помощи фрагментов, автором был проведен проведено применение метода и моделирование ответа искусственной нейронной сети.

Апробация работы

Достоверность научных положений и выводов, полученных в данной диссертационной работе, обеспечивается результатами экспериментальных исследований, успешным представлением основных положений в докладах на ведущих международных конференциях, согласованностью результатов диссертационной работы с результатами других авторов.

Основные результаты по теме диссертации изложены в 16 печатных изданиях и 20 докладах на международных и всероссийских конференциях. Получено 2 гранта на участие в международных конференциях в США, диплом за лучший устный доклад на III Всероссийской молодежной конференции «Нейробиология интегративных функций мозга», Санкт-Петербург, 23–25 октября 2017. Автор участвовала в качестве исполнителя работ по гранту РНФ 14-15-00918, «Технологии оптимизации и восстановления когнитивных функций человека виртуальной средой», 2014–2016 гг., а также в Программе ПРАН, 2014–2017, проект «Сенсорно-моторные механизмы деятельности человека в реальном и виртуальном пространстве» под руководством Шелепина Ю.Е.

Результаты диссертационной работы доложены на 14 устных и 6 стендовых докладах следующих научно-практических конференций:

Neuroscience International Conference 2019, Chicago, USA; Vision Science Society International Conference 2019, St. Pete Beach, Florida, USA; Computational and Mathematical Models in Vision (MODVIS) Workshop at Vision Sciences Society International Conference 2018, St. Pete Beach, Florida, USA; Mutual Benefits of Cognitive and Computer Vision (MBCC) Workshop at The Computer Vision Foundation (CVPR) 2018, Salt Lake City, Utah, USA; The 5th IEEE International Conference on Video and Audio Signal Processing in the Context of Neurotechnologies (SPCN) 2020, Taoyuan, Taiwan, 2017, 2018, 2020, 2021, St. Petersburg, Russia; WiML workshop at Neural Information Processing Systems (NeurIPS) 2018, Montreal, Canada; Всероссийская конференция с международным участием Интегративная Физиология 2019, Санкт-Петербург, Россия; XXIII Съезд Физиологического общества им. И. П. Павлова, 2018, Воронеж, Россия; Международная научная конференция Прикладная Оптика 2018, Санкт-Петербург, Россия; Современные аспекты

интегративной физиологии 2018, Санкт-Петербург, Россия; Симпозиум, посвященный 100-летию Физиологического общества им. И.П. Павлова 2017, Санкт-Петербург, Россия; Технологическая перспектива в рамках Евразийского пространства 2017, Санкт-Петербург, Россия.

Публикации по теме диссертации

Основные результаты по теме диссертации изложены в 16 печатных изданиях, 7 — в периодических научных изданиях, индексируемых Scopus и WoS, 7 — в тезисах докладов всероссийских и международных конференций, 2 — в других печатных изданиях.

Статьи в изданиях, индексируемых в WoS, Scopus:

1. Вахрамеева О. А., Хараузов А. К., Пронин С. В., Малахова Е. Ю., Шелепин Ю. Е. Зрительный прайминг при распознавании мелких изображений в сцене содержащей объекты разного размера // Физиология человека, 2016, том 42, № 5, с. 39–48
2. Малахова Е. Ю. Визуализация информации, кодируемой нейронами высших областей зрительной системы // Оптический журнал, август 2018 г., стр. 61–66.
3. Жукова О. В., Малахова Е. Ю., Шелепин Ю. Е. Джоконда и неопределенность распознавания улыбки человеком и искусственной нейронной сетью. // Оптический журнал. 2019. № 11.
4. Малахова Е. Ю. Пространство описания зрительной сцены в искусственных и биологических нейронных сетях. // Оптический журнал. 2020. № 10.
5. Nam, Y., Sato, T., Uchida, G. Malakhova, E., Ullman, Sh., Tanifuji, M. View-tuned and view-invariant face encoding in IT cortex is explained by selected natural image fragments. Scientific Reports. No. 11, 7827 (2021).
6. Малахова Е. Ю. Представление категорий посредством прототипов согласованной активности нейронов в сверточных нейронных сетях. // Оптический журнал. 2021. № 12.
7. Малахова К. Ю., Шелепин К.Ю., Шелепин Ю.Е. Обнаружение и распознавание изображений в условиях помехи // Оптический журнал. 2024. Т. 91. № 8. С. 60–74.

Тезисы:

1. Малахова Е. Ю., Жукова О. В., Шелепин Ю. Е. Джоконда - неопределенность восприятия улыбки человеком и искусственной нейронной сетью // Тез. докл. Международной научной конференции «Технологическая перспектива в рамках Евразийского пространства: новые рынки и точки экономического роста». Санкт-Петербург, 26–28 октября 2017. - С. 309–312.

2. Малахова Е. Ю. Методы визуализации репрезентации информации в искусственных и биологических нейронных сетях // Тез. докл. Международной научной конференции «Технологическая перспектива в рамках Евразийского пространства: новые рынки и точки экономического роста». Санкт-Петербург, 26–28 октября 2017. - С. 321–324.
3. Малахова Е. Ю. Представление информации в нейронных сетях при распознавании семантики изображений // Тез. докл. XXIII съезда Физиологического общества им. И. П. Павлова. Воронеж, 18–22 сентября 2017. - С. 1642–1644.
4. Жукова О. В., Малахова Е. Ю. Разработка искусственной нейронной сети глубокого обучения при распознавании лиц в условиях неопределенности // Тез. симпозиума, посвященного 100-летию Физиологического общества им. И. П. Павлова. Санкт-Петербург, 17–19 апреля 2017. - С. 37–38.
5. Малахова Е. Ю. Моделирование и анализ сверточных нейронных сетей для задачи детекции текста на естественных изображениях // Тез. симпозиума, посвященного 100-летию Физиологического общества им. И. П. Павлова. Санкт-Петербург, 17–19 апреля 2017. - С. 63–64.
6. Малахова Е. Ю. Формирование зрительных категорий в нейронных сетях // Материалы Всероссийской молодежной конференции с международным участием «СОВРЕМЕННЫЕ АСПЕКТЫ ИНТЕГРАТИВНОЙ ФИЗИОЛОГИИ». Санкт-Петербург, 9–11 октября 2018. - С. 124.
7. Малахова К. Representation of categories in filters of deep neural networks // In Proceedings: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. 2018. - С. 1973–1975.

Другие печатные издания:

8. Коллективная монография: под ред. Шелепина Ю. Е., Чихмана В. Н. Нейротехнологии, Глава 13: Обработка зрительной информации в искусственных и биологических нейронных сетях // СПб.: Изд-во ВВМ, 2018. С. 336–346.
9. Коллективная монография: под ред. Шелепина Ю. Е., Огородниковой Е. А., Соловьёва Н. А., Якимовой Е. Г. Neural Networks and Neurotechnologies, Глава: Gioconda’s Smile — from biological to artificial neural networks // СПб.: Изд-во ВВМ, 2019. С. 204–210.

Глава 1. Исследование процесса распознавания зрительных образов высшими отделами зрительной системы приматов

В **Главе 1** представлен обзор литературы, современных задач и подходов в исследовании механизмов распознавания зрительных образов, выполняемых высшими отделами зрительной системы приматов. Основное внимание уделено следующим аспектам:

- структурно-функциональной организации зрительной системы, обеспечивающей распознавание образов, иерархическим принципам обработки информации в вентральном зрительном пути;
- широко используемым методам анализа функциональных характеристик нейронов, таким как индексы чувствительности и селективности, построение кривых настройки и др;
- ограничению классических методов при изучении высокоуровневой нейронной активности и перспективности использования методов искусственного интеллекта (включая сверточные нейронные сети) для моделирования и интерпретации данных.

Также описываются способы интеграции экспериментальной нейрофизиологии с вычислительным моделированием, что позволяет создавать репрезентативные модели процессов зрительного восприятия и исследовать представление информации в моделях.

1.1. Структурно-функциональная организация зрительной системы, обеспечивающая распознавание образов

Зрительная система отвечает за обработку информации, поступающей со зрительных анализаторов, и играет решающую роль в нашей способности распознавать и интерпретировать окружающий мир.

Основы нейрофизиологических механизмов организации нейронных сетей изучались еще в начале прошлого века Б. П. Бабкиным, который разработал концепцию о «временной связи» между нейронами коры головного мозга (Бабкин, 1904). В дальнейшем идеи нейропластичности и обучения были развиты Д. О. Хеббом, который сформулировал принцип долговременного усиления синапсов в процессе обучения, известный как «правило Хебба» (Hebb, 1949).

Развитие этих идей привело к детальному изучению сенсорных систем, включая зрительную. Исследованы размер и характеристики рецептивных поля различных областей зрительной системы, Хьюбел Д. и Визель Т (Hubel, Wiesel, 1962). обнаружили, что нейроны стриарной коры

(V1) избирательно реагируют на специфические характеристики зрительных стимулов, такие как ориентация линий, и классифицировали их на простые и сложные клетки, заложив фундамент для понимания иерархической обработки информации в зрительной системе (Hubel, Wiesel, 1962). Показано увеличение размера рецептивного поля в других областях по сравнению со стриарной корой и их перекрытие всей области центральной ямки поля зрения (Шелепин, 1982).

Изучение структуры нейронных сетей в зрительной системе позволило установить связь между характеристиками зрительной системы и статистическими свойствами окружающего мира (Глезер, Цуккерман, 1961; Цуккерман, 1975; Глезер, 1985; Field, 1989, 1999). В ходе исследований были описаны функциональные единицы зрительной коры, формирующие упорядоченные структуры с определенными пространственно-частотными характеристиками, описана их связь с воспринимаемыми изображениями.

Один из ранних подходов к изучению нейронной активности заключался фокусировался на частоте разрядки отдельных нейронов (Adrian, 1928). Более поздние исследования расширили эту идею, предположив, что информация может быть закодирована во временных паттернах спайков внутри серий импульсов, генерируемых отдельными нейронами (Подвигин, 1979; Softky, Koch, 1993), или в группах нейронов (Averbeck, Latham, Pouget, 2006). Показана временная устойчивость репрезентаций, кодируемых нейронами (Bondar et al., 2009; McMahon et al., 2014).

Понимание структурной и функциональной организации зрительного пути имеет важное значение как для фундаментальной науки, так и для клинического применения и может помочь в разработке целевых вмешательств и диагностических инструментов для этих состояний. Например, спектральный подход к моделированию механизмов зрительного восприятия (Campbell, Robson, 1968) сделал возможным разработку нового метода для исследования нейрофизиологических механизмов зрительного восприятия человека в области клиники и эргономики – визоконтрастометрии (Шелепин, Колесникова, Левкович, 1985).

Известно, что в процессе восприятия и распознавания объектов участвуют структуры вентрального пути зрительной системы (Кок, 1967; Шелепин, 1973; Mishkin, Ungerleider, 1982; Goodale, Milner, 1992). Вентральный путь представляет собой ряд взаимосвязанных областей мозга, которые располагаются от первичной зрительной коры (V1) до нижневисочной коры (ITC). Он включает в себя области V2, V4, а также заднюю и переднюю области ITC, обычно называемые задней нижневисочной корой (PIT) и передней нижневисочной корой (AIT), соответственно. Вентральный поток в основном участвует в обработке информации, связанной с характеристиками объекта, такими как форма, цвет и текстура (Кезели, 1983; Lee, 1991; Tanaka, 1996).

Зрительная информация в вентральном пути обрабатывается иерархически, при этом на каждом последующем этапе из входного сигнала извлекаются более сложные характеристики (Felleman, Van Essen, 1991). Такая иерархическая организация позволяет вентральному потоку постепенно преобразовывать необработанную зрительную информацию в высокоуровневое представление, которое может быть использовано для восприятия и распознавания объектов (Riesenhuber, Poggio, 2000; DiCarlo, Zoccolan, Rust, 2012). Например, нейроны в V1 чувствительны к простым признакам, таким как ориентация и пространственная частота, а нейроны в более высоких областях, таких как V4 и IT, реагируют на более сложные признаки, такие как кривизна и части объекта (Pasupathy, Connor, 2001; Kiani et al., 2007). Недавние исследования также подчеркнули роль рекуррентных связей в вентральном потоке, которые могут поддерживать интеграцию информации на разных этапах обработки и способствовать распознаванию объектов (Lamme, Roelfsema, 2000; Kravitz et al., 2013).

Широкое развитие получила область вычислительных нейронаук, позволяющая сопоставить результаты нейрофизиологических исследований зрения с математическим моделированием. Красильниковым Н. Н. был проведён ряд работ о механизмах согласованной фильтрации при распознавании образов (Красильников, 1986, 2001, 2011). Этот метод позволил создать количественные модели зрительного восприятия, основанные на известных психофизических механизмах, где центральным элементом выступает согласованный фильтр, сформированный на основе заранее подготовленных шаблонов. Данная концепция прослеживается и в более поздних исследованиях, сопоставляющих искусственные нейронные сети и механизмы человеческого восприятия (Жеребко, Луцив, 1999; Малахова, 2018, 2020, 2021; Bashivan, Kar, DiCarlo, 2019; Ponce et al., 2019).

Интеграция экспериментальной нейронауки с вычислительными методами и машинным обучением (Hassabis et al., 2017) открывает новые перспективы в изучении процессов распознавания объектов. В частности, сверточные фильтры нейронных сетей (Жеребко, Луцив, 1999; Луцив, 2011), возникающие при обработке больших обучающих выборок, можно рассматривать как аналог согласованных фильтров. Искусственные нейронные сети зарекомендовали себя как важный инструмент для моделирования и интерпретации сложных паттернов нейронной активности.

Современные исследования в области глубокого обучения позволили создать модели, воспроизводящие нейрофизиологические механизмы восприятия и распознавания образов. Они не только способствуют изучению работы зрительной системы, но и находят применение в разработке новых алгоритмов обработки изображений. В данной работе рассматриваются экспериментальные и теоретические подходы к декодированию нейрональной активности, включая анализ кодирования информации как на уровне отдельного нейрона, так и популяций.

Особое внимание уделяется использованию методов искусственного интеллекта и машинного обучения для моделирования и интерпретации этих процессов.

Распознавание образов. Под распознаванием образов понимается способность присваивать метки (например, имя существительное) предъявленным стимулам, начиная от конкретных названий в задаче идентификации и заканчивая общими в задаче категоризации (DiCarlo, Zoccolan, Rust, 2012). Важным свойством данного процесса является способность выполнять такие задачи при условии преобразований, сохраняющих идентичность объекта (например, при изменении его положения, размера, позы и фонового контекста), и без дополнительной контекстуальной информации, специфической для объекта или места (например, автомобиль должен распознаваться, даже если предъявлен без привычного контекста – дороги). Это свойство инвариантности восприятия к трансформациям объекта базируется на сложных нейрофизиологических механизмах обработки зрительной информации.

Был внесен значительный вклад в исследование механизмов инвариантности зрительного восприятия как ключевого принципа распознавания объектов (Глезер, 1966; Шелепин, 1973; Яковлев, 1982; Шелепин, Чихман, Фореман, 2008). Установлено, что отдельные нейроны нижневисочной коры преимущественно реагируют на сложные стимулы, такие как изображения рук и лиц (Gross, Rocha-Miranda, Bender, 1972). В ряде исследований показано, что при распознавании объектов в коре формируются представления, чувствительные к углу обзора, но инвариантные к другим характеристикам (Logothetis, Sheinberg, 1996). Предполагается, что инвариантность достигается за счет непрерывного кодирования признаков соседствующими нейрональными колонками (Tanaka, 1996), а кодирование объектов может осуществляться через различные комбинации активных и неактивных колонок, соответствующих отдельным признакам (Tsunoda et al., 2001) и др.

Процесс распознавания образов на отдельных изображениях, расположенных в центре зрительного поля, и поведенческая реакция на них составляет ~250 мс у обезьян (Fabre-Thorpe, Richard, Thorpe, 1998) и ~350 мс у людей (Thorpe, Fize, Marlot, 1996; Rousselet, Fabre-Thorpe, Thorpe, 2002). При этом в наборе стимулов изображения могут предъявляться последовательно со скоростью менее ~100 мс на изображение. Таким образом, зрительный образ проходит путь по зрительной системе менее чем за 200 мс (DiCarlo, Zoccolan, Rust, 2012). При этом сперва происходит определение присутствия объекта, а затем его узнавание (Liu, Harris, Kanwisher, 2002). Например, метод вызванных потенциалов в электроэнцефалограмме позволяет выделить компонент N170 (Bentin et al., 1996), который представляет собой отрицательное отклонение, возникающее примерно через 170 мс после просмотра изображения лиц. Однако обнаружение

лица на стимуле может быть зарегистрировано всего через 100 мс после предъявления (Liu, Harris, Kanwisher, 2002).

Моделирование и декодирование нейронной активности. Методы машинного обучения применяются для декодирования нейронной активности и прогнозирования поведенческих результатов (Naselaris et al., 2011). Эти подходы продемонстрировали способность предсказывать различные когнитивные состояния и поведенческие реакции, такие как зрительное восприятие (Kamitani, Tong, 2005), моторное планирование (Cunningham et al., 2008) и вспоминание (Rutishauser, Mamelak, Schuman, 2006).

Искусственные нейронные сети все чаще используются для моделирования и понимания нейронной активности благодаря своей способности улавливать сложные закономерности и нелинейные взаимосвязи. Они используются в различных контекстах, включая моделирование свойств ответа отдельных нейронов (Carandini, Heeger, 2011), характеристику паттернов активности на уровне популяции (Cadieu et al., 2014) и понимание вычислительных принципов, лежащих в основе нейронной обработки (Marblestone, Wayne, Kording, 2016). Отдельное внимание в работе уделяется сверточным нейронным сетям, более подробно рассмотренных в последующих разделах.

Модели, такие как искусственные нейронные сети, по своей природе являются упрощенным представлением объектов или процессов, включая мозговую активность. Их ключевая ценность заключается в том, насколько точно они соответствуют исследуемому процессу в пределах выбранной области моделирования. Допустимо не охватывать другие аспекты феномена, но важно понимать, что выводы, сделанные на ее основе, будут ограничены областью ее применимости.

С научной точки зрения, использование моделей неизбежно связано с упрощениями, которые делают сложные явления управляемыми. Однако их фундаментальным аспектом должна оставаться фальсифицируемость, то есть возможность проверки и потенциального опровержения (Поппер, 2010). Это означает, что модели не должны стремиться идеально отразить реальность, а должны быть полезным инструментом для решения конкретных задач и проходить проверку путем сопоставления с экспериментальными данными. В таком случае ограниченность моделей становится их функциональным аспектом, а не недостатком.

Несмотря на значительный прогресс в моделировании и понимании процессов, происходящих в мозге приматов, остаются значительные сферы для улучшения. Сложность биологических нейронных сетей и ограничения современных методов записи затрудняют полный охват нейронной активности. Кроме того, разработка новых, более сложных и биологически

правдоподобных алгоритмов машинного обучения и ИНС может дать дальнейшее понимание принципов, лежащих в основе кодирования информации в мозге (Kriegeskorte, Douglas, 2018).

Будущие исследования могут быть направлены на разработку новых экспериментальных парадигм и методов записи, которые позволят изучать нейронную активность с более высоким разрешением и на больших популяциях нейронов. Кроме того, интеграция различных научных дисциплин, таких как вычислительная нейронаука, машинное обучение и экспериментальные исследования, имеет решающее значение для понимания кодирования информации в головной мозге приматов.

1.2. Подходы к интерпретации нейрональной активности. Исследование кодирования информации в мозге при помощи искусственных нейронных сетей

Понимание того, как информация кодируется и обрабатывается в мозге, является центральным вопросом нейронауки. За прошедшие годы были разработаны различные методы изучения взаимосвязи регистрируемой нейрональной активности и процесса восприятия и обработки информации, начиная от микроэлектродной одноклеточной записи и заканчивая крупномасштабными исследованиями нейронных популяций. С развитием искусственного интеллекта эти методы дополняются вычислительными подходами, позволяющими глубже анализировать нейрональные данные.

Можно выделить два направления применения методов искусственного интеллекта в исследованиях нейрофизиологии зрительной системы:

- улучшение экспериментального процесса;
- интерпретация функции нейронов.

Регистрация ответа нейронов. Нейрональный ответ обычно характеризуется количеством и динамикой импульсации, производимой клеткой. Для вычисления реакции нейрона на стимул учитывается набор факторов, часть из которых приводится ниже:

- *Время от подачи стимула до отклика регистрируемого нейрона (латентное время отклика нейрона).* Для определения интервала, в котором активность нейрона будет рассматриваться как ответ на стимул, необходимо учитывать время распространения активации по нервной системе.
- *Фоновая активность,* или частота импульсации нейрона в состоянии покоя. Измеряется как количество спайков в интервал времени до проведения стимуляции.

- *Вариативность в ответе при повторном предъявлении* (для оценки стабильности ответа нейрона). Так, например, отсутствие корреляции в ответах на один и тот же стимул говорит о невозможности сопоставления стимула и реакции.
- *Адаптация* – снижение количества спайков при многократном предъявлении идентичных стимулов.
- *Состояние регистрируемых клеток,* отслеживаемое по динамике фоновой активности.

Значительные изменения в фоновой активности во время эксперимента могут говорить о возможном перераспределении активности в нейронной сети, деградации функции клетки и низкой достоверности полученных данных. Также должно учитываться состояние электрода, места, глубины и угла введения, состояние тканей, наличие клеток в непосредственной близости к электроду.

Анализ активности отдельной клетки проводится путем соотнесения частоты возникновения потенциалов действия (импульсов, или спайков) на единицу времени в условиях предъявления стимула и при отсутствии стимуляции. Значимым показателем считается наличие потенциала действия; его амплитуда, продолжительность и форма при этом не игнорируются как несущественные.

За меру активности нейрона принимается количество *импульсов за определенный период времени* (чаще всего спайки/с). Полученное значение *сравнивается с фоновой активностью* нейрона, измеренной в период до поступления сигнала в область регистрации. Если активность во время предъявления стимула была выше фоновой, то считается, что данный стимул возбуждает нейрон, если количество импульсов, наоборот, снижается, то такой стимул считается ингибирующим. Рассчитанный таким образом ответ может быть негативным, так как стимул (либо категория стимулов) снижает количество импульсов относительно фоновой активности нейрона:

$$r = r(t) - r(t'),$$

где $r(t)$ – обозначает активность нейрона в ответ на предъявление сигнала в промежутке времени t, $r(t')$ – фоновая активность в промежутке времени t'.

Таким образом, в качестве меры активности нейрона – нейронального ответа r рассматривается количество импульсов за определенный интервал между временем t и t_i:

$$r(t) = \sum_{i=1}^{n} h\,(t - t_i).$$

Здесь $h(t)$ представляет собой функцию ядра (kernel function), которая определяет вклад каждого отдельного спайка, произошедшего в момент t_i, в общее нейронное возбуждение $r(t)$. В зависимости от выбора $h(t)$, функция может представлять дельта-импульсы (в случае

дискретного представления), гауссовы функции (для сглаженного представления активности) или экспоненциальный спад (для учета постсинаптического потенциала). Выбор $h(t)$ влияет на то, насколько точно и гладко будет представлена активность нейрона во времени.

Во время регистрации состояние клетки постоянно меняется, так же, как и состояние локального окружения и крупномасштабных сетей, в которые эта клетка входит. Это приводит к вариабельности ответа, когда один и тот же стимул провоцирует различное количество потенциалов действия, и, в свою очередь, делает нецелесообразной интерпретацию поведения нейрона по одному измерению. Тем не менее реакция на стимул может быть аппроксимирована средним количеством импульсов – при условии проведения достаточного количества предъявлений на достаточно малом отрезке времени регистрации (Dayan, Abbott, 2001), аналогично методу синхронных накоплений при регистрации вызванных потенциалов в электроэнцефалограмме (Brazier, 1948). За счет многократной подачи стимулов сигнал, связанный со стимулом, растет значительно быстрее, чем шум спонтанной активности, попадающий в случайную фазу. Таким образом усиливается выделение сигнала из шума.

В неокортексе нейроны различных слоев могут иметь схожий ответ и по данному функциональному признаку часто рассматриваются как одна базовая вычислительная единица – модуль (Glezer, 1995), соответствующий нейрональной колонке (Mountcastle, 1957). В случае, когда нейрональная колонка выступает в роли единицы анализа, нейрональная активность усредняется в пределах отдельной колонки.

Методы исследования функциональных характеристик нейронов. Исследование функциональной роли нейрона или популяции клеток проводится посредством сопоставления характеристик входного сигнала с нейрональным ответом. Ранние исследования основывались на использовании микроэлектродов для регистрации активности одиночных нейронов (single-unit recordings). Этот метод позволил установить, что нейроны зрительной коры кодируют информацию через специфические пространственно-временные паттерны возбуждения. Например, работы (Hubel, Wiesel, 1962) продемонстрировали существование простых и сложных клеток в первичной зрительной коре (V1), реагирующих на определенные ориентации и движения световых полос.

Впоследствии, с развитием многоканальной регистрации, стало возможным регистрировать активность сразу нескольких нейронов (multi-unit recordings), что позволило исследовать кодирование информации на уровне нейронных популяций.

Современные технологии, такие как оптическая кальциевая визуализация (two-photon imaging) и электрофизиологические массивы (Neuropixels), позволяют регистрировать

активность сотен и тысяч нейронов одновременно. Это расширило наше понимание нейронных кодов, используемых зрительной системой.

Таким образом, расширение методов регистрации нейрональной активности открыло новые возможности для исследования зрительной системы, однако возросший объем данных потребовал более сложных подходов к их обработке и интерпретации. В этом контексте методы искусственного интеллекта стали мощным инструментом для анализа большого количества данных и моделирования нелинейных процессов.

Методы искусственного интеллекта, особенно глубокие нейросети, применяются для обработки многомерных данных, выявления скрытых корреляций в ответах нейронов и построения предсказательных моделей реакции популяций нейронов на зрительные стимулы. Например, сверточные нейросети используются для моделирования процессов в зрительной системе, поскольку их архитектура напоминает организацию слоев зрительной коры. На Рис. 1 приведен пример применения сверточных сетей для моделирования различных аспектов зрительного восприятия.

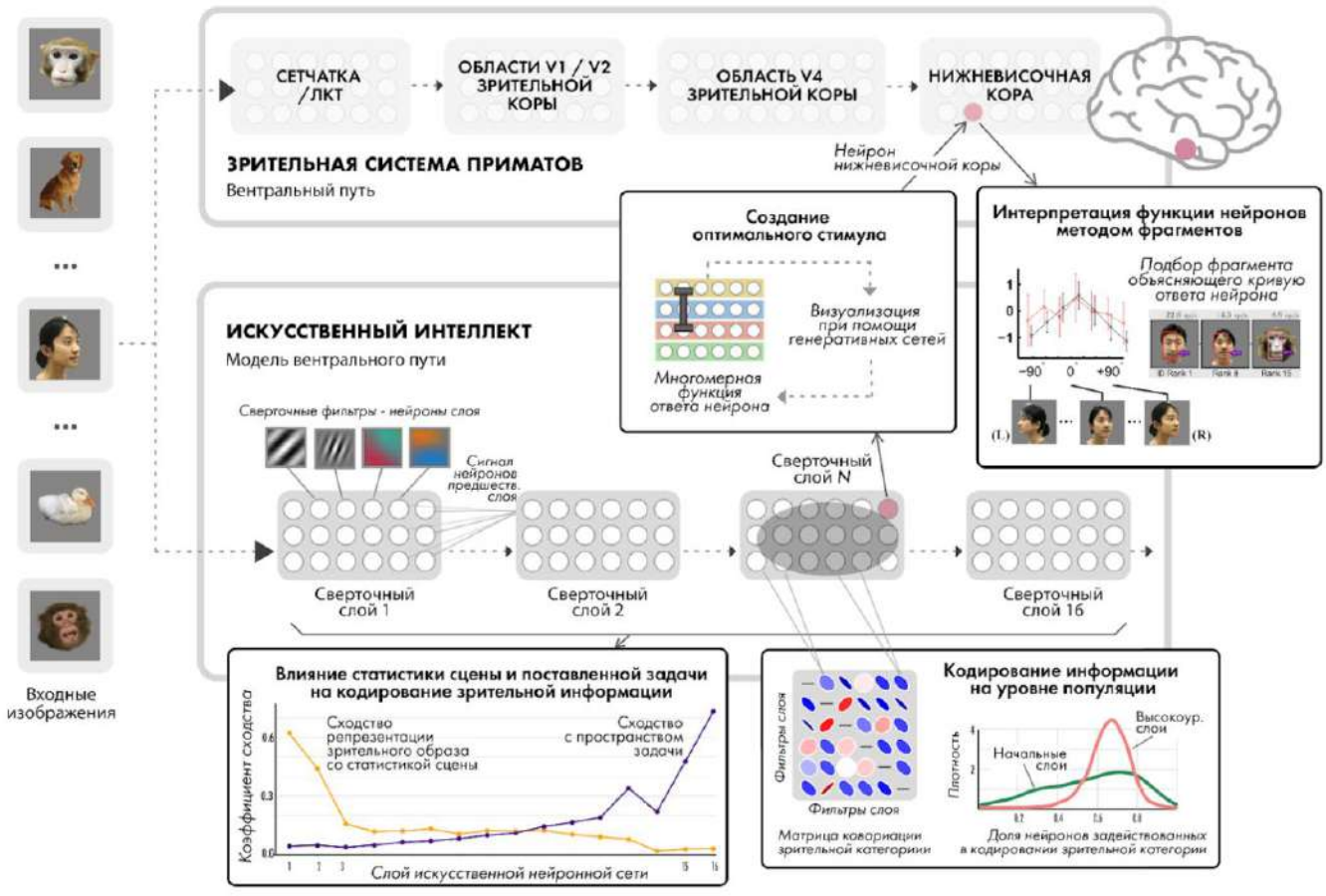

Рисунок 1. Исследование процесса распознавания образов в зрительной системе и искусственных нейронных сетях

Изучение ответа нейронов с целью описания их функциональных свойств во многом опирается на классические методы анализа ответа нейронов. В этом случае отдельный нейрон (либо же группа нейронов) рассматривается как детектор и проводится оценка его реакции на присутствие признака. Рассмотрим широко распространенные способы охарактеризовать функцию нейронов.

Индекс чувствительности. Применяется для определения способности детектора к различению сигнала, содержащего определенную информацию, и фонового шума. Эффективность детектора в указанной задаче разделения рассчитывается по формуле:

$$d' = \frac{\mu_S - \mu_N}{\sqrt{\frac{1}{2}(\sigma_S^2 + \sigma_N^2)}},$$

где μ_S– средний ответ; σ_S^2 – среднее квадратичное отклонение нейрона на предъявление сигнала; аналогично μ_N, σ_N^2 для шума. При этом предполагается, что распределения целевого сигнала и шума подчиняются нормальному закону и имеют равные стандартные отклонения. Высокое значение индекса говорит о лучшей различимости сигнала. Индекс чувствительности широко используется для выделения нейронов, реагирующих на определенную категорию (Aparicio, Issa, DiCarlo, 2016; Issa, DiCarlo, 2012; Ohayon, Freiwald, Tsao, 2012).

Метрика избирательности (селективности). Другой распространенной характеристикой функции нейрона является его избирательность по отношению к определенной категории (Aparicio, Issa, DiCarlo, 2016; Bell et al., 2011; Freiwald, Tsao, Livingstone, 2009; Tsao et al., 2006). Для категории S_C она рассчитывается следующим образом:

$$S_C = \frac{\mu_C - \mu_{NC}}{\mu_C + \mu_{NC}},$$

где μ_C – средний ответ нейрона для категории; μ_{NC} – средний ответ вне категории. Нейрон считается селективным к категории, если $\mu_C > 0$, а $\mu_{NC} < 0$ либо если $\mu_C < 0$ и $\mu_{NC} < 0$, но при этом активность подавляется в меньшей степени при предъявлении изображений категории, чем вне ее. Для расчета относительной избирательности к нескольким классам указанная формула применяется поочередно к каждой категории, все остальные стимулы, при этом, рассматриваются как экземпляры вне категории.

Существуют также вариации оценки избирательности. Например, для оценки предпочитаемой ориентации линии у нейронов первичной зрительной коры используется следующая формула:

$$K = \frac{\sqrt{(\sum_i R(\theta_i)\sin\alpha)^2 + (\sum_i R(\theta_i)\cos\alpha)^2}}{\sum_i R(\theta_i)},$$

где $R(\theta_i)$ обозначает ответ (спайков/сек) для i-го стимула с ориентацией θ_i, для стимула в виде полосы $\alpha = 2\theta_i$, для градиента $\alpha = \theta_i$. Результирующий вектор показывает предпочитаемую ориентацию стимула. Данный коэффициент принимает значения от 0 до 1. K = 0 означает, что

нейрон отвечает одинаково на все ориентации и K = 1 означает, что нейрон отвечает только на одну ориентацию. Наличие чувствительности к ориентации стимула определяется как K ≥ 0,1 (Batschelet, 1981; Levick, Thibos, 1982; Naito et al., 2013; Suematsu, Naito, Sato, 2012; Sun et al., 2004).

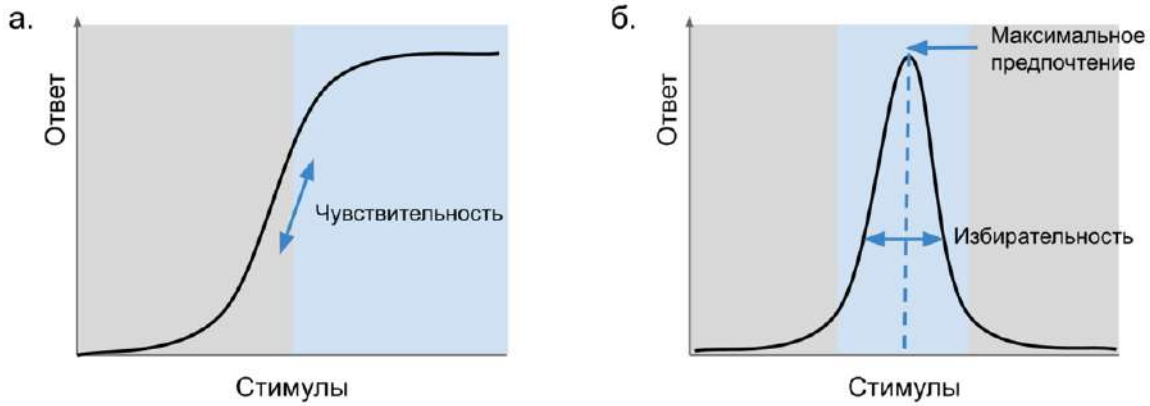

Рисунок 2. Схематическая иллюстрация метрик чувствительности (а) и избирательности (б) по отношению к предъявляемым стимулам.

Чувствительность и избирательность можно условно изобразить в виде функций (Рис. 2). При измерении чувствительности, полученное значение позволяет оценить разделимость классов, однако не говорит об избирательности детектора к проверяемому классу.

Кривая избирательности (tuning curve). Является расширенным аналогом метрики избирательности и используется для анализа и визуализации свойств нейрона. Для ее построения предъявляется набор стимулов и регистрируется частота вызванных импульсов. Средняя активация, выраженная как функция параметра, задающего стимул, и называется кривой избирательности (настройки).

Следует учитывать, что форма кривой при оценке избирательности может иметь различную форму, например, быть похожей на распределение Гаусса, бимодальное, трапезоидное, геометрическое и т. д. Форма кривой зависит от выбранного параметра и его влияния на нейрональный ответ. Например, клетки первичной зрительной коры реагируют на ориентацию стимула. В таком случае параметром является угол наклона линии.

Было показано, что для ряда первичных кортикальных областей сенсорных систем кривая настройки позволяет подобрать простую функцию, описывающую поведение нейрона. В частности, ответ клеток моторной коры хорошо аппроксимируется косинусоидой или другой гладкой функцией (Doya et al., 2011).

Нейроны высших областей реагируют на более сложные стимулы, что делает невозможной их интерпретацию через простые функции, как в случае с ориентационной избирательностью.

Однако кривая настройки дает возможность выразить нейрональный ответ через ранжирование стимулов, обращаясь таким образом к избирательности нейрона.

Сопоставление метрик избирательности к лицам. В данном разделе рассматривается два подхода к оценке избирательности, используемые в дальнейшем для идентификации искусственных нейронов, избирательных к изображениям лиц, аналогично нейронам нижневисочной коры.

Первый подход, Face Selectivity Index (FSI), основан на вычислении разницы средних значений активаций нейронов при предъявлении изображений лиц и других объектов:

$$\mathrm{FSI} = \frac{(\mathrm{MeanFaces} - \mathrm{MeanNonFaces})}{(\mathrm{MeanFaces} + \mathrm{MeanNonFaces})},$$

где *MeanFaces* – средняя активация нейрона на изображения с лицами, а *MeanNonFaces* – средняя активация на изображения без лиц. Эта метрика определяет относительное различие между средними значениями, нормализуя разницу на их сумму, что позволяет учесть абсолютный уровень активации нейрона. Такой подход применяется в нейрофизиологических исследованиях (см. *метрика избирательности* в разделе 1.2) и позволяет определить, насколько данный нейрон реагирует на лица по сравнению с другими объектами.

Второй подход, Category Selectivity Index (CS), сначала нормализует все активации, приводя их к единому масштабу относительно минимального и максимального значений в выборке, а затем вычисляет разницу средних нормализованных активаций между категориями:

$$CS = \frac{\sum Response_{norm}Face}{n_{Face}} - \frac{\sum Response_{norm}NonFace}{n_{NonFace}},$$

Здесь $Response_{norm}$ – это нормализованное значение активации нейрона, приводимое к диапазону от 0 до 1 по минимальному и максимальному значению в данных. Этот метод позволяет учесть вариативность активации нейронов и сделать метрику более устойчивой к абсолютным уровням их ответа.

Для визуализации и сравнения двух метрик – индекса избирательности к категории (CS) и индекса избирательности к лицам (FSI) – была проведена симуляция, в рамках которой генерировались два распределения, имитирующие ответ нейрона на две различных категории. Исследовались их свойства при различных значениях математического ожидания () и стандартного отклонения (). В ходе эксперимента параметры изменялись в диапазоне: от 0 до 10, с шагом 1; от 0 до 1, с шагом 0,1.

На Рис. 3 представлены примеры с разными параметрами распределений, где верхние графики показывают временные ряды активаций, а нижние – гистограммы распределения значений. Анализ результатов показывает, что FSI реагирует на различие между категориями даже в случаях, когда распределения сильно перекрываются, тогда как CS остается низким при

пересечении распределений и возрастает только при их четком разделении. Это говорит о том, что FSI больше ориентирован на разницу средних значений, независимо от степени пересечения двух категорий, а CS учитывает не только средние значения, но и распределение активаций. В ситуациях, когда одна категория отличается от другой только сдвигом среднего, обе метрики работают схоже, но если категории различаются за счет дисперсии или частичного перекрытия, FSI может давать высокое значение, даже если значительная часть выборок совпадает, тогда как CS более консервативен и чувствителен к степени реального разделения данных. Эти результаты помогают понять, в каких условиях каждая из метрик лучше отражает избирательность нейронов к лицам.

Таким образом, при выборе метрики для анализа нейрофизиологических данных следует руководствоваться целью исследования. Например, FSI удобен для количественной оценки, насколько в среднем нейрон реагирует на лица по сравнению с другими стимулами, даже если распределения активаций частично перекрываются. Это полезно, например, при составлении «кривой настройки» нейрона, где важно определить, как изменяется его отклик в зависимости от свойств стимула. CS, напротив, чувствителен к степени разделения распределений и позволяет выявлять нейроны, которые стабильно активируются только при предъявлении лиц, но не реагируют на другие категории. В этом смысле различие между FSI и CS схоже с подходами, использованными в разделе работы 3.2, где стимулы создавались с помощью двух разных функций потерь: *Evoked*, максимизирующей ответ целевого нейрона независимо от других, и *Softmax*, фокусирующейся на выборочной активации одного нейрона с минимизацией отклика остальных.

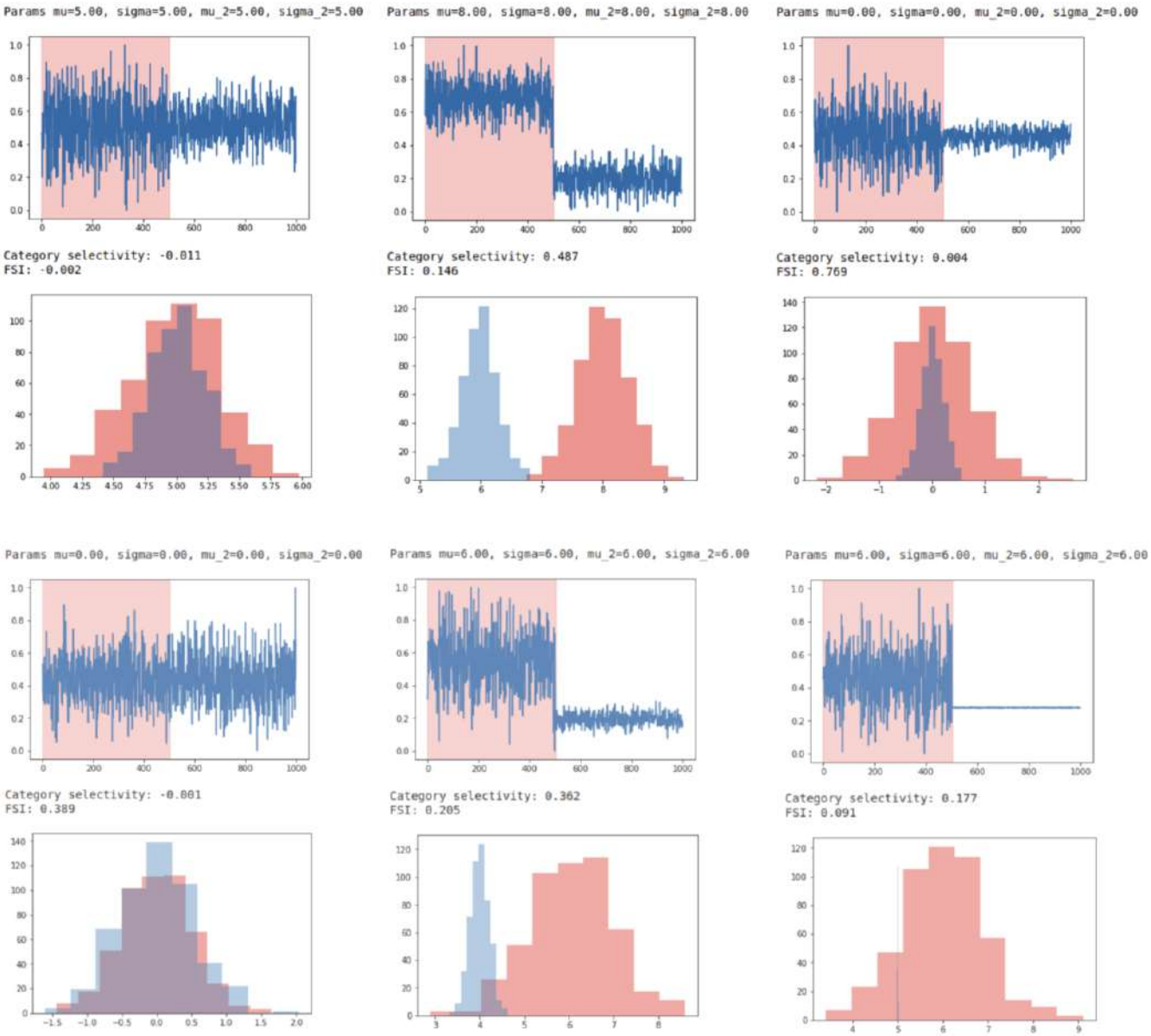

Рисунок 3. Симуляция ответа нейрона на две категории при различных параметрах μ и σ. Верхние графики показывают динамику активации нейронов в зависимости от предъявляемой категории, где красным красным цветом обозначена категория А. Нижние графики представляют гистограммы распределения активаций для каждой категории (А – красным, Б – синим). Приведены значения метрик Category Selectivity (CS) и Face Selectivity Index (FSI), отражающие степень различимости двух распределений

Вышеприведенные метрики позволяют провести анализ характеристик отдельных функциональных единиц. Тем не менее, для понимания того, каким образом происходит распознавание и узнавание объектов, недостаточно изучения функции отдельных нейронов. Представление категории элементов в отделах зрительной системы называется репрезентацией и является отдельным направлением исследований.

Проведение нейрофизиологического эксперимента для изучения зрительной системы. Начиная с первичных отделов зрительной коры повышается размер зрительного поля, охватываемого отдельными нейронами (Рис. 4). Если функция нейронов первичной зрительной коры может быть описана через чувствительность и избирательность к простейшим стимулам, таким как линия определенной наклонности, то последующие области агрегируют информацию и отвечают на все более сложные стимулы – формы, текстуры, объекты, достигая инвариантного представления категорий объектов в нижневисочной коре.

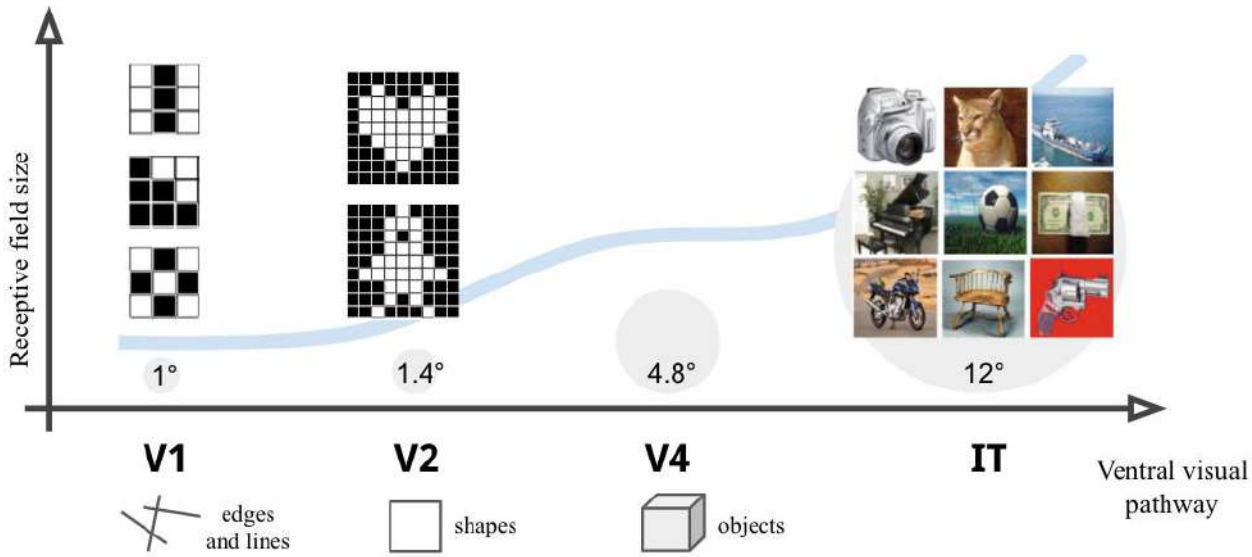

Рисунок 4. Размер зрительного поля, охватываемого нейронами областей зрительной системы, и повышение сложности исследования функции нейронов

Основная парадигма исследований, принятая в нейронауке, включает следующие этапы: подготовка и предъявление стимулов, регистрация ответа, анализ данных и объяснение функции области или нейрона. Этот подход хорошо работает на ранних стадиях зрительного восприятия, где рецептивные поля клеток являются небольшими.

В случае если задачей исследования является поиск уникального стимула, вызывающего максимально возможную активацию нейрона, то необходимо принимать во внимание количество возможных итераций – предъявлений стимулов испытуемому. Предположим, что мы работаем с черно-белыми изображениями размером 3x3 пикселя и пытаемся найти идеальный стимул для нейрона в области V1. Поскольку изображения черно-белые, то каждый пиксель может иметь одно из двух состояний. Общее количество возможных стимулов для изображения размером 3x3 будет: $2^{(3x3)} = 2^9 = 512$. Однако если размер изображения увеличивается до 6x6 пикселей, то общее количество возможных стимулов уже составляет: $2^{(6x6)} = 2^{36} \approx 68{,}7$ млрд. Для черно-белого изображения 60x60 количество комбинаций составляет $2^{(60x60)}$ – чрезвычайно большое число, невозможное для простого перебора. Указанные расчеты предполагают двоичное представление

для каждого пикселя (черный или белый) и не учитывают ряд важных факторов, таких как расположение стимула относительно фовеа, чувствительность зрительной системы к статистике натуральных сцен и др.

Приведенные расчеты являются условными, но дают представление о сложности задачи исследования в нейрофизиологии высших отделов зрительной коры и подтверждают необходимость поиска более эффективных подходов.

Вычислительные модели распознавания объектов в вентральном зрительном пути

Интерпретация нейрофизиологических данных является важным этапом в исследованиях мозга. Выше приведены классические подходы к интерпретации, включающие различные метрики, а также корреляционный анализ, который используется для определения степени связи между активностью нейронов и предъявленными стимулами. Альтернативным подходом является моделирование. Разработанные компьютерные модели представляют определенные характеристики системы и могут быть использованы как для имитации работы мозга, как и для изучения свойств модели.

Кратко рассмотрим основные вычислительные модели распознавания объектов в вентральном зрительном пути. До появления современных сверточных нейронных сетей разрабатывались многочисленные модели имитирующие определенные аспекты работы зрительной системы. Например, модель Hierarchical Models and X (HMAX) (Riesenhuber, Poggio, 2000) фокусировалась на иерархической обработке сигнала в вентральном зрительном пути посредством моделирования двух типов клеток – простых и сложных, – аналогичных тем, что найдены в V1, а также вычислений более высокого уровня, необходимых для формирования абстрактных представлений признаков. Простые нейроны избирательны к определенной ориентации, в то время как сложные, несмотря на наличие избирательности, покрывают более широкий спектр пространственных частот, проявляя свойство инвариантности еще на ранних этапах обработки сигнала (Lian et al., 2021).

Последующие исследования представляют наличие рекуррентных связей для имитации механизмов обратной связи (Spoerer, McClure, Kriegeskorte, 2017). Эти модели согласуются с данными о рекуррентной обработке в вентральном потоке (Lamme, Roelfsema, 2000) и предполагают, что петли обратной связи могут помочь в задачах, требующих задействования контекстной информации или наличия внимания.

Также проводится изучение кросс-модальной интеграции, основывающееся на предположении, что вентральный поток не является изолированным, а взаимодействует с другими сенсорными модальностями (Amedi et al., 2007). Модели, учитывающие это, позволяют

объяснить некоторые проявления кросс-модальной пластичности, наблюдаемые в клинических условиях (Pascual-Leone, Hamilton, 2001).

Сверточные нейронные сети, как модель вентрального зрительного пути

Сверточные нейронные сети (CNN, СНС), принадлежащие к классу методов глубокого обучения, выделяются в отдельное направление на основе архитектуры моделей. Они расширяют идею иерархического извлечения признаков и проявляют высокое сходство с уровнями обработки в вентральном пути (Cadieu et al., 2014; Khaligh-Razavi, Kriegeskorte, 2014; Yamins et al., 2014). (Более подробно архитектура сверточных нейронных сетей будет рассмотрена в следующем разделе.) Имитация работы мозга не является основной задачей при обучении СНС, однако при распознавании образов проявляется эмерджентно.

При моделировании нейрофизиологических данных при помощи сверточных нейронных сетей можно выделить два последовательных этапа исследований, различающихся по своим целям и уровню сложности анализа.

На первом этапе СНС используются для создания моделей, способных предсказывать нейрональные ответы на различные стимулы (Yamins et al., 2014). Основная задача здесь заключается в имитации работы мозга – построении искусственных систем, которые воспроизводят наблюдаемые в экспериментах паттерны нейронной активности. Успешное решение этой задачи создает фундамент для более глубокого понимания принципов работы мозга.

Второй этап представляет собой более продвинутый уровень исследования, где уже обученные модели становятся самостоятельным объектом изучения (Kriegeskorte, 2015). На этом этапе анализируются свойства и характеристики моделей, которые сложно или невозможно исследовать непосредственно на биологических системах. Такой подход позволяет получать новые знания о принципах обработки информации и организации нейронных сетей без проведения экспериментов. Например, можно детально исследовать влияние различных параметров стимулов на активность искусственных нейронов или анализировать внутренние представления, формируемые в процессе обработки информации.

Исследования показывают, что иерархическая организация сверточных нейронных сетей демонстрирует функциональное сходство со зрительной системой приматов (Рис. 5). В работах (Cadieu et al., 2014; Khaligh-Razavi, Kriegeskorte, 2014, 2015; Kheradpisheh et al., 2016) показана схожесть репрезентаций глубоких слоев искусственных нейросетей и нижневисочной коры, как в случае анализа микроэлектродной регистрации нейронального ответа в коре макак, так и рассмотрении фМРТ данных человека. В исследовании (Güçlü, Gerven, 2015) выявлено соответствие между отдельными слоями искусственной нейронной сети и областями

вентральной зрительной коры, где наблюдается градиент возрастающей сложности обработки информации. Первичная зрительная область V1 соответствует начальным слоям сети (слои 1.6–1.8) и специализируется на обработке простых признаков, таких как контраст и края. Область V2 соотносится со средними нижними слоями (2.1–2.3) и обрабатывает признаки среднего уровня, включая контуры и текстуры. V4 соответствует средним верхним слоям (3.0–3.9) и настроена на обработку более сложных признаков, таких как формы и текстуры. Наконец, латеральная затылочная область LO соответствует глубоким слоям сети (5.0–5.2) и специализируется на обработке признаков высокого уровня, включая части объектов и целые объекты. При этом области демонстрируют частичное перекрытие функций, формируя иерархическую организацию обработки зрительной информации.

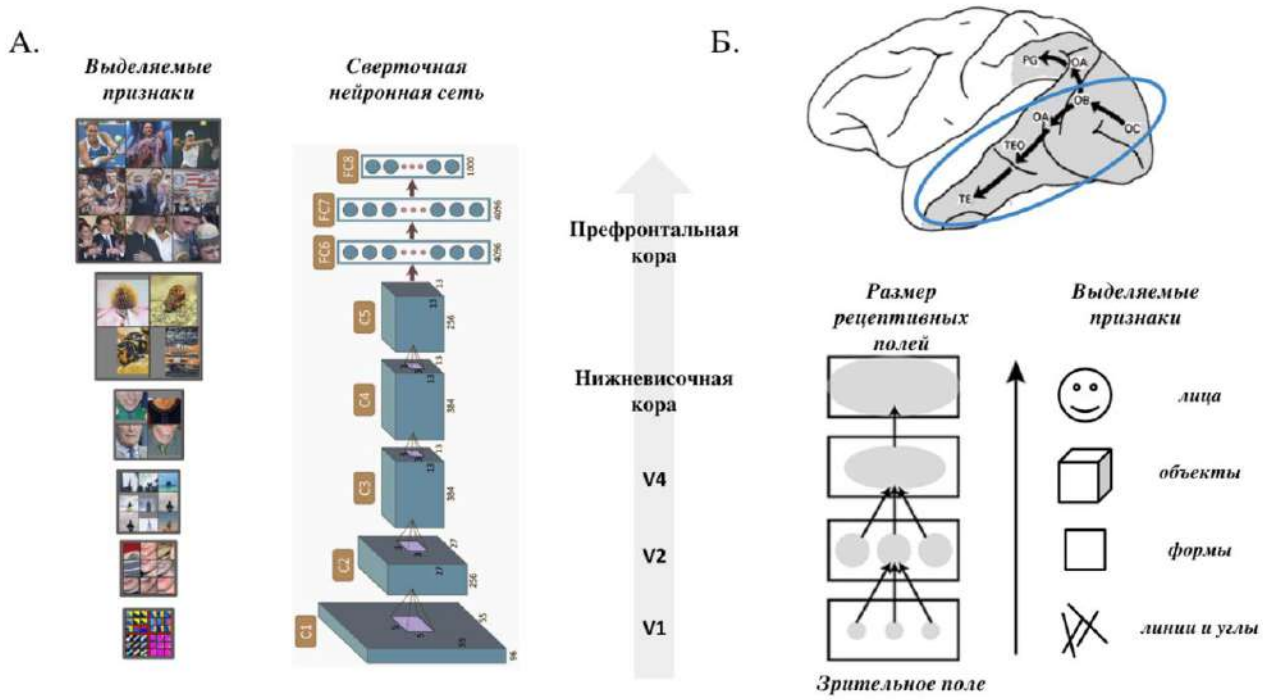

Рисунок 5. Иерархическая обработка сигнала в вентральном зрительном пути (Б) и сверточных нейронных сетях (А)

Широкое применение сверточных нейронных сетей для моделирования зрительных процессов, стало переломным моментом в нейронауках и открыло новые возможности к изучению свойств рецептивных полей, иерархической обработки сигнала и инвариантности в зрительных системах. Однако, с того времени в компьютерном зрении появились другие архитектуры, которые дополняют классические СНС или предлагают альтернативные подходы к работе со зрительной информацией. Среди наиболее ярких примеров можно выделить капсульные нейросети, трансформеры, диффузионные модели.

Трансформеры изначально были разработаны для задач обработки текстовой информации, но были адаптированы для работы с изображениями (например, Vision Transformers - ViTs) (Dosovitskiy et al., 2020). В отличие от СНС, сохраняющих пространственного расположение выделяемых признаков, трансформеры разделяют входное изображение на небольшие фрагменты, которые переводятся в вектора. При этом для сохранения пространственной информации используется позиционное кодирование. Основой работы трансформеров служат механизма самовнимания, взвешивающие значимость различных частей входного изображения. Архитектура этого класса моделей позволяет учитывать взаимодействия между фрагментами сцены, что важно для моделирования обработке глобального контекста зрительными областями коры, а также имитации механизмов пространственного внимания.

Капсульные сети (Sabour, Frosst, Hinton, 2017) строят иерархические, пространственные отношения между функциями с помощью «капсул», которые представляют собой небольшие группы нейронов, и используют динамическое маршрутизирование для передачи информации между уровнями иерархии. Эта архитектура обеспечивает инвариантность и робастность представлений, позволяя капсулам нижних уровней передавать информацию капсулам верхних уровней на основе согласованности прогнозов. Эти механизмы аналогичны модулярной и рекуррентной организации нейронных сетей в мозге, что делает капсульные сети интересными для нейронаук в контексте иерархического распознавания и интеграции информации.

Таким образом, переход от простого моделирования нейрональных ответов к использованию глубоких нейросетевых моделей как источника новых знаний представляет собой важное развитие в области нейрофизиологических исследований, расширяющее возможности изучения принципов работы мозга.

Архитектура сверточных нейронных сетей

С точки зрения архитектуры СНС обычно состоят из входного слоя, некоторого количества скрытых слоев и выходного слоя. Скрытые слои включают в себя сверточные слои, слои подвыборки, полносвязные слои и слои нормализации.

Сверточные слои. Сверточный слой представляет собой основной функциональный СНС, предназначенный для автоматического и адаптивного выделения пространственных признаков. Данный слой состоит из набора сверточных фильтров (ядер), применяющих операцию свертки к выходам с предыдущего слоя. Каждый фильтр выделяет отдельный признак, или пространственный паттерн, во входном сигнале. Параметры фильтра и то, какой признак он выделяет, формируются в процессе обучения сети.

Элементы первого сверточного слоя можно рассматривать как подобие простых клеток в первичной зрительной коре (V1), локальные пространственные частоты, ориентацию, текстуры.

Ранние слои обнаруживают низкоуровневые признаки, которые затем интегрируются последующими слоями для обнаружения признаков более высокого порядка.

Слои активации. Полученный результат применения фильтров свертки к входному сигналу подается на вход в функцию активации, которая представляет собой некоторую нелинейную функцию. Наиболее распространенной функцией активации в современных архитектурах является ReLU (от англ. rectified linear unit), производящая операцию отсечения отрицательной части входного сигнала.

Функции активации, подобные ReLU, вносят нелинейность в работу сверточных слоев. Их можно сопоставить с пороговой активацией в нейронах, когда требуется определенный уровень ответа на входной сигнал, чтобы возник потенциал действия.

Результат операции свертки входного сигнала одним из фильтров сверточного слоя с последующим применением функции активации называют *картой признаков*, или *картой активации*. Ответ сверточных фильтров на одно изображение либо набор изображений является *внутренним представлением* изображения в искусственной нейронной сети, кодированием входных данных посредством имеющейся структуры и весов на определенном уровне иерархии сети.

Слои подвыборки. Данные слои осуществляют нелинейную операцию уплотнения карты признаков, чаще всего за счет применения функции максимума. Таким образом, из некоторого пространственного региона входных значений остается только значение, отражающее наиболее сильный ответ.

Данную операцию можно сравнить с функциями сложных клеток в V1, чувствительных к одному и тому же признаку в небольшом пространственном диапазоне.

Полносвязные слои. Данные с последнего сверточного слоя уплощаются, утрачивая пространственную структуру, и передаются в полносвязную нейронную сеть, состоящую из некоторого количества слоев. Полносвязные слои объединяют признаки со всех частей изображения для выполнения задачи классификации.

Эти слои более схожи с нижневисочной корой, где нейроны проявляют избирательность к сложным формам и, как предполагается, интегрируют более простые признаки, обнаруженные ранее в вентральном потоке. Эти слои в CHC и нейроны в IT способствуют достижению высокоуровневого понимания зрительного сигнала на основе объединения более простых признаков.

Слои нормализации. Также стоит рассмотреть слои нормализации, которые нормализуют активации в сети таким образом, чтобы они имели приблизительно нулевое среднее и единичную дисперсию. Это улучшает сходимость задачи оптимизации, позволяя быстрее и стабильнее проводить обучение за счет регуляции внутренней статистики выходов слоя. По сути, они

выступают в роли регуляризаторов, не позволяя какому-либо одному признаку оказывать непропорционально большое влияние и обеспечивая тем самым сбалансированный вклад всех признаков.

Операцию нормализации можно сопоставить с механизмом латерального торможения, которое в биологических системах часто служит для выделения существенных особенностей путем снижения активности нейронов, реагирующих на менее значимые стимулы, что позволяет усиливать сигнал по сравнению с шумом. Если латеральное торможение акцентирует внимание на специфических, локальных признаках, подавляя другие, то слои нормализации работают на более глобальном уровне, также представляют собой механизм, регулирующий и обусловливающий внутренние представления сети.

Сверточные нейронные сети учатся классифицировать изображения на основе признаков, присутствующих в сцене. Отдельные функциональные единицы (фильтры или нейроны) СНС часто называют детекторами признаков, исходя из их способности определять наличие определенного пространственного паттерна на изображении. Внутреннее представление изображения для СНС, таким образом, формируется за счет активности существующих детекторов признаков.

Различные техники фокусируются на исследовании того, как изображение или понятие категории (класса) представлено в сети, что необходимо для понимания процесса классификации, осуществляемого сетью, повышения эффективности передачи знаний между моделями, контроля и интерпретируемости знаний, кодируемых моделью.

Исследование характеристик активности нейронов в искусственных нейросетях

В отличие от живых систем, после завершения обучающей стадии ИНС имеют фиксированную структуру, что приводит к детерминированному ответу при предъявлении одного и того же стимула. За меру ответа, аналогичную нейрональной активности отдельной клетки, чаще всего принимается значение a, полученное после взвешивания входного сигнала, суммирования и применения функции активации A:

Отдельным вопросом является необходимость сопоставления активности искусственного нейрона с фоновой активностью. Так как функционирование ИНС осуществляется только в процессе обработки подаваемого сигнала, то понятие фоновой активации как таковой отсутствует. Однако в некоторых случаях имеет смысл проведение оценки среднего ответа нейрона на предъявление различных стимулов.

Методы визуализации функции нейронных сетей и их элементов. Модели глубокого обучения часто критикуются за отсутствие интерпретируемости. Для решения этой проблемы

был разработан ряд подходов, позволяющих качественно и количественно описать поведение моделей.

Визуализация относится к наиболее распространенным способам интерпретации нейрональной активности. Данный подход является промежуточным шагом при составлении описания функциональной роли одного нейрона или их группы, популяции: в ходе эксперимента предъявляется набор стимулов и регистрируется вызванный ответ. Затем отбираются изображения, вызывающие максимальный отклик, либо же, наоборот, ингибирующие активность, строится график, иллюстрирующий меру активности в зависимости от предъявленного стимула.

Указанный метод строится на предположении, что найденная зависимость отражает функциональную роль нейрона в осуществлении задачи распознавания образов. Также предполагается, что изображения характеризуются неким признаком, присутствие которого вызывает повышение активности, и наоборот. Не исключается, что признак может быть негативным, то есть заключаться в отсутствии чего-либо. Также не допускается, что признаки могут быть целостными, то есть не поддающимися описанию одним конкретном словом или зрительным образом.

Следуя указанной парадигме, поиск изображений, вызывающих наиболее сильную активацию, позволит сделать заключения о функции нейрона. При проведении нейро- и психофизиологических экспериментов исследователями формируется база изображений, предположительно отражающих указанную функцию, по которой проводится оценка. При работе с ИНС производится поиск по обширным наборам данных, без предварительного отбора.

Подобный подход называется максимизацией активации и может быть сформулирован в виде задачи оптимизации как поиск такого изображения, которое бы максимизировало активацию a выбранного нейрона. Или для сверточных нейронных сетей:

$$x^* = \underset{x}{\operatorname{argmax}}(a_i^l(\theta, x)),$$

где x начальное изображение; x^* – последующая итерация изображения; a_i^l – активация i-го элемента в слое l. При этом – параметры нейронной сети, которые переводят входное изображение размерности $H \times W \times C$ (высота, ширина, количество каналов) в вероятностное распределение тренировочных классов.

Так как поставленная задача является оптимизационной, то вместо перебора изображений она может быть решена при помощи методов генерации и постепенной модификации. Подобные техники относятся к наиболее прогрессивным методам интерпретации работы ИНС и реализуют в себе различные подходы к синтезированию изображений (Erhan et al., 2009; Simonyan, Vedaldi,

Zisserman, 2013; Wei et al., 2015; Nguyen, Yosinski, Clune, 2016; Nguyen et al., 2016; Nguyen, 2017; Olah, Mordvintsev, Schubert, 2017).

Генеративно-состязательные сети. Другим подходом к пониманию работы нейронных сетей является применение генеративно-состязательных сетей (GAN) (Goodfellow et al., 2014) и вариационные автоэнкодеры (VAE) (Kingma, Welling, 2014) с целью создания визуализаций и других интерпретаций, помогающих понять кодирование информации в модели глубокого обучения.

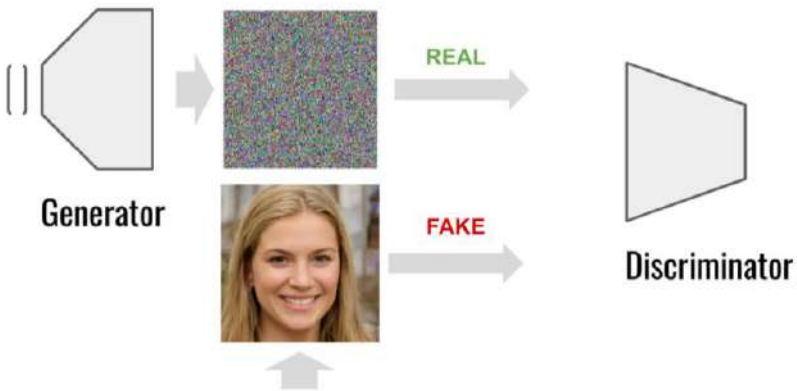

Рисунок 6. Обучение генеративной сети. Слева – модель генератор, справа – дискриминатор. Дискриминатору предъявляется два изображения – реальное (из набора обучающих данных) и сгенерированное

Генеративно-состязательные сети (GANs) представляют собой класс генеративных моделей, которые состоят из двух взаимодействующих нейронных сетей: генератора и дискриминатора.

Генератор принимает на вход случайный шум и пытается преобразовать его в данные, которые внешне неотличимы от реальных. Дискриминатор, в свою очередь, принимает на вход как реальные данные, так и сгенерированные и пытается отличить одни от других (Рис. 6). В процессе обучения генератор становится все совершеннее в создании ложных стимулов, а дискриминатор – в их определении, в результате чего обе сети взаимно обучают друг друга.

Важным моментом является то, что помимо возможности создавать реалистичные изображения и другие виды данных, GANs могут быть применены и для интерпретации работы других сетей. Например, с целью визуализации внутренних представлений сверточных нейронных сетей. В работе (Bau et al., 2018), GANs использовались для создания визуализаций признаков, которые CNN изучила для задачи классификации изображений. Это позволило получить более глубокое понимание того, как работают отдельные нейроны и как они связаны с конечной задачей.

VAE, в свою очередь, использовались для анализа и интерпретации работы рекуррентных нейронных сетей. (Karpathy, Johnson, Fei-Fei, 2015) использовали VAE для генерации идеального

стимула для рекуррентной сети, которые отражали внутренние представления модели о временных рядах.

Исследование репрезентаций в обученных сетях. Искусственные нейронные сети применяются для моделирования зрительного восприятия как на уровне оценки поведенческих реакций (например, категоризации изображений и принятия решений (Lake et al., 2015; Dodge, Karam, 2017; Geirhos et al., 2018; Baker et al., 2018; Gangopadhyay, Das, 2019), подверженности зрительным иллюзиям (Williams, Yampolskiy, 2018; Watanabe et al., 2018)), так и на более глубоком уровне понимания представления информации и характеристик работы нейронов в различных слоях сети (Cadieu et al., 2014; Kriegeskorte, 2015; Kheradpisheh et al., 2016; Kietzmann, McClure, Kriegeskorte, 2018).

Однако при моделировании зачастую не учитываются различия в кодировании информации, вызванные постановкой задачи во время обучения искусственного агента. Большинство работ фокусируется на изучении репрезентации категорий объектов и отдельных зрительных образов. Анализ и сопоставление откликов осуществляется на уровне отдельных нейронов, что, тем не менее, не говорит о принципиальной схожести пространств представления зрительной информации.

Исследования в области нейрофизиологии живых организмов показали, что функция, выполняемая нейронами, определяется количеством ресурсов, выделяемых для обработки сенсорного сигнала, мерой иерархичности системы и важностью задачи для организма. Начиная с сетчатки, работа клеток оптимизирована под статистику зрительной сцены и выполняемую задачу (Lettvin et al., 1959; Bialek, Owen, 1990; Atick, Redlich, 1992; Gollisch, Meister, 2010), затем сохранение информация о статистике сигнала прослеживается и в последующих звеньях зрительного пути – таламусе (Dan, Atick, Reid, 1996), первичной зрительной коре (Olshausen, Field, 1996; Bell, Sejnowski, 1997), области V2 и выше (Hyvärinen, Gutmann, Hoyer, 2005; Cadieu, Olshausen, 2012). В то же время известно, что обширную область высших отделов зрительной системы как человека, так и ряда человекоподобных обезьян, занимают клетки, отвечающие за распознавание таких важных для выживания категорий, как лица (Gross, Rocha-Miranda, Bender, 1972; Tanaka, 1996; Kanwisher, McDermott, Chun, 1997; McCarthy et al., 1997), тела (Downing et al., 2001), места обитания (Epstein, Kanwisher, 1998). Таким образом, характеристики входного сигнала позволяют объяснить функциональные свойства начальных этапов зрительного восприятия.

Направленное создание стимульного материала во время проведения экспериментальной сессии. Новым направлением в нейрофизиологии является разработка методов повышения качества процесса исследования при помощи алгоритмов и методов искусственного интеллекта.

В отличие от классического экспериментального подхода с предъявлением статичных изображений, более современные методы включают использование компьютерных моделей для создания стимулов (FaceGen, 2023), разработку виртуальных сред для контролируемого воздействия (Bundell, 2019; Pizzoli et al., 2019; Riva, Serino, 2020), а также применение алгоритмов для оптимизации стимуляции на основе обратной связи (DiMattina, Zhang, 2013; Gardner et al., 2015; Lancaster et al., 2018; Lorenz, Hampshire, Leech, 2017).

Подобные работы, направленные на создание и внедрение новых методик зрительной стимуляции, ставят своей задачей усовершенствование самого процесса экспериментального поиска. Например, методы байесовской оптимизации могут быть применены для управления ходом эксперимента и поиска оптимальных настроек экспериментального оборудования, алгоритмы активного поиска позволяют максимизировать получаемую информацию за счет адаптивного подбора стимулов на основе регистрируемого психофизического ответа, и т. д. Такой подход позволяет сократить время эксперимента и снизить уровень шума в полученных данных.

Обсуждение

Исследование процессов распознавания зрительных образов в высших отделах зрительной системы приматов подтверждает сложность и высокую степень организации зрительного восприятия. Зрительная система организована иерархически, где нейроны первичных зрительных областей (V1) реагируют на простейшие признаки, такие как ориентация линий и частотные характеристики. Постепенно информация обрабатывается и интегрируется в вентральном пути, что позволяет выделять более сложные признаки, включая формы и текстуры. На завершающих этапах вентрального пути (в нижневисочной коре) достигается инвариантное представление категорий объектов, что позволяет организму эффективно распознавать зрительные образы, несмотря на изменения в фоновом контексте, положении или размере объекта.

Одним из ключевых вопросов является то, как нейроны высших областей формируют устойчивые категории объектов и каким образом достигается инвариантность восприятия. Исследования показывают, что эта способность связана с механизмами интеграции информации на разных уровнях зрительной системы, а также с возможностью контекстно-зависимой модуляции нейронных ответов. Это особенно важно для объяснения устойчивости восприятия к изменениям в освещении, ракурсе и других параметрах стимула.

В данной главе рассматриваются *различные подходы к исследованию функциональных характеристик нейронов*. Индекс чувствительности применяется для определения способности нейрона к различению сигнала от фонового шума. Метрики избирательности позволяют оценить,

насколько нейрон предпочитает определенную категорию стимулов. Кривые настройки используются для анализа и визуализации свойств нейрона при предъявлении набора стимулов.

Проведено *сравнение метрик избирательности Face Selectivity Index (FSI) и Category Selectivity Index (CS)* на основе симулированных данных. Сравнение показало различную чувствительность методов к особенностям распределения активаций: FSI хорошо выявляет разницу между средними откликами, даже если категории частично пересекаются, в то время как CS более точно отражает степень разделения категорий, что важно учитывать при выборе метода анализа нейронных откликов в биологических и искусственных системах.

Методологический анализ показывает, что традиционные подходы к изучению нейронной активности, имеют как сильные, так и слабые стороны. К их преимуществам относится возможность количественной оценки реакции нейронов на конкретные стимулы и построение кривых настройки, что особенно эффективно для понимания роли нейронов на ранних этапах обработки информации. Однако эти методы обнаруживают существенные ограничения при исследовании высокоуровневых областей мозга, где нейроны реагируют на сложные комбинации признаков и проявляют контекстно-зависимые свойства. Это обуславливает необходимость разработки новых подходов, основанных на улучшенных технологиях исследования функции нейронов.

В главе подробно рассмотрены сверточные нейронные сети как модели вентрального зрительного пути. Продемонстрировано их функциональное сходство с биологическими системами по структуре и принципам обработки информации. Показано, что иерархическая организация сверточных нейронных сетей соответствует градиенту возрастающей сложности обработки информации в зрительной коре.

Однако функциональное соответствие искусственных нейронных сетей и вентрального зрительного пути остается предметом дискуссий. С одной стороны, CHC демонстрируют успешное воспроизведение поведенческих и нейрофизиологических данных, включая инвариантность представлений. С другой стороны, их обучение и представления информации могут отличаться от биологических механизмов: например, в мозге процесс обучения происходит без явного наличия обратно распространяемого градиента, а активность нейронов может зависеть от рекуррентных связей и модуляции внимания.

Тем не менее использование методов машинного обучения открывает новые возможности в области нейрофизиологии, включая задачи моделирования и интерпретирования работы биологических нейронных систем. Эти методы помогают выявлять ключевые характеристики, на которые реагируют отдельные нейроны, а также создавать контролируемые стимулы для нейрофизиологических экспериментов. Открытым остается вопрос о том, насколько результаты, полученные в таких моделях, могут быть перенесены на живую нервную систему. Это

подчеркивает важность дальнейшего изучения свойств моделей, понимания их ограничений и возможной сферы применения, а также необходимость тщательного сопоставления модельных данных с экспериментальными результатами.

Описаны различные методы интерпретации нейрональной активности, включая техники визуализации и применение генеративно-состязательных сетей. Эти подходы помогают выявлять ключевые характеристики, на которые реагируют отдельные нейроны, а также создавать контролируемые стимулы для нейрофизиологических экспериментов.

Современные исследования зрительного восприятия требуют интеграции нейрофизиологических методов, вычислительного моделирования и машинного обучения. Комбинированные подходы, такие как использование искусственных нейронных сетей для генерации гипотез о кодировании информации в мозге и их последующая проверка в нейрофизиологических экспериментах, открывают новые возможности в изучении нейронных механизмов восприятия.

Перспективы дальнейших исследований включают разработку биологически правдоподобных вычислительных моделей, интеграцию экспериментальных данных в обучение нейросетей и усовершенствование методов анализа популяционной активности нейронов. Особое внимание следует уделить созданию подходов, объединяющих разные методы, а также новым способам визуализации и интерпретации нейронной активности, что позволит глубже понять принципы обработки информации в зрительной системе.

Глава 2. Нейрофизиологические исследования работы нейронной сети нижневисочной коры в процессе распознавания образов объектов

В **Главе 2** представлено описание и анализ данных нейрофизиологических экспериментов, в которых проводилась регистрирация нейронов нижневисочной коры приматов в ответ на предъявление набора изображений. Основное внимание уделено следующим аспектам:

- методологии сбора данных, включая дизайн эксперимента, стимульный материал, регистрацию и предварительную обработку нейрональной активности;
- характеристике импульсной активности нейронов, ее зависимости от пространственной организации и функциональных особенностей зрительной системы;
- анализу избирательности нейронов к различным категориям зрительных стимулов, а также закономерностям распределения нейрональной активности;
- интерпретации полученных результатов с точки зрения понимания кодирования признаков в высокоуровневых областях зрительной системы.

Полученные данные формируют эмпирическую основу для последующего моделирования процессов зрительного восприятия и анализа принципов нейронной обработки зрительной информации.

2.1. Сбор экспериментальных данных

Исследования функциональных характеристик нейронов высших отделов зрительной системы проводились на приматах (макаки), зрительная система которых во многом схожа со зрительной системой человека. Данные были получены методом внеклеточной микроэлектродной записи в серии экспериментов в лаборатории Танифуджи М. в RIKEN Brain Science Institute (Япония). Регистрация нейрональной активности производилась в области нижневисочной коры у трех макак в серии экспериментов (Sato, Uchida, Tanifuji, 2009; Sato et al., 2013). Экспериментальный протокол был одобрен комитетом по экспериментальным животным института RIKEN и соответствовал рекомендациям института RIKEN и Национальных институтов здоровья (NIH).

Расположение электродов. Электроды располагались в передней части нижневисочного отдела коры (Anterior Inferotemporal Cortex, область TEad), захватывающей заднюю стенку верхнетеменной борозды (Superior Temporal Sulcus), предполагая вовлеченность области в обработку сложных зрительных стимулов, включая распознавание лиц и тел.

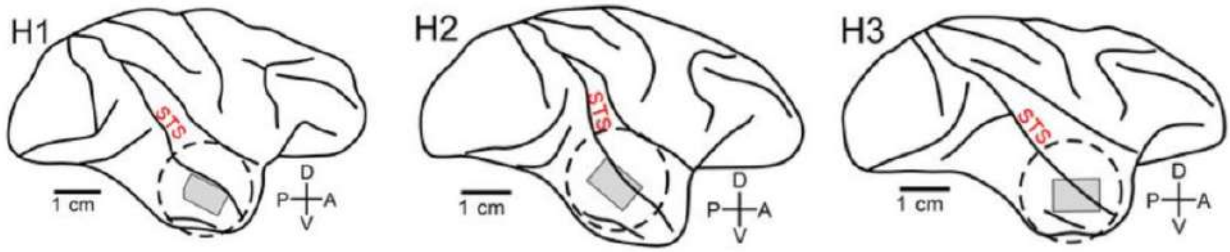

Рисунок 7. Схема размещения микроэлектродов у трех макак-резусов (H1, H2, H3) в Эксперименте 1. Адаптировано из: (Sato et al., 2013)

Регистрация нейронального ответа производилась в 190 точках записи у трех макак (у H1 – 33 точки записи, H2 – 133, у H3 – 24 соответственно) (Sato, Uchida, Tanifuji, 2009; Sato et al., 2013). Размещение электродов приведено на Рис. 7. Пунктирной линией изображено положение камеры, серым прямоугольником – расположение массивов микроэлектродов внутри камеры. Верхняя височная борозда обозначена аббревиатурой STS.

Регистрация активности нейронов выполнялась при помощи массивов из восьми микроэлектродов Utah (Blackrock Neurotech), внедренных в кору головного мозга перпендикулярно ее поверхности. Полученный сигнал усиливался и пропускался через полосовой фильтр с частотой 500 Гц – 3 кГц, после чего оцифровывался с частотой 25 кГц. Наличие спайка фиксировалось, в случае если электрический сигнал превышал среднее значение более чем на 3,5 стандартных отклонения.

Стимульный материал. В экспериментах использовался набор изображений *Takayuki-1550* (Sato, Uchida, Tanifuji, 2009), состоящий из 1550 стимулов, в числе которых следующие категории: лица людей (535 изображений), лица макак (285), тела макак (50), тела людей (50), другие животные (200), цветы и растения (120), пейзажи и природные объекты (120), искусственно созданные предметы и объекты (150), перевернутые лица обезьян и людей (40). Включение в набор данных широкого спектра категорий позволяет более точно определить избирательность нейронов к тем или иным признакам. Пример стимулов приведен на Рис. 8. Усредненное изображение стимула для категории с лицами и других изображений приведено на Рис. 9.

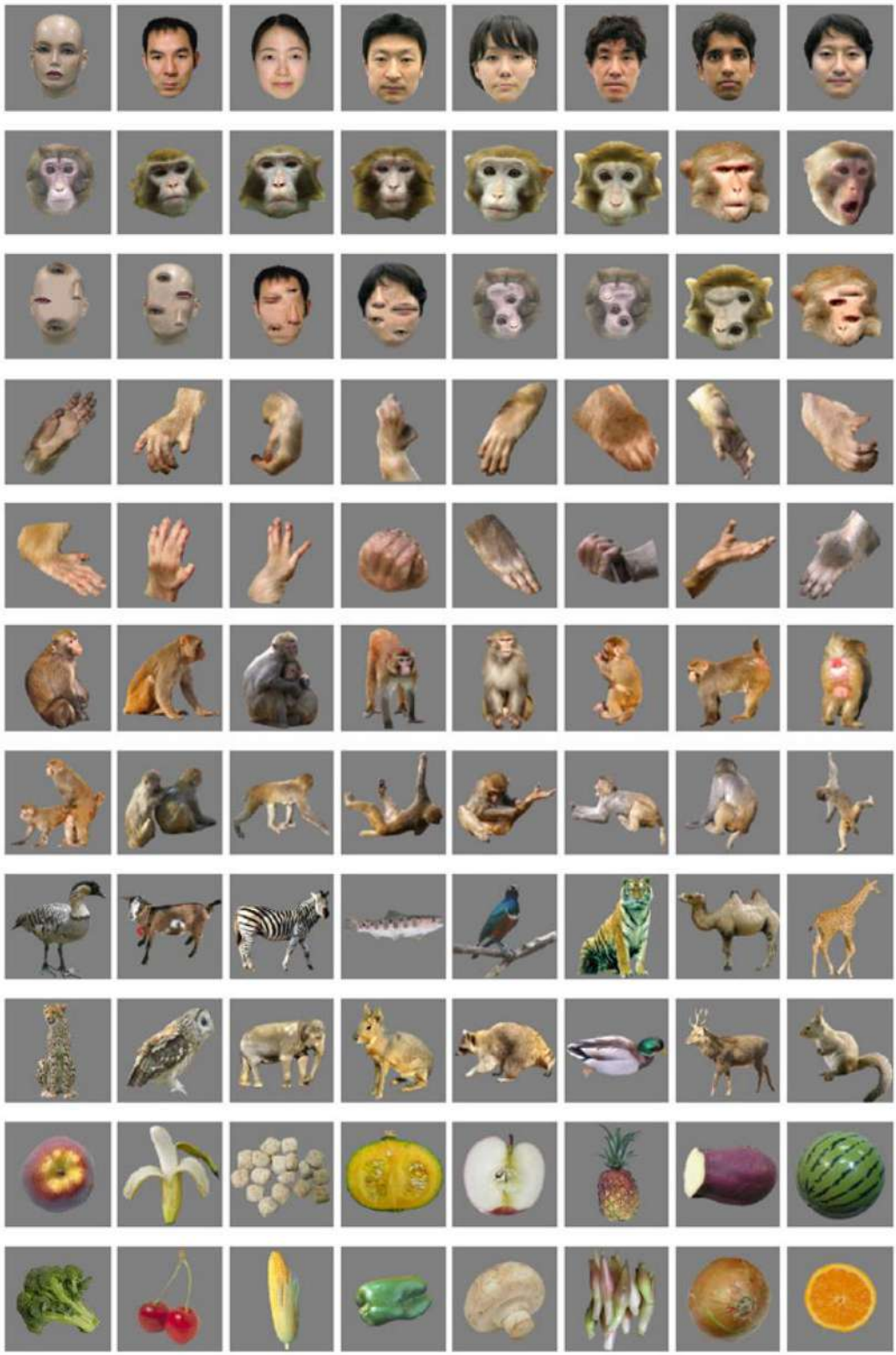

Рисунок 8. Пример изображений из набора стимулов, адаптировано из (Sato et al., 2013)

Дизайн эксперимента. Стимулы предъявлялись на сером фоне через 21-дюймовый ЭЛТ-монитор, установленный на расстоянии 57 см от глаз животных. Размер предъявляемых объектов и лиц составлял 20° зрительного угла для субъектов H1 и H3, и 10° для субъекта H2. В связи с тем, что животные находились под анестезией во время записи нейронной активности, положение стимулов автоматически смещалось на 0,15° зрительного угла каждые 12 мс.

Протокол стимуляции включал предъявление каждого стимула на 100 мс с последующим межстимульным интервалом в 200 мс.

Размер стимульных изображений составлял 200 × 200 пикселей (угловой размер – 20°). Весь набор изображений предъявлялся в 12 блоках в псевдослучайной последовательности. Для исключения эффекта подавления адаптации (1509 различных изображений, каждое повторялось 12 раз) изображения внутри блока предъявлялись в случайном порядке.

Для нейронов нижневисочной области *периодом фоновой активности* рассматривался интервал с *50 мс* предшествующих предъявлению стимула до 50 мс после его предъявления, отражая спонтанную активность нейронов (Nam et al., 2021).

Период *с 70 мс до 220 мс* рассматривался как *период релевантного нейронального ответа*. Для получения *вызванного ответа* из частоты пульсации в релевантный период вычиталась частота пульсации в предварительный период. Вызванный нейрональный ответ был *усреднен по блокам* с повторениями стимулов. Активность учитывалась *в целом по нейрональной колонке* – данные, полученные с одного микроэлектрода усреднялись по 8-ми каналам.

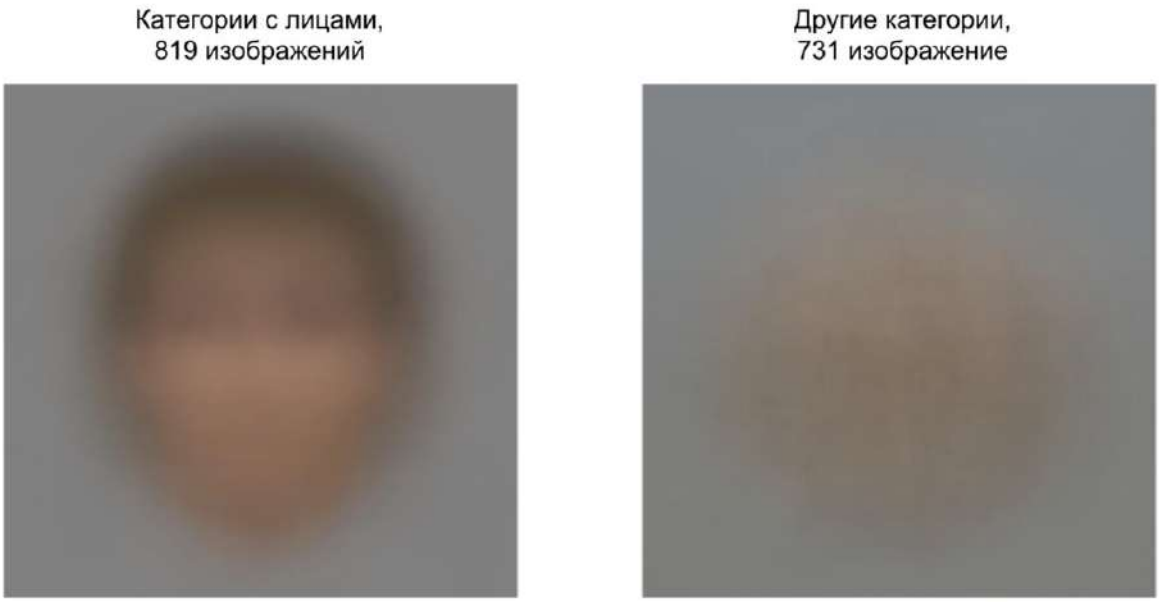

Рисунок 9. Усредненные стимулы для категорий с изображениями лиц и всеми другими категориями

Имплантация микроэлектродов (Sato, Uchida, Tanifuji, 2009; Sato et al., 2013). Перед первоначальной операцией по имплантации записывающей камеры обезьянам было проведено МРТ-сканирование. Реконструировалась боковая проекция мозга с указанием борозд и извилин из корональных срезов МРТ-изображений для определения положения записывающей камеры.

Во время начальной операции имплантировался пост для фиксации головы и титановая записывающая камера (внутренний диаметр 18 мм). Титановый пост для фиксации головы был

прикреплен к верхней части черепа. После закрепления поста два болта из нержавеющей стали для записи электроэнцефалограммы (ЭЭГ) были имплантированы над дуральной поверхностью левой и правой лобных долей. В удалении от стальных болтов для записи ЭЭГ был имплантирован перевернутый Т-образный титановый болт (Т-болт) для электрического заземления. Центр камеры приблизительно совпадал с центром передней средней височной борозды по оси AP, а треть расстояния от верхнего края камеры приходилась на верхнюю височную борозду (по оси DV). Подобное размещение соответствовало области TE (TEad) нижневисочной коры. Обычно положение центра камеры находилось на 15–20 мм впереди от ушной перекладины. Поверхностный рисунок кровеносных сосудов использовался в качестве эталона для картирования мест проникновения электродов.

Во время первоначальной операции по имплантации электродов для фиксации головы обезьян анестезировали внутрибрюшинной инъекцией пентобарбитала натрия (35 мг/кг). Глубокий наркоз поддерживали дополнительными внутривенными инъекциями пентобарбитала натрия (5–10 мг). Температура тела поддерживалась на уровне 36,6°C. На протяжении всей операции отслеживалась ЭКГ.

В первый день проведения экспериментальной сессии, во время обнажения поверхности коры внутри камеры записи, обезьян искусственно вентилировали следующей смесью: N_2O (70%), O_2 (30%), изофлуран (1–2%). На протяжении всей процедуры проводился мониторинг ЭКГ и ЭЭГ, и поддерживалась глубокая анестезия, за счет регулирования концентрации изофлурана. Также во время всех экспериментов поддерживался концентрация CO_2 между 3,5 и 4,5%, а также температура тела на уровне 37,6°C.

Во время электрофизиологической записи обезьяны были анестезированы внутривенным введением бромида векурония (0,067 мг/кг/ч) и искусственно вентилировались смесью, содержащей 70% N_2O, 30% O_2, и до 0,5% изофлурана. В целях недопущения болевых ощущений непрерывно в течение всего эксперимента внутривенно вводился цитрат фентанила (0,83 мкг/кг/ч). Также проводились ЭЭГ, ЭКГ, отслеживались концентрация CO_2 в выдыхаемом воздухе (от 4,0 до 5,0%) и поддерживалась температура тела на уровне 37,6°C.

2.2. Статистический анализ нейрональной активности клеток нижневисочной коры

В данном разделе приводятся результаты статистического анализа импульсной активности нейронов нижневисочной коры в ответ на предъявление стимулов Takayuki-1550. В частности, рассматриваются общие характеристики импульсной активности нейронов, а также различия в активности в ответ на предъявление стимулов различных категорий.

Для изучения функциональных характеристик зрительной системы базовой единицей анализа рассматривается нейрональный модуль, соответствующий функциональной единице нижневисочной коры.

2.3.1. Характеристики импульсной активности

Регистрация вызванного нейронального ответа производилась в 190, однако из них в анализе рассматривались только 143 точках записи, остальные были исключены по техническим причинам. Относительное расположение электродов на коре изображено на Рис. 10 Нейрональный отклик (активность), вызванный стимульным воздействием, рассчитывался как относительная величина – разница в количестве импульсов в секунду по сравнению с фоновой активностью.

Одной из важных характеристик нейронального ответа, существенно влияющей на качество анализа, является стабильность активации при повторном предъявлении стимула. Так как каждое изображение было предъявлялось 12 раз в 12 экспериментальных блоках, было возможно провести оценку стабильности нейронального ответа. Анализ был выполнен при помощи *t*-критерия Стьюдента для множественных сравнений, проверяющего гипотезу о различии средних двух зависимых групп – ответов на предъявление идентичного стимула в точке регистрации для четных и нечетных экспериментальных блоков. Общее количество точек регистрации с *p*-значением $< 0{,}05$ (вероятностью отрицания нулевой гипотезы) составило 78. Таким образом, в указанной группе активность нейронов достоверно вызывала схожий ответ при повторных предъявлениях одного и того же стимула.

Полученные t-статистики и p-значения применялись для выделения точек регистрации со «стабильным» меж экспериментальными блоками нейрональным откликом (p-значение < 0.05).

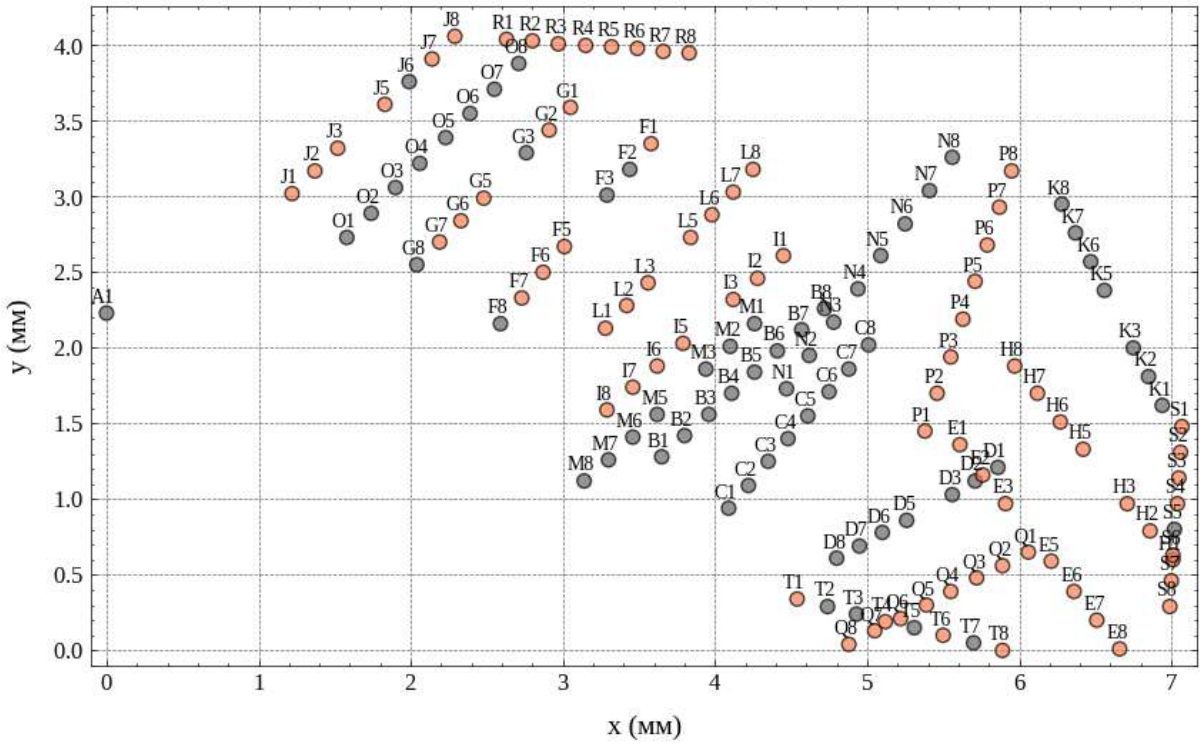

Рисунок 10. Относительные координаты электродов, регистрирующих нейронную активность. Оранжевым цветом обозначены точки регистрации со «стабильным» блоками ответом

Проведен анализ схожести ответов в зависимости от евклидова расстояния между точками регистрации для каждого наблюдателя. Показано, что существует обратная зависимость: чем ближе располагаются друг к другу электроды, тем выше корреляция их активности (коэф. корр. Спирмена = 0,41, $p \leq 0.001$ на данных с точек записи со стабильным ответом, 0,35 на всех данных) (Рис. 11). Для расчета коэффициента корреляции был выбран подход Спирмена, так как так как он позволяет выявлять монотонные зависимости, не предполагая линейности между переменными. Оптимальной моделью аппроксимации оказалась либо степенная зависимость, характерная для процессов с иерархической организацией и локальными взаимодействиями (R^2 = 0.248 на данных с точек записи со стабильным ответом, R^2=0.171 на всех данных), либо логарифмическая зависимость (R^2 = 0,239 и R^2 = 0,169 соответственно). Объясняющая способность экспоненциальной (R^2 = 0,196 и R^2 = 0,145) и линейной (R^2 = 0,179 и R^2 = 0,127) была существенно ниже. Это согласуется с тем, что кортикальные нейроны формируют локально связанные ансамбли, в то время как с увеличением расстояния эти связи ослабевают, снижая синхронизацию.

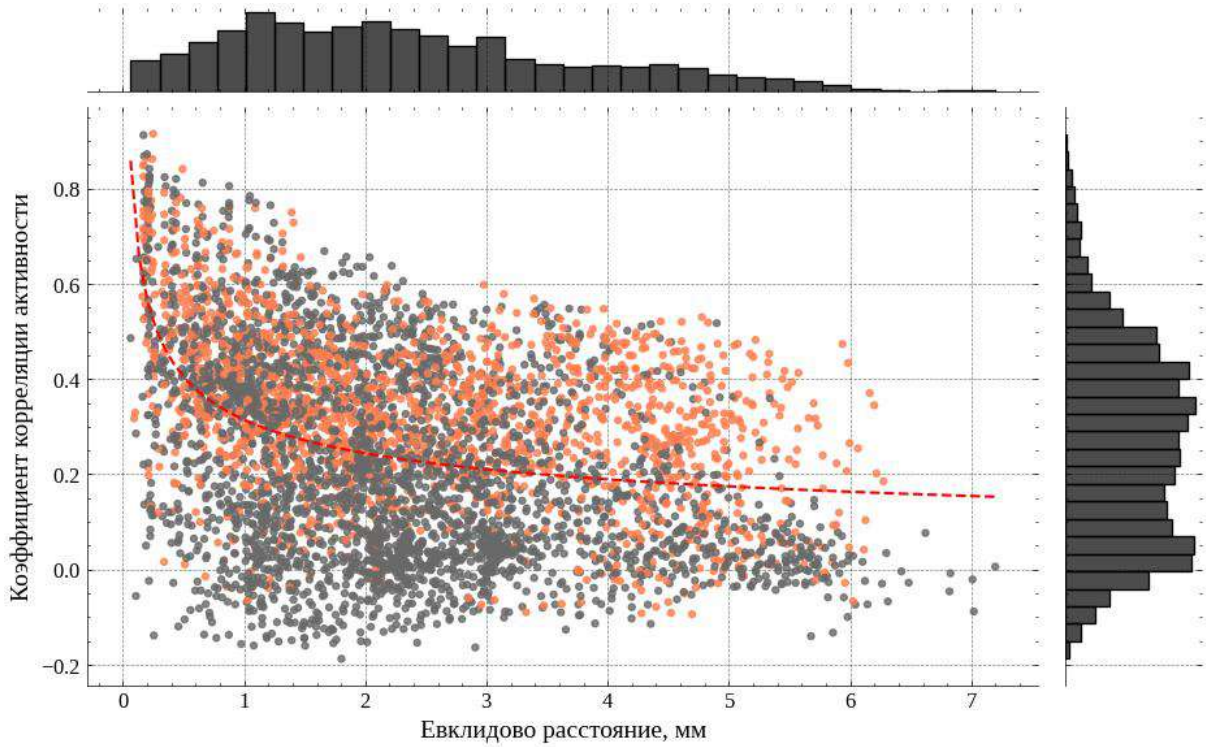

Рисунок 11. Сходство в активности нейронов в зависимости от расстояния, на котором они располагаются: по оси абсцисс указано евклидово расстояние в мм, по оси ординат – корреляция импульсной активности. Красной пунктирной линией отмечены значения степенной аппроксимации. Оранжевым цветом обозначены точки регистрации со «стабильным» ответом

Анализ вариативности в ответах нейрональных колонок (Рис. 12) выявил зависимость между средней вызванной активностью точки регистрации и среднеквадратическим отклонением в активности: чем выше средний вызванный ответ, тем больше разброс в возможных ответах (коэф. корр. Спирмена = -0.58, p-значение $\leq$ 0,001). Зависимость имеет возрастающий характер, однако для однозначного определения рода зависимости недостаточно данных, так как различные функции схожим образом аппроксимируют данные: степенная (R^2 = 0.22); логарифмическая (R^2 = 0.13), линейная (R^2 = 0.22). Тем не менее возрастающий характер можно объяснить тем, что нейроны с сильной активностью не только имеют более высокий средний ответ, но и демонстрируют большую изменчивость. Также представляет интерес область со средней активностью 5–15 спайков/с, где при относительно небольшой средней активности отмечается широкий разброс в стандартном отклонении ответов (от 10 до 40 единиц). Такая картина может отражать сложную природу функционирования этих нейронов.

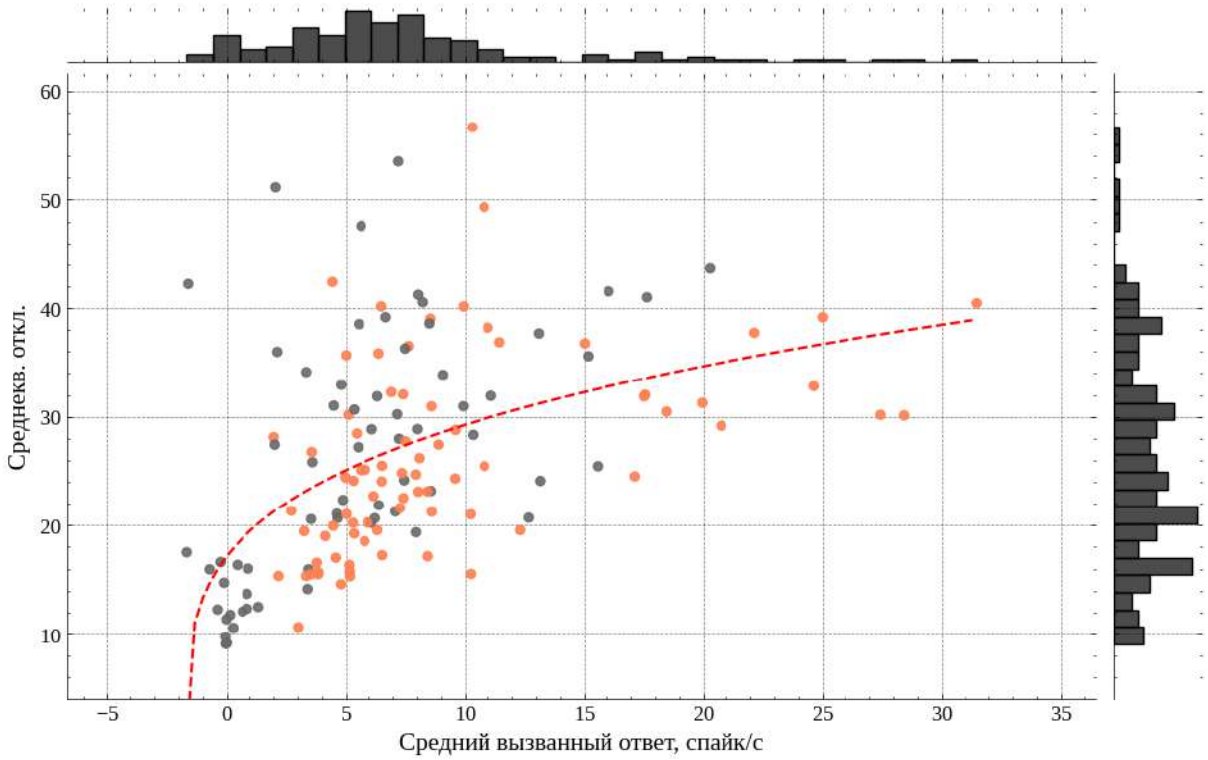

Рисунок 12. Вариативность в ответах нейрональных колонок в зависимости от количества регистрируемых спайков. По оси абсцисс – средняя вызванная активация нейрональной колонки. По оси ординат – стандартное отклонение в ответах. Оранжевым цветом обозначены точки регистрации со стабильным ответом. Красной пунктирной линией отмечены значения степенной аппроксимации

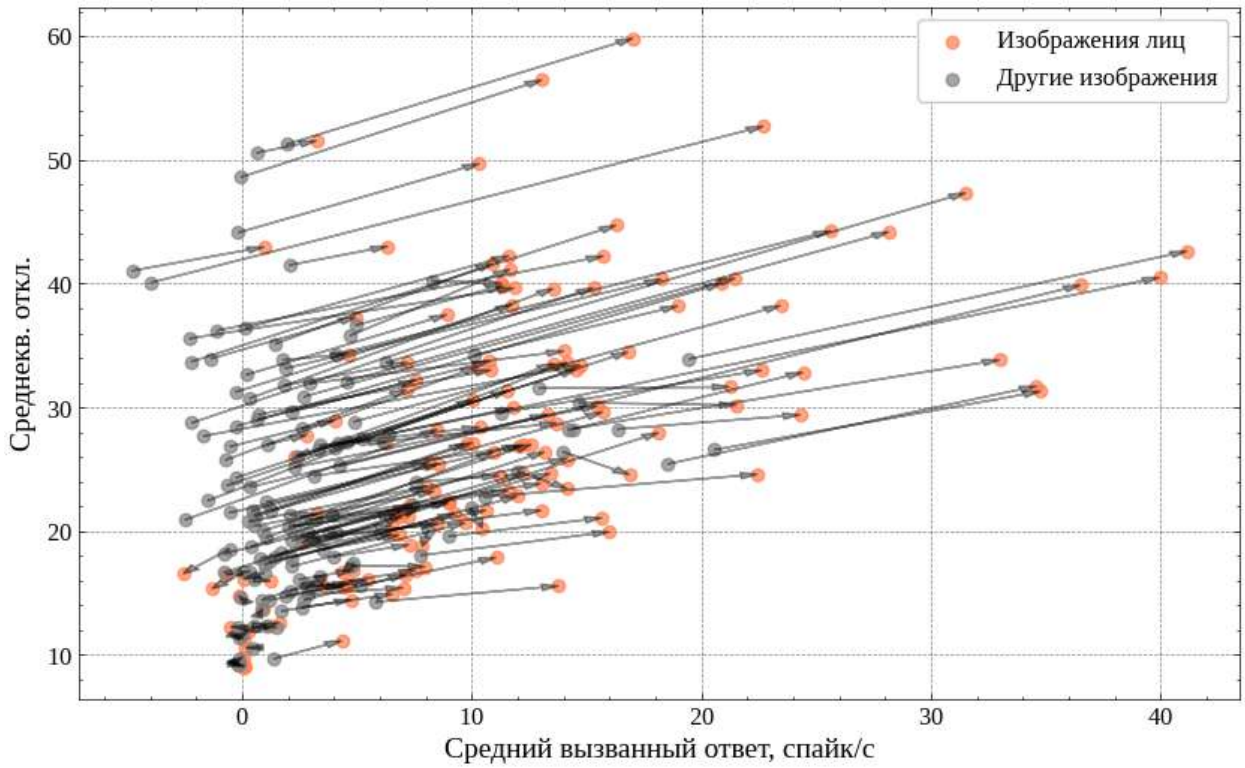

Рисунок 13. Изменение ответа нейрональных колонок при предъявлении изображений лиц и других стимулов. По оси абсцисс – средняя вызванная активация, по оси ординат – стандартное

отклонение в ответах. Оранжевым цветом обозначены точки регистрации ответа на изображения лиц, серым – на стимулы без лиц. Линиями соединены точки, соответствующие одной и той же нейрональной колонке в разных условиях, что позволяет оценить направленность изменений в средней активности и вариативности ответов

Были проанализированы средние вызванные ответы и стандартные отклонения в двух условиях: при предъявлении изображений, принадлежащим к категориям «лица» и «другие изображения». Каждому нейрону соответствуют две точки на графике (Рис. 13) – одна для изображений лиц, другая для других стимулов, соединенные линией, что позволяло оценить направленность изменения ответа. Разница между условиями была количественно оценена с помощью теста Уилкоксона для зависимых выборок. Анализ показал, что у 130 из 143 нейронов средний отклик был выше на изображений лиц, в то время как только 13 демонстрировали обратный эффект (разница в среднем 8,04 спайков/с, SD = 6,88, $p < 0,001$). Аналогично, у 118 нейрональный ансамблей стандартное отклонение увеличивалось при предъявлении стимулов с изображениями лиц, тогда как у 25 оно возрастало на другие изображения (разница 3,51, SD = 3,70, $p < 0,001$). Эти результаты подтверждают выраженную тенденцию к повышенной активности нейронов при восприятии лиц, сопровождающуюся увеличением вариативности их ответов.

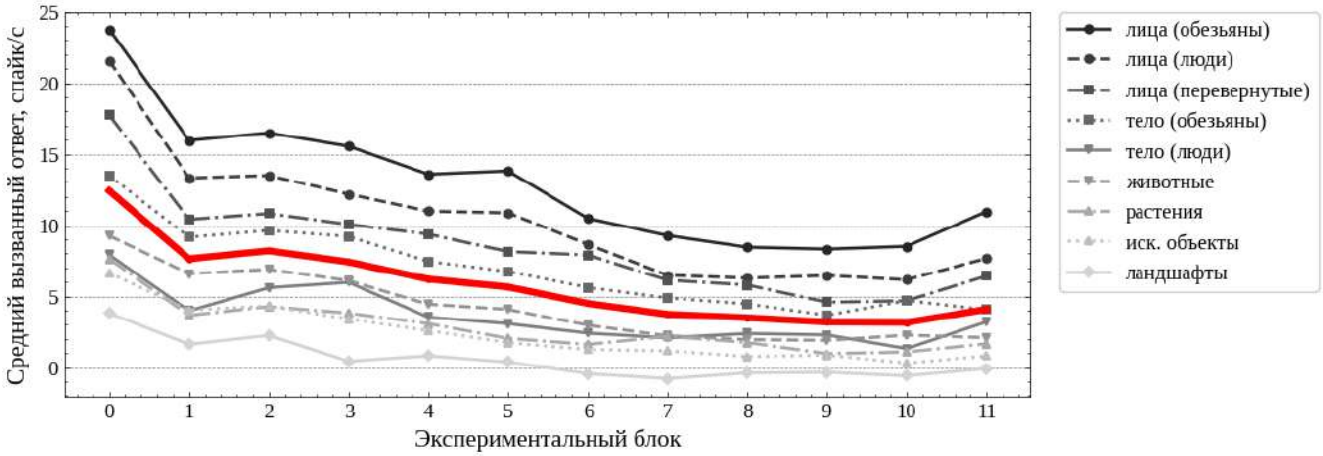

Рисунок 14. Средний ответ на изображения различных категорий. По оси абсцисс – номер экспериментального блока, по оси ординат – вызванный ответ. Красной сплошной линией обозначен усредненный ответ для всех категорий

По мере проведения эксперимента наблюдалось снижение как среднего вызванного ответа, так и различий в ответе на изображения, относящиеся к разным категориям (Рис.14). Из чего можно заключить, что первые шесть блоков являются наиболее информативными для изучения функциональных характеристик нейронов.

2.3.2. Различия в активности в ответ на предъявление стимулов различных категорий

Проведен анализ нейронального ответа по отношению к предъявлению стимулов различных категорий. На Рис. 15 изображено распределение ответов в каждой из категорий, а также средний вызванный ответ. Максимальный ответ наблюдается для изображений лиц обезьян (14.93 спайка/с), далее следуют лица людей (12.29 спайка/с) и перевернутые лица (9.51 спайка/с). Последовательное снижение ответа наблюдается для изображений тела обезьян (7.83 спайка/с), других животных (5.58 спайка/с), тела людей (4.84 спайка/с), растений (3.91 спайка/с), искусственных объектов (3.65 спайка/с) и ландшафтов (1.37 спайка/с).

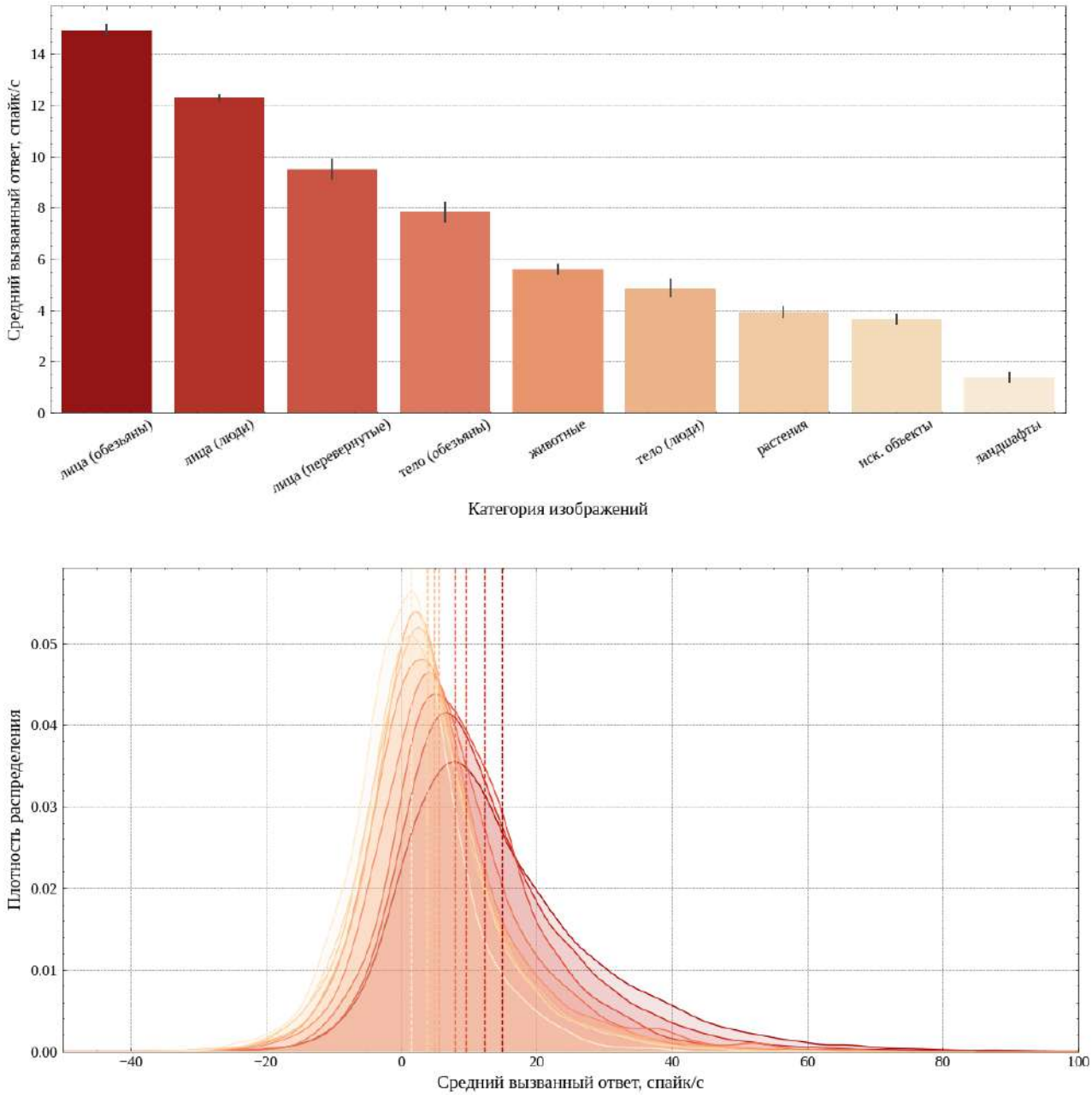

Рисунок 15. Средний вызванный ответ по категориям (вверху) и распределение внутри каждой из категорий (внизу)

Анализ распределений ответов показывает их частичное перекрытие между категориями, при этом категории лиц образуют обособленную группу с более высокими значениями. Сохранение высокого уровня ответа даже на перевернутые изображения лиц указывает на устойчивость механизмов детекции лиц к пространственным трансформациям стимула.

Для количественной оценки избирательности для каждой точки регистрации рассчитан индекс избирательности по отношению к каждой из девяти категорий. Так как категории не сбалансированы (содержат различное количество изображений), была произведена нормализация на размер категории. Затем для каждой точки регистрации был вычислен абсолютный индекс избирательности по каждой категории.

Общее количество точек регистрации с уровнем избирательности $K \geq 0{,}1$ составило 100, с избирательностью $K \geq 0{,}2$ – 59 точек. Максимальная избирательность наблюдается к изображениям лиц обезьян (71 точка) и лиц людей (39 точек), что подтверждает специализацию этой области в обработке лицевой информации. Однако лишь 2 точки регистрации проявляют избирательность выше 0,1 ко всем трем категориям, связанным с лицами (лица обезьян, лица людей, перевернутые лица), что свидетельствует о преимущественной специализации отдельных нейронных популяций.

Выраженная избирательность $K \geq 0{,}2$ отмечена в 7 точках для лиц обезьян, что предполагает существование специализированных нейронных ансамблей, ориентированных на биологически значимые стимулы. В то же время избирательность к искусственным объектам и ландшафтам сопровождается средним отрицательным индексом (-0,1 и -0,14 соответственно), указывая на направленное снижение активности нейронов при предъявлении этих категорий стимулов.

Таблица 1. Избирательность нейронов по различным категориям стимулов

Категория	**K >= 0.1**	**K >= 0.2**	**Средний коэф. K**
лица (люди)	39	0	0.06
ландшафты	100	25	-0.14
растения	64	7	-0.09
животные	41	3	-0.06
иск. объекты	79	10	-0.1
тело (люди)	46	4	-0.07
тело (обезьяны)	34	3	-0.01
лица (обезьяны)	71	7	0.1
лица (перевернутые)	8	0	0.01

В Табл. 1 приведено количество нейронов с индексом избирательности ≥ 0.1 и ≥ 0.2, а также среднее значение индекса избирательности для каждой категории стимулов. Положительные

значения индекса указывают на преимущественное усиление активности нейронов при данной категории, отрицательные – на снижение активности.

На Рис. 16 представлена обратная функция распределения избирательности нейронов по отношению к различным категориям стимулов. Данная функция для каждого значения селективности *x* показывает вероятность того, что нейрон будет иметь селективность больше либо равную *x*. По вертикальной оси отложена вероятность (от 0 до 1), по горизонтальной - значение селективности нейрона. Смещение кривой вправо указывает на более высокую селективность нейронов к соответствующей категории стимулов, то есть их предпочтительную активацию в ответ на предъявление стимулов данной категории. Как видно из графика, для некоторых категорий практически отсутствуют положительно-селективные нейроны.

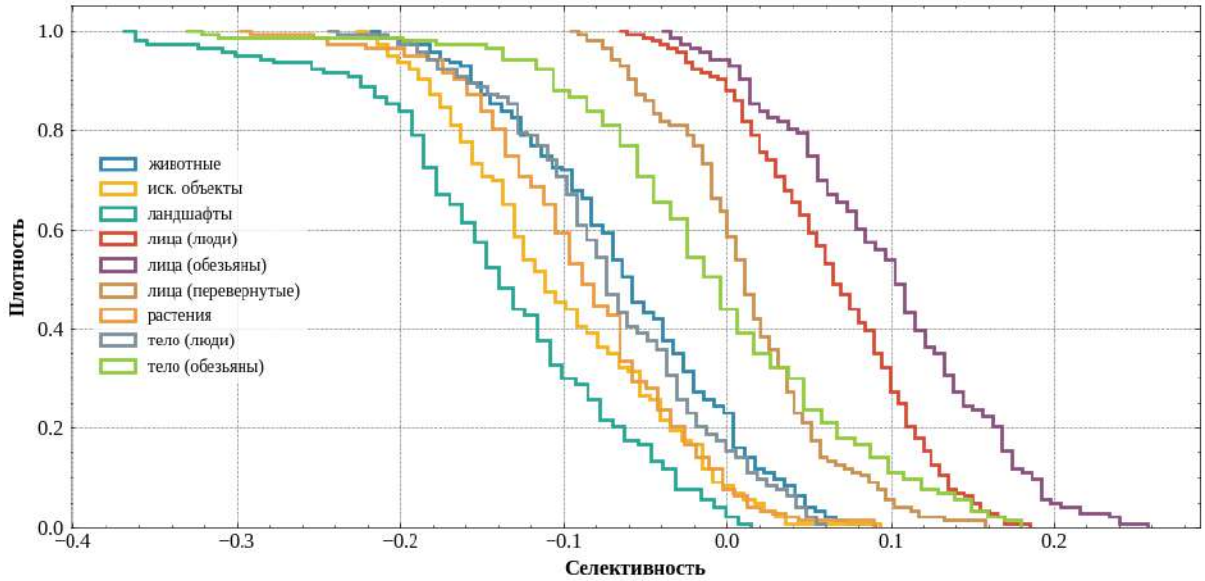

Рисунок 16. Распределение селективности нейронов по категориям стимулов: обратная функция распределения

При этом невозможно утверждать, что нейрональные колонки имеют ярко выраженную избирательность по отношению только к одной из категорий. Для примера рассмотрим одну из точек регистрации *E8* (Рис. 17-19). На Рис. 17 отражена средняя активация по категориям для точки регистрации *E8*. Видно, что нейроны в среднем сильнее отвечают на изображения лиц обезьян, но при этом реагируют в том числе и на изображения тел обезьян и лиц людей.

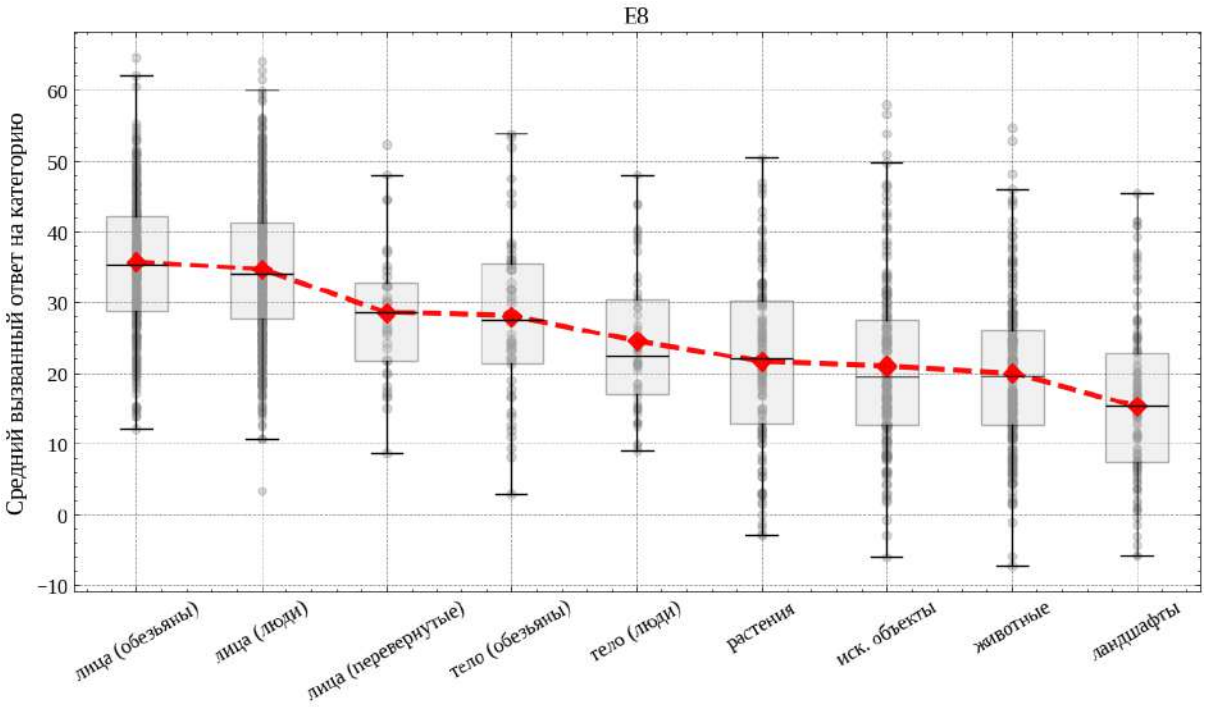

Рисунок 17. Ответ нейрональной колонки E8 на различные категории изображений. Серыми точками обозначены ответы на отдельные изображения, красной пунктирной линией – усредненный по категориям ответ

Для этой же точки регистрации стимулы были проранжированы в соответствии со средним вызванным ответом и отмечены цветом в соответствии с категорией (Рис. 18). Из них для удобства визуализации отобрано только первых 100 значений.

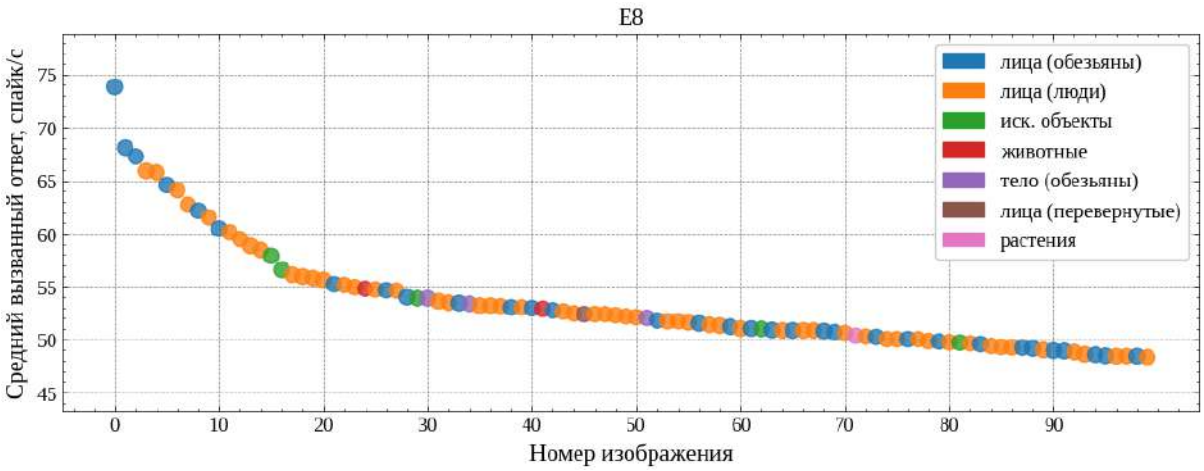

Рисунок 18. Стимулы, отсортированные по среднему вызванному ответу для точки регистрации E8

Несмотря на высокую избирательность к категории лиц, стимулы из других категорий также вызывают активацию. Таким образом, селективность не является исключительной, а семантическая категория объекта не позволяет в полной мере объяснить функциональные характеристик нейрональной колонки.

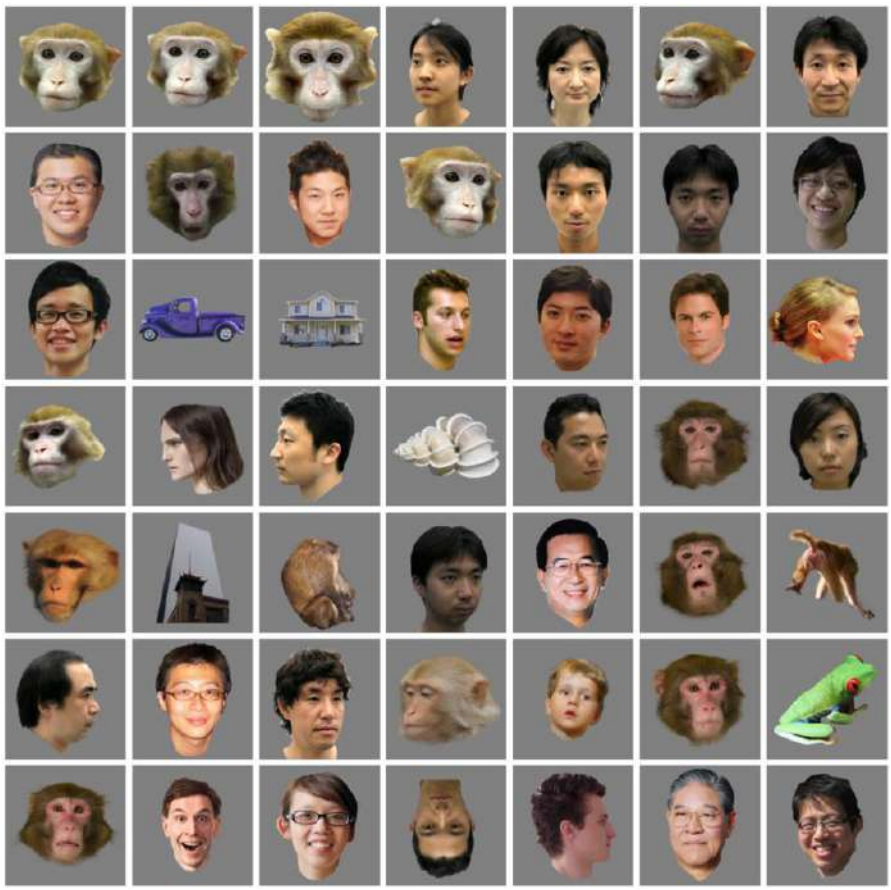

Рисунок 19. Пример 49 стимулов, вызывающих наибольшую среднюю активацию в точке регистрации E8

Полученные данные позволяют построить матрицу схожести категорий (Рис. 20) в пространстве активности нейронов нижневисочной коры. Наибольшая степень схожести активности наблюдается между различными изображениями лиц (лица обезьян, лица людей, перевернутые лица), что подтверждает специализацию данной области на обработку изображений лиц. В то же время, активность, вызванная изображениями животных, растений, искусственных объектов и ландшафтов, негативно коррелирует с активностью, вызванной лицами, что свидетельствует о раздельной обработке этих классов стимулов. Интересно, что внутри этих групп наблюдается высокая внутригрупповая корреляция, указывающая на их меньшую дифференциацию нейронными популяциями.

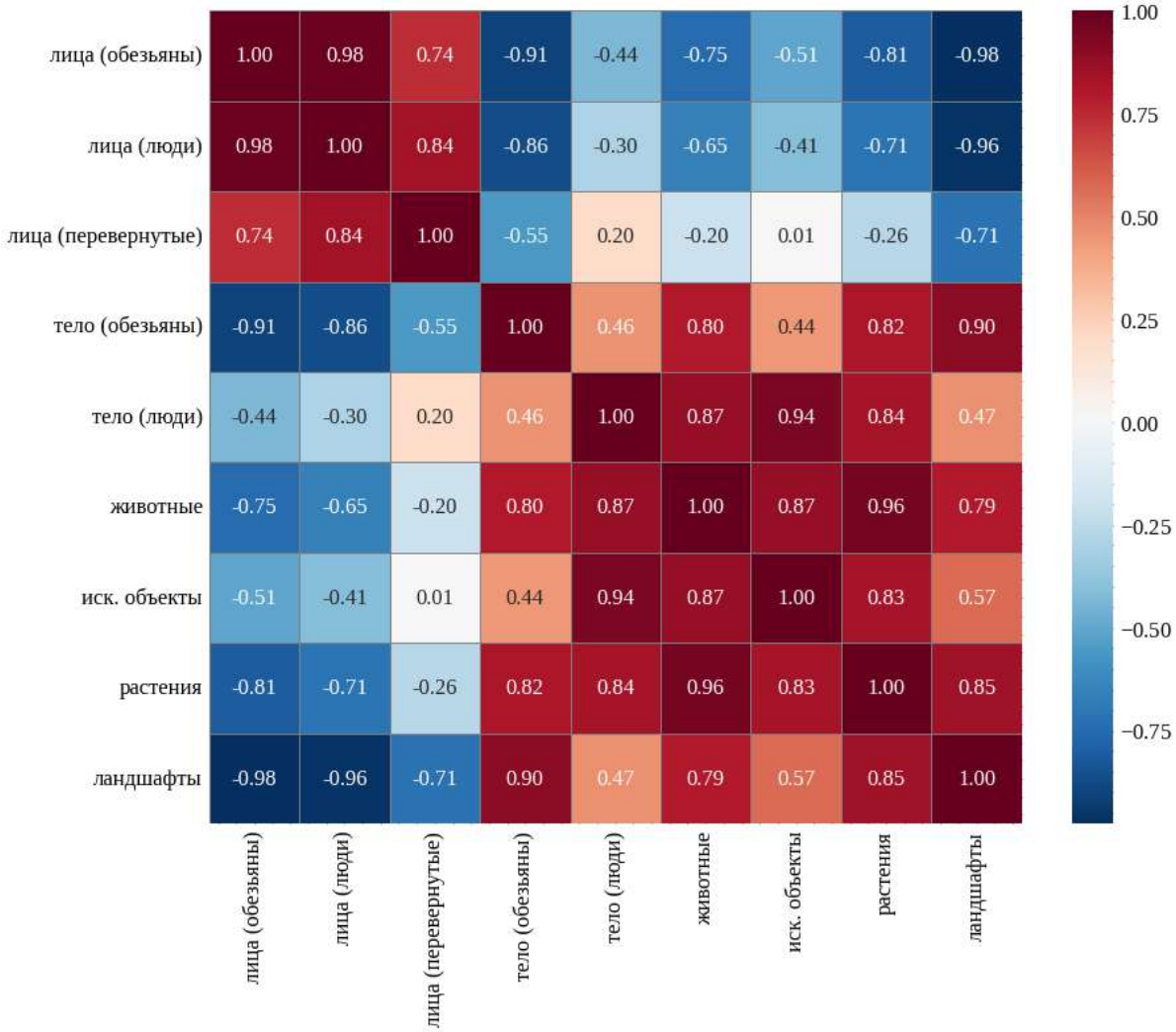

Рисунок 20. Матрица схожести категорий на основе усредненного по нейрональным колонкам ответа на категорию

Обсуждение результатов

В данной главе представлен процесс сбора нейрофизиологических данных и проведен статистический анализа импульсной активности нейронов нижневисочной коры при предъявлении зрительных стимулов. Основные результаты:

Проведен анализ отклика нейрональных ансамблей в 143 точках записи, расположенных в области TEad нижневисочной коры, у трех макак резус в ответ на предъявление 1550 изображений, принадлежащих к 9 категориям.

Анализ нейронных ответов показал, что большинство зарегистрированных нейронов *демонстрируют усиленную активность при предъявлении стимулов с изображениями лиц* по сравнению с другими образами. Средний вызванный ответ был выше для лиц у 130 из 143 нейронов, а стандартное отклонение – у 118 нейронов, что указывает на повышенную модуляцию активности. Нейрональные колонки с высоким уровнем средней активации демонстрируют

расширенный диапазон ответов, что проявляется в положительной связи между средней вызванной активностью и ее вариативностью (коэф. корр. = 0,58, $p \leq 0,001$).

Расширенный диапазон ответов, может отражать более высокую чувствительность к различиям в свойствах стимулов и свидетельствовать об участии нейронов в кодировании более широкого спектра признаков, обеспечивая большую гибкость в обработке зрительной информации. Подобный эффект наблюдается в системах с избыточным кодированием, где высокая вариативность позволяет нейронной сети динамически перестраивать свои ответы в зависимости от входного сигнала.

Показано, что *пространственная организация нейрональной активности проявляет закономерности, характерные для локально организованных ансамблей*. Установлено, что корреляция нейронной активности (коэф. корр. = 0,41, $p \leq 0,001$) снижается с увеличением расстояния между точками регистрации по нелинейной зависимости, описывающейся степенным (R^2=0.25) либо же логарифмическим законом (R^2=0.24). Это указывает на наличие структурированной, иерархически организованной нейронной активности в TEad, что согласуется с наблюдениями в других областях коры (Bassett, Bullmore, 2006; Yu et al., 2011).

Сопоставимая эффективность описания полученных данных как логарифмической, так и степенной зависимостью отражает давнюю научную дискуссию, продолжающуюся уже более полувека. Этот вопрос остается открытым, поскольку обе модели способны адекватно описывать нейронные отклики, но различаются в интерпретации механизмов их формирования. Логарифмическая зависимость традиционно используется для объяснения компрессии сенсорных сигналов в нейросетях, тогда как степенная функция лучше описывает нелинейные процессы масштабирования активности и отражает постепенное затухание связи без выхода на плато. Наблюдаемое равенство в их предсказательной способности подчеркивает необходимость дальнейшего анализа.

Полученные результаты перекликаются с идеей распределенного кодирования информации в высших отделах зрительной коры: несмотря на ослабление синхронизации на больших расстояниях, функциональная связь между удаленными участками сохраняется, позволяя координировать обработку сложных зрительных стимулов. В этом контексте можно предположить, что межнейронные взаимодействия формируют сложную систему, в которой поддерживается баланс между локальной специализацией и глобальной координацией.

Нижневисочная кора демонстрирует выраженную категориальную избирательность, при этом 100 из 143 точек регистрации (70%) проявляют значимую избирательность ($K \geq 0,1$), а у 59 точек (41%) полученный индекс выше 0,2. Максимальная избирательность наблюдается к лицам обезьян (71 точка) и людей (39 точек), что подтверждает специализацию этой области в обработке информации о лицах, однако только 2 точки превышают порог $K \geq 0,1$ для всех трех

категорий, связанных с лицами, указывая на селективность к определенным типам изображений лиц. Выраженная избирательность (K ≥ 0,2) отмечена в 7 точках для лиц обезьян, что ожидаемо, учитывая, что исследование проводилось на обезьянах, и их зрительная система эволюционно адаптирована к обработке социально значимых стимулов.

Ответ нейрональных колонок на искусственные объекты и ландшафты описывается негативной избирательностью (значения индекса K = -0,10 и -0,14 соответственно) по сравнению с усредненным ответом на другие категории. В частности, 10 точек регистрации (K ≥ 0.2) демонстрируют отрицательную избирательность к искусственным объектам, а 25 точек (K ≥ 0.2) к натуральным сценам (ландшафтам).

Выявленная выраженная категориальная избирательность нейронов нижневисочной коры подтверждает ее ключевую роль в обработке информации о лицах. Тот факт, что 70% точек регистрации демонстрируют высокую категориальную избирательность, что значительная часть популяции нейронов данной области участвует в обработке зрительных категорий. Важно отметить, что избирательность нейронов носит градуальный, а не бинарный характер.

При этом выявленная отрицательная избирательность нейронов нижневисочной коры к искусственным объектам и ландшафтам указывает на отсутствие либо подавление нейронной активности в ответ на стимулы этих категорий. Такой эффект может отражать механизм активной динамической фильтрации информации, не относящейся к ключевым категориям, обрабатываемым данной областью.

Подобное снижение ответа могло бы быть объяснено конкурентным взаимодействием между нейронными ансамблями. Если нейроны, избирательно реагирующие на лица, активируются, это может сопровождаться угнетением активности в соседних областях, не вовлеченных в обработку этой категории, и наоборот. Подобная динамика на уровне популяции позволила бы проводить перераспределение ресурсов в коре, а также модулирование обработки информации на уровне крупномасштабных сетей.

В ходе эксперимента выявлено постепенное снижение как среднего вызванного ответа, так и различий между категориями стимулов, что может свидетельствовать об адаптации зрительной системы. Однако этот эффект также может быть обусловлен утомляемостью, снижением внимания или другими факторами. В связи с этим временной интервал, в течение которого можно надежно анализировать категориальную избирательность, оказывается ограниченным. Наиболее информативными для исследования оказались первые шесть экспериментальных блоков, за исключением, возможно, первых одного-двух блоков, где мог проявляться эффект новизны.

Полученные результаты согласуются с текущими представлениями о функции нижневисочной коры как области, ответственной за распознавание лиц и придающей приоритет биологически и социально значимым стимулам, что подтверждает ее центральную роль в обработке зрительной информации. Выбор нейрональной колонки как базовой единицы анализа позволил изучить функциональную организацию коры на мезоскопическом уровне, рассмотреть возможное влияние как локальной специализации, так и расширенной динамики нейронной сети.

Интересным продолжением работы могло бы стать изучений возможной внутренней структуры категориальной избирательности. В частности, определение, существует ли определенная организация избирательности к различным категориям стимулов. В настоящее время многие исследования фокусируются на бинарной избирательности (например, индекс избирательности к лицам), однако в наблюдаемых нейронных колонках может прослеживаться более сложная, градиентная избирательность к различным категориям, таким как тела, объекты или сцены.

Дополнительный анализ мог бы выявить сниженную, но систематическую избирательность к определенным категориям, которая остается незамеченной при бинарном разделении стимулов. Например, если категории лицо и тело демонстрируют более высокий индекс избирательности по сравнению с другими категориями, но при этом оказываются ближе друг к другу в нейронном представлении, чем лицо и ландшафт, это могло бы указывать на скрытую иерархию категориального кодирования. Подобное исследование могло бы помочь выявить принципы представления дополнительной (контекстуальной) информации в нейронных сетях, которая не проявляется в явной селективности, но может играть важную роль в обработке сигнала.

Для дальнейшего понимания принципов обработки информации необходимо интегрировать экспериментальные данные с методами машинного обучения, а также исследовать временную динамику нейрональных ответов и взаимодействие различных отделов зрительной коры. Такой комплексный подход позволит глубже понять механизмы работы высших отделов зрительной системы, раскрыть закономерности адаптации и перераспределения нейронных ресурсов, а также создать более точные модели обработки зрительной информации.

Глава 3. Исследование функции зрительной коры головного мозга при помощи комплекса сверточных и генеративно-состязательных сетей

В **Главе 3** представлено исследование возможностей применения сверточных и генеративно-состязательных сетей для изучения функций нейронов высших областей зрительной системы. Исследуются возможности применения сверточных и генеративно-состязательных нейронных сетей для изучения функций нейронов нижневисочной коры приматов. В основе подхода лежит моделирование нейрональных ответов с использованием глубоких сверточных сетей, анализируемых методами интерпретации искусственного интеллекта. Генеративно-состязательные сети применяются для поиска стимулов, вызывающих специфические реакции в моделируемых нейронах, что позволяет изучать их избирательность.

Подход предполагает создание изображений, направленных на максимизацию активации целевых нейронов (модельных или биологических). Нахождение такого изображения, в свою очередь, помогло бы понять функциональное назначение нейрона. Создание стимула начинается с генерации случайного вектора в пространстве признаков. Вектор затем переводится в изображение и предъявляется наблюдателю, регистрируется ответ. Затем осуществляется пошаговое изменение вектора и, соответственно, стимула таким образом, чтобы усилить признаки, наиболее значимые для целевого нейрона, что приводит к постепенному формированию характерного паттерна. Этот процесс позволяет выявить особенности избирательности нейрона, показывая, какие элементы структуры или текстуры оказывают наибольшее влияние на его активацию.

Реализация метода визуализации при помощи генеративно-состязательных сетей предполагает оптимизацию стимула на основе ответа наблюдателя. В технических науках этот процесс осуществляется через сверточные нейронные сети или другие модели, непосредственно оптимизирующие подаваемый входной сигнал. Такая оптимизация может выполняться быстро, проходя большое количество итераций. Однако в реальных экспериментах количество предъявляемых наблюдателю стимулов существенно ограничено. Поэтому целесообразным является введение прокси-модели, обученной на ответах нижневисочной коры и представляющей собой аппроксимацию нейрональных колонок этой области мозга. Создание стимулов осуществляется посредством взаимодействия генеративных моделей и моделей наблюдателя, после чего созданный материал предъявляется непосредственно биологическому наблюдателю. Такой подход позволяет значительно ускорить процесс экспериментального поиска и сократить количество необходимых предъявлений живому субъекту.

3.1. Моделирование ответа нейронов нижневисочной коры

Для задачи моделирования использовались данные, описанные в Главе 2, отражающие активность 143 кортикальных колонок нижневисочной коры у макак. Так как изображения были представлены в 12 блоках, для получения общего нейронального ответа использовался усредненный по блокам отклик.

Данные были случайным образом разбиты на тренировочное и тестовое множества, содержащие 1240 и 310 наблюдений соответственно. Эффективность обучения модели оценивалась на тестовом множестве. По результатам предъявления тестовых стимулов оценивалась возможность предсказывать ответ в каждой отдельно взятой нейрональной колонке.

Архитектура модели состояла из сверточного блока и надстройки из полносвязных слоев (Рис. 21). Сверточный блок включал в себя часть предобученных слоев сети VGG16 и служил для извлечения высокоуровневых признаков.

Количество предобученных слоев и место усечения сети определись схожестью признаков, кодируемых в слое, с ответом популяции нейронов. Так как нейрофизиологические данные регистрировались в области, отвечающей за распознавание лиц, было проведено сопоставление двух моделей архитектуры: классической VGG, обученной на наборе данных ImageNet, и специализированной VGGFace, натренированной на распознавании лиц.

На указанные сверточные слои надстраивался полносвязный блок, состоящий из отдельных подсетей в количестве, равном количеству кортикальных колонок, чья деятельность моделировалась сетью (см. Рис. 21). Создание множественных подсетей позволило снизить корреляцию между предсказываемыми значениями, что в большей мере соответствует статистике нейробиологических данных.

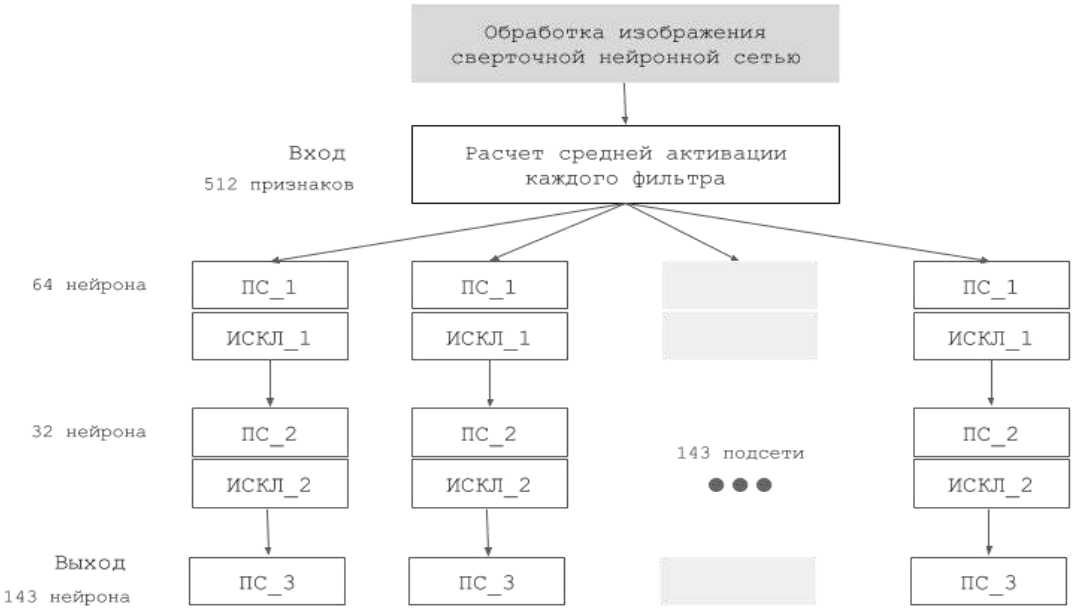

Рисунок 21. Схема архитектуры сети для предсказания вызванной активации. ПС – полносвязный слой, ИСКЛ – слой исключения

Всего было проведено 200 эпох обучения, в каждой из который модель проходила через полный тренировочный набор данных.

По результатам предъявления тестовых стимулов оценивалась возможность предсказывать ответ в каждой из 143 моделируемых колонок. Коэффициент корреляции в среднем по всем колонкам составил 0,68, *p*-значение < 0,01 (Рис. 22, слева). Однако качество прогнозирования варьируется (Рис. 22, справа) и существенно зависит (коэф. корр. = 0,7, *p*-значение < 0,01) от стабильности активации нейронов в ответ на повторное предъявление, измеряемой как корреляция ответов между четными и нечетными блоками.

Детальный анализ распределения коэффициентов корреляции для слоя *block5_conv2* (Рис. 23) показывает значительную вариабельность качества предсказаний для разных колонок, со средним значением 0,45. При этом выделяются колонки с особенно высокой точностью предсказания: E1 (0,82), E2 (0,80), H8 (0,76), M5 (0,73) и E7 (0,72). Такой разброс в точности предсказаний может быть обусловлен как различиями в стабильности ответов самих нейронов, так и особенностями их функциональной специализации. Наличие колонок с высокими коэффициентами корреляции (>0,7) свидетельствует о принципиальной возможности точного моделирования нейрональных ответов с помощью искусственных нейронных сетей для определенных популяций нейронов.

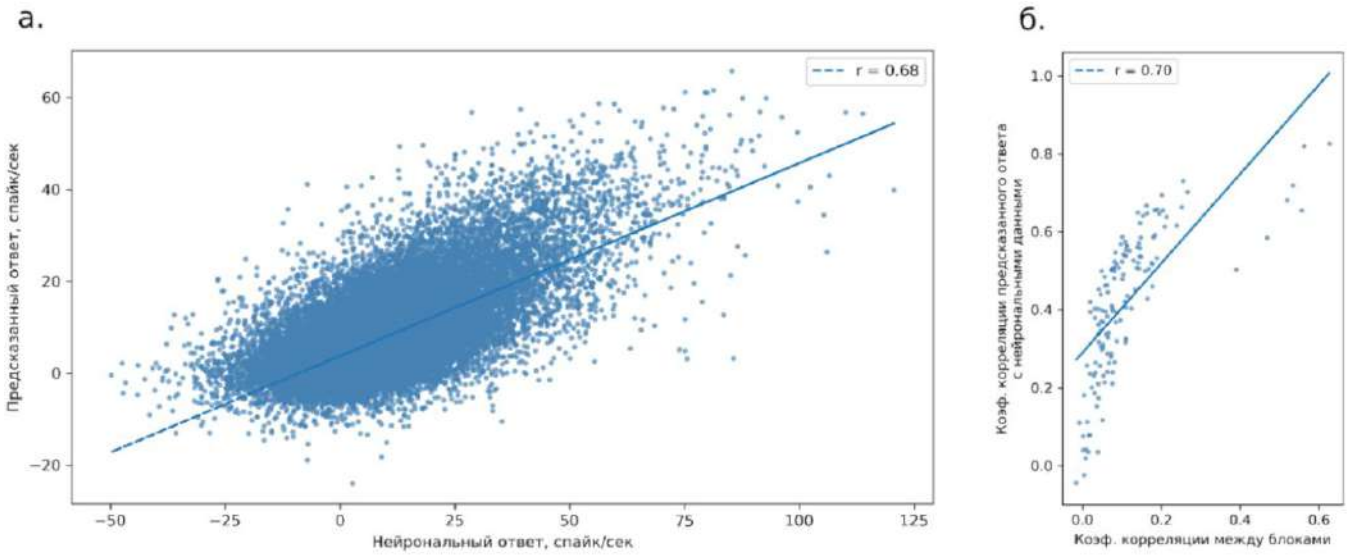

Рисунок 22. Эффективность модели на тестовых данных. а) Сопоставление предсказанного значения вызванной активации с действительным значением на тестовых данных для всех кортикальных колонок. По оси абсцисс – усредненный по изображению нейрональный ответ (количество спайков в секунду), по оси ординат – предсказанная активация (количество спайков в секунду). б) Зависимость качества предсказания от стабильности ответов нейронов: по оси абсцисс указано значение корреляции регистрируемого нейронального ответа, посчитанное между четными и нечетными блоками эксперимента, по оси ординат – корреляция предсказанных сетью ответов с оригинальными данными

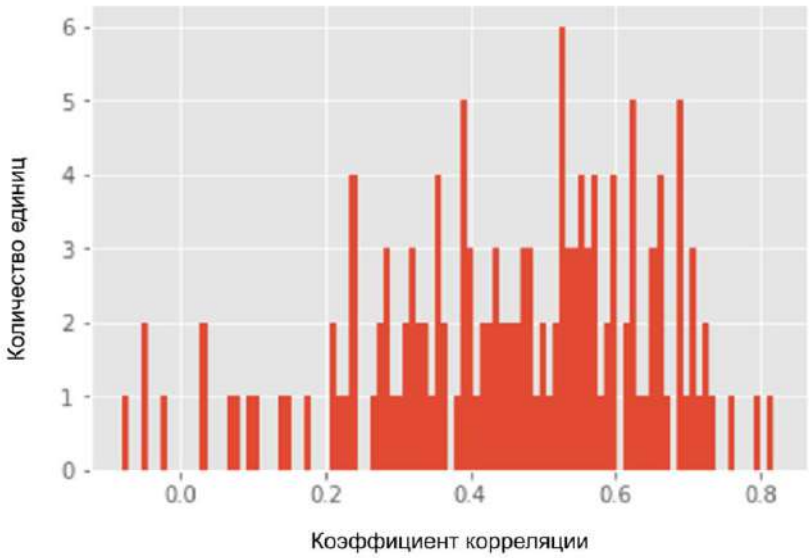

Рисунок 23. Распределение коэффициента корреляции предсказанного ответа с нейрональным для различных колонок (точек регистрации)

После обучения модели был проведен анализ ее свойств методами, применяемыми для интерпретации систем искусственного интеллекта: в частности, функции отдельного нейрона визуализировались через набор стимулов, вызывающих его активацию. Это осуществлялось посредством генерации оптимального входного сигнала, что снимало ограничения, связанные с

пространством поиска, как в случае с перебором изображений из базы данных. Для создания изображений использовалась система Plug-n-Play Generative Networks (PPGN) (Nguyen et al., 2016). В отличие от других подходов к визуализации (Mordvintsev, Olah, Tyka, 2015; Mahendran, Vedaldi, 2016; Olah, Mordvintsev, Schubert, 2017), данный метод, помимо уровня активации выбранного нейрона, учитывает также требование к натуралистичности результата, что позволяет избежать искусственных стимулов, при которых нейронная сеть с высокой уверенностью детектирует объект в изображениях, расцениваемых человеком как шум, либо как объект, очевидно принадлежащий другому классу (Szegedy et al., 2014; Nguyen, Yosinski, Clune, 2015).

3.2. Применение генеративно-состязательных сетей в целях создания оптимального стимула для модели кортикальной колонки

Возможно выделить два подхода в исследовании функции нейронов при помощи генеративно-состязательных сетей – постэкспериментальный анализ (офлайн) и анализ во время проведения эксперимента (онлайн).

Офлайн-подход предусматривает три шага:

1. Обучение модели. Для этого на данных нейрофизиологического эксперимента создается модель, позволяющая с высокой точностью предсказывать вызванный ответ для определенных кортикальных колонок.
2. Создание изображений. При оперировании в высокоуровневом пространстве описания производится поиск изображений, вызывающих определенный отклик у обученной модели.
3. Интерпретация результатов. Найденные изображения могут анализироваться на предмет наличия признаков как в пространстве изображений, так и в высокоуровневом пространстве описания.

Онлайн-подход является продолжением офлайн-подхода и возможен к применению во время самой экспериментальной сессии. В этом случае добавляются дополнительные этапы:

1. Обучение модели.
2. Создание изображений.
3. Предъявление изображений и регистрация нейронального ответа. Предъявление набора изображений позволяет лучше понять, какие признаки вызывают активацию нейронов.
4. Промежуточная интерпретация результатов и корректировка работы модели. Собранные нейтральные ответы могут использоваться для переобучения модели и ее корректировки с учетом вновь полученных данных.

5. Интерпретация результатов.

Шаги 1–4 могут повторяться в цикле до окончания времени, отведенного на экспериментальную сессию.

Полученная модель, таким образом, имитирует поведение отдельных кортикальных нейронов или групп нейронов *in silico*. Затем модель исследуется путем поиска оптимального стимула, вызывающего требуемый отклик. Для этого используются: модель, генерирующая изображения (деконволюционная нейронная сеть, обученная по принципу генеративно-состязательных сетей), и модель, ограничивающая пространство поиска для создания естественно выглядящих изображений (сверточная нейронная сеть). Общая схема решения представлена на Рис. 24, процесс поиска латентного кода приведен на Рис. 25.

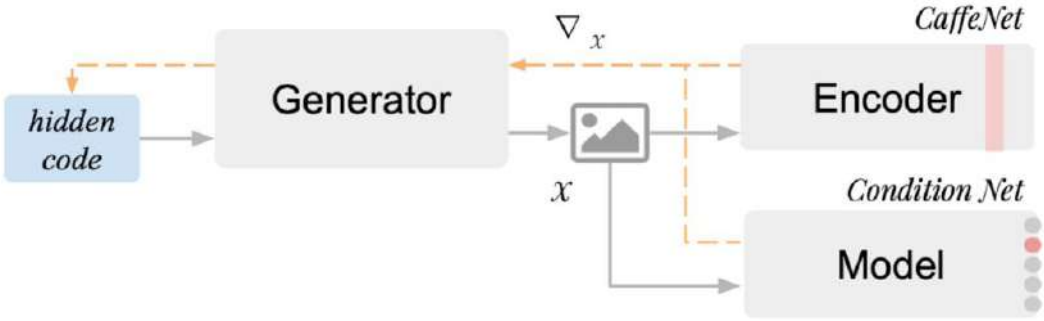

Рисунок 24. Схема решения для исследования функции нейронов. В процессе генерации участвует комплекс нейросетей, состоящий из модели-генератора Generator (шумоподавляющего автокодировщика), сверточной нейронной сети Encoder (кодировщика) и целевой нейронной сети Model

Генеративная модель (Generator) создает изображения на основе полученного вектора (hidden code), изображение подается в исследуемую сеть (Model) и сеть, выступающую в роля регуляризатора (CaffeNet, слой fc6). В основе подхода лежит сэмплирование из совместного распределения, задаваемого двумя моделями:

$$p(x, y) = p(x)p(y|x),$$

где $p(y|x)$ представляет результат активации выбранного слоя в модели-объекте визуализации, а $p(x)$ локализует область поиска, обеспечивая создание естественно выглядящих изображений.

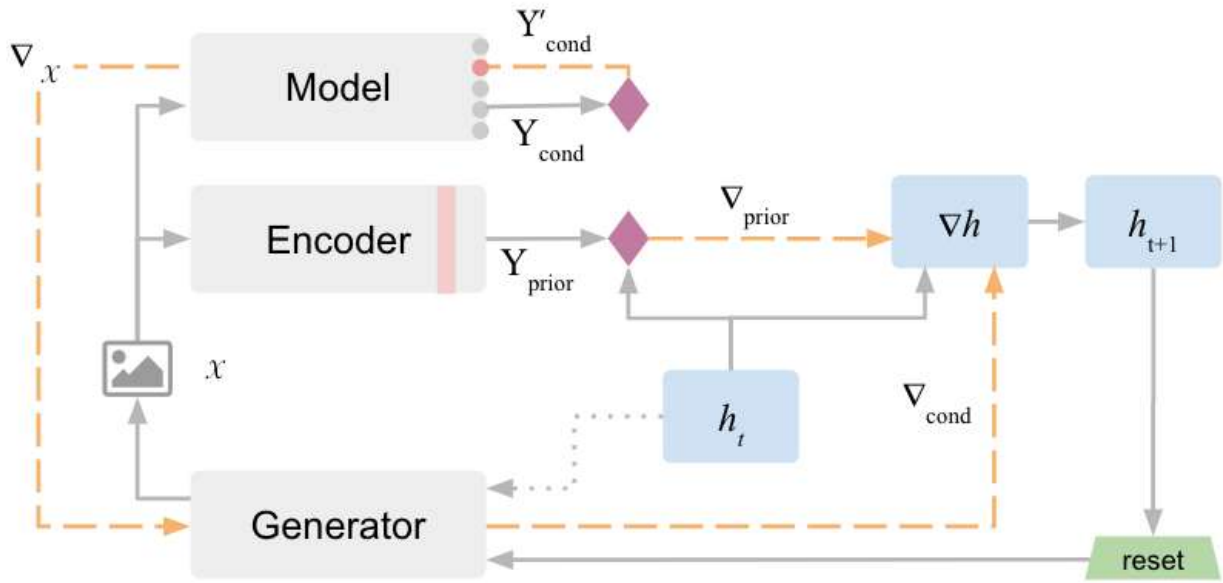

Рисунок 25. Детальная схема поиска латентного кода h для активации выбранного нейрона

Детальное описание итерации оптимизации включает следующие шаги:

1. На вход модели Generator подается вектор значений h_t, содержащий случайное высокоабстрактное представление изображения. На основе h_t при помощи предобученных слоев генератора, выполняющих операцию деконволюции, создается RGB-изображение x.
2. Изображение x пропускается через модель Encoder, выступающую в роли регуляризатора. Отслеживается активация в выбранном слое регуляризатора Y_{prior}. Кодирование признаков должно приблизительно соответствовать уровню кодирования признаков в модели Model, так как вектор h будет постоянно смещаться в сторону регуляризационного пространства.
3. Одновременно с этим изображение x пропускается через целевую модель зрительной коры Model, в выходном слое которой регистрируется активация, Y_{cond}. Рассчитывается Y'_{cond} как $Y'_{cond} = \log(\mathrm{p}(\mathrm{y} = y_c|x))$, либо через другую заданную функцию, определяющую желаемый паттерн активации нейронов.
4. Вычисляется градиент для модели-регуляризатора:

$$\nabla_{prior} = d\log(p(h)) / dh = Y_{prior} - h.$$

5. Вычисляется градиент для изображения: $\nabla_x = d\,Y'_{cond} / dx$.
6. Вычисляется градиент для целевой модели: $\nabla_{cond} = \nabla_x / dh$.

7. Вычисляется градиент для кода h:

$$\nabla h = \epsilon_1 \nabla_{prior} + \epsilon_2 \nabla_{cond} + N,$$

где ϵ_1 и ϵ_2 – коэффициенты, определяющие вес регуляризатора и условия; N – случайный шум.

8. Последним шагом рассчитывается следующая итерация кода h:

$$h_{t+1} = h + lr / mean(\nabla h),$$

где lr – скорость обучения (learning rate).

Целевая модель была выбрана с учетом ее способности отражать нейронные процессы в нижневисочной коре, опираясь на корреляцию ее внутренних представлений с нейронными данными. Особое внимание уделялось выбору архитектуры, предобученных весов и метода агрегации активаций, что позволило добиться более точного соответствия нейронным откликам (см. Табл. 2).

Таблица 2. Основные компоненты целевой модели Model и их краткое описание

Архитектура базовой модели	VGG16	Была выбрана модель VGG16, которая показывает высокую точность в категоризации объектов, имеет простую архитектуру и во многом аналогична модели AlexNet (CaffeNet), применяемой как регуляризатор в PPGN
Веса	ImageNet	Использовалась VGG16, обученная на наборе данных ImageNet, так как было обнаружено больше элементов, ответы которых коррелирует с нейронными данными, по сравнению с той же моделью, но обученой на наборе данных Celeb A (VGGFace – модель, обученная на восприятие лиц)
Слой усечения	*block5_conv2*	Скрытые слои *block4_conv3, block5_conv1, block5_conv2* показали наибольшую корреляцию с нейрональными данными, что может свидетельствует о том, что указанные слои имеют сопоставимое по угловому охвату рецептивное поле и сложность выделяемых признаков
Функция агрегации активации слоя	Максимальное значение для карты активации	Нейроны нижневисочной коры имеют большое рецептивное поле, покрывающее значительную площадь стимульных изображений. Таким образом, для каждого сверточного фильтра стоит учитывать карту активации целиком и в качестве функции усреднения выбирать максимальное значение, так как оно отражает предпочтения фильтра лучше, чем среднее или значение центрального элемента
Дополнительные слои	Отдельные подсети	Для каждой нейрональной колонки (из 143) тренировалась отдельная подсеть, базирующаяся на выбранном слое усечения и предсказывающая нейрональный ответ

Выбор слоя для моделирования ответа нейрональных колонок. Для выбора слоя, наиболее соответствующего обработке информации в нейронных колонках, были проанализированы корреляции между предсказанными и истинными нейронными откликами на различных уровнях сверточной сети. На Рис. 26 показаны коэффициенты корреляции между активациями слоев модели и экспериментально зарегистрированными нейронными ответами. Наибольшее соответствие наблюдается для слоев block5_conv3 и block5_conv2 (коэф. корр. = 0,68), что свидетельствует о схожести их представлений с этапами обработки зрительных стимулов в нижневисочной коре. На основе этих данных была произведена усеченная версия модели, использующая выбранный слой для дальнейшего предсказания нейронных ответов.

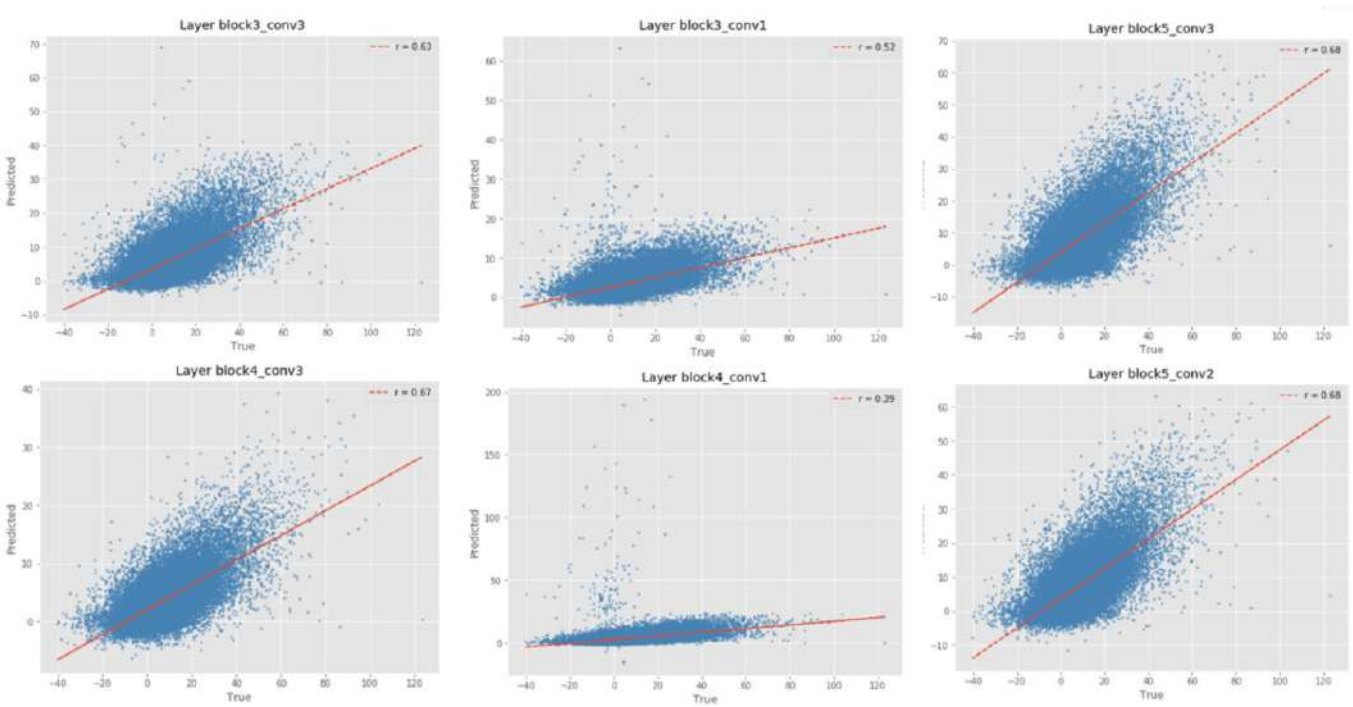

Рисунок 26. Корреляция ответов нейронов и слоев искусственной нейросети

На Рис. 27 представлены матрицы сходства изображений, вычисленные на основе корреляции нейронных ответов (слева) и активаций обученной модели (справа). Красные области отражают положительную корреляцию между реакциями на пары изображений, тогда как синие – отрицательную. Сравнение двух матриц позволяет оценить, насколько глубинная нейросеть воспроизводит структуру восприятия стимулов, характерную для нейронов. Визуальное сходство паттернов корреляции указывает на способность модели улавливать особенности представления объектов, аналогичные тем, что наблюдаются в нейронных данных.

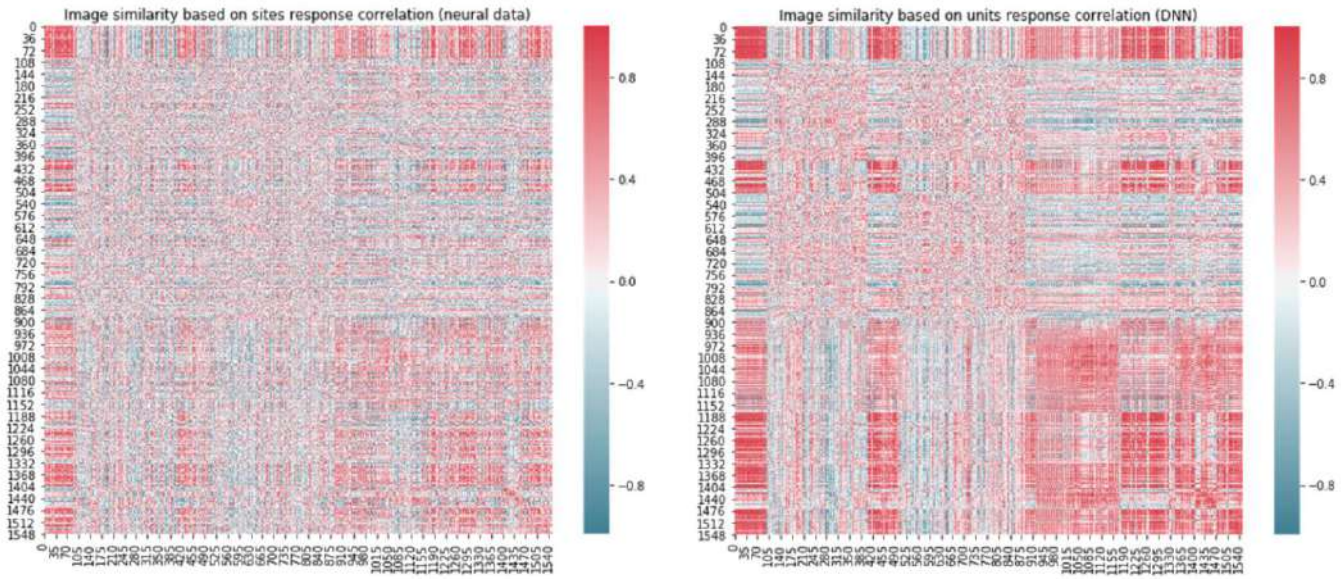

Рисунок 27. Матрица сходства изображений: на основе нейрональных ответов (слева) и на основе ответов обученной модели (справа)

Функция ошибки. Функция ошибки, также известная как функция потерь или целевая функция, используется в задачах оптимизации, включая обучение нейронных сетей и генерацию изображений с помощью генеративных моделей, для оценки расхождения между текущим и желаемым результатом. В данном контексте она измеряет степень соответствия определенной точки в латентном пространстве искомому решению. Оптимизация функции ошибки, например с помощью градиентного спуска или других методов, направлена на поиск в скрытом пространстве, чтобы минимизировать или максимизировать значение целевой функции, приближая модель к желаемому результату.

Целевая функция как в моделировании, так и в создании стимулов во время экспериментальной сессии может задаваться произвольно. Например, исследователю может быть интересно исключительно активация одного выбранного нейрона, либо же, в ином случае, интерес представляет определенный паттерн активации, вызываемый на наборе регистрируемых нейронов, так называемый контроль ответа популяции нейронов (Bashivan, Kar, DiCarlo, 2019).

В данной работе было рассмотрено и экспериментально реализовано два подхода:

- *Evoked*: максимизация ответа целевого нейрона и игнорирование активации на других регистрируемых нейронах.
- *Softmax*: выборочная активация одного из нейронов, при этом минимизация ответа на других нейронах.

Необходимо отметить, что стратегия *Evoked* целесообразна, когда необходимо изучить характеристики ответа отдельного нейрона. В то же время подход *Softmax* не только стремится к

активации выбранного нейрона, но и одновременно подразумевает отсутствие активации в других регистрируемых зонах. Это достигается за счет функции softmax, которая перераспределяет входные данные таким образом, что одно значение (или несколько) оказывается значительно больше остальных, а сумма всех значений остается равной единице. *Softmax* подходит для задач, требующих выявления различий в функциях избирательности наблюдаемых нейронов.

В процессе оптимизации случайного стимула (Рис. 28) осуществляется его пошаговое видоизменение, направленное на максимизацию активации модельного нейрона. На каждом этапе оптимизации изображение модифицируется таким образом, чтобы усилить признаки, наиболее значимые для данного нейрона, что приводит к постепенному формированию характерного паттерна. Этот процесс позволяет выявить особенности избирательности нейрона, показывая, какие элементы структуры или текстуры оказывают наибольшее влияние на его активацию. В результате визуализация полученного стимула отражает предпочтительные параметры нейрона-детектора.

Рисунок 28. Пример пошаговой оптимизации изображения с целью максимизации ответа выбранного элемента модели нижневисочной коры

Приведем пример оригинальных экспериментальных стимулов и сгенерированных изображений, ассоциирующихся с активацией кортикальных колонок № 25 и № 27 (Рис. 29). Заметно, что сгенерированные стимулы фокусируются преимущественно на глазах, а также на абстрактных формах, которые, вероятно, отражают другие признаки, выделяемые рассматриваемым нейрональным участком. Например, во всех стимулах, созданных для колонки

№ 25 (Рис. 29 а, б) прослеживается наличие диагональной линии, направленной с левого верхнего угла в правый нижний.

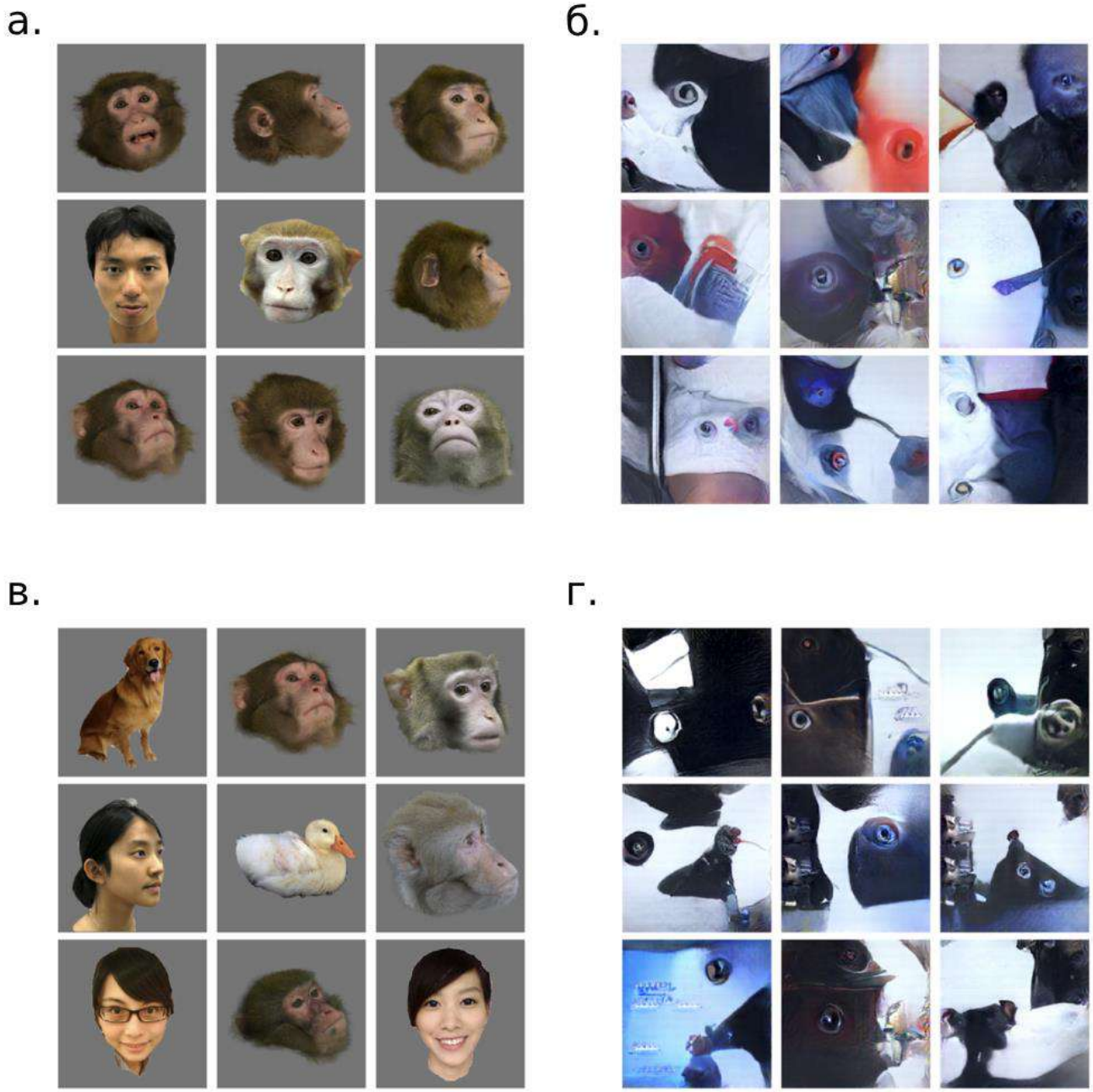

Рисунок 29. Сопоставление стимулов, вызывающих высокую активацию нейрональной колонки и изображений, сгенерированных при помощи модели для интерпретации функций нейрона: а) топ-9 экспериментальных стимулов для кортикальной колонки № 25; б) топ-9 сгенерированных изображений, вызывающих реакцию в № 25; в) топ-9 экспериментальных стимулов кортикальной колонки № 27; г) топ-9 сгенерированных изображений, вызывающих реакцию в № 27

Таким образом в офлайн-подходе показано как предложенный метод выделяет признаки, вызывающие высокий уровень активации целевого нейрона, и создает изображения, обладающие данными признаками.

3.3. Генерация стимульного материала во время экспериментальной сессии

Предложенный подход к моделированию нейронального отклика и визуализации кодируемых признаков может быть применен как в пост-экспериментальном анализе, так и во время экспериментальной сессии. В качестве следующего шага валидации предложенного подхода был проведен эксперимент, в котором модель обучалась на данных основного блока с естественными изображениями. Затем обученная модель использовалась для генерации стимульного материала в дополнительных блоках экспериментальной сессии. Если обученная модель действительно эффективно предсказывает нейронные отклики, а метод визуализации с использованием генеративно-состязательных сетей применим к данной модели, то стимулы, полученные в результате такой визуализации, при предъявлении их животному в той же экспериментальной сессии, должны вызывать усиленный нейрональный отклик.

Регистрация нейрональной активности в ответ на предъявление сгенерированных изображений производилась в передней части нижневисочной коры (TEad) макаки при помощи массива Utah, включающего 8 электродов (Site 1–8), каждый из которых содержал 8 каналов. Для анализа полученные сигналы усреднялись по каналам каждого электрода, а затем преобразовывались в вызванный ответ.

Стимульный материал предъявлялся в трех экспериментальных блоках (12 повторений для каждого стимула). В первом блоке использовались исключительно естественные изображения из набора Takayuki-1550 (1509 изображений). В последующих двух блоках предъявлялись как естественные (из основного блока), так и сгенерированные изображения. После каждого блока проводилось обучение модели на основе собранных данных, а затем созданные моделью стимулы использовались для предъявления в следующем блоке. Таким образом, стимулы второго блока включали изображения, полученные после обучения модели на первом блоке, а стимулы третьего блока формировались на основе обновленной модели после второго блока.

Время презентации стимула составляло 100 мс, затем следовала фиксация 200 мс. Начальные 50 мс до и 50 мс после вывода стимула на экран учитывались как период спонтанной активности (базовый период), период с 70 мс до 220 мс после появления стимула рассматривался как период вызванного нейронального ответа. Более подробно о методике предъявления стимульного материала изложено в разделе 2.1.

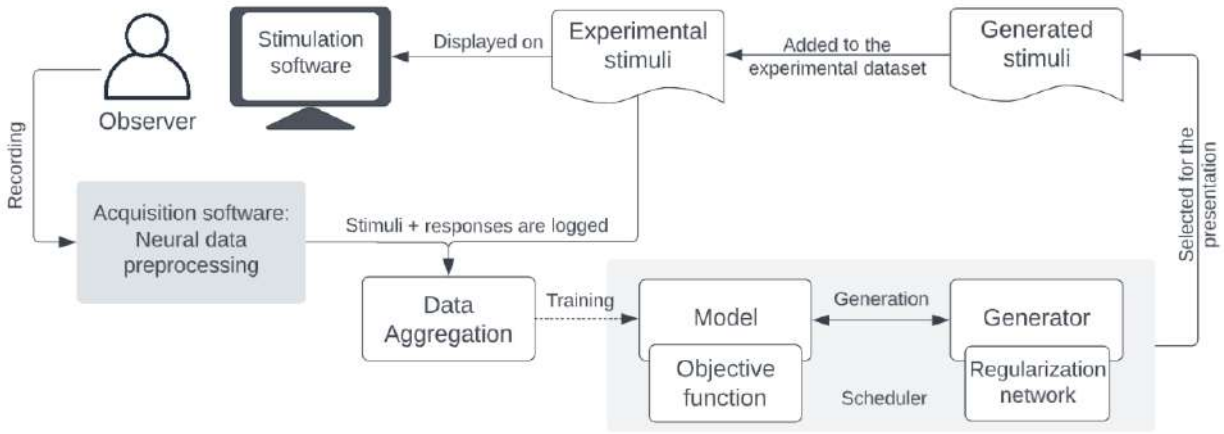

Рисунок 30. Программное решение для создания стимульного материала во время экспериментальной сессии

Для создания стимульного материала было разработано программное решение, реализующее предложенный подход (Рис. 30) и состоящее из следующих модулей:

- Наблюдатель (Observer). Участник исследования (в данном случае примат).
- Программа стимуляции (Stimulation software). Программное обеспечение, осуществляющее предъявление стимулов. Разработано на Matlab в лаборатории М. Танифуджи.
- Программное обеспечение регистрации нейронального ответа (Acquisition software). Отвечает за получение сигнала с электродов и первичную обработку данных. Разработано на Matlab в лаборатории М. Танифуджи.
- Экспериментальные стимулы. Натуральные изображения, подобранные исследователями лаборатории М. Танифуджи и предъявляемые на экране с целью регистрации нейронального ответа (Takayuki-1550). Они также добавляются в тренировочный набор данных для обучения генеративной модели.
- Модуль сбора тренировочных данных (Data Aggregation). Модуль, сопоставляющий стимулы и соответствующие реакции для формирования тренировочного набора.
- Блок оптимизации, включающий в себя следующие компоненты:
    - Модель (Model). Визуализируемая модель, обученная на данных экспериментальной сессии и представляющая наблюдателя.
    - Функция ошибки (Objective function). Модуль, определяющий насколько хорошо сгенерированное изображение соответствует целевой функции и осуществляющий оптимизацию высокоуровневого описания, кодирующего изображение.
    - Генератор (Generator). Модуль, генерирующий стимулы на основе полученного высокоуровневого описания.

- Регуляризатор (Regularization network). Дополнительная модель, используемая для улучшения качества или устойчивости генерируемых стимулов. Добавляется к высокоуровневому описанию перед его подачей в генератор.
- Планировщик (Scheduler). Отвечает за контроль процесса оптимизации и формирование набора изображения для следующего экспериментального блока.

- Сгенерированные стимулы (Generated stimuli): Результат работы генеративного модуля, передаваемый на презентацию.

Структура экспериментального блока была организована таким образом, чтобы обеспечить как проверку стабильности нейрональных ответов, так и тестирование эффективности сгенерированных стимулов (Рис. 31). Каждый блок содержал 240 изображений, которые предъявлялись в общей сложности 12 раз для получения надежной оценки нейрональных ответов. Усреднялся контраст изображений, а также проводилось размытие краев изображения. Стимулы предъявлялись на нейтральном сером фоне, края изображения размывались.

Основная часть блока состояла из 120 контрольных изображений, отобранных из базового набора стимулов. Оставшиеся 120 изображений были разделены поровну между двумя целевыми областями регистрации по 60 изображений на каждую. Внутри изображений, созданных для целевой области, были представлены две функции потерь – Evoked, направленная на максимизацию активации в данной конкретной точке регистрации, и Softmax, которая стремилась вызвать активацию в данной точке, при этом минимизируя ответ в других.

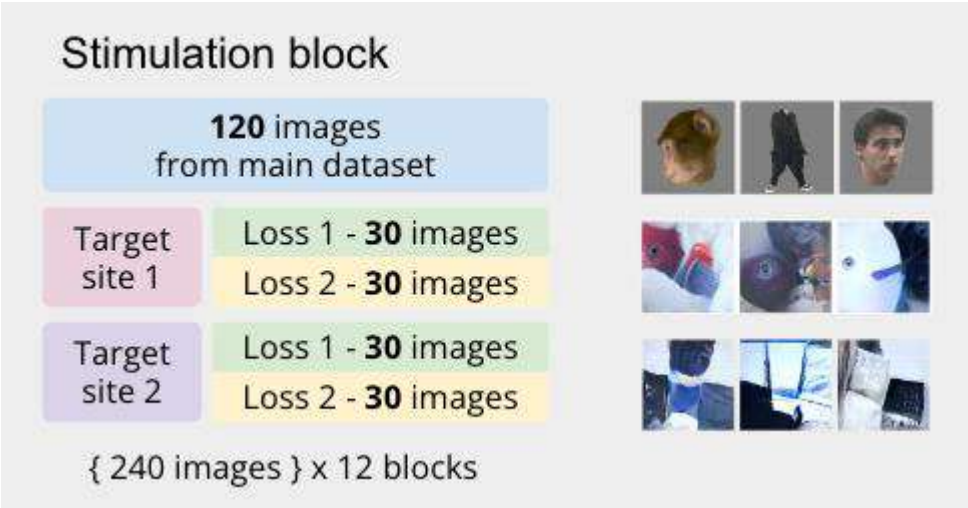

Рисунок 31. Структура экспериментального блока

Такая структура позволяла не только оценивать эффективность различных подходов к генерации стимулов, но и контролировать возможные эффекты утомления и адаптации нейронов в ходе длительной экспериментальной сессии.

После предъявления изображений проводился анализ их ответа на естественные и сгенерированные стимулы без прерывания хода эксперимента. Полученный нейрональный ответ усреднялся по 12 блокам. На основании 120 изображений проводилось его сопоставление с ранее наблюдаемым ответом во время предварительной части эксперимента.

Результаты предъявления первого блока сгенерированных изображений

Данные нейрональной активности, полученные во время основного экспериментального блока (Takayuki-1550), проходили предварительную обработку и подавались для обучения целевой модели. Проводилось несколько итераций генерации (Рис. 32) с различными функциями потерь. Созданные стимулы отбирались по результатам предсказанного вызванного отклика – изображения с максимальным ответом модели добавлялись к стимульному материалу. Цикл повторялся после каждого блока, его выполнение занимало приблизительно 5 минут на лабораторном оборудовании.

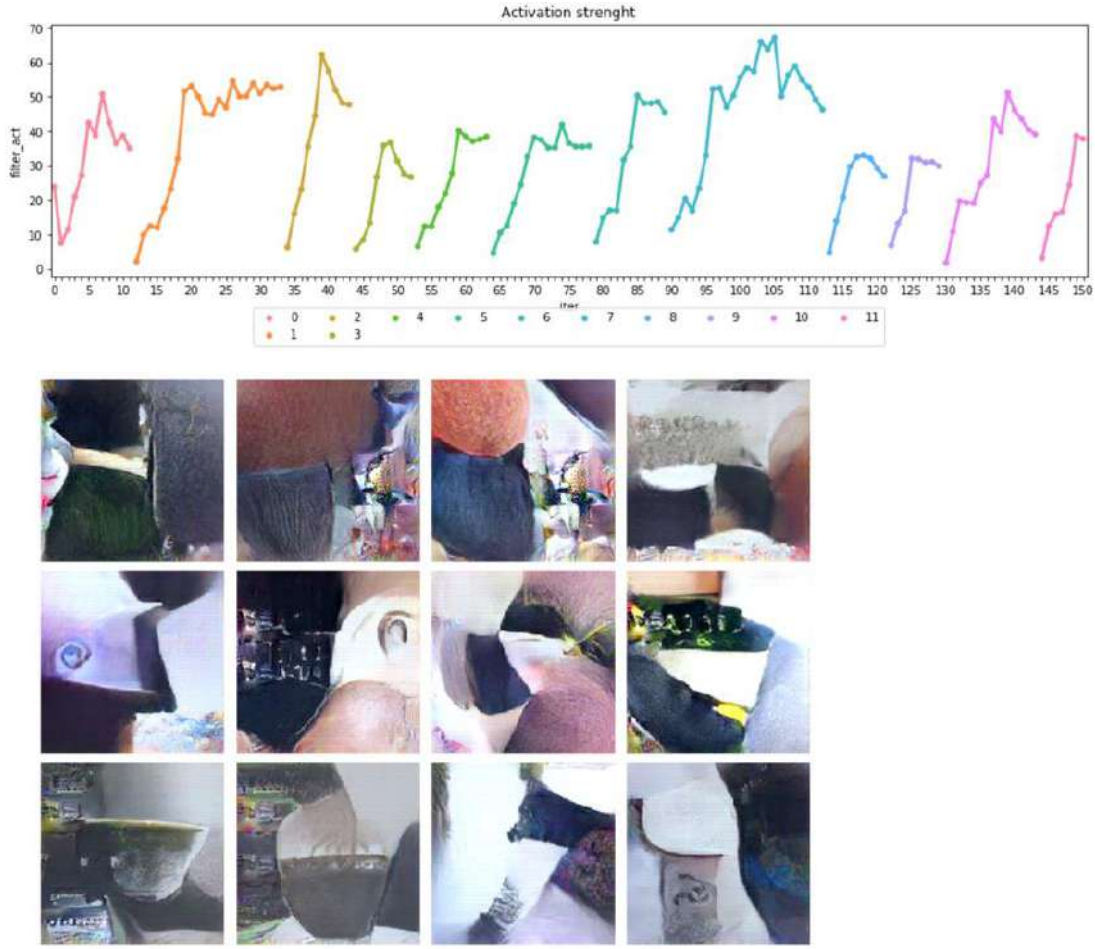

Рисунок 32. Пример создания набора стимулов. Сверху траектории создания изображения, где на каждой итерации предсказывается ответ. Снизу – изображение последней итерации

Приведены примеры сгенерированных изображений, предъявляемых в блоке 1. Заметно, что изображения, созданные с использованием функции потерь *Evoked (*Рис. 33) имеют больший

контраст и насыщенность, чем полученные посредством применения *Softmax* (Рис. 34). При предъявлении изображений в качестве стимулов их общий контраст нормализовывался, а края изображения размывались, как и у других стимулов. Некоторые созданные изображения содержат фрагменты, схожие с признаками лиц, однако большинство из них представляет собой абстрактные паттерны.

Рисунок 33. Изображения, сгенерированные в блоке 1 с использованием функции ошибки Evoked

Стимулы, созданные с применением функции *Evoked*, были направлены на усиление нейронного ответа в области *Site 1*, стремясь к максимизации общей активации в данной области. В отличие от этого, функция *Softmax* использовалась для генерации стимулов, обеспечивающих избирательную активацию исключительно в *Site 1*, при этом минимизируя отклик в других областях (*Site 2-8)*.

Рисунок 34. Изображения, сгенерированные в блоке 1 с использованием функции ошибки Softmax

Проведен анализ вызванного нейронного ответа и его сопоставление с откликом в основном блоке для оценки стабильности нейронной активности. Дополнительно была проверена предсказательная способность модели путем сравнения ее прогнозов с зарегистрированными нейронными ответами. Корреляционные диаграммы рассеяния демонстрируют нейрональную активность в области регистрации *Site 1* для первого блока сгенерированных изображений:

- Ответ на стимул между четными и нечетными блоками предъявления (Рис. 35 а). Коэффициент корреляции составляет 0,74, что указывает на умеренную положительную связь.
- Корреляция предсказанных моделью ответов и реальных ответов нейронов на стимульные изображения (Рис. 35 б). Коэффициент корреляции составляет 0,49, что указывает на умеренную положительную связь.
- Ответ, полученный во время эксперимента, и предсказанный ответ на изображения, сгенерированные с использованием функции ошибки *Evoked* (Рис. 35 в). Несмотря на

более слабый ответ в целом, отмечается умеренная положительная связь (коэф. корр. = 0,42).

- Ответ, полученный во время эксперимента, и предсказанный ответ на изображения, сгенерированные с использованием функции ошибки *Softmax* (Рис. 35 г). Стимулы, созданные с использованием данного метода, вызывают отклик меньшей амплитуды. Тем не менее прослеживается слабая положительная связь (коэф. корр. = 0,31).

Таким образом, результаты первого блока показали, что в дополнительном блоке сохранилось сходство отклика на естественные изображения (коэф. корр. = 0,74). При этом обученная модель обладала предсказательной способностью, что было проверено путем регистрации ответа модели на изображения дополнительного блока, а затем сопоставлением с нейрональными данными (коэф. корр. = 0,49). Стоит отметить, что данный показатель ниже, чем при обучении модели и проверке ее эффективности методом валидации на тестовом подмножестве, где точность предсказания достигала коэффициента корреляции 0,73, что, с одной стороны, подчеркивает необходимость экспериментальной валидации модельных предсказаний, а с другой – указывает на возможные дополнительные факторы, влияющие на нейронные отклики в реальных условиях. Эти факторы могут включать динамическую изменчивость нейронных состояний, адаптацию к стимулам, а также влияние фоновой активности и общего контекста восприятия.

Таким образом, несмотря на снижение точности предсказаний по сравнению с тестовым подмножеством, сам факт наличия корреляции между модельными прогнозами и реальными нейронными откликами подтверждает работоспособность подхода.

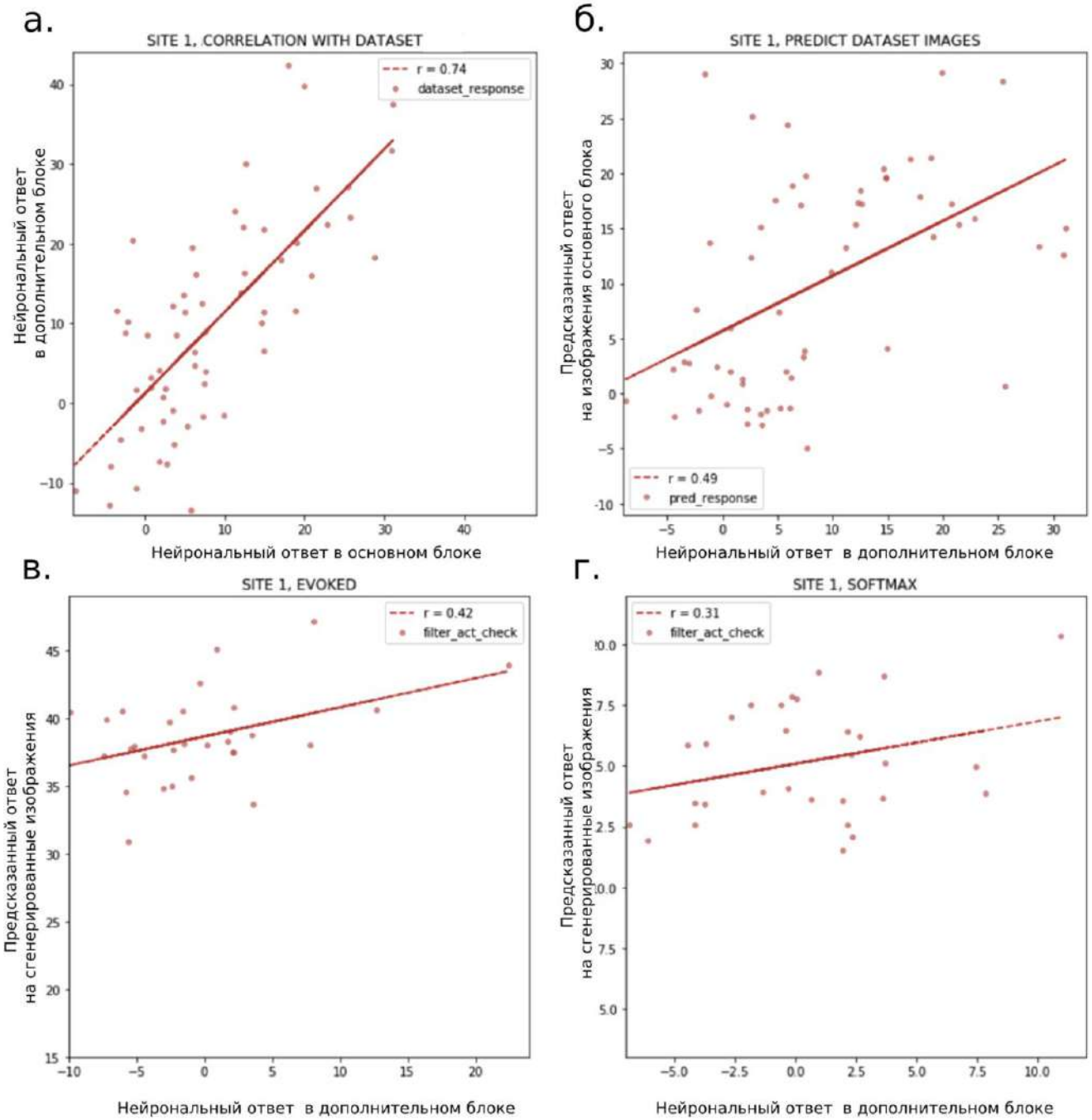

Рисунок 35. Регистрация нейрональной активации в области Site 1, блок 1: а) корреляция ответов между четными и нечетными блоками; б) корреляция предсказания модели и ответов нейронов на сгенерированные изображения; в) использование функции ошибки Evoked для генерации изображения; г) применение функции ошибки Softmax

Анализ нейрональной активности в точке записи Site 6 демонстрирует в целом более слабый ответ (Рис. 36). Обученная модель, несмотря на хорошие результаты на тренировочных данных (Рис. 36 б; коэф. корр. = 0,41), показывает неспособность корректно предсказать ответ на генерируемые стимулы (Рис. 36 в, г). Тем не менее, несмотря на меньшую силу ответа, часть созданных стимулов вызывает положительный отклик.

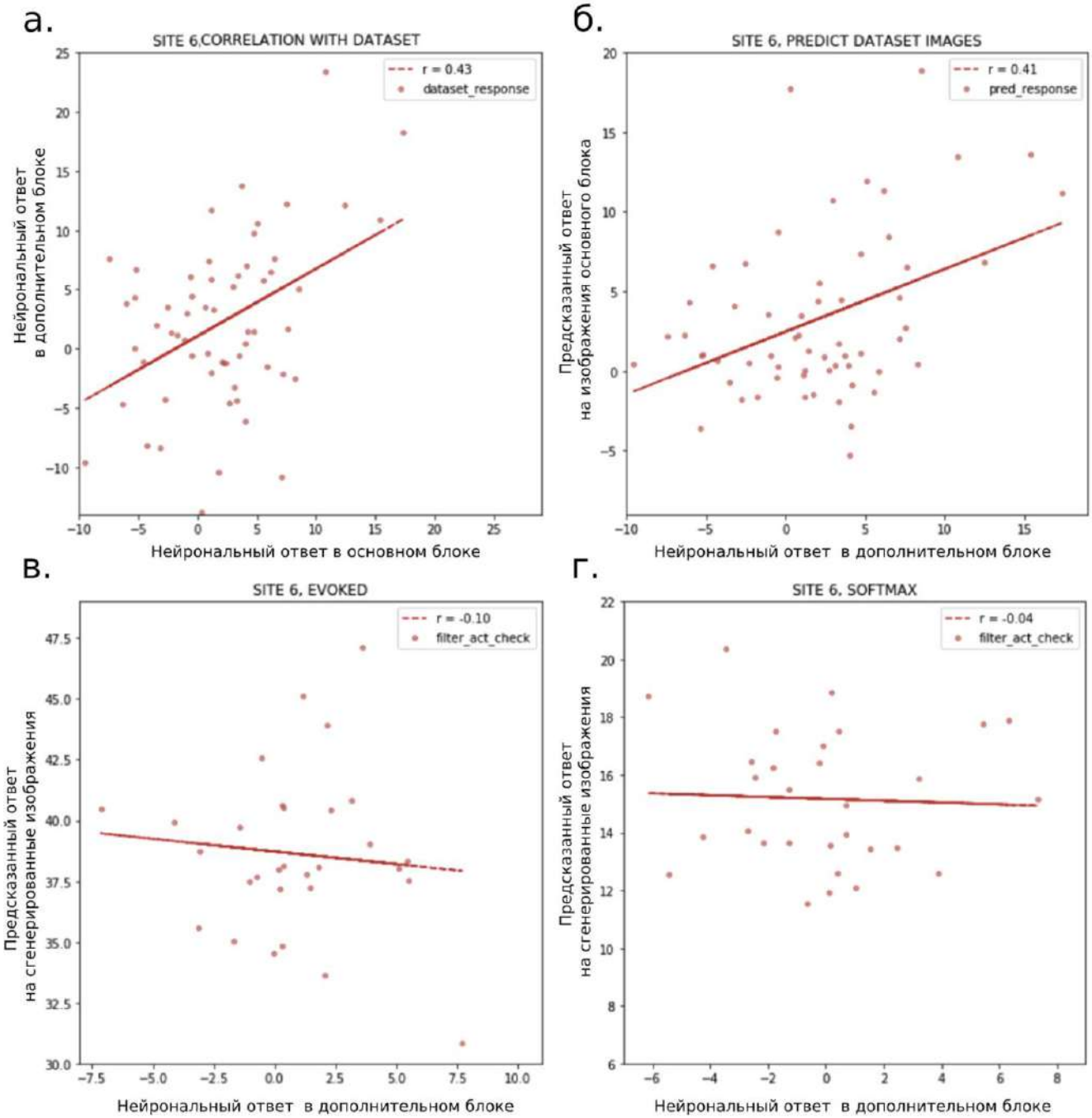

Рисунок 36. Регистрация нейрональной активации в области Site 6, блок 2: а) корреляция ответов между четными и нечетными блоками; б) корреляция предсказания модели и ответов нейронов на сгенерированные изображения; в) использование функции ошибки Evoked для генерации изображения; г) применение функции ошибки Softmax.

Сравнительный анализ распределений нейрональных ответов на естественные и сгенерированные стимулы (Рис. 37) подтверждает, что хотя в целом сгенерированные стимулы (показаны оранжевым) вызывают более слабый нейрональный ответ по сравнению с естественными изображениями (показаны голубым), некоторые из них достигают уровня активации, сопоставимого с естественными стимулами.

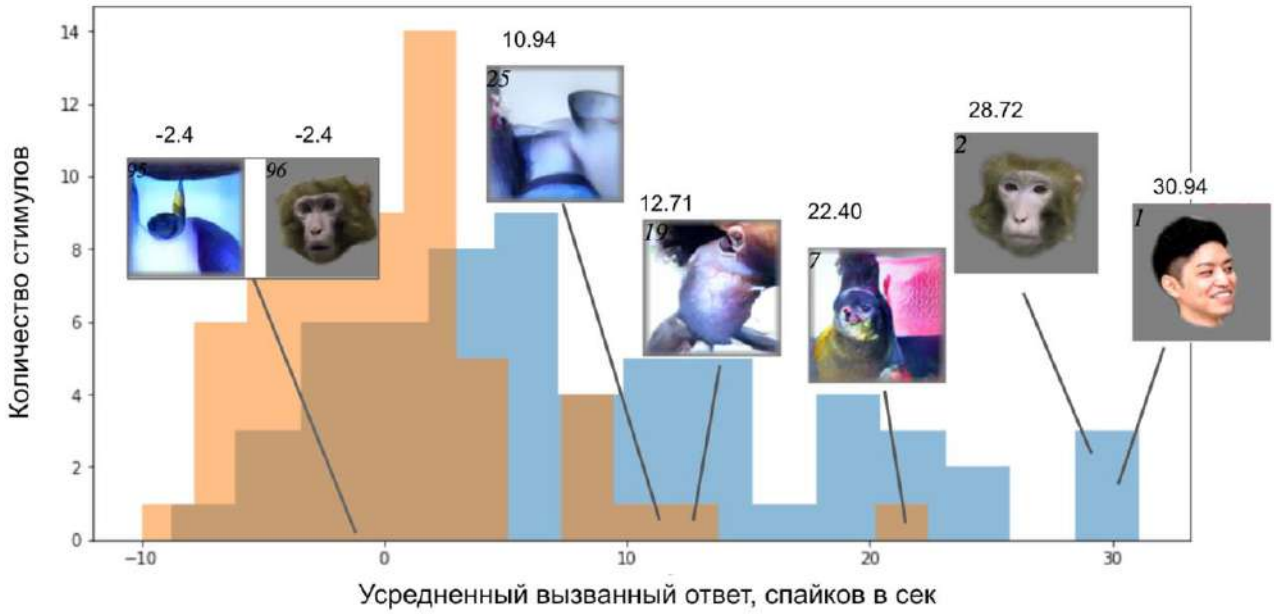

Рисунок 37. Распределение вызванного нейронального ответа для естественных и сгенерированных стимулов в точке регистрации Site 1, блок 1. Голубым цветом показано распределение ответов на естественные стимулы, оранжевым – на сгенерированные изображения

На Рис. 38 представлены 30 стимулов, которые вызвали наиболее сильный вызванный отклик в точке регистрации *Site 5*. Значения, представленные над каждым изображением, обозначают степень усредненного вызванного ответа. Можно заметить, что большинство изображений относятся к категории «лиц», причем в равной мере представляющие как лица обезьян, так и лица людей. Несмотря на явную избирательность к лицам, среди стимулов есть и абстрактные сгенерированные изображения, которые также вызвали высокий уровень активации. Среди 30 стимулов: 21 изображение с лицами, 2 изображения относятся к категории «не лица» (например, архитектурные элементы и животные), 7 изображений являются сгенерированными, что может свидетельствовать о реакции на определенные признаки, выделенные моделью в процессе генерации, а не исключительно на биологически значимые стимулы.

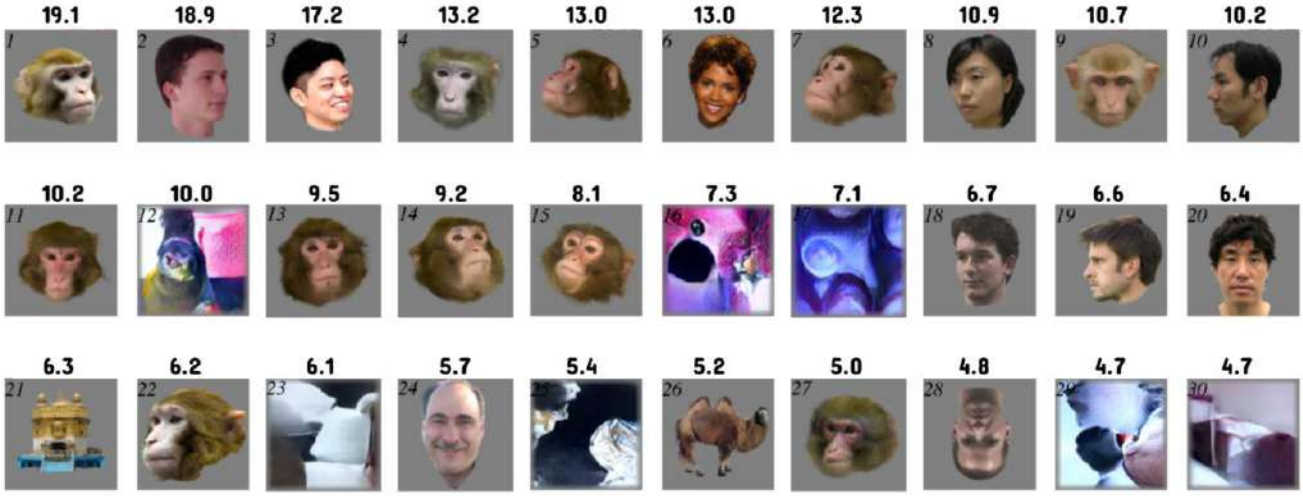

Рисунок 38. Пример 30 изображений, вызывающих максимальную активацию в Site 5, блок 1, с усредненным вызванным ответом, указанным над каждым изображением

Результаты предъявления второго блока сгенерированных изображений

Данные, полученные из области *Site 6* во втором блоке эксперимента (Рис. 39), показывают общий пониженный уровень нейронального ответа активности. Несмотря на это, среди естественных изображений прослеживается избирательность к лицам. Особый интерес представляет тот факт, что сгенерированные изображения, имеющие абстрактный характер и не содержащие явных черт лиц, способны вызывать сопоставимые нейрональные ответы с фотографиями реальных лиц.

Это наблюдение позволяет предположить, что нейроны данной области реагируют не столько на целостные образы лиц, сколько на определенные геометрические паттерны и комбинации признаков, которые были выделены генеративной моделью.

По результатам предъявления второго блока сгенерированных изображений наблюдается общее снижение нейрональной активности, что может быть связано с развитием утомления животного и адаптацией к стимуляции. При этом характер ответов существенно различается между двумя областями регистрации.

В области *Site 1*, несмотря на умеренное сходство ответов (коэф. корр. = 0.56) второго блока сгенерированных изображений и основного блока, а также сохранение предсказательной способности модели (коэф. корр. = 0.45), диапазон ответов на сгенерированные стимулы значительно уже (-10 до 10 спайков/сек) по сравнению с естественными изображениями (-10 до 40 спайков/сек). Интересно отметить, что функция *Softmax* показывает лучшие результаты (коэф. корр. = 0.35) по сравнению с *Evoked* (коэф. корр. = -0.09), что может указывать на ее большую устойчивость к эффектам утомления.

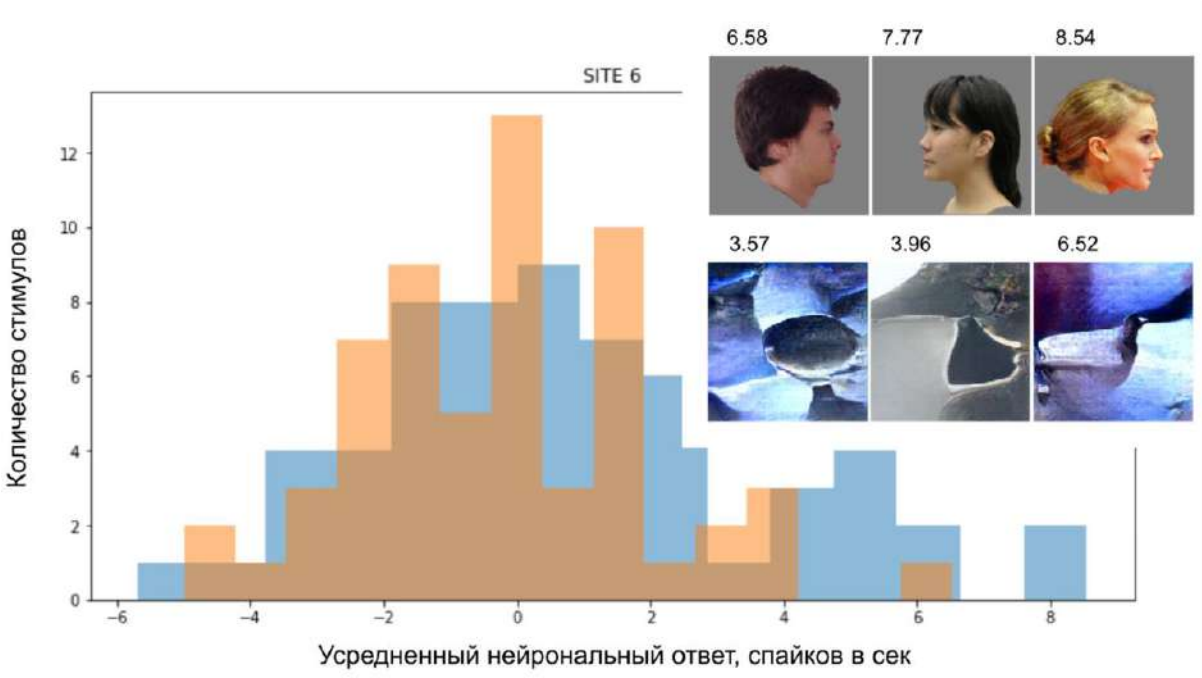

Рисунок 39. Распределение усредненного вызванного нейронального ответа для естественных и сгенерированных стимулов в точке регистрации Site 6, блок 2

В области *Site 6*, которая не является целевой, наблюдается умеренное сходство ответов (коэф. корр. = 0,61) и умеренная предсказательная способность модели (коэф. корр. = 0,45). При этом диапазон нейронных ответов значительно сужается: если для естественных стимулов он составлял от -5 до 35 спайков/сек, то для сгенерированных стимулов – всего от -6 до 6 спайков/сек. Это свидетельствует о снижении вариативности откликов при использовании искусственно созданных изображений.

Общая тенденция второго блока заключается в том, что сгенерированные стимулы вызывают более слабые ответы по сравнению с естественными изображениями. Причиной тому может быть переобучение модели, так как между экспериментальными блоками проводилось дообучение модели наблюдателя на основе результатов, полученных в предыдущем блоке. Однако набор обучающих данных мог быть недостаточным (всего 240 ед.) и искажающим статистику за счет включения значительного количества сгенерированных изображений.

Результаты применения метода показывают его принципиальную эффективность и открывают возможности к созданию целевых стимулов в нейрофизиологических исследованиях, а также новое понимание механизмов кодирования информации в высших отделах зрительной коры. В отличие от других подходов к визуализации функции искусственных нейронных сетей данный метод помимо уровня активации выбранного нейрона учитывает также требование к натуралистичности результата, что важно для проведения нейрофизиологических экспериментов. Это позволяет избежать создания стимулов, при которых нейронная сеть с

высокой уверенностью детектирует объект в изображениях, расцениваемых человеком как шум, либо как объект, очевидно принадлежащий другому классу (Nguyen, Yosinski, Clune, 2015).

При этом по сравнению с традиционным подходом пространство создаваемых изображений не ограничено базой стимулов и позволяет исследовать описания, кодируемые нейронами, вне времени проведения эксперимента, а также создавать стимульный материал во время проведения экспериментальной сессии.

Обсуждение результатов

В данной главе проведено исследование применения сверточных и генеративно-состязательных нейронных сетей для моделирования нейронального ответа и изучения функций высших отделов зрительной системы. Полученные результаты расширяют понимание методологических аспектов нейрофизиологических исследований, и открывают новые возможности в изучении фундаментальных принципов обработки информации в зрительной системе. Основные результаты:

Моделирование нейрональных ответов с помощью сверточных нейронных сетей продемонстрировало высокую точность предсказаний (коэф. корр. = 0,68, $p < 0,01$), что подтверждает соответствие между глубокой архитектурой искусственных сетей и процессами обработки информации в нижневисочной коре. Выявлено соответствие между глубокими слоями сети (block5_conv2) и нейронными ответами данной области. Это наблюдение поддерживает гипотезу о иерархической организации зрительной системы и указывает на универсальность принципов обработки зрительной информации. При этом выявленная зависимость точности предсказаний от стабильности нейрональных данных (коэф. корр. = 0,7, p-значение < 0,01) подчеркивает критическую важность качества экспериментальных данных и необходимость их тщательной предварительной обработки.

Разработанный *подход к генерации стимулов на основе анализа нейрональных ответов* позволил создавать оптимизированные изображения, вызывающие целенаправленную активацию нейронных популяций. В отличие от традиционных методов, использующих заранее подготовленные наборы стимулов, предложенный метод позволяет создавать естественно выглядящие изображения, оптимизированные для активации конкретных нейронных популяций. Это открывает новые возможности для изучения функциональной специализации нейронов и механизмов кодирования информации в зрительной системе.

Применение генеративно-состязательных сетей (GAN) для изучения функций нейронов реализовано в двух вариантах: постэкспериментальный анализ (офлайн) и анализ во время эксперимента (онлайн). Офлайн-метод позволил исследовать свойства нейронов по собранным

данным, тогда как онлайн-метод использовался для динамической коррекции стимуляции на основе нейронных ответов во время экспериментальной сессии. *Анализ во время эксперимента (онлайн-подход)* дает возможность проверять гипотезы о функциях нейронов непосредственно в ходе регистрации, корректируя параметры стимуляции на основе получаемых ответов.

Разработанное программное решение для генерации стимульного материала продемонстрировало эффективность в реальных экспериментальных условиях. Сравнительный анализ функций потерь *Evoked* и *Softmax* (коэф. корр. = 0,42 и 0,31 соответственно) выявил их различные преимущества и ограничения. Функция *Evoked* оказалась более эффективной для получения сильных нейрональных ответов, в то время как *Softmax* может быть полезна для дифференциации функциональных свойств различных нейронов. Также функция *Softmax* продемонстрировала большую устойчивость к утомлению нейронов во втором блоке эксперимента (коэф. корр. = 0,35 против -0,09 для *Evoked*).

Важно отметить выявленные ограничения метода, включая: риск переобучения модели при межблочном дообучении; временные ограничения на обработку данных между экспериментальными блоками (около 5 минут); снижение нейрональных ответов в ходе длительных экспериментальных сессий; и вариабельность эффективности метода для различных нейрональных популяций.

Эти ограничения указывают на необходимость дальнейшего развития методологии, в частности: разработки более устойчивых алгоритмов обучения, возможности оптимизации изображений непосредственно на основе нейронального ответа, улучшения методов предварительной валидации генерируемых стимулов.

Интересным наблюдением является то, что сгенерированные стимулы часто содержат фрагменты, напоминающие естественные объекты (например, элементы лиц), но их присутствие не является определяющим фактором силы нейронального ответа. Это наблюдение поднимает фундаментальные вопросы о природе признаков, кодируемых нейронами высших отделов зрительной коры, и может указывать на то, что традиционные представления о специализации этих областей требуют пересмотра.

Существенным достижением исследования является успешная интеграция методов искусственного интеллекта в реальный нейрофизиологический эксперимент. Это особенно важно, учитывая скорость развития подходов машинного обучения и искусственного интеллекта, и возможности, предоставляемые ими для изучения таких сложных систем как головной мозг.

Глава 4. Модельные исследования работы нейронных сетей в процессе распознавания образов объектов

В **Главе 4** основное внимание уделено изучению механизмов, посредством которых статистические свойства сигналов и решаемые задачи формируют функциональные представления информации на различных уровнях обработки. Предлагаются подходы к исследованию представления зрительной информации в нейронных сетях, которые могут быть применены как к искусственным сетям сверточной архитектуры (в том числе моделирующим нейрональные колонки), так и данным, полученным посредством регистрации нейрональной активности областей головного мозга.

4.1. Исследование пространства представления информации посредством оценки схожести пространств описания

Моделирование сложных биологических систем, таких как зрительная кора головного мозга, является одним из основных инструментов современной нейронауки. Несмотря на существенные различия между биологическими и искусственными нейронными сетями, последние предоставляют уникальную возможность для изучения принципов обработки информации в нервной системе. Важно понимать, что цель такого моделирования заключается не в точном воспроизведении всех аспектов биологической системы, а в выявлении и исследовании ключевых принципов ее функционирования.

В процессе изучения механизмов зрительного восприятия ключевым вопросом является понимание того, как информация трансформируется на различных этапах обработки. Особый интерес представляет анализ того, как исходный сигнал постепенно преобразуется в функциональные представления, соответствующие решаемым системой задачам. Для исследования этих трансформаций предлагается метод, основанный на оценке схожести пространств описания, который позволяет проследить, как меняется представление информации на разных уровнях нейронной сети.

Методы

Представление (кодирование) изображений на различных этапах обработки зрительного сигнала изучается посредством сопоставления сходства объектов, спроецированных в это пространство. Близким подходом является анализ репрезентативного сходства (RSA), в котором паттерны активности, вызванные каждым из множества стимулов (например, различными изображениями, словами и т. д.), сравниваются друг с другом (Kriegeskorte, Mur, Bandettini,

2008). В результате этих сравнений формируется матрица репрезентативного несходства (RDM), которая описывает различия между нейронными репрезентациями каждой пары стимулов.

Особенностью представленного метода является применение косинусного коэффициента в качестве меры сходства, а также использование референтных матриц, что в конечном итоге позволяет изучить каким образом статистика сигнала и постановка задачи – функциональное пространство – определяют представление информации в нейронных сетях.

Исследование пространства представления информации производилось посредством измерения сходства между объектами, спроецированными в это пространство. В случае работы со скрытыми слоями нейронной сети пространство признаков задавалось набором фильтров слоя, а присутствие признака определялось через уровень активации фильтра.

Метод исследования оценки схожести пространств описания

Предложенный подход, включает в себя следующие шаги:

1. Для набора изображений, принадлежащих к различным зрительным категориям, задаются либо рассчитываются референтные матрицы сходства объектов в пространстве сигнала (*S*) и задачи (*T*). В данном случае, в пространстве сигнала сходство рассчитывалось как Евклидово расстояние в пиксельном пространстве; в пространстве задачи, изображения, принадлежащие к одному классу, имели сходство равное единице, в то время как сходство элементов, относящихся к разным категориям, считалось нулевым.
2. Набор изображений пропускается через сверточную сеть, получая для каждого изображения карту активаций $X'_{L\times L\times N}$ на каждом слое сети, где L – размеры карты активации, N – количество фильтров определенного слоя.
3. Полученные карты активации усредняются по первым двум измерениям и центрируются по среднему значению для фильтра, рассчитанному заранее, получая вектор размерности N: $X_N = X'_N - \mu(X_N)$.
4. Подсчитывается сходство между парами объектов с помощью косинусного коэффициента и формируются матрицы сходства К объектов внутри слоя:

$$\mathrm{K} = \frac{AB}{\|A\|\|B\|}.$$

5. Рассчитывается коэффициент корреляции между матрицами сходства в пространстве задачи и пространстве сигнала $R_{task} = corr(K,T)$ и $R_{signal} = corr(K,T)$.

Для расчета сходства объектов была выбрана метрика косинусный коэффициент, который позволяет определить присутствие и меру выраженности общих признаков, игнорируя непересекающиеся характеристики. Это особенно важно в случае разреженного кодирования, т.

е. активации малого числа нейронов из общего количества. Свойство разреженного кодирования наблюдается в работе сенсорных отделов коры головного мозга, а также в высших слоях сверточных нейронных сетей.

Модель нейронной сети. Моделирование проводится с применением нейронной сети классической сверточной архитектуры VGG16 (Simonyan & Zisserman, 2015), обученной задаче распознавания 1000 категорий изображений ImageNet (Deng, et al., 2009). Начальные блоки VGG16 состоят из сверточных слоев – набора фильтров (ядер свертки), отображающих входной сигнал в значение. Каждый подобный фильтр представляет ось в многомерном пространстве, образуемым всеми элементами слоя.

Данные. Анализ проводился на 100 случайно выбранных категориях ImageNet. Каждая категория включала в себя 50 изображений из валидационного множества, не участвующего в обучении сети. Для каждого изображения были собраны активации всех 13 сверточных слоев, представленных в виде тензоров *(S x S x N)*, где S – размеры карты активации, полученные последовательным применением фильтра ко всему объему входной информации, а N – количество фильтров определенного слоя. Указанный тензор затем преобразовывался в одномерный массив (вектор) для проведения дальнейших операций.

Метрика сходства. В качестве меры сходства двух изображений используется косинусный коэффициент. Данный коэффициент отражает геометрическое сходство векторов, учитывая меру их сонаправленность, но не конкретную величину значений. Вектора одинаковой ориентации будут иметь коэффициент, равный 1; ортогонально направленные вектора – коэффициент равный 0; в то время как противоположно направленные – -1, вне зависимости от векторной величины. Таким образом, в случае если мера сходства между двумя объектами равна нулю, представления этих объектов независимы друг от друга.

Результаты

Для каждой пары наблюдений рассчитывается их мера сходства, затем, все возможные комбинации представляются в виде матрицы сходства (мер конвергенции). Результирующие матрицы, полученные для разных этапов обработки сигнала, сопоставляются с референсными матрицами. В качестве референсных матриц рассматриваются матрицы мер конвергенции, рассчитанные для пространства сигнала (пиксельное представление изображения) и пространства задачи (класс изображения в общем пространстве классов). Оценка сходства двух матриц производилась путем расчета коэффициента корреляции между ними.

В пространстве сигнала сходство в статистических характеристиках изображения определяло близость объектов. На Рис. 40 приведена матрица мер конвергенции, для 500 изображений из 10 случайно выбранных категорий (50 изображений в категории) валидационного множества

ImageNet. Как видно, элементы некоторых категорий являются подобными между собой, другие – значительно отличаются по статистическим характеристикам. Из указанной подвыборки наиболее гомогенной являлась категория «мыс» со средним по парным коэффициентом сходства K равным 0.36, наименее консистентной – категория «солнцезащитные очки» с коэффициентом K=0.02; на Рис. 41 приведены примеры изображений данных категорий.

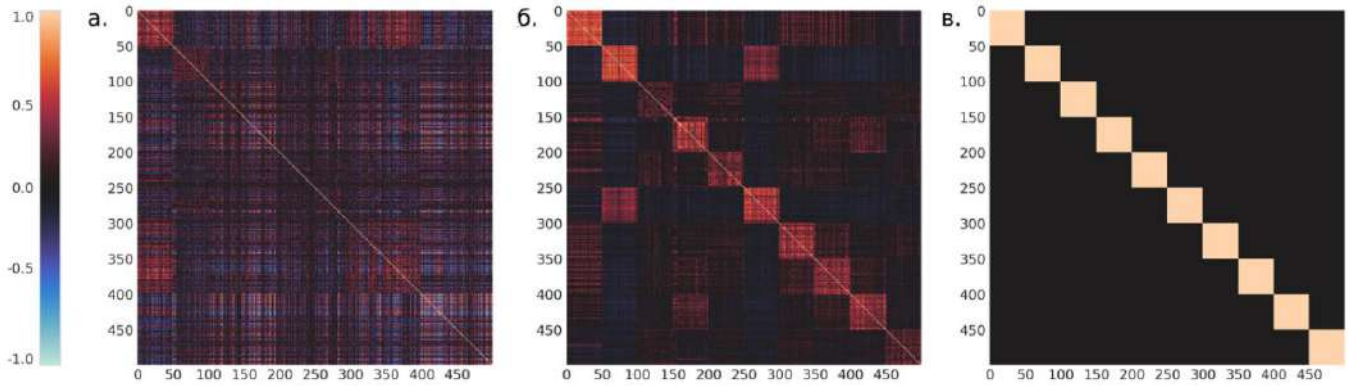

Рисунок 40. Матрица мер конвергенции для 500 изображений в (а) пространстве сигнала, (б) последнем сверточном слое и (в) пространстве задачи

Пространство задачи задавалось функцией принадлежности к классу – сходство между элементами одного класса считалось максимальным и приравнивалось единице, в то время как сходство элементов, относящихся к разным категориям, считалось нулевым (Рис. 40). Стоит заметить, что данный подход не отражает семантической близости категорий для человека (так как разные породы собак, например, считались настолько же отдаленными друг от друга, как собака от аэроплана, например), однако используется в обучении ИНС и, таким образом, обуславливает пространство описания задачи для нейронной сети.

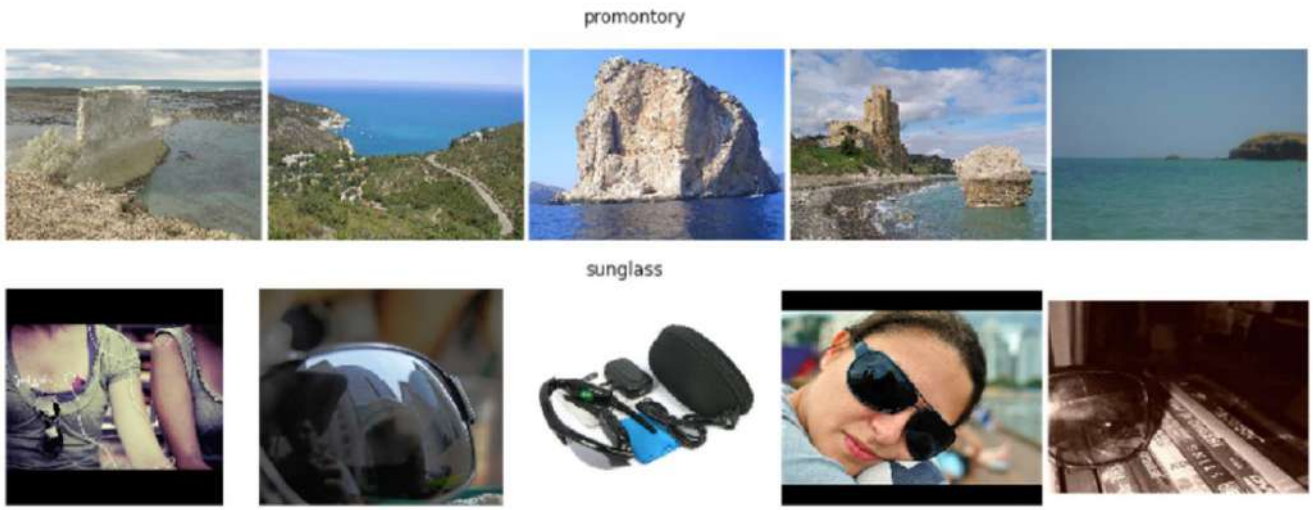

Рисунок 41. Примеры изображений валидационного множества

Далее аналогичная матрица мер конвергенции была рассчитана для всех сверточных слоев нейронной сети. Так как каждый слой производит нелинейное преобразование входа, попарное

сходство между объектами могло существенно меняться после каждого преобразования. В итоге было получено 16 матриц, отражающих сходство элементов валидационного множества в каждом из 13 сверточных и 3 полносвязных слов. Данные матрицы были сопоставлены с референтными для определения соответствия пространству сигнала и пространству задачи.

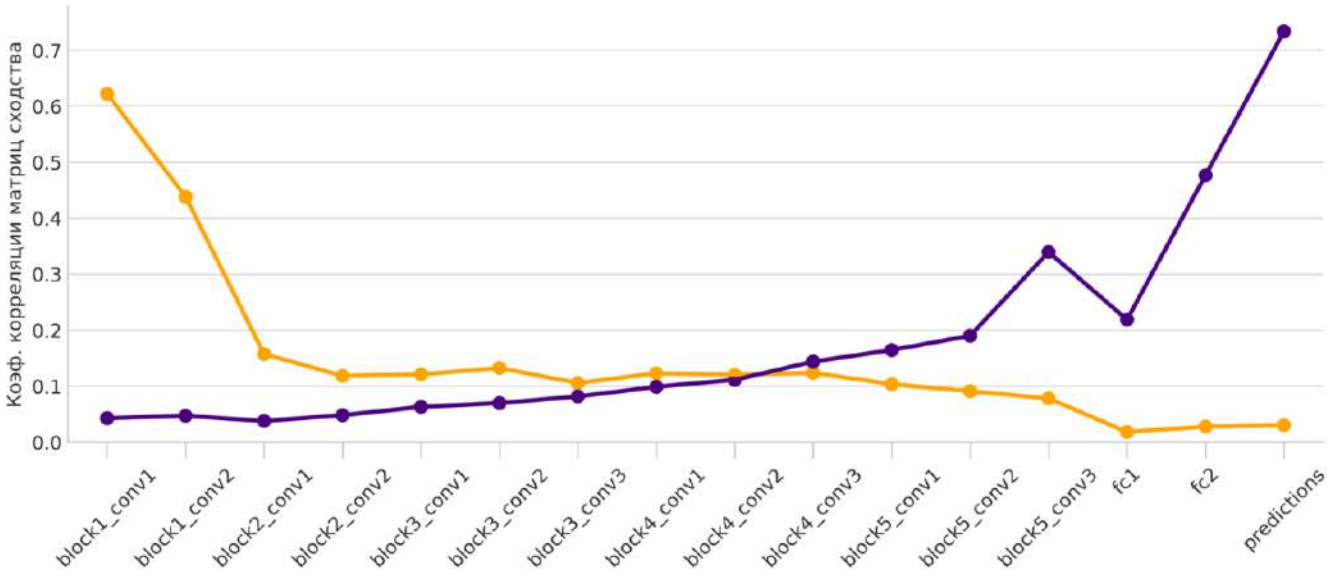

Рисунок 42. Влияние оппонентных факторов (сигнала и задачи) на формирование представлений информации в слоях нейронной сети. По оси абсцисс – слой нейронной сети VGG16, по оси ординат – корреляция между матрицей сходства изображений внутри слоя и референтной матрицей. Фиолетовым цветом обозначена корреляция с матрицей сигнала, оранжевым – с матрицей задания

На Рис. 42 приведены результаты сопоставления представления изображений в скрытых слоях с каждой из референтных матриц. В начальных слоях сети репрезентация изображений имеет значительное сходство с пространством сигнала (коэф. корр. = 0,62, $p \leq 0,01$ для первого сверточного слоя). Однако уже после первых двух уровней нелинейных преобразований указанное сходство значительно снижается (коэф. корр. = 0,16, $p \leq 0,01$). В то же время мера сходства с пространством задачи является несущественной на начальных этапах (коэф. корр. = 0,04, $p \leq 0,01$ для первого слоя), но постепенно возрастает для всех последующих слоев, достигая значения 0,34, $p \leq 0,01$ для последнего сверточного слоя. В первом полносвязном слое сходство с пространством задачи упало более чем на 0,1 по сравнению с последним сверточным слоем и составило 0,22, $p \leq 0,01$. Однако, с каждым последующим шагом оно существенно прирастало, достигнув значения 0,74 в слое предсказания.

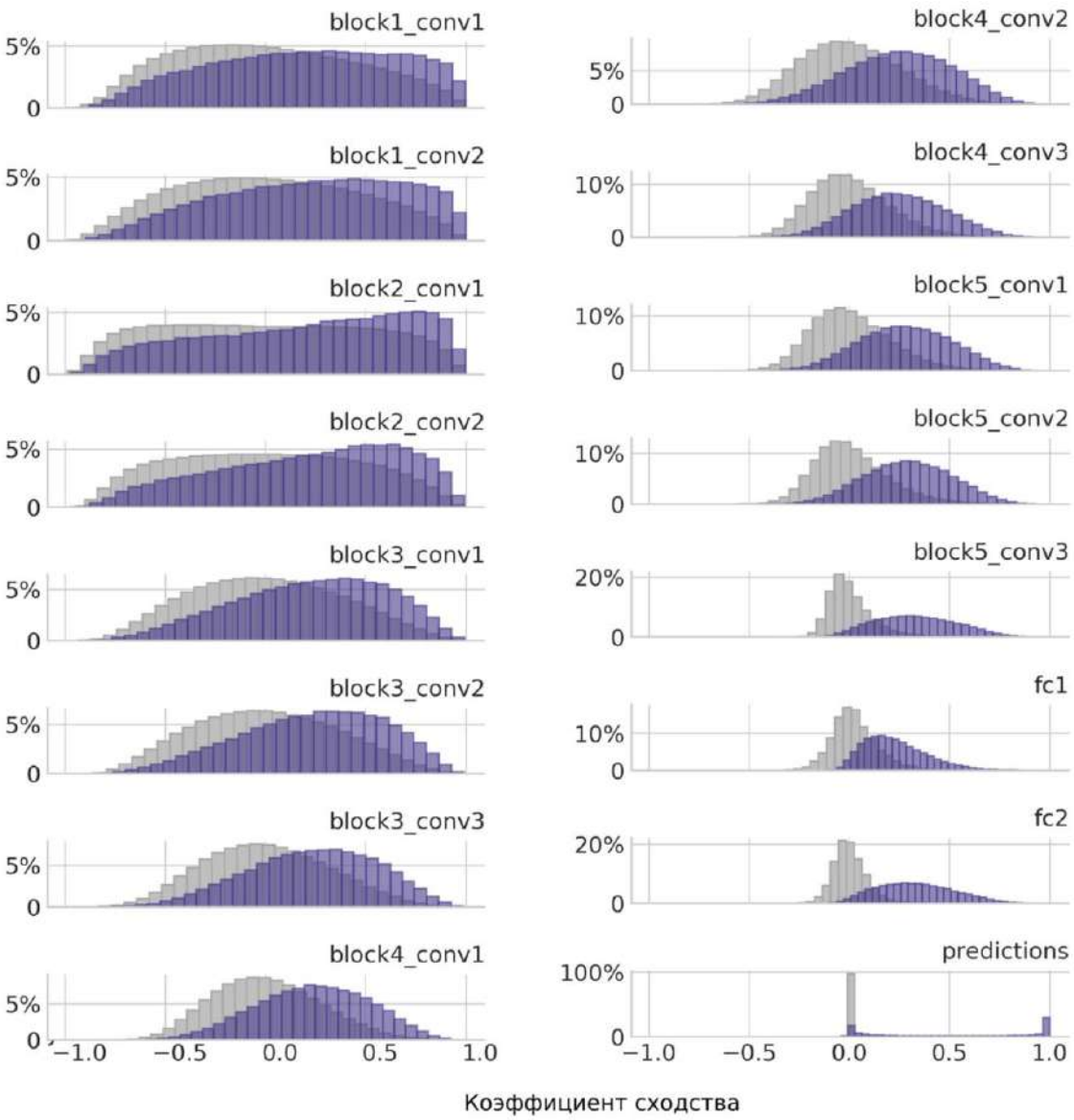

Рисунок 43. Гистограмма распределения коэффициентов сходства изображений для слоев нейронной сети. По оси абцисс – значение косинусного коэффициента, по оси ординат – процент от общего числа наблюдений. Синим цветом обозначены значения для объектов, принадлежащих к одной категории, серым – к разным

На протяжении всего процесса обработки (Рис. 43) повышается уровень сходства между репрезентациями экземпляров одного класса (K=0.05±0.27 в пространстве сигнала; 0.12±0.44 для первого сверточного слоя; 0.18±0.11 для последнего сверточного слоя). Одновременно с этим, увеличивается количество ортогональных репрезентаций, т. е. тех, где коэффициент сходства равен 0, что говорит о декорреляции представлений.

Тем не менее, нельзя заключить об окончательном формировании пространства, согласованного с задачей, в сверточных слоях нейронной сети. При рассмотрении матрицы мер конвергенции для последнего слоя (Рис. 40), заметно, что сохраняется близость между некоторыми классами. В Табл. 3 приведены примеры классов с максимальным средним

сходством репрезентаций в последнем сверточном слое и значение для них же в первом слое. Для ряда классов, изначально имеющих низкий уровень сходства, но близкое семантическое значение, схожесть репрезентаций так же возрастает, несмотря на поставленную задачу различения этих классов.

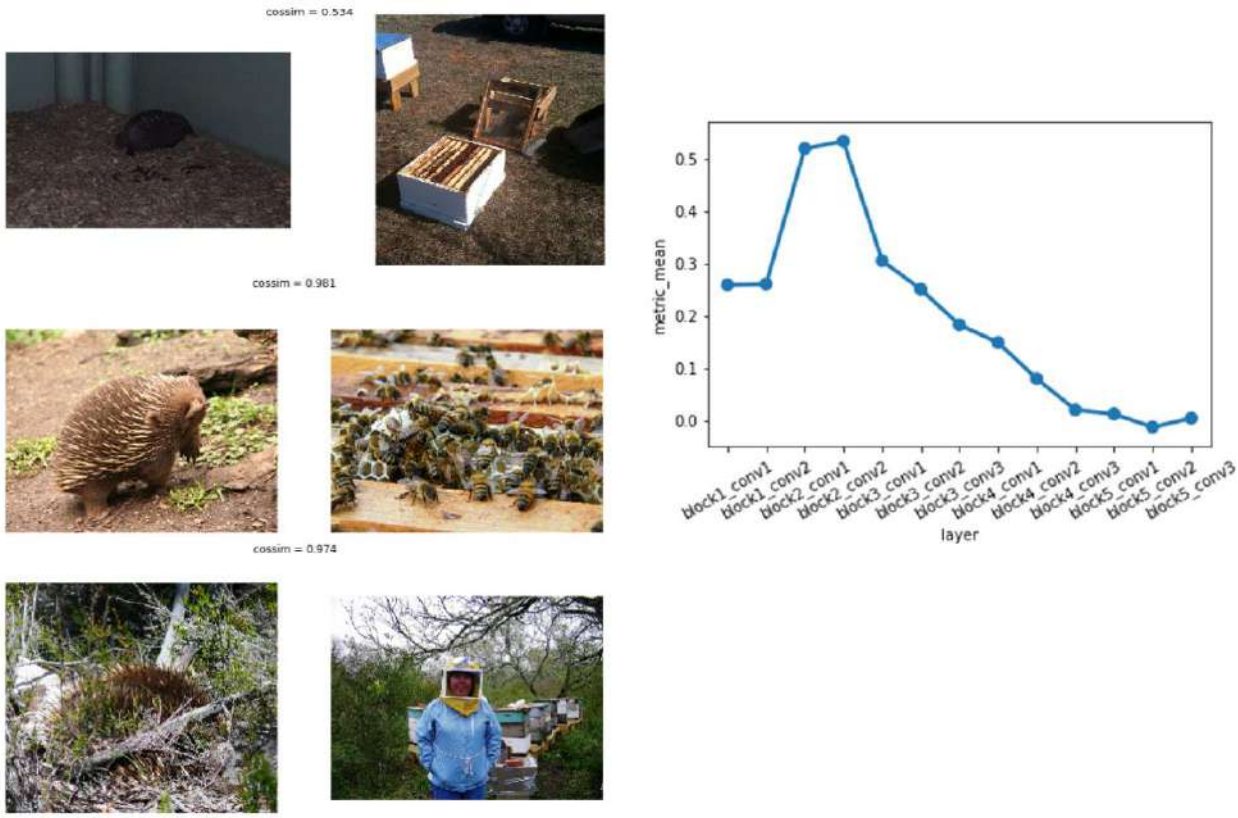

Рисунок 44. Пример двух категорий (ехидна и пасека), имеющих высокий уровень сходства изображений на начальных этапах обработки сигнала, но полностью декоррелированных впоследствии

В качестве другого примера можно привести два класса (Рис. 44), имеющих схожую низкоуровневую статистику сигнала (K=0.26 для первого слоя, 0.52 для третьего), но при этом полностью декоррелированных в последних слоях нейронной сети (K=0.02 и ниже, начиная с десятого). К третьему слою репрезентации изображений этих классов становятся более схожими. Эта тенденция, в целом, подтверждается гистограммой значений данного слоя (Рис. 42), что говорит о сходстве объектов, представленных в указанном пространстве, а следовательно, может говорить о формировании на данном уровне сети детекторов, выделяющих наиболее универсальные базовые зрительные элементы.

Увеличение схожести репрезентаций для пар категорий вроде «Луговой тетерев/Перепел» или «Норвич терьер/Австралийский терьер» в высших слоях сети (с ~0.1-0.2 до ~0.3-0.4) вероятно отражает способность сети формировать инвариантные признаки, характерные для родительского класса объектов (Табл. 3). В начальных слоях изображения этих категорий могут

сильно различаться из-за вариаций в освещении, позе, фоне и других низкоуровневых характеристиках. Однако по мере продвижения по слоям сети происходит абстрагирование от этих вариаций и выделение существенных признаков класса - например, характерной формы клюва у птиц или особенностей строения морды у определенных пород собак, общей натуральной сцены и других сопутствующих контекстов.

Таблица 3. Пример значительных изменений в коэффициенте сходства между категориями в процессе обработки зрительного сигнала

		Средний коэффициент сходства между элементами	
Категория 1	**Категория 2**	**Первый сверточный слой**	**Последний сверточный слой**
Луговой тетерев	*Перепел*	0,26	0,34
Касатка	*Мыс*	0,52	0,31
Норвич терьер	*Австралийский терьер*	0,12	0,36
Схипперке (порода собак)	*Грюнендаль (порода собак)*	0,09	0,37
Цикадки	*Сетчатокрылые*	0,18	0,31

Это наблюдение согласуется с тем, как сеть учится игнорировать несущественные различия и фокусироваться на ключевых характеристиках объекта. Данный процесс можно рассматривать как формирование своего рода «образа категории», где схожие черты внутри семантически близких классов начинают давать похожие паттерны активации, несмотря на существенные различия в исходных изображениях. Задача сети, таким образом, становится двунаправленной – выделение родительского класса, а также определение характеристик, максимально дифференцирующих схожие классы.

Проведенный анализ демонстрирует, что представление информации в нейронной сети претерпевает существенные изменения от начальных слоев к конечным, постепенно трансформируясь от пространства, близкого к статистике входного сигнала, к пространству, организованному в соответствии с решаемой задачей. Особый интерес представляют промежуточные слои, где происходит наиболее сложная трансформация представлений, что открывает новые перспективы для понимания механизмов обработки зрительной информации в биологических системах.

4.2. Исследование представления зрительной категории на уровне популяции нейронов слоя нейронной сети

Традиционные подходы к анализу нейронных сетей часто фокусируются на изучении активности отдельных нейронов или фильтров. Однако такой подход может упускать важные аспекты распределенного кодирования информации, характерного для биологических систем. В данном разделе представляется метод, позволяющий исследовать представление зрительных категорий на уровне согласованной активности всей популяции нейронов слоя, что более соответствует реальным механизмам обработки информации в мозге.

В качестве объекта исследования в работах по интерпретации кодирования информации зачастую выступает активность одного нейрона. Например, работа (Bau et al., 2017) посвящена поиску и применению фильтров, кодирующих определенный семантический концепт. Однако без дополнительных ограничений, которые принуждают классическую СНС заключать представление концепции в пределах одного нейрона (Higgins et al., 2018; Kågebäck, Mogren, 2017), такое предположение упускает из виду сложность кодирования образов на уровне слоя нейронной сети. В некоторых задачах применяется оценка коллективной активности нейронов, например, для расчета схожести изображений (Dosovitskiy, Brox, 2016; Gao et al., 2017), либо для переноса стиля с одного изображения на другое (Gatys, Ecker, Bethge, 2016), однако в задачи данных методик не входит изучение кодирования зрительных категорий.

Методы

Модель и набор данных. Расчеты проводились на СНС классической архитектуры VGG16, обученной на полном наборе данных ImageNet. Архитектура VGG16 включает 13 сверточных, 5 слоев обобщения и 3 полностью связанных слоя. В ходе анализа модели были представлены 10 случайно выбранных ImageNet категорий (~1300 изображений на категорию) и полный набор данных из тестового множества (1000 категорий, по 50 изображений в каждой). Изображения подавались в сеть и активации всех сверточных слоев регистрировались в виде многомерных векторов *(k x dim x dim x n)*, где k − количество изображений, *dim* обозначает длину или ширину карты признаков, а n равно количеству фильтров в слое.

Предложенный подход может быть использован и для биологических нейросетей, где для его применения требуется регистрация активности популяций нейронов с высоким пространственно-временным разрешением, например, с использованием многоканальной электрофизиологии или оптической кальциевой визуализации. В этом случае многомерные вектор активаций может иметь вид *(k x t x n)*, где k − количество изображений, t обозначает число временных отсчетов записи, а n равно количеству регистрируемых нейронов.

Метод исследования представления зрительной категории на основе усредненной ковариационной матрицы (прототипа)

Предлагаемый подход учитывает не только согласованность активаций фильтров, но и сложность распределенного паттерна их взаимодействия на разных уровнях обработки – от низкоуровневых статистических характеристик изображения до абстрактных и семантических признаков.

Для его реализации используется ковариационная матрица активаций фильтров, которая позволяет выявлять закономерности в обработке изображений нейросетью. Она описывает степень и направление линейной связи между различными фильтрами внутри слоя, представляя собой симметричную матрицу, где каждое значение отражает уровень их взаимосвязи.

Для более надежного анализа применяется прототип, основанный на усреднении ковариационных матриц группы изображений, что снижает влияние случайных факторов и позволяет выделить устойчивые зависимости в ответах популяции нейронов.

Основные этапы метода:

1. Формируется набор изображений, принадлежащих к представляющим интерес категориям.
2. Изображения подаются в сверточную сеть и для каждого изображения регистрируются карты активаций $X'_{l \times l \times n}$ на каждом слое сети, где l – размеры карты активации, n – количество фильтров определенного слоя. Активации уплощаются по первым двум измерениям, получая размерность $(m \times n)$, где m равно представляющим интерес $l \times l$.
3. Полученные карты активации центрируются $X_i = X'_i - \mu(X_i)$, где i является индексом строки матрицы, соответствующим i-тому фильтру, в $\mu(X_i)$ обозначает средний ответ фильтра, рассчитанный заранее.
4. Рассчитывается матрица ковариации C для ответов фильтров, $C = \beta XX^T$, где β – коэффициент $1 / (m - 1)$. Получаемая матрица имеет размерность $(m \times m)$ и элемент C_{ij} соответствует ковариации активаций i-того и j-того фильтров.
5. Формируется *прототип категории* путем усреднения всех матриц C изображений, принадлежащих к одной категории.
6. При обработке нового изображения, рассчитывается его близость к имеющимся прототипам.

Результаты

На основе ответов модели на обучающие данные были сформированы прототипы категорий путем усреднения матриц ковариаций всех изображений, принадлежащих к классу. Затем ранее

не видимые сетью изображения десяти классов пропускались через модель и им присваивались метки на основе их сходства с эталонными прототипами (рассчитанными на всем объеме обучающих данных), аналогично широко распространенному в статистике методу эталонов или ближайших соседей (Fix, Hodges, 1989). Сходство с прототипом оценивалось двумя способами: Евклидовой метрикой и через коэффициент корреляции. Изображению присваивалась метка класса ближайшего прототипа, и результаты присвоения сравнивались с реальными метками.

В качестве оценки эффективности было проведено сравнение двух подходов к созданию прототипов посредством выполнения задачи классификации и отнесения нового изображения к одной из категорий на основе: (1) среднего вектора активации, как например в метрике схожести (Dosovitskiy, Brox, 2016) изображений и (2) ковариационной матрицы активаций фильтров.

Результаты показывают (Рис. 45), в начальных слоях (block1_conv1 – block2_conv1) точность классификации остается низкой для обоих методов (~0.3–0.4), что свидетельствует о слабой дискриминативной способности ранних представлений. Однако начиная с block2_conv2 наблюдается резкий рост точности для прототипов, построенных на основе ковариационной матрицы, тогда точность классификации при помощи прототипов, основанных на средних активациях фильтров, растет более медленно. В последнем сверточном слое (block5_conv3) точность обоих подходов выравнивается, что может свидетельствовать о том, что на поздних этапах сети оба метода приводят к достаточно устойчивым представлениям классов.

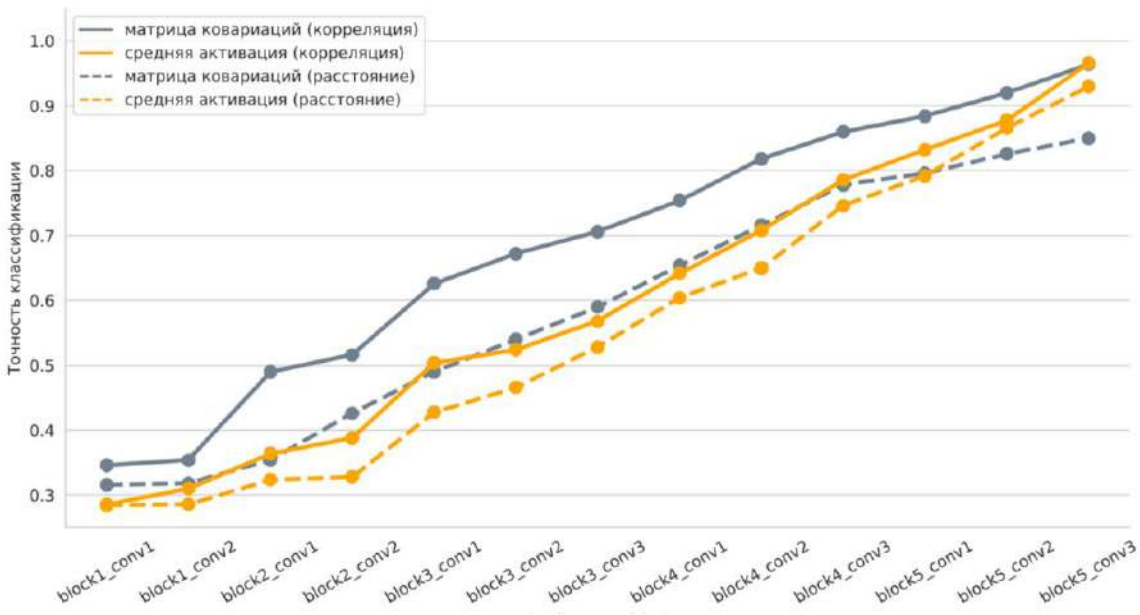

Рисунок 45. Точность классификации на основе близости к прототипу для 10 классов, отобранных из валидационного набора данных. Серым цветом обозначен подход с использованием ковариационной матрицы для формирования прототипов; желтым - подход на основе средних активаций фильтров

Классификация, использующая ковариационные матрицы в качестве прототипа категории, превосходит классификацию, основанную на векторе средних активаций, практически на всех слоях сети, особенно на промежуточных уровнях (block2_conv2 – block4_conv2), где разрыв в точности достигает 15–20%.

Важно отметить сопоставимую точность классификации в последнем слое. Подобные результаты могут свидетельствовать о том, что на поздних слоях сети информация о категориях концентрируется в отдельных нейронах. На средних слоях ковариация активаций фильтров играет важную роль, поскольку сеть еще не сформировала четкие представления классов, и взаимосвязи между активациями помогают лучше различать категории. Однако на последних слоях информацию организуется в сети таким образом, что отдельные нейроны или небольшие группы нейронов начинают явно кодировать классы, и дополнительная информация о ковариации становится ненужной. Это объясняет, почему метод ковариационной матрицы дает преимущество на средних слоях, но в конце оба метода показывают схожие результаты.

Результаты подтверждают, что учет попарных взаимодействий фильтров в слое дает более информативное представление о категории, чем усредненный ответ отдельных нейронов, особенно при исследовании промежуточных этапов обработки зрительного сигнала.

Оценка возможности аппроксимации прототипа. Доступ к обучающим данным не всегда возможен. Более того, новые поступающие данные могут не всегда наилучшим образом отражают то, что сеть видела и чему она научилась ранее. Поэтому важно понимать, насколько хорошо прототип, построенный на частичных или ранее не виденных данных, схож с эталонным прототипом, рассчитанных на всем объеме обучающих данных.

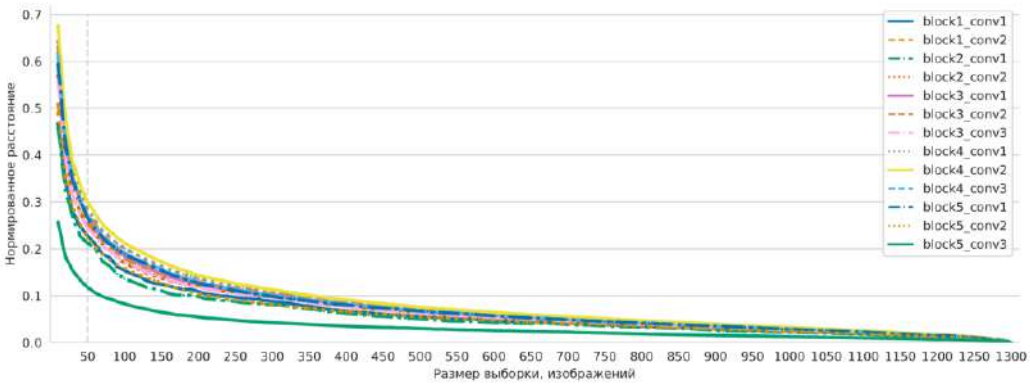

Рисунок 46. Аппроксимация прототипа категории на основании частичных или новых данных. Влияние количества изображений, используемых для составления прототипа на его отдаленность (нормированное Евклидово расстояние) от эталонного прототипа

Рис. 46 иллюстрирует, как количество изображений, отобранных из обучающих данных, влияет на качество прототипа. Для большинства слоев достаточно 50–100 изображений для

достижения относительно хорошего приближения (среднее нормированное расстояние 0,23 от эталонного прототипа). Чем более высокоуровневый слой, тем меньшее количество данных требуется для формирования прототипа.

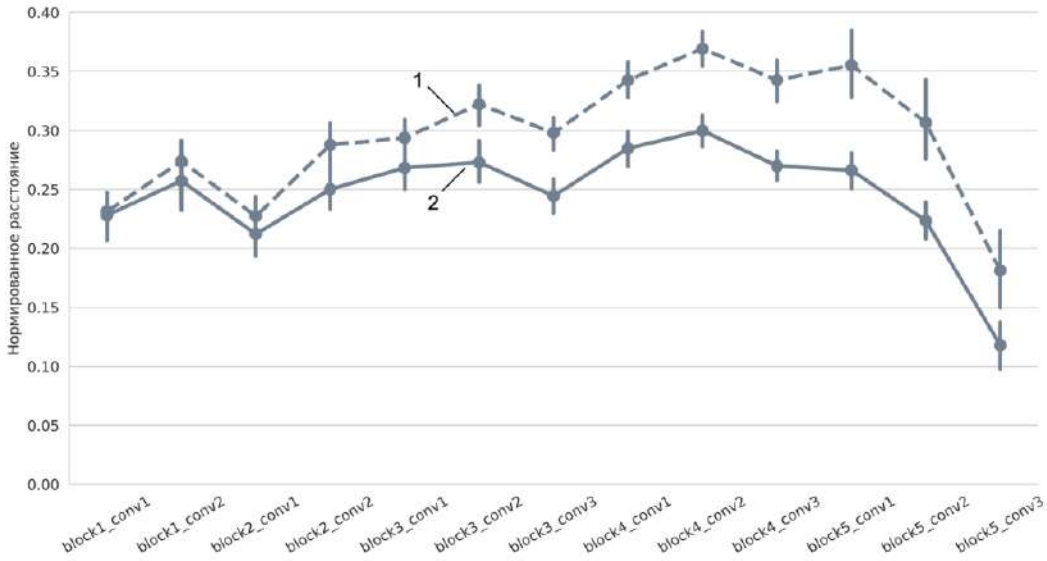

Рисунок 47. Сопоставление качества аппроксимации прототипа при использовании обучающих и новых данных. Нормированное расстояние указано от прототипа, рассчитанного на полном наборе тренировочных данных. Для аппроксимации использована выборка размером 50 изображений. Сплошная линия для обучающих данных; пунктирная линия для тех же категорий из валидационного набора данных

На основе проведенного анализа зафиксируем количество изображений 50 объектами и рассмотрим, каким образом влияет использование ранее не виденных сетью данных на формирование прототипа. На Рис. 47 показаны расстояния для оценки прототипа с использованием обучающих (ранее виденных) и валидационных (ранее не виденных) данных.

В промежуточных слоях наблюдается рост ухудшение качества аппроксимации, особенно для валидационных данных (block2_conv2 – block5_conv2), что свидетельствует о сложности сформировавшегося представления категории и затруднительности его исследования посредством предъявления малого количества стимулов, как новых, так и ранее наблюдаемых. Относительно малая выборка (50 изображений) не передает всех аспектов категории, закодированных в представлении нейронной сети в процессе наблюдения огромного разнообразия входной информации.

Стоит отметить, что на последнем слое (block5_conv3) расстояние от эталонного прототипа существенно снижается. Несмотря на малую выборку подаваемых данных, извлекаемая из изображений информация сходна с эталонным прототипом, что может свидетельствовать о переходе к семантической природе признаков, когда шум и разнообразие зрительной сцены отсеиваются и не более представлены в кодировании категории. Прототипы, построенные для

последнего слоя, являются устойчивыми даже при малом количестве данных для аппроксимации.

Таким образом метод построения прототипов на основе ковариационных матриц показал, что коллективная активность фильтров слоя несет более информативное представление категории, чем усредненные активации отдельных нейронов.

1. На ранних слоях (block1_conv1 – block2_conv1) точность классификации остается низкой для обоих методов, что свидетельствует о слабой дискриминативной способности ранних представлений.
2. На средних слоях (block2_conv2 – block4_conv2) прототипы, основанные на ковариации, демонстрируют значительно лучшую точность классификации (разница до 15–20%), что указывает на важность взаимодействия фильтров в этот период обработки данных.
3. На поздних слоях (block5_conv3) различия между методами нивелируются: информация о категориях концентрируется в отдельных нейронах, а ковариационная структура становится менее значимой.

Таким образом, представления категорий на средних слоях сети оказываются наиболее распределенными и зависят от коллективной активности фильтров, тогда как на поздних слоях информация схлопывается в отдельные детекторы.

Результаты применения предложенного метода демонстрируют преимущества анализа коллективной активности нейронов над подходами, основанными на изучении отдельных элементов. Использование ковариационных матриц для формирования прототипов категорий не только повышает точность классификации, но и позволяет лучше понять принципы распределенного кодирования информации в нейронных сетях. Это открывает новые возможности для сопоставления механизмов обработки информации в искусственных и биологических системах.

4.3. Представление категории в скрытых слоях сверточной нейронной сети

Важным аспектом понимания представления образов в нейронных сетях является оценка сложности представления, а также понимание общей картины по популяции рассматриваемых нейронов. В данном разделе используются два дополнительных подхода для детального анализа кодирования категорий.

Метод главных компонент (англ. Principal Components Analysis, PCA) помогает оценить сложность представления категории: если для объяснения значительной доли дисперсии

требуется большое число главных компонент, значит, представление категории является сложным и распределенным.

Анализ значимых направлений, который, в отличие от PCA, не учитывает согласованность активности фильтров, позволяет определить, какие именно нейроны (фильтры) слоя играют ключевую роль в кодировании категории. Это дает представление об индивидуальном вкладе каждого фильтра в процесс распознавания.

Модель и набор данных. В данном исследовании использовалась сверточная нейронная сеть VGG16, предварительно обученная на наборе данных ImageNet. Анализ значимых направлений проводился на 50 000 тренировочных изображениях, пропущенных через сеть, при этом ответы каждого фильтра усреднялись. Для анализа главных компонентов (PCA) также использовался этот же набор изображений: на их основе формировался прототип категории с учетом ковариационной матрицы, после чего к нему применялся PCA-анализ.

Определение значимых направлений

Одним из аспектов в изучении представлений категорий в нейронных сетях является понимание того, какие нейроны (или фильтры) играют наибольшую роль при обработке изображений данной категории. В отличие от методов, изучающих согласованную активность популяции нейронов, данный подход фокусируется на выявлении наиболее информативных направлений в многомерном пространстве признаков, что позволяет оценить, какие элементы слоя сети оказывают наибольшее влияние на различение категорий. Этот анализ является схожим с оценкой избирательности нейрона (Рис. 2).

Каждый элемент слоя нейросети определяет направление в многомерном пространстве, где входные изображения проецируются на набор детекторов признаков. Активации фильтров в этом пространстве можно рассматривать как координаты, а каждый фильтр определяет отдельное измерение (направление). Анализ значимых направлений позволяет выделить те признаки, которые модель использует для различения категорий, оценивая индивидуальный вклад фильтров в кодирование информации.

Значимым считалось направление, ответ которого на специфическую категорию (обучающие и валидационные данные) существенно отличался от среднего ответа фильтра, т. е. непересекающиеся доверительные интервалы уровня 0,95. *Средний ответ фильтра* рассчитывался на основе предъявления 50000 изображений из валидационного набора данных ImageNet. Для получения среднего ответа фильтра первоначально проводилась операция подвыборки, выделяющая максимальную активацию детектора на всем пространстве изображения, а затем было подсчитано среднее по всем изображениям, принадлежащим к классу.

Метод значимых направлений позволил оценить вклад отдельных фильтров в представление категории. Результаты анализа показали, что количество значимых направлений (как общее, так и относительное) представляющих категорию внутри слоя, постепенно увеличивается от начальных слоев и глубже в сеть.

При расчете на тренировочных данных до 93% фильтров демонстрируют отличную от базовой линии реакцию активности на изображения – все элементы сети максимально задействованы в процессе распознавания. Однако для валидационных данных это количество по крайней мере на почти 30% меньше (57%–70%), что говорит о тонкой подстройке к обучающим изображениям и оптимизации вычислительных ресурсов сети, в то время как новые данные демонстрируют меньшее количество изученных признаков и, таким образом, менее различимы в сформированном пространстве признаков.

Значимые направления полезны для понимания того, как отдельные фильтры участвуют в представлении категории, тем не менее они не дают информации о характере совместной активности единиц слоя. Чтобы лучше понять, как происходит кодирование категории внутри слоя, необходимо проводить интегрированный анализ, рассматривающий как индивидуальную, так и коллективную функцию нейронов.

Оценка сложности категории при помощи метода главных компонент

Ответы, вызванные набором изображений одной категории, образуют облако точек данных, спроецированных на оси пространства слоя сети. Анализ главных компонент этих точек данных помогает понять структуру репрезентаций категорий. В отличие от метода значимых направлений, где общее количество направлений равнялось числу рассматриваемых элементов (осей пространства), метод главных компонент выявляет ортогональные направления с наибольшей дисперсией, уменьшая размерность данных. Он позволяет оценить сложность категории: чем больше главных компонент требуется для объяснения значительной части дисперсии, тем более разнородными и сложными являются представления данной категории в пространстве признаков.

Для матрицы данных X главные направления являются ее собственными векторами, а проекции данных на эти оси, называемые главными компонентами, могут быть выражены собственными значениями матрицы. Было проведено сингулярное разложение *прототипов*, построенных на основе матрицы ковариаций.

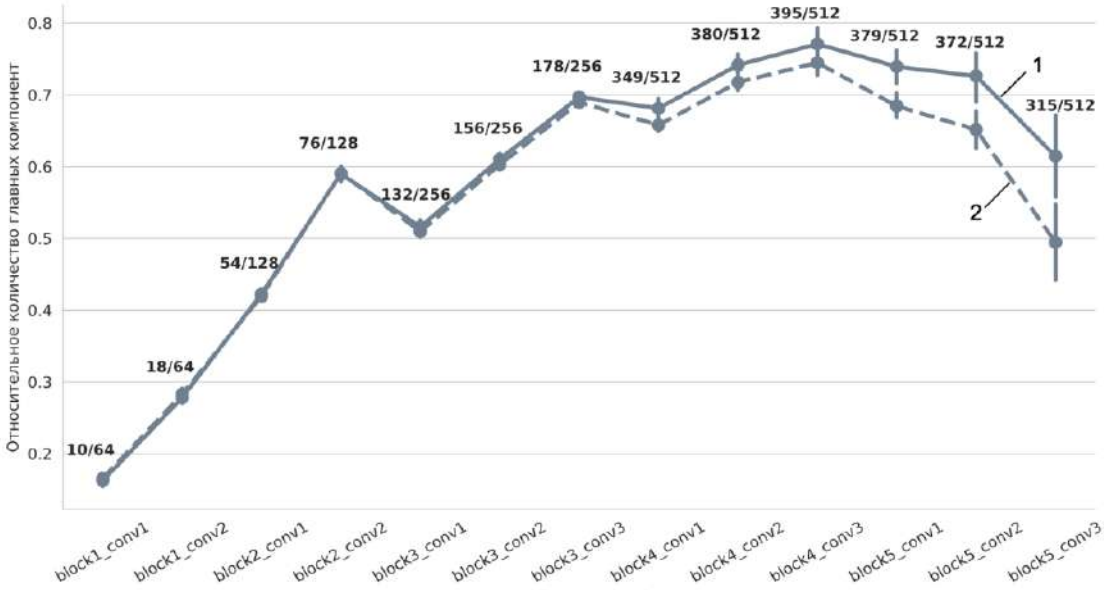

Рисунок 48. Сложность представления категории: доля главных компонент, необходимая для покрытия 95% дисперсии ответа слоя. Сплошная линия для обучающих данных; пунктирная линия для тех же категорий из валидационного набора данных

Анализ сложности категорий при помощи метода главных компонент позволил количественно оценить сложность представления категории, измеряя количество осей, необходимых для объяснения дисперсии ответов нейросети.

Результаты метода главных компонент (Рис. 48) представлений категорий показывают, что:

4. На первом слое репрезентации категорий сжаты в небольшое количество главных компонент (10 направлений объясняют 95% дисперсии), что указывает на их низкую сложность.
5. Количество необходимых главных компонент увеличивается на 15% с каждым последующим этапом обработки (первые 4 слоя, block1_conv1–block2_conv2), что говорит о возрастающей сложности представления категорий. На средних слоях (5–12) требуется гораздо больше главных компонент для описания представления категории (среднее 68%, максимум 77% в 10-м слое), что подтверждает гипотезу о высокой степени распределенности информации.
6. На поздних слоях количество необходимых компонент вновь снижается, что соответствует переходу к более семантическому кодированию, достигая 62% в последнем слое, говоря о происходящем сжатии информации до нескольких осей, наиболее важных для классификации.

Анализ сложности представлений категорий показывает, что число главных компонент значительно увеличивается на более поздних стадиях обработки, что означает, что

представление категории распространяется на большинство фильтров слоя, формируя распределенные представления. Полученные результаты свидетельствуют об ограниченной применимости методов, когда один фильтр рассматривается в качестве детектора категории (Bau et al., 2017).

4.4. Интерпретация функции нейронов посредством фрагментов натуральных изображений

Исследование популяционного кодирования выявило сложную картину распределенных представлений, где категории описываются через согласованную активность множества нейронов. Однако для полного понимания принципов работы такой системы важно также понимание того, какие конкретно зрительные признаки вносят вклад в формирование этих распределенных представлений. Это приводит к необходимости совмещения различных подходов с целью изучения функций отдельных нейронов через выражаемые визуальные характеристики, и сохраняя при этом связь с общей картиной распределенного кодирования.

Процессы, происходящие в первичной зрительной коре, являются хорошо изученными поэтому логичным следующим шагом является применение этого знания для изучения более высокоуровневых этапов обработки зрительной информации.

Предлагается метод аппроксимации функции нейронов через набор признаков в пространстве локальных ориентаций и цветов, извлеченных из фрагментов натуральных изображений разработанный в лаборатории Танифуджи М. в RIKEN Brain Science Institute (Nam et al., 2021). Автором диссертационной работы проводится применение метода к искусственной нейронной сети для сопоставления полученных данных с регистрацией нейрональной активности у приматов.

Методы

Танифуджи М. и коллегами была проведена запись ответов нейронных колонок из 190 участков передней нижневисочной коры трех анестезированных макак при предъявлении 1509 стимульных изображений, включавших: 287 лиц с контролируемым ракурсом (по 7 ракурсов для 41 лица), 532 лица с неконтролируемым ракурсом, 690 изображений из других категорий. Для анализа были отобраны 88 участков со стабильным ответом (коэф. корр. > 0,5 между усредненными ответами на четные и нечетные блоки с коррекцией Спирмена-Брауна) и избирательностью к лицам (индекса избирательности к лицам > 1/3). Из них дополнительно выделены 39 участков, демонстрирующих значимую вариативность ответов в зависимости от ракурса лица ($p < 10^{-6}$, тест ANOVA).

Метод интерпретации и аппроксимации функции нейронов при помощи фрагментов, включал следующие основные шаги (Рис. 49):

Создание словаря фрагментов. Для аппроксимации функций нейронов использовался метод, основанный на фрагментах естественных изображений. Собрано 7753 естественных изображения из базы данных VOC 2010 (Everingham et al., 2010), которые затем разделены на 560 000 фрагментов различного размера (от 8×8 до 20×20 пикселей). Каждый фрагмент в пиксельном пространстве RGB преобразован в набор из семи каналов: четыре канала локальных ориентаций (0°, 45°, 90°, 135°), полученные с помощью фильтров Габора, и три цветовых канала (красный, зеленый, синий). Подобное представление имитирует кодирование информации в ранних зрительных областях (V1/V2) и используется как «признак-кандидат» для последующего анализа.

Рисунок 49. Схема ИНС для интерпретации и аппроксимации функции нейронов через набор комбинаций признаков в пространстве локальных ориентаций и цветов

Представление стимулов. Каждое изображение из набора стимулов (1509) преобразовано в аналогичное представление на уровне V1/V2. Все изображение целиком обработано с помощью фильтров Габора и операций локального максимума, в результате чего получено представление в виде набора из семи каналов (четыре ориентационных и три цветовых). Это обеспечивает сопоставимость представлений фрагментов и стимулов в одном пространстве признаков.

Сопоставление признаков и предсказание нейрональных ответов. Для каждой пары «признак-кандидат – стимул» выполнена процедура, при которой представление фрагмента-кандидата систематически «сканировало» все представление стимула по принципу скользящего окна. В каждой позиции *(x, y)* вычислялось евклидово расстояние между фрагментом и соответствующим участком стимула, формировалась двумерная карта расстояний, отражающая степень сходства в каждой позиции, после чего определялся глобальный минимум расстояния, что соответствует инвариантности IT-нейронов к положению признака. Для учета инвариантности к масштабу предусмотрен поиск оптимального размера участка стимула в пределах заданного диапазона *b*, что реализовано через изменение размера либо фрагмента, либо участка стимула для нахождения наилучшего соответствия. Минимальное евклидово расстояние преобразовывалось в предсказанный нейрональный ответ с использованием радиальной базисной функции.

Отбор оптимальных признаков. Затем для каждого фрагмента-кандидата сформирован вектор предсказанных ответов на все стимулы. Для оценки соответствия между предсказанными и реальными нейрональными ответами рассчитаны два коэффициента корреляции: глобальная корреляция (*rglobal*) – корреляция между предсказанными и нейрональными ответами на все 1222 стимула, и локальная корреляция (*rlocal*) – корреляция только для 532 лиц. В качестве оптимального признака для данного нейронального участка выбирался фрагмент с наивысшей глобальной корреляцией среди тех, что показали значимые результаты для обоих типов корреляции (α=0,05).

Проверка модели. Для подтверждения эффективности идентифицированных признаков использованы два основных метода. Первый – кросс-валидация, при которой 1222 стимула разделены на равные обучающую и тестовую выборки, признаки идентифицированы с использованием обучающей выборки, а эффективность предсказания оценена на тестовой выборке. Второй метод – предсказание настройки на ракурс, когда использованы признаки, идентифицированные на основе ответов на лица с неконтролируемым ракурсом и не-лица, эти признаки применены для предсказания ответов на лица с контролируемым ракурсом, и рассчитана корреляция между предсказанными и реальными кривыми настройки на ракурс.

Анализ инвариантности к ракурсу. Построено 29-мерное пространство признаков (несмотря на наличие 39 участков с настройкой на ракурс, некоторые участки имели идентичные признаки), где каждое измерение представляло предсказанный ответ одного идентифицированного признака. Для проверки инвариантности к ракурсу лица разделены на целевые (одна идентичность) и нецелевые (все остальные идентичности), найден вектор проекции *w*, максимально разделяющий целевые и нецелевые лица, эффективность разделения количественно оценена с помощью показателя AUC (площадь под ROC-кривой), а обобщающая способность проверена с использованием перекрестной проверки с исключением одного ракурса.

Результаты

Анализ данных показал, что признаки, кодируемые нейронами IT-коры, могут быть успешно аппроксимированы комбинациями локальных ориентаций и цветов, происходящими из фрагментов естественных изображений. Построенная модель (Рис. 50) объяснила до 81,0% вариабельности нейрональных ответов (показатель объясненной дисперсии составил 65,7%) и точно предсказала избирательность нейронов к определенным ракурсам лиц, причем 33 из 39 участков (84,6%) показали значимую корреляцию ($p < 0,05$) между нейронными и предсказанными кривыми настройки на поворот лица (Рис. 51), несмотря на отсутствие лиц с контролируемым ракурсом в обучающей выборке. Выделенные признаки могут быть охарактеризованы двумя категориями: (1) состоящие из множества локальных компонентов, позволяющие захватывать одни и те же части лица независимо от ракурса (n=12) и (2) локальные, с небольшим числом компонентов, реагирующие на детали вроде линии волос (n=17).

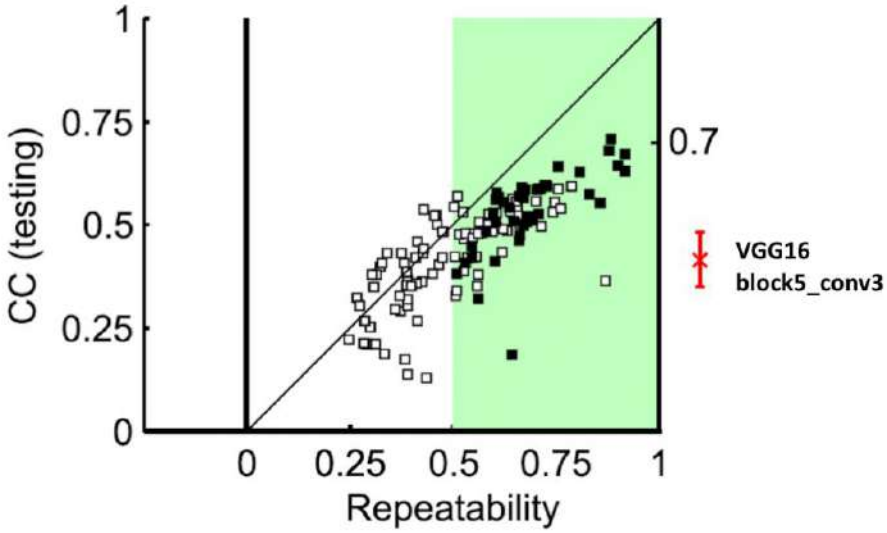

Рисунок 50. Эффективность предложенного метода фрагментов в предсказании нейронального ответа. Горизонтальная ось - повторяемость каждой колонки лиц, оцениваемая по корреляции между четными и нечетными усредненными ответами на изображения. Среди 152 колонок с избирательностью к лицам > 0.3 (черные прямоугольники), были отобраны 88 колонок с повторяемостью > 0,5 для проведения теста перекрестной валидации. Красным

обозначена эффективность применения метода для интерпретации активаций ИНС. Источник: (Nam et al., 2021)

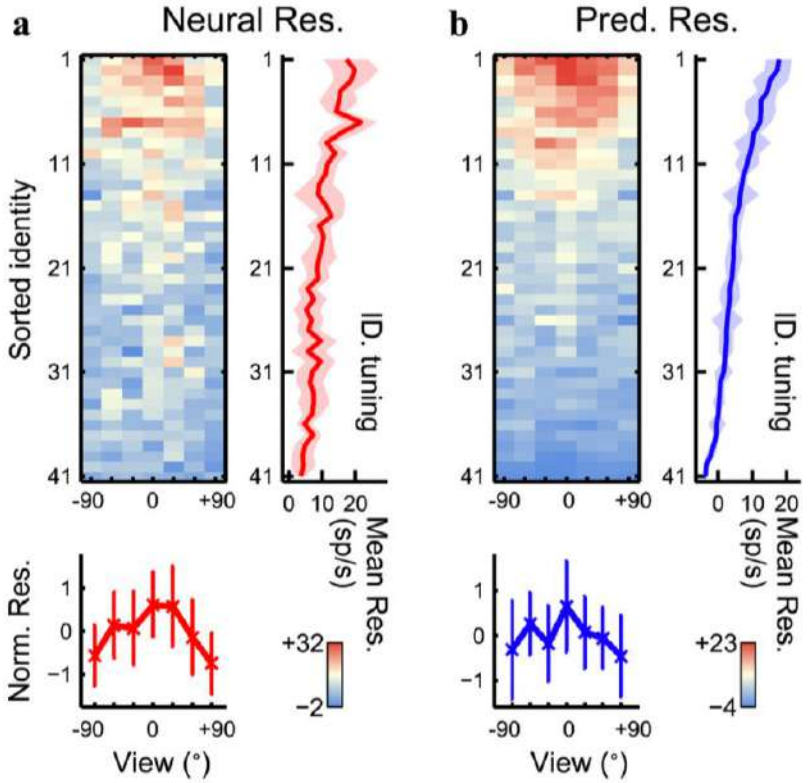

Рисунок 51. (A) Нейронные ответы на 287 стимулов лица с контролем поворота лица, записанные с нейрональной колонки S1_a. (B) Предсказанные ответы на основе фрагментов. Вертикальная ось - изображения 41 человека в порядке убывания предсказанных ответов. Горизонтальная ось, 7 видов от левого (- 90) до правого (+90) профилей, взятых через каждые 30°. Кривая справа показывает кривую настройки к личности, полученную усреднением по поворотам лица. Кривая внизу - кривую настройки на поворот лица. Источник: (Nam et al., 2021)

Применение подхода фрагментов к глубоким нейронным сетям. Использованный подход может быть применен и к сверточным нейронным сетям. Была показана возможность описания признаков, кодируемых глубокими слоями (VGG-16, тринадцатый слой), при помощи фрагментов (коэф. корр. = 0,42±0,13; $p < 0.05$ в 510 фильтрах из 512 (Рис. 50).

Для оценки эффективности подхода в условиях, исключающих «биологический шум», метод был применен к трем сверточным нейронным сетям VGG16 и VGG-Face. Сеть VGG-Face (Parkhi, Vedaldi, Zisserman, 2015) обучалась задаче распознавания 2622 известных людей на наборе данных из 2,6 млн изображений, собранных в интернете. В каждой архитектуре мы определили слой, функционально соответствующий нижневисочной коре головного мозга, где были обнаружены нейроны с избирательной реакцией на лица. Их активность регистрировалась на наборе стимулов Takayuki-1550, содержащем разнообразные стимулы с изображениями лиц и без них. Для анализа активности использовался метод интерпретации и аппроксимации функциональных свойств нейронов через набор фрагментов, что позволило выявить

специфические паттерны, на которые реагируют искусственные нейроны. При оптимальном согласовании ожидалось увидеть корреляцию между фрагментом изображения и силой отклика искусственного нейрона, приближающуюся к единице.

Для оценки этой избирательности к лицам было отобрано 5000 изображений из набора данных COCO (Lin et al., 2014) – 1000 изображений из категории «Человек» («person») и 4000 изображений из других категорий. Все изображения были поданы в сеть, и для каждого сверточного ядра была собрана средняя активация для расчета индекса избирательности к лицам.

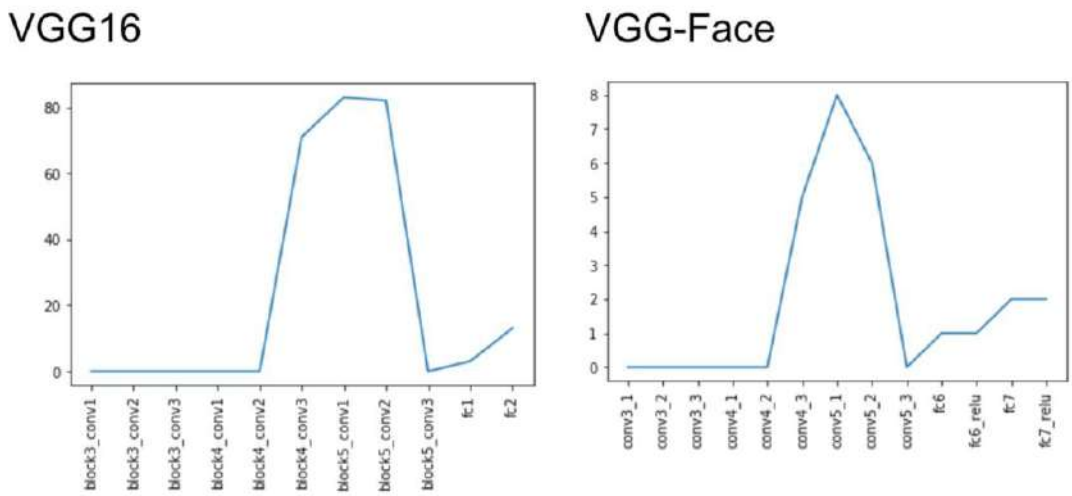

Рисунок 52. Количество нейронов, избирательных к категории лица по метрике CS

Сравнение распределения фильтров, избирательных к изображениям лиц (Рис. 52), показало, что в VGG16 наибольшее количество таких нейронов сосредоточено в слоях block4_conv3 и block5_conv2, где их число достигает ≈80. В более ранних сверточных слоях избирательные нейроны практически отсутствуют, а после block5_conv3 их число резко сокращается, достигая минимума в полносвязных слоях. Однако относительное количество избирательных нейронов, существенно увеличивается по мере продвижения по сети (Рис.53). В отличие от VGG16, в VGG-Face количество избирательных нейронов значительно ниже на всех слоях, с максимумом около 7 нейронов в conv5_3. Обе сети имеют наибольшее количество избирательных нейронов в слое 11 (block5_conv1, либо же conv5_1)

Распределение индекса избирательности к лицам (FSI) в VGG16 и VGG-Face демонстрирует различия в характере обработки изображений лиц. В VGG16 FSI достигает максимальных значений (≈50 000) в средних сверточных слоях (block4_conv3 и block5_conv2), но затем резко снижается после block5_conv3, падая до ≈10 000 в полносвязных слоях. В VGG-Face максимальный FSI выше, достигая ≈80 000 в аналогичных слоях, но его снижение в глубоких слоях менее выражено, и в полносвязных слоях он остается на уровне ≈10 000, тогда как в VGG16

он падает значительно резче. Это говорит о том, что VGG-Face лучше сохраняет избирательность к лицам на поздних этапах обработки.

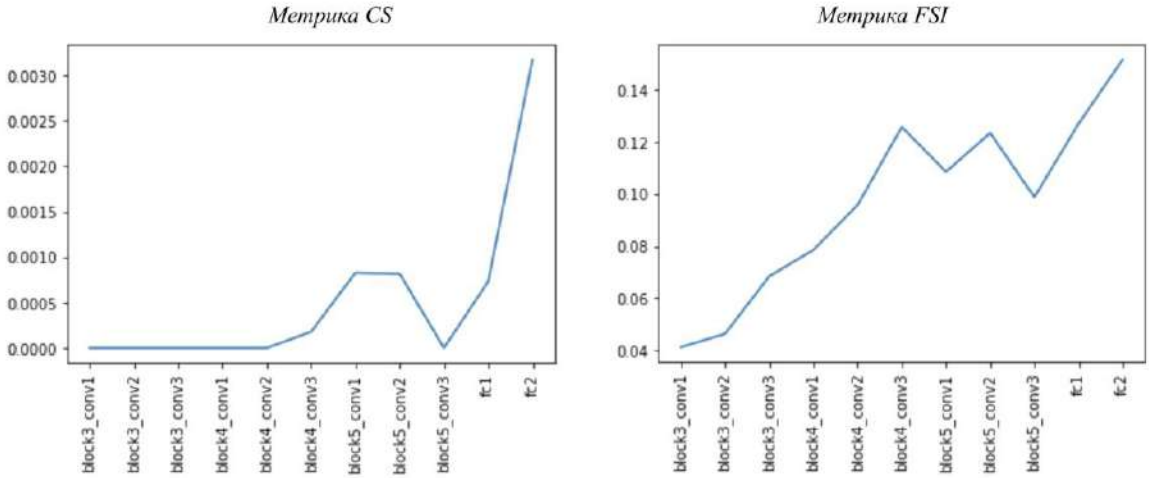

Рисунок 53. Относительное количество нейронов сети VGG16, избирательных к категории лица по метрике CS и FSI

Интересным наблюдением является то, что, несмотря на то что VGG-Face была обучена исключительно на изображениях лиц, а в обучающем наборе данных VGG16 категория лиц отсутствовала намеренно, в ее сверточных слоях наблюдается меньшее количество нейронов, явно специализированных на лица. В VGG16 такие нейроны появляются как универсальный механизм обработки естественных изобрадений, тогда как в VGG-Face детекторы, по всей видимости, выполняют другую функцию – не детекцию лиц, а их различение.

По результатам анализа была выбрана модель VGG16 и слой 11 (block5_conv1). Слой block5 показал наибольшее количество положительно избирательных нейронов (n = 40) среди других сверточных слоев, таких как block5_conv2 (n = 39, второе место) или block5_conv3 (n = 33). При этом в block5_conv3 был найден нейрон с более высоким FSI, однако он менее коррелирует с нейронными данными. Следует отметить, что слои 11–13 являются последними сверточными слоями и могут рассматриваться как «высший» уровень обработки зрительной информации из-за значительных нелинейных преобразований и больших рецептивных полей.

На Рис. 54 приведен пример средних усредненной активации фильтров слоя block3_conv1 и слоя block5_conv1 в ответ на предъявление изображений лиц.
Более яркие области указывают на более высокий уровень активации нейронов. В block3_conv1 наблюдается широкая активация по всей поверхности лица. Это связано с тем, что на данном уровне сверточной сети происходит пространственно-частотный анализ изображения, схожий с локальным преобразованием Фурье. Фильтры реагируют преимущественно на локальные особенности, такие как градиенты и текстуры, выполняя пространственно-частотный анализ.

Этот уровень обработки остается преимущественно низкоуровневым и не отражает специфичности категории стимула. Это объясняет, почему мы видим «призрачные» очертания лица при усреднении активаций всех фильтров – сеть активно кодирует пространственные частоты, характерные для лица. Если бы в качестве стимула использовался другой объект с выраженной структурой, например, сложный узор или текст, то при усреднении мы могли бы увидеть его контуры, отражающие реакцию фильтров на характерные градиенты и частотные компоненты изображения.

В block5_conv1 активация приобретает более локализованный и разреженный характер. В отличие от предыдущего слоя, где активации распределены более равномерно, здесь появляются четко выделенные области, которые могут соответствовать характерным структурным элементам изображения. Эти активации не просто анализируют статистику сигнала, как в block3_conv1, а скорее указывают на конкретные формы или объекты, что можно рассматривать как начальный этап семантического анализа.

Дальнейший анализ фокусировался именно на фильтрах с высокой избирательностью к лицам привлекают по двум причинам: 1) исследование признаков, выделяемых этими фильтрами, позволяет больше понять о механизмах лиц от объектов искусственными нейросетями, 2) учитывая, что подход с фрагментами был разработан для анализа нейрональной активности колонок нижневисочной коры, имеет смысл применить метод к искусственным нейронам, демонстрирующим аналогичные свойства.

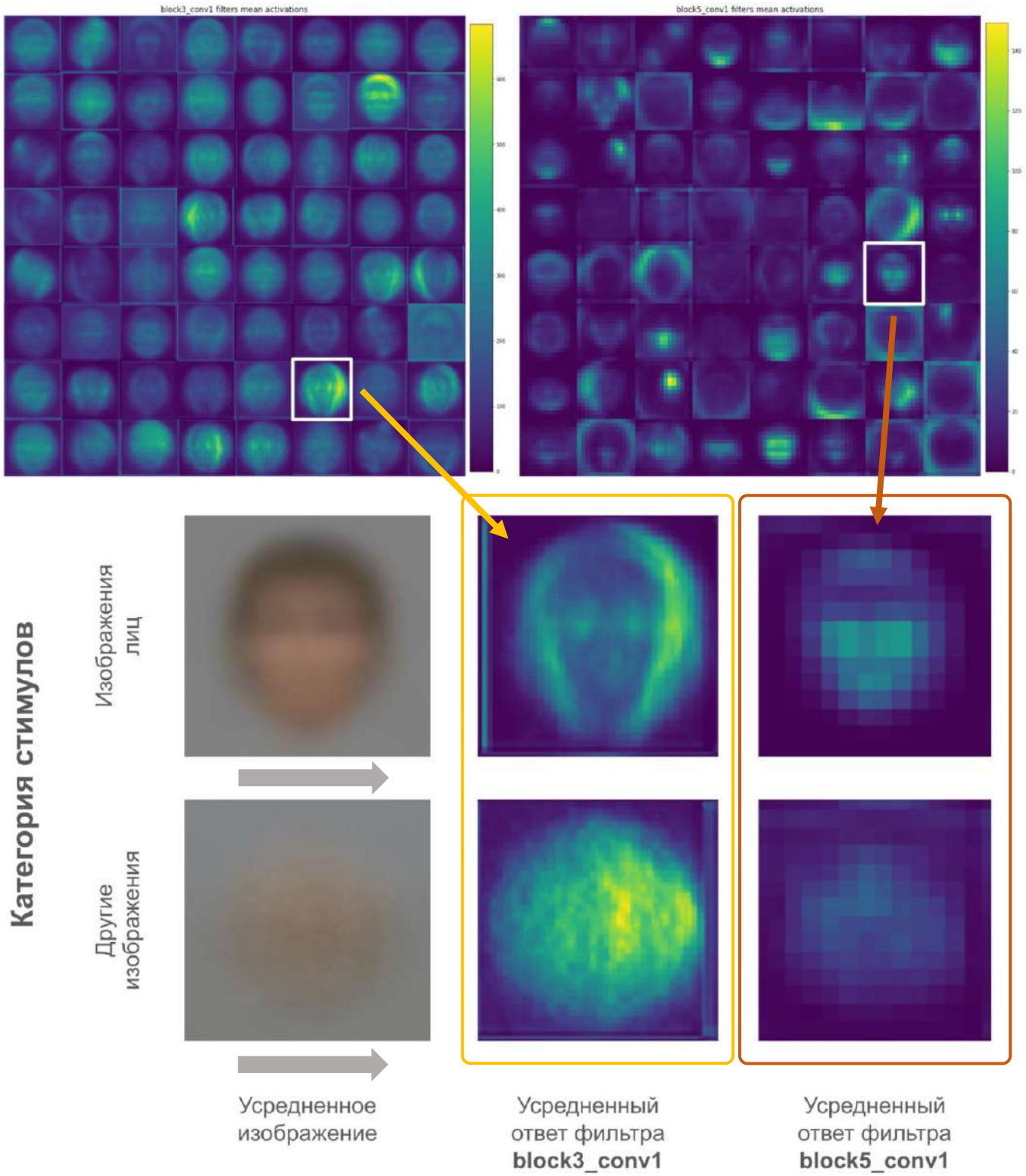

Рисунок 54. Сверху: Усредненная активация нейронов слоев block3_conv1 (слева) и block5_conv1 (справа) в ответ на предъявление изображений лиц. Более яркий цвет обозначает усиленный уровень активации. Снизу: пример активации одного из фильтров слоя при предъявлении изображений двух категорий

Среди сверточных фильтров слоя 11 VGG16 (block5_conv1) были выделены наиболее избирательные к лицам фильтры по двум показателям: по индексу избирательности к лицам FSI

– 104 (0.43), 12 (0.36), 413 (0.33), по избирательности распределения – 104 (0.05), №12 (0.02), 403 (0.02), причем фильтры 104 и 12 были выделены по обеим метрикам. На Рис. 55 приведен пример ответа фильтра 104 на стимулы, содержащие изображения лиц, где наблюдается высокая активация, тогда как его отклик на изображения без лиц (область без фона) практически отсутствует. Интересно, что на перевернутые лица (зеленый фон) активация заметно ниже, что может свидетельствовать о специфичности нейрона к конфигурации лиц в их стандартной ориентации. В стимулах 410–510, а также 1180–1330 представлены лица обезьян.

После выбора целевого слоя и фильтра (104) в сеть были поданы 1509 стимулов и собраны активации от 100352 нейронов (ширина: 14, высота: 14, глубина: 512). Поскольку все нейроны в двухмерной карте активации используют одно и то же сверточное ядро, но кодируют разные области зрительного поля, анализ признаков проводился на уровне фильтра, а не отдельных нейронов.

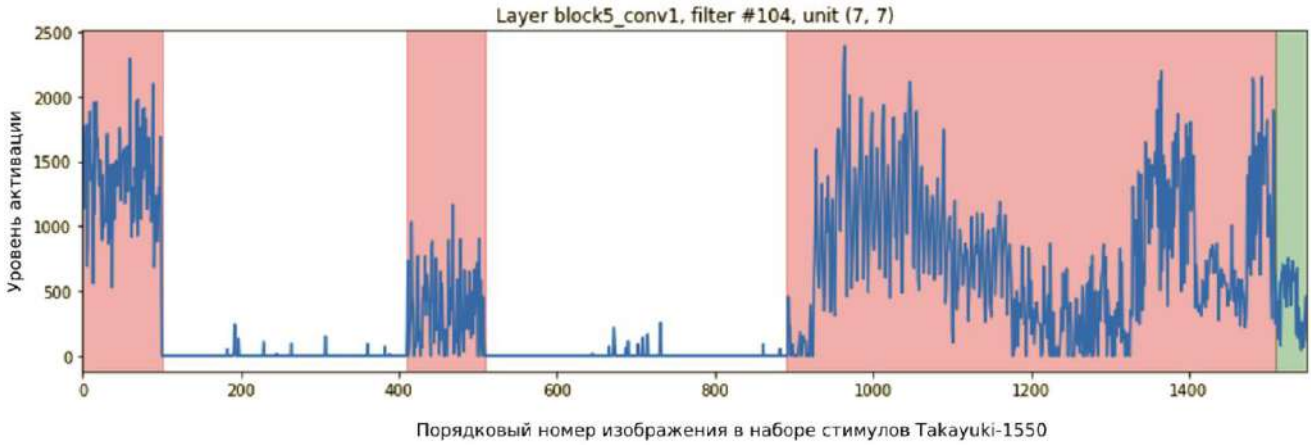

Рисунок 55. Ответ искусственного нейрона (block5_conv1, фильтр 104, центральный элемент) на стимулы Takayuki-1550. Красным фоном выделены стимулы содержащие изображения лица, зеленым цветом – перевернутые лица

Для каждого из 512 ядер были усреднены активации от 196 нейронов (14 × 14), чтобы получить отклик, и составлены векторы откликов по всем стимулам. Затем были определены зрительные признаки (фрагменты), наилучшим образом описывающие эти векторы, как при анализе нейронных откликов. Корреляция между предсказанными и фактическими векторами откликов составила 0,501 ± 0,103 (макс.: 0,786, мин.: 0,221, $p < 0,001$ во всех случаях), что указывает на то, что подход с фрагментами может извлекать признаки, кодируемые высшими слоями сверточных нейронных сетей.

Данный подход продемонстрировал, что комбинирование визуальных признаков, полученных при помощи фрагментов, позволяет объяснить инвариантное к ракурсу распознавание лиц в нижневисочной коре, а также продемонстрировать, что высшие уровни обработки в вентральном

зрительном пути могут быть аппроксимированы вычислениями, аналогичным работе неглубокой искусственной нейронной сети.

4.5. Распознавание образов в условиях неопределенности

Зрительное восприятие в условиях неопределенности представляет собой процесс обработки информации при наличии помех, неоднозначности сигналов или неполных данных. Исследования показывают, что зрительная система обладает механизмами адаптивной фильтрации, которые позволяют минимизировать влияние внутреннего шума и фоновых помех (Шелепин, Чихман, Фореман, 2008). Например, адаптация к статистическим свойствам изображения является ключевым механизмом, позволяющим зрительной системе повысить точность восприятия в условиях изменяющихся характеристик сигнала (Barlow, 1961). Бинокулярная интеграция играет важную роль в увеличении точности восприятия: объединение информации от обоих глаз снижает уровень внутреннего шума, что ведет к улучшению детекции зрительных стимулов примерно в 1,4 раза по сравнению с монокулярным зрением (Campbell, Green, 1965). Этот эффект согласуется с моделью бинокулярного суммирования, согласно которой независимые шумы в зрительных каналах комбинируются, что приводит к снижению порогов чувствительности (Глезер, Цуккерман, 1959).

Другим важным аспектом устойчивости зрительного восприятия является влияние фоновых помех и уровня контраста. Вероятностная структура шума в изображении влияет на способность зрительной системы выделять целевые объекты. В частности, было продемонстрировано, что зрительная система способна компенсировать шум за счет динамической регуляции чувствительности к различным пространственным частотам (Pelli, Farell, 1999). Кроме того, значительное подавление низкоуровневого шума происходит на первичной стадии восприятия изображения (Тибилов, Васильев, 2020; Ala-Laurila, Rieke, 2014). В техническом зрении, как и в живой природе инженеры целенаправленно размывают изображение для уменьшения шума дискретизации (Красильников, 1986).

Современные нейросетевые модели, несмотря на свою эффективность в обработке изображений, демонстрируют ограниченную устойчивость к изменяющимся условиям восприятия. Было показано, что сверточные сети, натренированные на ImageNet, демонстрируют сильную зависимость от текстурных признаков, что снижает их способность к генерализации в условиях измененного контекста (Geirhos et al., 2018). Однако добавление специально обученного шума позволяет улучшить устойчивость сетей к вариациям изображений, приближая их характеристики к биологическим системам (Jin, Dundar, Culurciello, 2016; Liu et al., 2021)

4.5.1. Внутренний шум и помехи в зрительном восприятии

Зрительная система человека функционирует в условиях постоянного присутствия внутреннего шума (Faisal, Selen, Wolpert, 2008), который возникает на разных уровнях обработки информации. Этот шум включает флуктуации фоторецепторной активности, вариабельность нейронного ответа и стохастические процессы, связанные с принятием решений. Несмотря на это, зрение остается устойчивым к помехам за счет работы адаптивных механизмов подавления шума и интеграции сенсорных данных.

Одним из главных источников неопределенности является *фотонный шум*, обусловленный дискретной природой света. Количество фотонов, попадающих на сетчатку, подвержено флуктуациям, что ограничивает чувствительность зрения при низкой освещенности. Однако сетчатка обладает механизмами адаптации, позволяющими компенсировать эту неопределенность. На следующем уровне обработки информации важную роль играет нейронный шум, который возникает в процессе передачи и кодирования сигналов. Электрофизиологические эксперименты показывают высокую вариабельность ответа отдельных нейронов, прямо пропорциональную его средней частоте разрядов (Tolhurst, Movshon, Dean, 1983), но при этом ансамблевая активность множества нейронов компенсирует индивидуальные колебания (Tolhurst, Movshon & Dean, 1983; Gur, Snodderly, 2006)

Еще один тип неопределенности связан с шумом принятия решений, возникающим на когнитивных уровнях обработки информации. В экспериментах с праймингом показано, что даже неосознаваемые стимулы могут оказывать влияние на интерпретацию последующего зрительного образа (Dehaene et al., 1998). Это говорит о том, что мозг постоянно использует вероятностные стратегии для интерпретации неоднозначных сигналов, снижая эффект случайных флуктуаций на низких уровнях обработки.

В серии психофизических экспериментов (Вахрамеева и др., 2016) изучались пороги восприятия зрительных стимулов в условиях шума и влияние прайминга на точность распознавания. В первом эксперименте испытуемым предъявлялись изображения с различным уровнем шума, и они должны были определить наличие объекта. Было показано, что вероятность правильного распознавания снижается при уровне шума выше 40%, а при 50% и более восприятие становится случайным. Однако при умеренном уровне шума (до 20%) точность оставалась выше 80%, что подтверждает высокую устойчивость зрительной системы.

Во втором эксперименте анализировалось влияние прайминга, при котором испытуемым сначала предъявлялся слабый или зашумленный вариант стимула, а затем тестовое изображение. Было выявлено, что предварительное предъявление схожего объекта увеличивает точность распознавания в среднем на 15% при шуме 30–50%. Максимальный эффект наблюдался при задержке 200–400 мс между праймером и тестовым стимулом, что свидетельствует о

вероятностных механизмах обработки информации и важности временной динамики активации нейронных ансамблей.

Несмотря на высокий уровень неопределенности на каждом этапе обработки, зрительная система обладает механизмами компенсации шума, такими как:

- *Адаптивная фильтрация.* Сетчатка и корковые области используют латеральное торможение и механизмы усиления контрастов, что улучшает детекцию границ объектов в зашумленных условиях.
- *Бинокулярная интеграция.* Совмещение информации с двух глаз позволяет устранить часть стохастического шума и повысить точность пространственного восприятия.
- *Контекстуальное усиление.* Влияние фоновой информации на обработку целевого стимула помогает зрительной системе восполнять недостающие данные и устранять двусмысленность.

4.5.2. Исследование помехоустойчивости зрительной системы с использованием диффузионных моделей

Для анализа механизмов подавления шума в зрительном восприятии была использована диффузионная модель, позволяющая имитировать процесс восстановления недостающей или искаженной информации и создавать детализированные и фотореалистичные изображения. Основной задачей исследования было изучение, какие признаки изображения сохраняются при постепенном добавлении и устранении шума, и как этот процесс соотносится с нейрофизиологическими механизмами обработки зрительной информации.

Методы

Создание зашумленных стимулов. В ходе исследования применяли диффузионную модель архитектуры Stable Diffusion (Rombach et al., 2022) для задачи направленной генерации изображений, соответствующих текстовому запросу. В отличие от сверточных нейронных сетей, кодирующих статистику зрительных категорий напрямую (см. разделы 4.1, 4.2, 4.3) и показывающих высокое сходство с кодированием информации зрительной корой приматов (Khaligh-Razavi, Kriegeskorte, 2014; Yamins, DiCarlo, 2016; Schrimpf et al., 2018) диффузионные модели работают со зрительными образами косвенным образом, обучаясь вычитать шум из поступающих изображений. На Рис. 56 представлен пример создания изображения диффузионными моделями, путем пошагового вычитания шума.

Рисунок 56. Пример создания изображения диффузионными моделями, путем пошагового вычитания шума

Были установлены следующие параметры: в качестве сэмплера, определяющего график восстановления зашумленного изображения, использовался k-LMS; количество шагов вывода варьировалось между 15, 30 и 50; применялась техника безклассификаторного направления, позволяющая улучшить релевантность изображения к заданному тексту.

Для каждого из указанных наборов шагов было создано 30 различных вариантов, определяемых ключом (seed). На графиках приведены усредненные данные для 90 различных траекторий генерации, полученных из одного текстового запроса, но с различными ключами. Количество шагов переведено в относительное значение для сопоставления траекторий с 15, 30 и 50 шагами.

Изображение, созданное на последнем шаге генерации был выбран в качестве эталонного изображения, предполагая, что все зрительные признаки, необходимые для распознавания зрительной сцены, на нем уже сформированы в достаточном виде. Далее, рассчитывалось сходство между отдельно взятым шагом генерации и эталоном. Подобный способ позволил нам количественно оценить выраженность зрительных признаков на всех этапах создания изображения.

Моделирование восприятия изображений в условиях неопределенности. В качестве входных данных использовались изображения, подвергавшиеся разным уровням шумового воздействия, моделирующим вариабельность сенсорного входа в биологической системе. Стимулы,

созданные при помощи диффузионных сетей, имеют количественное описание выраженности зрительных признаков.

Точность классификации в условиях неопределенности исследовалась с применением сверточной нейронной сети ResNet-50 (He et al., 2016). Проводилась оценки вероятности, с которой модель назначила корректный класс, изображению, созданному на определенном шаге расшумления.

Чтобы оценить выраженность зрительных признаков на промежуточных этапах генерации изображения, мы применили метод оценки структурного сходства изображений (Zhou Wang et al., 2004). *Индекс структурного сходства (SSIM)* используется для количественной характеристики воспринимаемого качества цифрового изображений и видео. Значение индекса равное 1 указывает на полное сходство двух изображений. Метод построен с учетом особенностей зрительного восприятия и рассматривает ухудшение изображения как воспринимаемое изменение структурной информации, а также маскирование яркости и контраста. Под структурной информацией понимается идея о том, что пространственно-близлежащие пиксели имеют сильные взаимозависимости, несущие важную информацию о структуре объектов в зрительной сцене. В случайном шуме, как на начальном этапе генерации, структурная информация отсутствует.

Результаты

В ранних этапах обработки диффузионная модель преимущественно восстанавливала низкочастотные компоненты изображения, соответствующие крупномасштабной организации объектов в сцене (Рис. 56). Это соответствует принципу обработки информации в зрительной системе, где на начальных уровнях коры преимущественно кодируются базовые характеристики, такие как ориентация и частотные компоненты. Более сложные детали изображения восстанавливались на поздних стадиях обработки, что согласуется с тем, как высшие зрительные области мозга интегрируют информацию.

Несмотря на постепенную детализацию образов (Рис. 56), результаты распознавания создаваемых изображений, аналогично психофизиологическим исследованиям, показали наличие скачкообразного перехода, отождествляемого с феноменом «инсайта» (Шелепин, Пронин, Шелепин, 2015). На этом этапе создаваемые стимулы получают практически однозначную интерпретацию принадлежности к определенной категории. Требуется более 60% итераций, чтобы изображения начали корректно классифицироваться. По достижению 80% происходит качественный скачок и изображения практически всегда (> 0.8) классифицируются корректно.

Результаты анализа изображений, создаваемых на промежуточных шагах генерации представлены на Рис. 57. Видно, что на первые 40% итераций структурная информация, измеряемая посредством SSIM, практически отсутствует. Во второй половине процесса происходит наиболее интенсивное формирование структурных элементов и других характеристик зрительной сцены.

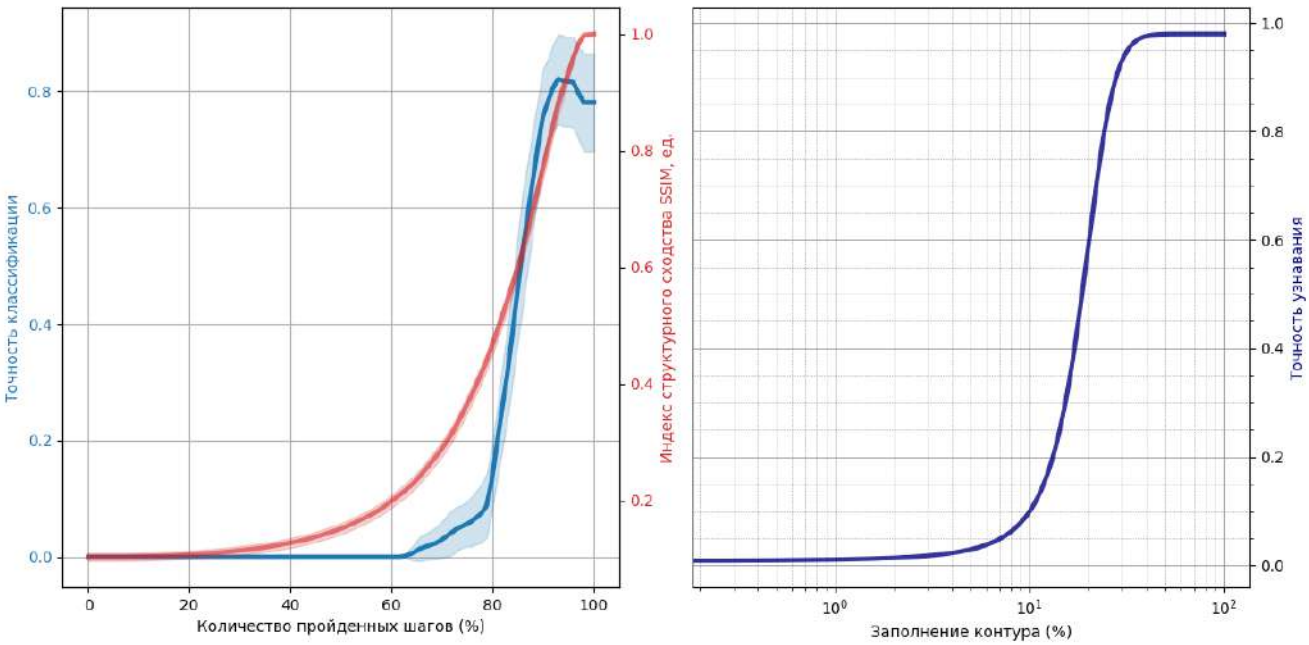

Рисунок 57. Зависимость точности распознавания объектов от степени неопределенности сигнала. (слева) Распознавание категории объекта нейронной сетью на созданном диффузионной моделью изображении. Синим цветом обозначена вероятность корректного распознавания класса. Красным - структурное сходство изображений с конечным результатом. (справа) Зависимость правильного распознавания объекта наблюдателем от процента заполнения контура изображения объекта

Представленная на Рис. 57 психометрическая кривая распознавания неполных изображений (Голлин-тест) отражает зависимость правильного распознавания от процента заполнения контура изображения объекта. Форма S образных кривых очень похожа и справа и слева. Но положение перегиба на оси абсцисс различно, так как связано с различным отношением сигнал/шум в анализируемых изображениях.

Проводя эмпирическую оценку (например, Рис. 56), можно заметить, что фундаментальные признаки сцены прослеживаются наблюдателем несколько раньше, чем фиксируются индексом SSIM, что может свидетельствовать о ограниченной применимости индекса.

Скачкообразный переход в точности распознавания может быть обусловлен тем, что сверточные нейронные сети в первую очередь опираются на локальные признаки изображения

(Geirhos et al., 2018), а не на его глобальную структуру. Хотя увеличение рецептивных полей в более глубоких слоях позволяет учитывать более обширный контекст, обработка сигнала в искусственных сетях остается привязанной к локальным статистическим характеристикам, таким как текстуры и градиенты. В отличие от этого, зрительная система человека интегрирует информацию на нескольких уровнях, что позволяет гибко переключаться между локальными деталями и общей структурой объекта.

Кроме того, фундаментальные различия в обучении биологической и искусственной нейронных систем могут играть важную роль в формировании этого перехода. Человеческая зрительная система с рождения функционирует в условиях постоянного зашумления сенсорного сигнала, где шум не является исключением, а представляет собой неотъемлемую часть процесса обработки информации. Мозг использует вероятностные стратегии, адаптивные механизмы подавления помех и интеграцию контекста, что позволяет стабильно распознавать объекты даже при значительных искажениях. В отличие от этого, искусственные нейросети, хоть и обучаются с зашумлением, возможно, не сталкиваются с теми типами помех, которые характерны для реального восприятия. В большинстве случаев добавление шума в процессе обучения нейросетей носит формальный характер и может не отражать сложные закономерности реальной среды, где шум часто коррелирован с освещением, движением и глубинными особенностями сцены.

Таким образом, скачкообразный переход в искусственных нейросетях может быть связан как с особенностями обработки локальных признаков, так и с условиями обучения. В искусственных системах появление выраженных локальных текстур может становиться доминирующим фактором при классификации, тогда как в биологической системе в момент инсайта происходит перестройка восприятия: внимание переключается с локальных деталей на глобальную форму объекта, что позволяет идентифицировать его целостный образ даже в условиях высокой неопределенности.

4.5.3. Распознавание мимики в условиях неопределенности

Исследования показывают, что восприятие эмоций зависит от множества факторов, включая освещение, выраженность мимической экспрессии, контекста и индивидуальных особенностей зрительной системы наблюдателя. Предполагается, что нейроны нижневисочной коры чувствительны к лицевой мимике (см. 4.4) и играют роль в оценке направления взгляда, положения уголков губ, глаз, бровей, степени их изгиба и взаимного расположения частей лица, что важно для опознания мимики. В данном разделе внимание уделяется изучению роли высших зрительных центров в обработке амбивалентных мимических выражений.

Методы

Для исследования механизмов распознавания улыбки в условиях неопределенности проведено два параллельных эксперимента (Жукова, Малахова, Шелепин, 2019): нейрофизиологическое исследование с использованием функциональной магнитно-резонансной томографии и моделирование работы искусственной нейронной сети сверточной архитектуры.

В нейрофизиологическом исследовании приняли участие 24 испытуемых (8 мужчин и 16 женщин) в возрасте 19–33 лет. В качестве стимулов использованы изображения виртуальных лиц, созданные в программном обеспечении FaceGen (FaceGen, 2023). Эти изображения были разделены на две группы:

- Надпороговые условия: лица с явно выраженной мимикой радости и грусти, а также с небольшими поворотами головы.
- Пороговые (неопределенные) условия: нейтральные выражения лиц без четких признаков эмоций.

Испытуемым последовательно предъявлялись изображения в четырех фазах стимуляции, различавшихся по наличию эмоций и задаче (определение эмоции или поворота головы). Регистрация мозговой активности проводилась на 1,5T MR-сканере Siemens Symphony, анализ данных осуществлялся методом BOLD-картирования в программе SPM8.

В моделировании использовалась сверточная нейросеть VGG-Face (Parkhi, Vedaldi, Zisserman, 2015), предварительно обученная на 2,6 млн изображений лиц. Для задачи распознавания эмоций сеть была дообучена на специализированном наборе данных, включавшем 1000 изображений виртуальных лиц с разной степенью выраженности эмоций. Анализ работы сети включал изучение карт активации сверточных фильтров и классификацию лиц в условиях четко выраженной и неопределенной мимики.

Результаты

В условиях надпорогового восприятия точность распознавания эмоций человеком составляла 83%, а поворотов головы 95%. Однако при предъявлении пороговых изображений вероятность распознавания мимики снижалась до случайного уровня (50%), указывая на высокую неопределенность восприятия. ФМРТ-данные показали, что при восприятии выраженной мимики и движений головы активируются фронтальные и теменные области, а также сеть базового режима (медиальная префронтальная кора, височно-теменной стык). В условиях неопределенности наблюдалась активация фузиформной извилины, нижневисочной коры и зон, связанных с избирательным вниманием, что указывает на попытки мозга компенсировать недостаточность информации.

Моделирование при помощи сверточной нейронной сети показало, что в надпороговых условиях сеть классифицировала улыбку с точностью 97%, что сопоставимо с человеческим

восприятием. Однако при пороговых изображениях (неявные эмоции, слабые контуры) вероятность корректного распознавания снижалась до 50%, аналогично испытуемым. Анализ карт активации показал, что сеть использует многоуровневую фильтрацию: начальные слои выделяют простые признаки (границы, цветовые градиенты), а глубокие – сложные паттерны, характерные для лиц. Однако в условиях неопределенности сеть демонстрировала нестабильные активации, что указывает на недостаток семантической интерпретации, присущей биологической зрительной системе.

Распознавание улыбки Джоконды оказалось особенно сложным. Вероятность ее классификации как «улыбка» составила 0,69, что подтверждает субъективную неопределенность восприятия этого изображения как человеком, так и нейросетью.

Обсуждение результатов

В данной главе представлен комплексный анализ механизмов обработки зрительной информации в нейронных сетях в модельных исследованиях и посредством анализа нейрофизиологических данных. Основные результаты:

Предложен новый метод исследования кодирования зрительной информации посредством оценки схожести пространств описания. Анализ функциональных пространств нейронных сетей показал, что на ранних этапах обработки информация сохраняет сходство со статистическими характеристиками *входного сигнала* (коэф. корр. = 0,62), тогда как на более высоких уровнях представления становятся оптимизированными для категоризации (коэф. корр. = 0,34). Выявлен *«переходный» этап обработки* в промежуточных слоях, где наблюдается резкое снижение корреляции как со входными характеристиками (с 0,62 до 0,16), так и с *пространством задачи* (около 0,1), что указывает на радикальную трансформацию представлений. Таким образом влияние задачи распространяется и на сверточные слои, затрагивая последние 4 слоя из 13, в то время как статистика сцены формирует представления в первых двух-трех слоях. Однако воздействие этих фундаментальных факторов не оказывается монотонным и не распространяется на промежуточные слои.

Схожие идеи о начальных этапах обработки зрительного сигнала, где информация проходит стадию декорреляции для устранения избыточности и оптимального распределения статистических характеристик сцены, были изложены в работах (Глезер, Цуккерман, 1961; Barlow, 1961; Field, 1994; Vinje, Gallant, 2000; Pitkow, Meister, 2012). Однако, на более высоких уровнях обработки происходит консолидация представлений. В исследованиях (Шелепин, 1984; Field, 1987, 1999; Olshausen, Field, 2004) отмечалось, что увеличение связности представлений наблюдаемого в высокоуровневых областях мозга способствует формированию инвариантных кодов и более эффективной обработке сложных зрительных сцен.

Свойство оппонентности и его фундаментальное влияние обработку информации можно рассматривать как универсальный принцип перераспределения информации, обеспечивающее баланс между точностью представления сигнала и обобщением, необходимым для решения задач биологическими и искусственными системами.

Наличие переходного этапа трансформация представлений может отражать переструктурирование информации перед ее интеграцией в семантические категории. Интересно, что аналогичная неопределенность представлений наблюдается в биологической зрительной системе, например, в промежуточных областях вентрального пути, следующих за первичной/вторичной зрительной корой, где нейроны демонстрируют сложные и вариативные характеристики ответа.

Исследование семантической близости категорий выявило, что в высших слоях увеличивается схожесть представлений внутри некоторых пар категорий, несмотря на обучение сети их различению. Например, для пород собак *коэффициент схожести* возрастает с 0,09 до 0,37, а для некоторых видов птиц – с 0,26 до 0,33, что указывает на формирование инвариантных признаков образов, устойчивых к несущественным вариациям в поставленной задаче.

Применено два подхода для изучения свойств представления категории на различных этапах обработки сигнала искусственной нейронной сетью на основе анализа ответа модели на 50000 изображений.

Анализ значимых направлений, участвующих в кодировании категории, показал, что их количество увеличивается с глубиной слоя, причем на тренировочных данных до 93% фильтров активно участвуют в распознавании, тогда как на валидационных данных этот показатель снижается на ~30% (до 57–70%). Данное наблюдение свидетельствует о наличии эффекта переобучения, при котором сеть адаптируется к специфическим особенностям обучающего набора, но менее эффективно обобщает знания на новые данные. Это также указывает на недоиспользование «мощностей» модели, так как значительная часть фильтров остается слабо активированной при обработке новых данных, что может свидетельствовать о неравномерном распределении информативности признаков и потенциале для улучшения обобщающей способности сети.

Предложен *метод построения прототипов категорий* на основе ковариационной матрицы, выявляющий важные аспекты представления информации популяцией нейронов. Показано преимущество метода по сравнению с анализом индивидуальных активаций, особенно на средних слоях (block2_conv2 – block4_conv2), где предложенный метод обеспечивает прирост точности определения категории до 15–20%. Это дополнительно свидетельствует о том, что на промежуточных этапах обработки зрительных образов, информация о категориях распределена в виде взаимосвязей между активациями нейронов, а не закодирована в отдельных элементах.

Однако на поздних слоях (block5_conv3) происходит консолидация представлений, и категории начинают кодироваться более локализовано, снижая значимость согласованной активности.

Анализ главных компонент, задействованных в кодировании категории, показал, что в начальных слоях для описания 95% дисперсии достаточно 10 главных компонент, тогда как в высших слоях требуется до 300 компонент.

Таким образом, *популяционный анализ активности* показал, что для понимания категоризации в нейронных сетях важно учитывать отклики не только отдельных нейронов, но и паттерны активности на уровне целых слоев. Если традиционные подходы предполагают, что на высших уровнях обработки зрительного сигнала категории представлены относительно компактно, полученные данные указывают на обратное: чем выше уровень представления, тем сложнее и «многомернее» становится категория. Например, вместо представления о том, что все изображения кошек активируют один компактный набор нейронов, а изображения собак – другой, с четкой границей между ними, нейронные представления могут перекрываться и изменяться в зависимости от контекста, позволяя системе адаптировать границы категорий, обобщать ранее невиданные объекты и учитывать разнообразие их характеристик.

Если распознавание объектов происходит за счет совместной активности нейронных ансамблей, это означает, что регистрируемые ответы одиночных нейронов не отражают полную картину организации представлений. Для более точного понимания таких процессов необходима регистрация нейронной активности с высоким пространственным разрешением и методы многомерного анализа популяционной активности, что позволило бы выявлять скрытые пространственно-временные закономерности, недоступные при анализе отдельных сигналов.

Рассмотрена возможность *построения прототипа представления зрительной категории на ограниченном наборе изображений.* Показано, что для большинства слоев достаточно 50–100 изображений для приближения к эталонному прототипу, однако на промежуточных слоях (block2_conv2 – block5_conv2) точность аппроксимации снижается, особенно если используются ранее не предъявляемые стимулы. Это свидетельствует о сложности сформированных представлений и ограниченности малых выборок в передаче всех воспринимаемых аспектов категории. На последнем слое (block5_conv3) расстояние до эталонного прототипа уменьшается, указывая на переход к семантическому кодированию, при котором сеть выделяет ключевые признаки, игнорируя шум и вариативность зрительного сигнала.

Исследование прототипов категорий выявило, что формирование обобщенного представления требует относительно небольшого количества примеров, но само представление становится значительно сложнее на более высоких уровнях сети. Рост числа главных компонент, необходимых для описания категорий, предполагает, что высокоуровневое кодирование

использует многомерное представление, недоступное для прямой интерпретации. Это ставит вопросы о том, насколько интерпретируемыми могут быть высшие уровни обработки.

При этом различия между категориями уменьшаются, а внутрикатегориальное разнообразие растет. Это проявляется в том, что представления разных объектов внутри одной категории не схлопываются в единый центр, а остаются распределенными, охватывая множество возможных вариаций. Например, нейронные активности для разных пород собак или различных видов птиц начинают перекрываться несмотря на то, что сеть обучалась различать их. Границы между категориями становятся нечеткими, что может свидетельствовать о том, что система кодирования адаптируется к возможному расширению категорий и учету новых стимулов.

Проведен анализ избирательности к лицам в искусственных нейросетях и определено, что в VGG16 наблюдается больше нейронов, специализированных на обработку лиц, по сравнению с VGG-Face, обученной распознаванию личностей, несмотря на отсутствие категории «лицо» в обучающем наборе. Показано, что, несмотря на обучение VGG-Face исключительно на лицах, в VGG16 наблюдается больше нейронов, избирательных к лицам по метрике *CS*, даже при отсутствии категории «лицо» в обучающем наборе. В VGG16 такие нейроны сосредоточены в слоях block4_conv3 и block5_conv2 (до ≈80 нейронов), тогда как в VGG-Face их максимум составляет около 7 в conv5_3. Это свидетельствует о том, что VGG16 формирует универсальные механизмы обработки зрительных признаков, тогда как VGG-Face адаптируется к различению индивидуальных лиц, а не к выделению лиц в целом.

Рассмотрен *метод аппроксимации функций нейронов через фрагменты изображений* показавший, что функции нейронов нижневисочной коры можно аппроксимировать через набор признаков, включающих локальные ориентации и цвета, извлеченные из фрагментов натуральных изображений (коэф. корр. = 0,51 для всех стимулов; 0,37 для лиц). Применение этого подхода к искусственным нейросетям позволило выявить аналогичные закономерности в обработке зрительной информации, демонстрируя возможность аппроксимации глубоких слоев (VGG-16, слой 13) через комбинации локальных ориентаций и цветов (коэф. корр. = 0,44).

Таким образом, несмотря на сложные и разреженные представления на высших этапах обработки сигнала, базовые характеристики активности в этих слоях удается достаточно аппроксимировать посредством комбинаций простых признаков. С одной стороны, этот результат открывает перспективу создания неглубоких интерпретируемых моделей, в которых сложные вычисления представлены в виде набора четко интерпретируемых зрительных компонентов. С другой стороны, он подчеркивает ограниченность анализа отдельных нейронов, поскольку, рассматривая их изолированно, можно прийти к упрощенным выводам о характере обработки информации. Это усиливает необходимость применения продвинутых методов

нейронной регистрации и анализа популяционной активности, позволяющих захватить глобальные закономерности, которые невозможно выявить, изучая нейроны по отдельности.

Проведено исследование процесса распознавания образов в условиях неопределенности с использованием диффузионных моделей. Показано, что восстановление изображения в модели проходит через последовательные этапы, соответствующие разным уровням обработки информации в зрительной системе: на ранних шагах восстанавливаются низкочастотные компоненты и глобальная структура сцены, тогда как высокочастотные детали появляются на более поздних стадиях. Анализ точности распознавания зашумленных изображений точности классификации показал наличие скачкообразного перехода, аналогичного феномену «инсайта» в психофизиологических экспериментах. После 60% итераций изображения начинают распознаваться с высокой вероятностью, а после 80% шагов классификация становится практически однозначной.

Рассмотренные методы и полученные результаты дополняют друг друга, предоставляя возможность изучить проблему кодирования зрительной информации с разных перспектив: от анализа общей трансформации пространств представления до интерпретации базовых функциональных единиц этой системы. Такой многоуровневый подход не только углубляет понимание принципов работы как искусственных, так и биологических нейронных сетей, но и позволяет выявить как общие закономерности в организации систем зрительного восприятия, так и ключевые различия, а также ограничения в применении моделей.

Выводы

1. *Активность нейронов нижневисочной коры* головного мозга приматов (область TEad), усиливается при предъявлении изображений лица по сравнению с изображениями других объектов. Средний вызванный ответ на лица был выше в 130 из 143 точек регистрации активности нейрональных колонок. В 118 точках регистрации возрастало стандартное при предъявлении тестов в виде изображений лиц (коэф. корр. = 0,58, $p \leq 0{,}001$).
2. *Индекс избирательности* к категории лиц был высок и составил K = 0,06 для изображений лиц людей и K = 0,10 для изображений лиц обезьян. Показано, что искусственные объекты и ландшафты демонстрируют низкий уровень избирательности по сравнению с другими категориями, что проявляется в средних значениях индекса K = -0,10 и -0,14 соответственно.
3. *Анализ пространственной организации* показал, что нейрональная активность в нижневисочной коре имеет высокую корреляция ответов, характерную для локально организованных ансамблей (коэф. корр. = 0,41, $p \leq 0{,}001$), которая снижается с увеличением расстояния между точками регистрации и имеет нелинейный характер и может быть описана как степенной (R^2=0.25) так логарифмической зависимостью (R^2=0.24).
4. *Моделирование нейрональных ответов с помощью сверточных нейронных сетей* продемонстрировало высокую точность предсказаний реакций нейрональных колонок нижневисочной коры (коэф. корр. = 0,68, $p < 0{,}001$). Также выявлено соответствие между глубокими слоями сети (block5_conv2) и ответами нейрональных колонок. Полученные результаты указывают на сходство в механизмах обработки информации биологическими и искусственными нейронными сетями, выполняющими задачу зрительного распознавания. При этом точность модели зависела от стабильности ответа нейрональных колонок (коэф. корр. = 0,7, $p < 0{,}01$), рассчитанного на основе реакции в четных и нечетных блоках стимуляции.
5. Предложен и разработан *адаптивный метод генерации стимулов на основе анализа ответов нейронов нижневисочной коры*, позволяющий создавать оптимизированные изображения, вызывающие целенаправленную активацию нейронных ансамблей. В отличие от традиционных подходов, подобный метод «генерации стимулов на основе обратной связи» дает возможность тестирования функциональной специализации нейронов без ограничения фиксированными наборами стимулов. *Анализ двух подходов к созданию стимулов*, показал различия в эффективности их применения: функция потерь

Evoked (коэф. корр. = 0,42) обеспечивала максимальную активацию нейронов, в то время как *Softmax* (коэф. корр. = 0,31) оказалась в среднем менее эффективной, но продемонстрировала большую устойчивость к «утомлению», а точнее снижению общего уровня нейронального ответа нижневисочной коры во втором блоке эксперимента (коэф. корр. = 0,35 против -0,09 для *Evoked*).

6. Применение разработанного метода *предсказания ответов нейронов посредством фрагментов естественных изображений* показало, что ответ нейронов нижневисочной коры можно предсказать на основе набора признаков V1/V2, включающих локальные ориентации и цвета (коэф. корр. = 0,51 для всех стимулов; 0,37 для лиц). Применение этого подхода к искусственным нейросетям позволило выявить аналогичные закономерности в обработке зрительной информации, демонстрируя возможность предсказания ответа глубоких слоев (VGG-16, слой 13) через комбинации локальных ориентаций и цветов (коэф. корр. = 0,44).
7. *В модельных исследованиях установлены оппонентные отношения* между пространством описания сигнала и пространством постановки задачи, а также их влияние на представление информации на разных этапах зрительной обработки. Начальные слои сохраняют структуру, близкую к статистике входного сигнала (коэф. корр. = 0,62), тогда как последние сверточные слои адаптируются под задачу (коэф. корр. = 0,34), хотя традиционно рассматриваются как этап выделения зрительных характеристик. В завершающих полносвязных слоях происходит консолидация представлений (коэф. корр. = 0,74), подчиняясь структуре задачи. Обнаружен фазовый характер обработки: сначала резкий разрыв связей с входными данными, затем этап трансформации представлений и распределенного кодирования, а на заключительном этапе – консолидация признаков под задачи распознавания.
8. *Исследование устойчивости к шуму в задаче распознавания образа* показало наличие скачкообразного перехода у искусственных нейронных сетей, аналогично достижению порогового уровня восприятия, например феномену «инсайта» в экспериментах с Голлин-тестом. После 60% итераций изображения начинают распознаваться с высокой вероятностью, а после 80% шагов классификация становится практически однозначной. Однако причины подобного перехода у искусственных нейронных сетей могут объясняться различными механизмами кодирования признаков о категории объекта.

Заключение

В диссертационной работе исследуется процесс формирования и распознавания зрительных образов в высших отделах зрительной системы приматов. Основное внимание уделяется механизмам кодирования зрительной информации нейронами нижневисочной коры, особенно при восприятии лиц и других категорий изображений. Разрабатываются и применяются новые методы изучения функциональных характеристик нейронов и нейронных сетей, а также совершенствуется сам процесс экспериментального поиска.

Одной из задач данного исследования является преодоление разрыва между классическими нейрофизиологическими методами, которые фокусируются на свойствах отдельных нейронов, и современными вычислительными подходами, позволяющими анализировать большие и сложные массивы данных. Такой синтез дает более целостное понимание того, каким образом высшие отделы зрительной системы формируют зрительные представления и обрабатывают сложные зрительные сцены. Помимо исследования ответа нейронов нижневисочной коры, особое внимание уделяется нейронной активности на уровне популяции, а также пониманию роли и задачи каждого этапа в процессе распознавания как единого целого – от пространства сигнала до решения задач, поставленных перед наблюдателем.

Методологически значимым является разработанный подход, совмещающий современные генеративные модели искусственного интеллекта с нейрофизиологическим экспериментом, позволяющий целенаправленно создавать оптимальные стимулы для активации конкретных нейронных популяций, что открывает новые возможности для исследования функциональной специализации нейронов.

Использование сверточных нейронных сетей в качестве моделей биологических систем иллюстрирует, как в ходе обучения на реальных или синтетических данных формируются иерархические представления о зрительных объектах. Важным результатом стало понимание степени влияния обучающих данных и поставленной задачи на формирование внутренних представлений в нейронной сети. Показано наличие продолжительной переходной зоны, где исходные статистические признаки сцены и целевые задачи категоризации перестраиваются в новое многомерное пространство. На этом этапе, по сути, происходит перераспределение информации внутри популяции нейронов.

Также показано снижение уровня вовлеченности нейронов в обработку новой информации, по сравнению с ранее предъявляемой, что может иметь важные импликации как для понимания механизмов научения в биологических системах, так и для разработки более эффективных алгоритмов машинного обучения с улучшенной способностью к обобщению.

Исследование кодирования категорий показало, что вопреки ожиданиям о повышении компактности описания образов по мере продвижения по иерархической системе, структура представлений на этих уровнях оказывается многомерной. В частности, для описания 95% дисперсии ответа слоя требуется использовать до 80% главных компонент от максимально возможного их числа. Это демонстрирует значительное усложнение представлений образов с увеличением глубины обработки, вопреки представлению о прогрессивном упрощении и консолидации информации в иерархических системах.

Применение метода предсказания нейронных ответов посредством фрагментов естественных изображений подтвердило, что функции нейронов нижневисочной коры могут быть успешно аппроксимированы комбинациями локальных ориентаций и цветов. Аналогичные закономерности обнаружены при применении этого подхода к искусственным нейронным сетям, что указывает на сходство в механизмах обработки зрительной информации.

Проведено исследование процесса распознавания образов в условиях неопределенности с применением современных диффузионных моделей. Установлено, что точность распознавания зашумленных изображений моделью зависит от меры накопления структурной информации, и подчиняется характерному скачкообразному переходу, наблюдаемому при пороговом восприятии в сенсорных системах, когда при преодолении определенного информационного уровня классификация трансформируется из случайной в практически однозначную.

Наконец, сопоставление результатов, полученных на биологических нейронах и искусственных нейронных сетях, раскрывает как фундаментальные сходства, так и существенные различия в принципах организации этих систем. Построенная модель нейронных откликов позволила провести количественный анализ характеристик кодирования информации, выявив ключевые закономерности организации зрительного восприятия.

В основе моделей для исследования механизмов распознавания зрительных образов использованы сверточные нейронные сети, поскольку они демонстрируют высокую эффективность в задачах компьютерного зрения и имеют принципиальные сходства с архитектурой зрительной системы приматов. Однако их применение имеет ряд ограничений. Несмотря на способность воспроизводить ключевые этапы иерархической обработки зрительной информации, сверточные сети остаются упрощенными моделями и не учитывают ряд фундаментальных свойств биологических нейронных систем. В частности, в них отсутствуют латеральные связи между нейронами, механизмы обратной связи, а также рекуррентные процессы, играющие важную роль в обработке информации в коре головного мозга.

Помимо структурно-функциональных различий и различий в выполняемой задаче, возможны также и расхождения в статистике стимульного материала, используемого для обучения моделей и предъявляемого в нейрофизиологических экспериментах. Несмотря на то, что сверточные

нейронные сети обучаются на огромных наборах изображений (миллионы экземпляров), экспериментальные стимулы нейрофизиологических исследований, оптимизируемые под получение стабильного нейронального ответа – нейтральный фон, контролируемый контраст, и т. д., – значительно отличаются от изображений, на которых тренируются модели, что может влиять на способность моделей предсказывать ответ на подобные стимулы. Помимо того, обучающие наборы данных для СНС существенно отличаются от естественного зрительного окружения, смещая распределение статистики зрительного сигнала. Указанные факторы могут приводить к несоответствиям в формируемых представлениях внутри сетей.

Кроме того, в искусственных нейронных сетях описания формируются исходя из целевой функции, которая оптимизировалась в ходе обучения. Например, сеть, обученная на классификации изображений, будет формировать описания признаков, полезных для различения классов. А сеть, обученная на предсказании движения, будет выделять динамические особенности видеоряда. В биологических нейронных сетях, таких как зрительная система, описания также формируются исходя из решаемых организмом задач. Например, в высших отделах зрительной коры приматов есть нейроны, избирательно реагирующие на лица. Это обусловлено важностью распознавания лиц для социального взаимодействия. Однако среда, в которой находятся биологические организмы повседневно, существенно отличается от искусственно создаваемых обучающих множеств. Различны их статистические характеристики, но еще более различны задачи, стоящие перед биологические организмами и ставящиеся перед искусственными системами.

Таким образом, для правильной интерпретации кодирования информации важно анализировать не только стимулы и ответы, но и функциональную роль сети в решении поставленной перед ней задачи. Это позволит получить более глубокое понимание принципов обработки информации.

Стоит отметить, что в практически во всех текущих исследования изучение кодирования информации в вентральном пути проводится посредством предъявления и анализа статических изображений. Однако в реальных условиях объекты динамичны, и важную роль играют временные аспекты обработки информации. Возможным направлением дальнейших исследований представляется изучение динамических аспектов распознавания образов. Понимание того, как меняется представление объекта в зрительной системе по мере поступления новой информации, смене контекста, либо выполняемой задачи, позволит получить более полную картину механизмов распознавания объектов в динамических условиях реального мира.

Список сокращений и условных обозначений

ИСКЛ – слой исключения

ИНС – искусственная нейронная сеть

(ф)МРТ – (функциональная) магнитно-резонансная томография

ПС – полносвязный слой

СНС – сверточная нейронная сеть

ЭКГ – электрокардиография

ЭЭГ – электроэнцефалограмма

AIT – передняя нижневисочная кора (Anterior Inferior Temporal Cortex)

AMTS – передняя средняя височная борозда (Anterior Middle Temporal Sulcus)

AUC ROC – площадь под кривой ошибок (Area Under The ROC Curve)

BA – поле Бродмана (Brodmann area)

CNN – сверточная нейронная сеть (Convolutional Neural Network)

CS – индекс избирательности к категории (Category Selectivity Index)

DV – дорсо-вентральная ось (Dorsoventral axis)

EEG – электроэнцефалограмма

FSI – индекс избирательности к лицам (Face Selectivity Index)

GAN – генеративно-состязательная сеть (Generative Adversarial Network)

HMAX – Hierarchical Models and X (модель иерархической обработки зрительной информации)

IT/ITC – нижневисочная кора (Inferior Temporal Cortex)

LO – латеральная затылочная область (Lateral Occipital)

MRI – магнитно-резонансная томография

NIH – Национальные институты здоровья (National Institutes of Health)

PCA – метод главных компонент (Principal Component Analysis)

PPGN – Plug-n-Play Generative Networks

PIT – задняя нижневисочная кора (Posterior Inferior Temporal Cortex)

RDM – матрица репрезентативного несходства (Representational Dissimilarity Matrix)

ReLU – выпрямленная линейная функция активации (Rectified Linear Unit)

RSA – анализ репрезентативного сходства (Representational Similarity Analysis)

SD – стандартное отклонение (Standard Deviation)

SSIM – индекс структурного сходства (Structural similarity index measure)

STS – верхняя височная борозда (Superior Temporal Sulcus)

TEad – область нижневисочной коры (dorsal part of anterior TE)

V1 – первичная зрительная кора

V2 – вторичная зрительная кора

V4 – четвертая область зрительной коры

VAE – вариационный автоэнкодер (Variational Autoencoder)

ViT – Модель трансформер для изображений (Vision Transformer)

Список литературы

1. Бабкин Б. П. Опыт систематического изучения сложно-нервных (психических) явлений у собаки (Диссертация) // 1904. СПб: ВМА, 95 с.
2. Вахрамеева О. А., Харауэов А. К., Пронин С. В., Малахова Е. Ю., Шелепин Ю. Е. Зрительный прайминг при распознавании мелких изображений в сцене содержащей объекты разного размера // Физиология человека, 2016, том 42, № 5, с. 39–48.
3. Глезер В. Д. Зрение и мышление. Л.: Наука, 1985.
4. Глезер В. Д. Механизмы опознания зрительных образов. Л.: Наука, 1966. 204 с.
5. Глезер В. Д., Дудкин К. Н., Подвигин Н. Ф., Невская А. А., Праздникова Н. В. Зрительное опознание и его нейрофизиологические механизмы. М.: Наука, 1975. 272 с.
6. Глезер, В. Д., Цуккерман, И. И. Информация и зрение. Издательство: М., 1961. 182 с.
7. Глезер В. Д., Цуккерман И. И. О дублировании каналов связи в зрительном анализаторе // Биофизика. 1959. Т. 4. № 5. С. 620–621.
8. Жеребко А. К., Луцив В. Р. Согласованная фильтрация в естественных и искусственных нейронных сетях // Оптический журнал. 1999. Т. 66. № 9. С. 69–72.
9. Жукова О. В., Малахова Е. Ю., Шелепин Ю. Е. Джоконда и неопределенность распознавания улыбки человеком и искусственной нейронной сетью // Оптический журнал. 2019. Т. 86. № 11. С. 40–49.
10. Кезели А. Р. Нейрофизиологические механизмы цветового зрения. Тбилиси: Мецниереба, 1983. 184 с.
11. Кок Е. П. Зрительные агнозии: Синдромы расстройств высших зрительных функций при односторонних поражениях височно-затылочной и теменно-затылочной области мозга. Л.: Медицина, 1967.
12. Красильников Н. Н. Влияние шумов на контрастную чувствительность и разрешающую способность приемной телевизионной трубки // Техника телевидения. 1958. № 25. С. 26–43.
13. Красильников Н. Н. Теория передачи и восприятия изображений. Теория передачи изображений и ее приложения. М., 1986.
14. Красильников Н. Н. Цифровая обработка 2D и 3D изображений. БХВ-Петербург, 2011.
15. Красильников Н. Н. Цифровая обработка изображений. М.: Вузовская книга, 2001. 320 с.
16. Луцив В. Автоматический анализ изображений: Объектно-независимый структурный подход. LAP LAMBERT Academic Publishing, 2011. 308 с.
17. Малахова Е. Ю. Визуализация информации, кодируемой нейронами высших областей зрительной системы // Оптический журнал. 2018. Т. 85. № 8. С. 61–66.
18. Малахова Е. Ю. Представление категорий посредством прототипов согласованной активности нейронов в сверточных нейронных сетях // Оптический журнал. 2021. № 12.

19. Малахова Е. Ю. Пространство описания зрительной сцены в искусственных и биологических нейронных сетях // Оптический журнал. 2020. Т. 87. № 10.

20. Подвигин Н. Ф. Динамические свойства нейронных структур зрительной системы. Л.: Наука, 1979. 158 с.

21. Поппер К. «Логика научного исследования». М.: АТС: Астрель, 2010.

22. Праздникова Н. В., Глезер В. Д., Данилова В. Ф. Два вида обобщения в эволюции зрительного мозга // Журн. эволюцион. биохим. и физиол. 1985. Т. 21. № 5. С. 435–442.

23. Праздникова Н. В., Данилова В. Ф., Мешкенайте В. И. Роль полей 7, 21 дорсальной и вентральной зон латеральной супрасильвиевой области коры головного мозга кошки в зрительном опознании // Сенсорные системы. 1989. Т. 3. № 3. С. 292–301.

24. Тибилов А. С., Васильев В. Н. Формирование сигнала биполяра палочек при малых освещенностях // Оптический Журнал. 2020. Т. 87. № 12. С. 76–84.

25. Цуккерман И. И. Статистическая структура изображений и особенности зрительного восприятия // Переработка информации в зрительной системе. Л.: Наука, 1975.

26. Шевелев И. А. Нейроны зрительной коры. Адаптивность и динамика рецептивных полей. М.: Наука, 1984. 232 с.

27. Шелепин Ю. Е. Локализация областей зрительной коры кошки, дающих инвариантный ответ при изменении размера изображения // Нейрофизиология. 1973. Т. 5. № 2. С. 115–121.

28. Шелепин Ю. Е. Ориентационная избирательность и пространственно-частотные характеристики рецептивных полей нейронов затылочной коры кошки // Нейрофизиология. 1981. Т. 13. № 3. С. 227–232.

29. Шелепин Ю. Е. Пространственно-частотные характеристики рецептивных полей нейронов латеральной супрасильвиевой области коры больших полушарий кошки // Нейрофизиология. 1982. Т. 14. № 6.

30. Шелепин Ю. Е. Сопоставление топографических и пространственно-частотных характеристик латеральной супрасильвиевой и стриарной коры кошки // Нейрофизиология. 1984. Т. 16. № 1. С. 35–41.

31. Шелепин Ю. Е., Колесникова Л. Н., Левкович Ю. И. Визоконтрастометрия. Наука, 1985.

32. Шелепин Ю. Е., Чихман В. Н., Фореман Н. Анализ исследований восприятия фрагментированных изображений: целостное восприятие и восприятие по локальным признакам // Российский физиологический журнал им. И. М. Сеченова. 2008. Т. 94. № 7. С. 758–776.

33. Шелепин К. Ю., Пронин С. В., Шелепин Ю. Е. Распознавание фрагментированных изображений и возникновение «инсайта» // Оптический журнал. 2015. Т. 82. № 10. С. 70–78.

34. Яковлев В. В. Различия в описании зрительного образа на уровне заднетеменной и нижневисочной коры обезьян // Физиология. 1982. С. 754–757.

35. Adrian E. D. The Basis of Sensation // Jason W. Brown Library. 1928.

36. Ala-Laurila P., Rieke F. Coincidence detection of single-photon responses in the inner retina at the sensitivity limit of vision // Curr Biol. 2014. Т. 24. № 24. С. 2888–2898.

37. Amedi A., Stern W. M., Camprodon J. A., Bermpohl F., Merabet L., Rotman S., Hemond C., Meijer P., Pascual-Leone A. Shape conveyed by visual-to-auditory sensory substitution activates the lateral occipital complex. Nature Neuroscience. 2007. Т. 10, № 6. С. 687–689.

38. Aparicio P. L., Issa E. B., DiCarlo J. J. Neurophysiological Organization of the Middle Face Patch in Macaque Inferior Temporal Cortex // J. Neurosci. 2016. Т. 36. № 50. С. 12729–12745.

39. Atick J. J., Redlich A. N. What Does the Retina Know about Natural Scenes? // Neural Computation. 1992. Т. 4. № 2. С. 196–210.

40. Averbeck B. B., Latham P. E., Pouget A. Neural correlations, population coding and computation // Nat Rev Neurosci. 2006. Т. 7. № 5. С. 358–366.

41. Baker N., Lu H., Erlikhman G., Kellman P. J. Deep convolutional networks do not classify based on global object shape // PLOS Computational Biology. 2018. Т. 14, № 12. С. e1006613.

42. Barlow H. B. Possible Principles Underlying the Transformations of Sensory Messages // Sensory Communication. 1961.

43. Bashivan P., Kar K., DiCarlo J. J. Neural population control via deep image synthesis // Science. 2019. Т. 364. № 6439.

44. Batschelet E. Circular statistics in biology. London, New York: Academic Press, 1981.

45. Bau D., Zhou B., Khosla A., Oliva A., Torralba A. Network dissection: quantifying interpretability of deep visual representations // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI: IEEE, 2017. С. 3319–3327.

46. Bau D., Zhu J.-Y., Strobelt H., Zhou B., Tenenbaum J. B., Freeman W. T., Torralba A. GAN dissection: visualizing and understanding generative adversarial networks // arXiv. 2018.

47. Bell A. H., Malecek N. J., Morin E. L., Hadj-Bouziane F., Tootell R. B. H., Ungerleider L. G. Relationship between functional magnetic resonance imaging-identified regions and neuronal category selectivity // The Journal of Neuroscience: The Official Journal of the Society for Neuroscience. 2011. Т. 31, № 34. С. 12229–12240.

48. Bell A. J., Sejnowski T. J. The «independent components» of natural scenes are edge filters // Vision Res. 1997. Т. 37. № 23. С. 3327–3338.

49. Bentin S., Allison T., Puce A., Perez E., McCarthy G. Electrophysiological studies of face perception in humans // Journal of Cognitive Neuroscience. 1996. Т. 8, № 6. С. 551–565.

50. Bialek W., Owen W. G. Temporal filtering in retinal bipolar cells. Elements of an optimal computation? // Biophys J. 1990. Т. 58. № 5. С. 1227–1233.

51. Bondar I. V., Leopold D. A., Richmond B. J., Victor J. D., Logothetis N. K. Long-term stability of visual pattern selective responses of monkey temporal lobe neurons // PLOS ONE. 2009. Т. 4, № 12. С. e8222.

52. Brazier M. A. B. Physiological Mechanisms Underlying the Electrical Activity of the Brain // Journal of Neurology, Neurosurgery & Psychiatry. 1948. Т. 11. № 2. С. 118–133.

53. Bundell S. Exploring the human brain with virtual reality // Nature. 2019.

54. Cadieu C. F., Hong H., Yamins D. L. K., Pinto N., Ardila D., Solomon E. A., Majaj N. J., DiCarlo J. J. Deep neural networks rival the representation of primate IT cortex for core visual object recognition // PLOS Computational Biology. 2014. Т. 10, № 12. С. e1003963.

55. Cadieu C. F., Olshausen B. A. Learning Intermediate-Level Representations of Form and Motion from Natural Movies // Neural Computation. 2012. Т. 24. № 4. С. 827–866.

56. Campbell F. W., Green D. G. Monocular versus binocular visual acuity // Nature. 1965. Т. 208. № 5006. С. 191–192.

57. Campbell F. W., Robson J. G. Application of fourier analysis to the visibility of gratings // The Journal of Physiology. 1968. Т. 197. № 3. С. 551–566.

58. Carandini M., Heeger D. J. Normalization as a canonical neural computation // Nat Rev Neurosci. 2011. Т. 13. № 1. С. 51–62.

59. Cunningham J. P., Yu B. M., Gilja V., Ryu S. I., Shenoy K. V. Toward optimal target placement for neural prosthetic devices // Journal of Neurophysiology. 2008. Т. 100, № 6. С. 3445–3457.

60. Dan Y., Atick J. J., Reid R. C. Efficient Coding of Natural Scenes in the Lateral Geniculate Nucleus: Experimental Test of a Computational Theory // J. Neurosci. 1996. Т. 16. № 10. С. 3351–3362.

61. Dayan P., Abbott L. F. Theoretical neuroscience: computational and mathematical modeling of neural systems. Cambridge, Mass: Massachusetts Institute of Technology Press, 2001. 460 с.

62. Dehaene S., Naccache L., Le Clec'H G., Koechlin E., Mueller M., Dehaene-Lambertz G., van de Moortele P. F., Le Bihan D. Imaging unconscious semantic priming // Nature. 1998. Т. 395, № 6702. С. 597–600.

63. DiCarlo J. J., Zoccolan D., Rust N. C. How does the brain solve visual object recognition? // Neuron. 2012. Т. 73. № 3. С. 415–434.

64. DiMattina C., Zhang K. Adaptive stimulus optimization for sensory systems neuroscience // Front. Neural Circuits. 2013. Т. 7.

65. Dodge S., Karam L. A study and comparison of human and deep learning recognition performance under visual distortions: 26th International Conference on Computer Communications and Networks, ICCCN 2017 // 2017 26th International Conference on Computer Communications and Networks, ICCCN 2017. 2017.

66. Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., et al. An image is worth 16x16 words: transformers for image recognition at scale // arXiv. 2020.

67. Dosovitskiy A., Brox T. Generating Images with Perceptual Similarity Metrics based on Deep Networks // Advances in Neural Information Processing Systems. Curran Associates, Inc., 2016.

68. Downing P. E., Jiang Y., Shuman M., Kanwisher N. A cortical area selective for visual processing of the human body // Science. 2001. Т. 293, № 5539. С. 2470–2473.

69. Doya K., Ishii S., Rao R. P. N., Pouget A. Bayesian Brain: Probabilistic Approaches to Neural Coding // MIT Press. 2011. 326 с.

70. Epstein R., Kanwisher N. A cortical representation of the local visual environment // Nature. 1998. Т. 392. № 6676. С. 598–601.

71. Erhan D., Bengio Y., Courville A., Vincent P. Visualizing higher-layer features of a deep network // Technical Report, Université de Montréal, 2009.

72. Everingham M., Van Gool L., Williams C. K. I., Winn J., Zisserman A. The Pascal Visual Object Classes (VOC) challenge // International Journal of Computer Vision. 2010. Т. 88, № 2. С. 303–338.

73. Fabre-Thorpe M., Richard G., Thorpe S. J. Rapid categorization of natural images by rhesus monkeys // Neuroreport. 1998. Т. 9. № 2. С. 303–308.

74. FaceGen // Википедия. 2023.

75. Faisal A. A., Selen L. P. J., Wolpert D. M. Noise in the nervous system // Nat Rev Neurosci. 2008. Т. 9. № 4. С. 292–303.

76. Felleman D. J., Van Essen D. C. Distributed hierarchical processing in the primate cerebral cortex // Cerebral Cortex. 1991. Т. 1, № 1. С. 1–47.

77. Field, D. J. Relations between the statistics of natural images and the response properties of cortical cells // J Opt Soc Am A. 1987. Т. 4. № 12 С. 2379–2394.

78. Field, D. J. What is the goal of sensory coding? // Neural Computation. 1994. Т. 6. № 4. С. 559–601.

79. Field D. J. Matched filters, wavelets and the statistics of natural scenes // Journal of Optical Technology - J Opt Technol TR. 1999. Т. 66. С. 788–796.

80. Field D. J. What The Statistics Of Natural Images Tell Us About Visual Coding // Human Vision, Visual Processing, and Digital Display. SPIE, 1989. С. 269–276.

81. Fix E., Hodges J. L. Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties // International Statistical Review / Revue Internationale de Statistique. 1989. Т. 57. № 3. С. 238–247.

82. Freiwald W. A., Tsao D. Y., Livingstone M. S. A face feature space in the macaque temporal lobe // Nat Neurosci. 2009. Т. 12. № 9. С. 1187–1196.

83. Fukushima K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position // Biol. Cybernetics. 1980. Т. 36. № 4. С. 193–202.

84. Gangopadhyay P., Das J. Do Primates and Deep Artificial Neural Networks Perform Object Categorization in a Similar Manner? // J. Neurosci. 2019. Т. 39. № 6. С. 946–948.

85. Gao F., Wang Y., Li P., Tan M., Yu J., Zhu Y. DeepSim: deep similarity for image quality assessment // Neurocomputing. 2017. Т. 257. С. 104–114.

86. Gardner J. R., Song X., Weinberger K. Q., Barbour D., Cunningham J. P. Psychophysical detection testing with Bayesian active learning // Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence UAI'15. Arlington, Virginia, USA: AUAI Press, 2015. С. 286–297.

87. Gatys L. A., Ecker A. S., Bethge M. Image Style Transfer Using Convolutional Neural Networks // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016. С. 2414–2423.

88. Geirhos R., Janssen D. H. J., Schütt H. H., Rauber J., Bethge M., Wichmann F. A. Comparing deep neural networks against humans: object recognition when the signal gets weaker // arXiv. 2018.

89. Geirhos R., Rubisch P., Michaelis C., Bethge M., Wichmann F. A., Brendel W. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness // arXiv. 2018.

90. Glezer V. D. Vision and Mind: Modeling Mental Functions // 1st ed. New York: Psychology Press, 1995. С. 290.

91. Gur M., Snodderly D. M. High Response Reliability of Neurons in Primary Visual Cortex (V1) of Alert, Trained Monkeys // Cerebral Cortex. 2006. Т. 16. № 6. С. 888–895.

92. Gross C. G., Rocha-Miranda C. E., Bender D. B. Visual properties of neurons in inferotemporal cortex of the macaque // Journal of Neurophysiology. 1972. Т. 35, № 1. С. 96–111.

93. He K., Zhang X., Ren S., Sun J. Deep residual learning for image recognition // 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016. С. 770–778.

94. Hebb D. O. The organization of behavior; a neuropsychological theory. Oxford, England: Wiley, 1949. xix, 335 c.

95. Hubel D. H., Wiesel T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex // The Journal of Physiology. 1962. Т. 160, № 1. С. 106–154.

96. Jin J., Dundar A., Culurciello E. Robust Convolutional Neural Networks under Adversarial Noise // 2016.

97. Karpathy A., Johnson J., Fei-Fei L. Visualizing and Understanding Recurrent Networks // 2015.

98. Khaligh-Razavi S.-M., Kriegeskorte N. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation // PLOS Computational Biology. 2014. Т. 10. № 11. С. e1003915.

99. Kheradpisheh S. R., Ghodrati M., Ganjtabesh M., Masquelier T. Deep networks can resemble human feed-forward vision in invariant object recognition // Scientific Reports. 2016. Т. 6, № 1. С. 32672.

100. Kiani R., Esteky H., Mirpour K., Tanaka K. Object category structure in response patterns of neuronal population in monkey inferior temporal cortex // Journal of Neurophysiology. 2007. Т. 97, № 6. С. 4296–4309.

101. Kietzmann T. C., McClure P., Kriegeskorte N. Deep Neural Networks in Computational Neuroscience // 2018. С. 133504.

102. Kingma D. P., Welling M. Auto-Encoding Variational Bayes // arXiv e-prints. International Conference on Learning Representations // 2014.

103. Kravitz D. J., Saleem K. S., Baker C. I., Ungerleider L. G., Mishkin M. The ventral visual pathway: an expanded neural framework for the processing of object quality // Trends in Cognitive Sciences. 2013. Т. 17, № 1. С. 26–49.

104. Kriegeskorte N. Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing // Annual Review of Vision Science. 2015. Т. 1. № 1. С. 417–446.

105. Kriegeskorte N., Douglas P. K. Cognitive computational neuroscience // Nat Neurosci. 2018. Т. 21. № 9. С. 1148–1160.

106. Kriegeskorte N., Mur M., Bandettini P. Representational similarity analysis - connecting the branches of systems neuroscience // Frontiers in Systems Neuroscience. 2008. Т. 2.

107. Lake B. M., Zaremba W., Fergus R., Gureckis T. M. Deep neural networks predict category typicality ratings for images // Proceedings of the 37th Annual Meeting of the Cognitive Science Society, CogSci 2015. 2015. С. 1243–1248.

108. Lamme V. A., Roelfsema P. R. The distinct modes of vision offered by feedforward and recurrent processing // Trends Neurosci. 2000. Т. 23. № 11. С. 571–579.

109. Lancaster J., Lorenz R., Leech R., Cole J. H. Bayesian optimization for neuroimaging pre-processing in brain age classification and prediction // Frontiers in Aging Neuroscience. 2018. Т. 10.

110. LeCun Y., Boser B., Denker J. S., Henderson D., Howard R. E., Hubbard W., Jackel L. D. Backpropagation applied to handwritten zip code recognition // Neural Computation. 1989. Т. 1, № 4. С. 541–551.

111. Lee B. B. Spectral Sensitivity in Primate Vision // Vision and Visual Dysfunction: Limits of Vision. CRC Press, 1991. С. 191–201.

112. Lettvin J. Y., Maturana H. R., McCulloch W. S., Pitts W. H. What the frog’s eye tells the frog’s brain // Proceedings of the IRE. 1959. Т. 47, № 11. С. 1940–1951.

113. Levick W. R., Thibos L. N. Analysis of orientation bias in cat retina // J Physiol. 1982. Т. 329. С. 243–261.

114. Lian Y., Almasi A., Grayden D. B., Kameneva T., Burkitt A. N., Meffin H. Learning receptive field properties of complex cells in V1 // PLoS Computational Biology. 2021. Т. 17, № 3. С. e1007957.

115. Liu J., Harris A., Kanwisher N. Stages of processing in face perception: an MEG study // Nat Neurosci. 2002. Т. 5. № 9. С. 910–916.

116. Liu A., Liu X., Yu H., Zhang C., Liu Q., Tao D. Training robust deep neural networks via adversarial noise propagation // Trans. Img. Proc. 2021. Т. 30. С. 5769–5781.

117. Logothetis N. K., Sheinberg D. L. Visual object recognition // Annu Rev Neurosci. 1996. Т. 19. С. 577–621.

118. Lorenz R., Hampshire A., Leech R. Neuroadaptive Bayesian Optimization and Hypothesis Testing // Trends in Cognitive Sciences. 2017. Т. 21. № 3. С. 155–167.

119. Mahendran A., Vedaldi A. Visualizing Deep Convolutional Neural Networks Using Natural Pre-images // Int J Comput Vis. 2016. Т. 120. № 3. С. 233–255.

120. Marblestone A. H., Wayne G., Kording K. P. Toward an Integration of Deep Learning and Neuroscience // Frontiers in Computational Neuroscience. 2016. Т. 10.

121. McCarthy G., Puce A., Gore J. C., Allison T. Face-specific processing in the human fusiform gyrus // Journal of Cognitive Neuroscience. 1997. Т. 9, № 5. С. 605–610.

122. McMahon D. B. T., Bondar I. V., Afuwape O. A. T., Ide D. C., Leopold D. A. One month in the life of a neuron: longitudinal single-unit electrophysiology in the monkey visual system // Journal of Neurophysiology. 2014. Т. 112, № 7. С. 1748–1762.

123. McMahon D. B. T., Jones A. P., Bondar I. V., Leopold D. A. Face-selective neurons maintain consistent visual responses across months // Proceedings of the National Academy of Sciences. 2014. Т. 111. С. 8251–8256.

124. Mishkin M., Ungerleider L. G. Contribution of striate inputs to the visuospatial functions of parieto-preoccipital cortex in monkeys // Behav Brain Res. 1982. Т. 6. № 1. С. 57–77.

125. Mordvintsev A., Olah C., Tyka M. Inceptionism: Going Deeper into Neural Networks // 2015.

126. Mountcastle V. B. Modality and Topographic Properties of Single Neurons of Cat's Somatic Sensory Cortex // Journal of Neurophysiology. 1957. Т. 20. № 4. С. 408–434.

127. Naito T., Okamoto M., Sadakane O., Shimegi S., Osaki H., Hara S.-I., Kimura A., Ishikawa A., Suematsu N., Sato H. Effects of stimulus spatial frequency, size, and luminance contrast on orientation tuning of neurons in the dorsal lateral geniculate nucleus of cat // Neuroscience Research. 2013. Т. 77, № 3. С. 143–154.

128. Nam Y., Sato T., Uchida G., Malakhova E., Ullman S., Tanifuji M. View-tuned and view-invariant face encoding in IT cortex is explained by selected natural image fragments // Scientific Reports. 2021. Т. 11, № 1. С. 7827.

129. Naselaris T., Kay K. N., Nishimoto S., Gallant J. L. Encoding and decoding in fMRI // NeuroImage. 2011. Т. 56, № 2. С. 400–410.

130. Nguyen A., Dosovitskiy A., Yosinski J., Brox T., Clune J. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks // Advances in Neural Information Processing Systems. 2016. Т. 29. С. 3387–3395.

131. Nguyen A., Yosinski J., Clune J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images // 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2015. С. 427–436.

132. Nguyen A., Yosinski J., Clune J. Multifaceted Feature Visualization: Uncovering the Different Types of Features Learned By Each Neuron in Deep Neural Networks // ArXiv, 2016.

133. Nguyen A. M. AI Neuroscience: Visualizing and Understanding Deep Neural Networks. University of Wyoming, 2017. 226 с.

134. Ohayon S., Freiwald W. A., Tsao D. Y. What makes a cell face selective? The importance of contrast // Neuron. 2012. Т. 74. № 3. С. 567–581.

135. Olah C., Mordvintsev A., Schubert L. Feature Visualization // Distill. 2017. Т. 2. № 11. С. e7.

136. Olshausen B. A., Field D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images // Nature. 1996. Т. 381. № 6583. С. 607–609.

137. Olshausen B. A., Field D. J. Sparse coding of sensory inputs // Curr Opin Neurobiol. 2004. № 14. С. 481–487.

138. Parkhi O. M., Vedaldi A., Zisserman A. Deep Face Recognition // Procedings of the British Machine Vision Conference 2015. Swansea: British Machine Vision Association, 2015. С. 41.1–41.12.

139. Pascual-Leone A., Hamilton R. The metamodal organization of the brain // Prog Brain Res. 2001. Т. 134. С. 427–445.

140. Pasupathy A., Connor C. E. Shape representation in area V4: position-specific tuning for boundary conformation // J Neurophysiol. 2001. Т. 86. № 5. С. 2505–2519.

141. Pelli D. G., Farell B. Why use noise? // J Opt Soc Am A Opt Image Sci Vis. 1999. Т. 16. № 3. С. 647–653.

142. Pitkow X., Meister M. Decorrelation and efficient coding by retinal ganglion cells // Nat. Neurosci. 2012. Т. 15. № 4. С. 628–635.

143. Pizzoli S. F. M., Mazzocco K., Triberti S., Monzani D., Alcañiz Raya M. L., Pravettoni G. User-centered virtual reality for promoting relaxation: an innovative approach // Frontiers in Psychology. 2019. Т. 10. С. 479.

144. Ponce C. R., Xiao W., Schade P. F., Hartmann T. S., Kreiman G., Livingstone M. S. Evolving images for visual neurons using a deep generative network reveals coding principles and neuronal preferences // Cell. 2019. Т. 177, № 4. С. 999–1009.e10.

145. Riesenhuber M., Poggio T. Models of object recognition // Nat Neurosci. 2000. Т. 3. № 11. С. 1199–1204.

146. Riva G., Serino S. Virtual Reality in the Assessment, Understanding and Treatment of Mental Health Disorders // J Clin Med. 2020. Т. 9. № 11. С. 3434.

147. Rombach R., Blattmann A., Lorenz D., Esser P., Ommer B. High-resolution image synthesis with latent diffusion models // 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022. С. 10674–10685.

148. Rousselet G. A., Fabre-Thorpe M., Thorpe S. J. Parallel processing in high-level categorization of natural images // Nat Neurosci. 2002. Т. 5. № 7. С. 629–630.

149. Rutishauser U., Mamelak A. N., Schuman E. M. Single-trial learning of novel stimuli by individual neurons of the human hippocampus-amygdala complex // Neuron. 2006. Т. 49. № 6. С. 805–813.

150. Sabour S., Frosst N., Hinton G. E. Dynamic Routing Between Capsules // 2017.

151. Sato T., Uchida G., Tanifuji M. Cortical Columnar Organization Is Reconsidered in Inferior Temporal Cortex // Cerebral Cortex. 2009. Т. 19. № 8. С. 1870–1888.

152. Sato T., Uchida G., Lescroart M. D., Kitazono J., Okada M., Tanifuji M. Object representation in inferior temporal cortex is organized hierarchically in a mosaic-like structure // Journal of Neuroscience. 2013. Т. 33, № 42. С. 16642–16656.

153. Schrimpf M., Kubilius J., Hong H., Majaj N. J., Rajalingham R., Issa E. B., Kar K., et al. Brain-Score: which artificial neural network for object recognition is most brain-like? // bioRxiv. 2018.

154. Simonyan K., Vedaldi A., Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps // CoRR. 2013.

155. Softky W. R., Koch C. The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs // J. Neurosci. 1993. Т. 13. № 1. С. 334–350.

156. Spoerer C. J., McClure P., Kriegeskorte N. Recurrent Convolutional Neural Networks: A Better Model of Biological Object Recognition // Frontiers in Psychology. 2017. Т. 8.

157. Suematsu N., Naito T., Sato H. Relationship between orientation sensitivity and spatiotemporal receptive field structures of neurons in the cat lateral geniculate nucleus // Neural Networks. 2012. Т. 35. С. 10–20.

158. Sun C., Chen X., Huang L., Shou T. Orientation bias of the extraclassical receptive field of the relay cells in the cat's dorsal lateral geniculate nucleus // Neuroscience. 2004. Т. 125, № 2. С. 495–505.

159. Szegedy C., Zaremba W., Sutskever I., Bruna J., Erhan D., Goodfellow I., Fergus R. Intriguing properties of neural networks // 2nd International Conference on Learning Representations, ICLR 2014. 2014.

160. Tanaka K. Inferotemporal cortex and object vision // Annu Rev Neurosci. 1996. Т. 19. С. 109–139.

161. Thorpe S., Fize D., Marlot C. Speed of processing in the human visual system // Nature. 1996. Т. 381. № 6582. С. 520–522.

162. Tolhurst D. J., Movshon J. A., Dean A. F. The statistical reliability of signals in single neurons in cat and monkey visual cortex // Vision Res. 1983. Т. 23. № 8. С. 775–785.

163. Tsao D. Y., Freiwald W. A., Tootell R. B. H., Livingstone M. S. A cortical region consisting entirely of face-selective cells // Science (New York, N.Y.). 2006. Т. 311, № 5761. С. 670–674.

164. Tsunoda K., Yamane Y., Nishizaki M., Tanifuji M. Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns // Nature Neuroscience. 2001. Т. 4, № 8. С. 832–838.

165. Vinje W. E., Gallant J. L. Sparse coding and decorrelation in primary visual cortex during natural vision // Science. 2000. №287. С. 1273-1276.

166. Watanabe E., Kitaoka A., Sakamoto K., Yasugi M., Tanaka K. Illusory motion reproduced by deep neural networks trained for prediction // Frontiers in Psychology. 2018. Т. 9.

167. Wei D., Zhou B., Torralba A., Freeman W. Understanding intra-class knowledge inside CNN // arXiv. 2015.

168. Williams R. M., Yampolskiy R. V. Optical Illusions Images Dataset // ArXiv, 2018.

169. Yamins D. L. K., DiCarlo J. J. Using goal-driven deep learning models to understand sensory cortex // Nat Neurosci. 2016. Т. 19. № 3. С. 356–365.

170. Yamins D. L. K., Hong H., Cadieu C. F., Solomon E. A., Seibert D., DiCarlo J. J. Performance-optimized hierarchical models predict neural responses in higher visual cortex // Proceedings of the National Academy of Sciences. 2014. Т. 111, № 23. С. 8619–8624.

171. Zhou Wang, Bovik A. C., Sheikh H. R., Simoncelli E. P. Image quality assessment: from error visibility to structural similarity // IEEE Transactions on Image Processing. 2004. Т. 13, № 4. С. 600–612.