Перспективы и ограничения нейросетевых моделей в нейронауках

А. А. Онучин[1]

Нейросетевые модели — стремительно набирающий популярность метод исследования и описания сложных мозговых процессов. Именно поэтому вопрос биологического правдоподобия и реалистичности подобных моделей является крайне актуальным. В данной работе мы обсудим существующие нейросетевые модели и примеры нейробиологического моделирования на их основе. **Ключевые слова:** искусственные нейросети, моделирование сложных систем, вычислительные нейронауки

1. Введение

В когнитивных теориях были популярны ‘блочные’ модели. Это идея, навеянная компьютерной архитектурой фон Неймана, укоренилась на десятилетия в репертуаре большинства исследователей. ‘Блоки’ соответствуют функциональным модулям в нервной системе (НС) (например, звуковому восприятию), а стрелки между ними указывают на направление и порядок информационных взаимодействий. Подобные модели получались на основе данных поведенческих корреляционных исследований со здоровыми испытуемыми - например, в задачах по исследованию различий восприятия речи и других звуков, это послужило основой для выделения различных модулей восприятия речи и восприятия нелингвистических акустических стимулов. Хотя некоторые из этих моделей были уточнены в исследованиях на людях с неврологическими нарушениями, большая их часть формулировалась без какой бы то ни было оглядки на нейробиологию мозга.

Значимым шагом к более строгой формулировке моделей такого типа было создание *локальных нейросетевых моделей* (ЛН), заполнивших ‘блоки’ одиночными искусственными нейронами, которые, как предполагалось, могут локально описывать содержание опыта. Взаимнооднозначное соответствие между искусственными нейронами и функциональными модулями делает крайне элементарным переход от блочных моделей к нейросетевым, однако, предположение о том, что отдельные нейроны могу быть ответственны за целостные функции не выдерживает

[1] *Онучин Арсений Андреевич* — студент 5-го курса ф-та Психологии МГУ, e-mail: arseniyonuchin04.09.97@gmail.com.

Onuchin Arsenii Andreevich — 5th year student, Lomonosov Moscow State University, Faculty of Psychology

никакой критики [1]. Помимо этого, подобные модели не позволяют моделировать процессы обучения и не дают никакой динамики поведения моделируемой системы.

1.1. Аттракторные нейросетевые модели

Благодаря нейроанатомическим исследованиям стало ясно, что связи в коре характеризуются высокой плотностью внутренних контактов и множественными повторами межнейрональных связей. Это позволило предположить, что такая архитектура связана с ассоциативной памятью, что послужило основой для создания семейства моделей называемых *аттракторными нейросетями* (АН). В АН нейроны контактируют с большинством нейронов сети, что разительно отличается от архитектуры, используемой в стандартных искусственных нейросетях. Реализация процессов обучения в АН основана на правиле Хебба. Согласно ему, связь усиливается там, где пресинаптический нейрон активируется в некотором временном окне до активации постсинаптического нейрона и наоборот. Данное правило реализуется механизмами долговременной потенциации и долговременной депривации в мозге. Предполагается, что нейронные цепи и циклы, сформированные в процессе обучения, функционируют как распределенные сетевые представления перцептивных, когнитивных или «смешанных» состояний. Следовательно, то, что корковые нейроны работают совместно в группах [1] и что функционально они распределены по тем же группам, может быть смоделировано АН. Это позволяет моделировать процессы памяти и обучения, что было невозможно в ЛН.

В АН возможна полная активация нейронального ансамбля, за счет лишь частичной стимуляции (аналогично гештальту): восприятие объекта возможно по его части. Подобная устойчивость достигается за счет сетевой архитектуры, которая стремится вернуться к заученному состоянию. Например, сеть Хопфилда позволяет моделировать механизмы ассоциативной памяти. Экспериментальные данные показывают, что даже при половине вышедших из строя нейронов в ней вероятность правильного ответа на стимул стремится к 1. Сеть Хопфилда полносвязна и состоит из N нейронов. В каждый момент времени t всякий нейрон может находиться в одном из двух состояний $S_i(t) \in \{-1, 1\}$, ответственных за модельное ‘возбужденине’ и ‘торможение’. Динамика i-го нейрона описывается дискретной динамической системой $S_i(t) := \mathrm{sign}[\sum_{j=1}^{N} J_{i,j} S_i(t-1)]$, где $J_{i,j}$ — матрица весовых коэффициентов, определяющих взаимодействие i, j нейронов. При этом в сети отсутствуют петлевые связи: $J_{i,i} = 0$. Обучение такой сети сводится к задаче минимизации некоторого функционала и подбору значений матрицы J.

АН могут быть использованы для моделирования широкого спектра нейробиологических феноменов и когнитивных процессов: от распознавания до поиска пути.

1.2. Многослойные персептроны

В отличие от АН, где сеть имела вид полного графа, в случае с *многослойными персептронами* (МП) архитектура сводится к наличию контактов только между нейронами из соседних слоёв. Такие сети состоят из нейронов соединенных последовательно, слой за слоем. Всего слоёв выделяют три: входной, скрытый и выходной, что навеянно нейроанатомией сетчатки.

Такие модели дают представление объектов 'плотно' упакованным и не разреженным: они распределены по всем нейронам скрытого слоя и *вектор активации* по всему слою является нейросетевым аналогом объекта. Плотное распределение сигнала по сети разительно отличается от того, что мы видим в АН. Обучение реализуется методом градиентного спуска, когда после каждого прохождения сигнала через сеть результат сверяется с эталоном, считается мера ошибки и относительно нее корректируются все веса в сети.

По мере обнаружения ограничений данной модели для реализации механизмов памяти, было предложено добавлять в архитектуру дополнительные слои, что расширяет потенциальную применимость в моделировании когнитивных процессов.

1.3. Глубокие нейросетевые модели

Дальнейшее развитие пошло по пути увеличения числа слоев в МП, что получило развитие в теории глубоких нейросетей (ГН). Нейробиологическая мотивация, стоящая за идеей роста числа слоёв состоит в схожести с нейроанатомической архитектурой зрительной коры.

ГН претерпели череду эволюций, приведших к появлению несовпадающих архитектур. Например, появились сверточные нейросети, которые включают топографические проекции между (не обязательно всеми) слоями сети, чтобы упрощать процесс обработки смежных входных данных: для различных нейронов выходного слоя используются одна и та же матрица весов, которую также называют *ядром свёртки*, что отличается от наборов индивидуальных весов у каждого нейрона в стандартной модели глубокой нейросети. Подобные улучшения и модификации структуры многослойного персептрона позволили достичь человеческого уровня (качественного) решения таких задач, как классификация объектов или распознование речи.

О проблемах. Узконаправленность ГН и невозможность сочетать несколько разномодальных задач в одной сети. ГН склонны к неуместным обобщениям, например, в задачах классификации, выдавая ответ на сильно зашумленных и неидентифицируемых данных [3]. Существует целая серия работ о неустойчивости подобных моделей к некоторым незначительным пертурбациям и шумам во входных данных, которые человек с легкостью бы отфильтровал [2]. Такие сети не имеют динамики и активируются также плотно, как и обычные МП, чем отличаются от разреженной работы мозга.

Список литературы

[1] Abeles M., "Corticonics: Neural circuits of the cerebral cortex.", *Cambridge University Press*, 1991.

[2] Carlini N., Wagner D., *Towards evaluating the robustness of neural networks*, 2017 ieee symposium on security and privacy (sp), 2017, 39-57 c.

[3] Nguyen A., Yosinski J., Clune J., "Understanding neural networks via feature visualization: A survey", *Springer*, 2019, 55-76

Perspectives and constraints on neural network models of neurobiological processes

Onuchin Arsenii Andreevich

Artificial and natural neural network models are a new toolkit which could be potentially have been used for clarifying of complex brain functions. To attend this goal, such models need to be neurobiologically realistic. In this work we discuss different types of neural models and also identify aspects under which their biological credibility can be improved.

Keywords: neural networks, complex systems modeling, computational neuroscience.

References

[1] Abeles M., "Corticonics: Neural circuits of the cerebral cortex.", *Cambridge University Press*, 1991

[2] Carlini N., Wagner D., *Towards evaluating the robustness of neural networks*, 2017 ieee symposium on security and privacy (sp), 2017, 39-57 c.

[3] Nguyen A., Yosinski J., Clune J., "Understanding neural networks via feature visualization: A survey", *Springer*, 2019, 55-76