Не умирай, нейрон: российские математики доработали выводы Нобелевского лауреата

Ученые из Института AIRI и Сколтеха изучили, как выглядит ландшафт функции энергии, и выяснили, что она не подходит для многих современных нейронных сетей. Исследователи предложили способ исправить ситуацию в исходной системе и новый способ построения функции энергии, благодаря которому можно избежать образования плоских областей в ландшафте памяти. Результаты работы помогут другим учёным получать гарантии при теоретическом анализе моделей ассоциативной памяти.

Когда мы сохраняем фотографии, заметки, музыку в своем телефоне или компьютере, как правило мы пользуемся памятью с произвольным доступом (RAM). В такой памяти данные хранятся в отдельных ячейках, каждая из которых обладает уникальным адресом. Получить доступ к ячейке можно только зная адрес. В качестве простого примера можно рассмотреть список (list), реализованный в языке Python, с несколькими строками, содержащими информацию о жизни Ивана Петровича: [‘16 февраля 2010 года Иван Петрович отправился на рыбалку и поймал здоровенного карася.’, ‘Однажды, вкручивая лампочку, Иван Петрович подумал, что Ленин все-таки не зря был вождем мировой революции.’, ‘В 2023 поздней весной у Ивана Петровича родился внук Сергей.’]. В этом примере первое ‘воспоминание’ о карасе имеет адрес 0, а оставшиеся — 1 и 2. Хоть адрес никак не связан с содержанием, именно его нужно вспомнить, чтобы извлечь информацию об одном из событий.

Человеческая память устроена иначе. Иван Петрович скорее вспомнит о своем легендарном карасе, подумав о феврале, или о 2010 году, или о зимней рыбалке, но уж точно не о ячейке под номером 0. Память такого типа называется ассоциативной или контекстно-адресуемой памятью. Главное отличие от памяти с произвольным доступом в том, что извлечение информации происходит не по абстрактному адресу, а по приближенному содержанию воспоминания (в случае с Иваном Петровичем ‘весна’ -> ‘внук’, ‘зима’ -> ‘карась’, ‘Ленин’ -> ‘лампочка’). Многочисленные опыты физиологов и психологов (например, Скиннер, Павлов) показали, что живые существа обучаются, формируя ассоциации именно с содержанием импульса, поэтому контекстно-адресуемая память выглядит более естественной с биологической точки зрения (A Brief History of Intelligence). Желание узнать больше о принципах ассоциативной памяти привело к появлению большого

пласта эмпирических исследований и упрощенных математических моделей. Одну из первых и наиболее удачных моделей такого типа предложил Джон Хопфилд.

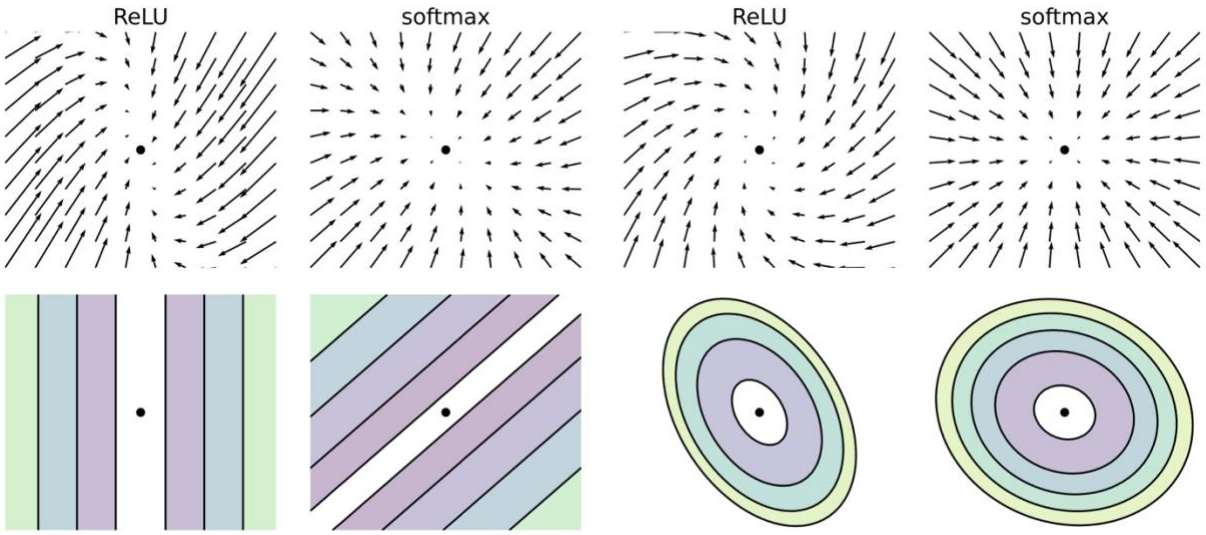
В работе 1982 года Джон Хопфилд предложил модель памяти на основе спинового стекла. Воспоминания описываются векторами с дискретными компонентами, принимающими значения +1 и -1 – спин вверх и спин вниз. Процесс извлечения воспоминаний соответствует движению к ближайшему локальному минимуму функции энергии системы, которая описывает взаимодействия между соседними спинами. Двумя годами позже Хопфилд представил вторую модель ассоциативной памяти на основе системы обыкновенных дифференциальных уравнений. В этой модели воспоминания уже описывались векторами с вещественными числами.

В последнее время исследователи в области искусственного интеллекта стали активно изучать память систем, и непрерывность превратилась в одну из наиболее важных характеристик, необходимых для обучения эффективных ИИ-моделей. Например, в одной из недавних популярных статей авторы обнаружили связь между механизмом внимания и ассоциативной памятью Хопфилда. В этом же году Джон Хопфилд и его соавтор Дмитрий Кротов сформулировали новую версию модели ассоциативной памяти, которую сейчас называют “обобщенной моделью Хопфилда”. Эта модель не только объединила все прошлые, включая дискретную (1982) и непрерывную (1984) память Хопфилда, но и позволила естественным образом связать ассоциативную память с другими нейросетевыми моделями. Например, с механизмом внимания, сверточными сетями, сетями с вентелем (gated mechanism, GRU) и множеством других современных архитектур.

С технической точки зрения обобщенная модель Хопфилда описывается особым классом обыкновенных дифференциальных уравнений, обладающих функцией Ляпунова, также называемой функцией энергии. Функция энергии призвана гарантировать существование стационарных состояний, которые моделируют воспоминания. Более точно, предполагается, что начальные условия соответствуют неполному или искаженному содержанию воспоминания. В процессе эволюции согласно обыкновенному дифференциальному уравнению, система приходит к состоянию, которое не меняется с течением времени. Это состояние и описывает воспоминание наиболее близкое к начальным условиям.

Ученые из Института AIRI и Сколтеха изучили, как выглядит ландшафт функции энергии, и выяснили, что она не подходит для многих современных нейронных сетей. Оказывается, что в случае, если какие-то активационные функции нейронов насыщаются, энергия становится плоской. Представьте себе оживленный город, в котором много фонарей и светофоров. Когда уличное освещение гаснет,

определенные участки городского пространства темнеют, замедляется движение транспорта. Нейронные сети состоят из цепочек искусственных нейронов, которые имитируют нейроны головного мозга человека. Так же, как и фонари, нейроны могут гаснуть или, как говорят ученые, «умирать». Мертвые нейроны могут привести к образованию плоских областей в энергетическом ландшафте нейросетей. После этого узнать, что происходит в этих областях, становится невозможно. Из-за этого невозможными становятся изучение устойчивости системы и восстановление динамических переменных из уравнения.



Функция энергии в работе Хопфилда и Кротова

Функция энергии в работе группы «Вычислительный интеллект»

Математики из научной группы «Вычислительный интеллект» AIRI обнаружили, что даже в таких сложных ситуациях можно собрать полезную информацию о состоянии сети, посмотрев на матрицу вторых производных (гессиан) функции Лагранжа, связанной с активационными функциями нейронов. Кроме того, ученые доказали, что если устойчивое состояние стабильно, то и вся плоская область вокруг него может считаться стабильной. Это означает, что даже если некоторые нейроны мертвы, сеть может функционировать эффективно.

Исследователи предложили способ исправить ситуацию в исходной системе и новый способ построения функции энергии, благодаря которому можно избежать образования плоских областей в ландшафте памяти. Результаты работы помогут другим ученых получать гарантии при теоретическом анализе моделей ассоциативной памяти.

«В 1984 году Джон Хопфилд сформулировал модель ассоциативной памяти для узкого класса нейронных сетей с обратимыми активационными функциями. В 2020 году Джон Хопфилд и Дмитрий Кротов переформулировали модель 1984 года так, чтобы формально исключить обратные активационные функции. Ученые из AIRI обнаружили, что это не решило проблему и функция энергии, предложенная в 2020 году, также не может быть использована для описания нейронных сетей с современными активационными функциями (например, ReLU, GELU, softmax). Ассоциативная память требует анализа устойчивости. А еще для дальнейшего развития ИИ необходимо сформулировать математическое обоснование принципов его работы. Мало получить рабочее решение. Ученый должен знать, как и почему оно функционирует определенным образом»

Иван Оселедец — д.ф.-м.н., профессор РАН, генеральный директор Института AIRI, профессор Сколтеха

Вопросы: pr@airi.net

Институт AIRI *— автономная некоммерческая организация, занимающаяся фундаментальными и прикладными исследованиями в области искусственного интеллекта. На сегодняшний день более 180 научных сотрудников AIRI задействовано в исследовательских проектах Института для работы совместно с глобальным сообществом разработчиков, академическими и индустриальными партнерами.*