УДК 004.891.2

ОГРАНИЧЕНИЯ РОСТА МОЩНОСТЕЙ НЕЙРОННЫХ СЕТЕЙ: ФИЗИЧЕСКИЕ И ЭНЕРГЕТИЧЕСКИЕ АСПЕКТЫ

Касьян В. А., студент гр.253501, Ахметов Р. Я., студент гр.253502, Сенько Н. С., студент гр.253502, Внук О.М., магистрант гр. 225941, Владымцев В. Д., ассистент каф. Информатики

Белорусский государственный университет информатики и радиоэлектроники г. Минск, Республика Беларусь

Владымцев В. Д. – ассистент каф. Информатики

Аннотация. В данной научной работе рассматриваются ограничения роста мощностей нейронных сетей, вызванные физическими и энергетическими ограничениями современных компьютеров. Проводится анализ влияния энергопотребления, роста стоимости и роста производительности компьютеров на развитие нейронных сетей, а также предлагаются пути оптимизации и возможные решения для снижения энергоемкости и стоимости обучения нейронных сетей.

Ключевые слова: Нейронные сети, ограничения роста, мощность, обучение нейронных сетей, энергоэффективность, архитектура нейронных сетей, вычислительные ресурсы, технологические ограничения, энергопотребление оптимизация нейронных сетей, производительность.

Появление нейронных сетей стало одним из наиболее важных достижений в области искусственного интеллекта за последнее время. Нейронные сети - это алгоритм машинного обучения, который имитирует структуру человеческого мозга при выявлении закономерностей в данных. Они продемонстрировали замечательную эффективность в самых разных приложениях, таких как распознавание изображений и речи, обработка естественного языка и автономное вождение.

Со временем, развитие микроэлектронной промышленности привело к значительному увеличению количества транзисторов на интегральных схемах. Этот рост количества транзисторов был сопровожден их уменьшением и повышением тактовой частоты. Закон Мура, сформулированный в 1965 году Гордоном Муром, предсказывал, что количество транзисторов на интегральной схеме будет удваиваться примерно каждые два года. В течение нескольких десятилетий этот прогноз оставался достаточно точным.

Большие модели — это нейронные сети с большим количеством параметров. Из-за своих размеров эти модели требуют значительного объема вычислительных ресурсов для обучения и оптимизации, что делает вычислительную мощность решающим фактором при работе с ними. На заре разработки нейронных сетей обучение модели было трудоемким и ресурсоемким процессом, что делало его непрактичным для большинства исследователей. Однако с появлением параллельных вычислений и использованием графических процессоров (GPU) обучение моделей стало значительно быстрее. Это привело к тому, что все больше исследователей смогли присоединиться к этой области и расширили сферу своих применений. Возросшая доступность больших моделей также привела к более сложным и детальным исследованиям, при этом такие модели тестируются и дорабатываются. В конечном счете это привело к повышению производительности и более точным прогнозам в различных областях, включая обработку естественного языка, распознавание изображений и речи.

Тем не менее, существуют физические и энергетические ограничения, связанные с дальнейшим ростом мощности нейронных сетей. Как уже упоминалось ранее, закон Мура сталкивается с проблемами, вызванными уменьшением размеров транзисторов и увеличением тактовой частоты. Эти ограничения влияют на производительность и энергопотребление вычислительных систем, используемых для обучения и работы с нейронными сетями.

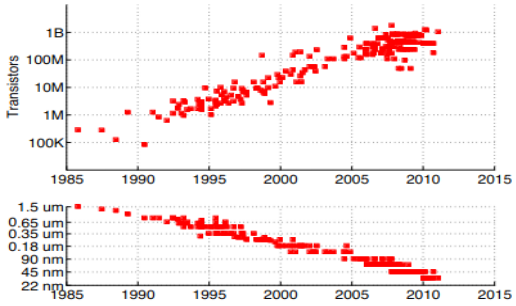

Рисунок 1 – Рост количества транзисторов и их уменьшение с течением времени

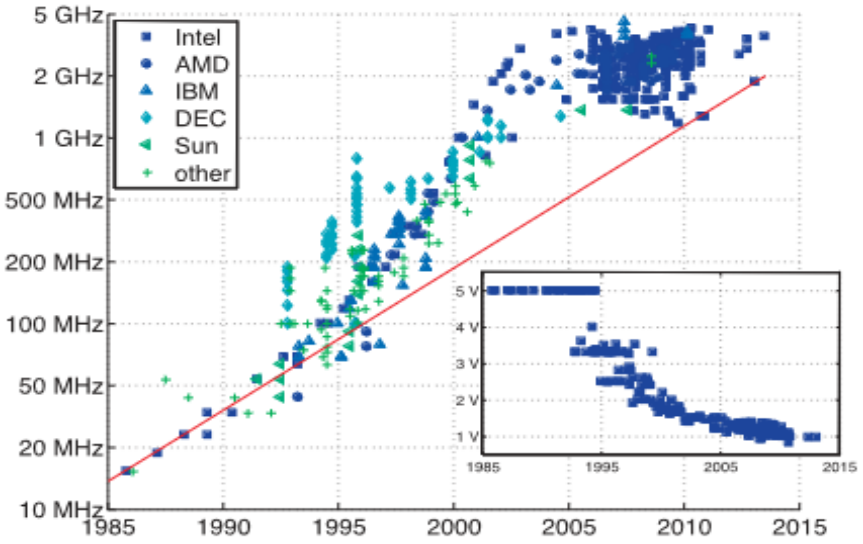

Рисунок 2 – Увеличение тактовой частоты с течением времени

Для преодоления этих проблем, исследователи и разработчики активно изучают новые подходы и технологии, такие как квантовые вычисления, нейроморфные компьютеры и специализированные архитектуры для обработки нейронных сетей, например, Tensor Processing Units (TPU) от Google. Кроме того, исследуются методы оптимизации архитектуры нейронных сетей, снижения их энергопотребления и увеличения эффективности обучения.

Вместе с тем, разработка алгоритмов машинного обучения, требующих меньше вычислительных ресурсов и энергии, также является важным направлением исследований в области нейронных сетей. Прогресс в этой области позволит сократить затраты на обучение моделей и сделать их доступными для широкого круга пользователей, что в свою очередь может способствовать дальнейшему расширению применения нейронных сетей в различных отраслях.

Использование графических процессоров (GPU) в машинном обучении стало крупным достижением за последние годы. Обеспечивая высокую степень распараллеливания вычислений, графические процессоры могут обрабатывать большие объемы данных и выполнять множество вычислений одновременно, что делает их хорошо подходящими для задач с высокой вычислительной нагрузкой, таких как обучение нейронных сетей.

На заре вычислений на графических процессорах исследователи часто использовали игровые графические процессоры, предназначенные для потребителей, для обучения своих моделей. Однако по мере того, как нейронные сети становились все крупнее и сложнее, для удовлетворения спроса на более быстрые и мощные вычисления стало необходимо специализированное оборудование, такое как тензорные процессоры Google (TPU). Это открыло новую эру в машинном обучении, характеризующуюся разработкой специализированных аппаратных и программных средств, предназначенных для полного использования мощности графических процессоров и других высокопроизводительных вычислительных ресурсов.

Со временем вычислительная мощность, необходимая для обучения сложных нейронных сетей, возросла. Это обусловлено необходимостью достижения самых современных результатов в области обработки естественного языка, компьютерного зрения и распознавания речи. Более

крупные модели требуют больше энергии для обучения, но прогресс в алгоритмах и растущая доступность данных позволили исследователям продолжать раздвигать границы возможного. Однако сам объем требуемой вычислительной мощности может стать проблемой для многих организаций. Чтобы решить эту проблему, необходимы дальнейшие исследования для разработки более эффективных алгоритмов, которые позволяют получать самые современные результаты с меньшими затратами энергии. Кроме того, были бы полезны более доступные вычислительные ресурсы, такие как облачные сервисы или специализированное оборудование.

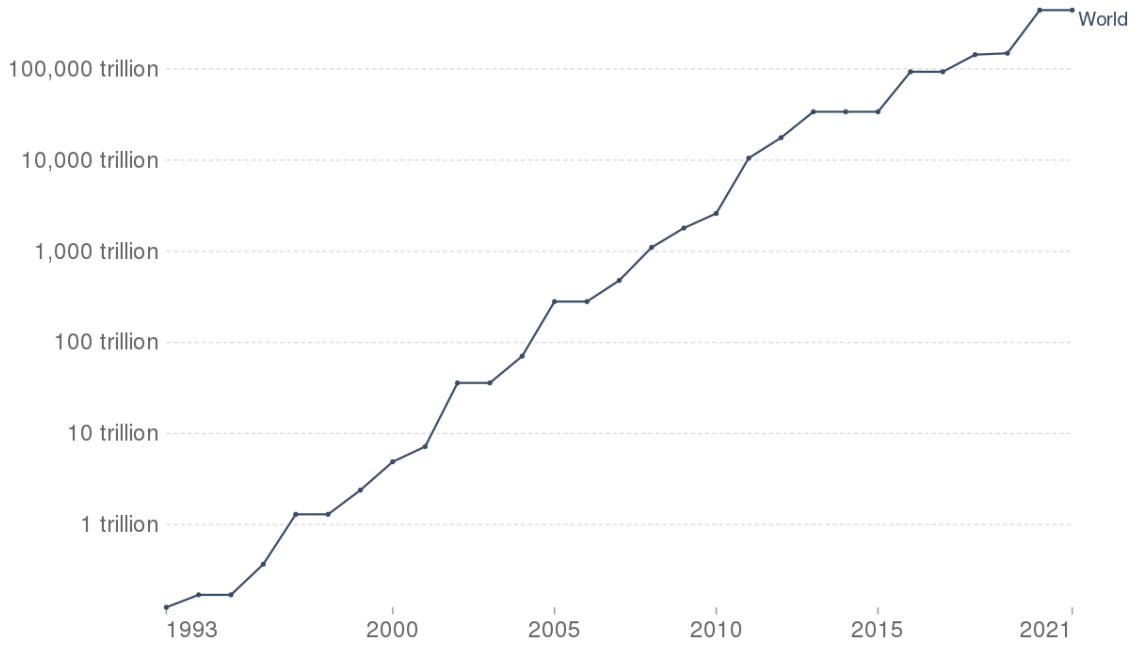

Рисунок 3 – Количество операций с плавающей точкой в секунду, которые могут выполнить суперкомпьютеры

Чтобы дать некоторое представление о количестве вычислительной мощности, необходимой для обучения нейронных сетей, давайте рассмотрим модель BERT в качестве примера. BERT (Bidirectional Encoder Representations от Transformers) - это модель обработки естественного языка с 340 миллионами параметров. Google обучал БЕРТА, используя кластер из 4096 процессоров в течение четырех дней. Это служит ярким примером того, как вычислительная мощность, необходимая для обучения моделей, значительно увеличилась в масштабе.

Чтобы привести более свежий пример, GPT-3 (Generative Pre-trained Transformer 3), модель обработки естественного языка, представленная OpenAI в 2020 году, имеет ошеломляющие 175 миллиардов параметров, что делает ее одной из крупнейших моделей на сегодняшний день. Обучение такой масштабной модели требует значительного объема вычислительной мощности и ресурсов, что делает ее доступной только для организаций, обладающих необходимой инфраструктурой. Поскольку нейронные сети продолжают увеличиваться в размерах и усложняться, важно учитывать последствия их энергопотребления и изучать способы сделать их более устойчивыми и эффективными.

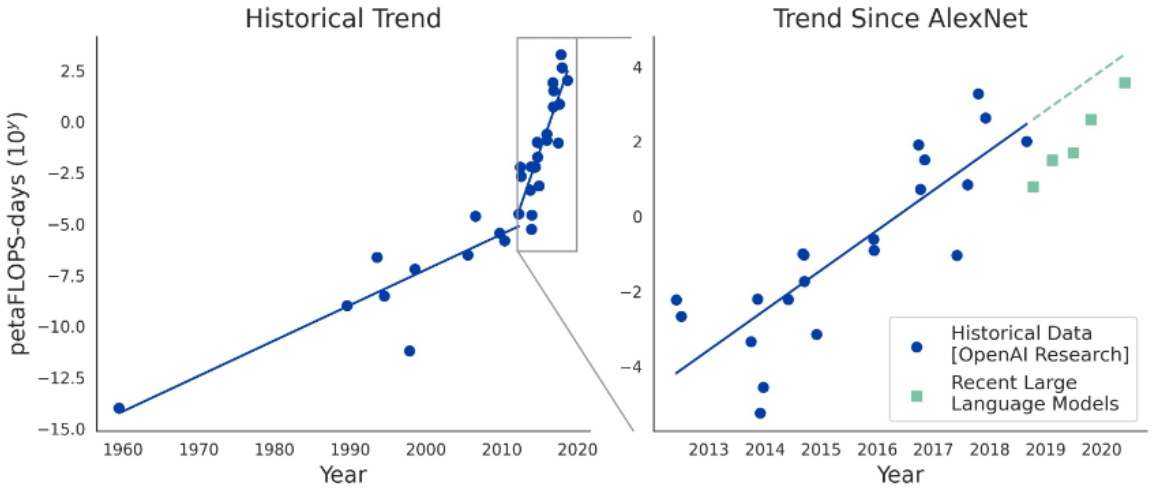

Рисунок 4 – Увеличение количества вычислений для обучения нейронных сетей с течением времени

Тенденция к созданию более крупных моделей в машинном обучении привела к значительному увеличению объема вычислительной мощности, необходимой для обучения этих моделей. Развитие параллельных вычислений и использование графических процессоров сделали возможным обучение моделей, которые когда-то считались невозможными. Поскольку нейронные сети продолжают развиваться, мы можем ожидать, что спрос на вычислительную мощность будет продолжать расти. Это поднимает важные вопросы о воздействии машинного обучения на окружающую среду и необходимости разработки более энергоэффективного оборудования.

Также стоит упомянуть проблему чрезмерного энергопотребления нейронных сетей. Нейронные сети, особенно крупные модели, требуют значительного количества энергии для обучения и функционирования, что может негативно сказаться на окружающей среде и экономической составляющей их использования. С 2000 года стоимость обучения нейронных сетей резко выросла, и этот рост продолжается до сих пор. Несмотря на то, что нейронные сети были изобретены еще в 1950-х годах, они долгое время не были особенно популярны из-за ограничений вычислительной техники и недостаточно большого количества данных для обучения. Однако с развитием технологий и появлением больших объемов данных, нейронные сети стали все более популярными, и сегодня они используются в самых разных областях: от распознавания речи до анализа изображений и машинного перевода.

Одним из главных факторов, определяющих стоимость обучения нейронных сетей, является объем данных, необходимых для обучения. С увеличением объема данных растет и стоимость их сбора, обработки и хранения. Кроме того, обучение нейронной сети требует высокопроизводительных компьютеров с большим количеством вычислительных ресурсов, которые также являются дорогостоящими.

Еще одним фактором, влияющим на стоимость обучения нейронных сетей, является квалификация специалистов, занимающихся этой работой. Обучение нейронных сетей требует высокой квалификации и опыта в области алгоритмов машинного обучения и вычислительной техники, что в свою очередь ведет к высоким затратам на оплату труда таких специалистов.

Стоимость обучения нейронных сетей также зависит от сложности задачи, которую необходимо решить. Некоторые задачи, такие как распознавание речи или анализ изображений, требуют большого количества данных и сложных алгоритмов обработки, что ведет к более высокой стоимости обучения.

Несмотря на рост стоимости обучения нейронных сетей, их популярность продолжает расти, поскольку они являются мощным инструментом для решения самых разных задач. Более того, с развитием технологий и увеличением количества данных, точность и эффективность нейронных сетей продолжает увеличиваться, что делает их еще более привлекательными для использования в различных областях.

Вместе с тем, рост стоимости обучения нейронных сетей также стимулирует развитие новых методов и технологий, направленных на уменьшение затрат на обучение. Например, в последние годы активно развивается область автоматизированного машинного обучения (AutoML), которая позволяет автоматически настраивать и оптимизировать нейронные сети без необходимости вручную настраивать их параметры. Это позволяет сократить время и затраты на обучение и делает использование нейронных сетей более доступным для широкого круга пользователей.

Таким образом, хотя стоимость обучения нейронных сетей продолжает расти с 2000 года, их популярность и значимость не уменьшаются. С появлением новых технологий и методов,

направленных на сокращение затрат на обучение, они становятся все более доступными для использования в различных областях и могут стать ключевым инструментом в решении сложных задач.

Потребляемая мощность стала тем фактором, который заставляет разработчиков суперкомпьютеров переосмысливать их архитектуру. Поскольку отдельные узлы суперкомпьютера потребляют все больше электроэнергии и выделяют все больше тепла, их необходимо разносить в пространстве и интенсивно охлаждать. Без использования экзотических систем охлаждения перегрев приведет к тому, что суперкомпьютеры просто не смогут выполнять нужные исследователям приложения. К сожалению, стоимость построения нестандартных охлаждающих систем может быть вполне сопоставима со стоимостью самого суперкомпьютера, а их обслуживание обойдется еще дороже.

Система	Число процессоров	Показатели устойчивости и готовности
ASC Q	8192	Среднее время между перерывами в работе: 6,5 часа, 114 незапланированных случаев выхода из строя в течение месяца. Источники сбоев: система хранения, процессоры, оперативная память
ASC White	8192	Среднее время между отказами: 5 часов (2001) и 40 часов (2003). Источники сбоев: система хранения, процессоры, оборудование третьих фирм.
PSC Lemieux	3016	Среднее время между перерывами в работе: 9,7 часа. Уровень готовности: 98,33%
Google (предположительно)	450000	600 перезагрузок в день; 2-3% оборудования, подлежащего замене в течение года. Источники сбоев: система хранения и память. Уровень готовности: ~100%

Источник информации: D.A. Reed

Рисунок 5 – таблица зависимости количества сбоев в работе суперкомпьютера от числа процессоров

Как показано во врезке «Энергетическая эффективность Green Destiny», суперкомпьютер с низким энергопотреблением оказался удивительно устойчивым в работе. За два года не было зарегистрировано ни одного отказа. Согласно результатам опроса, опубликованным в 2001 году компанией Contingency Planning Research, стоимость часа простоя такого компьютера варьируется от 90 тыс. долл. при выполнении им операций, связанных с организацией продаж по каталогу, до 6,5 млн долл. при выполнении брокерских операций. Нет никаких гарантий того, что суперкомпьютер никогда не сломается, и это наглядно проиллюстрировано в таблице. А также общая стоимость владения такой техникой значительно превышает первоначальную стоимость ее приобретения.

Машинное обучение находится на пути к тому, чтобы потреблять всю поставляемую энергию, и такая модель является дорогостоящей, неэффективной и неустойчивой.
В значительной степени это объясняется тем, что данная область является новой, чрезвычайно интересной и быстро развивающейся. Она разрабатывается для того, чтобы открыть новые горизонты в плане точности или возможностей. Сегодня это означает большие модели и большие обучающие наборы, что требует экспоненциального роста вычислительных возможностей и потребления огромного количества энергии в центрах обработки данных как для обучения, так и для выводов. Кроме того, умные устройства начинают появляться повсюду.

Но цифры потребляемой мощности начинают пугать людей. На недавней конференции по автоматизации проектирования технический директор AMD Марк Пейпермастер представил слайд, показывающий энергопотребление систем ML в сравнении с мировым производством энергии:

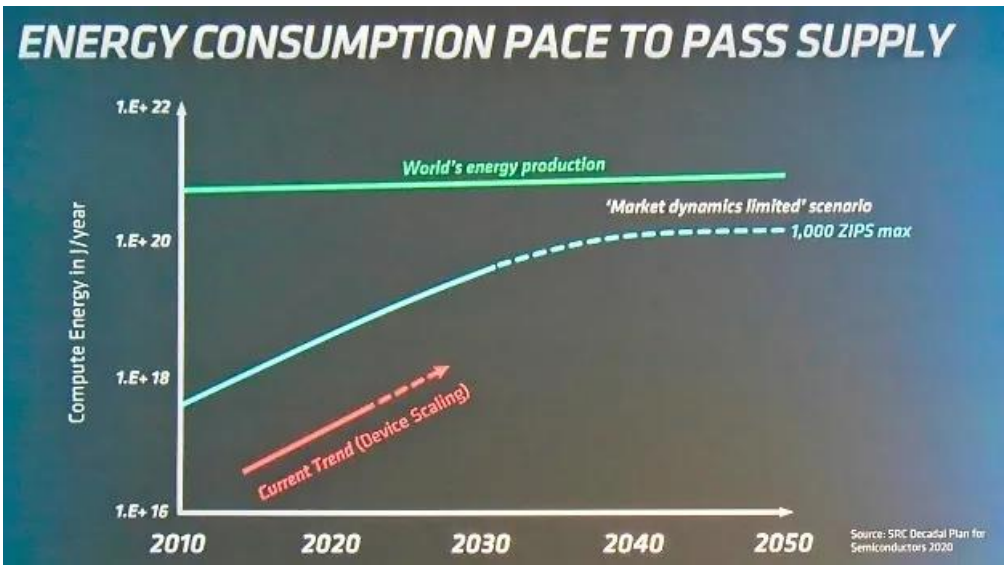

Рисунок 6 – слайд с презентации Марка Пейпермастера

Компания Papermaster не единственная, кто бьет тревогу. "Мы забыли, что движущей силой инноваций в течение последних 100 лет была эффективность", - говорит Стив Тейг, генеральный директор Perceive. "Именно это привело к появлению закона Мура. Сейчас мы живем в эпоху анти эффективности".

А Аарт де Геус, председатель совета директоров и генеральный директор компании Synopsys, от имени планеты Земля призвал сделать что-то с этим. "Тот, у кого есть мозги, чтобы понять, должен иметь сердце, чтобы помочь".

Почему потребление энергии растет так быстро? "Вычислительные потребности нейронных сетей ненасытны", - говорит Ян Братт, научный сотрудник и старший директор по технологиям компании Arm. Чем больше сеть, тем лучше результаты и тем больше проблем вы можете решить". Энергопотребление пропорционально размеру сети. Поэтому энергоэффективный вывод абсолютно необходим для внедрения все более сложных нейронных сетей и расширенных сценариев использования, таких как приложения для работы с голосом и зрением в реальном времени".

К сожалению, не все заботятся об эффективности. "Если посмотреть на то, что пытаются сделать компании-гиперскейлеры, то они пытаются получить более качественное и точное распознавание голоса, речи, рекомендательные движки", - говорит Тим Векслинг, старший вице-президент по продуктам и развитию бизнеса компании Mythic. "Это денежный вопрос. Чем выше точность, тем больше клиентов они могут обслужить и получить большую прибыль". Если посмотреть на обучение в центре обработки данных и вывод этих очень больших моделей NLP, то именно там потребляется много энергии. И я не знаю, есть ли реальная мотивация для оптимизации энергопотребления в этих приложениях".

Но некоторым людям не все равно. "Существует некоторое коммерческое давление, направленное на снижение углеродного воздействия этих компаний, не прямое денежное, а скорее на то, что потребитель примет только углеродно-нейтральное решение", - говорит Александр Уэйкфилд, ученый из Synopsys. "Это давление со стороны "зеленой" энергетики, и если один из этих поставщиков заявит, что он нейтрален к выбросам углекислого газа, то больше людей, скорее всего, будут использовать его".

Но не вся энергия потребляется в области облачных вычислений. Растет число умных периферийных устройств, которые также вносят свой вклад в эту проблему. "Существуют миллиарды устройств, составляющих IoT, и в какой-то момент в недалеком будущем они будут потреблять больше энергии, чем мы производим в мире", - говорит Марси Вайнштейн, директор по стратегическому и техническому маркетингу компании Aspinity. "Они потребляют энергию для сбора, передачи и выполнения любых действий с данными, которые они собирают".

С увеличением мощности и сложности современных нейронных сетей, важность энергоэффективности и оптимизации становится все более актуальной. Решение экологических и экономических проблем, связанных с ростом энергопотребления нейронных сетей, требует совместных усилий со стороны исследователей, инженеров и промышленности. Одним из направлений оптимизации является проработка новых архитектур нейронных сетей, которые могут обеспечить более эффективное использование вычислительных ресурсов и снижение энергопотребления. Например, применение разреженных нейронных сетей или использование квантовых компьютеров для обработки нейросетевых моделей может значительно уменьшить

затраты на энергию. Кроме того, исследователи активно разрабатывают методы сжатия и ускорения нейронных сетей, которые позволяют уменьшить размер моделей и, таким образом, снизить потребление ресурсов. Это включает техники, такие как кластеризация весов, прореживание и квантизация. Важным аспектом является также разработка и применение более энергоэффективного оборудования, такого как специализированные архитектуры для обработки нейронных сетей (ASIC), которые могут существенно снизить энергопотребление по сравнению с традиционными графическими процессорами (GPU). С другой стороны, увеличение стоимости нейронных сетей может стать препятствием для их широкого внедрения, особенно для малых и средних предприятий. В этом контексте важно разработать стратегии и программы, направленные на увеличение доступности нейросетевых технологий для всех участников рынка, включая предоставление обучающих материалов, программ финансирования и поддержки разработчиков. Исследования в области энергоэффективных нейронных сетей и оптимизации их стоимости являются важными шагами для обеспечения долгосрочной устойчивости и широкого распространения нейронных сетей в различных отраслях и приложениях. При этом необходимо продолжать разрабатывать и внедрять новые методы обучения, которые будут использовать меньше вычислительных ресурсов и энергии. Это может включать использование переноса обучения, обучения с подкреплением, мета обучения и других подходов, которые могут сократить время и стоимость обучения нейросетевых моделей. Создание открытых исходных кодов и общедоступных ресурсов, таких как предварительно обученные модели, базы данных и инструменты для работы с нейронными сетями, также будет способствовать доступности нейросетевых технологий для широкого круга пользователей. Это может стимулировать инновации и разработку новых решений на основе нейронных сетей, которые будут учитывать экологические и экономические ограничения.

В заключение, преодоление физических и энергетических ограничений роста мощности нейронных сетей потребует совместных усилий со стороны исследователей, инженеров, промышленности и законодателей. Разработка новых алгоритмов, технологий и аппаратных решений, а также учет экологических и экономических факторов, будет ключом к долгосрочной устойчивости и успешному развитию этой перспективной области искусственного интеллекта.

Список использованных источников:

1. Medium[Электронный ресурс]. – Режим доступа: URL:https://medium.com/illumination/gpt-3-vs-gpt-4-987872f48ecf (дата обращения: 11.04.2023).

2. Medium[Электронный ресурс]. – Режим доступа: URL:https://medium.com/illumination/meta-llama-vs-chatgpt-a-detailed-comparison-9794ccedd41c (дата обращения: 11.04.2023).

3. Stadtherr M. A. High performance computing: Are we just getting wrong answer faster? //CAST division awards banquet, Miami Beach, Florida. – 1998.

4. Lohn A., Musser M. AI and Compute: How Much Longer Can Computing Power Drive Artificial Intelligence Progress //Center for Security and Emerging Technology. https://doi. org/10.51593. – 2022.

5. Ву-Чун Фен, Кирк Камерон Green500: рейтинг энергетической эффективности [Электронный ресурс]. // Открытые системы: изд.научн. Журн. - Режим дотупа: URL:https://www.osp.ru/os/2008/01/4839411?ysclid=lfil0bl72k217199330 (дата обращения: 11.04.2023).

6. Markov I. L. Limits on fundamental limits to computation //Nature. – 2014. – Т. 512. – №. 7513. – С. 147-154.

UDC 004.891.2

LIMITATIONS OF NEURAL NETWORK CAPACITY GROWTH: PHYSICAL AND ENERGY ASPECTS

Kasyan V. A, Akhmetov R. Y, Senko N. S

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Vladymtsev V. D. – Assistant of the Department of Informatics

Annotation. This scientific work explores the limitations of neural network power growth caused by the physical and energy limitations of modern computers. The analysis of the influence of power consumption, cost growth and growth of computer performance on the development of neural networks is carried out, and optimization ways and possible solutions to reduce the power consumption and cost of training of neural networks are proposed.

Keywords. Neural networks, growth constraints, power, neural network training, energy efficiency, neural network architecture, computing resources, technological constraints, energy consumption optimization of neural networks, performance.