*59-я Научная Конференция Аспирантов, Магистрантов и Студентов БГУИР, Минск, 2023*

**ОГРАНИЧЕНИЯ РОСТА МОЩНОСТЕЙ НЕЙРОННЫХ СЕТЕЙ: ФИЗИЧЕСКИЕ И ЭНЕРГЕТИЧЕСКИЕ АСПЕКТЫ**

*Касьян В. А. студент группы 253501, Ахметов Р. Я. студент группы 25302,  
 Сенько Н. С. студент группы 25302, Владымцев В. Д. – ассистент каф. Информатики*

*Белорусский государственный университет информатики и радиоэлектроники*

*г. Минск, Республика Беларусь*

**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.*

Нейронные сети являются ключевым инструментом в области искусственного интеллекта и применяются в различных областях, от компьютерного зрения до голосового и текстового понимания. Увеличение эффективности и сложности нейронных сетей связано с увеличением вычислительных ресурсов и производительности компьютеров.

Однако, существует гипотеза, что рост мощностей нейронных сетей может попасть в серьезные физические и энергетические ограничения, вызванные ограничениями современных компьютеров. Например, увеличение энергопотребления, рост стоимости оборудования и инфраструктуры, а также замедление темпов роста производительности компьютеров могут привести к ограничению развития нейронных сетей или даже достижению практического предела их мощности.

Поэтому важно проводить анализ этих физических и энергетических ограничений и искать способы оптимизации и улучшения эффективности нейронных сетей. Это может включать в себя исследование новых архитектур нейронных сетей, использование более эффективных алгоритмов обучения, а также исследование возможности использования более энергоэффективных компьютерных систем для обучения и применения нейронных сетей.

Цель данной научной работы заключается в анализе физических и энергетических ограничений, с которыми сталкивается развитие нейронных сетей, и в предложении путей оптимизации и возможных решений для снижения энергоемкости и стоимости обучения нейронных сетей. В работе рассматриваются три ключевых аспекта, влияющих на ограничение мощностей нейронных сетей: энергопотребление, рост стоимости и рост производительности компьютеров.

Машинное обучение продолжает направляться к возрастанию потребления энергии, что делает модель дорогостоящей, неэффективной и неустойчивой. Большая часть компаний в погоне за качеством изменяет нейросети количественно: увеличивает их размер, что приводит к появлению невероятно ресурсоемких моделей.[1]

Такое поведение компаний приводит к тому, что мощности, потребляемые ими на тренировку и использование нейросетей растут экспоненциально, что в конечном итоге приводит к дефициту электроэнергии, не говоря о воздействии на экологию. Однако есть и путь улучшения качества нейросетей и их обучения.[2] Модели, разработанные с использованием оптимизированного и эффективного обучения, оказываются более точными и менее зависимыми от ресурсов. Нейронная сеть Vicuna 13B, основанная на LLaMA 13B, имеет размер в 13,5 раз меньше, чем GPT-3, но качество генерации текста сравнимое с GPT-3. Кроме того, GPT-4, имеющая примерно в 5000 раз больше параметров, отстает в точности отдельных генераций текстов, что демонстрирует, что правильно обученная, сравнительно небольшая модель может работать наравне с гораздо более крупными моделями. Такие модели также обладают высокой энергоэффективностью: Vicuna можно запустить на настольном компьютере или даже смартфоне, в то время как GPT-3 и GPT-4 требуют больших вычислительных кластеров для работы.

С ростом мощностей нейронных сетей увеличиваются и затраты на их обучение и эксплуатацию. Этот рост затрат связан с несколькими факторами, включая аппаратное обеспечение, энергопотребление, стоимость хранения данных и затраты на человеческие ресурсы.

С развитием машинного обучения возрастает спрос на мощные графические процессоры (GPU), специализированные интегральные схемы (ASIC) и другие высокопроизводительные вычислительные устройства. Это приводит к увеличению инвестиций в разработку и производство таких устройств, а также к снижению стоимости их производства на единицу вычислительной мощности.

Кроме того, энергопотребление нейронных сетей растет с увеличением их мощностей и сложности. Большие центры обработки данных, которые используются для обучения и эксплуатации нейронных сетей, потребляют огромное количество электроэнергии, что ведет к увеличению затрат на электричество и негативному воздействию на окружающую среду.

Для накопления и хранения больших объемов данных, необходимых для обучения нейронных сетей, требуются значительные инвестиции в инфраструктуру хранения данных, такую как облачные хранилища и серверы. В связи с этим увеличиваются затраты на поддержание и расширение такой инфраструктуры, что также сказывается на общих затратах на нейронные сети.

Разработка, обучение и эксплуатация нейронных сетей требуют высококвалифицированных специалистов в области машинного обучения, программирования и других технических дисциплин. Рост мощностей нейронных сетей и увеличение их сложности могут привести к увеличению затрат на оплату труда и обучение специалистов. В связи с этим многие компании инвестируют в разработку и обучение своих сотрудников, что также может повлиять на общие затраты.

Несмотря на все эти вызовы, индустрия машинного обучения и искусственного интеллекта продолжает развиваться и прогрессировать. Исследователи и разработчики постоянно стремятся к созданию более эффективных и энергоэффективных алгоритмов и архитектур нейронных сетей. В результате появляются новые методы обучения, которые требуют меньше вычислительных ресурсов и энергии.

Также растет использование аппаратных решений, специально предназначенных для работы с нейронными сетями, таких как специализированные интегральные схемы (ASIC) и вычислительные модули, например, Tensor Processing Units (TPU). Эти устройства предоставляют значительный прирост производительности и энергоэффективности по сравнению с традиционными чипами, что позволяет сократить затраты на обучение и эксплуатацию нейронных сетей.

С учетом роста количества транзисторов на интегральной схеме (Закон Мура), разработчики нейронных сетей могут использовать все больше вычислительных ресурсов для улучшения алгоритмов и обучения моделей. Однако наблюдается тенденция, подобная Закону Мура для нейросетей, закон Касьяна-Ахметова-Сенько (Закон КАС): удвоение требуемой мощности для обработки нейросетей происходит каждые 6 месяцев. Такая тенденция создает дополнительное давление на промышленность, которая уже едва ли может поддерживать увеличение мощностей по закону Мура, чего тем более не хватает для выполнения Закона КАС.

В целом, с увеличением мощностей нейронных сетей растут и стоимость их обучения и эксплуатации. Однако, благодаря активному развитию технологий и аппаратного обеспечения, инновационным методам обучения и оптимизации алгоритмов, индустрия машинного обучения и искусственного интеллекта продолжает находить пути для сокращения затрат и повышения эффективности своих решений.

Выводы:

Современные нейронные сети стали значительно мощнее и сложнее, что приводит к росту их энергопотребления, стоимости и производительности компьютеров. Рост нейросетевых моделей происходит по Закону КАС: удвоение требуемой мощности для обработки нейросетей происходит каждые 6 месяцев.

Экологические и экономические проблемы, связанные с ростом энергопотребления, привлекают внимание исследователей и инженеров, которые ищут методы оптимизации энергопотребления, включая снижение мощности моделей, использование более энергоэффективных алгоритмов и оборудования, а также разработку аппаратных решений, специализированных для работы с нейронными сетями. Важность такой оптимизации будет расти с усугублением экологических и энергетических проблем. Увеличение стоимости нейронных сетей связано с ростом затрат на аппаратное обеспечение, энергетические затраты, хранение данных и человеческие ресурсы.

**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.