YaFSDP позволяет сократить до 20% ресурсов GPU и ускоряет процесс на 26%

Яндекс разработала и опубликовала в открытом доступе на GitHub новый инструмент YaFSDP, призванный помочь компаниям, работающим с искусственным интеллектом, оптимизировать ресурсы при обучении крупных языковых моделей (LLM). Ключевые преимущества YaFSDP :

Является наиболее эффективным публично доступным средством для оптимизации использования памяти GPU и улучшения связи между графическими процессорами при обучении LLM.

Обеспечивает до 26% более высокую скорость обучения по сравнению с предыдущими версиями инструмента FSDP.

«Обучение LLM – это трудоёмкий и ресурсоёмкий процесс», – отметили в Яндексе. «Инженеры по машинному обучению и компании, разрабатывающие собственные LLM, тратят значительное время и ресурсы GPU – что равно деньгам – на обучение этих моделей. Чем больше модель, тем больше времени и затрат требуется на её обучение».

Компания оценивает, что использование YaFSDP для обучения модели с 70 миллиардами параметров может сэкономить ресурсы примерно 150 GPU, что составляет около 0,5–1,5 миллиона долларов в месяц, в зависимости от поставщика виртуальных GPU или платформы.

Используя передовые модели LLaMa от Meta, известные своими инновациями и поддержкой открытого ИИ, Яндекс продемонстрировала впечатляющие результаты своего инструмента YaFSDP:

На базе LLaMa 2 70B достигнуто финальное ускорение обучения на 21%

На LLaMa 3 70B ускорение составило 26%

Эти показатели свидетельствуют о высокой производительности YaFSDP в оптимизации ресурсов GPU и памяти при тренировке крупных языковых моделей.

Разработка YaFSDP — очередной вклад Яндекса в открытую экосистему ИИ. Ранее компания выпустила такие популярные инструменты, как:

CatBoost — Продвинутая библиотека градиентного бустинга на деревьях решений с открытым исходным кодом

YTsaurus — основная система для хранения и обработки данных Яндекса

AQLM — fддитивное квантование для языковых моделей

Petals — децентрализованный вывод и точная настройка больших языковых моделей

Многие крупные технологические компании также делают ИИ основой своих продуктов, например, недавно Apple анонсировала свои услуги Apple Intelligence в рамках предстоящего обновления iOS 18.

Публикация YaFSDP под открытой лицензией демонстрирует приверженность Яндекса принципам открытого ИИ и стремление внести весомый вклад в развитие отрасли, предоставляя сообществу передовые наработки. Это позволит другим компаниям и исследователям извлечь выгоду из более быстрого и экономичного обучения языковых моделей.