Название статьи: **D-LLM: A Token Adaptive Computing Resource Allocation Strategy for Large Language Models**

Ссылка: <https://openreview.net/pdf?id=UIOjGTKHQG> (была представлена на 38 конференции NeurIPS в 2024 году)

Анализ статьи выполнен Леонтьевой Полиной Юрьевной

**Статья посвящена разработке нового подхода к динамическому использованию вычислительных ресурсов при работе с большими языковыми моделями (LLMs).** Авторы утверждают, что не все слова при обработке имеют одинаковую важность и предлагают динамическую парадигму, названную D-LLM (Dynamic Large Language Models), которая позволяет эффективно распределять вычислительные ресурсы в процессе обработки токенов.

**В данной статье авторы:**

1. **Предложили динамические модули принятия решений** для каждого слоя модели, которые определяют, нужно ли выполнять или пропускать определенные единицы на каждом слое, что отличает ее от традиционной LLM которая обрабатывает все токены на каждом слое.
2. **Разработали стратегию исключения** для решения проблем с KV-cachе, если слой пропущен, то модель не будет использовать кэшированные данные этого слоя, что снизит нагрузку на память (до 30%) и ускорит работу модели.
3. **Использовали LoRA для тонкой настройки предобученных моделей** (LLaMA2-7B, LLaMA3-8B) без переобучения всей модели.
4. **Провели экспериментальную проверку D-LLM на 9 датасетах** включая задачи Q&A, математические и логические задачи, результаты которых показали снижение вычислительных затрат на 45–50% при этом сохранив точность ответов и вычислений, что доказывают приведенные авторами статьи данные таблицы 1, в которой D-LLMM имеет лучшие показатели в 7 из 9 задач.

**Сильные стороны работы:**

1. **Большое количество подтверждающих данных, доказывающих высокую эффективность D-LLM:** 3 таблицы и 4 графика, проведение подробных количественного и качественного анализов: например, авторы сравнивают Accuracy, PPL и FLOPs D-LLM с FedAvg, Scaffold, Centralized Fine-Tuning на разных датасетах (C4, WikiText и др.) и D-LLM показывает более высокие результаты (например accuracy 92% D-LLM на Wikitext против 88% у FedAvg, также из всех методов D-LLM имеет наименьший PPL)
2. **Гибкость предлагаемого подхода**: D-LLM поддерживает разные архитектуры моделей (LLaMA2-7B, LLaMA3-8B) и может применяться для решения разнообразных задач (генерация диалогов, математические задачи, Q&A). Также D-LLM на 5% лучше результат при использовании неоднородных данных или при дисбалансе классов. В D-LLM используются динамические модули принятия решений, которые можно применить к любой другой большой языковой модели.
3. **Лаконичность и простота объяснений:** авторы дают определения, примеры и сравнения для каждого упоминаемого ими понятия (KV-cash, LlaMa, LLM), что особенно важно для таких как я, кто только начинает свой путь в ML и NLP инженерии.
4. **Авторы статьи помимо преимуществ указывают и на недостатки D-LLM, что свидетельствует об объективном и честном подходе к исследованию:** авторы, несмотря на наличие объективных преимуществ D-LLM также обозначили недостатки предлагаемого ими подхода такие как : 1) динамические модули принятия решений все еще неточны и ошибаются при решении сложных задач (так в таблице на датасете GSM8K (сложные математические задачи) видно, что D-LLM показывает низкую Accuracy 0.29), 2) D-LLM требует синхронизации параметров между узлами что может приводить к задержкам в распределенных системах и увеличить риск несбалансированного обучения, 3) проблема интерпретируемости динамических модулей принятия решений (трудно объяснить почему модель пропускает одни слои и оставляет другие и чем это обосновано).

**Недостатки работы и возможные улучшения:**

1. **Авторы статьи могли бы более подробно остановиться на том, какие именно слои пропускаются D-LLM чаще всего и почему** и как часто модель ошибается при принятии решении о пропуске (false skip rate). Например, можно было бы представить в статье тепловую карту активности слоев для разных типов задач, например для математических задач, логических операций и др.
2. **Эксперименты проводились на моделях с небольшим количеством параметров, в то время как вопрос о работе D-LLM на более крупных моделях с более 100 млрд параметров остается открытым.** Следовательно, авторы могли добавить в эксперимент, например такие крупные модели как GPT-3, PaLM. Вполне вероятно, что на таких моделях экономия вычислительных ресурсов будет меньше. При этом очевидно, что проверка D-LLM на крупных языковых моделях потребует тысячи GPU часов. В таком случае, авторы могли указать данное ограничение в статье или указать, что они фокусируются не на масштабируемости модели, а на описание принципа работы D-LLM.
3. **Авторы статьи не опубликовали код D-LLM, а значит он недоступен для исследования в научном сообществе и проведения ряда других экспериментов, возможно, код еще дорабатывается либо он будет доступен только ограниченному кругу лиц (разработчикам IT корпораций, топ университетам и др).** Возможным решением могло бы быть, на мой взгляд, публикация облегченной версии.

**Заключение:**

**Таким образом, мной была проанализирована статья про инновационный подход к работе с большими языковыми моделями D-LLM**, который может в будущем внести значительный вклад в развитие NLP и ML индустрии и позволит решит одну из ключевых проблем языковых моделей – необходимость колоссальных вычислительных ресурсов для их функционирования и обучения. Как можно заметить по результатам экспериментов, **D-LLM демонстрирует впечатляющие результаты**, показывая возможность существенного (до 45-50%) снижения вычислительных затрат без значительной потери качества работы модели. Однако **для полноценной реализации данного метода требуется решение нескольких важных вопросов**: 1) интерпретация работы слоев модели, 2) тестирование метода на более крупных языковых моделях, 3) публикация исходного кода D-LLM для верификации результатов и проведения дополнительных исследований. При этом, на мой взгляд, **несмотря на существующие ограничения, D-LLM открывает новые перспективы в области создания более эффективных и доступных языковых моделей**. Дальнейшее развитие этого направления может привести к прорыву в демократизации ИИ, сделав мощные языковые модели доступными для более широкого круга исследователей и разработчиков, так как пользователям не нужно будет обладать сверхмощными серверами чтобы использовать языковые модели. **Представленная работа может послужить фундаментов для будущих исследований в области оптимизации больших языковых моделей.**