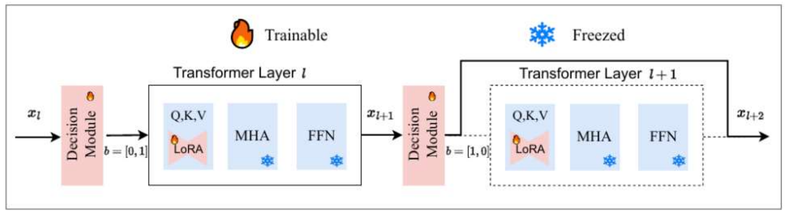
ЛЕТО С AIRI

D-LLM: Оптимизация вычислений для LLM (research proposal)

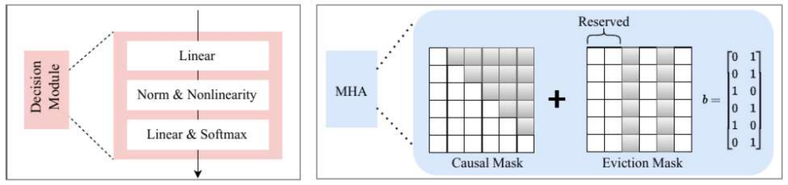
Леонтьева Полина Юрьевна

Мотивация

Большие языковые модели (LLM) требуют огромных вычислительных ресурсов, что существенно ограничивает их доступность, при этом существующие подходы неэффективно обрабатывают все токены одинаково, игнорируя различия в их важности. Для решения данной актуальной проблемы китайские исследователи разработали и апробировали D-LLM – динамическую большую языковую модель.



(a) The framework of D-LLMs.



(b) The Dynamic Decision Module. (c) The KV-cache Eviction Mask.

Рис 1. Структура предлагаемой D-LLM [1, P. 5]

Постановка задачи

Задача исследования - разработать динамическую стратегию распределения вычислений в LLM, которая сократит FLOPs на 45-50% за счёт выборочного пропуска слоёв, сохраняя точность на базовых NLP-задачах.

Задача research proposal- разработка решений для устранения ключевых ограничений D-LLM и улучшения модели

Описание подхода

Исследование состоит из 3 основных этапов:

1) **Разработки динамического модуля принятия решений. В модуле** каждый слой модели определяет нужно ли обрабатывать токен или пропустить его, используется LoRA для тонкой настройки предобученных моделей (LLaMA2-7B, LLaMA3-8B).

2) **Стратегии исключения**: если слой пропущен, то модель не будет использовать кэшированные данные этого слоя, что снизит нагрузку на память (до 30%) и ускорит работу модели

3) **Проведения экспериментов и проверка метрик качества модели** (9 датасетов: Q&A, Math, Alpaca и др.)

Разработка датасета

Авторы исследования разработали специализированный датасет для обучения и оценки модели D-LLM, объединив данные из 9 различных источников. В состав вошли текстовые корпуса (C4, WikiText - 60% данных), математические задачи (GSM8K - 20%) и логические задания (HotpotQA - 20%). Особое внимание уделялось балансировке: 50% коротких (до 512 токенов) и 50% длинных контекстов, с равномерным распределением по уровням сложности.

Качественные результаты экспериментов

1) Модель сохранила сопоставимую с базовыми LLM точность на 7 из 9 тестовых датасетов
2) Модель корректно идентифицирует и пропускает обработку служебных слов и избыточных повторов, сохраняя при этом ключевые смысловые токены.
3) На сложных задачах, требующих многошаговых рассуждений (математические вычисления, логические выводы), модель демонстрировала фрагментарность мышления и теряла цепочку аргументации.

Количественные результаты

1) Уменьшение FLOPs на 45–50%
2) Экономия памяти KV-кэша на 30%
3) Accuracy: 92% (WIkiText) , 89% (Q&A), 0.29 (мат задачи, выше чем у остальных моделей)
4) D-LLM демонстрирует снижение перплексии на 15%
5) Ускорение инференса в 1.8 раз

Dataset	*MoD* [59]		*Sh. Lla. PPL* [35]		*Sh. Lla. Tay.* [35]		*Ada-Inf.* [17]		*D-LLM*	
Q&A	PPL(↓)	FLOPs(↓)	PPL(↓)	FLOPs(↓)	PPL(↓)	FLOPs(↓)	PPL(↓)	FLOPs(↓)	PPL(↓)	FLOPs(↓)
Alpaca	10.32	0.56	7.09	0.66	7.65	0.66	319	0.65	6.01	0.59
SAMSum	4.47	0.56	4.39	0.66	4.66	0.66	874	0.56	3.18	0.55
Math	Acc.(↑)	FLOPs(↓)	Acc.(↑)	FLOPs(↓)	Acc.(↑)	FLOPs(↓)	Acc.(↑)	FLOPs(↓)	Acc.(↑)	FLOPs(↓)
GSM8K	0.08	0.56	0.10	0.66	0.18	0.66	0.00	0.83	0.29	0.59
MaWPS	0.33	0.56	0.52	0.66	0.39	0.66	0.00	0.90	0.74	0.56
Com. Sen.	Acc.(↑)	FLOPs(↓)	Acc.(↑)	FLOPs(↓)	Acc.(↑)	FLOPs(↓)	Acc.(↑)	FLOPs(↓)	Acc.(↑)	FLOPs(↓)
BoolQ	0.64	0.56	0.67	0.66	0.73	0.66	0.71	0.61	0.73	0.52
PIQA	0.49	0.56	0.76	0.66	0.83	0.66	0.55	0.63	0.84	0.52
SIQA	0.58	0.56	0.75	0.66	0.81	0.66	0.80	0.64	0.82	0.54
OBQA	0.42	0.56	0.63	0.66	0.81	0.66	0.78	0.76	0.80	0.53
MMLU	0.28	0.56	0.47	0.66	0.53	0.66	0.41	0.60	0.53	0.55

Рис 2. Сравнение производительности D-LLM с MoD, Sh.Lia, Ada-Inf в различных задачах [1, P. 7]

Выводы research proposal

В рамках research proposal были сформированы выводы, в которых предлагаются следующие возможные решения для улучшения модели в качестве перспективы для дальнейших исследований:

1) Представить тепловую карту активности слоев для разных типов задач

2) Провести эксперименты на более крупных языковых моделях

3) Провести дообучение (fine-tuning) модели на узкоспециализированных датасетах (GSM8K,синтетические данные с CoT разметкой)

[1] Yikun Jiang, Huanyu Wang, Lei Xie, D-LLM: A Token Adaptive Computing Resource Allocation Strategy for Large Language Models

[2] https://github.com/Jyk-122/D-LLM (репозиторий модели)

@POLLYLEO6