Исследование функции оценки токенов модели Transformer в контексте разреженности нейронных сетей

Егор Прокопов

Университет ИТМО, AI Talent Hub

email: egorprokopov216@gmail.com

Введение и мотивация

Архитектура **трансформер** является основой большинства современных **LLM**, а также **VLM**. Однако обучение (а в случае обработки изображений - и инференс), упирается в **квадратичную сложность вычислений и памяти** по длине последовательности.

Гипотеза этой работы заключается в том, что возможен **механизм отбора токенов**, который позволит оценивать, какие токены следует отправлять на обработку слоем трансформера, а какие - нет. В идеале, это позволит значительно уменьшить стоимость обработки последовательности, а также увеличить длину обрабатываемого контекста.

Связанные работы

Законы масштабирования, ставшие теоретическим обоснованием обучения по настоящему больших моделей на сотни миллиардов параметров, связывают качество обучения модели с количеством вычислений, данных и количеством параметров модели. Архитектура **Mixture-of-Experts** (MoE) позволила улучшить качество модели посредством увеличения количества параметров почти без увеличения количества вычислений на токен. Это было достигнуто благодаря идеям разреженных активаций - разные эксперты обрабатывают разные токены.

Разреженность в обработке текста может быть достигнута посредством пропуска токеном некоторых слоёв трансформера. В работе **Mixture-of-Recursions** (MoR) было показано, что для некоторых токенов последовательности необходима многократная обработка трансформерным слоём, но для большинства достаточно меньшего числа преобразований.

В свою очередь текст и изображения - две совершенно разные модальности. Главное, в текущем контексте, отличие заключается в том, что из изображений возможно удалить несколько пикселей, при этом не потеряв никакой информации. Таким образом, мы можем перекодировать изображение в такую последовательность, удаление элементов которой позволит сжать это изображение без потери семантики (а не просто пропускать некоторые слои, как в MoR). Из конкретных механизмов рассматривается **быстрое преобразование Фурье** (FFT) с низкочастотным фильтром и **сингулярное разложение** (SVD).

Токенизация изображений

Алгоритм FFT не справляется с задачей токенизации изображений, поэтому его выход обрабатывается небольшой свёрточной нейронной сетью. Функцией оценки токенов в этом токенизаторе является низкочастотный фильтр.

Алгоритм сингулярного разложения выступает бутылочным горлышком производительности, поскольку вычисления выполняются на CPU, а не на GPU. Поэтому, вместо использования самого алгоритма, вычисление матриц сингулярных векторов заменяется на два свёрточных блока, а вместо вычисления сингулярных чисел используется полносвязная нейронная сеть. Она же и является функцией оценки токенов.

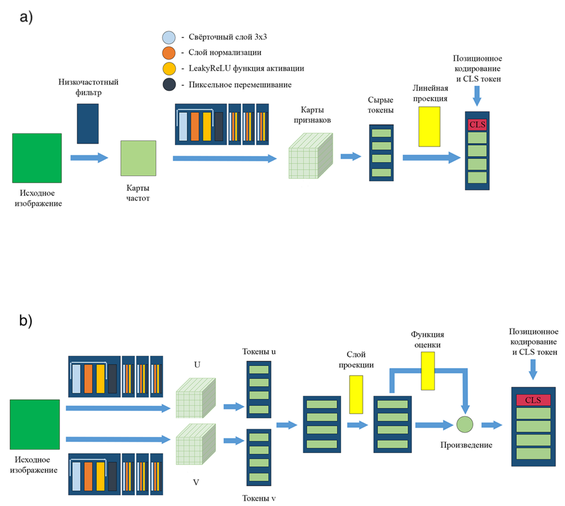

Рис. 1: *a) Токенизатор на основе FFT с низкочастотным фильтром. b) Токенизатор на основе SVD (SVD-TEF) с функцией оценки токенов*

Токенизатор на основе SVD предоставляет более гибкий инструментарий гейтинга токенов, чем токенизатор на основе FFT: выход функции оценки токенов можно рассматривать как объясненную дисперсию изображения. Это позволяет фильтровать токены с помощью накопленной объясненной дисперсии, благодаря чему длина последовательности после токенизации изображения не фиксирована.

Функция оценки токенов в трансформерах

Идея большинства работ, так или иначе связанных с фильтрацией токенов в трансформере, сводится к использованию идеи маршрутизатора MoE. Только вместо предсказания эксперта, этот маршрутизатор выносит вердикт: пропустить токен на слой трансформера или же отфильтровать.

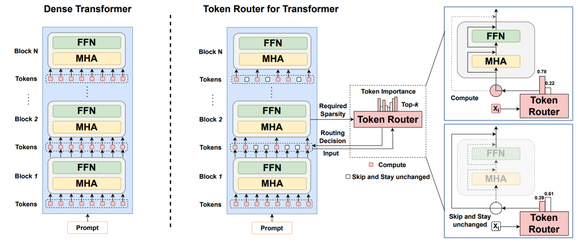

Рис. 2 *Взято из статьи* ***FTP: A Fine-Grained Token-wise pruner for Large Language Models via Token Routing***. Слева - обычный плотный трансформер. Справа - разреженный трансформер с фильтрацией (маршрутизацией) токенов.

В этой работе рассматривается другая идея: пусть нейронная сеть не выступает фильтром сама по себе, а будет давать оценку объясненной дисперсии каждого токена. Фильтрация токенов же происходит с помощью накопленной дисперсии (наподобие SVD-токенизатора изображений).

Анализ результатов и выводы

В работе с текстом для тестов была взята модель **GPT-2**, а дообучение происходило на датасете wikitext. Несмотря на хорошие результаты во время обучения, при фильтрации даже самого малого числа токенов, модель (GPT2-TEF) показывала серьезную деградацию качества.

	GPT2	t=1.00	t=0.99	t=0.95
Perplexity	**16.8**	21.0	509.3	588.6
seq_length	**512**	512	505.	482.2

Таблица 1 *Результаты обучения GPT2 и GPT2-TEF. В качестве целевой метрики выбрана perplexity. seq_length - средняя длина последовательности. t - порог фильтрации по накопленной обьясненной дисперсии.*

Наилучший результат в токенизации изображений удалось достичь при использовании токенизатора на основе сингулярного разложения с функцией оценки токенов (SVD-TEF).

	ViT	t=0.99	t=0.95	t=0.90	**t=0.80**	t=0.70	t=0.50	t=0.10
F1	0.79	0.85	0.85	0.85	**0.85**	0.83	0.77	0.62
seq_length	257	253.6	242.2	228.9	**192.5**	166.0	120.2	23.1

Таблица 2 *Результаты токенизации изображений SVD TEF и ViT. В верхней строке представлены сравниваемые модели.*