Исследование эффективных трансформеров

М. Д. Ковалева

Московский физико-технический институт

14 декабря 2022 г.

Постановка задачи

Проблема

Трансформеры полагаются на обучаемый механизм внимания. К сожалению, обычный рансформер масштабируется квадратично с количеством токенов L во входной последовательности, что является непомерно дорогим для большого L.

Цель работы —

сравнение нескольких недавних методов улучшения моделей трансформеров: performer и informer

Задачи работы

1) изучить статьи,
2) реализовать модели: baseline, informer, performer,
3) провести эксперименты по сравнению потребления памяти скорости моделей на данных танзакций

[1] *Martins, Pedro Henrique, Zita Marinho, and André FT Martins.* "∞ – *former* : *InfiniteMemoryTransformer*." *Informer*" *model*

[2] *Choromanski, Krzysztof, et al.* "Rethinking attention with performers."arXiv preprint arXiv:2009.14794 (2020). - "Performer"model

[3] *Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I.* Attention is all you need. - "Full attention"model

Performer

Классический attention

$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V = D^{-1}AV$, $A = exp(\frac{QK^T}{\sqrt{d_k}})$, $D = diag(A1_L)$, где $A \in \mathbb{R}^{LxL}$ - матрица внимания. Сложность метода зависит от ее размера, который квадратичен по длине входной последовательности.

Основная идея performer

Приблизить матрицу A с помощью positive Orthogonal Random features (FAVOR+) механизма произзведение матрицц меньшей разметности как:
$A = Q'(K')^T$, где $Q', K' \in \mathbb{R}^{Lxr}$
$A(i,j) = \mathbf{K}(q_i^T, k_j^T)$, где q_i, k_j это i-тый и j-тый строки в матрицах Q, K и $\mathbf{K}$ - ядро $\mathbb{R}^d x \mathbb{R}^d \to \mathbb{R}_+$ определенное отображением $\phi : \mathbb{R}^d \to \mathbb{R}_+^r$ как $\mathbf{K}(x,y) = \mathbb{E}[\phi(x)^T, \phi(y)]$. (В матрицах Q', K' строки задатся $\phi(q_i^T)^T$ и $\phi(k_i^T)^T$ соответственно)

Преимущества

Сложность метода становится линейной по длине входной последовательности.

Informer

Основная идея

Использование неограниченной долговременной памяти (Long term memory, LTM), которая позволяет модели обрабатывать произвольно длинные контексты. Чтобы сделать LTM неограниченным, используется структура внимания в непрерывном пространстве (continuous-space attention framework). В этом подходе входная последовательность представляется в виде непрерывного сигнала, выраженного в виде линейной комбинации N радиальных базисных функций.

Преимущества

1. Контекст может быть представлен с использованием числа базовых функций N, меньшего, чем количество токенов, что снижает вычислительные затраты на внимание
2. N может быть фиксированным, что позволяет представлять неограниченный контекст в памяти без увеличения его сложности для внимания.

Список литературы

1. *Martins, Pedro Henrique, Zita Marinho, and André FT Martins.* "∞ – *former* : *InfiniteMemoryTransformer*.
2. *Choromanski, Krzysztof, et al.* "Rethinking attention with performers."arXiv preprint arXiv:2009.14794 (2020).
3. *Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I.* Attention is all you need.