2022 年 6 月 17 日 統計数理研究所 オープンハウス

最近のセルフアテンションの計算量削減手法の紹介

三原 千尋　総合研究大学院大学 統計科学専攻 博士課程(5年一貫制)5年

概要 ごく最近のセルフアテンションの計算量削減手法 3 つを筆者の視点で比較しつつ紹介する。結論からいうと、Performer と Skyformer は通常のセルフアテンションをよく近似できることを根拠にしているがこれらは直接比較しづらく、Informer はそもそも議論の出発点が異なる。

Transformer は原理的に離れた単語間の依存関係を直接扱える点が利点だが、「i 単語目から j 単語目にどれだけ注意すべきか (セルフアテンション)」をすべて計算するために系列長 n に対し $\mathcal{O}(n^2)$ の計算量を要する。長い系列の依存関係を扱うには結局この計算量がボトルネックになる。

Transformer のセルフアテンション

$$y_i = \sum_{j=1}^{n} a_{i,j} v_j = \sum_{j=1}^{n} \underset{j}{\mathrm{softmax}}\left(\frac{q_i \cdot k_j}{\sqrt{d}}\right) v_j = \sum_{j=1}^{n} \left[\frac{\exp(q_i \cdot k_j/\sqrt{d})}{\sum_{j'=1}^{n} \exp(q_i \cdot k_{j'}/\sqrt{d})}\right] v_j$$

各単語の特徴ベクトル v_i を、前後の文脈を反映した特徴ベクトル y_i にしたい。
そのための重み $a_{i,j}$ が i 単語目から j 単語目へのセルフアテンションである。
セルフアテンション層は各単語の特徴 v_i 及び $a_{i,j}$ を与える q_i, k_i を学習する。
※ 「単語」「文脈」といっているが自然言語に限らず系列データに適用できる。
時系列を各時点までの特徴ベクトル列にする場合は未来方向には注意しない。

セルフアテンションの計算量を削減する手法の提案は最近盛んであり、以下の路線がみられる(筆者による分類であり排他的・網羅的ではない)。

- **スパース路線** 「i 単語目から j 単語目にどれだけ注意すべきか」の行列 $(a_{i,j})_{i,j}$ の要素をすべて計算するのではなく、一部省略する。
  – 例. Informer [Zhou et al., 2021]
- **低ランク近似路線** 「i 単語目から j 単語目にどれだけ注意すべきか」の行列 $(a_{i,j})_{i,j}$ を得ようとはするが、低ランク近似する。
  – 例. Skyformer [Choromanski et al., 2021]
- **要約(交換)路線** 「注意の重みを乗じること」と「和をとること」を交換する。いい換えると、「i 単語目から文章全体にどれだけ注意すべきか」というように文章を要約する。原理的に計算量が $\mathcal{O}(n)$ になる。
  – 例. Performer [Chen et al., 2021]

これらの計算量削減版 Transformer の有効性は主にベンチマークタスクにおける精度で経験的に示されている。その手法をとる理論的根拠も示されているが、切り口が三者三様であり、手法間の定性的な比較は難しい。上の例に示した各手法は理論的根拠を以下のようにおいている。

Informer [Zhou et al., 2021]

手法の概要 セルフアテンションの行列 $(a_{i,j})_{i,j}$ のうち一様分布からのカルバック・ライブラー距離が大きい行 (=重要な「注意」を含んでいる見込みが高い) のみを計算し、他の行は計算せずに一様分布とする方針を取る。しかし、KL距離を計算しようとすると結局セルフアテンションの行列のすべての要素が必要になるので、サンプリングによってその上限を見積もっている (なお、Informer 自体はセルフアテンションのスパース化以外の計算量削減も盛り込んでいる)。

理論的根拠 あるベクトルの Softmax と一様分布とのKL距離が一定確率でここまでに収まるという上限を、ベクトルの一部の要素のサンプリングから見積もることができる。

これは一様分布との KL 距離が小さい行を省くことが前提になっている。

Performer [Choromanski et al., 2021]

手法の概要 カーネル法の計算量削減手法として知られる Random Feature を応用する。具体的に、セルフアテンションの式中の $\exp(q_i \cdot k_j/\sqrt{d})$ の箇所を何らかの特徴空間における内積 $\phi(q_i) \cdot \phi(k_j)$ で表現できれば式中の和と交換できるが、Performer では以下の特徴写像 (RPF: Positive Random Feature) でこれを表現した。

$$\phi_{\mathrm{RPF}}(x) = \frac{\exp(-\|x\|^2/2)}{\sqrt{m}}\big(\exp(w_1 \cdot x), \cdots, \exp(w_m \cdot x)\big)$$

$w_1, \cdots, w_m$ は d 次元標準正規分布からサンプリングしたランダムベクトル。

理論的根拠 RPF によるセルフアテンションは $\phi(q_i) \cdot \phi(k_j)$ 通常の Softmax によるセルフアテンションとの2乗誤差が q_i, k_j, d, m に依存する値で上から抑えられる。

これは Transformer と同じ表現力が期待できるという議論と捉えられる。

Skyformer [Chen et al., 2021]

手法の概要 セルフアテンションにカーネル法における Nyström 近似を適用する。先行手法の Nyströmformer [Xiong et al., 2021] では Nyström 近似を直接適用していたが、Skyformer では対称行列に拡張して Nyström 近似をより適切に適用している。具体的に、$(z_1, \cdots, z_n, z_{n+1}, \cdots, z_{2n}) = (q_1, \cdots, q_n, k_1, \cdots, k_n)$ に対して以下のグラム行列 B を考え、いくつかのインデックスをサンプリングして Nyström 近似 $\overline{B}$ を得て、この右上 $n \times n$ ブロックをセルフアテンションとする。

$$B = \begin{pmatrix} \exp(-\frac{\|z_1 - z_1\|^2}{2\sqrt{d}}) & \exp(-\frac{\|z_1 - z_2\|^2}{2\sqrt{d}}) & \cdots \\ \exp(-\frac{\|z_2 - z_1\|^2}{2\sqrt{d}}) & \exp(-\frac{\|z_2 - z_2\|^2}{2\sqrt{d}}) & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

指数型でなくガウシアンカーネルなのは計算の安定化のためとある。

理論的根拠 この手法で得られるセルフアテンションの行列は通常のセルフアテンションの行列 $(a_{i,j})_{i,j}$ とのスペクトルノルム (最大固有値) がほとんどの確率で小さくなることがいえる。

Nyström 近似が固有値分解に基づくためスペクトルノルムでの議論になっている。これも Transformer と同じ表現力が期待できるという議論と捉えられるが、Performer と直接比較できる形にはなっていない。

いずれの手法も GitHub に動くコードがあるのでそのリポジトリの URL も示す。

参考文献

[Zhou et al., 2021] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, Wancai Zhang. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. in *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI 2021).*
https://www.aaai.org/AAAI21Papers/AAAI-7346.ZhouHaoyi.pdf
https://github.com/zhouhaoyi/Informer2020

[Choromanski et al., 2021] Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J Colwell, Adrian Weller. Rethinking Attention with Performers. In *Proceedings of the 9th International Conference on Learning Representations (ICLR 2021).*
https://openreview.net/forum?id=Ua6zuk0WRH
https://github.com/lucidrains/performer-pytorch

[Chen et al., 2021] Yifan Chen, Qi Zeng, Heng Ji, Yun Yang. Skyformer: Remodel Self-Attention with Gaussian Kernel and Nyström Method. in *Advances in Neural Information Processing Systems 34 (NeurIPS 2021).*
https://proceedings.neurips.cc/paper/2021/hash/10a7cdd970fe135cf4f7bb55c0e3b59f-Abstract.html
https://github.com/pkuzengqi/Skyformer