Attention

Xie Zejian
11810105@mail.sustech.edu.cn

Department of Finance, SUSTech

May 20,2021

Contents

2 Bound of the "Losing Rank" Rate

2.1 Structure and Path Decomposing

For one head,

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK'}{\sqrt{d}}\right)V$$

where $Q = W^Q\mathbf{X}$ and $K = W^K\mathbf{X}$.

Note that the core step of attention is the softmax function and it has the shift invariance property, namely

$$\text{softmax}\left(\mathbf{A} + \mathbf{x1}'\right) = \text{softmax}(\mathbf{A})$$

The property above is the basis of path decomposition theorem.

The output of SAN is given by

$$SA_h(\mathbf{X}) = \mathbf{P_h}(\mathbf{XW_{V,h}} + \mathbf{eb}'_{\mathbf{V,h}})$$

where $\mathbf{P_h} \in \mathbb{R}^{n\times n}$ is being softmax, thus $\mathbf{P_h}\mathbf{e} = \mathbf{e}$ and hence

$$SA_h(\mathbf{X}) = \mathbf{P_h}\mathbf{XW_{V,h}} + \mathbf{eb}'_{\mathbf{V,h}}$$

Concat the H hands, we have

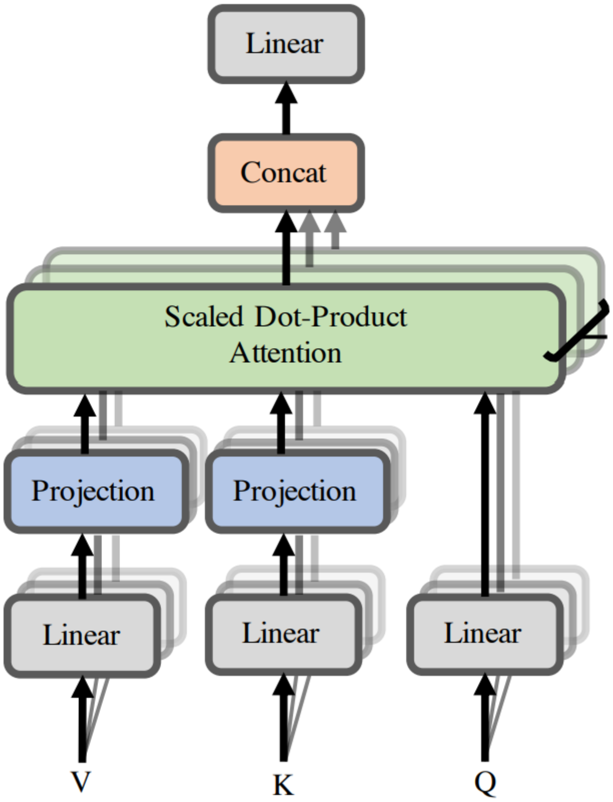


Figure 1: Attention Mechanism

$$
\begin{aligned}
SA(\mathbf{X}) &= [SA_1(\mathbf{X}) \quad ... \quad SA_H(\mathbf{X})] \begin{bmatrix} W_{O,1} \\ ... \\ W_{O,H} \end{bmatrix} \\
&= \sum_h \left(\mathbf{P_h X W_{V,h} W_{O,h}} + \mathbf{e b'_{V,h} W_{O,h}} \right)
\end{aligned}
$$

Let $\mathbf{W_h} := \mathbf{W_{V,h} W_{O,h}}$ and $\mathbf{b_{O,h}} := \mathbf{b'_{V,h} W_{O,h}}, \mathbf{b_O} = \sum_h \mathbf{b_{O,h}}$, then

$$
SA(\mathbf{X}) = \sum_h \mathbf{P_h X W_h} + \mathbf{e b'_O}
$$

Stack SA layer, we have

$$
\begin{aligned}
SA \circ SA(\mathbf{X}) &= \sum_h \mathbf{P_h} (\sum_{\mathbf{h}} \mathbf{P_h X W_h} + \mathbf{eb'}) \mathbf{W_h} + \mathbf{eb'} \\
&= \sum_{h_1, h_2} \mathbf{P_{h_1} P_{h_2} X W_{h_1} W_{h_2}} + \mathbf{eb'}
\end{aligned}
$$

where $\mathbf{b}$ isn't the same all the time, but we omitted the script by defining new $\mathbf{b}$ with summation. Hence

Theorem 2.1 (Path decomposition)**.**

$$
SAN(\mathbf{X}) = \sum_{path \in [H]^L} \mathbf{P_{path} X W_{path}} + \mathbf{eb'}
$$

where $SAN = SA_L \circ ... \circ SA_1$ (L layers with H heads). Note if we include dummy heads, aka, skip layers, the cardinality should be $(H+1)^L$ instead of H^L.

2.2 Bound

Definition 2.1.

- Composite norm of a matrix is defined as

$$
\|\mathbf{X}\|_{1,\infty} = \sqrt{\|\mathbf{X}\|_1 \|\mathbf{X}\|_\infty}
$$

- The residual of the matrix is defined as

$$
\text{res}\,(\mathbf{X}) = \mathbf{X} - \mathbf{1x'} \text{ where } \mathbf{x} = \arg\min_{\mathbf{x}} \|\mathbf{X} - \mathbf{1x'}\|_2
$$

We want to prove that $\|\text{res}\,(\text{SAN}(\mathbf{X}))\|_{1,\infty} \to 0$ as $L \to \infty$ where L is the number of layers in transformer.

Theorem 2.2 (bound of a single head single layer)**.** *When $\left\|\mathbf{W}^l_{QK}\right\|_1 \|\mathbf{W}^l_V\|_{1,\infty} \leq \beta$, we have*

$$
\|res\, SAN(\mathbf{X})\|_{1,\infty} \leq \left(\frac{4\beta}{\sqrt{d_{qk}}} \right)^{\frac{3^L - 1}{2}} \|res\, \mathbf{X}\|^{3^L}_{1,\infty}
$$

If the theorem of above is true, we can use path decomposition theorem to prove that the right side of this inequality → 0, then we proved that the transformer structure will lead to "rank losing" of input matrix. In NLP tasks, it means that the input will be embedded into the same word vector as $L \to \infty$.

2.3 Complexity of Self-Attention

Theorem 2.3 (self-attention is low rank)**.**

$$\mathbb{P}\left\{\left\|\widetilde{\mathbf{P}}\boldsymbol{\omega}' - \mathbf{P}\boldsymbol{\omega}'\right\| \leq \varepsilon \left\|\mathbf{P}\boldsymbol{\omega}'\right\|\right\} > 1 - o(1)$$

where rank $\left(\widetilde{\mathbf{P}}\right) = \Theta(\log n)$.

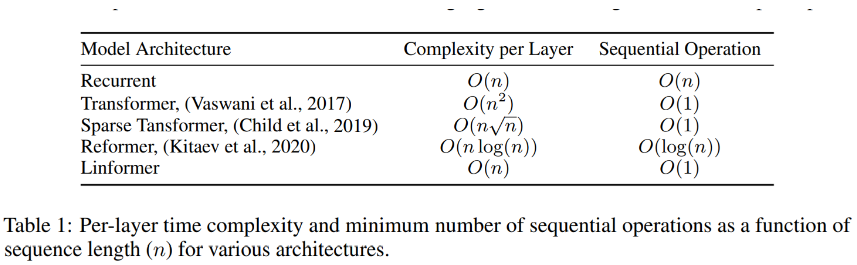
Proof.

$$P = \text{softmax}\left\{\frac{A}{\sqrt{d_{qk}}}\right\} = \exp\sqrt{\mathbf{d_{qk}}\mathbf{A}} \cdot \mathbf{D}_{\mathbf{A}}^{-1}$$

By JL lemma, we claim the approximate matrix can be $\widetilde{\mathbf{P}} = \mathbf{P}\mathbf{R}'\mathbf{R}$ where $\mathbf{R}$ is random $k \times n$ matrix with $\mathcal{N}(0, \frac{1}{k})$ entries. Then we can SVD $\widetilde{P}$ to achieve linear complexity

$$\widetilde{P} \simeq \sum_{i=1}^{k} \sigma_i \mathbf{u_i}\mathbf{v_i'}$$

□



Model Architecture	Complexity per Layer	Sequential Operation
Recurrent	$O(n)$	$O(n)$
Transformer, (Vaswani et al., 2017)	$O(n^2)$	$O(1)$
Sparse Tansformer, (Child et al., 2019)	$O(n\sqrt{n})$	$O(1)$
Reformer, (Kitaev et al., 2020)	$O(n\log(n))$	$O(\log(n))$
Linformer	$O(n)$	$O(1)$

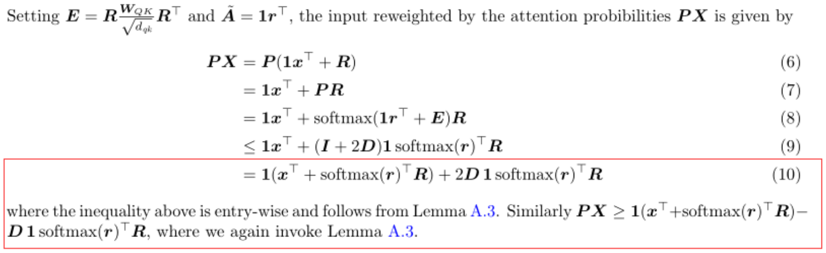
Table 1: Per-layer time complexity and minimum number of sequential operations as a function of sequence length (n) for various architectures.

Figure 2: Complexity of Transformer

According to the theorem 2.3, we can find a low rank matrix, i.e. rank $\tilde{P} = \Theta(\log n \cdot \log n)$ to replace P in any self-attention layer.

It means that the low rank property of the softmax function is the reason of the losing of rank.

2.4 Some Problem in the Papers



Setting $\boldsymbol{E} = \boldsymbol{R}\frac{\boldsymbol{W}_{QK}}{\sqrt{d_{qk}}}\boldsymbol{R}^\top$ and $\tilde{\boldsymbol{A}} = \mathbf{1}\boldsymbol{r}^\top$, the input reweighted by the attention probibilities $\boldsymbol{PX}$ is given by

$$\begin{aligned} \boldsymbol{PX} &= \boldsymbol{P}(\mathbf{1}\boldsymbol{x}^\top + \boldsymbol{R}) & (6)\\ &= \mathbf{1}\boldsymbol{x}^\top + \boldsymbol{PR} & (7)\\ &= \mathbf{1}\boldsymbol{x}^\top + \text{softmax}(\mathbf{1}\boldsymbol{r}^\top + \boldsymbol{E})\boldsymbol{R} & (8)\\ &\leq \mathbf{1}\boldsymbol{x}^\top + (\boldsymbol{I} + 2\boldsymbol{D})\mathbf{1}\,\text{softmax}(\boldsymbol{r})^\top\boldsymbol{R} & (9)\\ &= \mathbf{1}(\boldsymbol{x}^\top + \text{softmax}(\boldsymbol{r})^\top\boldsymbol{R}) + 2\boldsymbol{D}\,\mathbf{1}\,\text{softmax}(\boldsymbol{r})^\top\boldsymbol{R} & (10)\end{aligned}$$

where the inequality above is entry-wise and follows from Lemma A.3. Similarly $\boldsymbol{PX} \geq \mathbf{1}(\boldsymbol{x}^\top + \text{softmax}(\boldsymbol{r})^\top\boldsymbol{R}) - \boldsymbol{D}\,\mathbf{1}\,\text{softmax}(\boldsymbol{r})^\top\boldsymbol{R}$, where we again invoke Lemma A.3.

1. The proof in A.1, shown in Figure **??**.
2. The norm in the definition of residual. If a norm does not satisfy triangle inequality, can it represent distance?
3. Why the input residual is smaller than 1?
4. The problems in conclusion part.