Vision Transformer Compression and Architecture Exploration with Efficient Embedding Space Search Supplement Material

Daeho Kim[1][0000-0003-2921-3168] and Jaeil Kim[2,⋆][0000-0002-9799-1773]

[1] Department of Artificial Intelligence, Kyungpook National University
[2] School of Computer Science and Engineering, Kyungpook National University
{kdaeho27,threeyears}@gmail.com

1 Proof of Theorem 4

In this section, we prove Theorem 4. We compare Pre-LN Transformer and Res-Post-LN transformer in Table 1. σ indicates an activation function, and $W^{1,l} \in \mathbb{R}^{d_m \times d}$ and $W^{2,l} \in \mathbb{R}^{d \times d_m}$ are weight parameters in MLP. To calculate analytically, the hidden dimension d_m and the feature dimension d are considered identical in MLP. We denote the output feature $Y^{\text{res}}_{\text{l},i}$ at i-th position of the l-th layer, where $Y^{\text{res}}_{\text{l},i}$ is a real-valued tensor of dimension d with $i = 1, 2, \ldots, n$ and $l = 1, 2, \ldots, L$.

Lemma 1. *For the Res-Post-LN transformer with initialized weights,* $\mathbb{E}(\|Y^{res}_{l+1,i}\|_2^2) \leq 2[(2l-1)d^2 + ld]$ *for all* $l > 0$ *and* i.

Proof of Lemma 1 We first estimate the expectation of $\|Y^{\text{res}}_{l,i}\|_2^2$ for $l > 0$. Since $\|\text{LN}(\cdot)\|_2^2 = d$ by Lemma 2 in [1],

$$\begin{aligned}
\mathbb{E}(\|Y^{\text{res}}_{l,i}\|_2^2) &= \mathbb{E}(\|Y^{\text{res}}_{1,i} + \sum_{k=1}^{l-1}(\text{LN}(Y^{\text{res},2}_{k,i}) + \text{LN}(Y^{\text{res},6}_{k,i}))\|_2^2) \\
&\leq \mathbb{E}(\|Y^{\text{res}}_{1,i} + \sum_{k=1}^{l-1}\text{LN}(Y^{\text{res},2}_{k,i}) + \sum_{k=1}^{l-1}\text{LN}(Y^{\text{res},6}_{k,i})\|_2^2) \\
&\leq \mathbb{E}(\|Y^{\text{res}}_{1,i}\|_2^2) + \mathbb{E}(\|\sum_{k=1}^{l-1}\text{LN}(Y^{\text{res},2}_{k,i})\|_2^2) + \mathbb{E}(\|\sum_{k=1}^{l-1}\text{LN}(Y^{\text{res},6}_{k,i})\|_2^2) \\
&\leq 2(l-1)d
\end{aligned}$$

⋆ Corresponding author

Table 1: Pre-LN Transformer v.s. Res-Post-LN Transformer

Pre-LN Transformer	Res-Post-LN Transformer
$Y_l^{\text{pre},1} = \text{LN}(Y_l^{\text{pre}})$	$Y_l^{\text{res},1} = \text{LN}(Y_l^{\text{res}})$
$Y_l^{\text{pre},2} = \text{MSA}(Y_l^{\text{pre},1})$	$Y_l^{\text{res},2} = \text{MSA}(Y_l^{\text{res}})$
	$Y_l^{\text{res},3} = \text{LN}(Y_l^{\text{res},2})$
$Y_l^{\text{pre},3} = Y_l^{\text{pre}} + Y_l^{\text{pre},2}$	$Y_l^{\text{res},4} = Y_l^{\text{res}} + Y_l^{\text{res},3}$
$Y_l^{\text{pre},4} = \text{LN}(Y_l^{\text{pre},3})$	$Y_l^{\text{res},5} = \text{LN}(Y_l^{\text{res},4})$
$Y_l^{\text{pre},5} = \sigma(Y_l^{\text{pre},4}W^{1,l} + b^{1,l})W^{2,l} + b^{2,l}$	$Y_l^{\text{res},6} = \sigma(Y_l^{\text{res},5}W^{1,l} + b^{1,l})W^{2,l} + b^{2,l}$
	$Y_l^{\text{res},7} = \text{LN}(Y_l^{\text{res},6})$
$Y_{l+1}^{\text{pre}} = Y_l^{\text{pre},5} + Y_l^{\text{pre},3}$	$Y_{l+1}^{\text{res}} = Y_l^{\text{res},4} + Y_l^{\text{res},7}$
Final LN : $Y_{\text{Final}}^{\text{pre},1} = \text{LN}(Y_{l+1}^{\text{pre}})$	Final LN : $Y_{\text{Final}}^{\text{res},1} = \text{LN}(Y_{l+1}^{\text{res}})$

In a similar way, we have

$$\begin{aligned}
\mathbb{E}(\|Y_{l+1,i}^{\text{res}}\|_2^2) &= \mathbb{E}(\|Y_{l,i}^{\text{res},4}\|_2^2) + \mathbb{E}(\|Y_{l,i}^{\text{res},7}\|_2^2) + 2\mathbb{E}(Y_{l,i}^{\text{res},7}Y_{l,i}^{\text{res},4\ \text{T}}) \\
&= \mathbb{E}(\|Y_{l,i}^{\text{res}}\|_2^2) + \mathbb{E}(\|Y_{l,i}^{\text{res},\ 3}\|_2^2) + \mathbb{E}(\|Y_{l,i}^{\text{res},7}\|_2^2) + 2\mathbb{E}(Y_{l,i}^{\text{res},7}Y_{l,i}^{\text{res},4\ \text{T}}) \\
&\leq 2(l-1)d + 2d + 2(2(l-1)d^2 + d^2) \\
&\leq 2[(2l-1)d^2 + ld]
\end{aligned}$$

Then, we have $\mathbb{E}(\|Y_{l+1,i}^{\text{res}}\|_2^2) \leq 2[(2l-1)d^2+ld]$, which is bounded by $\mathbb{E}(\|Y_{l+1,i}^{\text{res}}\|_2^2) \leq \mathcal{O}(ld^2)$.

Proof of Therorem 4 The loss of the Res-Post-LN Transformer can be described as:

$$\tilde{\mathcal{L}}(Y_{\text{Final},1}^{\text{res}}, \ldots, Y_{\text{Final},n}^{\text{res}}) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(Y_{\text{Final},i}^{\text{res}})$$

Using back-propagation, the gradient of $\mathcal{L}(Y_{\text{Final}}^{\text{res}}, i)$ with respect to the last layer weights $W^{2,L}$ can be written as

$$\begin{aligned}
\frac{\partial \mathcal{L}(Y_{\text{Final},i}^{\text{res}})}{\partial W_{pq}^{2,L}} &= \frac{\partial \mathcal{L}(Y_{\text{Final},i}^{\text{res}})}{\partial Y_{\text{Final},i}^{\text{res}}} \frac{\partial Y_{\text{Final},i}^{\text{res}}}{\partial Y_{L+1,i}^{\text{res}}} \frac{\partial Y_{L+1,i}^{\text{res}}}{\partial Y_{L,i}^{\text{res},7}} \frac{\partial Y_{L,i}^{\text{res},7}}{\partial Y_{L,i}^{\text{res},6}} \frac{\partial Y_{L,i}^{\text{res},6}}{\partial W_{pq}^{2,L}} \\
&= \frac{\partial \mathcal{L}(Y_{\text{Final},i}^{\text{res}})}{\partial Y_{\text{Final},i}^{\text{res}}} \mathbf{J}_{LN}(Y_{L+1,i}^{\text{res}}) \mathbf{J}_{LN}(Y_{L,i}^{\text{res},6})(0, 0, \ldots, [\sigma(Y_{L,i}^{\text{res},5}W^{1,L})]_p, \\
&\quad \ldots, 0, 0)^T
\end{aligned}$$

Here, $b^{2,l}$ is initialized to be 0 and $[\sigma(Y_{L,i}^{\text{res},5}W^{1,L})]_p$ denotes the $p-$th element of $\sigma(Y_{L,i}^{\text{res},5}W^{1,L})$. The absolute value can be bounded by as follows:

$$\begin{aligned}|\frac{\partial \mathcal{L}(Y_{\text{Final},i}^{\text{res}})}{\partial W_{pq}^{2,L}}| &\leq \|\frac{\partial \mathcal{L}(Y_{\text{Final},i}^{\text{res}})}{\partial Y_{\text{Final},i}^{\text{res}}}\|_2 \|\mathbf{J}_{LN}(Y_{L+1,i}^{\text{res}})\|_2 \|\mathbf{J}_{LN}(Y_{L,i}^{\text{res},6})\|_2 \\ &\quad \|(0,0,\ldots,[\sigma(Y_{L,i}^{\text{res},5}W^{1,L})]_p,\ldots,0)\|_2 \\ &= \|\frac{\partial \mathcal{L}(Y_{\text{Final},i}^{\text{res}})}{\partial Y_{\text{Final},i}^{\text{res}}}\|_2 \|\mathbf{J}_{LN}(Y_{L+1,i}^{\text{res}})\|_2 \|\mathbf{J}_{LN}(Y_{L,i}^{\text{res},6})\|_2 \\ &\quad |[\sigma(Y_{L,i}^{\text{res},5}W^{1,L})]_p|\end{aligned}$$

According to Lemma 2 in [1], $\|Y_{L,i}^{\text{res},5}\|_2^2 = d$ and $[Y_{L,i}^{\text{res},5}W^{1,L}]_p$ obeys normal distribution $N(0,1)$, we can have the following inequality using Chernoff bound

$$\Pr[|[Y_{L,i}^{\text{res},5}W^{1,L}]_p| \geq a_0] \leq \exp(-\frac{a_0^2}{2})$$

$$\Pr[\sigma([Y_{L,i}^{\text{res},5}W^{1,L}]_p)^2 \geq 2\ln 100d] \leq \frac{0.01}{d}$$

We have $\sigma([Y_{L,i}^{\text{res},5}W^{1,L}]_p)^2 \leq 2\ln 100d$ with probability at least 0.99, for all $p = 1,2,\ldots,d$. Since with probability $1-\delta(\epsilon)$, $\frac{|\|Y_{L+1,i}^{\text{res}}\|_2^2 - \mathbb{E}\|Y_{L+1,i}^{\text{res}}\|_2^2|}{\mathbb{E}\|Y_{L+1,i}^{\text{res}}\|_2^2} \leq \epsilon$, we have $\|Y_{L+1,i}^{\text{res}}\|_2^2 \leq (1+\epsilon)\mathbb{E}\|Y_{L+1,i}^{\text{res}}\|_2^2$. Using Lemma 5 in [1], we have

$$\begin{aligned}\Pr[\|Y_{L+1,i}^{\text{res}}\|_2^2 &\leq \alpha_0 \mathbb{E}\|Y_{L+1,i}^{\text{res}}\|_2^2] \\ &\leq \frac{(1+\epsilon)\mathbb{E}\|Y_{L+1,i}^{\text{res}}\|_2^2 - \mathbb{E}\|Y_{L+1,i}^{\text{res}}\|_2^2}{(1+\epsilon-\alpha_0)\mathbb{E}\|Y_{L+1,i}^{\text{res}}\|_2^2} \\ &= \frac{\epsilon}{1+\epsilon-\alpha_0}\end{aligned}$$

which equals

$$\Pr[\|Y_{L+1,i}^{\text{res}}\|_2^2 \geq \alpha_0 \mathbb{E}\|Y_{L+1,i}^{\text{res}}\|_2^2] \geq 1 - \frac{\epsilon}{1+\epsilon-\alpha_0}$$

Using union bound, we have with probability $0.99 - \delta(\epsilon) - \frac{\epsilon}{1+\epsilon-\alpha_0}$

$$\begin{aligned}|\frac{\partial \mathcal{L}(Y_{\text{Final},i}^{\text{res}})}{\partial W_{pq}^{2,L}}|^2 &= O(\big[\|\mathbf{J}_{LN}(Y_{L+1,i}^{\text{res}})\|_2^2 \|\mathbf{J}_{LN}(Y_{L,i}^{\text{res},6})\|_2^2 |[\sigma(Y_{L,i}^{\text{res},5}W^{1,L})]_p|^2\big]) \\ &\leq O(\frac{2d^2\ln 100d}{\|Y_{L+1,i}^{\text{res}}\|_2^2 \|Y_{L,i}^{\text{res},6}\|_2^2}) \leq O(\frac{d^2\ln d}{\alpha_0 \mathbb{E}\|Y_{L+1,i}^{\text{res}}\|_2^2 \mathbb{E}\|Y_{L,i}^{\text{res},6}\|_2^2}) \leq O(\frac{\ln d}{\alpha_0 L d})\end{aligned}$$

So we have

$$|\frac{\partial \tilde{\mathcal{L}}}{\partial W_{pq}^{2,L}}|^2 = |\frac{1}{n}\sum_{i=1}^{n}\frac{\partial \mathcal{L}(x_{\text{Final},i}^{\text{res}})}{\partial W_{pq}^{2,L}}|^2 = O(\frac{\ln d}{\alpha_0 L d})$$

Thus

$$\|\frac{\partial \tilde{\mathcal{L}}}{\partial W^{2,L}}\|_F = \sqrt{\sum_{p,q=1}^{d} |\frac{\partial \tilde{\mathcal{L}}}{\partial W_{pq}^{2,L}}|^2} \leq O(\sqrt{\frac{d \ln d}{\alpha_0 L}})$$

We can have the following result with probability at least $0.99 - \delta(\epsilon) - \frac{\epsilon}{0.9+\epsilon}$ taking $\alpha_0 = \frac{1}{10}$, for the Res-Post-LN transformer

$$\|\frac{\partial \tilde{\mathcal{L}}}{\partial W^{2,L}}\|_F \leq O(\sqrt{\frac{d \ln d}{L}})$$

Since d and d_m are identical, we have

$$\|\frac{\partial \tilde{\mathcal{L}}}{\partial W^{2,L}}\|_F \leq O(\sqrt{\frac{d_m \ln d_m}{L}})$$

The above inequality corresponds to Theorem 4 in the main paper.

References

1. Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., Liu, T.: On layer normalization in the transformer architecture. In: International Conference on Machine Learning. pp. 10524–10533. PMLR (2020)