Supplementary material: On Layer Norm in the Transformer Architecture

Ruibin Xiong[†*1 2] **Yunchang Yang**[*3] **Di He**[4 5] **Kai Zheng**[4] **Shuxin Zheng**[5] **Chen Xing**[6] **Huishuai Zhang**[5] **Yanyan Lan**[1 2] **Liwei Wang**[4 3] **Tie-Yan Liu**[5]

1. Experimental Settings

1.1. Machine Translation

Experiment on Section 3 The training/validation/test sets of the IWSLT14 German-to-English (De-En) task contain about 153K/7K/7K sentence pairs, respectively. We use a vocabulary of 10K tokens based on a joint source and target byte pair encoding (BPE) (Sennrich et al., 2015). All of our experiments use a Transformer architecture with a 6-layer encoder and 6-layer decoder. The size of embedding is set to 512, the size of hidden nodes in attention sub-layer and position-wise feed-forward network sub-layer are set to 512 and 1024, and the number of heads is set to 4. Label smoothed cross entropy is used as the objective function by setting $\epsilon = 0.1$ (Szegedy et al., 2016), and we apply dropout with a ratio 0.1. The batch size is set to be 4096 tokens. When we decode translation results from the model during inference, we set beam size as 5 and the length penalty as 1.2.

Experiment on Section 4 The configuration of IWLST14 De-En task is the same as in Section 3. For the WMT14 En-De task, we replicate the setup of (Vaswani et al., 2017), which consists of about 4.5M training parallel sentence pairs, and uses a 37K vocabulary based on a joint source and target BPE. Newstest2013 is used as the validation set, and Newstest2014 is used as the test set. One of the basic configurations of the Transformer architecture is the `base` setting, which consists of a 6-layer encoder and 6-layer decoder. The size of the hidden nodes and embeddings are set to 512. The number of heads is 8. Label smoothed cross entropy is used as the objective function by setting $\epsilon = 0.1$. The batch size is set to be 8192 tokens per GPU on 16 NVIDIA Tesla P40 GPUs.

1.2. Unsupervised Pretraining

We follow Devlin et al. (2018) to use English Wikipedia corpus and BookCorpus for the pre-training. As the dataset BookCorpus (Zhu et al., 2015) is no longer freely distributed. We follow the suggestions from Devlin et al. (2018) to crawl and collect BookCorpus[1] on our own. The concatenation of two datasets includes roughly 3.4B words in total, which is comparable with the data corpus used in Devlin et al. (2018). We first segment documents into sentences with Spacy[2]; Then, we normalize, lower-case, and tokenize texts using Moses (Koehn et al., 2007) and apply BPE(Sennrich et al., 2016). We randomly split documents into one training set and one validation set. The training-validation ratio for pre-training is 199:1. All experiments are conducted on 32 NVIDIA Tesla P40 GPUs.

The base model in Devlin et al. (2018) consists of 12 Transformer layers. The size of hidden nodes and embeddings are set to 768, and the number of heads is set to 12.

1.3. GLUE Dataset

MRPC The Microsoft Research Paraphrase Corpus (Dolan & Brockett, 2005) is a corpus of sentence pairs automatically extracted from online news sources, with human annotations for whether the sentences in the pair are semantically equivalent, and the task is to predict the equivalence. The performance is evaluated by the accuracy.

RTE The Recognizing Textual Entailment (RTE) datasets come from a series of annual textual entailment challenges (Bentivogli et al., 2009). The task is to predict whether sentences in a sentence pair are entailment. The performance is evaluated by the accuracy.

Fine-tuning on GLUE tasks We use the validation set for evaluation. To fine-tune the models, following Devlin et al.

[*]Equal contribution [†]Works done while interning at Microsoft Research Asia [1]CAS Key Laboratory of Network Data Science and Technology, Institute of Computing Technolog, Chinese Academy of Sciences [2]University of Chinese Academy of Sciences [3]Center for Data Science, Peking University, Beijing Institute of Big Data Research [4]Key Laboratory of Machine Perception, MOE, School of EECS, Peking University [5]Microsoft Research [6]College of Computer Science, Nankai University. Correspondence to: Shuxin Zheng <shuxin.zheng@microsoft.com>, Di He <dihe@microsoft.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

[1]https://www.smashwords.com

[2]https://spacy.io

(2018); Liu et al. (2019), we search the optimization hyper-parameters in a search space including different batch sizes (16/32), learning rates ($1e^{-5}$ - $1e^{-4}$) and number of epochs (3-8). We find that the validation accuracy are sensitive to random seeds, so we repeat fine-tuning on each task for 6 times using different random seeds and compute the 95% confidence interval of validation accuracy.

2. Proof of Lemma 1

Proof. Denote $X = (X_1, X_2, ..., X_d)$ in which X_i are i.i.d. Gaussian random variables with distribution $N(0, \sigma^2)$. Denote $\rho_X(x)$ as the probability density function of X_1. Then $\mathbb{E}(\|\text{ReLU}(X)\|_2^2) = \sum_{i=1}^d \mathbb{E}[\text{ReLU}(X_i)^2] = \sum_{i=1}^d \mathbb{E}[\text{ReLU}(X_i)^2|X_i \geq 0]\mathbb{P}(X_i \geq 0) = \frac{d}{2}\mathbb{E}[\text{ReLU}(X_1)^2|X_1 \geq 0] = \frac{d}{2}\mathbb{E}[X_1^2|X_1 \geq 0] = \frac{d}{2}\int_{-\infty}^{+\infty} x^2 \rho_{X|X>0}(x)dx = \frac{d}{2}\int_0^{+\infty} x^2 2\rho_X(x)dx = \frac{1}{2}\sigma^2 d.$ □

3. Proof of Lemma 2

Proof. At initialization, the layer normalization is computed as $\text{LN}(v) = \frac{v-\mu}{\sigma}$. It is easy to see that layer normalization at initialization projects any vector v onto the $d-1$-sphere of radius $\sqrt{d}$ since $\|\text{LN}(v)\|_2^2 = \|\frac{v-\mu}{\sigma}\|_2^2 = \frac{\sum_{k=1}^d (v_k-\mu)^2}{\sigma^2} = d$.

We first estimate the expected l_2 norm of each intermediate output $x_{l,i}^{post,1}, \cdots, x_{l,i}^{post,5}$ for $l > 0$. Using Xavier initialization, the elements in $W^{V,l}$ are i.i.d. Gaussian random variables sampled from $N(0, 1/d)$. Since $\|x_{l,i}^{post}\|_2^2 = d$ by the definition of Layer Normalization when $l > 0$, we have

$$\begin{aligned}
\mathbb{E}(\|x_{l,i}^{post,2}\|_2^2) =& \mathbb{E}(\|x_{l,i}^{post}\|_2^2) + \mathbb{E}(\|x_{l,i}^{post,1}\|_2^2) \\
&+ 2\mathbb{E}(x_{l,i}^{post,1} {x_{l,i}^{post}}^\top) && (1)\\
=& \mathbb{E}(\|x_{l,i}^{post}\|_2^2) + \mathbb{E}(\|x_{l,i}^{post,1}\|_2^2) \\
&+ \frac{2}{n}\mathbb{E}(\sum_{j=1}^n x_{l,j}^{post} W^{V,l} {x_{l,i}^{post}}^\top) && (2)\\
=& \mathbb{E}(\|x_{l,i}^{post}\|_2^2) + \mathbb{E}(\|x_{l,i}^{post,1}\|_2^2) && (3)\\
=& \mathbb{E}(\|x_{l,i}^{post}\|_2^2) + \mathbb{E}(\|\frac{1}{n}\sum_{i=1}^n x_{l,i}^{post}\|_2^2) && (4)\\
\leq& 2d && (5)
\end{aligned}$$

and $\mathbb{E}(\|x_{l,i}^{post,2}\|_2^2) = \mathbb{E}(\|x_{l,i}^{post}\|_2^2) + \mathbb{E}(\|x_{l,i}^{post,1}\|_2^2) = \mathbb{E}(\|x_{l,i}^{post}\|_2^2) + \mathbb{E}(\|\frac{1}{n}\sum_{i=1}^n x_{l,i}^{post}\|_2^2) \geq \mathbb{E}(\|x_{l,i}^{post}\|_2^2) = d$.

Similarly, we have $\|x_{l,i}^{post,3}\|_2^2 = d$ by the definition of Layer Normalization. Again, for the ReLU activation function, the elements in $W^{1,l}$ and $W^{2,l}$ are i.i.d. Gaussian random variables sampled from $N(0, 1/d)$. According to Lemma 1, we have

$$\begin{aligned}
\mathbb{E}(\|x_{l,i}^{post,4}\|_2^2) =& \mathbb{E}(\|\text{ReLU}(x_{l,i}^{post,3}W^{1,l})W^{2,l}\|_2^2) && (6)\\
=& \mathbb{E}(\mathbb{E}(\mathbb{E}(\|\text{ReLU}(x_{l,i}^{post,3}W^{1,l})W^{2,l}\|_2^2 \\
&|x_{l,i}^{post,3}, W^{1,l})|x_{l,i}^{post,3})) && (7)\\
=& \mathbb{E}(\mathbb{E}(\|\text{ReLU}(x_{l,i}^{post,3}W^{1,l})\|_2^2|x_{l,i}^{post,3})) && (8)\\
=& \mathbb{E}(\frac{1}{2}\|x_{l,i}^{post,3}\|_2^2) = \frac{d}{2} && (9)
\end{aligned}$$

Based on this, we can estimate the scale of $\mathbb{E}(\|x_{l,i}^{post,5}\|_2^2)$ as follows.

$$\begin{aligned}
\mathbb{E}(\|x_{l,i}^{post,5}\|_2^2) =& \mathbb{E}(\|x_{l,i}^{post,3}\|_2^2) + \mathbb{E}(\|x_{l,i}^{post,4}\|_2^2) \\
&+ 2\mathbb{E}(x_{l,i}^{post,3} {x_{l,i}^{post,4}}^\top) && (10)\\
=& \mathbb{E}(\|x_{l,i}^{post,3}\|_2^2) + \mathbb{E}(\|x_{l,i}^{post,4}\|_2^2) \\
&+ \frac{2}{n}\mathbb{E}(\sum_{j=1}^n \text{ReLU}(x_{l,j}^{post,3}W^{1,l})W^{2,l} {x_{l,i}^{post,3}}^\top) && (11)\\
=& \mathbb{E}(\|x_{l,i}^{post,3}\|_2^2) + \mathbb{E}(\|x_{l,i}^{post,4}\|_2^2) = d + \frac{d}{2} = \frac{3}{2}d && (12)
\end{aligned}$$

Using similar technique we can bound $\mathbb{E}(\|x_{l,i}^{pre}\|_2^2)$ for the Pre-LN Transformer.

$$\begin{aligned}
\mathbb{E}(\|x_{l,i}^{pre,3}\|_2^2) =& \mathbb{E}(\|x_{l,i}^{pre}\|_2^2) + \mathbb{E}(\|x_{l,i}^{pre,2}\|_2^2) \\
&+ 2\mathbb{E}(x_{l,i}^{pre,2} {x_{l,i}^{pre}}^\top) && (13)\\
=& \mathbb{E}(\|x_{l,i}^{pre}\|_2^2) + \mathbb{E}(\|x_{l,i}^{pre,2}\|_2^2) \\
&+ \frac{2}{n}\mathbb{E}(\sum_{j=1}^n x_{l,j}^{pre,1} W^{V,l} {x_{l,i}^{pre}}^\top) && (14)\\
=& \mathbb{E}(\|x_{l,i}^{pre}\|_2^2) + \mathbb{E}(\|x_{l,i}^{pre,2}\|_2^2) && (15)\\
=& \mathbb{E}(\|x_{l,i}^{pre}\|_2^2) + \mathbb{E}(\|\frac{1}{n}\sum_{i=1}^n x_{l,i}^{pre,1}\|_2^2) && (16)
\end{aligned}$$

It is easy to see that we have $\mathbb{E}(\|x_{l,i}^{pre}\|_2^2) \leq \mathbb{E}(\|x_{l,i}^{pre,3}\|_2^2) \leq \mathbb{E}(\|x_{l,i}^{pre}\|_2^2) + d$. And similar to (10)-(12),

$$\begin{aligned}
\mathbb{E}(\|x_{l+1,i}^{pre}\|_2^2) =& \mathbb{E}(\|x_{l,i}^{pre,3}\|_2^2) + \mathbb{E}(\|x_{l,i}^{pre,5}\|_2^2) \\
&+ 2\mathbb{E}(x_{l,i}^{pre,3} {x_{l,i}^{pre,5}}^\top) && (17)\\
=& \mathbb{E}(\|x_{l,i}^{pre,3}\|_2^2) + \mathbb{E}(\|x_{l,i}^{pre,5}\|_2^2) && (18)\\
=& \mathbb{E}(\|x_{l,i}^{pre,3}\|_2^2) + \frac{1}{2}d && (19)
\end{aligned}$$

Combining both, we have $\mathbb{E}(\|x_{l,i}^{pre}\|_2^2) + \frac{1}{2}d \leq \mathbb{E}(\|x_{l+1,i}^{pre}\|_2^2) \leq \mathbb{E}(\|x_{l,i}^{pre}\|_2^2) + \frac{3}{2}d$. Then we have $(1 + \frac{l}{2})d \leq \mathbb{E}(\|x_{l,i}^{pre}\|_2^2) \leq (1 + \frac{3l}{2})d$ by induction.

□

4. Proof of Lemma 3

The proof of Lemma 3 is based on Lemma 4.1:

Lemma 4. 1. *Let $\alpha \in \mathbb{R}^d$ be a vector such that $\|\alpha\|_2 = 1$, then the eigenvalue of $I - \alpha^\top \alpha$ is either 1 or 0.*

Proof. Let $\{e_1, ..., e_d\}$ be unit vectors such that $e_1 = \alpha$ and $e_i \perp e_j$ for all (i, j). Then we have $e_1(I - \alpha^\top \alpha) = e_1 - e_1\alpha^\top \alpha = e_1 - \alpha = 0$ and $e_i(I - \alpha^\top \alpha) = e_i - e_i\alpha^\top \alpha = e_i$ for $i \neq 1$. So e_i are all the eigenvectors of $I - \alpha^\top \alpha$, and their corresponding eigenvalues are $(0, 1, 1, ..., 1)$. Hence we complete our proof. □

Proof of Lemma 3. Denote $y = x(I - \frac{1}{d}\mathbf{1}^\top \mathbf{1})$, where $\mathbf{1} = (1, 1, ..., 1) \in \mathbb{R}^d$, then the layer normalization can be rewritten as

$$\mathrm{LN}(x)_i = \frac{y_i}{\sqrt{\frac{1}{d}\sum_{j=1}^d y_j^2}} \tag{20}$$

We explicitly calculate the Jacobian of layer normalization as

$$\frac{\partial \mathrm{LN}(x)_i}{\partial y_j} = \frac{\partial}{\partial y_j}\left(\frac{y_i}{\sqrt{\frac{1}{d}\sum_{k=1}^n y_k^2}}\right) \tag{21}$$

$$= \frac{\delta_{ij}\sqrt{\frac{1}{d}\sum_{k=1}^n y_k^2} - y_i \frac{\frac{1}{d} y_j}{\sqrt{\frac{1}{d}\sum_{k=1}^n y_k^2}}}{\frac{1}{d}\sum_{k=1}^n y_k^2} \tag{22}$$

$$= \sqrt{d}\frac{\delta_{ij}\|y\|_2^2 - y_i y_j}{\|y\|_2^{\frac{3}{2}}} = \frac{\sqrt{d}}{\|y\|_2}\left(\delta_{ij} - \frac{y_i y_j}{\|y\|_2^2}\right) \tag{23}$$

where $\delta_{ij} = 1$ when $i = j$ and $\delta_{ij} = 0$ when $i \neq j$. In the matrix form,

$$\frac{\partial \mathrm{LN}(x)}{\partial y} = \frac{\sqrt{d}}{\|y\|_2}\left(I - \frac{y^\top y}{\|y\|_2^2}\right) \tag{24}$$

and

$$\mathbf{J}_{LN}(x) = \frac{\partial \mathrm{LN}(x)}{\partial x} \tag{25}$$

$$= \frac{\partial \mathrm{LN}(x)}{\partial y}\frac{\partial y}{\partial x} \tag{26}$$

$$= \sqrt{d}\frac{1}{\|y\|_2}\left(I - \frac{y^\top y}{\|y\|_2^2}\right)\left(I - \frac{1}{d}\mathbf{1}^\top \mathbf{1}\right). \tag{27}$$

Since the eigenvalue of the matrix $(I - \frac{y^\top y}{\|y\|_2^2})$ and $(I - \frac{1}{d}\mathbf{1}^\top \mathbf{1})$ are either 1 or 0 (by Lemma 4.1), we have $\|(I - \frac{y^\top y}{\|y\|_2^2})\|_2 = \mathcal{O}(1)$ and $\|(I - \frac{1}{d}\mathbf{1}^\top \mathbf{1})\|_2 = \mathcal{O}(1)$. So the spectral norm of $\mathbf{J}_{LN}(x)$ is

$$\|\mathbf{J}_{LN}(x)\|_2 = \mathcal{O}\left(\frac{\sqrt{d}}{\|y\|_2}\right) = \mathcal{O}\left(\frac{\sqrt{d}}{\|x\|_2}\right) \tag{28}$$

□

5. Proof of Theorem 1

The proof of Theorem 1 is based on Lemma 4.2:

Lemma 4. 2. *Let Y be a random variable that is never larger than B. Then for all $a < B$,*

$$\Pr[Y \leq a] \leq \frac{\mathbb{E}[B - Y]}{B - a} \tag{29}$$

Proof. Let $X = B - Y$, then $X \geq 0$ and Markov's inequality tells us that

$$\Pr[X \geq B - a] \leq \frac{\mathbb{E}[X]}{B - a} \tag{30}$$

Hence

$$\Pr[Y \leq a] \leq \frac{\mathbb{E}[B - Y]}{B - a} \tag{31}$$

□

Proof of Theorem 1. We prove Theorem 1 by estimating each element of the gradient matrix. Namely, we will analyze $\frac{\partial \tilde{\mathcal{L}}}{\partial W_{pq}^{2,L}}$ for $p, q \in \{1, ..., d\}$. The loss of the post-LN Transformer can be written as

$$\tilde{\mathcal{L}}(x_{L+1,1}^{post}, ..., x_{L+1,n}^{post}) = \frac{1}{n}\sum_{i=1}^n \mathcal{L}(x_{L+1,i}^{post}) \tag{32}$$

Through back propagation, for each $i \in \{1, 2, ..., n\}$ the gradient of $\mathcal{L}(x_{L+1,i})$ with respect to the last layer's parameter $W^{2,L}$ in the post-LN setting can be written as:

$$\frac{\partial \mathcal{L}(x_{L+1,i}^{post})}{\partial W_{pq}^{2,L}} = \frac{\partial \mathcal{L}(x_{L+1,i}^{post})}{\partial x_{L+1,i}^{post}}\frac{\partial x_{L+1,i}^{post}}{\partial x_{L,i}^{post,5}}\frac{\partial x_{L,i}^{post,5}}{\partial x_{L,i}^{post,4}}\frac{\partial x_{L,i}^{post,4}}{\partial W_{pq}^{2,L}} \tag{33}$$

$$= \frac{\partial \mathcal{L}(x_{L+1,i}^{post})}{\partial x_{L+1,i}^{post}}\mathbf{J}_{LN}(x_{L,i}^{post,5})\frac{\partial x_{L,i}^{post,4}}{\partial W_{pq}^{2,L}} \tag{34}$$

$$= \frac{\partial \mathcal{L}(x_{L+1,i}^{post})}{\partial x_{L+1,i}^{post}}\mathbf{J}_{LN}(x_{L,i}^{post,5})(0, 0, ..., [\mathrm{ReLU}(x_{L,i}^{post,3}W^{1,L})]_p, ..., 0)^\top \tag{35}$$

Here $[\text{ReLU}(x_{L,i}^{post,3}W^{1,L})]_p$ means the p-th element of $\text{ReLU}(x_{L,i}^{post,3}W^{1,L})$. So the absolute value of $\frac{\partial \mathcal{L}(x_{L+1,i}^{post})}{\partial W_{pq}^{2,L}}$ can be bounded by

$$|\frac{\partial \mathcal{L}(x_{L+1,i}^{post})}{\partial W_{pq}^{2,L}}|\leq \|\frac{\partial \mathcal{L}(x_{L+1,i}^{post})}{\partial x_{L+1,i}^{post}}\|_2 \|\mathbf{J}_{LN}(x_{L,i}^{post,5})\|_2 \|(0,0,...,[\text{ReLU}(x_{L,i}^{post,3}W^{1,L})]_p,...,0)^\top\|_2 \tag{36}$$

$$=\|\frac{\partial \mathcal{L}(x_{L+1,i}^{post})}{\partial x_{L+1,i}^{post}}\|_2 \|\mathbf{J}_{LN}(x_{L,i}^{post,5})\|_2 |[\text{ReLU}(x_{L,i}^{post,3}W^{1,L})]_p| \tag{37}$$

which implies

$$|\frac{\partial \mathcal{L}(x_{L+1,i}^{post})}{\partial W_{pq}^{2,L}}|^2\leq \|\frac{\partial \mathcal{L}(x_{L+1,i}^{post})}{\partial x_{L+1,i}^{post}}\|_2^2 \|\mathbf{J}_{LN}(x_{L,i}^{post,5})\|_2^2 |[\text{ReLU}(x_{L,i}^{post,3}W^{1,L})]_p|^2 \tag{38}$$

Since all the derivatives are bounded, we have $\|\frac{\partial \mathcal{L}(x_{L+1,i}^{post})}{\partial x_{L+1,i}^{post}}\|_2^2= \mathcal{O}(1)$. So

$$|\frac{\partial \mathcal{L}(x_{L+1,i}^{post})}{\partial W_{pq}^{2,L}}|^2 =\mathcal{O}(\left[\|\mathbf{J}_{LN}(x_{L,i}^{post,5})\|_2^2 |[\text{ReLU}(x_{L,i}^{post,3}W^{1,L})]_p|^2\right]) \tag{39}$$

Since $\|x_{L,i}^{post,3}\|_2^2= d$, $[x_{L,i}^{post,3}W^{1,L}]_p$ has distribution $N(0,1)$, using Chernoff bound we have

$$\Pr[|[x_{L,i}^{post,3}W^{1,L}]_p|\geq a_0]\leq \exp(-\frac{a_0^2}{2}).$$

So

$$\Pr[\text{ReLU}([x_{L,i}^{post,3}W^{1,L}]_p)^2\geq 2\ln 100d]\leq \frac{0.01}{d}.$$

Thus with probability at least 0.99, for all $p=1,2,...,d$ we have $\text{ReLU}([x_{L,i}^{post,3}W^{1,L}]_p)^2\leq 2\ln 100d$.

Since with probability $1-\delta(\epsilon)$, $\frac{|\|x_{L,i}^{post,5}\|_2^2-\mathbb{E}\|x_{L,i}^{post,5}\|_2^2|}{\mathbb{E}\|x_{L,i}^{post,5}\|_2^2}\leq\epsilon$, we have $\|x_{L,i}^{post,5}\|_2^2\leq (1+\epsilon)\mathbb{E}\|x_{L,i}^{post,5}\|_2^2$. Using Lemma 4.2, we have

$$\Pr[\|x_{L,i}^{post,5}\|_2^2\leq\alpha_0\mathbb{E}\|x_{L,i}^{post,5}\|_2^2] \tag{40}$$

$$\leq\frac{(1+\epsilon)\mathbb{E}\|x_{L,i}^{post,5}\|_2^2-\mathbb{E}\|x_{L,i}^{post,5}\|_2^2}{(1+\epsilon-\alpha_0)\mathbb{E}\|x_{L,i}^{post,5}\|_2^2} \tag{41}$$

$$=\frac{\epsilon}{1+\epsilon-\alpha_0} \tag{42}$$

for an arbitrary constant $\alpha_0>0$, which equals

$$\Pr[\|x_{L,i}^{post,5}\|_2^2\geq \alpha_0\mathbb{E}\|x_{L,i}^{post,5}\|_2^2]\geq 1-\frac{\epsilon}{1+\epsilon-\alpha_0} \tag{43}$$

So according to union bound, with probability at least $0.99-\delta(\epsilon)-\frac{\epsilon}{1+\epsilon-\alpha_0}$ we have $|\frac{\partial \mathcal{L}(x_{L+1,i}^{post})}{\partial W_{pq}^{2,L}}|^2= \mathcal{O}(\left[\|\mathbf{J}_{LN}(x_{L,i}^{post,5})\|_2^2|[\text{ReLU}(x_{L,i}^{post,3}W^{1,L})]_p|^2\right])\leq \mathcal{O}(\frac{2d\ln 100d}{\|x_{L,i}^{post,5}\|_2^2})\leq \mathcal{O}(\frac{d\ln d}{\alpha_0\mathbb{E}\|x_{L,i}^{post,5}\|_2^2})=\mathcal{O}(\frac{\ln d}{\alpha_0})$. So we have

$$|\frac{\partial \tilde{\mathcal{L}}}{\partial W_{pq}^{2,L}}|^2=|\frac{1}{n}\sum_{i=1}^n\frac{\partial \mathcal{L}(x_{L+1,i}^{post})}{\partial W_{pq}^{2,L}}|^2 \tag{44}$$

$$\leq\frac{1}{n}\sum_{i=1}^n|\frac{\partial \mathcal{L}(x_{L+1,i}^{post})}{\partial W_{pq}^{2,L}}|^2=\mathcal{O}(\frac{\ln d}{\alpha_0}) \tag{45}$$

and

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

.

The loss of the pre-LN Transformer can be written as

$$\tilde{\mathcal{L}}(x_{Final,1}^{pre},...,x_{Final,n}^{pre})=\frac{1}{n}\sum_{i=1}^n\mathcal{L}(x_{Final,i}^{pre}) \tag{46}$$

Using the same technique, in the pre-LN setting the gradient of $\mathcal{L}(x_{Final,i}^{pre})$ with respect to the last layer's parameter $W^{2,L}$ can be written as

$$\frac{\partial \mathcal{L}(x_{Final,i}^{pre})}{\partial W_{pq}^{2,L}}=\frac{\partial \mathcal{L}(x_{Final,i}^{pre})}{\partial x_{Final,i}^{pre}}\frac{\partial x_{Final,i}^{pre}}{\partial x_{L+1,i}^{pre}}\frac{\partial x_{L+1,i}^{pre}}{\partial x_{L,i}^{pre,5}}\frac{\partial x_{L,i}^{pre,5}}{\partial W_{pq}^{2,L}} \tag{47}$$

$$=\frac{\partial \mathcal{L}(x_{Final,i}^{pre})}{\partial x_{Final,i}^{pre}}\mathbf{J}_{LN}(x_{L+1,i}^{pre})(0,0,...,[\text{ReLU}(x_{L,i}^{pre,4}W^{1,L})]_p,...,0)^\top \tag{48}$$

So the absolute value of each component of the gradient is bounded by

$$|\frac{\partial \mathcal{L}(x_{Final,i}^{pre})}{\partial W_{pq}^{2,L}}|\leq\|\frac{\partial \mathcal{L}(x_{Final,i}^{pre})}{\partial x_{Final,i}^{pre}}\|_2\|\mathbf{J}_{LN}(x_{L+1,i}^{pre})\|_2 \|(0,0,...,[\text{ReLU}(x_{L,i}^{pre,4}W^{1,L})]_p,...,0)\|_2 \tag{49}$$

$$=\|\frac{\partial \mathcal{L}(x_{Final,i}^{pre})}{\partial x_{Final,i}^{pre}}\|_2\|\mathbf{J}_{LN}(x_{L+1,i}^{pre})\|_2 |[\text{ReLU}(x_{L,i}^{pre,4}W^{1,L})]_p| \tag{50}$$

Since $\|x_{L,i}^{pre,4}\|_2^2= d$ and $[x_{L,i}^{pre,4}W^{1,L}]_p$ obeys distribution $N(0,1)$, using Chernoff bound we have

$$\Pr[|[x_{L,i}^{pre,4}W^{1,L}]_p|\geq a_0]\leq\exp(-\frac{a_0^2}{2}).$$

So

$$\Pr[\text{ReLU}([x_{L,i}^{pre,4}W^{1,L}]_p)^2 \geq 2\ln 100d] \leq \frac{0.01}{d}.$$

So with probability at least 0.99, for all $p = 1, 2, ..., d$ we have $\text{ReLU}([x_{L,i}^{pre,4}W^{1,L}]_p)^2 \leq 2\ln 100d$.

Since with probability $1-\delta(\epsilon)$, $\frac{|\|x_{L+1,i}^{pre}\|_2^2 - \mathbb{E}\|x_{L+1,i}^{pre}\|_2^2|}{\mathbb{E}\|x_{L+1,i}^{pre}\|_2^2} \leq \epsilon$, we have $\|x_{L+1,i}^{pre}\|_2^2 \leq (1+\epsilon)\mathbb{E}\|x_{L+1,i}^{pre}\|_2^2$. Using Lemma 5, we have

$$\Pr[\|x_{L+1,i}^{pre}\|_2^2 \leq \alpha_0 \mathbb{E}\|x_{L+1,i}^{pre}\|_2^2] \tag{51}$$
$$\leq \frac{(1+\epsilon)\mathbb{E}\|x_{L+1,i}^{pre}\|_2^2 - \mathbb{E}\|x_{L+1,i}^{pre}\|_2^2}{(1+\epsilon-\alpha_0)\mathbb{E}\|x_{L+1,i}^{pre}\|_2^2} \tag{52}$$
$$= \frac{\epsilon}{1+\epsilon-\alpha_0} \tag{53}$$

which equals

$$\Pr[\|x_{L+1,i}^{pre}\|_2^2 \geq \alpha_0 \mathbb{E}\|x_{L+1,i}^{pre}\|_2^2] \geq 1 - \frac{\epsilon}{1+\epsilon-\alpha_0} \tag{54}$$

According to union bound, with probability $0.99 - \delta(\epsilon) - \frac{\epsilon}{1+\epsilon-\alpha_0}$ we have $|\frac{\partial \mathcal{L}(x_{Final,i}^{pre})}{\partial W_{pq}^{2,L}}|^2 = \mathcal{O}(\left[\|\mathbf{J}_{LN}(x_{L+1,i}^{pre})\|_2^2 |[\text{ReLU}(x_{L,i}^{pre,4}W^{1,L})]_p|^2\right]) \leq \mathcal{O}(\frac{2d\ln 100d}{\|x_{L+1,i}^{pre}\|_2^2}) \leq \mathcal{O}(\frac{d\ln d}{\alpha_0 \mathbb{E}\|x_{L+1,i}^{pre}\|_2^2}) = \mathcal{O}(\frac{\ln d}{\alpha_0 L})$. So we have

$$|\frac{\partial \tilde{\mathcal{L}}}{\partial W_{pq}^{2,L}}|^2 = |\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \mathcal{L}(x_{Final,i}^{pre})}{\partial W_{pq}^{2,L}}|^2 = \mathcal{O}(\frac{\ln d}{\alpha_0 L}) \tag{55}$$

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

Take $\alpha_0 = \frac{1}{10}$, we have that with probability at least $0.99 - \delta(\epsilon) - \frac{\epsilon}{0.9+\epsilon}$, for the Post-LN Transformer we have $\|\frac{\partial \tilde{\mathcal{L}}}{\partial W^{2,L}}\|_F \leq \mathcal{O}(d\sqrt{\ln d})$ and for the Pre-LN Transformer we have $\|\frac{\partial \tilde{\mathcal{L}}}{\partial W^{2,L}}\|_F \leq \mathcal{O}(d\sqrt{\frac{\ln d}{L}})$ □

6. Extension to other layers

For simplicity, we denote $x_l = \text{Concat}(x_{l,1}, ..., x_{l,n}) \in \mathbb{R}^{nd}$ and $x_l^k = \text{Concat}(x_{l,1}^k, ..., x_{l,n}^k) \in \mathbb{R}^{nd}$ for $k = \{1, 2, 3, 4, 5\}$. Then in the Post-LN Transformer, the gradient of the parameters in the l-th layer (take $W^{2,l}$ as an example) can be written as

$$\frac{\partial \tilde{\mathcal{L}}}{\partial W^{2,l}} = \frac{\partial \tilde{\mathcal{L}}}{\partial x_{L+1}^{post}} \left(\prod_{j=l+1}^{L} \frac{\partial x_{j+1}^{post}}{\partial x_j^{post}}\right) \frac{\partial x_{l+1}^{post}}{\partial W^{2,l}},$$

where

$$\frac{\partial x_{j+1}^{post}}{\partial x_j^{post}} = \frac{\partial x_{j+1}^{post}}{\partial x_j^{post,5}} \frac{\partial x_j^{post,5}}{\partial x_j^{post,3}} \frac{\partial x_j^{post,3}}{\partial x_j^{post,2}} \frac{\partial x_j^{post,2}}{\partial x_j^{post}}.$$

The Jacobian matrices of the Post-LN Transformer layers are:

$$\frac{\partial x_{j+1}^{post}}{\partial x_j^{post,5}} = \begin{pmatrix} \mathbf{J}_{LN}(x_{j,1}^{post,5}) & & \\ & \ddots & \\ & & \mathbf{J}_{LN}(x_{j,n}^{post,5}) \end{pmatrix} \tag{56}$$

$$\frac{\partial x_j^{post,5}}{\partial x_j^{post,3}} = \begin{pmatrix} I & & \\ & \ddots & \\ & & I \end{pmatrix} + \begin{pmatrix} W^{2,j} & & \\ & \ddots & \\ & & W^{2,j} \end{pmatrix} \begin{pmatrix} \mathbf{J}_1^j & & \\ & \ddots & \\ & & \mathbf{J}_n^j \end{pmatrix} \begin{pmatrix} W^{1,l} & & \\ & \ddots & \\ & & W^{1,l} \end{pmatrix} \tag{57}$$

where

$$\mathbf{J}_i^j = \text{diag}\left(\sigma'\left(x_{j,i}^{post,3}\left(\mathbf{w}_1^{1,j}\right)^\top\right), ..., \sigma'\left(x_{j,i}^{post,3}\left(\mathbf{w}_d^{1,j}\right)^\top\right)\right) \in \mathbb{R}^{d\times d}$$

$$\frac{\partial x_j^{post,3}}{\partial x_j^{post,2}} = \begin{pmatrix} \mathbf{J}_{LN}(x_{j,1}^{post,2}) & & \\ & \ddots & \\ & & \mathbf{J}_{LN}(x_{j,n}^{post,2}) \end{pmatrix} \tag{58}$$

$$\frac{\partial x_j^{post,2}}{\partial x_j^{post}} = \begin{pmatrix} I & & \\ & \ddots & \\ & & I \end{pmatrix} + \begin{pmatrix} \frac{1}{n}W^{V,j} & \cdots & \frac{1}{n}W^{V,j} \\ \vdots & \ddots & \vdots \\ \frac{1}{n}W^{V,j} & \cdots & \frac{1}{n}W^{V,j} \end{pmatrix} \tag{59}$$

Using Hölder's inequality, we have

$$\mathbb{E}\|\frac{\partial x_{j+1}^{post}}{\partial x_j^{post}}\|_2 \leq \mathbb{E}\left[\|\frac{\partial x_{j+1}^{post}}{\partial x_j^{post,5}}\|_2 \|\frac{\partial x_j^{post,5}}{\partial x_j^{post,3}}\|_2 \|\frac{\partial x_j^{post,3}}{\partial x_j^{post,2}}\|_2 \|\frac{\partial x_j^{post,2}}{\partial x_j^{post}}\|_2\right] \tag{60}$$

$$\leq \sqrt{\mathbb{E}\left[\|\frac{\partial x_{j+1}}{\partial x_j^{post,5}}\|_2^2\right] \mathbb{E}\left[\|\frac{\partial x_j^{post,5}}{\partial x_j^{post,3}}\|_2^2 \|\frac{\partial x_j^{post,3}}{\partial x_j^{post,2}}\|_2^2 \|\frac{\partial x_j^{post,2}}{\partial x_j^{post}}\|_2^2\right]} \tag{61}$$

Since $\frac{\partial x_{j+1}}{\partial x_j^{post,5}} = diag(\mathbf{J}_{LN}(x_{j,1}^{post,5}), ..., \mathbf{J}_{LN}(x_{j,n}^{post,5}))$, we have $\sqrt{\mathbb{E}\left[\|\frac{\partial x_{j+1}^{post}}{\partial x_j^{post,5}}\|_2^2\right]} \approx \sqrt{\mathbb{E}\frac{d}{\|x_{j,1}^{post,5}\|_2^2}} \approx \sqrt{\frac{2}{3}}$ when

$\|x_{j,1}^{post,5}\|_2^2$ concentrates around its expectation $\mathbb{E}\|x_{j,1}^{post,5}\|_2^2$ which equals $\frac{3}{2}d$ according to Lemma 2. Therefore, when we estimate the norm of $\frac{\partial \tilde{\mathcal{L}}}{\partial W^{2,l}}$ for post-LN transformer, there exists a term $\mathcal{O}(\frac{2}{3}^{(L-l)/2})$, which exponentially decreases as l goes smaller. Similarly, in the pre-LN Transformer, the gradient can be written as

$$\frac{\partial \tilde{\mathcal{L}}}{\partial W^{2,l}} = \frac{\partial \tilde{\mathcal{L}}}{\partial x_{Final}^{pre}} \frac{\partial x_{Final}^{pre}}{\partial x_{L+1}^{pre}} \Big(\prod_{j=l+1}^{L} \frac{\partial x_{j+1}^{pre}}{\partial x_j^{pre}} \Big) \frac{\partial x_{l+1}^{pre}}{\partial W^{V,l}},$$

where

$$\frac{\partial x_{j+1}^{pre}}{\partial x_j^{pre}} = \frac{\partial x_{j+1}^{pre}}{\partial x_j^{pre,3}} \frac{\partial x_j^{pre,3}}{\partial x_j^{pre}}.$$

The Jacobian matrices of the Pre-LN Transformer layers are:

$$\frac{\partial x_{j+1}^{pre}}{\partial x_j^{pre,3}} = \begin{pmatrix} I & & \\ & \ddots & \\ & & I \end{pmatrix} + \begin{pmatrix} W^{2,j} & & \\ & \ddots & \\ & & W^{2,j} \end{pmatrix} \begin{pmatrix} \mathbf{J}_1^{(h')} & & \\ & \ddots & \\ & & \mathbf{J}_n^{(h')} \end{pmatrix} \begin{pmatrix} W^{1,j} & & \\ & \ddots & \\ & & W^{1,j} \end{pmatrix} \begin{pmatrix} \mathbf{J}_{LN}(x_{j,1}^{pre,3}) & & \\ & \ddots & \\ & & \mathbf{J}_{LN}(x_{j,n}^{pre,3}) \end{pmatrix} \quad (62)$$

$$\frac{\partial x_j^{pre,3}}{\partial x_j^{pre}} = \begin{pmatrix} I & & \\ & \ddots & \\ & & I \end{pmatrix} + \begin{pmatrix} \frac{1}{n}W^{V,j} & \dots & \frac{1}{n}W^{V,j} \\ \vdots & \ddots & \vdots \\ \frac{1}{n}W^{V,j} & \dots & \frac{1}{n}W^{V,j} \end{pmatrix} \begin{pmatrix} \mathbf{J}_{LN}(x_{j,1}^{pre}) & & \\ & \ddots & \\ & & \mathbf{J}_{LN}(x_{j,n}^{pre}) \end{pmatrix} \quad (63)$$

If l is sufficiently large, the norm of $\mathbf{J}_{LN}(x_{j,i}^{pre})$ and $\mathbf{J}_{LN}(x_{j,i}^{pre,3})$ are very small (of order $\mathcal{O}(\frac{1}{\sqrt{j}})$) as j is between $l+1$ and L, which means the eigenvalues of matrix $\frac{\partial x_{j+1}^{pre}}{\partial x_j^{pre,3}}$ and $\frac{\partial x_j^{pre,3}}{\partial x_j^{pre}}$ are close to 1. Then we can see that $\mathbb{E}\|\frac{\partial x_{j+1}^{pre}}{\partial x_j^{pre,3}}\|_2$ and $\mathbb{E}\|\frac{\partial x_j^{pre,3}}{\partial x_j^{pre}}\|_2$ are nearly 1, and the norm of $\frac{\partial \tilde{\mathcal{L}}}{\partial W^{2,l}}$ for pre-LN transformer is independent of l when l is large.

7. Examples of (ϵ, δ)-bounded random variables

In this section we give an example of (ϵ, δ)-bounded random variable. This example comes from Example 2.5 in (Wainwright, 2019) and we give a short description below.

If $Z = (Z_1, ..., Z_n)$ is a Gaussian vector with distribution $N(0, I_n)$, then $Y = \|Z\|_2^2 = \sum_{k=1}^n Z_k^2$ has distribution χ_n^2. And $\mathbb{E}Y = \sum_{k=1}^n \mathbb{E}Z_k^2 = n$

A random variable X with mean $\mu = \mathbb{E}[X]$ is called *sub-exponential* if there are non-negative parameters (ν, α) such that $\mathbb{E}[\exp(\lambda(X-\mu))] \leq \exp(\frac{\nu^2\lambda^2}{2})$ for all $|\lambda| < \frac{1}{\alpha}$. The next proposition comes from Proposition 2.2 in (Wainwright, 2019).

Proposition 1 (Sub-exponential tail bound)**.** *Suppose that X is sub-exponential with parameters (ν, α). Then*

$$\mathbb{P}[X - \mu \geq t] \leq \begin{cases} \exp(-\frac{t^2}{2\nu^2}) & \text{if } 0 \leq t \leq \frac{\nu^2}{\alpha}, \text{ and} \\ \exp(-\frac{t}{2\alpha}) & \text{for } t > \frac{\nu^2}{\alpha} \end{cases} \quad (64)$$

and from Example 2.5 in (Wainwright, 2019), the χ^2 variable Y is sub-exponential with parameters $(\nu, \alpha) = (2\sqrt{n}, 4)$. So we can derive the one-sided bound

$$\mathbb{P}\left[Y - n \geq n\epsilon\right] \leq \exp(-n\epsilon^2/8), \quad \text{for all } \epsilon \in (0, 1) \quad (65)$$

So Y is (ϵ, δ)-bounded with $\epsilon \in (0, 1)$ and $\delta = \exp(-n\epsilon^2/8)$.

8. Small learning rate experiment

Theoretically, we find that the gradients of the parameters near the output layers are very large for the Post-LN Transformer and suggest using large learning rates to those parameters makes the training unstable. To verify whether using small-step updates mitigates the issue, we use a very small but fixed learning rate and check whether it can optimize the Post-LN Transformer (without the learning rate warm-up step) to a certain extent. In detail, we use a fixed learning rate of $1e^{-4}$ at the beginning of the optimization, which is much smaller than the $\text{lr}_{max} = 1e^{-3}$ in the paper. Please note that as the learning rates during training are small, the training converges slowly, and this setting is not very practical in real large-scale tasks. We plot the validation curve together with other baseline approaches in Figure 6. We can see from the figure, the validation loss (pink curve) is around 4.3 in 27 epochs. This loss is much lower than that of the Post-LN Transformer trained using a large learning rate (blue curve). But it is still worse than the SOTA performance (green curve).

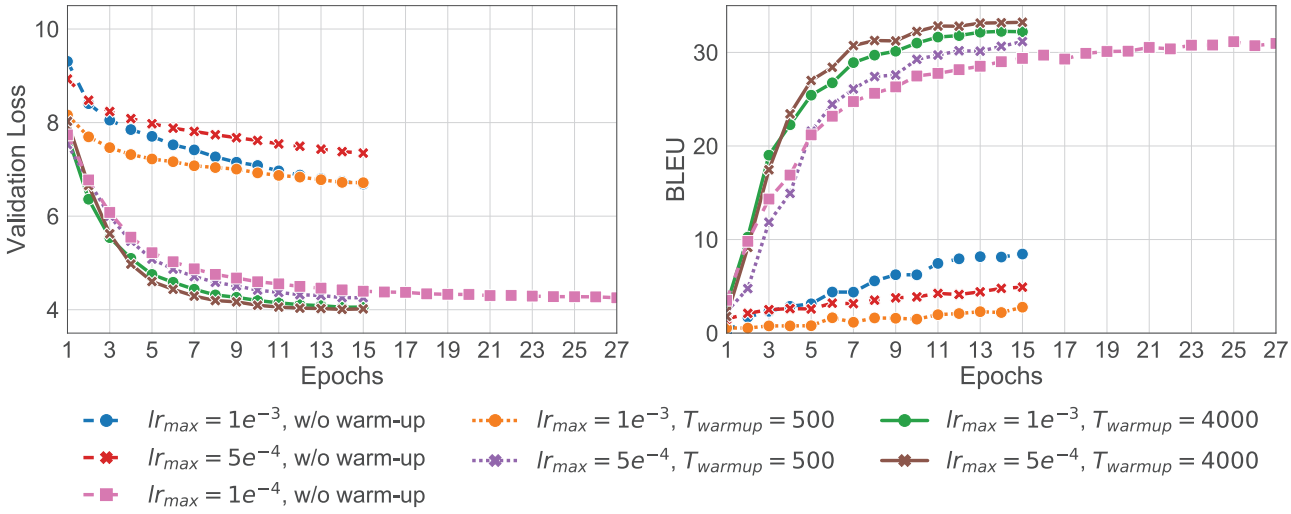

Figure 1. Performances of the models on the IWSLT14 De-En task.

References

Bentivogli, L., Dagan, I., Dang, H. T., Giampiccolo, D., and Magnini, B. The fifth PASCAL recognizing textual entailment challenge. 2009.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Dolan, W. B. and Brockett, C. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the International Workshop on Paraphrasing.*, 2005.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pp. 177–180, 2007.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Sennrich, R., Haddow, B., and Birch, A. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

Sennrich, R., Haddow, B., and Birch, A. Neural machine translation of rare words with subword units. In *ACL*, 2016.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *arXiv preprint arXiv:1506.06724*, 2015.