Understanding Transformer Models at Initialization

Yuchong Zhang & Jonah Mackey

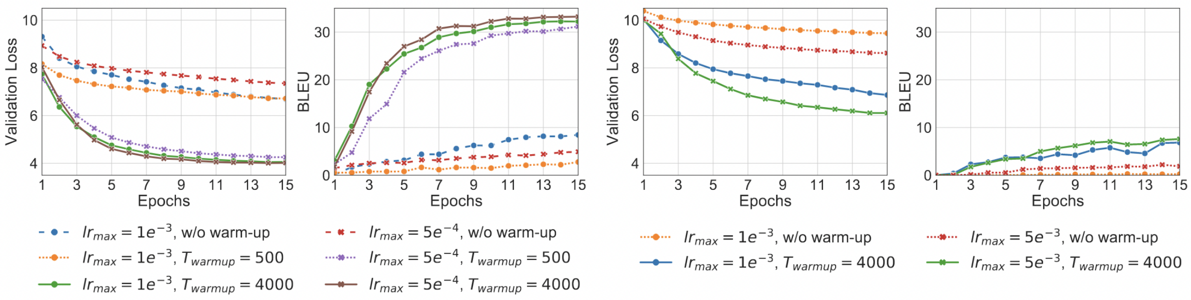
December 8, 2022

1 On Layer Normalization in the Transformer Architecture

In this section we will summarize the main paper [1].

1.1 Motivation

The authors conducted experiments on the IWSLT14 German to English (De-En) translation task. They train a vanilla transformer model with Adam and SGD optimizers, and for each optimizer they consider various learning rate warm-up settings. Results are shown in Figure 1.



(a) Loss/BLEU on the IWSLT14 De-En task (Adam)

(b) Loss/BLEU on the IWSLT14 De-En task (SGD)

Figure 1: Performances of the models optimized by Adam and SGD on the IWSLT14 De-En task.

From these experiments, the authors make the following observations:

- The learning rate warm-up stage is essential.
- The training dynamics are sensitive to the value of T_{warmup} (the number of warm-up steps).
- Adam performs better than SGD.

1.2 Main Contribution

The two main disadvantages of the warm-up stage are that it adds more hyper-parameters to worry about and it can slow down optimization. Motivated by this, the authors propose a modification to the transformer layer. The original architecture (Post-LN) follows "self-attention/feed-forward → addition → layer normalization", the authors proposed architecture (Pre-LN) follows "self-attention/feed-forward → addition → layer normalization". They claim that the Pre-LN transformer's training is more stable and can achieve results competitive with the Post-LN transformer without learning rate warm-up. They justify this theoretically with Theorem 1.2. For theoretical analysis, they consider a transformer model with single-head attention (rather than the standard multi-head attention) and make the following assumptions:

- The query and key weight matrices in all layers are initialized to be zero matrices.
- The input vectors are sampled from the same Gaussian distribution.

Definition 1.1. A random variable $Z \geq 0$ is called $(\epsilon, \delta)-$*bounded* if with probability at least $1-\delta$, $\frac{Z-\mathbb{E}Z}{\mathbb{E}Z} \leq \epsilon$, where $\epsilon > 0$ and $0 < \delta < 1$.

Theorem 1.2. (*Gradients of the last layer in the transformer*) *Assume that the l_2 norm of the unnormalized last layer outputs of the Post-LN transformer and Pre-LN transformer are $(\epsilon, \delta)-$bounded at all positions. Then with probability at least $0.99 - \delta - \frac{\epsilon}{0.9+\epsilon}$, for the Post-LN transformer with L layers, the gradient of the last layer weights, W, satisfies*

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

and for the Pre-LN transformer with L layers,

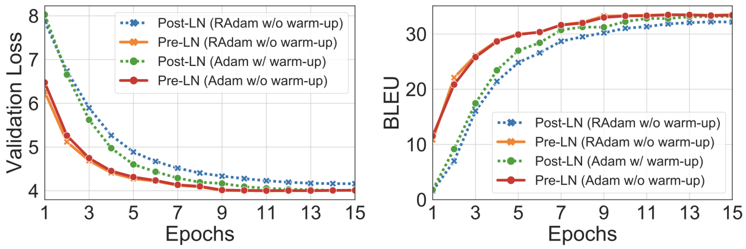
$$\|\frac{\partial \mathcal{L}}{\partial W}\|_F \leq \mathcal{O}\left(d\sqrt{\frac{\ln d}{L}}\right)$$

Note that in the above theorem, $\mathcal{L}$ denotes the loss for the whole sequence, and $\|\cdot\|_F$ denotes the Frobenius norm. The authors also verify this theory empirically. Based on their theory and experiments, the authors make the following conclusions:

- For the Post-LN transformer, the gradients at initialization are large. Thus, training is unstable without learning rate warm-up.
- For the Pre-LN transformer, the gradients at initialization are well-behaved. Thus, training is more stable and no longer requires learning rate warm-up.

1.3 Results

In the previous section, the authors concluded that the Pre-LN transformer does not require the warm-up stage during training. They verify this empirically on two widely used tasks: the IWSLT14 German-to-English (De-En) task and the WMT14 English-to-German (En-De) task. Their results are shown in Figure 2.



(a) Validation Loss (IWSLT) (b) BLEU (IWSLT)

Figure 2: Performances of the models on the IWSLT14 De-En task.

Their experiments show that the Pre-LN transformer trained without warm-up achieves results that are competitive with the Post-LN transformer with warm-up. Furthermore, the Pre-LN transformer converges faster than the Post-LN transformer.

2 Experiments

2.1 Motivation

In [1], the authors attempt to gain insights into the early stage of training for transformers by studying the gradients of Pre-LN and Post-LN transformers at initialization. For gradient-based optimization algorithms, the Hessian spectrum has been studied to draw insights to the optimization dynamics [2]. This motivated us to measure the Hessian spectrum (i.e., distribution of eigenvalues of the Hessian matrix) of Post-LN and Pre-LN transformers at initialization and see what can be said about their loss landscapes.

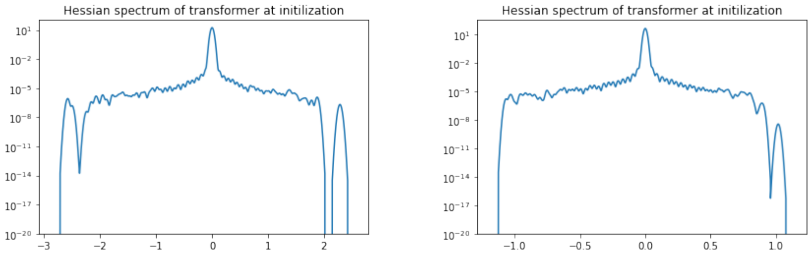
In addition, one of the main conclusions in [1] is that as the number of layers increases, the gradient norm of the last-layer weights decreases for the Pre-LN transformer and stays the same for the Post-LN transformer. This inspired us to measure the Hessian spectrum of Pre-LN and Post-LN transformers with different numbers of encoder/decoder layers, and see if any correlation can be observed.

2.2 Experimental Setting

We considered Pre-LN and Post-LN transformers with different numbers of encoder and decoder layers. For each model, we used 4 attention heads, 1024 hidden nodes in the position-wise feed-forward network, embedding dimension 512, and the Xavier initialization scheme, which is standard for transformers [3, 1]. For the loss function, we used label-smoothed cross entropy with = 0.1. We computed the loss on the German-to-English translation task for the Multi30k validation set, which consists of 1014 sentence pairs. Finally, we used the Lanczos algorithm to approximate the Hessian spectrum.

2.3 Results and Discussions

First, we consider Pre/Post-LN transformers with 6 layers in both the encoder and decoder, same as the setting used in [1]. The result is shown in Figure 3.



(a) Post-LN

(b) Pre-LN

Figure 3: Hessian spectrum of transformers with 6-layer encoder and 6-layer decoder

Clearly the shape of the spectrum of the Post-LN transformer is similar to that of the Pre-LN transformer. In either case, the spectrum is roughly symmetric around 0 and eigenvalues that are close to 0 have the largest density. However, the magnitude of the largest eigenvalue λ_{max} of the Hessian for the Post-LN transformer is larger than that for the Pre-LN transformer ($|\lambda_{max}|$ is around 3 for Post-LN and around 1 for Pre-LN.) Intuitively, the eigenvalues of the Hessian describe the curvature of the loss surface, so this result is saying that the loss landscape has larger curvature in the Post-LN case. More formally, previous works such as [2] has shown that if the loss function is quadratic, i.e., it is of the form

$$\mathcal{L}(\theta) = \frac{h}{2}(\theta - z)^2 + C$$

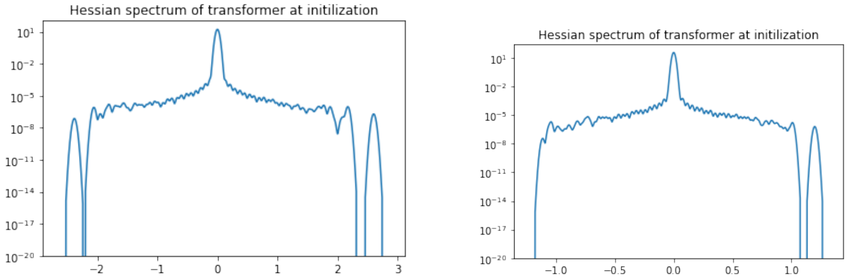
where θ denotes the model parameters, then the optimal learning rate of gradient descent is given by

$$\eta_{opt} = \frac{1}{\lambda_{max}}$$

where λ_{max} is the maximum eigenvalue of the Hessian of the loss. Our measurement result suggests that λ_{max} is larger for the Post-LN transformer. So if we make the additional assumption that the loss for transformer models at initialization can be well-approximated via its second-order Taylor expansion, then it follows that the optimal learning rate for Post-LN transformers is smaller at initialization, which in some sense explains why learning rate warm-up is needed when training Post-LN transformers.

However, we do not think this assumption should be naively made because a priori, there is no reason to assume that the loss function at initialization is convex or approximately quadratic. Moreover, one can argue that the discrepancy in $|\lambda_{max}|$ between Pre-LN and Post-LN transformers is not significant enough to explain why the initial learning rate for Post-LN transformers need to be so much smaller.

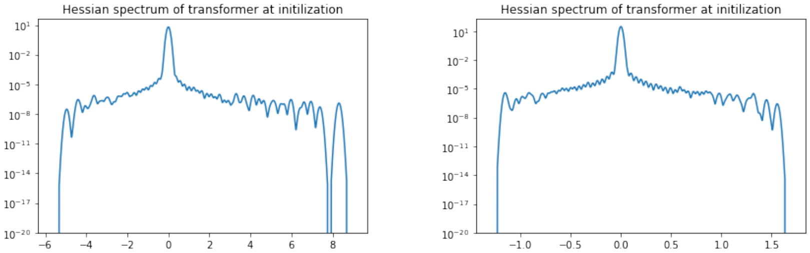
When we require the encoder to have the same number of layers as the decoder and vary that number, we observe roughly the same amount of discrepancy in $|\lambda_{max}|$ between Pre-LN and Post-LN transformers. However, a different pattern emerges when the encoder has a different number of layers from the decoder. We measured the Hessian spectrum for Post/Pre-LN transformers when the encoder has 6 layers and the decoder has 18 layers, and vice versa. The results are shown in Figures 4 and 5:



(a) Post-LN

(b) Pre-LN

Figure 4: Hessian spectrum of transformers with 6-layer encoder and 18-layer decoder



(a) Post-LN

(b) Pre-LN

Figure 5: Hessian spectrum of transformers with 18-layer encoder and 6-layer decoder

The results above suggest that while in both cases $|\lambda_{max}|$ of the Post-LN transformer is larger than that of the Pre-LN transformer, this discrepancy is more significant when the encoder has more layers than the decoder. Thus, it would be interesting to study how approximately quadratic the loss function is at initialization in both cases to see how well the theory in [2] can explain the optimization dynamics.

References

[1] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In *International Conference on Machine Learning*, pages 10524–10533. PMLR, 2020.

[2] Yann LeCun, Patrice Simard, and Barak Pearlmutter. Automatic learning rate maximization by online estimation of the hessian's eigenvectors. In S. Hanson, J. Cowan, and C. Giles, editors, *Advances in Neural Information Processing Systems*, volume 5. Morgan-Kaufmann, 1992.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. 2017.