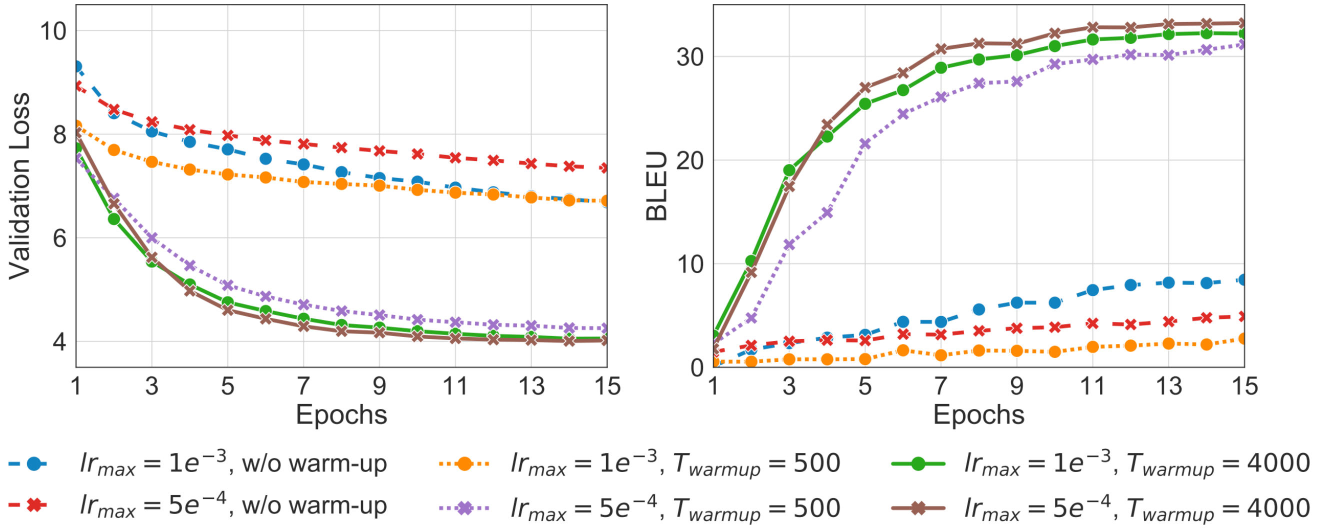
Transformer Details Not Described in The Paper

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

Transformer의 단점

- 학습이 까다롭다.
  - Bad local optima에 빠지기 매우 쉬움
  - 그런데 paper에서 이것을 언급하지 않음
    - #warm-up step, learning rate



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

- 오죽하면,,
  - Training Tips for the Transformer Model [Popel et al., 2018]
  - Transformers without Tears: Improving the Normalization of Self-Attention [Nguyen et al., 2019]
  - ON THE VARIANCE OF THE ADAPTIVE LEARNING RATE AND BEYOND [Liu et al., 2020]

On Layer Normalization in the Transformer Architecture [Xiong et al., 2020]

- Previous work:
  - Use Noam decay (warm-up and linear decay)
  - Rectified Adam (RAdam)

- Proposed:
  - Layer Norm의 위치에 따라 학습이 수월해짐
    - LN이 gradient를 평탄하게 바꾸는 효과

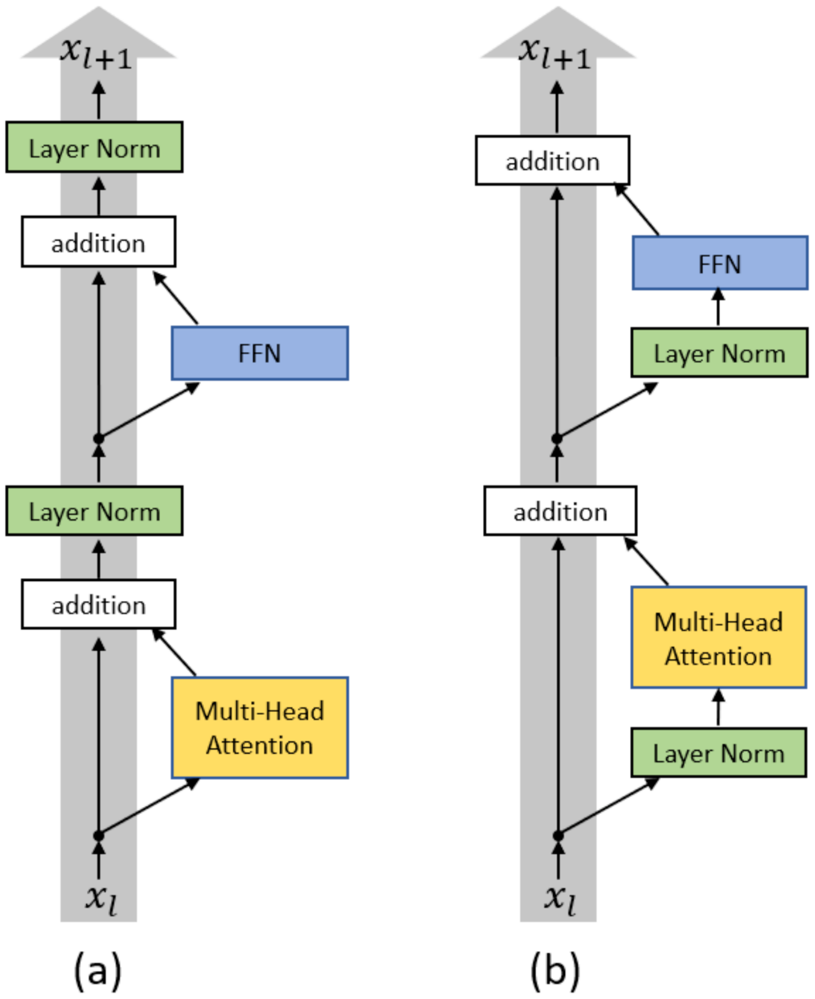


Figure 1: (a) Post-LN Transformer layer; (b) Pre-LN Transformer layer.

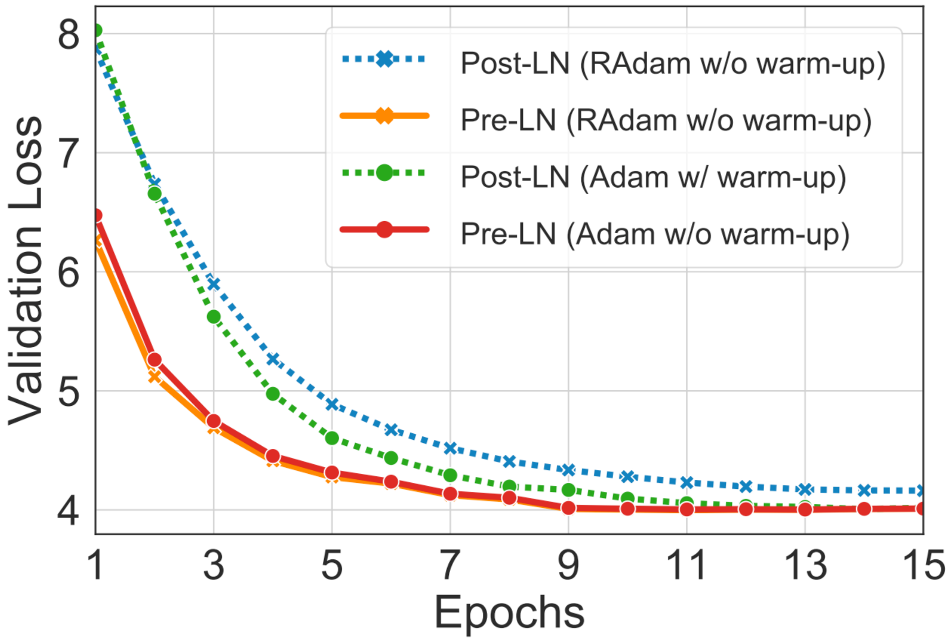
On Layer Normalization in the Transformer Architecture [Xiong et al., 2020]

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

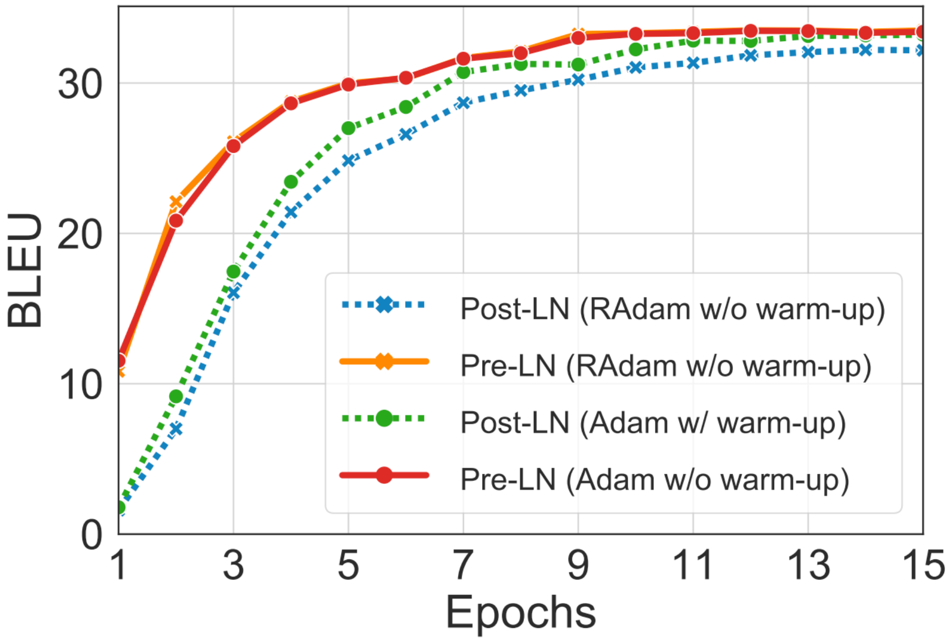
Post-LN Transformer	Pre-LN Transformer
$x_{l,i}^{post,1} = \text{MultiHeadAtt}(x_{l,i}^{post}, [x_{l,1}^{post}, \cdots, x_{l,n}^{post}])$	$x_{l,i}^{pre,1} = \text{LayerNorm}(x_{l,i}^{pre})$
$x_{l,i}^{post,2} = x_{l,i}^{post} + x_{l,i}^{post,1}$	$x_{l,i}^{pre,2} = \text{MultiHeadAtt}(x_{l,i}^{pre,1}, [x_{l,1}^{pre,1}, \cdots, x_{l,n}^{pre,1}])$
$x_{l,i}^{post,3} = \text{LayerNorm}(x_{l,i}^{post,2})$	$x_{l,i}^{pre,3} = x_{l,i}^{pre} + x_{l,i}^{pre,2}$
$x_{l,i}^{post,4} = \text{ReLU}(x_{l,i}^{post,3} W^{1,l} + b^{1,l}) W^{2,l} + b^{2,l}$	$x_{l,i}^{pre,4} = \text{LayerNorm}(x_{l,i}^{pre,3})$
$x_{l,i}^{post,5} = x_{l,i}^{post,3} + x_{l,i}^{post,4}$	$x_{l,i}^{pre,5} = \text{ReLU}(x_{l,i}^{pre,4} W^{1,l} + b^{1,l}) W^{2,l} + b^{2,l}$
$x_{l+1,i}^{post} = \text{LayerNorm}(x_{l,i}^{post,5})$	$x_{l+1,i}^{pre} = x_{l,i}^{pre,5} + x_{l,i}^{pre,3}$
	Final LayerNorm: $x_{Final,i}^{pre} \leftarrow \text{LayerNorm}(x_{L+1,i}^{pre})$

On Layer Normalization in the Transformer Architecture [Xiong et al., 2020]

- Evaluation Results



(a) Validation Loss (IWSLT)



(b) BLEU (IWSLT)

Summary

- Pre-Norm 방식을 통해 warm-up 및 LR 튜닝 제거 가능
  - LR decay는 여전히 필요

- 그 밖에도 Layer Norm을 대체하거나, weight initialization을 활용하여 좀 더 나은 성능을 확보할 수 있음