Improving Transformer Optimization Through Better Initialization

Xiao Shi Huang, Felipe Perez*, Jimmy Ba, Maksims Volkovs*

VECTOR INSTITUTE

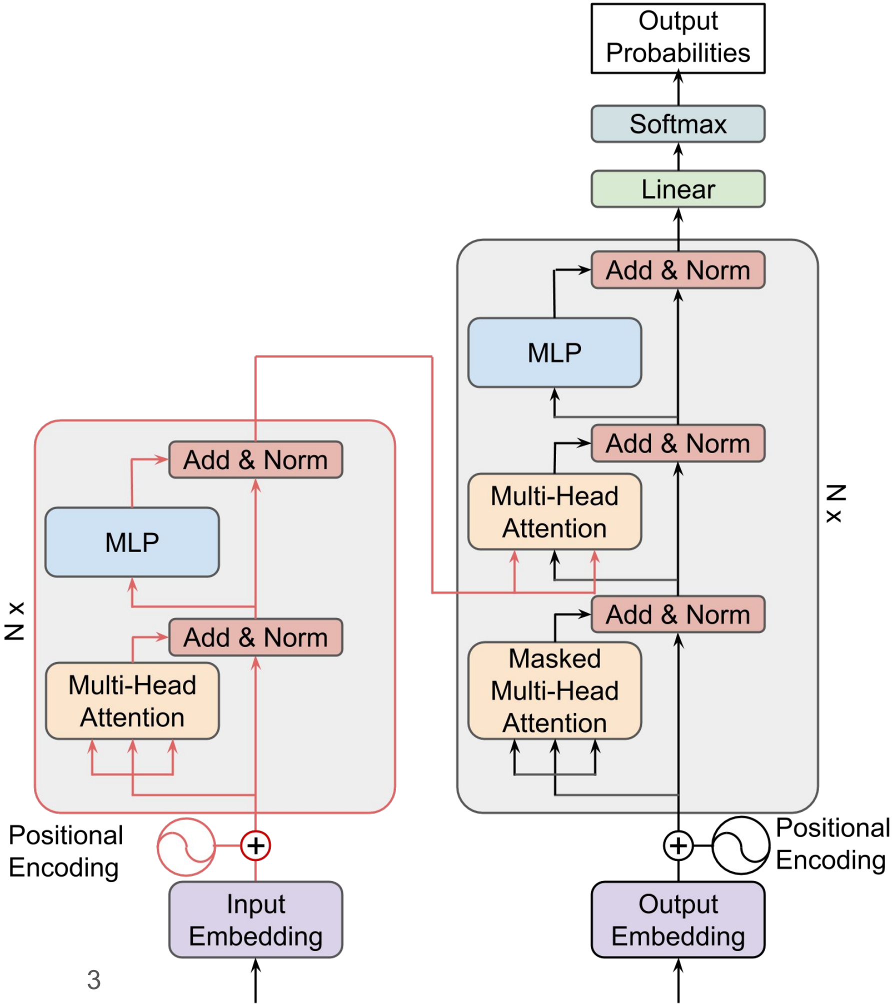
layer6

Mitacs

Computer Science
UNIVERSITY OF TORONTO

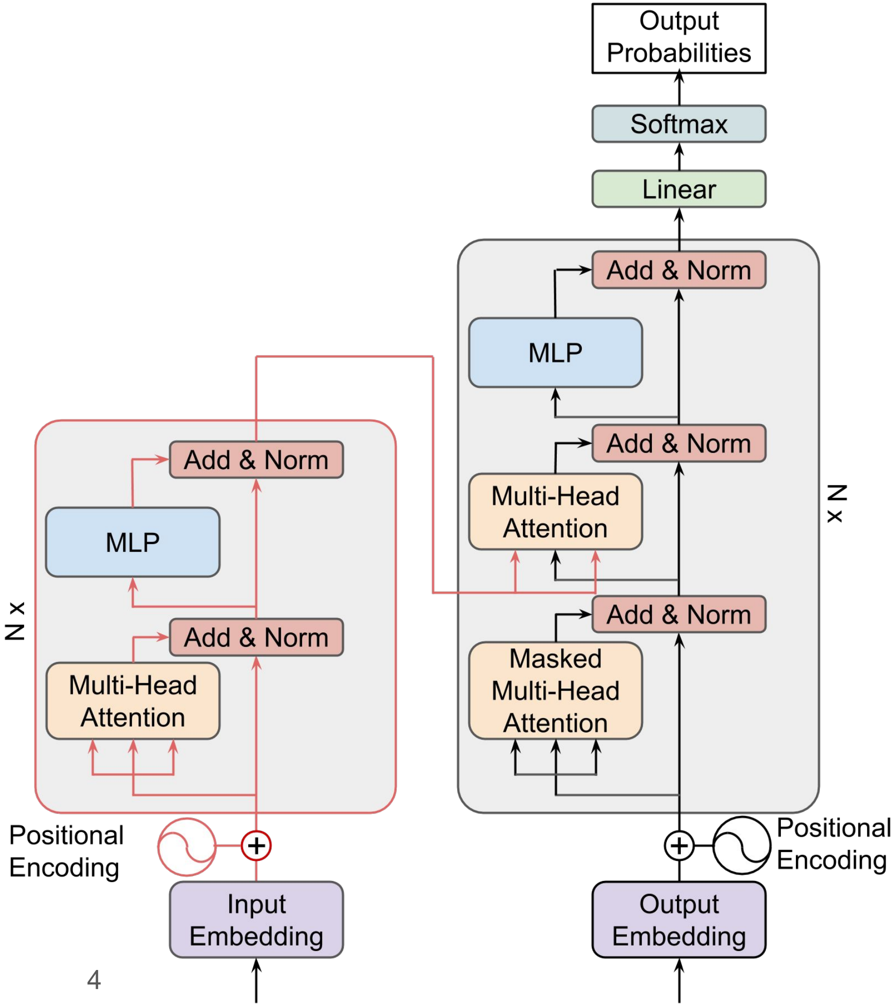
Agenda

- Transformer in Detail
- Removing Warmup: T-Fixup
- Experimental Results
- Summary



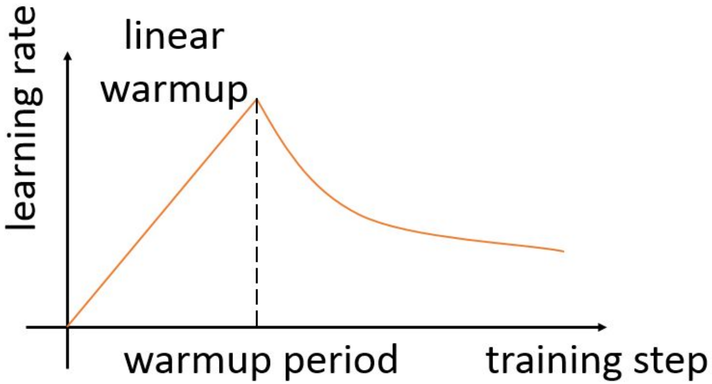
Transformer

- **Encoder-Decoder** architecture
- **Residual** backbone
- **Multi-Headed Attention** in ResBlock
- **LayerNorm** after every residual block

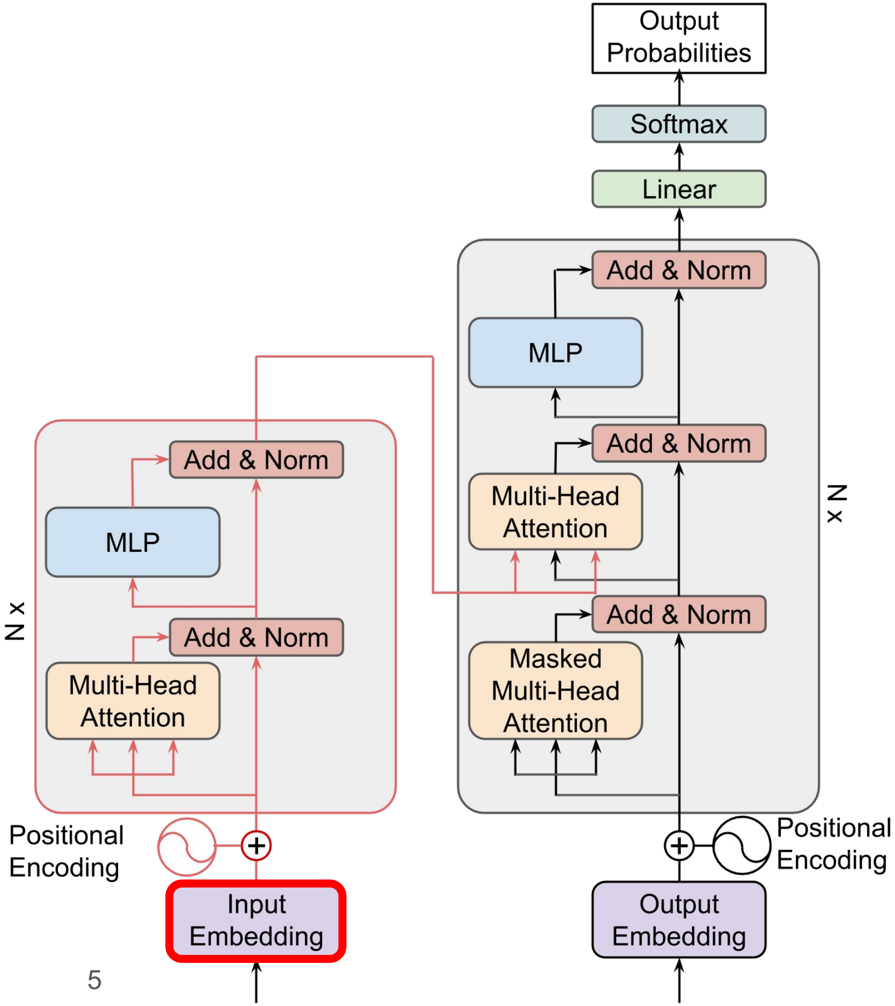


Training

- Adam optimizer
- Inverse square root learning rate decay
- Learning rate warmup
-

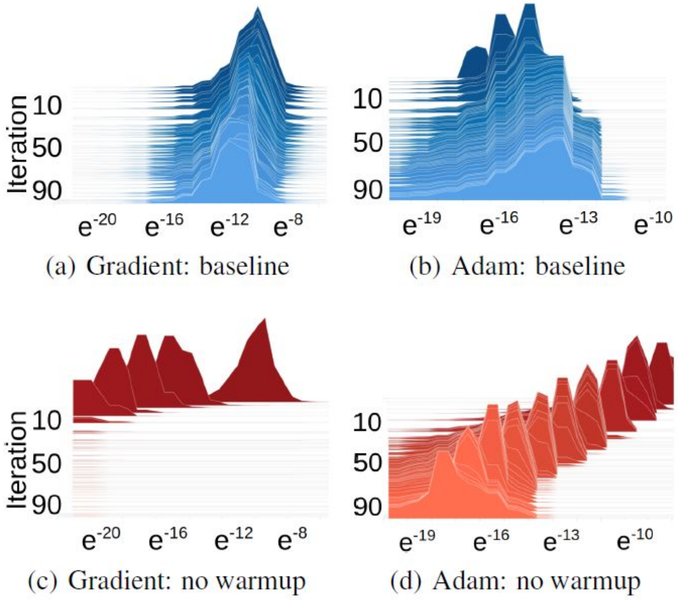


layer 6



Necessity of Warmup

- Gradient histogram

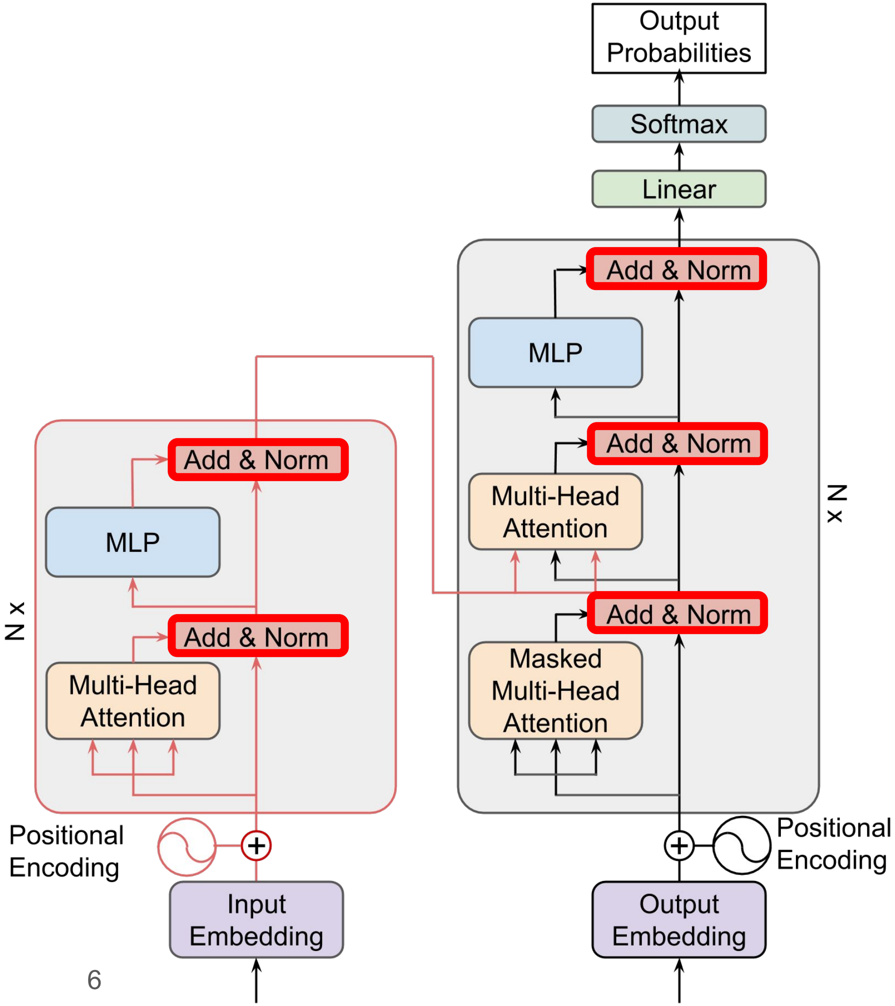


(a) Gradient: baseline

(b) Adam: baseline

(c) Gradient: no warmup

(d) Adam: no warmup



Necessity of Warmup

- LayerNorm in Backpropagation[2]

$$\left\| \frac{\partial \mathrm{LN}(\boldsymbol{x})}{\partial \boldsymbol{x}} \right\| = O\left(\frac{\sqrt{d}}{||\boldsymbol{x}||} \right)$$

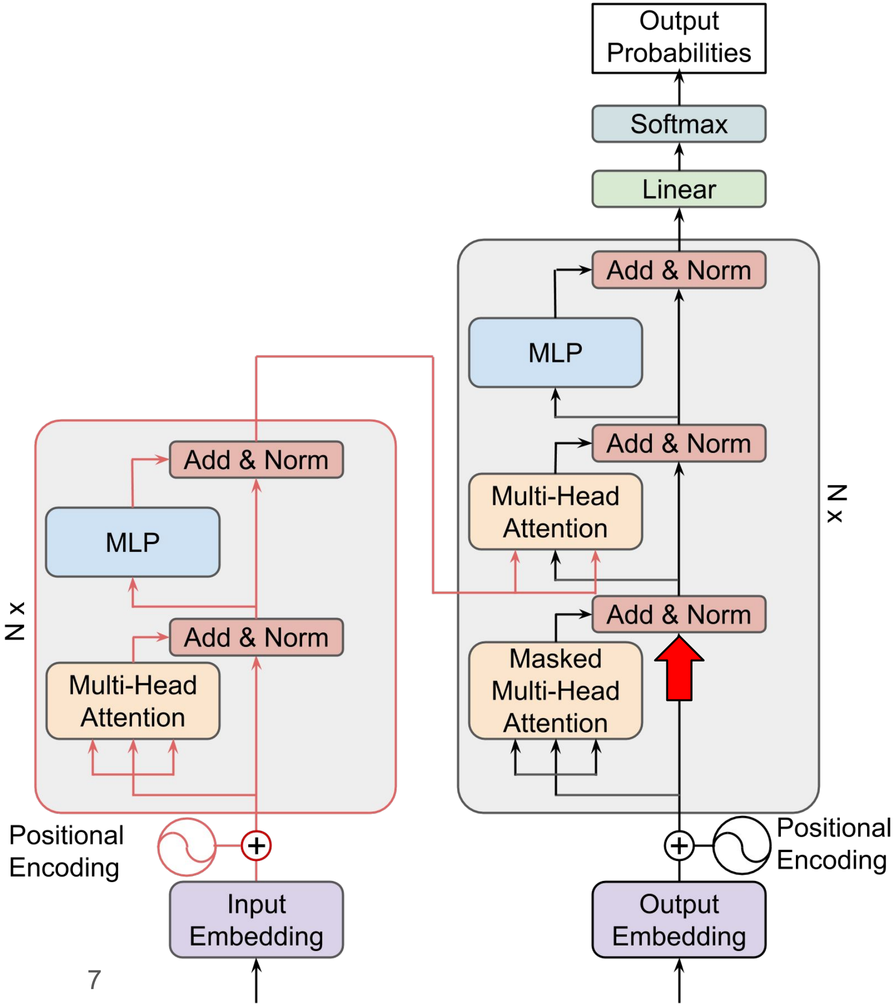
- x: input to Layer Normalization
- d: dimension of x

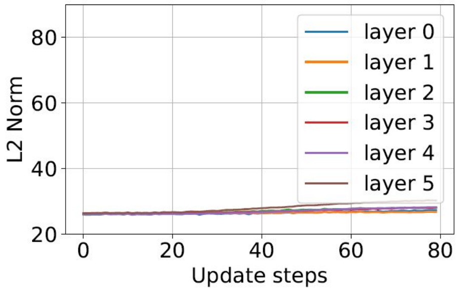
Error signal decreases with a large input

layer6

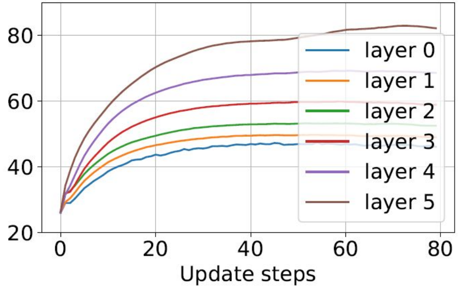
Necessity of Warmup

- LayerNorm in Backpropagation[2]



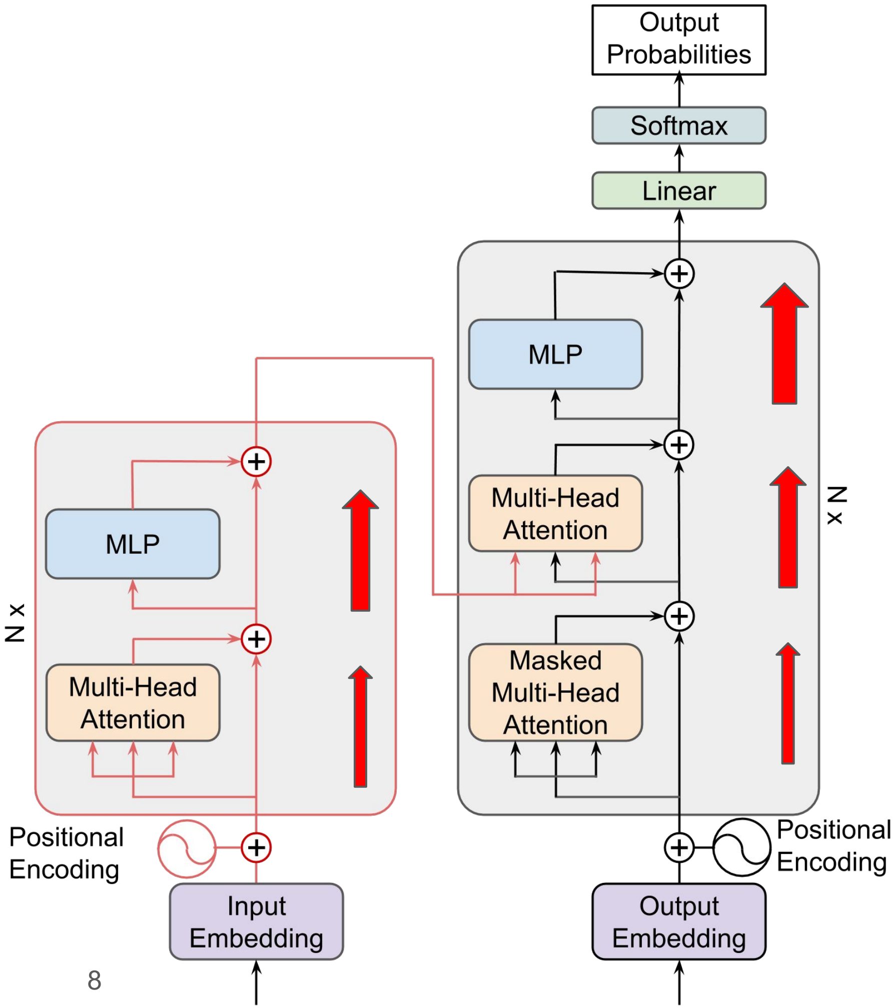


(a) Warmup



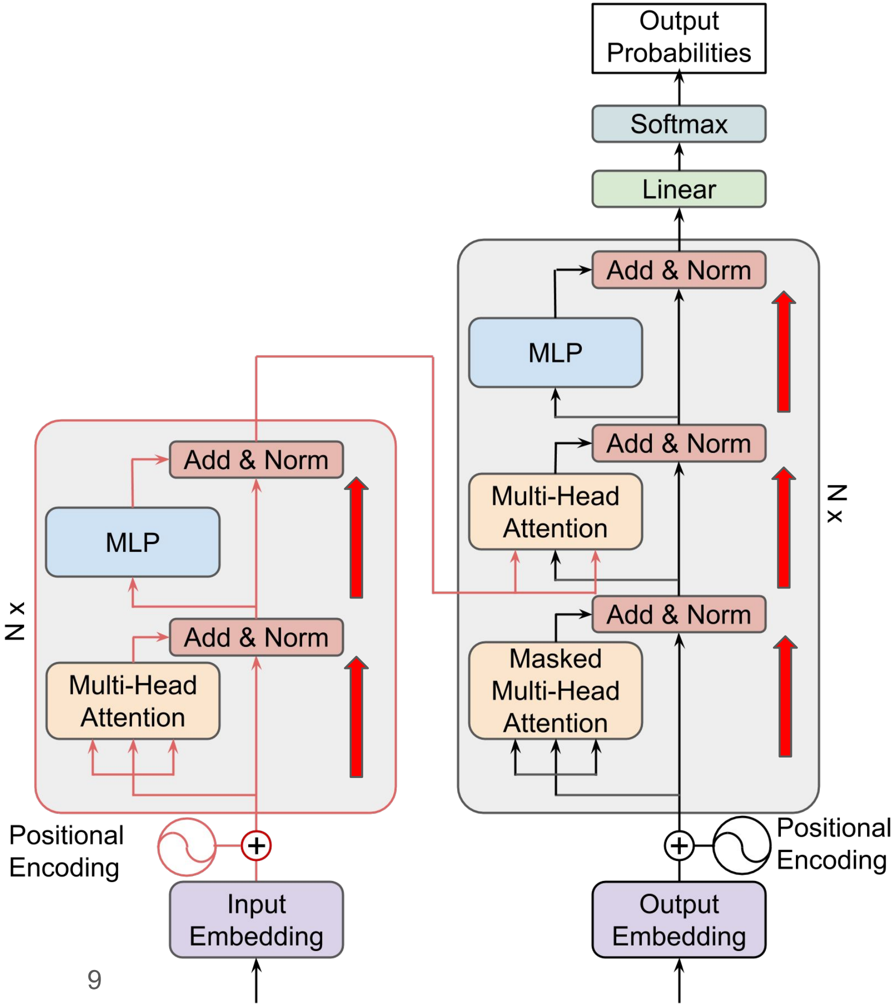
(b) No warmup

layer6



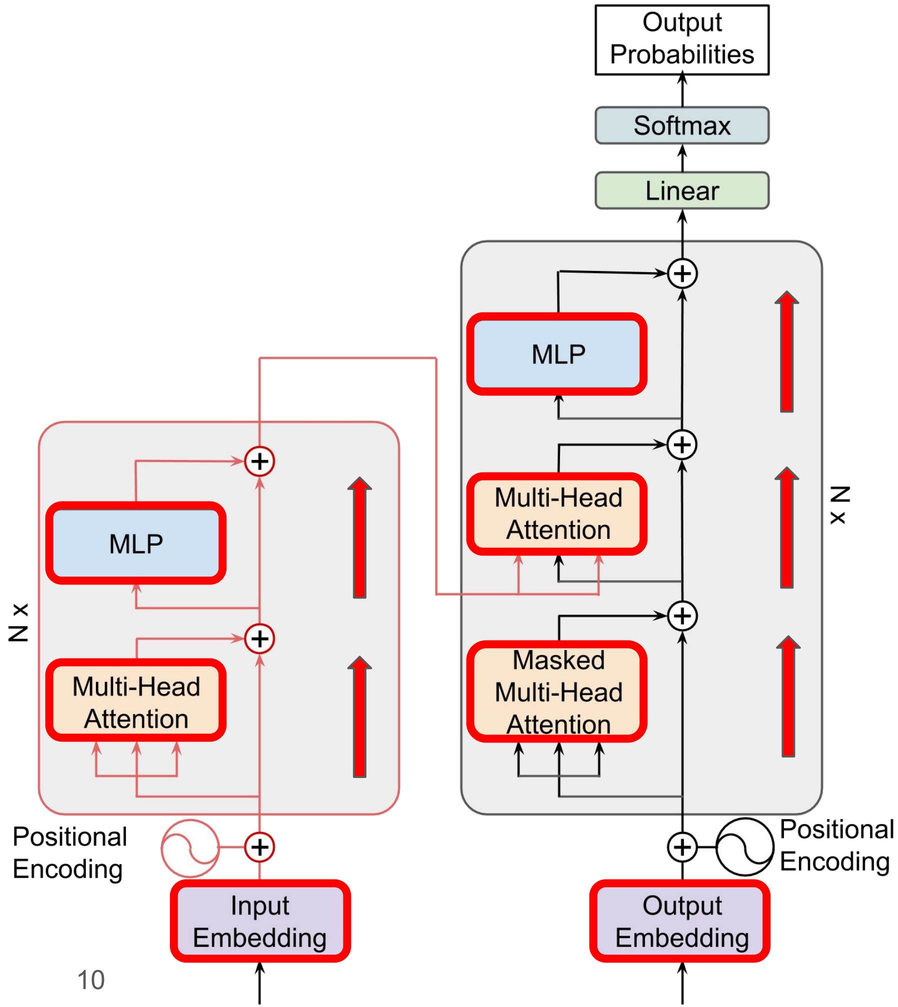
Removing Warmup

- Without LayerNorm:
  - Magnitude on backbone grows with layer depth



Removing Warmup

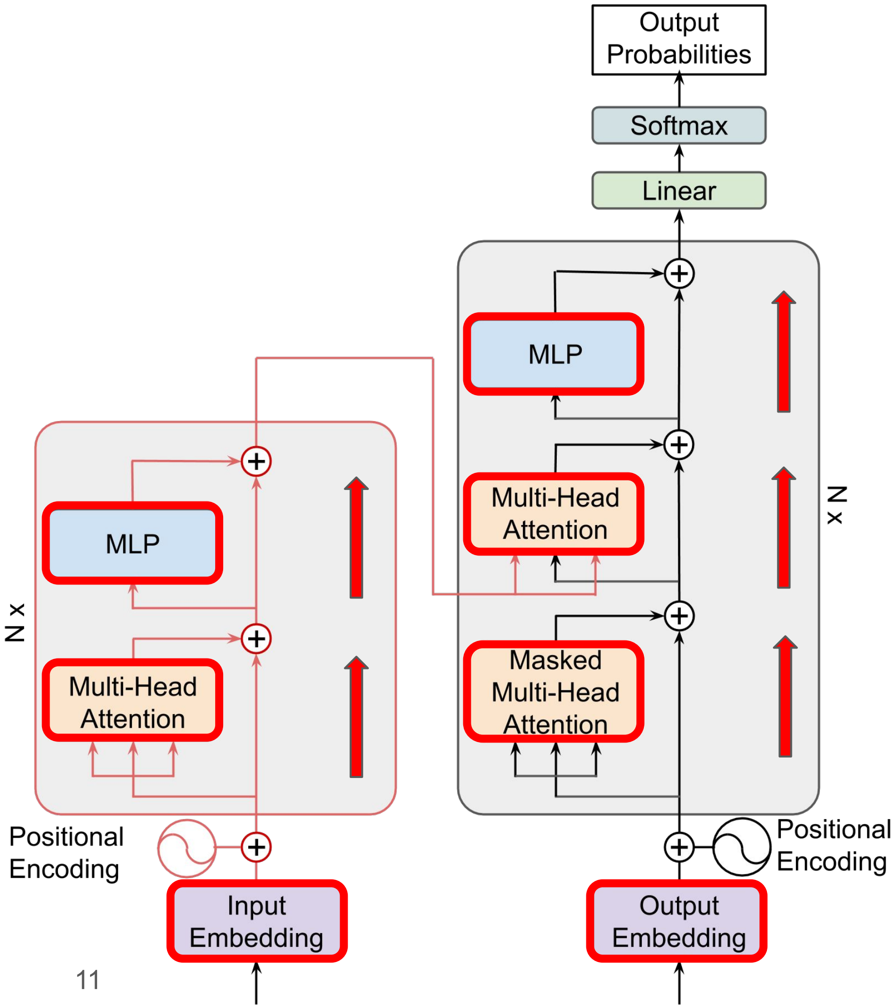
- Without LayerNorm:
    - Magnitude on backbone grows with layer depth
- With LayerNorm:
    - Reset to unit magnitude



Removing Warmup

- Without LayerNorm:
  - Magnitude on backbone grows with layer depth
- With LayerNorm:
  - Reset to unit magnitude
- Parameter-Controller Growth

layer6



Removing Warmup

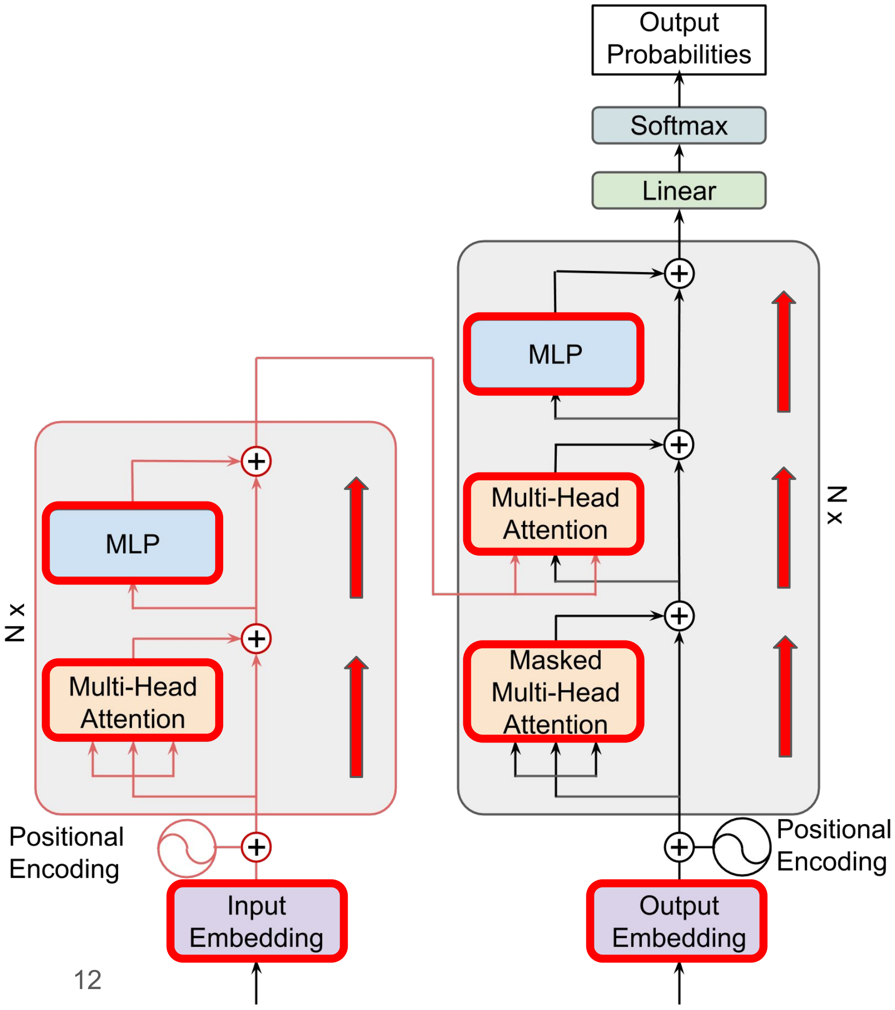
Goal: Control the total change on the output of the transformer after a gradient update.

Control output change in residual blocks:

- Feedforward blocks as in Fixup
- **Theorem:** For Attention blocks, this is controlled when:

$$\|v\|^2\|w\|^2 + \|w\|^2\|m\|^2 + \|v\|^2\|m\|^2 = \Theta(1/L)$$

v : Value projection matrix
w : mixing matrix
m : Value input
L : number of layers

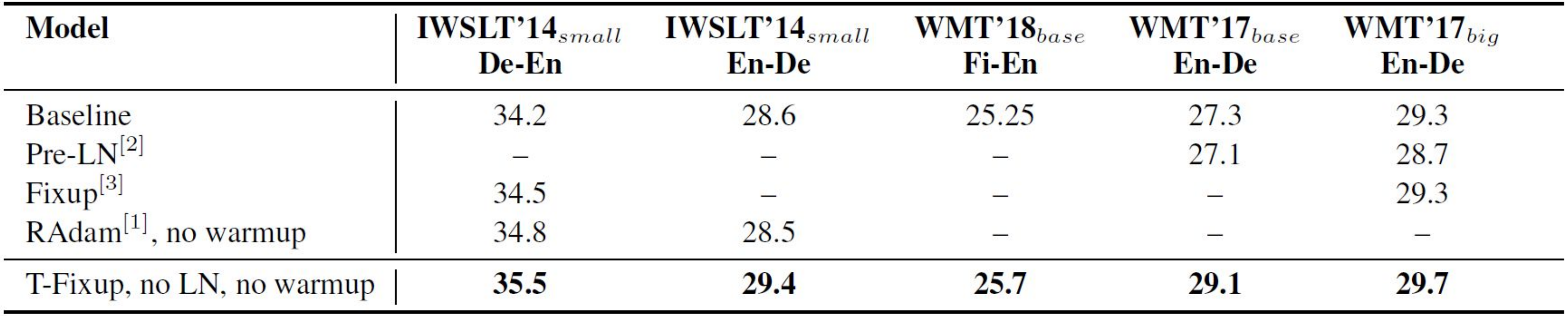


Removing Warmup

- T-Fixup Initialization
  - Xavier Initialization for all projection matrices
  - Gaussian initialization for embedding layers
  - Scale embedding layers and decoder parameters by $(9N)^{-1/4}$
  - Scale encoder parameters by $0.67N^{-1/4}$

Experimental Results

T-Fixup on Standard Transformer

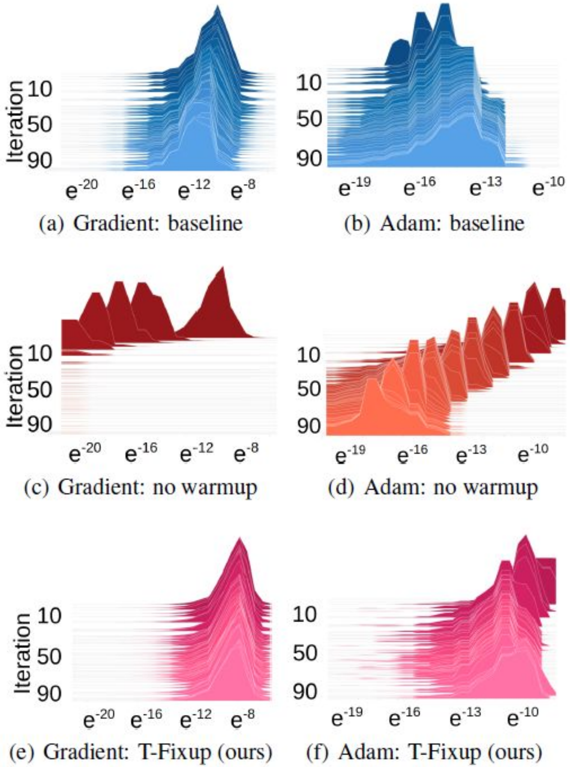


Model	**IWSLT'14$_{small}$** **De-En**	**IWSLT'14$_{small}$** **En-De**	**WMT'18$_{base}$** **Fi-En**	**WMT'17$_{base}$** **En-De**	**WMT'17$_{big}$** **En-De**
Baseline	34.2	28.6	25.25	27.3	29.3
Pre-LN[2]	–	–	–	27.1	28.7
Fixup[3]	34.5	–	–	–	29.3
RAdam[1], no warmup	34.8	28.5	–	–	–
T-Fixup, no LN, no warmup	**35.5**	**29.4**	**25.7**	**29.1**	**29.7**

Table 1. NMT Test BLEU Scores

- T-Fixup achieves consistently higher performance with less structure

T-Fixup on Standard Transformer: gradients



(a) Gradient: baseline

(b) Adam: baseline

(c) Gradient: no warmup

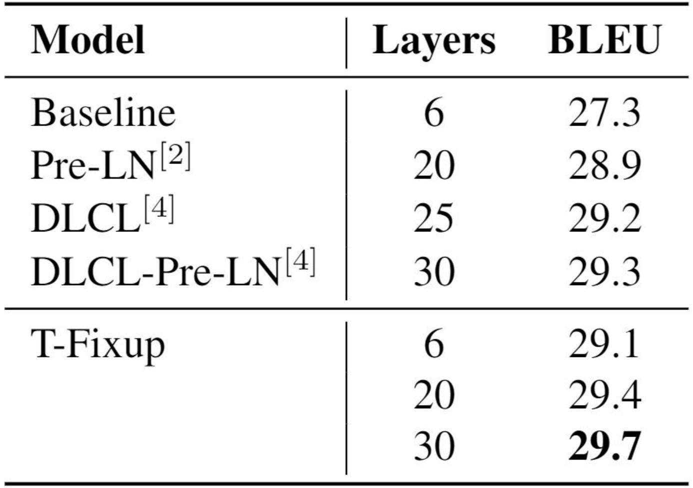
(d) Adam: no warmup

(e) Gradient: T-Fixup (ours)

(f) Adam: T-Fixup (ours)

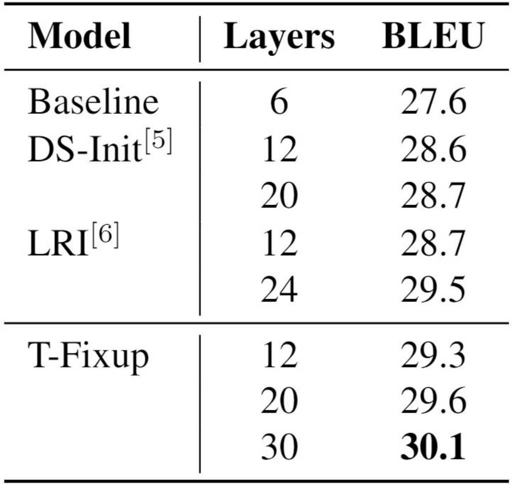
- Gradient and Adam Update Magnitudes
  - Vanilla Transformer Without Warmup
    - vanishing gradient
  - T-Fixup Without Warmup
    - stable error signal throughout training

T-Fixup on Deeper Transformer



Model	Layers	BLEU
Baseline	6	27.3
Pre-LN[2]	20	28.9
DLCL[4]	25	29.2
DLCL-Pre-LN[4]	30	29.3
T-Fixup	6	29.1
	20	29.4
	30	**29.7**

Table 2. WMT’17 En-De BLEU.



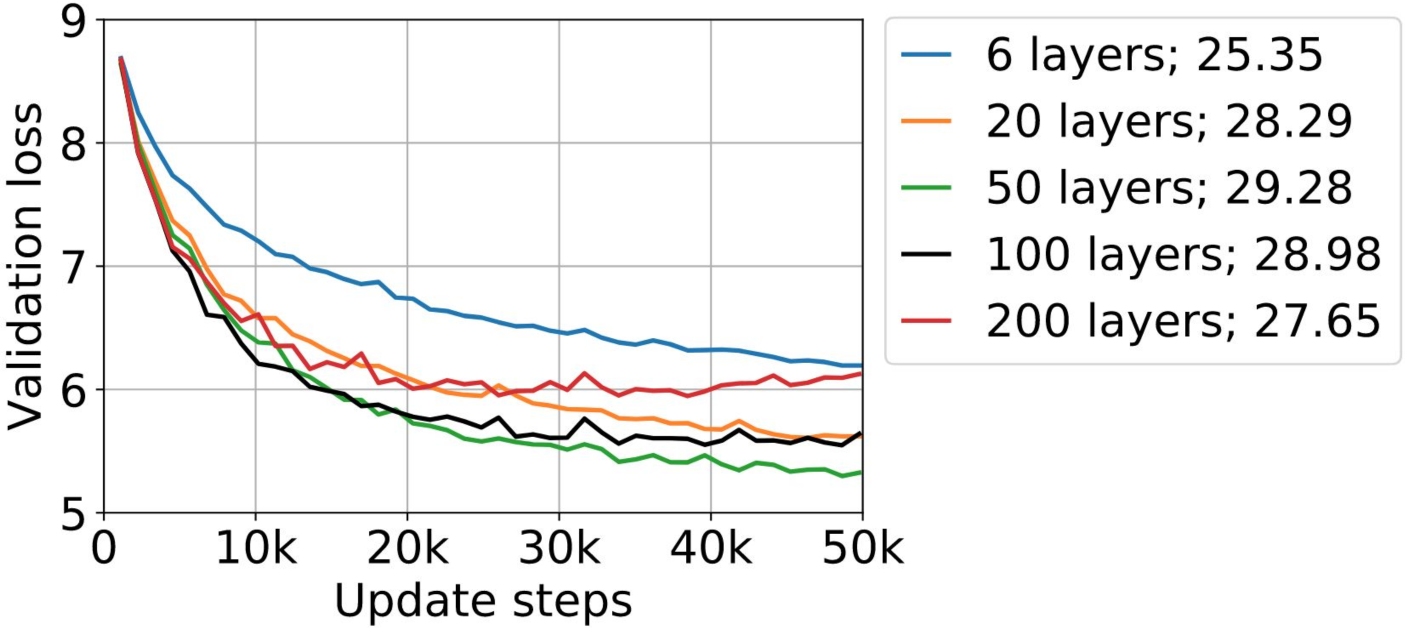
Model	Layers	BLEU
Baseline	6	27.6
DS-Init[5]	12	28.6
	20	28.7
LRI[6]	12	28.7
	24	29.5
T-Fixup	12	29.3
	20	29.6
	30	**30.1**

Table 3. WMT’14 En-De BLEU

- T-Fixup outperforms all competitive models with equal or less layers

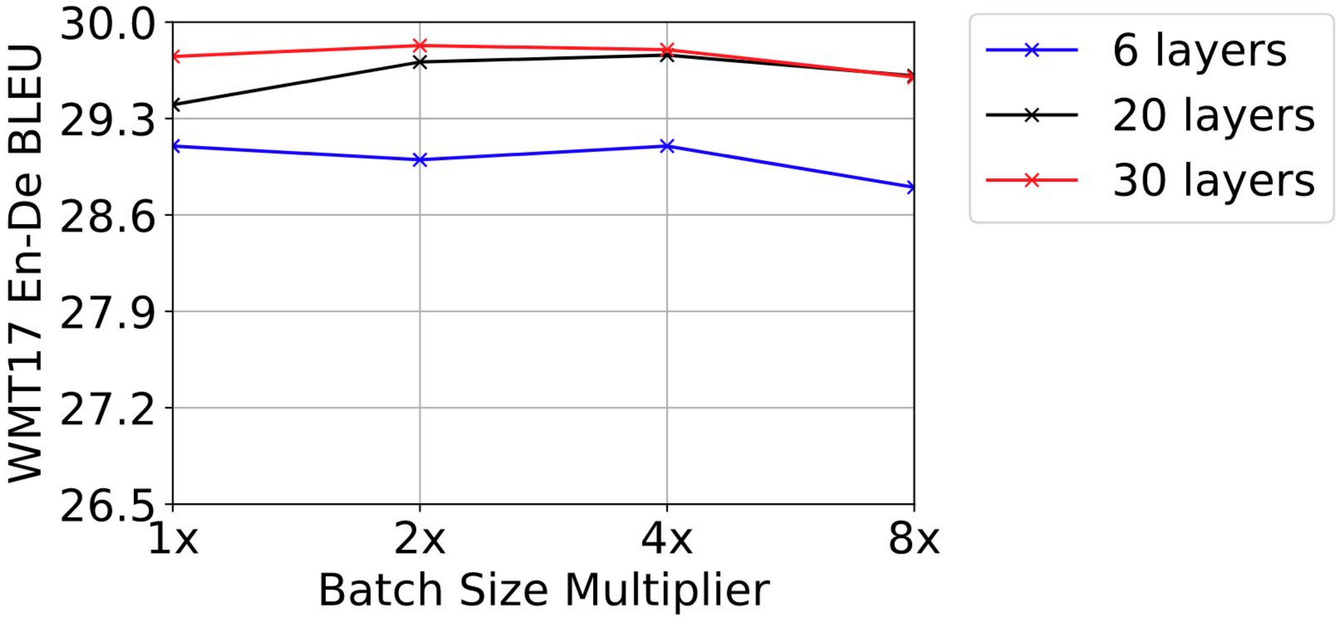
T-Fixup on Ultra-Deep Transformer

- IWSLT’14 De-En dataset, 64(embed)-128(MLP hidden)-2(head) Transformer



T-Fixup on Large Batch Training

- WMT'17 En-De Dataset, WMT$_{\text{base}}$ Transformer



Summary

Summary

- Requirement for learning rate warmup: Adam + LayerNorm
- T-Fixup Initialization
    - Superior performance on NMT
    - Ultra-Deep Transformer
- Future Work

Acknowledgement

Mitacs

VECTOR INSTITUTE

Computer Science
UNIVERSITY OF TORONTO

VELUT ARBOR ÆVO

Thank you!

Questions?

Contact: Xiao Shi (Gary) Huang
gary@layer6.ai

layer6

References

[1]: Liu, L. etc. *On the variance of the adaptive learning rate and beyond*. In ICLR, 2020

[2]: Xiong, R. etc. *On layer normalization in the transformer architecture*. In ICML, 2020

[3]: Zhang, H. etc. *Fixup initialization: residual learning without normalization*, In ICLR, 2019

[4]: Wang. Q. etc. *Learning deep transformer models for machine translation*. In ACL, 2019

[5]: Zhang, B. etc. *Improving deep transformer with depth-scaled initialization and merged attention*. In EMNLP, 2019

[6]: Xu. H. etc. *Why deep transformers are difficult to converge? From computation order to Lipschitz restricted parameter initialization*. In Arxiv