2024 ICLR

SIMPLIFYING TRANSFORMER BLOCKS

Bobby He & Thomas Hofmann*

Department of Computer Science, ETH Zurich

24.1.22

Presented by Yyyq

00

作者介绍

Published as a conference paper at ICLR 2023

DEEP TRANSFORMERS WITHOUT SHORTCUTS: MODIFYING SELF-ATTENTION FOR FAITHFUL SIGNAL PROPAGATION

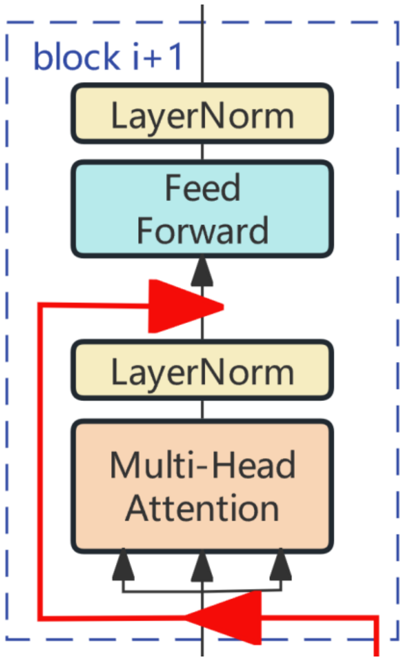
Bobby He[1] **James Martens**[2] **Guodong Zhang**[2] **Aleksandar Botev**[2]
Andrew Brock[2] **Samuel L. Smith**[2] **Yee Whye Teh**[1,2]
[1]University of Oxford, [2]DeepMind
Correspondence to: `bobby.he@stats.ox.ac.uk`, `jamesmartens@google.com`.

"It is possible to successfully train deep transformers without skip connections or normalisation layers."

01 问题描述

- 探索标准transformer块可以简化到什么程度
  - skip connections
  - projection/value matrices
  - sequential sub-blocks
  - normalisation layers
- 为什么研究？
  - 深度学习理论和实践之间存在差距
  - 训练和部署大型transformer模型的成本过高

02 相关工作



1、信号传播理论

- 信号传播理论有助于我们分析一个网络结构设计的是否良好。
- 随着网络的加深，对不同的输入已经无法区分，这显然不是一个良好的网络。

2、残差网络(ResNet)

- 残差块 Residual Block
- 跳跃连接 skip connection

3、残差降权的思想

- 剪枝操作，评估每个权重的重要性
- downweight the residual branch relative to the skip branch

4、Skipless

02

相关工作

1、信号传播理论

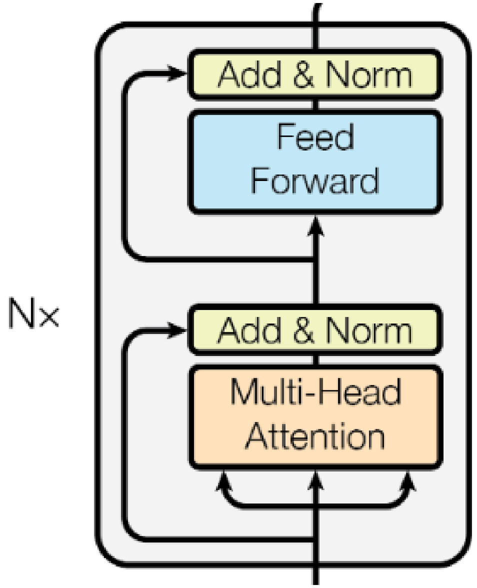
2、残差网络(ResNet)

3、残差降权的思想

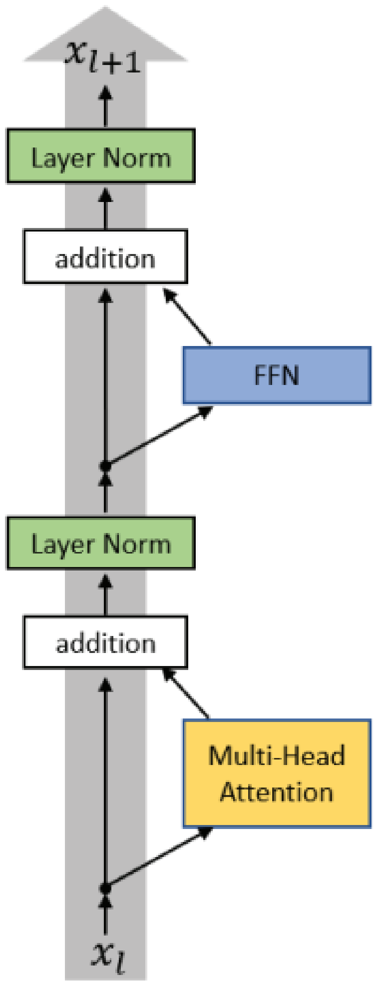
4、无残差架构 Skipless

- 在mlp和cnn中：非线性激活函数→更线性，即使没有跳跃连接，也可以实现良好的信号传播
- 在Attention中运用这个思想：Attention matrices need to be more “identity-like”
- 使用标准优化器，无残差架构会损失速度

03 预备知识

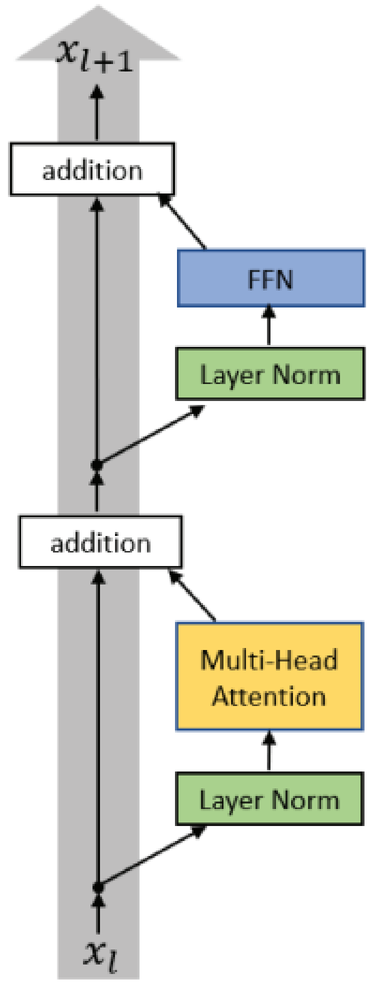


$$\mathrm{LayerNorm}(x + \mathrm{Sublayer}(x))$$



(a)

Post-LN

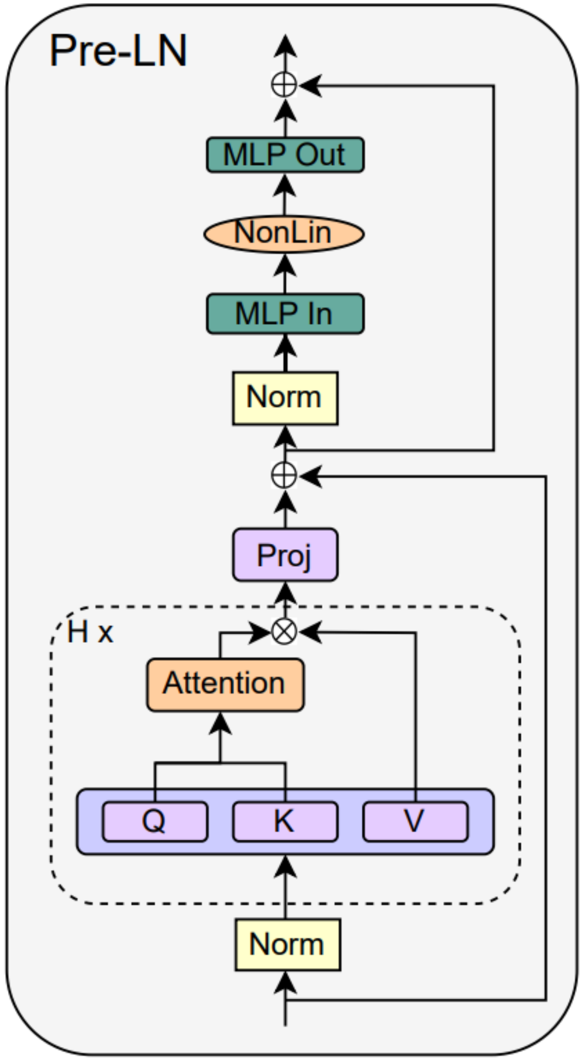


(b)

Pre-LN

[1] Xiong R, Yang Y, He D, et al. On layer normalization in the transformer architecture[C]//International Conference on Machine Learning. PMLR, 2020: 10524-10533.

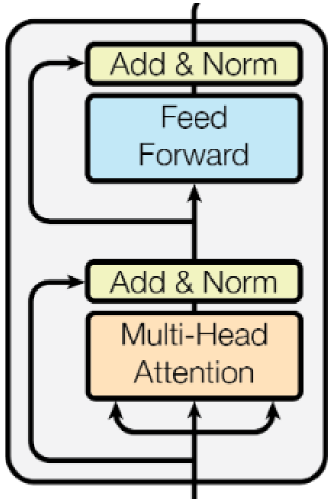
03 预备知识



$$\mathbf{X}_{\text{out}} = \alpha_{\text{FF}}\,\hat{\mathbf{X}} + \boxed{\beta_{\text{FF}}}\,\text{MLP}(\text{Norm}(\hat{\mathbf{X}})), \quad \text{where } \hat{\mathbf{X}} = \alpha_{\text{SA}}\,\mathbf{X}_{\text{in}} + \beta_{\text{SA}}\,\text{MHA}(\text{Norm}(\mathbf{X}_{\text{in}})). \tag{1}$$

$$\text{Attn}(\mathbf{X}) = \mathbf{A}(\mathbf{X})\mathbf{X}\boxed{\mathbf{W}^V} \quad \text{where } \mathbf{A}(\mathbf{X}) = \text{Softmax}\left(\frac{1}{\sqrt{d_k}}\mathbf{X}\mathbf{W}^Q{\mathbf{W}^K}^\top\mathbf{X}^\top + \mathbf{M}\right), \tag{2}$$

$$\text{MHA}(\mathbf{X}) = \text{Concat}\big(\text{Attn}_1(\mathbf{X}), \ldots, \text{Attn}_H(\mathbf{X})\big)\boxed{\mathbf{W}^P}, \tag{3}$$



α和β参数的默认值都是1。

*补充：Decoder-only的GPT 和 Encoder-only的BERT

➢ **Decoder-only 的 GPT（例如，GPT-2 和 GPT-3）：**

① 任务：<u>**GPT主要用于生成型任务**</u>，例如文本生成、对话生成等。它的模型结构允许生成下一个词或标记的概率分布，使其适用于自然语言生成任务。

② 定位： GPT强调对上下文的建模，通过自回归（autoregressive）的方式，从左到右逐步生成文本序列。

③ <u>mask矩阵的主对角线及以下为0，主对角线以上为 −∞；</u>

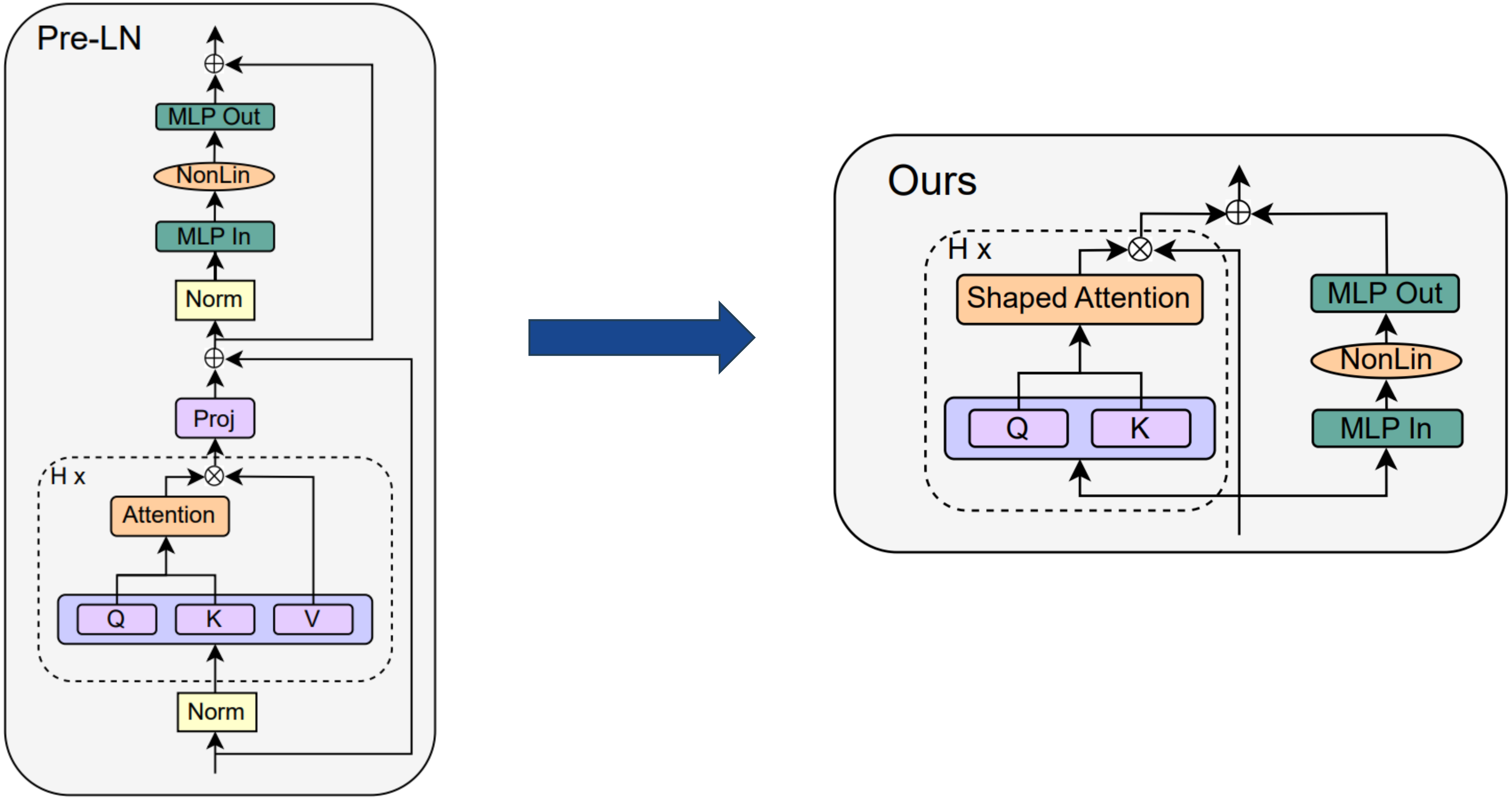
➢ **Encoder-only 的 BERT：**

① 任务： <u>**BERT主要用于预测型任务**</u>，例如语言模型的掩码语言建模（Masked Language Model，MLM）任务。在预训练阶段，BERT通过掩盖输入中的一些词并预测它们，学习词汇的上下文表示。

② 定位： BERT的关注点是双向上下文建模，它在预训练中通过双向编码获得更全面的上下文表示。

③ <u>mask矩阵全零。</u>

04 Simplifying Blocks



04 Simplifying Blocks：删除Attention子块的skip connection

$$\mathbf{A}(\mathbf{X}) = \text{Softmax}\left(\frac{1}{\sqrt{d_k}}\mathbf{X}\mathbf{W}^Q{\mathbf{W}^K}^\top\mathbf{X}^\top + \mathbf{M}\right),$$

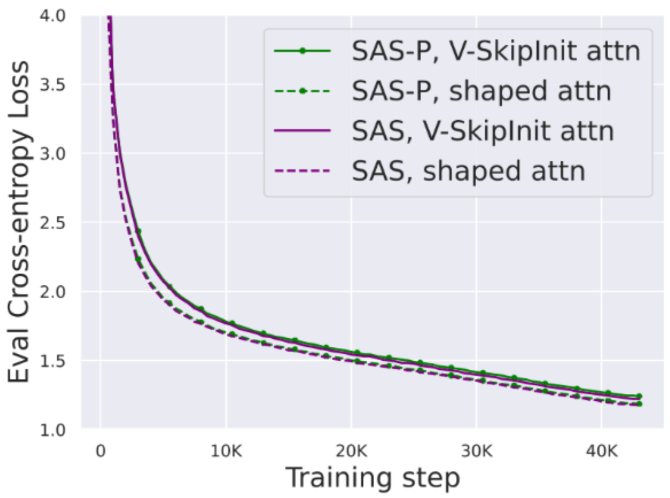
➢ 起点：Skipless模型

①.1 Value-SkipInit $\mathbf{A}(\mathbf{X}) \leftarrow (\alpha\mathbf{I}_T + \beta\mathbf{A}(\mathbf{X}))$

α和β是可训练的参数， α初始化为1， β初始化为0，I_T是单位矩阵

①.2 Shaped Attention $\mathbf{A}(\mathbf{X}) \leftarrow (\alpha\mathbf{I}_T + \beta\mathbf{A}(\mathbf{X}) - \gamma C).$

α，β，γ是可训练的参数， 初始化设为1；C是常量 = 初始化时候的A



[1] He B, Martens J, Zhang G, et al. Deep transformers without shortcuts: Modifying self-attention for faithful signal propagation[J]. arXiv preprint arXiv:2302.10322, 2023.
[2] NOCI L, LI C, LI M, et al. The Shaped Transformer: Attention Models in the Infinite Depth-and-Width Limit[J]. 2023.

04 Simplifying Blocks：删除Attention子块的skip connection

$$\mathbf{A}(\mathbf{X}) = \mathrm{Softmax}\left(\frac{1}{\sqrt{d_k}}\mathbf{X}\mathbf{W}^Q{\mathbf{W}^K}^\top\mathbf{X}^\top + \mathbf{M}\right),$$

- **起点：Skipless模型**

② 显式降低MLP分支的权重

$$\mathbf{X}_{\mathrm{out}} = \alpha_{\mathrm{FF}}\,\hat{\mathbf{X}} + \beta_{\mathrm{FF}}\,\mathrm{MLP}(\mathrm{Norm}(\hat{\mathbf{X}})),$$

$$\beta_{\mathrm{FF}} = O\left(\frac{1}{\sqrt{L}}\right) < 1 = \alpha_{\mathrm{FF}}.$$

$$\begin{aligned}
\mathbf{X}_{l+1} &= \mathbf{X}_l + f_l(N(\mathbf{X}_l)) + \beta g_l(N(\hat{\mathbf{X}}_l)) \\
&= \mathbf{X}_l + f_l\left(\frac{\mathbf{X}_l}{\sqrt{(1+\beta^2)l+1}}\right) + \beta g_l\left(\frac{\hat{\mathbf{X}}_l}{\sqrt{(1+\beta^2)l+2}}\right) \\
&= \mathbf{X}_0 + f_0(\mathbf{X}_0) + \beta g_0\left(\frac{\hat{\mathbf{X}}_0}{\sqrt{2}}\right) + f_1\left(\frac{\mathbf{X}_1}{\sqrt{2+\beta^2}}\right) + \beta g_1\left(\frac{\hat{\mathbf{X}}_1}{\sqrt{3+\beta^2}}\right) + \dots \\
&\quad + f_l\left(\frac{\mathbf{X}_l}{\sqrt{(1+\beta^2)l+1}}\right) + \beta g_l\left(\frac{\hat{\mathbf{X}}_l}{\sqrt{(1+\beta^2)l+2}}\right)
\end{aligned}$$

可以看到，原始的Pre-LN在初始化阶段，相当于输入 $\mathbf{X}_0$ 再加上一堆由前面各层输出组成的残差，且层越深的输出，降权幅度越大，保证了训练的稳定性。这里的 β 取值为1也没什么问题。

$$\begin{aligned}
\mathbf{X}_{l+1} &= \hat{\mathbf{X}}_l + \beta g_l(N(\hat{\mathbf{X}}_l)) \\
&= f_l(N(\mathbf{X}_l)) + \beta g_l(N(\hat{\mathbf{X}}_l)) \\
&= N(\mathbf{X}_l) + \beta g_l(\hat{\mathbf{X}}_l) \\
&= N(\mathbf{X}_l) + \beta g_l(N(\mathbf{X}_l)) \\
&= \frac{\mathbf{X}_l}{\sqrt{1+\beta^2}} + \beta g_l\left(\frac{\mathbf{X}_l}{\sqrt{1+\beta^2}}\right) \\
&= \frac{\mathbf{X}_0}{(1+\beta^2)^{(l+1)/2}} + \beta\left[\frac{g_0\left(\frac{\mathbf{X}_0}{\sqrt{1+\beta^2}}\right)}{(1+\beta^2)^{l/2}} + \frac{g_1\left(\frac{\mathbf{X}_1}{\sqrt{1+\beta^2}}\right)}{(1+\beta^2)^{(l-1)/2}} + \dots + g_l\left(\frac{\mathbf{X}_l}{\sqrt{1+\beta^2}}\right)\right]
\end{aligned}$$

可以看到，如果 β 仍然取1的话，随着层数变深，相当于对原始输入 $\mathbf{X}_0$ 做了指数级的降权，且残差部分也是对越浅的输出降权越大，这样是不利于稳定训练的。因此 β 也就是 β_{FF} 的初始化要很讲究，必须消除掉指数级的降权。论文中的 $\beta_{FF} = O(1/\sqrt{L})$ 就是一个很好的选择，将权重控制在了很接近1的范围内：

$$(1+\beta^2)^{(l+1)/2} = (1+O(1/L))^{(l+1)/2} \approx 1 + O\left(\frac{l+1}{2L}\right) \le 1 + O(1/2)$$

[1] https://zhuanlan.zhihu.com/p/677511164

04 Simplifying Blocks：删除Attention子块的skip connection

➢ 起点：Skipless模型

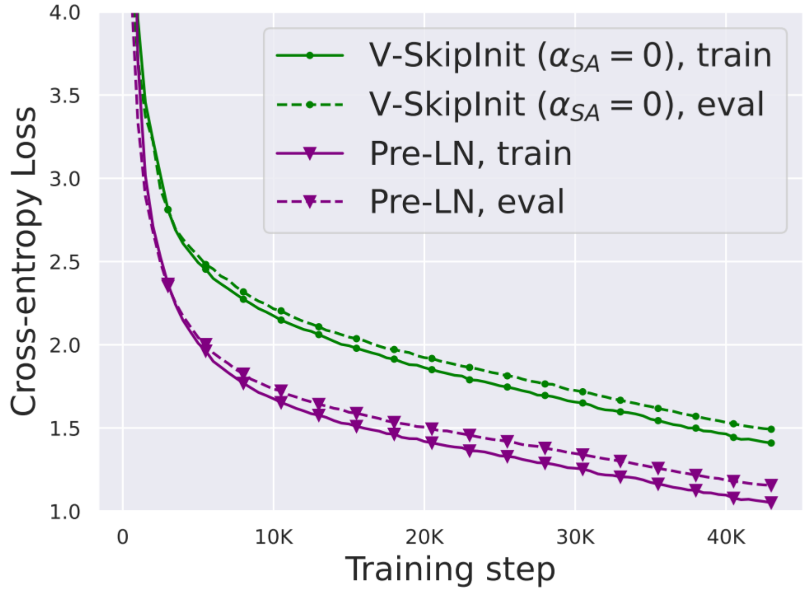
① Shaped Attention

$$\mathbf{A}(\mathbf{X}) \leftarrow (\alpha \mathbf{I}_T + \beta \mathbf{A}(\mathbf{X}) - \gamma C).$$

② 显式降低MLP分支的权重

$$\beta_{\mathrm{FF}} = O\left(\frac{1}{\sqrt{L}}\right) < 1 = \alpha_{\mathrm{FF}}.$$

③ 补回速度差距



Skipless训练慢 → Pre-LN有个隐式作用就是对残差分支做了降权

→ 论文[1]表明残差降权等价于降低学习率和缩小参数更新

→ 因此我们直接对参数降低学习率和缩小参数更新

[1] MARTENS J, BALLARD A, DESJARDINS G, et al. Rapid training of deep neural networks without skip connections or normalization layers using Deep Kernel Shaping.[J]. arXiv: Learning,arXiv: Learning, 2021.

04 Simplifying Blocks：删除Attention子块的skip connection

$$\mathrm{Attn}(\mathbf{X}) = \mathbf{A}(\mathbf{X})\mathbf{X}\boxed{\mathbf{W}^V},$$

$$\mathrm{MHA}(\mathbf{X}) = \mathrm{Concat}\big(\mathrm{Attn}_1(\mathbf{X}), \ldots, \mathrm{Attn}_H(\mathbf{X})\big)\boxed{\mathbf{W}^P}$$

- 起点：Skipless模型

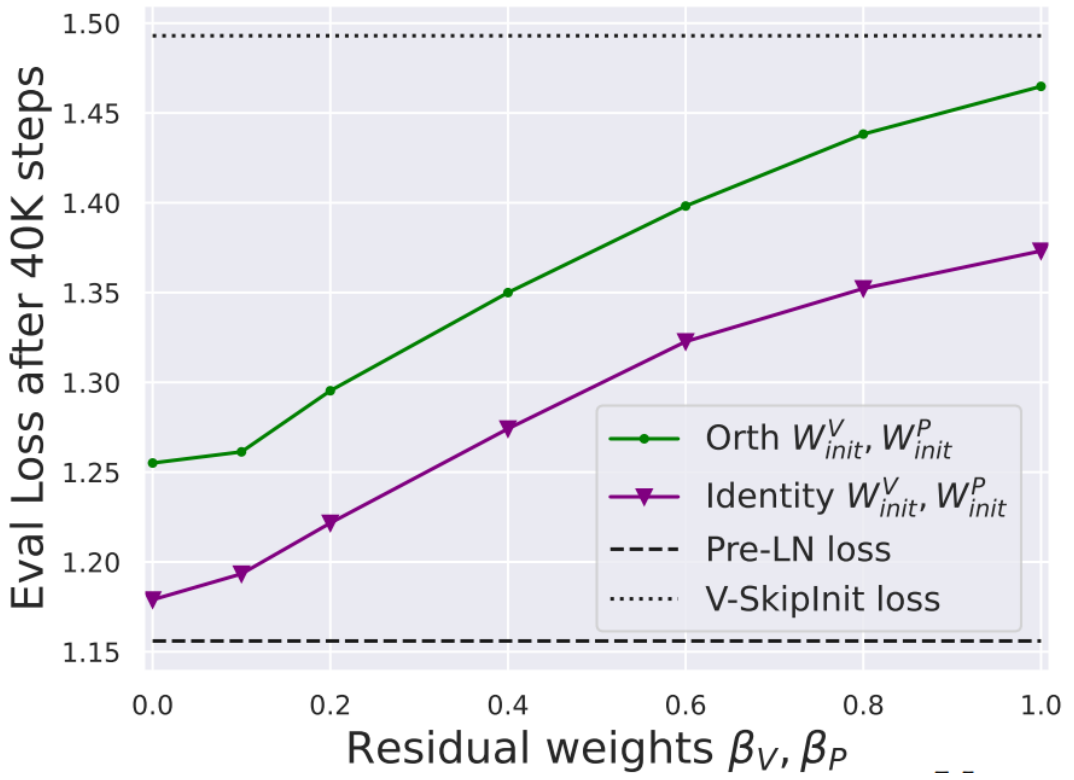
③ 补回速度差距

对参数降低学习率和缩小更新

$$\mathbf{W}^V = \alpha_V\, \mathbf{W}^V_{\text{init}} + \beta_V\, \Delta\mathbf{W}^V,$$

$$\mathbf{W}^P = \alpha_P\, \mathbf{W}^P_{\text{init}} + \beta_P\, \Delta\mathbf{W}^P,$$

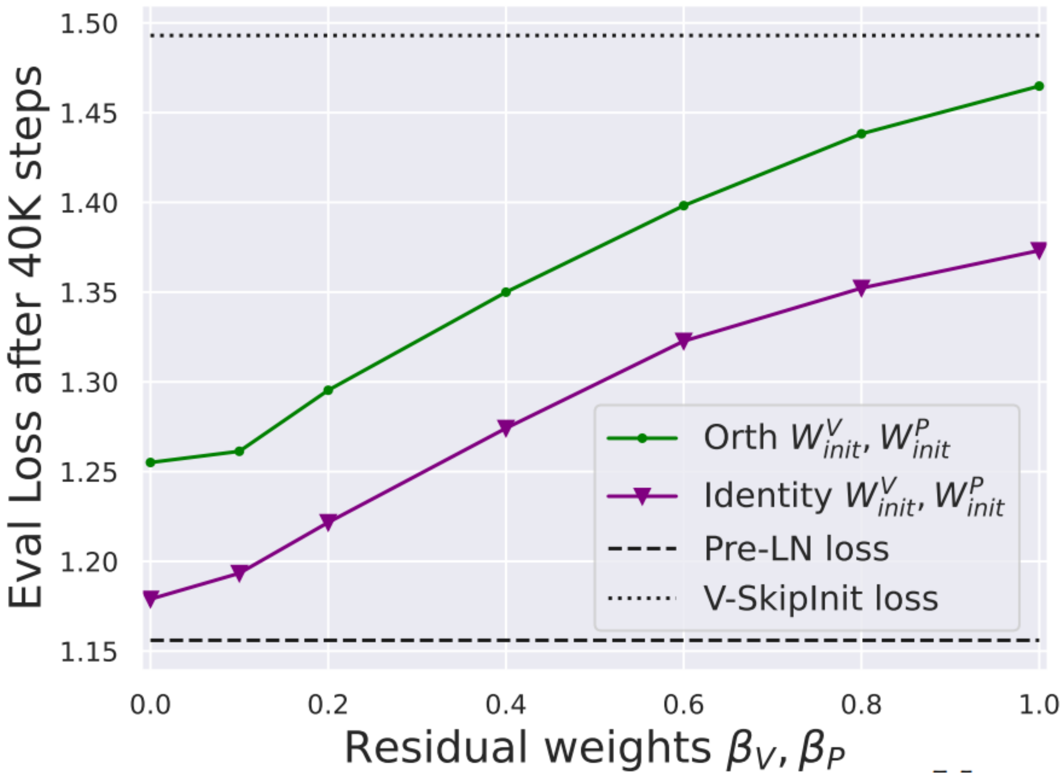
- W_{init}不更新，△W更新；
- α固定为1，β固定为$O(\frac{1}{\sqrt{L}})$的值

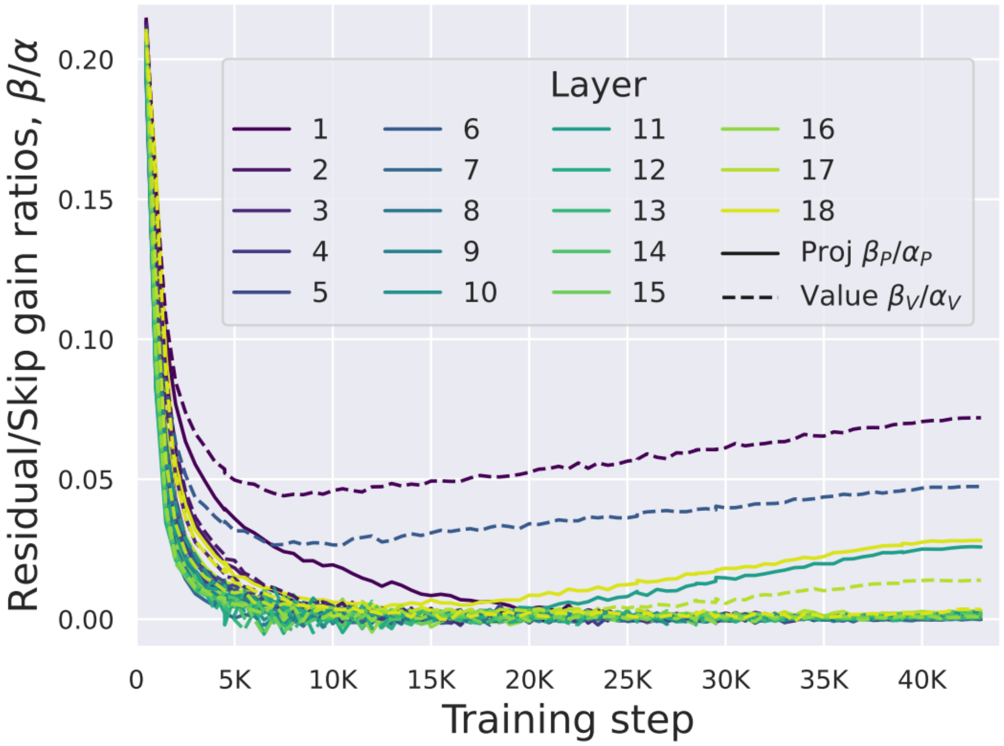


[1] MARTENS J, BALLARD A, DESJARDINS G, et al. Rapid training of deep neural networks without skip connections or normalization layers using Deep Kernel Shaping.[J]. arXiv: Learning,arXiv: Learning, 2021.

04 Simplifying Blocks：删除value和projection参数

➢ $\boldsymbol{\beta_V}$和$\boldsymbol{\beta_P}$设定为0



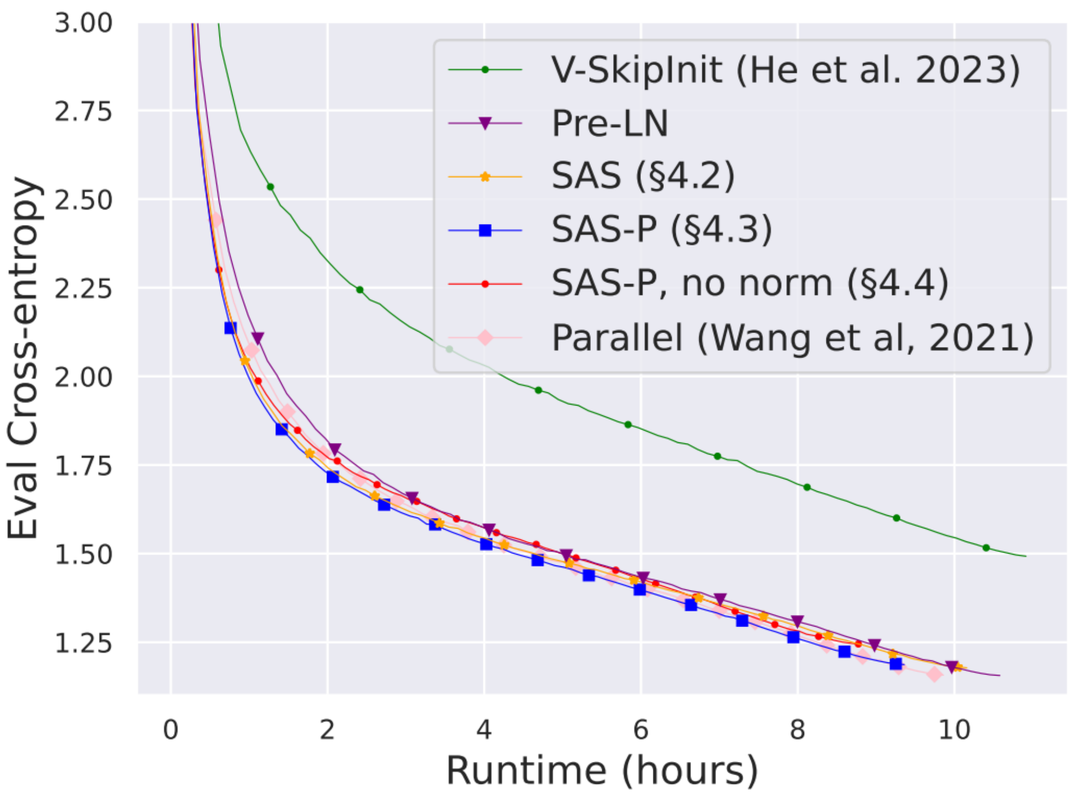


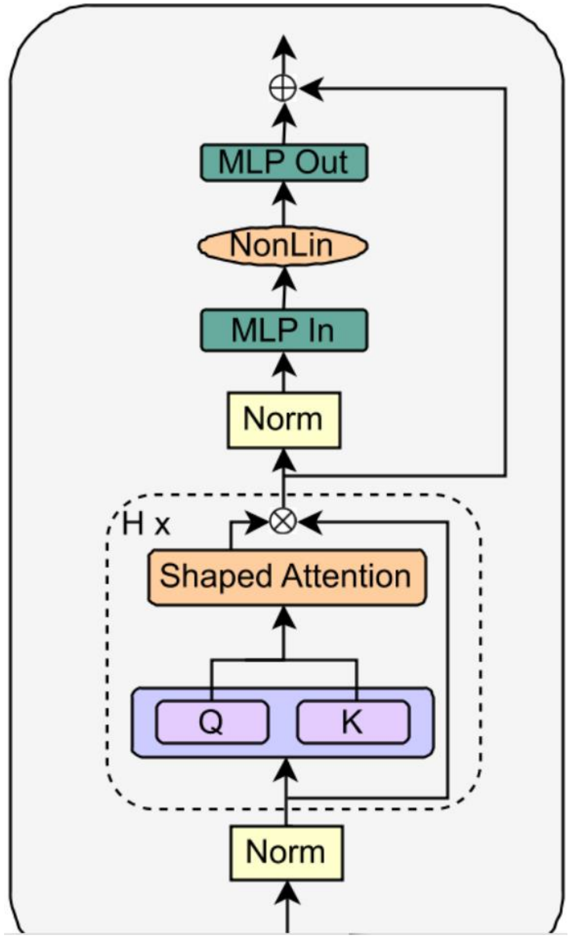
04 Simplifying Blocks：删除value和projection参数

➢ 起点：Skipless模型

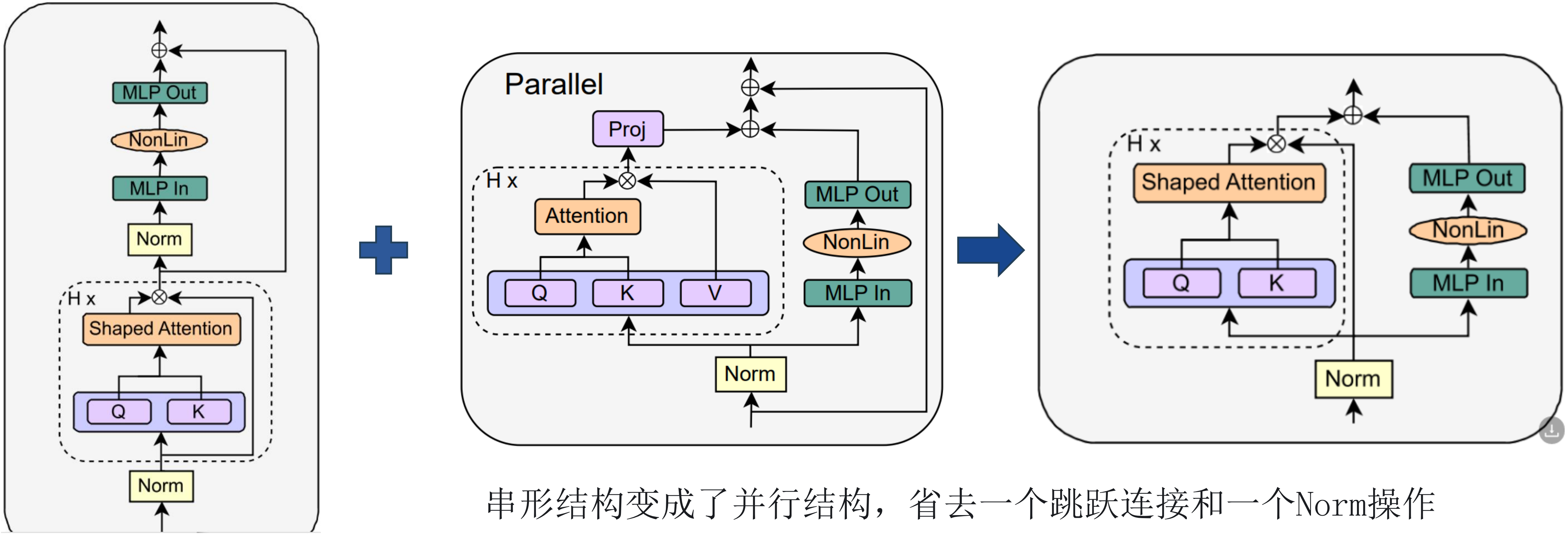
➢ Shaped Attention：$\mathbf{A}(\mathbf{X}) \leftarrow (\alpha \mathbf{I}_T + \beta \mathbf{A}(\mathbf{X}) - \gamma C).$

➢ $\boldsymbol{\beta_V}$和$\boldsymbol{\beta_P}$设定为0，参数减少了一半





04 Simplifying Blocks：删除MLP子块的skip connection



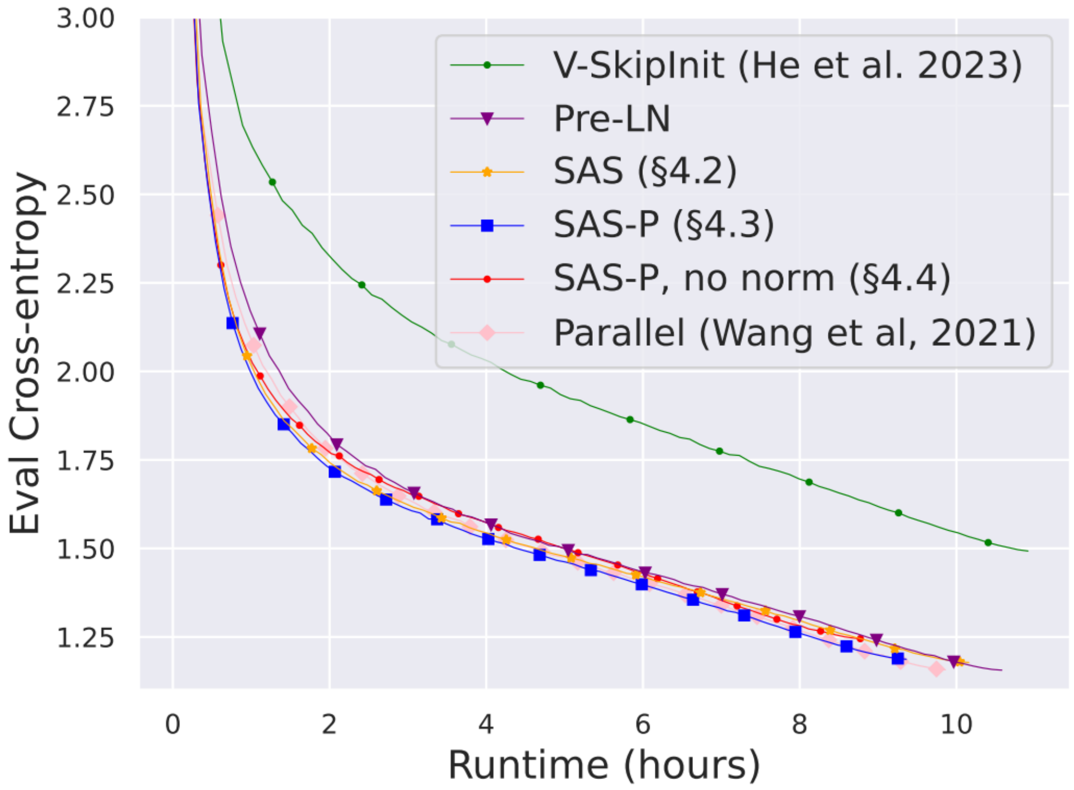
串形结构变成了并行结构，省去一个跳跃连接和一个Norm操作

提速15%

[1] Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/meshtransformer-jax, May 2021

[2] CHOWDHERY A, NARANG S, DEVLIN J, et al. PaLM: Scaling Language Modeling with Pathways[J].

04 Simplifying Blocks：删除归一化层

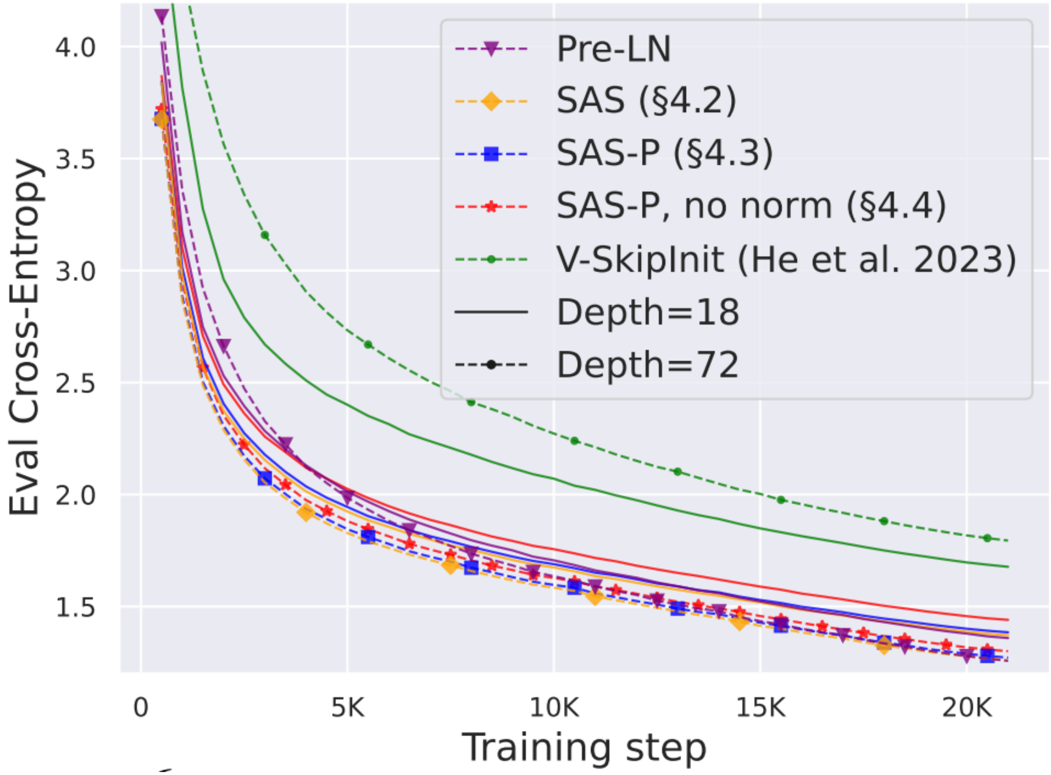


- **理论上：可删**

Norm的目的就是为了控制输入信号传播中的方差，间接做起到残差降权的作用

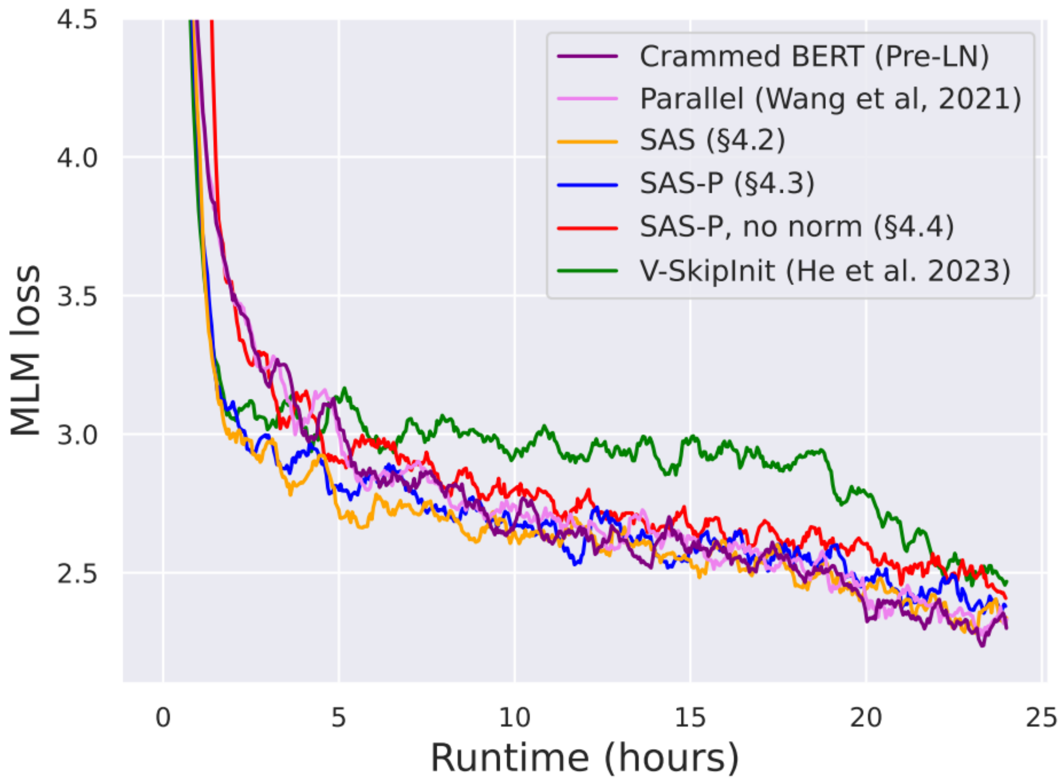
- **实验上：加速训练的作用无法解释，故保留**

05 实验分析：更大的深度



原深度：1亿多的参数量，只能算个“小模型”。
这里的实验把深度加到了72层，继续用上面的公式算出来参数量为5.5亿，也不算大，
这也正是有人质疑的地方：如果放到现在的千亿万亿级大模型上是否还凑效。

05 实验分析：推广到BERT



Block	GLUE	Params	Speed
Pre-LN (Crammed)	$78.9_{\pm.7}$	120M	1
Parallel	$78.5_{\pm.6}$	120M	1.05
V-SkipInit	$78.0_{\pm.3}$	120M	0.95
SAS (Sec. 4.2)	$78.4_{\pm.8}$	101M	1.09
SAS-P (Sec. 4.3)	$78.3_{\pm.4}$	101M	1.16
SAS-P, no norm	-	101M	1.20

参数量能节省16%，单次迭代速度快16%

谢谢观看

MANY THANKS！

24.1.23