1000层的Transformer，诞生了！

原创 卖萌酱 夕小瑶的卖萌屋 2022-03-02 14:36

卖萌屋今日学术精选

大家好，我是卖萌酱。

今天下午卖萌屋作者群里一位MILA实验室的大佬在临睡前（蒙特利尔时间凌晨0点半）甩出来一篇论文：

DeepNet: Scaling Transformers to 1,000 Layers

Hongyu Wang* Shuming Ma* Li Dong Shaohan Huang Dongdong Zhang Furu Wei†
Microsoft Research
https://github.com/microsoft/unilm

大佬表示太困了，肝不动了，于是卖萌酱左手抄起一罐咖啡，右手接过论文就开始肝了，必须第一时间分享给卖萌屋的读者小伙伴们！

论文链接：
https://arxiv.org/pdf/2203.00555.pdf

首先，把Transformer模型训深最大的问题是什么？

耗显存？

训练慢？

都不是！最大的问题是压根就不收敛啊...

所以这篇论文最关键的贡献就是提出了一种新的Normalization方式——DeepNorm，有效解决了Transformer训练困难的问题。

其实早在2019年，就有研究者针对Transformer训练困难的问题，提出了Pre-LN来提升Transformer的训练稳定性，但是随后有人发现，Pre-LN会导致模型底层的梯度比顶层的还要大，这显然是不合理的，因此往往训练出的模型效果不如传统的Post-LN。

尽管后续也有一些补丁来试图解决这些问题，但这些既有的尝试都只能让Transformer的模型深度最多训练到几百层，始终无法突破千层的天花板。

本文提出的DeepNorm，则成功打破了这个天花板。

```
def deepnorm(x):
    return LayerNorm(x * α + f(x))

def deepnorm_init(w):
    if w is ['ffn', 'v_proj', 'out_proj']:
        nn.init.xavier_normal_(w, gain=β)
    elif w is ['q_proj', 'k_proj']:
        nn.init.xavier_normal_(w, gain=1)
```

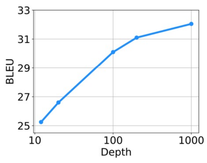
Architectures	Encoder		Decoder	
	α	β	α	β
Encoder-only (e.g., BERT)	$(2N)^{\frac{1}{4}}$	$(8N)^{-\frac{1}{4}}$	-	-
Decoder-only (e.g., GPT)	-	-	$(2M)^{\frac{1}{4}}$	$(8M)^{-\frac{1}{4}}$
Encoder-decoder (e.g., NMT, T5)	$0.81(N^4M)^{\frac{1}{16}}$	$0.87(N^4M)^{-\frac{1}{16}}$	$(3M)^{\frac{1}{4}}$	$(12M)^{-\frac{1}{4}}$

Figure 2: (a) Pseudocode for DEEPNORM. We take Xavier initialization (Glorot and Bengio, 2010) as an example, and it can be replaced with other standard initialization. Notice that α is a constant. (b) Parameters of DEEPNORM for different architectures (N-layer encoder, M-layer decoder).

DeepNorm

从以上DeepNorm伪代码实现中，可以看到这确实是simple but effective的方法，作者也给出了几个不同场景下的参数经验取值。

效果层面，作者在机器翻译benchmark上做了实验：



可以看到随着模型深度从10层到100层再到1000层，机器翻译BLEU指标持续上升。

Models	# Layers	# Params	WMT	OPUS	TED	Flores
M2M-100 (Fan et al., 2021)	48	12B	31.9	18.4	18.7	13.6
DEEPNET (ours)	200	3.2B	**33.9**	**23.0**	**20.1**	**18.6**

Table 3: BLEU scores for DEEPNET and M2M-100 on various evaluation sets.

而在与前人工作的比较上，200层的DeepNet（3.2B参数量）比Facebook M2M 48层的矮胖大模型（12B参数量）有足足5个点的BLEU值提升。

此外，作者表示将来会尝试将DeepNet往更多NLP任务上迁移（包括预训练语言模型），期待DeepNet能给NLP带来下一波春天！

上期回顾：
别再双塔了！谷歌提出DSI索引，检索效果吊打双塔，零样本超BM25！