Layer normalization 篇

来自： AiGC面试宝典

宁静致远

2023年09月29日 12:37

Layer normalization-方法篇

一、Layer Norm 篇

1.1 Layer Norm 的计算公式写一下?

$$\mu = E(X) \leftarrow \frac{1}{H}\sum_{i=1}^{H} x_i$$

$$\sigma \leftarrow Var(x) = \sqrt{\frac{1}{H}\sum_{i=1}^{H}(x_i - \mu)^2 + \epsilon}$$

$$y = \frac{x - E(x)}{\sqrt{Var(X) + \epsilon}} \cdot \gamma + \beta$$

gamma: 可训练的再缩放参数
beta: 可训练的再偏移参数

二、RMS Norm 篇 （均方根 Norm）

2.1 RMS Norm 的计算公式写一下?

$$RMS(x) = \sqrt{\frac{1}{H}\sum_{i=1}^{H} x_i^2}$$

$$x = \frac{x}{RMS(x)} \cdot \gamma$$

2.2 RMS Norm 相比于 Layer Norm 有什么特点?

RMS Norm 简化了 Layer Norm ，去除掉计算均值进行平移的部分。

对比LN，RMS Norm的计算速度更快。效果基本相当，甚至略有提升。

三、Deep Norm 篇

3.1 Deep Norm 思路?

Deep Norm方法在执行Layer Norm之前，up-scale了残差连接 (alpha>1)；另外，在初始化阶段down-scale了模型参数(beta<1)。

3.2 写一下 Deep Norm 代码实现?

```
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)
```

Deep Norm 有什么优点?

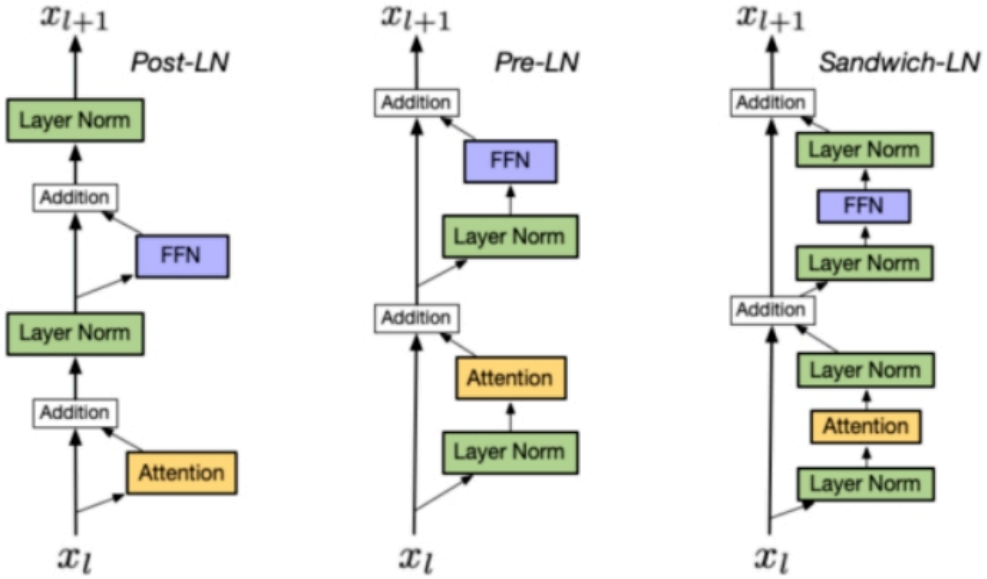
```
def deepnorm(x):
    return LayerNorm(x * α + f(x))

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)
```

Deep Norm可以缓解爆炸式模型更新的问题，把模型更新限制在常数，使得模型训练过程更稳定。

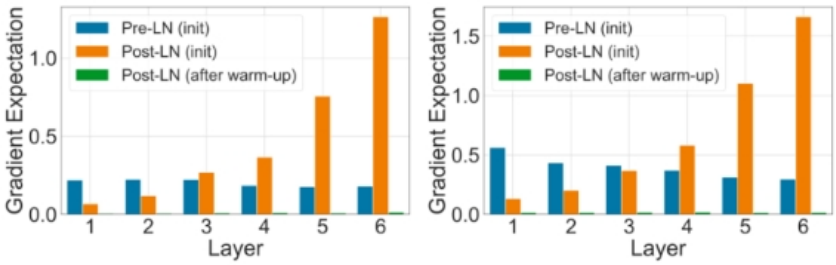
Layer normalization-位置篇

1 LN 在 LLMs 中的不同位置 有什么区别么？如果有，能介绍一下区别么？

回答：有，LN 在 LLMs 位置有以下几种：



1. Post LN：
    a. 位置：layer norm在残差链接之后
    b. 缺点：Post LN 在深层的梯度范式逐渐增大，导致使用post-LN的深层transformer容易出现训练不稳定的问题
2. Pre-LN：
    a. 位置：layer norm在残差链接中
    b. 优点：相比于Post-LN，Pre LN 在深层的梯度范式近似相等，所以使用Pre-LN的深层transformer训练更稳定，可以缓解训练不稳定问题
    c. 缺点：相比于Post-LN，Pre-LN的模型效果略差
3. Sandwich-LN：
    a. 位置：在pre-LN的基础上，额外插入了一个layer norm
    b. 优点：Cogview用来避免值爆炸的问题
    c. 缺点：训练不稳定，可能会导致训练崩溃。



(a) W^1 in the FFN sub-layers (b) W^2 in the FFN sub-layers

Layer normalization 对比篇

LLMs 各模型分别用了 哪种 Layer normalization?

模型	normalization
GPT3	Pre layer Norm
LLaMA	Pre RMS Norm
baichuan	Pre RMS Norm
ChatGLM-6B	Post Deep Norm
ChatGLM2-6B	Post RMS Norm
Bloom	Pre layer Norm
Falcon	Pre layer Norm

BLOOM在embedding层后添加layer normalization，有利于提升训练稳定性:但可能会带来很大的性能损失