Transformer相关——（6）Normalization方式

发表于 2021-08-17 | 更新于 2021-11-08 | 深度学习 | 字数总计: 1,465 | 阅读时长: 5分钟 | 阅读量: 3146

Transformer相关——（6）Normalization方式

引言

经过了残差模块后，Transformer还对残差模块输出进行了Normalization，本文对Normalization方式进行了总结，并回答为什么Transformer中选择使用Layer Normalization而不是Batch Normalization的问题。

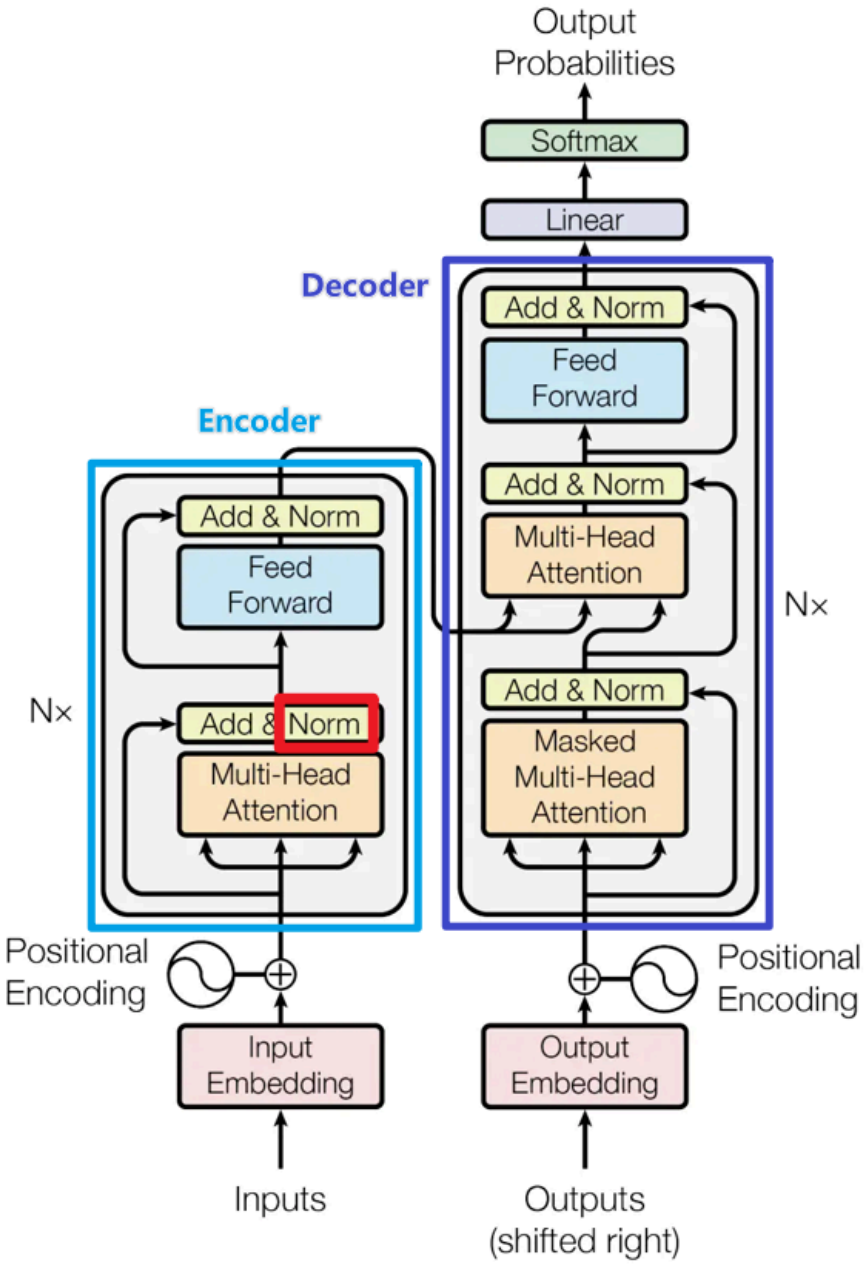


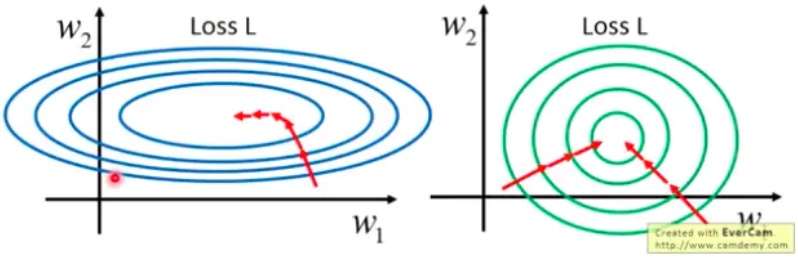
Figure 1: The Transformer - model architecture.

为什么要做Normalization?

> Normalization通过将一部分不重要的信息损失掉，以此来降低拟合难度以及过拟合的风险，从而加速模型收敛。其目的是让分布稳定下来（降低各个维度数据的方差）。

1 不同的特征具有不同数量级的数据，它们对线性组合后的结果的影响所占比重就很不相同，数量级大的特征显然影响更大。做

Normalization可以协调在特征空间上的分布，更好地进行梯度下降；



2 在神经网络中，特征经过线性组合后，还要经过激活函数，如果某个特征数量级过大，在经过激活函数时，就会提前进入它的饱和区间（比如sigmoid激活函数），即不管如何增大这个数值，它的激活函数值都在 1 附近，不会有太大变化，这样激活函数就对这个特征不敏感。 在神经网络用 SGD 等算法进行优化时，不同量纲的数据会使网络失衡，很不稳定。

Normalization方式

主要包括以下几种方法：BatchNorm（2015年）、LayerNorm（2016年）、InstanceNorm（2016年）、GroupNorm（2018年）。

> **BatchNorm**：batch方向做归一化，算NHW的均值，对小batchsize效果不好；BN主要缺点是对batchsize的大小比较敏感，由于每次计算均值和方差是在一个batch上，所以如果batchsize太小，则计算的均值、方差不足以代表整个数据分布；
>
> **LayerNorm**：channel方向做归一化，算CHW的均值，主要对RNN作用明显；
>
> **InstanceNorm**：一个channel内做归一化，算H*W的均值，用在风格化迁移；因为在图像风格化中，生成结果主要依赖于某个图像实例，所以对整个batch归一化不适合图像风格化中，因而对HW做归一化。可以加速模型收敛，并且保持每个图像实例之间的独立。
>
> **GroupNorm**：将channel方向分group，然后每个group内做归一化，算(C//G)HW的均值；这样与batchsize无关，不受其约束。在

> batchsize<16的时候, 可以使用这种归一化。
>
> **SwitchableNorm**：将BN、LN、IN结合，赋予权重，让网络自己去学习归一化层应该使用什么方法。
>
> **Weight Standardization**：权重标准化，2019年约翰霍普金斯大学研究人员提出。

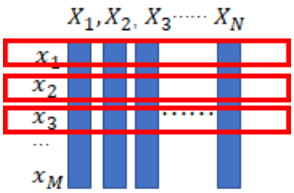
Batch Normalization-BN

针对一个Batch，在**同一维度的特征进行feature scaling**。

缺点

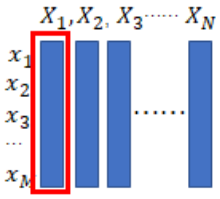
batch size较小的时候，效果差，因为其原理为用一个batch size的均值方差模拟整个数据分布的均值方差，如果batch size较小，其数据分布与整个数据分布差别较大。

在RNN中表现较差，因为RNN是逐步输入的。



Layer Normalization-LN

单独对一个样本的所有单词作缩放，与batch normalization的方向垂直，对RNN作用明显。



Instance Normalization-IN

一个batch，一个channel内做归一化。用在风格化迁移，因为在图像风格化中，生成结果主要依赖于某个图像实例，所以对整个batch归一化不适

合图像风格化中，因而对HW做归一化。可以加速模型收敛，并且保持每个图像实例之间的独立。

Group Normalization-GN

将channel方向分group，然后每个group内做归一化。与batchsize无关，不受其约束。

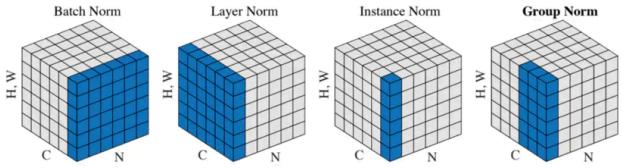


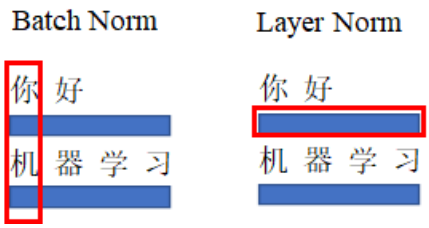
Figure 2. **Normalization methods**. Each subplot shows a feature map tensor, with N as the batch axis, C as the channel axis, and (H, W) as the spatial axes. The pixels in blue are normalized by the same mean and variance, computed by aggregating the values of these pixels.

为什么Transformer用Layer Normalization而不是Batch Normalization?

看了一些答案综合总结一下，可以从以下几个角度去解释：

1. BN是在同一维度进行归一化，但**对于一些问题来说，一个序列的输入同一"维度"上的信息可能不是同一个维度。**这么说可能有些绕，举一个NLP的例子来看：

   下面是一个batch size=2案例，按BN方式，在第一"维度"进行归一化的话就是将"你"和"机"的特征进行归一化，但这明显不是一个维度的信息。显然BN在此处使用是很不合理的。NLP中同一batch样本的信息关联不大（差异很大，但要学习的就是这种特征），更多应该概率句子内部（单个样本内部）维度的归一化。



2. 可以看作另外一个问题进行回答：为什么图像处理用batch normalization效果好，而自然语言处理用 layer normalization好?

> CV使用BN是认为不同卷积核feature map（channel维）之间的差异性很重要，LN会损失channel的差异性，对于batch内的不同样本，同一卷积核提取特征的目的性是一致的，所以使用BN仅是为了进一步保证同一个卷积核在不同样本上提取特征的稳定性。
>
> 而NLP使用LN是认为batch内不同样本同一位置token之间的差异性更重要，而embedding维，网络对于不同token提取的特征目的性是一致的，使用LN是为了进一步保证在不同token上提取的稳定性。

如何选择Normalization?

扩展总结一下如何选择Normalization方式：

取决于关注数据的哪部分信息。如果某个维度的信息差异很重要，需要被拟合，这个维度就不能进行归一化。

参考文献

Transformer架构详解

为什么要做 batch normalization

transformer 为什么使用 layer normalization，而不是其他的归一化方法？-佳雨

transformer 为什么使用 layer normalization，而不是其他的归一化方法？-Leo

各种归一化层（BatchNorm、LayerNorm、InstanceNorm、GroupNorm、Weight Standardization）及其Pytorch实现