Transformer相关——（5）残差模块

发表于 2021-08-17 | 更新于 2021-08-17 | 深度学习
| 字数总计: 1,163 | 阅读时长: 4分钟 | 阅读量: 1240

Transformer相关——（5）残差模块

引言

上一篇我们已经说完了Transformer中encoder的核心，这一篇我们来说一下mulit -head self-attention输出后为什么接了一个残差模块。我们先来看下残差模块解决了什么问题，然后再分析残差结构为什么可以解决这些问题。

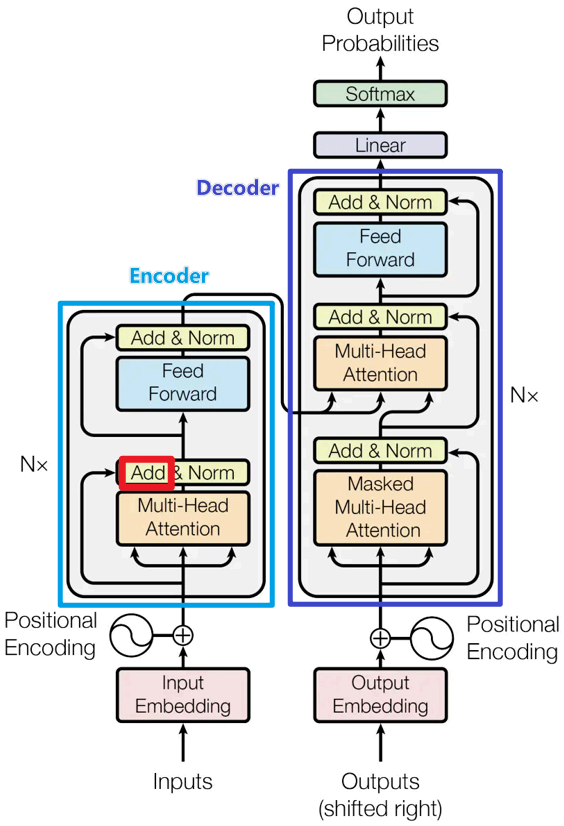

Figure 1: The Transformer - model architecture.

残差模块解决了什么问题?

1 一定程度上可以**缓解梯度弥散问题**:

> 现代神经网络一般是通过基于梯度的BP算法来优化，对前馈神经网络而言，一般需要前向传播输入信号，然后反向传播误差并使用梯度方法更新参数。
>
> 根据**链式法则**，当导数<1时，会导致反向传播中梯度逐渐消失，底层的参数不能有效更新，这也就是**梯度弥散(或梯度消失)**；当 导数>1 时，则会使得梯度以指数级速度增大，造成系统不稳定，也就是**梯度爆炸**问题。此问题可以被**标准初始化和中间层正规化方法有效控制**，这些方法使得深度神经网络可以收敛。

2 一定程度上**解决网络退化问题**:

> 在神经网络可以收敛的前提下，**随着网络深度增加，网络的表现先是逐渐增加至饱和，然后迅速下降。**
>
> 网络退化问题不是过拟合导致的，即便在模型训练过程中，同样的训练轮次下，退化的网络也比稍浅层的网络的训练错误更高，如下图所示。

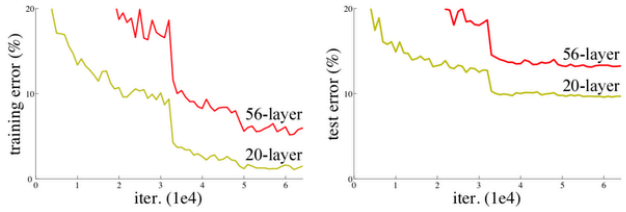

Figure 1. Training error (left) and test error (right) on CIFAR-10 with 20-layer and 56-layer "plain" networks. The deeper network has higher training error, and thus test error. Similar phenomena on ImageNet is presented in Fig. 4.

> 如果存在某个 K层的网络是当前最优的网络，那么可以构造一个更深的网络，其最后几层仅是该网络f第K 层输出的恒等映射(Identity Mapping)，就可以取得与 f一致的结果；也许 K还不是所谓"最佳层数"，那么更深的网络就可以取得更好的结果。总而言之，与浅层网络相比，更深的网络的表现不应该更差。因此，一个合理的猜测就是，**对神经网络来说，恒等映射并不容易拟合。**

3 一定程度上**缓解梯度破碎问题**：

> **在标准前馈神经网络中，随着深度增加，梯度逐渐呈现为白噪声(white noise)。**许多优化方法假设梯度在相邻点上是相似的，破碎的梯度会大大减小这类优化方法的有效性。另外，如果梯度表现得像白噪声，那么某个神经元对网络输出的影响将会很不稳定。

为什么残差模块可以解决这些问题?

残差模块的结构

残差网络通过加入 ，变得更加容易被优化。包含一个 shortcut connection 的几层网络被称为一个残差块（residual block）。

一个残差块（shortcut connections/skip connections）分为直接映射部分（x_l）和残差部分$F(x_l, W_l)$，可以表示为：

$$x_{l+1} = x_l + F(x_l, W_l)$$

示意图如下图所示：

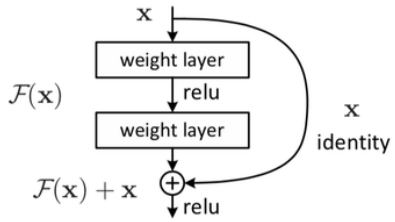

Figure 2. Residual learning: a building block.

残差模块如何解决上述问题?

1. 根据后向传播的链式法则可以看到，上述残差块中，因为增加了x**项(恒等映射)**，那么该网络求 x 的偏导的时候，**多了一项常数 1，所以反向传播过程，梯度连乘，也不会造成梯度消失。**

> **根据后向传播的链式法则，**
>
> $$\frac{\partial L}{\partial X_{Aout}} = \frac{\partial L}{\partial X_{Din}} \frac{\partial X_{Din}}{\partial X_{Aout}}$$
>
> **而** $X_{Din} = X_{Aout} + C(B(X_{Aout}))$
>
> **所以:**
>
> $$\frac{\partial L}{\partial X_{Aout}} = \frac{\partial L}{\partial X_{Din}} [1 + \frac{\partial X_{Din}}{\partial X_C} \frac{\partial X_C}{\partial X_B} \frac{\partial X_B}{\partial X_{Aout}}]$$

2

在前向传播时，**输入信号可以从任意低层直接传播到高层。由于包含了一个天然的恒等映射，一定程度上可以解决网络退化问题。**

3

The Shattered Gradients Problem: If resnets are the answer, then what is the question? 一文中提到在标准前馈神经网络中，随着深度增加，**神经元梯度的相关性(corelation)按指数级减少**（$\frac{1}{2^L}$）；同时，**梯度的空间结构也随着深度增加被逐渐消除**。这也就是梯度破碎现象。

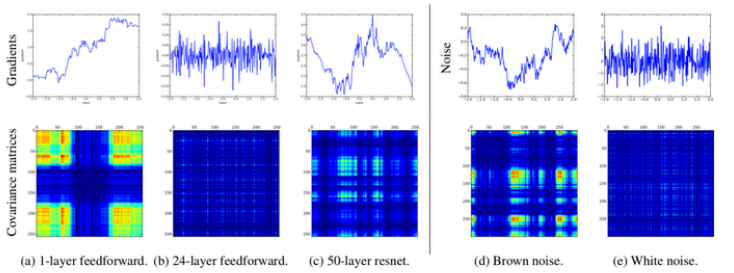

Figure 1: **Comparison between noise and gradients of rectifier nets with 200 neurons per hidden layer.** *Columns d–e:* brown and white noise. *Columns a–c:* Gradients of neural nets plotted for inputs taken from a uniform grid. The 24-layer net uses mean-centering. The 50-layer net uses batch normalization with $\beta = 0.1$, see Eq. (2).

相较标准前馈网络，**残差网络中梯度相关性减少的速度从指数级下降到亚线性级**(sublinearly, $\frac{1}{\sqrt{L}}$)，深度残差网络中，神经元梯度介于棕色噪声与白噪声之间(参见上图中的c,d,e)；残差连接可以**极大地保留梯度的空间结构**。残差结构缓解了梯度破碎问题。

参考文献

残差网络解决了什么，为什么有效?

论文解读 | Transformer 原理深入浅出

Transformer从零详细解读(可能是你见过最通俗易懂的讲解)