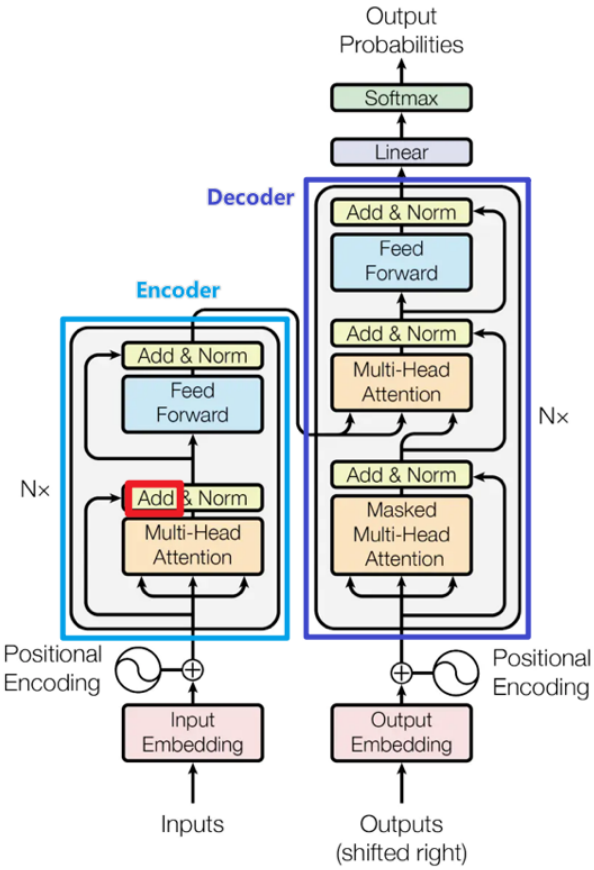
残差模块



1. 作用

- 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。

- 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。

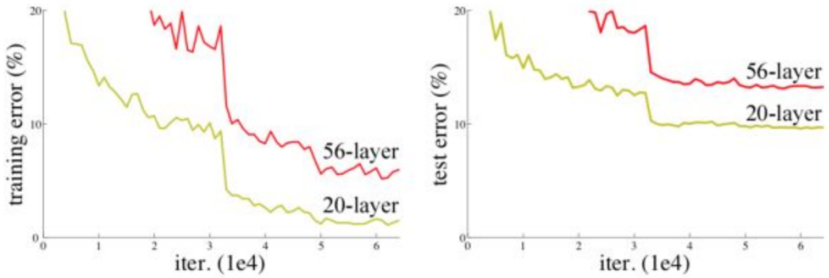


Figure 1. Training error (left) and test error (right) on CIFAR-10 with 20-layer and 56-layer "plain" networks. The deeper network has higher training error, and thus test error. Similar phenomena on ImageNet is presented in Fig. 4.

如果存在某个 K 层的网络f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。

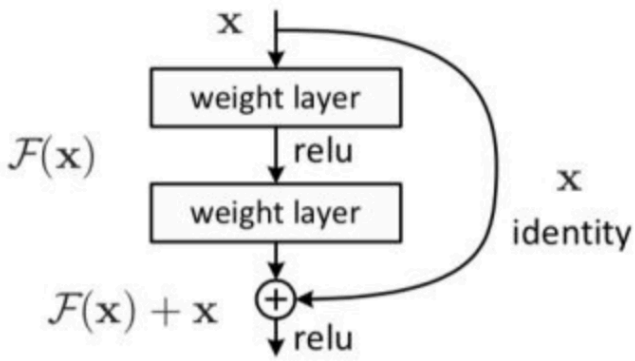
- 一定程度上**缓解梯度破碎**问题:

  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。

2. 结构

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

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



3. 原理

- 根据后向传播的链式法则可以看到，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，

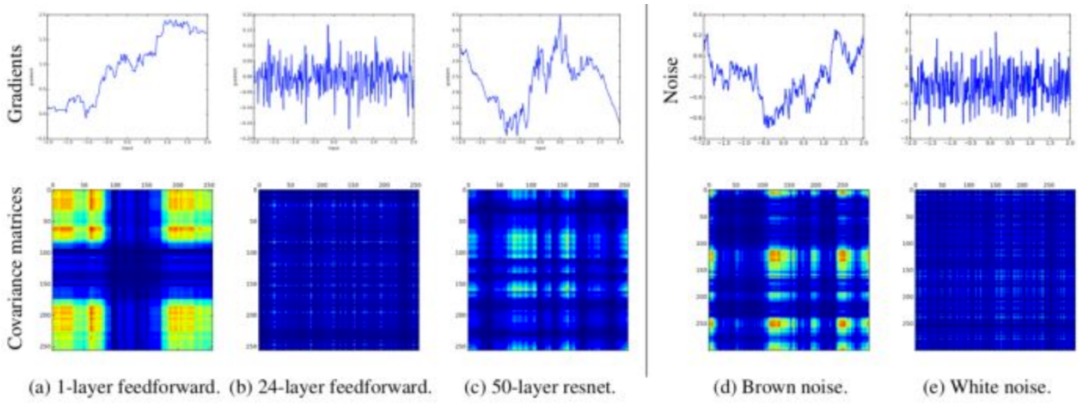
$$\frac{\partial L}{\partial X_{Aout}} = \frac{\partial L}{\partial X_{Din}} \frac{\partial X_{Din}}{\partial X_{Aout}}$$

而 $X_{Din} = X_{Aout} + C(B(X_{Aout}))$

所以:

$$\frac{\partial L}{\partial X_{Aout}} = \frac{\partial L}{\partial X_{Din}} [1 + \frac{\partial X_{Din}}{\partial X_C} \frac{\partial X_C}{\partial X_B} \frac{\partial X_B}{\partial X_{Aout}}]$$

- 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。
- 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。

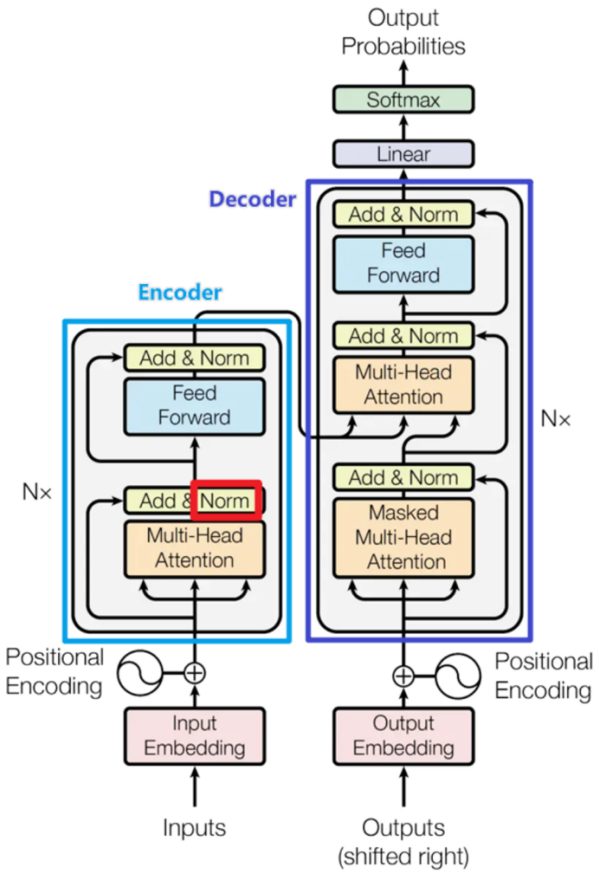


(a) 1-layer feedforward. (b) 24-layer feedforward. (c) 50-layer resnet. (d) Brown noise. (e) White noise.

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$(\frac{1}{\sqrt{L}})$。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。

Normalization



1. 原因

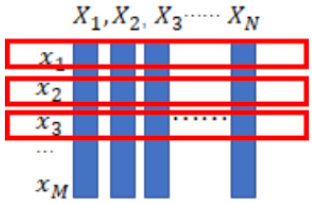
- Normalization 通过将一部分不重要的信息损失掉，以此来降低拟合难度以及过拟合的风险，从而加速模型收敛。其目的是让分布稳定下来（降低各个维度数据的方差）。
  a) 不同的特征具有不同数量级的数据，它们对线性组合后的结果的影响所占比重就很不相同，数量级大的特征显然影响更大。做 Normalization 可以协调在特征空间上的分布，更好地进行梯度下降。
  b) 在神经网络中，特征经过线性组合后，还要经过激活函数，如果某个特征数量级过大，在经过激活函数时，就会提前进入它的饱和区间（比如 sigmoid 激活函数），即不管如何增大这个数值，它的激活函数值都在 1 附近，不会有太大变化，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。

2. 方式

- 主要包括以下几种方法 :BatchNorm（2015 年）、LayerNorm（2016 年）、InstanceNorm（2016 年）、GroupNorm （2018 年）。
  a) 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。
     - 针对一个 batch，**在同一维度的特征进行 feature scaling**。
     - 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

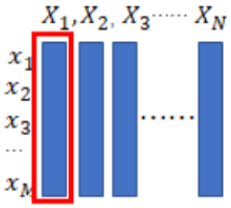
差模拟整个数据分布的均值方差，如果 batch size 较小，其数据分布与整个数据分布差别较大。

- 在 RNN 中表现较差，因为 RNN 是逐步输入的。



b) LayerNorm: channel 方向做归一化，算 CHW 的均值，主要对 RNN 作用明显。

- 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。



c) 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。

d) GroupNorm：将 channel 方向分 group，然后每个 group 内做归一化，算(Cl/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。

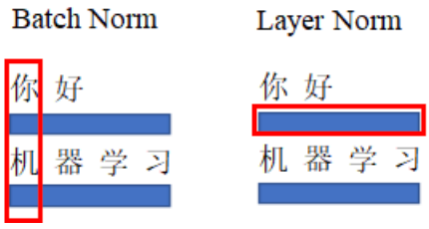
e) SwitchableNorm：将 BN、 LN、IN 结合，赋予权重，让网络自己去学习归一化层应该使用什么方法。

f) Weight Standardization：权重标准化，2019 年约翰霍普金斯大学研究人员提出。

3. 为什么 Transformer 用 Layer Normalization 而不是 Batch Normalization？

- 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。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。



- 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。

4. 如何选择？

- 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。