《Identity Mappings in Deep Residual Networks》

摘要：

1. **背景：**深度残差网络作为一种极深的网络框架，在精度和收敛等方面都展现出了很好的特性。（ResNet 论文发表后，受到深度学习界学者的肯定）
2. **本文的目的：**由于在 2015_ResNet 中，并没有对残差块(residual building blocks)为什么是**“恒等映射”**的构造做出具体的原因分析。因此，本文的研究方向就是分析残差块背后的计算传播方式，并进行“消融”实验来找出原因。
3. **结论：**比较发现，“恒等映射”的效果最优，并促使团队提出了一种的新的残差单元，能够使得训练变得更简单，进一步提高了网络的泛化能力。

什么是深度残差网络(ResNets)：

由很多个“残差单元”组成。每一个残差单元可以表示为：

$$y_l = h(x_l) + F(x_l, W_l) \qquad (1)$$

$$x_{l+1} = f(y_l) \qquad (2)$$

其中，x_l和x_{l+1}是第l个残差单元的输入和输出；F表示一个残差函数，例如堆积的两个3×3的卷积层；$W_l = \{W_{l,k|1\le k\le K}\}$是一组与第$l$个残差单元相关的权重(偏置项)，$K$是每一个残差单元中的层数，例如$K$为 2 或 3； $h(x_l) = x_l$代表一个恒等映射，$f(y_l)$为$Relu$激活函数。（以上说明，皆参考前述论文）

关于残差单元(Residual Units)的分析：（以2015_ResNet中的残差单元结构为例）

如果f也是一个恒等映射，将(2)式带入(1)式可得：

$$x_{l+1} = x_l + F(x_l, W_l) \qquad (3)$$

如果上式(3)应用于整个残差网络，则递归可得——对于任意深的单元L和任意浅的单元l：

$$x_L = x_l + \sum_{i=l}^{L-1} F(x_i, W_i) \qquad (4)$$

而上式(4)表明：(i) 对于任意深的单元L的特征x_L都可以表达为浅层单元l的特征x_l加上一个形如$\sum_{i=l}^{L-1} F$的残差函数，即**任意单元L和l之间都具有残差特性**。(ii) 对于任意深的单元L，它的特征 $x_L = x_0 + \sum_{i=0}^{L-1} F(x_i, W_i)$，即为之前所有残差函数输出的总和(加上$x_0$)。而正好相反的是，“plain network”中的特征x_L是一系列矩阵向量的乘积，即$\prod_{i=0}^{L-1} W_i x_0$ (忽略 BN 和 ReLU)

此外，(4)还具有良好的反向传播能力。假设损失函数为E，则从反向传播的链式法则中可得：

$$\frac{\partial E}{\partial x_l}=\frac{\partial E}{\partial x_L}\frac{\partial x_L}{\partial x_l}=\frac{\partial E}{\partial x_L}\left(1+\frac{\partial \sum_{i=l}^{L-1}F(x_i,W_i)}{\partial x_l}\right) \qquad (5)$$

上述(5)式说明，梯度$\frac{\partial E}{\partial x_l}$可以被分解成两个部分：其中$\frac{\partial E}{\partial x_L}$直接传递信息而不涉及任何权重层，而另一部分$\frac{\partial E}{\partial x_L}(\frac{\partial \sum_{i=l}^{L-1}F(x_i,W_i)}{\partial x_l})$表示通过权重层的传递。总之，$\frac{\partial E}{\partial x_L}$保证了信息能够直接传回任意浅层$l$，且在一个 mini-batch 中梯度$\frac{\partial x_L}{\partial x_l}$不可能出现消失的情况，因为通常$\frac{\partial \sum_{i=l}^{L-1}F}{\partial x_l}$对于一个 mini-batch 总的全部样本不可能都为-1。

关于恒等跳跃连接(Identity Skip Connections)的重要性分析：（

假设$h(x_l)=\lambda_l x_l$，则带入(3)式可得：

$$x_{l+1}=\lambda_l x_l+F(x_l,W_l) \qquad (6)$$

其中，λ_l是一个调节标量(为了简单起见，仍然假设f是恒等映射)。则通过方程的递归，可得类似(4)的公式：

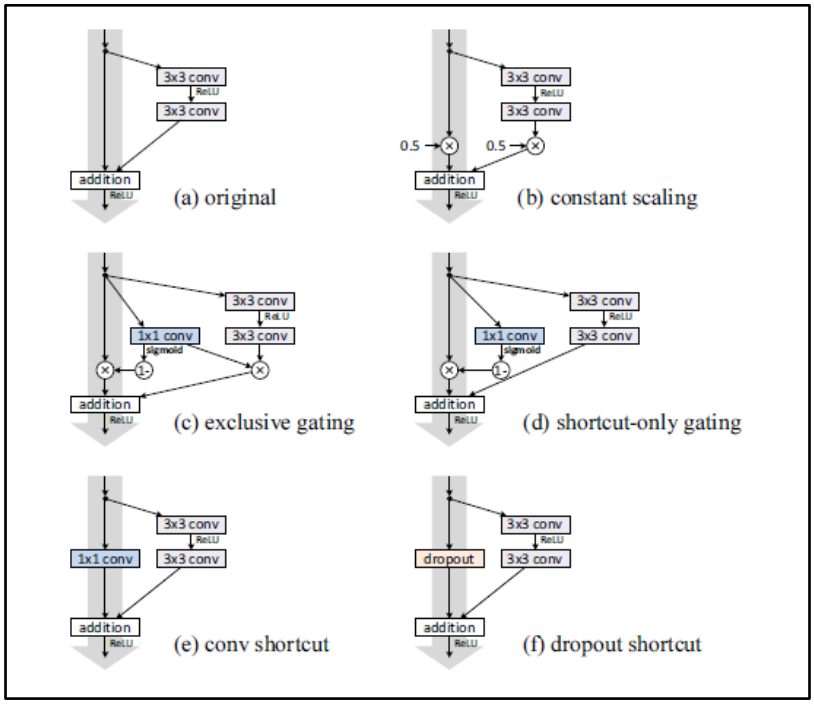
$$x_L=(\prod_{i=l}^{L-1}\lambda_i)x_l+\sum_{i=l}^{L-1}\hat{F}(x_i,W_i) \qquad (6)$$

其中，$\hat{F}$代表将这些标量合并到残差函数中。而类似(5)的反向传播过程可表示为：

$$\frac{\partial E}{\partial x_l}=\frac{\partial E}{\partial x_L}\left((\prod_{i=l}^{L-1}\lambda_i)+\frac{\partial \sum_{i=l}^{L-1}\hat{F}(x_i,W_i)}{\partial x_l}\right) \qquad (7)$$

而从(7)式可以看出：对于一个极深的网络(L很大)，如果对于所有的i都有$\lambda_i>1$，那么$\prod_{i=l}^{L-1}\lambda_i$将会呈指数型放大（**即出现梯度爆炸情况**）；如果$\lambda_i<1$，那么$\prod_{i=l}^{L-1}\lambda_i$将会呈指数型缩小或者消失（**即出现梯度消失情况**），从而阻断从跳跃连接传来的信号，并迫使它流向权重层。

关于“跳跃连接”的实验数据结论：



case	Fig.	on shortcut	on $\mathcal{F}$	error (%)	remark
original [1]	Fig. 2(a)	1	1	**6.61**	
constant scaling	Fig. 2(b)	0	1	fail	This is a plain net
		0.5	1	fail	
		0.5	0.5	12.35	frozen gating
exclusive gating	Fig. 2(c)	$1-g(\mathbf{x})$	$g(\mathbf{x})$	fail	init b_g=0 to −5
		$1-g(\mathbf{x})$	$g(\mathbf{x})$	8.70	init b_g=-6
		$1-g(\mathbf{x})$	$g(\mathbf{x})$	9.81	init b_g=-7
shortcut-only gating	Fig. 2(d)	$1-g(\mathbf{x})$	1	12.86	init b_g=0
		$1-g(\mathbf{x})$	1	6.91	init b_g=-6
1×1 conv shortcut	Fig. 2(e)	1×1 conv	1	12.22	
dropout shortcut	Fig. 2(f)	dropout 0.5	1	fail	

图(a)(b)(c)(d)(e)(f)是六种不同类型的跳跃连接方式，其中**(a)是恒等映射**，灰色箭头表示信息传播的最短路径。可以看出，跳跃连接中的操作 (缩放、门控、1×1 卷积以及 dropout) 都会阻碍信息的传递，对模型的优化造成困难。【直接证明了**恒等映射**是残差单元的最好跳跃连接方式。】

关于激活函数使用位置的分析：

通过重新调整激活函数($ReLU$或BN)来使得f为一个恒等映射。在2015_ResNet 中的原始残差连接的形状如下图(a)所示——BN在每一个权重层之后使用，然后再接一个$ReLU$，且在最后的元素相加之后还有一个$ReLU$(即$f = ReLU$)。

case	Fig.	ResNet-110	ResNet-164
original Residual Unit [1]	Fig. 4(a)	6.61	5.93
BN after addition	Fig. 4(b)	8.17	6.50
ReLU before addition	Fig. 4(c)	7.84	6.14
ReLU-only pre-activation	Fig. 4(d)	6.71	5.91
full pre-activation	Fig. 4(e)	**6.37**	**5.46**

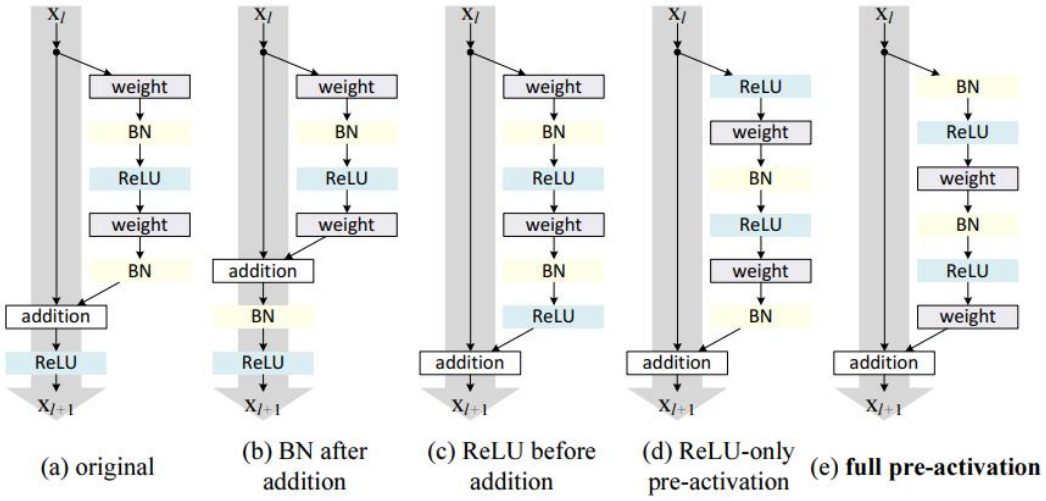


Figure 4. Various usages of activation in Table 2. All these units consist of the same components — only the orders are different.

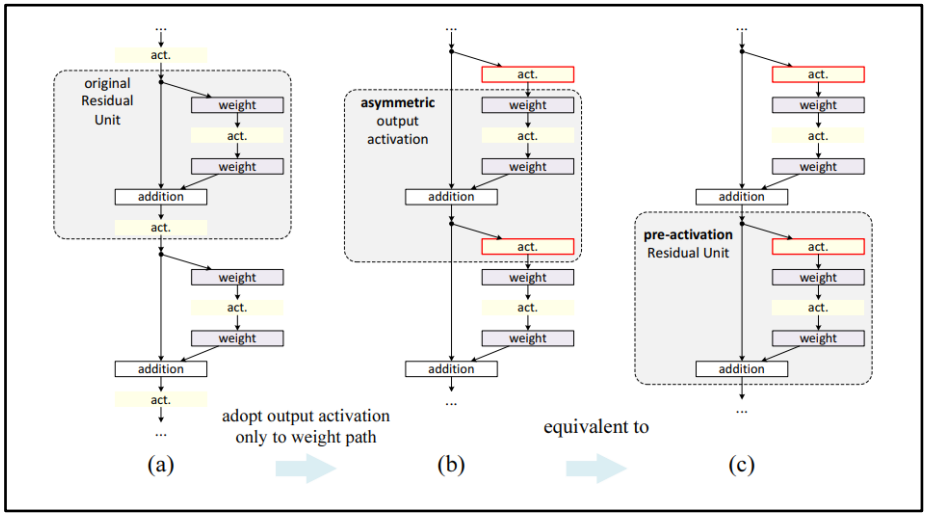
从图中可以看出，激活函数使用在不同位置时错误率存在差别，其中效果最好的是 full pre-activation（全前置激活）。

原因：

激活函数$x_{l+1} = f(y_l)$在两条路线上都能对下一个残差单元造成影响：$y_{l+1} = f(y_l) + F(f(y_l), W_l)$。而如果假设存在一种非对称方式，使得激活函数$\hat{f}$对于任意的$l$，都只对$F$路径造成影响，即$y_{l+1} = y_l + F(\hat{f}(y_l), W_l)$。通过给符号重命名，我们就可以得到：

$$x_{l+1} = x_l + F(\hat{f}(x_l), W_l) \qquad (9)$$

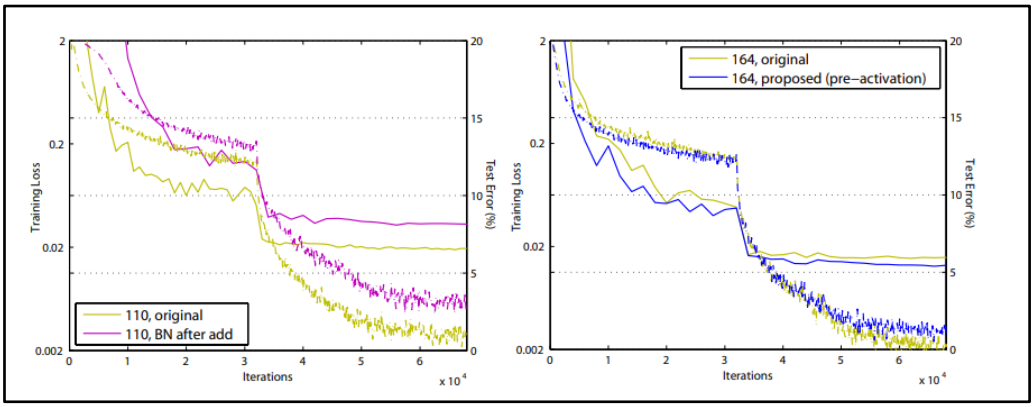
上述公式(9)表明了，如果一个新的附加激活$\hat{f}$是非对称的，那么就等同于将$\hat{f}$作为下一个残差单元的预激活(pre-activation)项，如下图所示。即不管是后激活(post-activation)还是预激活(pre-activation)，它们的区别仅是由元素级加法的存在而造成的。



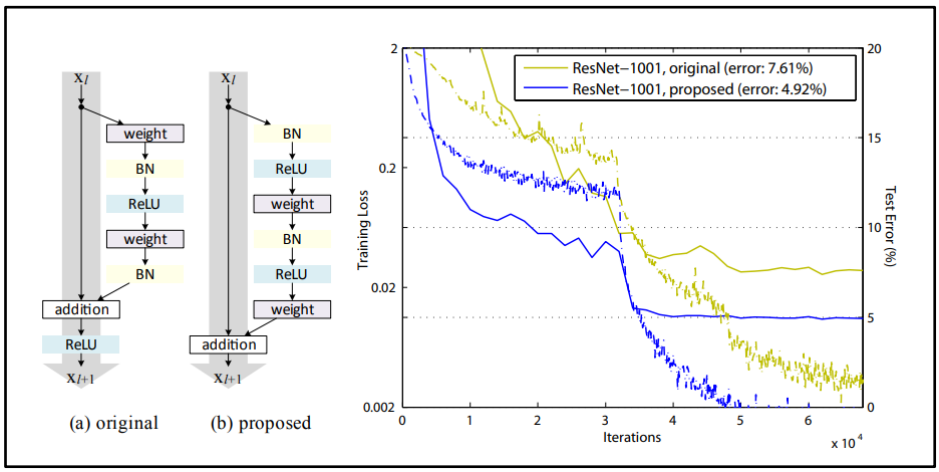
总之，预激活函数的影响有两个方面：

（1） 由于f也是恒等映射，它能使优化变得更加简单(与原始 ResNet 相比)。

（2） 在预激活中使用BN能够提高模型的正则化。



综上，本论文所提出的新的残差单元结构：



如图所示，新的残差单元(b)相比原始的残差单元(a)，区别仅是对激活函数做了前置化，但得到了更好的训练效果（效果提升巨大）。

总结：

本文主要是原作者们对自己所提出的残差单元结构的进一步改进和原因补充，使得之前论文中的 ResNet 在理论依据上更加饱满、充分，并在各项实验数据中都取得了更好的效果。