周报7 对ResNet残差网络的细研究

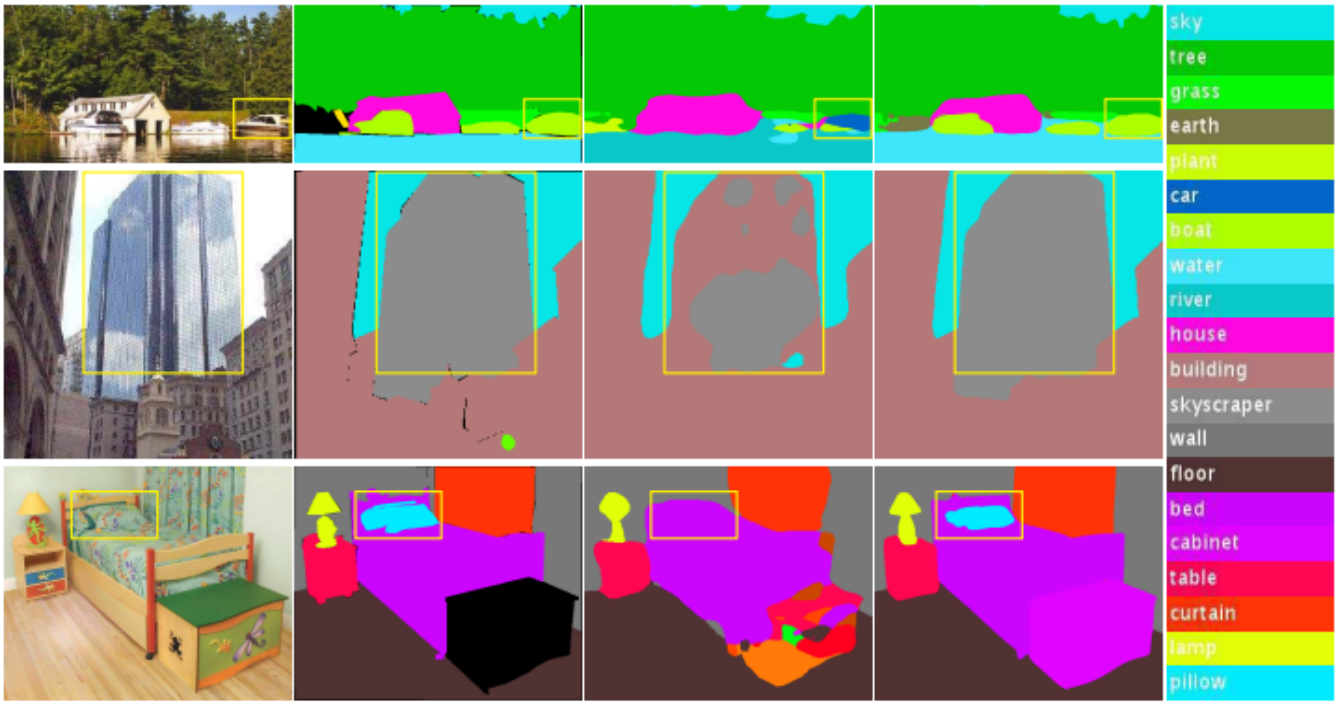
这周本来是想精读PSPNet的，但是在学习的过程中，先引入的ResNet就占用了很大一部分精力去钻研。

CVPR 2017: Pyramid Scene Parsing Network 主要内容

在阅读这篇论文时，我有了一个寻找创新突破的**思路**：作者是先通过观察之前的模型预测失败的cases（用FCN方法进行语义分割），然后分析其失败的原因，再思考如何解决这个问题，从而构建了更加优秀的PSPNet网络。因此，在我以后的科研中，如果想在某一方面进行突破时，也可以先确定一个baseline然后分析其失败的样例的内在原因或者**共性**，再想办法对其进行创新。

Important Observations

作者首先是对数据集ADE20K进行观察，其中含有150个stuff/object类别标签还有1038个场景描述。作者对其在FCN模型（baesline）训练结果进行了一些总结，总结为以下三点缺陷。



(a) Image (b) Ground Truth (c) FCN (d) PSPNet (e) ColorMap

Mismatched Relationship

在FCN的prediction results中，把上图第一行停在水面上的船识别为了汽车。从中作者总结出了FCN缺少收集语境信息的能力。在论文中有个名词一直没有懂 **co-occurrent visual patterns**，作者说“对于复杂的场景理解中语境关系是广泛且重要的，**There exist co-occurrent visual patterns**”。作者为此举了一个例子说明：飞机会大概率出现在跑道上或者在天空中滑行，而不会出现在公路上。我在网上也查找了资料，也没有对其有很好的解释说明。这里直译是“共同视觉模式”，但是我不是很理解它确切的含义。

Confusion Categories

在ADE20K dataset中存在许多迷惑性的分类。例如field and earth; mountain and hill; wall, house, building and skyscraper。它们有相同的外表。类别标注的专家有17.60%的pixel标错了。在FCN预测物体中，将上图第二行图片预测为一部分为skyscraper另一部分为building。真正的预测结果要么是skyscraper要么是building，而不是二者皆是。作者认为该问题可以利用类别之间的关系进行补救。

Inconspicuous Classes

一个场景中objects/stuff是任意大小的。一些小的东西同时又是很重要的东西例如streetlight是很难去发现的。相反一些大的物体超出了FCN的接受的范围，从而导致了不连续的预测。例如上图中的第三行没有解析出枕头。为了显著提高解析小物体或大物体的性能，应该把注意力放在包含不显著分类物体的子区域上。

作者从上述三点分析出：FCN中许多预测错误与语境关系和不同场地的全局关系有关。因此一个具有**global-scene-level**的deep network能够显著提升场景解析的性能。

Pyramid Pooling Module

针对以上的三点分析，作者提出了pyramid pooling module。在这之前，先要了解一下ResNet。

ResNet

Deep Residual Learning for Image Recognition

Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun
Microsoft Research
{kahe, v-xiangz, v-shren, jiansun}@microsoft.com

ResNet是2016年CVPR的一篇论文，这篇论文一出可以说是CV领域的里程碑，它拨开了神经网络的另一朵乌云。它也是现在非常流行的**残差网络**。

引入残差网络的原因

（当时我一直不理解为什么恒等映射不好拟合，而残差好拟合，读了论文后我理解了）

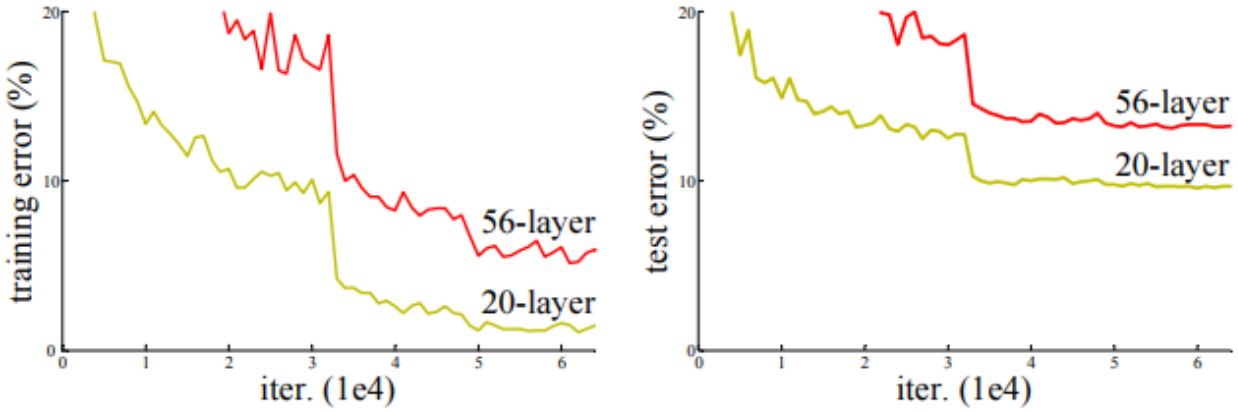


Figure 1. Training error (left) and test error (right) on CIFAR-10 with 20-layer and 56-layer "plain" networks. The deeper network has higher training error, and thus test error. Similar phenomena on ImageNet is presented in Fig. 4.

人们之前在对网络进行训练时总会遇到这样一个问题，网络层数越深，训练效果并没有预想中的那样好。后来人们发现，是因为在深度网络中，我们使用随机梯度下降的方法进行训练，由于网络层数增多，导致解空间更加复杂，因此采用SGD的方式很容易陷入局部最优解，从而导致网络进行了退化。也就是说假如目前 K 层的网络 f 是当前最优的网络，那么构造出更深的网络后，后面几层相当于只是该网络 f 第 K 层的恒等映射，因此人们发现一个问题就是**恒等映射不容易拟合。**

解决方法

对于深度残差网络，如果网络中后面的那些层已经是恒等映射，相当于就已经退化为了一个浅层网络。

因此**何凯明**他们考虑到假如神经网络非线性单元的输入和输出维度一致，可以将神经网络单元内要拟合的函数 $H(x)$ 拆分为两部分：

$$Z^l = H(a^{l-1}) = a^{l-1} + F(a^{l-1})$$

其中 $F(x)$ 是残差函数，在网络高层，学习一个恒等函数 $H(a^{l-1}) = a^{l-1}$ 等价于残差部分趋近于0，即 $F(a^{l-1}) = 0$。因此有了论文中的的shortcut结构：

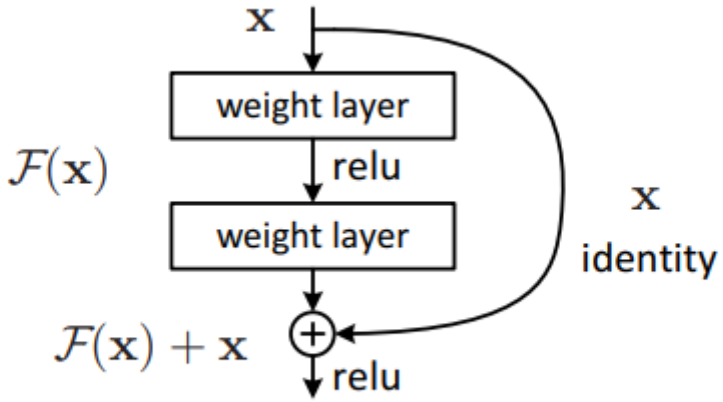


Figure 2. Residual learning: a building block.

但是此时依旧不能解决我的问题：**为什么神经网络学习残差会更加容易**?

Identity Mappings in Deep Residual Networks

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun

Microsoft Research

Abstract Deep residual networks [1] have emerged as a family of extremely deep architectures showing compelling accuracy and nice convergence behaviors. In this paper, we analyze the propagation formulations behind the residual building blocks, which suggest that the forward and backward signals can be directly propagated from one block to any other block, when using identity mappings as the skip connections and after-addition activation. A series of ablation experiments support the importance of these identity mappings. This motivates us to propose a new residual unit, which makes training easier and improves generalization. We report improved results using a 1001-layer ResNet on CIFAR-10 (4.62% error) and CIFAR-100, and a 200-layer ResNet on ImageNet. Code is available at: https://github.com/KaimingHe/resnet-1k-layers.

后来在何凯明的另一篇论文：Identity Mappings in Deep Residual Networks中找到了相应的解释

数学原理

对于不加激活函数的情况下，我们假设神经网络中任意两层 L 和 l（$L > l$），在残差网络中，我们可以得到：

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

递归地，我们可以得到：

$x_{l+2} = x_{l+1} + F(x_{l+1}, W_{l+1}) = x_l + F(x_{l+1}, W_{l+1}) + F(x_l, W_l)$

最终我们对于任意两层 L 和 l，可以有：

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

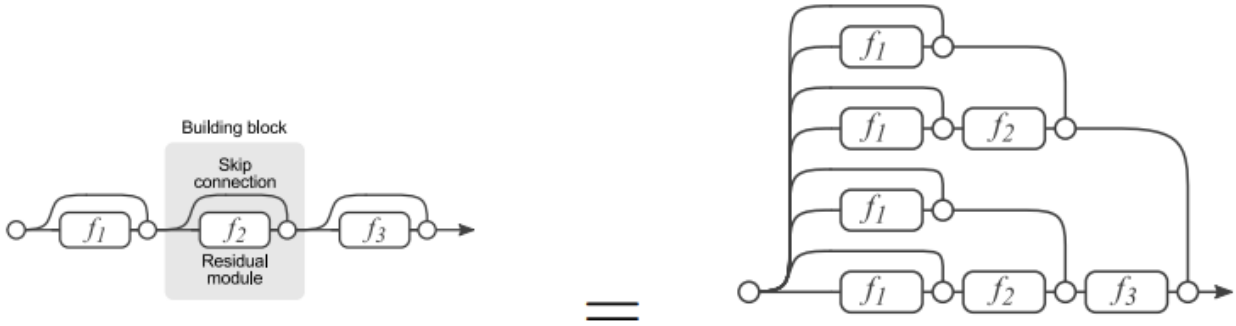
这里可以理解为**输入信号可以从任意低层直接传播到高层。**

根据链式法则，损失函数 ε 对 l 层的梯度可以展开为：

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

从上述公式可以看出，反向传播时，错误信号可以不经过任何中间权重矩阵变换直接传播到低层，一定程度上可以缓解梯度弥散问题，可以认为残差连接使得信息前后向传播更加顺畅。

后来Andreas Veit等人又从集成学习的角度上去研究了为什么残差网络效果更好。

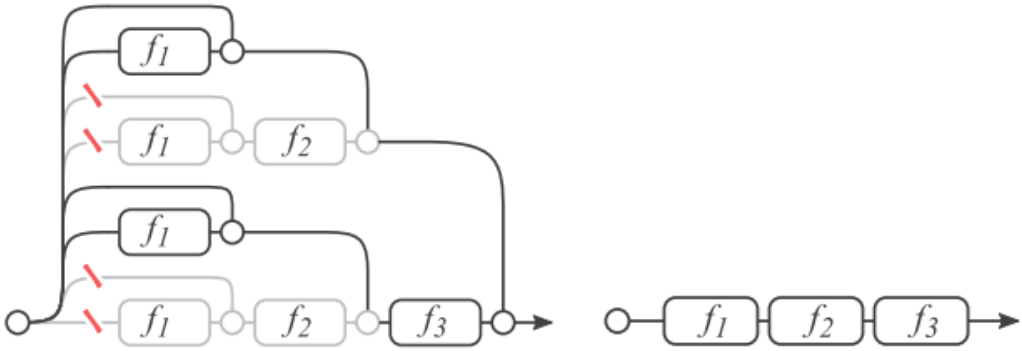


(a) Conventional 3-block residual network　　(b) Unraveled view of (a)

Figure 1: Residual Networks are conventionally shown as (a), which is a natural representation of Equation (1). When we expand this formulation to Equation (6), we obtain an *unraveled view* of a 3-block residual network (b). Circular nodes represent additions. From this view, it is apparent that residual networks have $O(2^n)$ implicit paths connecting input and output and that adding a block doubles the number of paths.

残差网络可以被看作是一系列路径集合组装而成的一个集成模型，其中不同的路径包含不同的网络层子集。作者们开展了几组实验，在测试时，删去残差网络的部分网络层或者交换某些网络模块的模块顺序，实验结果表明：在路径变化时，网络表现没有剧烈变化。

该实验结论表明：残差网络展开后的路径具有一定的独立性和冗余性，使得残差网络表现得像一个集成模型。作者还发现残差网络中主要在训练中贡献了梯度的是那些相对较短的路径。



(a) Deleting f_2 from unraveled view (b) Ordinary feedforward network

Figure 2: Deleting a layer in residual networks at test time (a) is equivalent to zeroing half of the paths. In ordinary feed-forward networks (b) such as VGG or AlexNet, deleting individual layers alters the only viable path from input to output.

这周解决了之前一个特别大的困惑，同时我还手动复现了ResNet代码，下周会继续精度PSPNet，同时进行代码复现。