一文简述ResNet及其多种变体

2018-04-22 机器之心

选自TowardsDataScience

作者：Vincent Fung

机器之心编译

参与：邹俏也、路雪

> 本文主要介绍了 ResNet 架构，简要阐述了其近期成功的原因，并介绍了一些有趣的 ResNet 变体。

在 AlexNet [1] 取得 LSVRC 2012 分类竞赛冠军之后，深度残差网络（Residual Network, 下文简写为 ResNet）[2] 可以说是过去几年中计算机视觉和深度学习领域最具开创性的工作。ResNet 使训练数百甚至数千层成为可能，且在这种情况下仍能展现出优越的性能。

因其强大的表征能力，除图像分类以外，包括目标检测和人脸识别在内的许多计算机视觉应用都得到了性能提升。

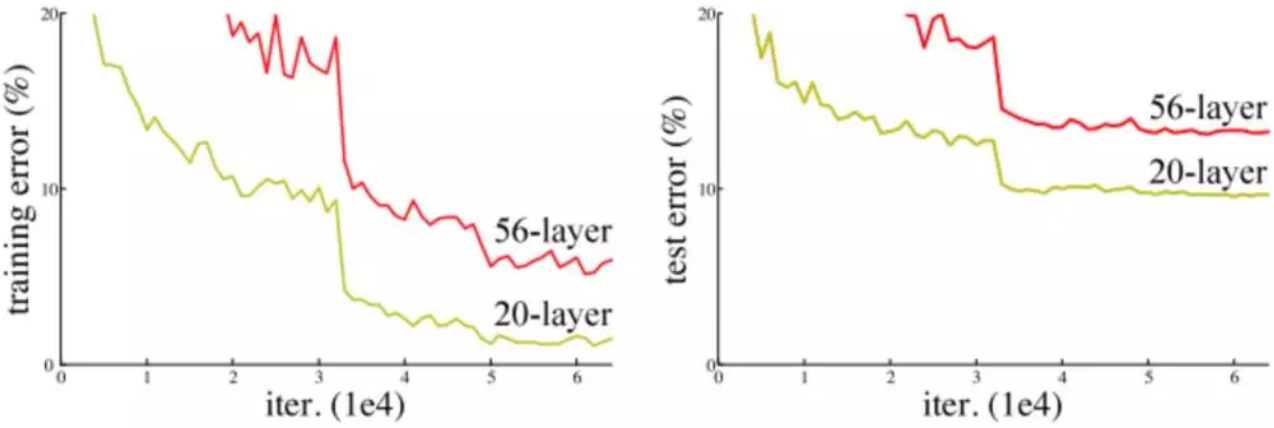
自从 2015 年 ResNet 让人们刮目相看，研究界的许多人在深入探索所其成功的秘密，许多文献中对该模型做了一些改进。本文分为两部分，第一部分为不熟悉 ResNet 的人提供一些背景知识，第二部分将介绍我最近阅读的一些论文，关于 ResNet 的不同变体和对 ResNet 架构的理解。

重新审视 ResNet

根据泛逼近定理（universal approximation theorem），只要给定足够的容量，单层的前馈网络也足以表示任何函数。但是，该层可能非常庞大，网络和数据易出现过拟合。因此，研究界普遍认为网络架构需要更多层。

自 AlexNet 以来，最先进的 CNN 架构已经越来越深。AlexNet 只有 5 个卷积层，而之后的 VGG 网络 [3] 和 GoogleNet（代号 Inception_v1）[4] 分别有 19 层和 22 层。

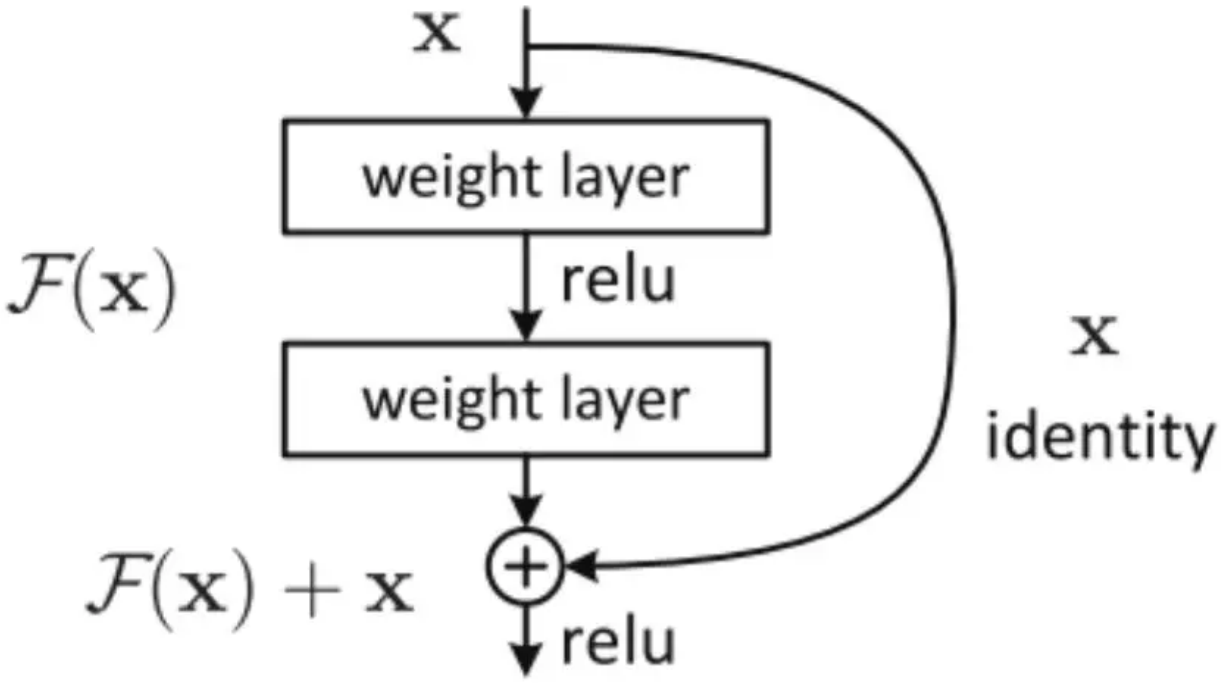
但是，网络的深度提升不能通过层与层的简单堆叠来实现。由于臭名昭著的梯度消失问题，深层网络很难训练。因为梯度反向传播到前面的层，重复相乘可能使梯度无穷小。结果就是，随着网络的层数更深，其性能趋于饱和，甚至开始迅速下降。



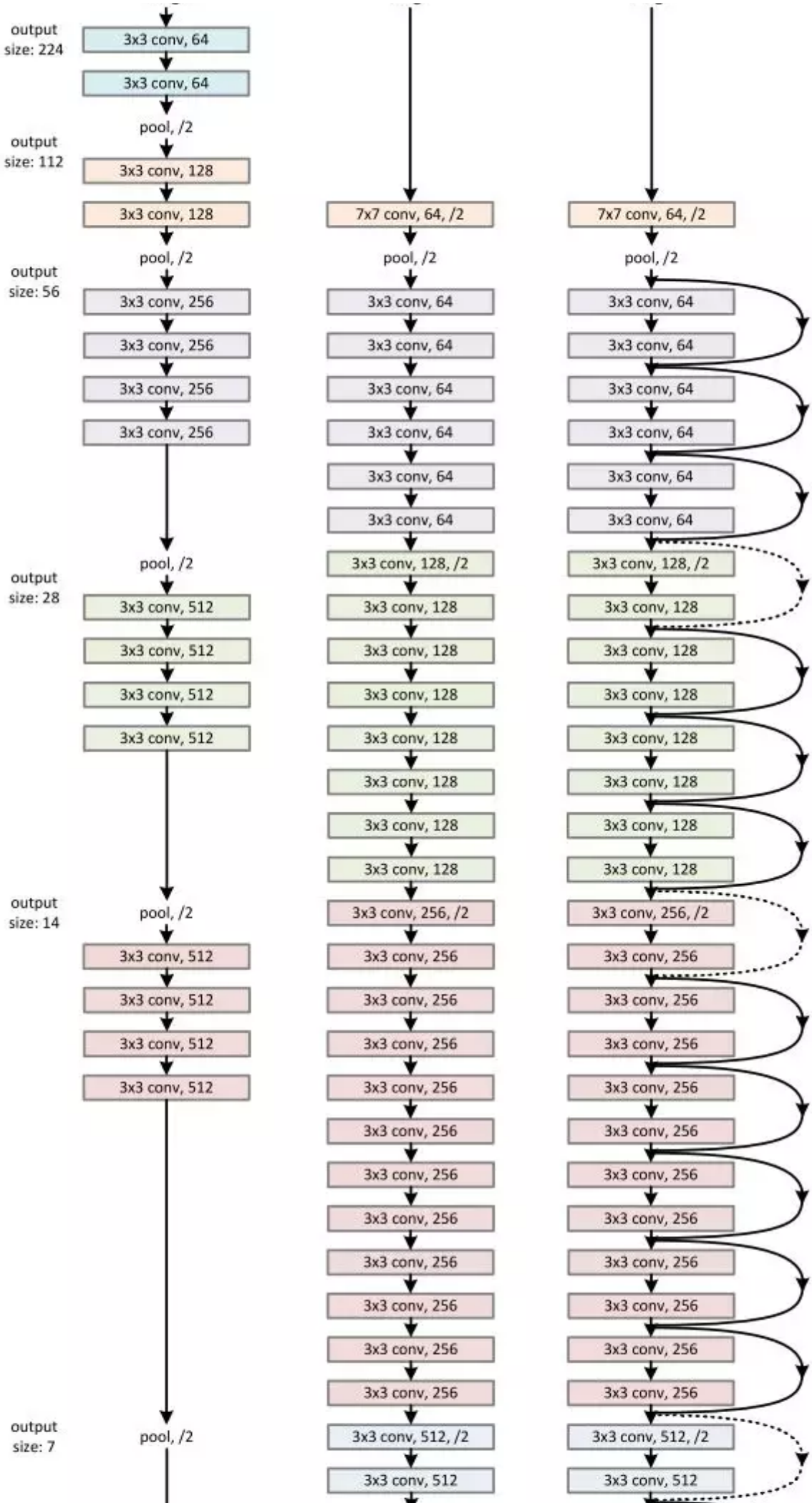
增加网络深度导致性能下降

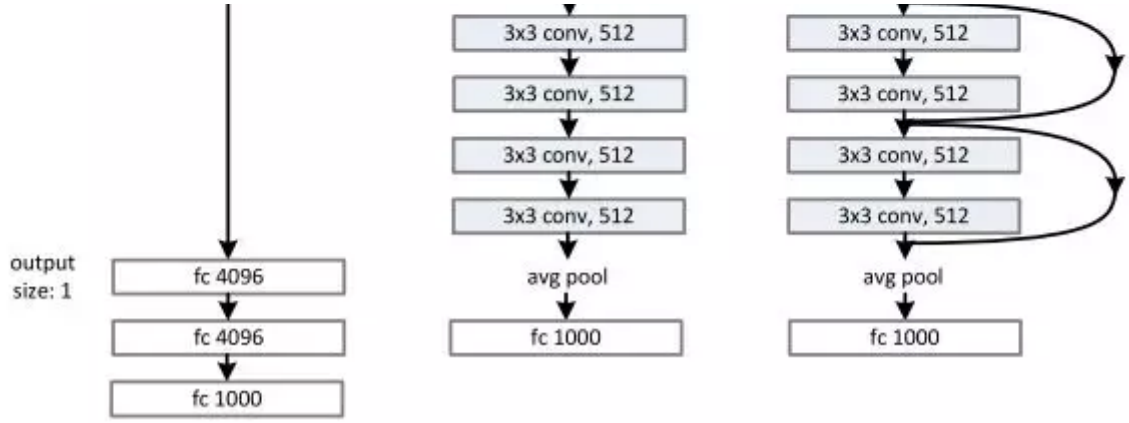
在 ResNet 出现之前有几种方法来应对梯度消失问题，例如 [4] 在中间层添加了一个辅助损失作为额外的监督，但其中没有一种方法真正解决了这个问题。

ResNet 的核心思想是引入一个所谓的「恒等快捷连接」（identity shortcut connection），直接跳过一个或多个层，如下图所示：



残差块


output size: 224
3x3 conv, 64
3x3 conv, 64
pool, /2
output size: 112
3x3 conv, 128
3x3 conv, 128
pool, /2
output size: 56
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
pool, /2
output size: 28
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
pool, /2
output size: 14
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
pool, /2
output size: 7
7x7 conv, 64, /2
pool, /2
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 128, /2
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 256, /2
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 512, /2
3x3 conv, 512
7x7 conv, 64, /2
pool, /2
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 128, /2
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 256, /2
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 512, /2
3x3 conv, 512



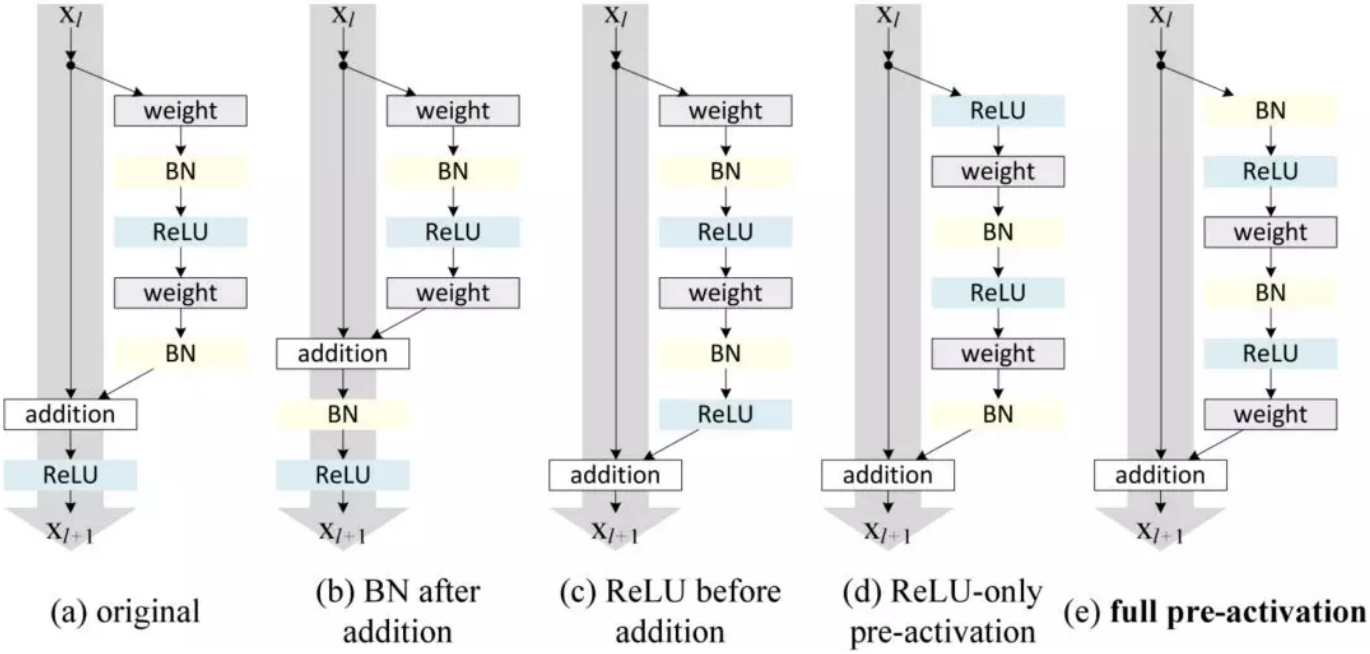
ResNet 架构

[2] 的作者认为，堆叠层不应降低网络性能，因为我们可以简单地在当前网络上堆叠恒等映射（该层不做任何事情），得到的架构将执行相同的操作。这表明较深的模型所产生的训练误差不应该比较浅的模型高。他们假设让堆叠层适应残差映射比使它们直接适应所需的底层映射要容易一些。上图中的残差块明确表明，它可以做到这一点。

事实上，ResNet 并不是第一个利用快捷连接的模型，Highway Networks [5] 就引入了门控快捷连接。这些参数化的门控制流经捷径（shortcut）的信息量。类似的想法可以在长短期记忆网络（LSTM）[6] 单元中找到，它使用参数化的遗忘门控制流向下一个时间步的信息量。ResNet 可以被认为是 Highway Network 的一种特殊情况。

然而，实验结果表明 Highway Network 的性能并不比 ResNet 好，这有点奇怪。Highway Network 的解空间包含 ResNet，因此它的性能至少应该和 ResNet 一样好。这表明，保持这些「梯度高速路」（gradient highway）的畅通比获取更大的解空间更为重要。

按照这种思路，[2] 的作者改进了残差块，并提出了一种残差块的预激活变体 [7]，梯度可以在该模型中畅通无阻地通过快速连接到达之前的任意一层。事实上，使用 [2] 中的原始残差块训练一个 1202 层的 ResNet，其性能比 110 层的模型要差。



残差块的变体

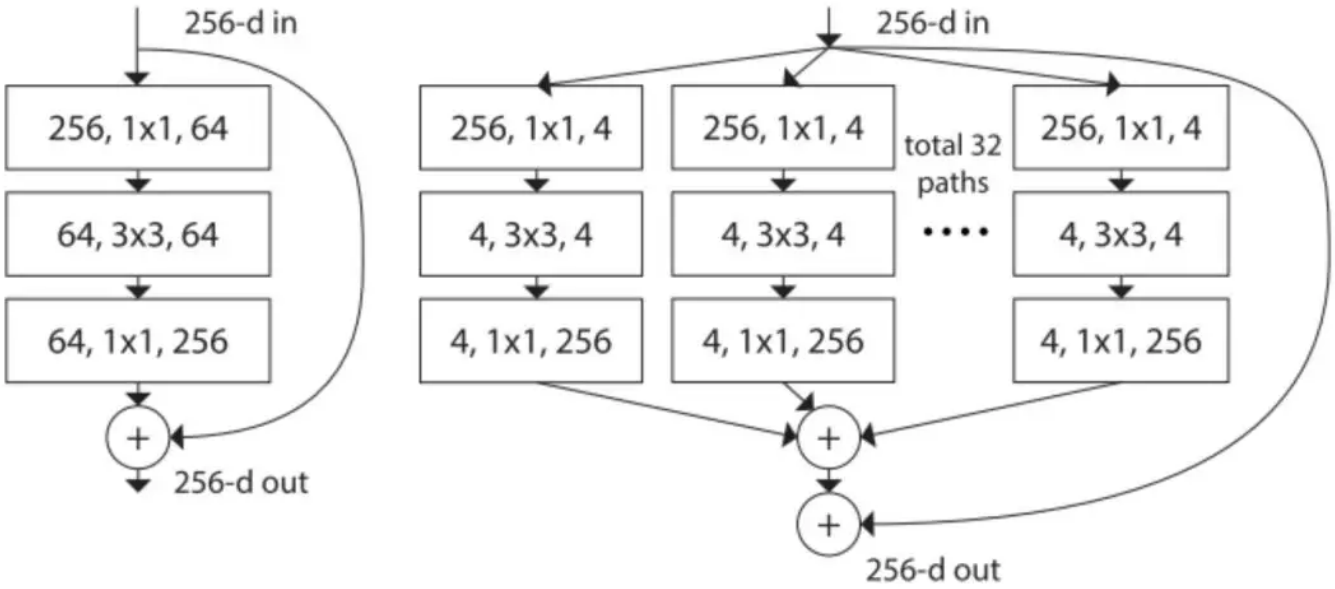
[7] 的作者在其论文中通过实验表明，他们可以训练出 1001 层的深度 ResNet，且性能超越较浅层的模型。他们的训练成果卓有成效，因而 ResNet 迅速成为多种计算机视觉任务中最流行的网络架构之一。

ResNet 的最新变体以及解读

随着 ResNet 在研究界的不断普及，关于其架构的研究也在不断深入。本节首先介绍几种基于 ResNet 的新架构，然后介绍一篇论文，从 ResNet 作为小型网络集合的角度进行解读。

ResNeXt

Xie et al. [8] 提出 ResNet 的一种变体 ResNeXt，它具备以下构建块：

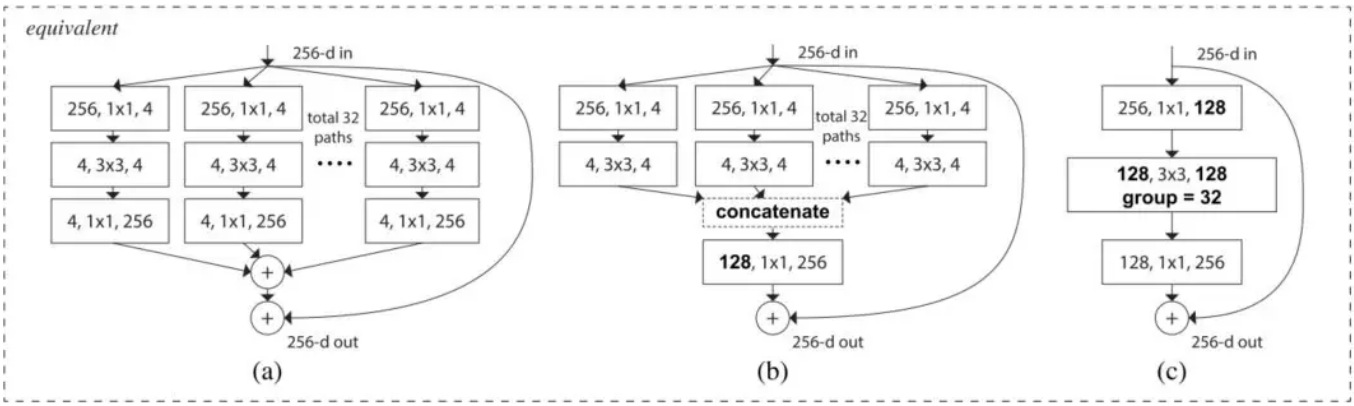


左：[2] 中 ResNet 的构建块；右：ResNeXt 的构建块，基数=32

ResNext 看起来和 [4] 中的 Inception 模块非常相似，它们都遵循了「分割-转换-合并」的范式。不过在 ResNext 中，不同路径的输出通过相加合并，而在 [4] 中它们是深度级联（depth concatenated）的。另外一个区别是，[4] 中的每一个路径互不相同（1x1、3x3 和 5x5 卷积），而在 ResNeXt 架构中，所有的路径都遵循相同的拓扑结构。

作者在论文中引入了一个叫作「基数」（cardinality）的超参数，指独立路径的数量，这提供了一种调整模型容量的新思路。实验表明，通过扩大基数值（而不是深度或宽度），准确率得到了高效提升。作者表示，与 Inception 相比，这个全新的架构更容易适应新的数据集或任务，因为它只有一个简单的范式和一个需要调整的超参数，而 Inception 需要调整很多超参数（比如每个路径的卷积层内核大小）。

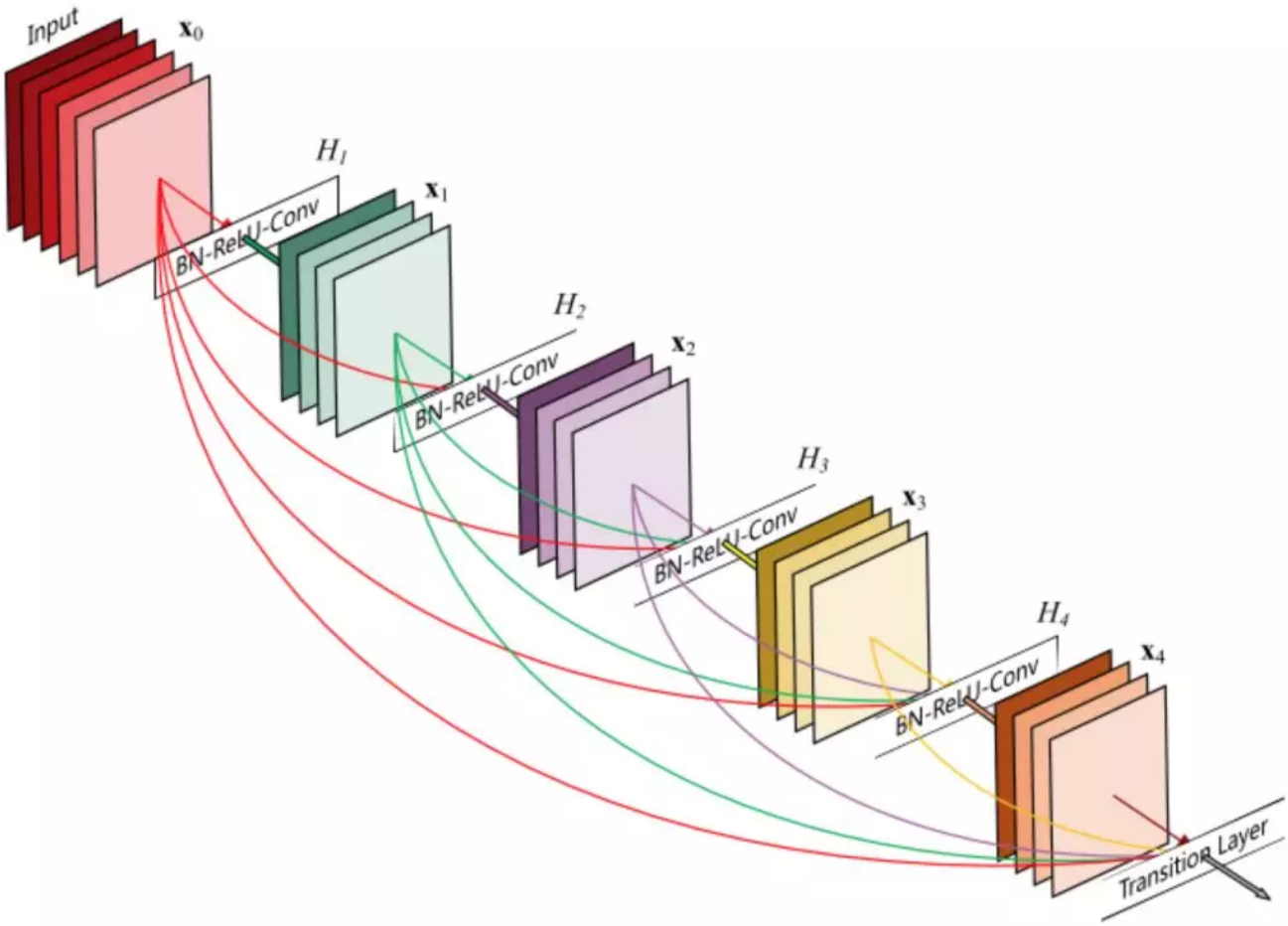
这个全新的结构有三种等价形式：



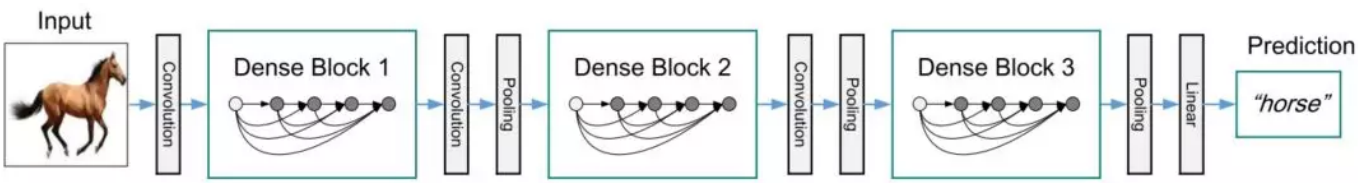
在实际操作中，「分割-变换-合并」范式通常通过「逐点分组卷积层」来完成，这个卷积层将输入的特征映射分成几组，并分别执行正常的卷积操作，其输出被深度级联，然后馈送到一个 1x1 卷积层中。

密集连接卷积神经网络

Huang 等人在论文 [9] 中提出一种新架构 DenseNet，进一步利用快捷连接，将所有层直接连接在一起。在这种新型架构中，每层的输入由所有之前层的特征映射组成，其输出将传输给每个后续层。这些特征映射通过深度级联聚合。



除了解决梯度消失问题，[8] 的作者称这个架构还支持特征重用，使得网络具备更高的参数效率。一个简单的解释是，在论文 [2] 和论文 [7] 中，恒等映射的输出被添加到下一个模块，如果两个层的特征映射有着非常不同的分布，那么这可能会阻碍信息流。因此，级联特征映射可以保留所有特征映射并增加输出的方差，从而促进特征重用。



遵循该范式，我们知道第 l 层将具有 k *（l-1）+ k_0 个输入特征映射，其中 k_0 是输入图像的通道数目。作者使用一个叫作「增长率」的超参数 (k) 防止网络过宽，他们还用了一个 1*1 的卷积瓶颈层，在昂贵的 3*3 卷积前减少特征映射的数量。整体架构如下表所示：

Layers	Output Size	DenseNet-121($k=32$)	DenseNet-169($k=32$)	DenseNet-201($k=32$)	DenseNet-161($k=48$)
Convolution	112×112	7×7 conv, stride 2			
Pooling	56×56	3×3 max pool, stride 2			
Dense Block (1)	56×56	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 6$
Transition Layer (1)	56×56	1×1 conv			
	28×28	2×2 average pool, stride 2			
Dense Block (2)	28×28	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 12$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 12$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 12$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 12$
Transition Layer (2)	28×28	1×1 conv			
	14×14	2×2 average pool, stride 2			
Dense Block (3)	14×14	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 24$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 32$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 48$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 36$
Transition Layer (3)	14×14	1×1 conv			
	7×7	2×2 average pool, stride 2			
Dense Block (4)	7×7	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 16$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 32$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 32$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \end{bmatrix} \times 24$
Classification Layer	1×1	7×7 global average pool			
		1000D fully-connected, softmax			

用于 ImageNet 的 DenseNet 架构

深度随机的深度网络

尽管 ResNet 的强大性能在很多应用中已经得到了证实，但它存在一个显著缺点：深层网络通常需要进行数周的训练时间。因此，把它应用在实际场景的成本非常高。为了解决这个问题，G. Huang 等作者在论文 [10] 中引入了一种反直觉的方法，即在训练过程中随机丢弃一些层，测试中使用完整的网络。

作者使用残差块作为他们网络的构建块。因此在训练期间，当特定的残差块被启用，它的输入就会同时流经恒等快捷连接和权重层；否则，就只流过恒等快捷连接。训练时，每层都有一个「生存概率」，每层都有可能被随机丢弃。在测试时间内，所有的块都保持被激活状态，并根据其生存概率进行重新校准。

从形式上来看，H_l 是第 l 个残差块的输出结果，f_l 是由第 l 个残差块的加权映射所决定的映射，b_l 是一个伯努利随机变量（用 1 或 0 反映该块是否被激活）。在训练中：

$$H_l = ReLU(b_l * f_l(H_{l-1}) + id(H_{l-1})).$$

当 b_l=1 时，该块为正常的残差块；当 b_l=0 时，上述公式为：

$$H_l = ReLU(id(H_{l-1})).$$

既然我们已经知道了 H_(l-1) 是 ReLU 的输出，而且这个输出结果已经是非负的，所以上述方程可简化为将输入传递到下一层的 identity 层：

$$H_l = id(H_{l-1}).$$

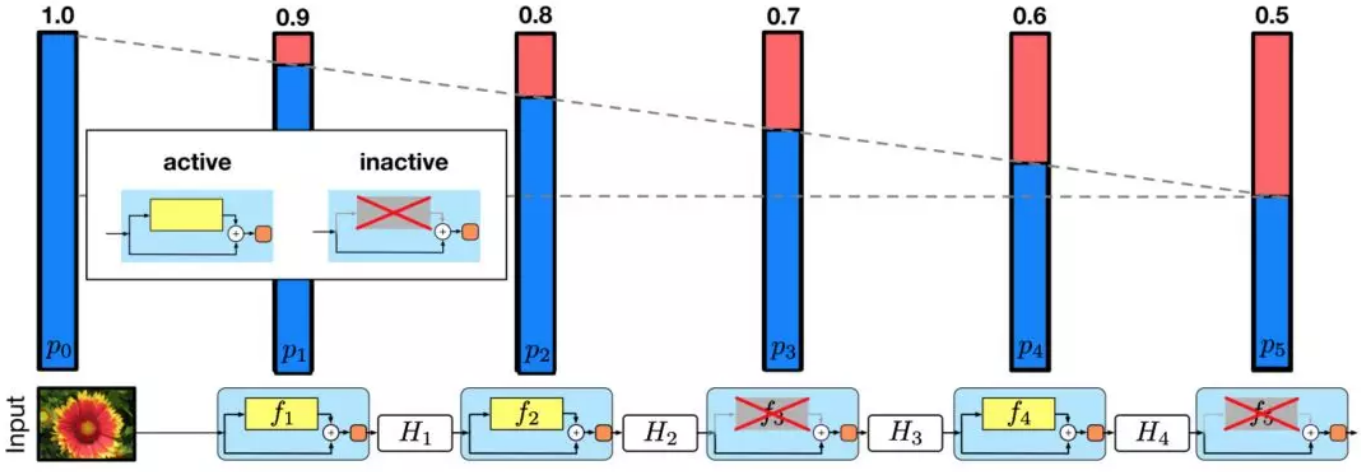
令 p_l 表示是第 l 层在训练中的生存概率，在测试过程中，我们得到：

$$H_l = ReLU(p_l * f_l(H_{l-1}) + id(H_{l-1})).$$

作者将线性衰减规律应用于每一层的生存概率，他们表示，由于较早的层提取的低级特征会被后面的层使用，所以不应频繁丢弃较早的层。这样，规则就变成：

$$p_l = 1 - \frac{l}{L}(1 - p_L).$$

其中 L 表示块的总数，因此 p_L 就是最后一个残差块的生存概率，在整个实验中 p_L 恒为 0.5。请注意，在该设置中，输入被视为第一层 (l=0)，所以第一层永远不会被丢弃。随机深度训练的整体框架如下图所示：



训练过程中，每一层都有一个生存概率

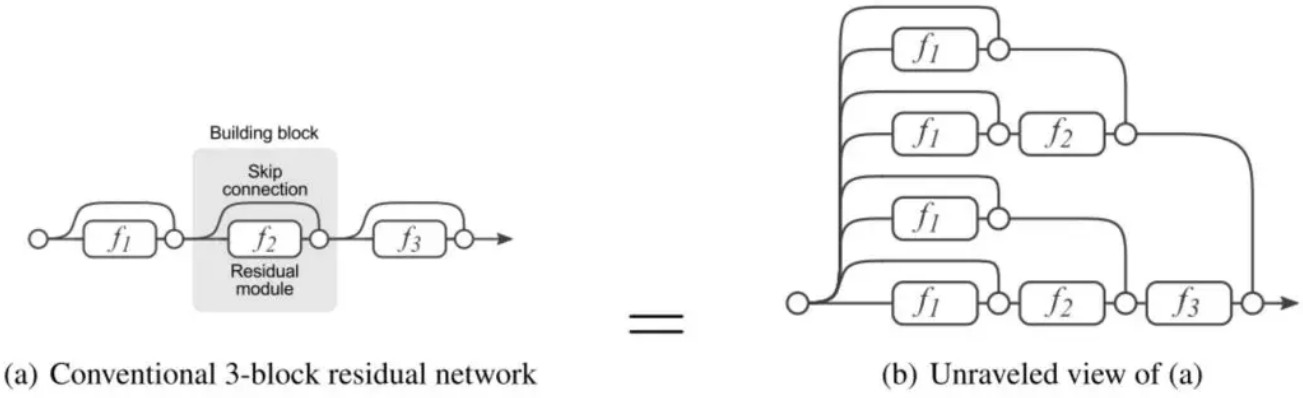
与 Dropout [11] 类似，训练随机深度的深度网络可被视为训练许多较小 ResNet 的集合。不同之处在于，上述方法随机丢弃一个层，而 Dropout 在训练中只丢弃一层中的部分隐藏单元。

实验表明，同样是训练一个 110 层的 ResNet，随机深度训练出的网络比固定深度的性能要好，同时大大减少了训练时间。这意味着 ResNet 中的一些层（路径）可能是冗余的。

作为小型网络集合的 ResNet

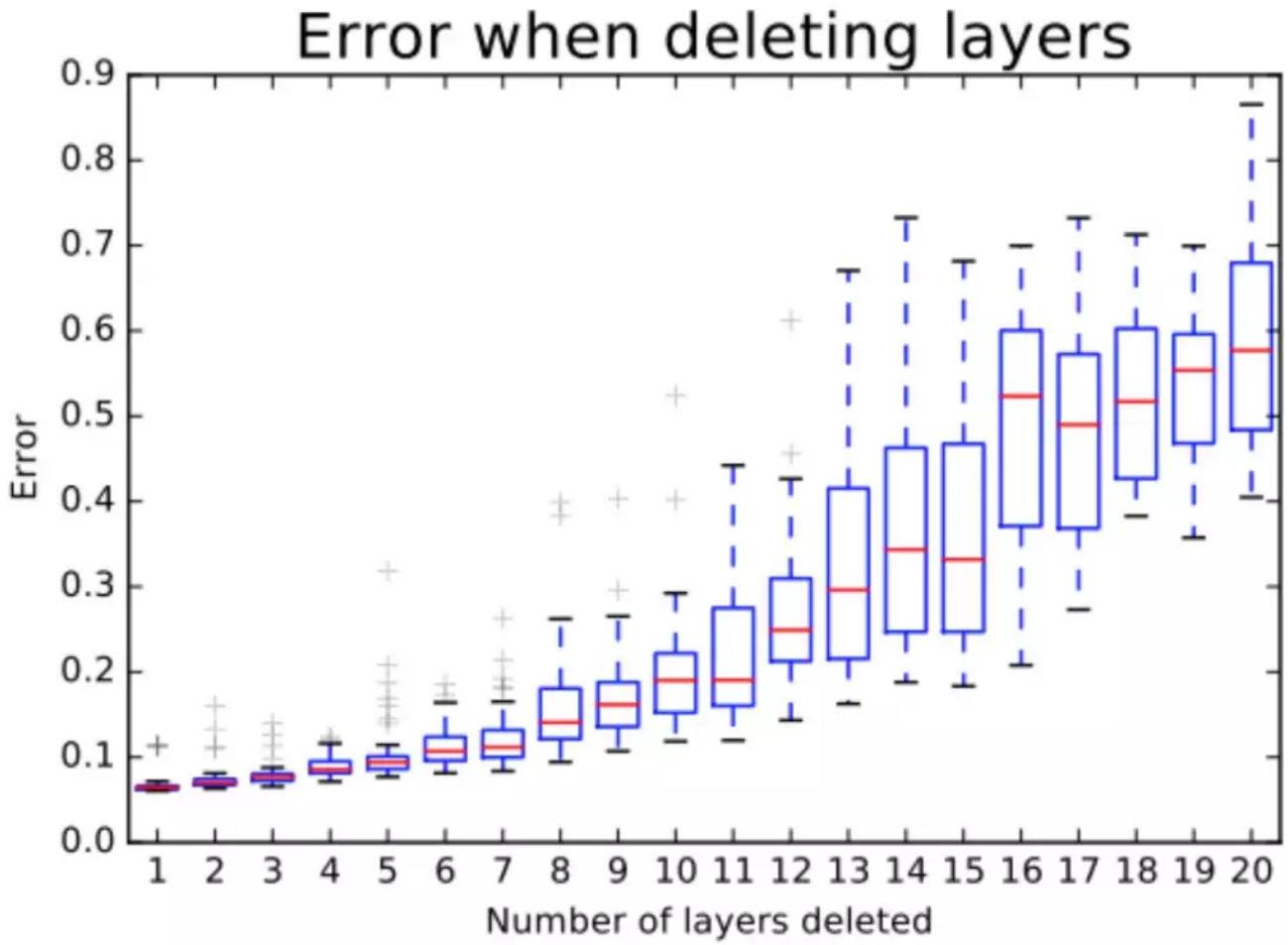
[10] 提出一种反直觉的方法，即在训练中随机丢弃网络层，并在测试中使用完整的网络。[14] 介绍了一种更加反直觉的方法：我们实际上可以删除已训练 ResNet 的部分层，但仍然保持相对不错的性能。[14] 还用同样的方式移除 VGG 网络的部分层，其性能显著降低，这使得 ResNet 架构更加有趣。

[14] 首先介绍了一个 ResNet 的分解图来使讨论更加清晰。在我们展开网络架构之后，很明显发现，一个有着 i 个残差块的 ResNet 架构有 2**i 个不同路径（因为每个残差块提供两个独立路径）。



根据上述发现，显然移除 ResNet 架构中的部分层对其性能影响不大，因为架构具备许多独立有效的路径，在移除了部分层之后大部分路径仍然保持完整无损。相反，VGG 网络只有一条有效路径，因此移除一个层会对该层的唯一路径产生影响。（如 [14] 中的实验所揭示的。）

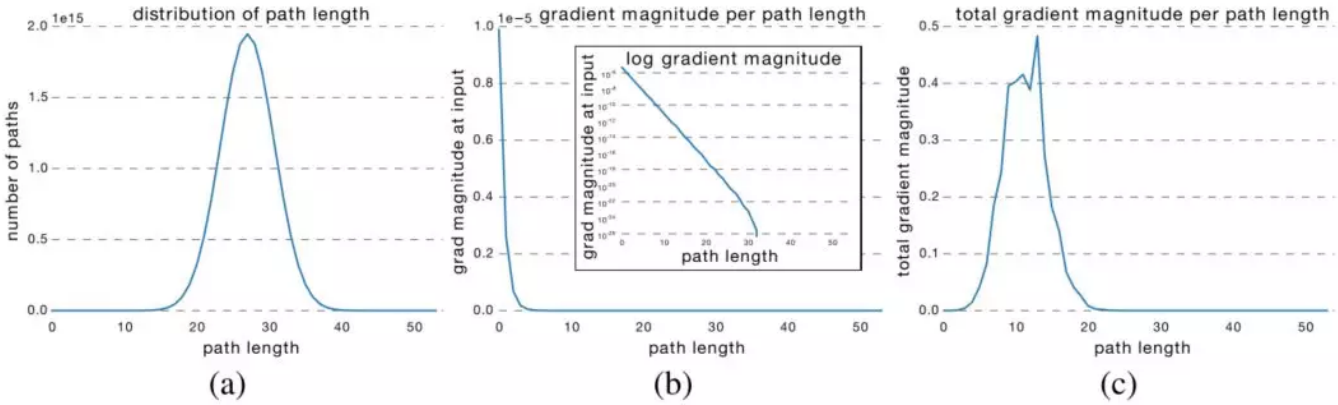
作者的另一个实验表明，ResNet 中不同路径的集合有类似集成的行为。他们在测试时删除不同数量的层，测试网络性能与删除层的数量是否平滑相关。结果表明，网络行为确实类似集成，如下图所示：



当被删除的层数增加时，误差值随之增长

最终，作者研究了 ResNet 中路径的特征：

很明显，路径的可能长度分布遵循二项分布，如下图 (a) 所示。大多数路径流经 19 到 35 个残差块。



为了研究路径长度与经过路径的梯度大小之间的关系，得到长度为 k 的路径的梯度大小，作者首先向网络输入了一批数据，并随机采样 k 个残差块。当梯度被反向传播时，它们在采样残差块中仅通过权重层进行传播。(b) 表明随着路径长度的增加，梯度大小迅速下降。

现在将每个路径长度的频率与其期望的梯度大小相乘，以了解每个长度的路径在训练中起到多大作用，如图 (c) 所示。令人惊讶的是，大多数贡献来自于长度为 9 到 18 的路径，但它们只占所有路径的一小部分，如 (a) 所示。这是一个非常有趣的发现，它表明 ResNet 并没有解决长路径的梯度消失问题，而是通过缩短有效路径的长度训练非常深层的 ResNet 网络。

结论

本文主要介绍了 ResNet 架构，简要阐述了其近期成功的原因，并介绍了几篇论文，它们叙述了一些有趣的 ResNet 变体，或提供了富有洞察力的解释。希望这篇文章有助于大家理解这项开创性的工作。

本文所有的图表均来自于参考文献中的原始论文。

References:

[1]. A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems,pages1097–1105,2012.

[2]. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385,2015.

[3]. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556,2014.

[4]. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,pages 1–9,2015.

[5]. R. Srivastava, K. Greff and J. Schmidhuber. Training Very Deep Networks. arXiv preprint arXiv:1507.06228v2,2015.

[6]. S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural Comput., 9(8):1735–1780, Nov. 1997.

[7]. K. He, X. Zhang, S. Ren, and J. Sun. Identity Mappings in Deep Residual Networks. arXiv preprint arXiv:1603.05027v3,2016.

[8]. S. Xie, R. Girshick, P. Dollar, Z. Tu and K. He. Aggregated Residual Transformations for Deep Neural Networks. arXiv preprint arXiv:1611.05431v1,2016.

[9]. G. Huang, Z. Liu, K. Q. Weinberger and L. Maaten. Densely Connected Convolutional Networks. arXiv:1608.06993v3,2016.

[10]. G. Huang, Y. Sun, Z. Liu, D. Sedra and K. Q. Weinberger. Deep Networks with Stochastic Depth. arXiv:1603.09382v3,2016.

[11]. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. The Journal of Machine Learning Research 15(1) (2014) 1929–1958.

[12]. A. Veit, M. Wilber and S. Belongie. Residual Networks Behave Like Ensembles of Relatively Shallow Networks. arXiv:1605.06431v2,2016.

原文链接：https://towardsdatascience.com/an-overview-of-resnet-and-its-variants-5281e2f56035