4-5 ResNet

王中雷

厦门大学王亚南经济研究院和经济学院, 2025

内容摘要

1. 研究动机

2. 网络结构

研究动机

1．赢得 ImageNet ILSVRC 2015 第一名

2．理论上讲，训练误差应该随着网络层数的增加而降低

3．然而，实际情况并非如此

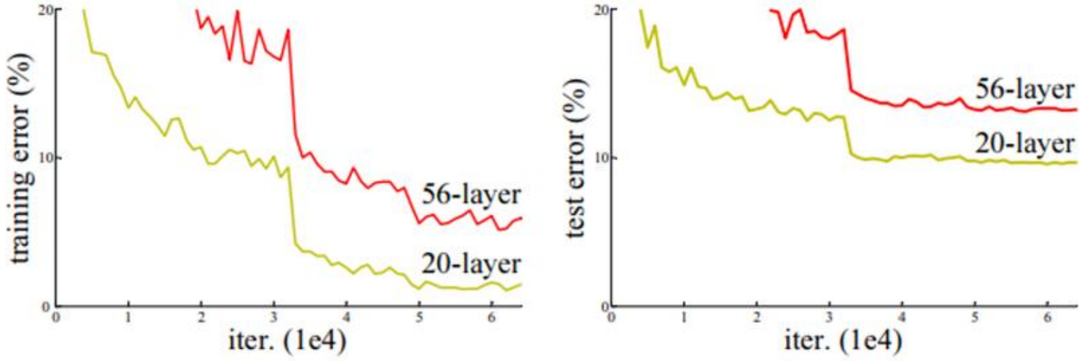


Figure 1. Training error (left) and test error (right) on CIFAR-10 with 20-layer and 56-layer "plain" networks. The deeper network has higher training error, and thus test error. Similar phenomena on ImageNet is presented in Fig. 4.

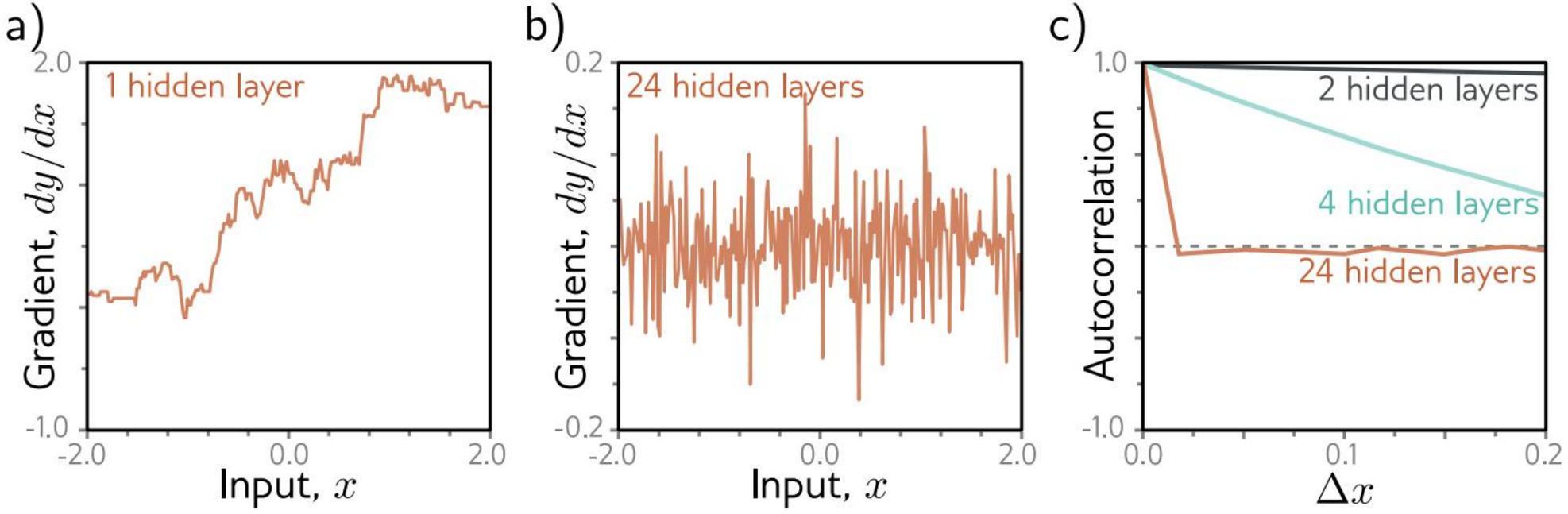
[K. He, X. Zhang, S. Ren and J. Sun, Deep Residual Learning for Image Recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778]

研究动机

1. 一般而言，复杂模型往往会对训练样本进行过拟合，导致在测试集上表现较差
2. 然而，过拟合往往会降低训练误差
3. 因此，问题是在于训练深度神经网络模型上，而不是来自于他们的泛化能力
4. 这个现象并没有理论解释，但一个合理的猜测是模型参数的初始化
   - 当我们稍微修改浅层参数时，梯度的表现往往会出乎意料
   - 利用合理的初始化，我们可以避免梯度消失或者梯度爆炸的问题
   - 然而，基于梯度的更新往往要求很小的步长（学习率）
     - 在实际中，步长（学习率）是一个（给定的）数值
     - 当代价函数表现得像重重叠叠的小山峰时，参数更新就可能出问题
   - 以上的猜测基于只有一维的输入和输出的神经网络数值结果得到

研究动机

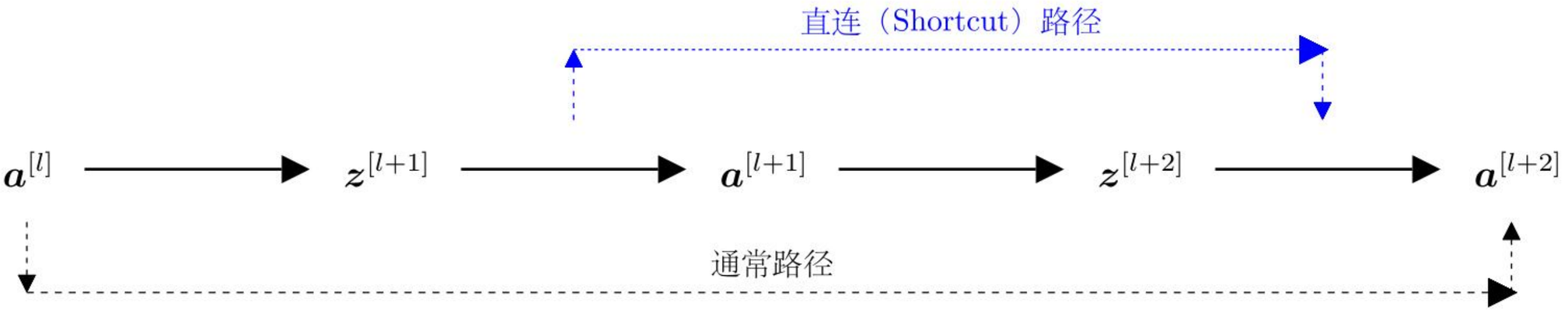
1. 以下图片来自于 Prince (2024) 的图 11.3



研究动机

1. 我们得到如下结论：
   - 对于浅层神经网络模型，模型输出对输入的偏导数变化较为平滑（缓慢）
   - 然而，对于深层神经网络模型，往往不那么平滑（变化较为剧烈）
2. 这种现象可以通过梯度间的“自相关函数”解释
   - 对于浅层神经网络模型，临近层的梯度之间往往存在相关性
   - 然而，对于深层神经网络模型，往往不是如此
3. 这被称作破碎的梯度（shattered gradients）现象

网络结构



$$z^{[l+1]} = W^{l+1}a^{[l]} + b^{[l+1]} \qquad a^{[l+1]} = \sigma^{[l+1]}(z^{[l+1]}) \qquad z^{[l+2]} = W^{l+2}a^{[l+1]} + b^{[l+2]} \qquad a^{[l+2]} = \sigma^{[l+2]}(z^{[l+2]})$$

1．直连（Shortcut）路径

$$a^{[l+2]} = \sigma^{[l+2]}(z^{[l+2]} + z^{[l+1]})$$

2．也就是说，$z^{[l+2]}$ 对$z^{[l+1]}$ 和$z^{[l+2]}$ 之间的“残差（residual）”进行建模

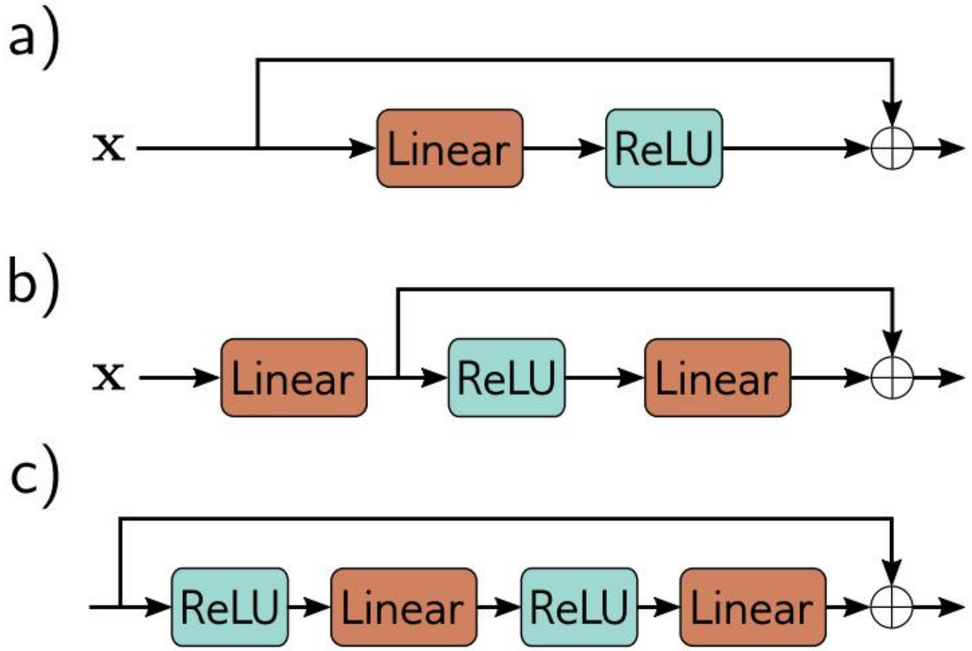
注意

1. 更一般地，向量$\boldsymbol{z}^{[l+2]}$ 和$\boldsymbol{z}^{[l+1]}$ 的长度可以不同
2. 这个问题可以通过引入模型参数$\boldsymbol{W}_s^{[l]}$ 进行解决

$$\boldsymbol{a}^{[l+2]} = \sigma^{[l+2]}(\boldsymbol{z}^{[l+2]} + \boldsymbol{W}_s^{[l]}\boldsymbol{z}^{[l+1]})$$

注意

1. 下图来自于 Prince (2024) 的图 11.5



注意

1．ReLU 激活函数的输出是非负的

2．我们如果将直连路径放在激活之后，那么我们会增加 “输入值”

3．因此，直连路径只能存在于两个线性变换之间

注意

1. 直连路径可被用于增加神经网络的深度
2. 因此，模型的训练可能出现（指数级）的梯度消失或者梯度爆炸
3. 我们可利用批量归一化（batch normalization）缓解这个问题带来的影响