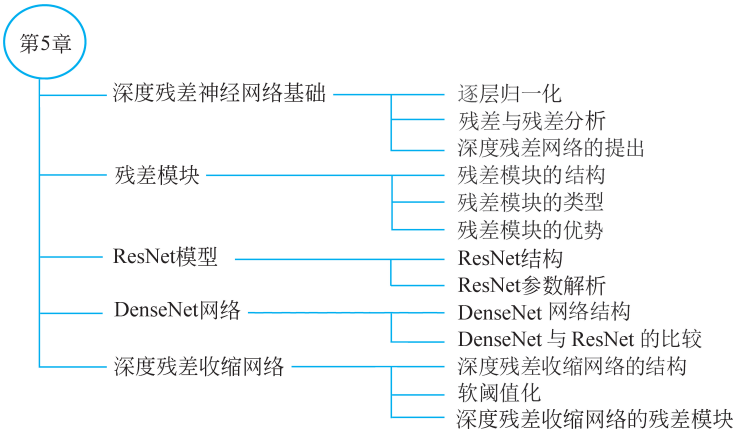
第 5 章

深度残差神经网络



为了解决多层神经网络训练出现的梯度消失、梯度爆炸和网络性能退化问题，研究者引入了数据标准化、权重初始化和 BN 层等技术。应用残差方法可以解决网络性能退化问题，以及避免神经网络层数加深导致损失函数值增大的问题。

5.1 深度残差神经网络基础

AlexNet、VGG、GoogLeNet 等网络模型的出现使神经网络的发展进入了几十层规模的阶段，而且网络的层数越深，越有可能获得优异的泛化能力。但是，当模型层数加大以后，网络变得越来越难以训练，这主要由梯度消失和梯度爆炸所造成。

5.1.1 逐层归一化

神经网络学习过程就是学习数据分布的过程，如果训练数据与测试数据的分布不同，则网络的泛化能力显著降低。

深度网络的训练是一个复杂的过程，只要网络的前面几层发生微小的改变，那么后面几层就被累积放大。一旦网络某一层的输入数据的分布发生改变，那么这一层网络就需要去学习新的数据分布，所以在训练过程中，如果训练数据的分布一直在发生变化，那么网络就要在每次迭代都去学习适应不同的分布，这样将降低网络的训练速度。网络在训练的过程中，除了输入层的数据外，后边各层的输入数据分布也一直在发生变化，对于中间各层，将数据分布的改变称为数据归一化。

逐层归一化(batch normalization，BN)是 2015 年由谷歌公司提出的一种模型正规化方

法，该方法通过对层间输入正规化来加速网络收敛。将传统机器学习中的数据归一化方法应用到深度神经网络中，对神经网络中隐藏层的输入进行归一化，从而使网络更容易训练。几种常用逐层归一化方法是批量归一化、层归一化、权重归一化和局部响应归一化。

网络除了输入层外，其他各层因为前层网络在训练时更新了参数，而引起后层输入数据分布的变化。如果在每一层输入时，加上预处理操作，将数据归一化至均值为 0、方差为 1，然后再输入后层计算，这样便解决了内部协变量偏移的问题了。事实上，在网络每一层输入时，插入了一个归一化层，也就是先做一个归一化处理，然后再进入网络的下一层。BN 层是一个可学习、有参数的网络层。

BN 方法就是在深度神经网络训练过程中使每一层神经网络的输入都保持相同的分布。因为深层神经网络在做非线性变换前的输入值（是指 $\boldsymbol{y}=\boldsymbol{W}\boldsymbol{x}+\boldsymbol{b}$，$\boldsymbol{x}$ 是输入，$\boldsymbol{b}$ 为偏置项）随着网络深度加深或者在训练过程中，其分布逐渐发生偏移或者变动。一般整体分布会逐渐向非线性函数的取值区间的上下限两端靠近，导致反向传播时低层神经网络的梯度消失，这是训练深层神经网络时收敛越来越慢的基本原因。BN 方法就是通过一定的规范化手段，将每层神经网络任意神经元输入值的分布拉回到均值为 0、方差为 1 的标准正态分布，其实就是将越来越偏的分布强制拉回比较标准的分布。这样就使得激活输入值落在非线性函数对输入比较敏感的区域，网络的输出就可以得到比较大的梯度，避免了梯度消失问题，而且梯度变大表明学习收敛速度加快，进而加快训练速度。

BN 就是对每一批数据进行归一化，对于训练中某一批数据 $\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$，这个数据可以是输入，也可以是网络中间某一层的输出。在 BN 方法出现之前，归一化操作一般都在数据输入层，对输入的数据进行求均值以及求方差做归一化，但是 BN 的出现打破了这一规定，使归一化处理可以在网络中任意一层进行。因为现在所用的优化方法大部分是最小批随机梯度下降（stochastic gradient descent，SGD），在 CNN 中，批就是训练网络所设定的图片数量的批量大小，所以通常将归一化操作又称为批量规范化。

1. BN 的主要步骤

BN 计算过程形式化描述如下：

输入：批量数据 $B=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_m\}$

学习参数：$\boldsymbol{\beta}$、$\boldsymbol{\gamma}$。

输出：$\boldsymbol{y}_i\leftarrow\boldsymbol{\gamma}\hat{\boldsymbol{x}}_i+\boldsymbol{\beta}=\mathrm{BN}_{\boldsymbol{\gamma},\boldsymbol{\beta}}(\boldsymbol{x}_i)$

BN 的主要步骤如下：

（1）计算每一批训练数据的均值：

$$\boldsymbol{\mu}_B\leftarrow\frac{1}{m}\sum_{i=1}^{m}\boldsymbol{x}_i$$

（2）计算每一批训练数据的方差：

$$\sigma_B^2\leftarrow\frac{1}{m}\sum_{i=1}^{m}(\boldsymbol{x}_i-\boldsymbol{\mu}_B)^2$$

（3）使用求得的均值和方差对该批次的训练数据进行规范化处理。其中ϵ是为了避免除数为 0 时所使用的微小正数：

$$\hat{\boldsymbol{x}}_i\leftarrow\frac{\boldsymbol{x}_i-\boldsymbol{\mu}_B}{\sqrt{\sigma_B^2+\epsilon}}$$

(4) 尺度变换和偏移：将 $\hat{\boldsymbol{x}}_i$ 乘以 $\boldsymbol{\gamma}$ 调整数值大小，再加上 $\boldsymbol{\beta}$ 增加偏移后得到 $\boldsymbol{y}_i$，$\boldsymbol{\gamma}$ 是尺度因子，$\boldsymbol{\beta}$ 是平移因子。这一步是 BN 的精髓，这是因为归一化后的 $\boldsymbol{x}_i$ 基本被限制在正态分布下，使网络的表达能力下降了。

$$\boldsymbol{y}_i \leftarrow \boldsymbol{\gamma}\hat{\boldsymbol{x}}_i + \boldsymbol{\beta} \equiv \mathrm{BN}_{\gamma,\beta}(\boldsymbol{x}_i)$$

在平移和缩放处理中，引入了可学习的重构参数 $\boldsymbol{\gamma}$ 和 $\boldsymbol{\beta}$，让网络可以学习并恢复出原始网络所要学习的特征分布，这就是算法关键之处：

$$\boldsymbol{y}^{(k)} = \boldsymbol{\gamma}^{(k)}\hat{\boldsymbol{x}}^{(k)} + \boldsymbol{\beta}^{(k)}$$

2. BN 层的作用

BN 层的作用如下：

(1) 加快网络训练和收敛的速度。

(2) 防止梯度爆炸或梯度消失。

(3) 防止过拟合。

BN 层一般用在线性层和卷积层后面，而不是放在非线性单元后。因为非线性单元的输出分布形状将在训练过程中变化，归一化无法消除其方差偏移，相反，全连接和卷积层的输出一般是一个对称、非稀疏的分布，更加类似高斯分布，对它们进行归一化处理可产生更加稳定的分布。例如，像 ReLU 这样的激活函数，如果输入的数据是一个高斯分布，经过变换的数据是小于 0 的被抑制，也就是分布小于 0 的部分直接变成 0，这样就更加接近高斯分布。

5.1.2　残差与残差分析

1. 残差定义

在数理统计中，残差是指实际观测值与估计值(拟合值)之间的差，残差蕴含了有关模型基本假设的重要信息。如果回归模型正确，则可以将残差看作观测值的误差。多次重复测量时，各次测得值与平均值(数学期望)之差称为残差。

残差是因变量的观测值 $\boldsymbol{y}_i'$ 与根据估计的回归方程求出的预测值 $\boldsymbol{y}_i$ 之差，用 $\boldsymbol{e}$ 表示。它反映了用估计的回归方程去预测 $\boldsymbol{y}_i'$ 而引起的误差。第 i 个观察值的残差为

$$\boldsymbol{e}_i = \boldsymbol{y}_i' - \boldsymbol{y}_i$$

残差(或残差平方和)反映数据的离散程度。

2. 残差分析

残差应符合模型的假设条件，且具有误差的一些性质。残差分析是指利用残差所提供的信息，来考察模型假设的合理性及数据的可靠性。显然，有多少对数据，就有多少个残差。

以某种残差为纵坐标，变量为横坐标作散点图，即为残差图，它是残差分析的重要工具之一，通常横坐标的选择有三种：

(1) 因变量的拟合值。

(2) 自变量。

(3) 当因变量的观测值为一时间序列时，横坐标可取观测时间或观测序号。

残差图的分布趋势可以帮助判明所拟合的线性模型是否满足有关假设条件，例如残差是否近似正态分布、是否方差齐次，变量间是否有其他非线性关系及是否还有重要自变量未进入模型等。当判明存在某种假设条件欠缺时，进一步的问题就是加以校正或补救。需要

分析具体情况，探索合适的校正方案，例如非线性处理、引入新的自变量，或考察误差是否有自相关性。残差图的示意图如图 5-1 所示。

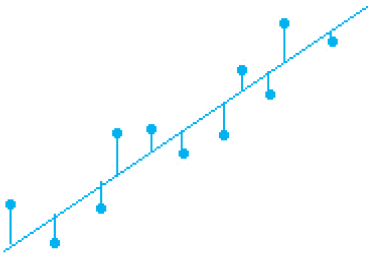


图 5-1 残差图的示意图

5.1.3 深度残差网络的提出

在深度神经网络中存在一个问题，网络层数加深，参数增多，网络表现能力理应更好，但随着深度的不断增加，将出现网络退化现象。这种退化现象表现在，随着神经网络层数加深，训练准确率将逐渐趋于饱和；如果层数继续加深，反而训练准确率下降，效果倒不好。这既不是梯度爆炸、梯度消失造成的，也不是过拟合造成的。

神经网络在误差反向传播时，反向连乘的梯度小于 1(或大于 1)，导致连乘的次数多了之后(网络层数加深)，传回首层的梯度过小甚至为 0(过大甚至无穷大)，这就是梯度消失/爆炸的概念。可在网络中加入 BN 层，通过规整数据的分布基本解决梯度消失/爆炸的问题，所以这个问题也不是导致深层网络退化的原因。过拟合在网络训练集上表现很好，在测试集上表现差(无论是在训练集还是测试集中，更深层次的网络表现均比浅层次的网络差，那显然就不是过拟合导致的)。网络退化现象由非线性激活函数 ReLU 的存在所造成，每次输入到输出的过程几乎是不可逆的，这也造成了许多不可逆的信息损失。也就是说，如果一个特征的一些有用的信息损失了，则其表现很难做到优秀。层数增多之后，信息在中间层损失掉了。

可以给深层神经网络添加一种回退到浅层神经网络的机制，当发现损失消失现象时就回退到浅层神经网络，使深层神经网络可以获得与浅层神经网络相当的模型性能。通过在输入和输出之间添加一条直接的捷径连接，可以使神经网络具有回退的能力。例如假设观察到第 13 层神经网络出现梯度消失现象，而第 10 层的网络模型并没有观测到梯度消失现象，那么可以考虑在最后的两个卷积层添加捷径连接。通过这种方式，网络模型可以自动选择是否经由这两个卷积层完成特征变换，还是直接跳过这两个卷积层而选择捷径连接，这就是深度残差网络的由来。

深度残差网络的出现打破了层次限制，使训练更深层次的神经网络成为可能。

5.2 残 差 模 块

在多层神经网络中加入残差模块，可以使其更加容易被优化，通常将增加了残差模块的神经网络称为残差神经网络。

5.2.1 残差模块的结构

1. 残差模块的基本组成

普通模块的结构如图 5-2 所示，残差模块的结构如图 5-3 所示。

(1) 残差模块比普通模块增加了右边的曲线，这条曲线为：快捷方式连接或身份映射。

(2) $H(\boldsymbol{x})=F(\boldsymbol{x},\boldsymbol{W}_i)+\boldsymbol{x}$。

(3) 模型需要学习的是 $F(\boldsymbol{x},\boldsymbol{W}_i)$这个残差，而不是普通模块的 $H(\boldsymbol{x})$，公式则可以变换为

$$F(\boldsymbol{x},\boldsymbol{W}_i)=H(\boldsymbol{x})-\boldsymbol{x}$$

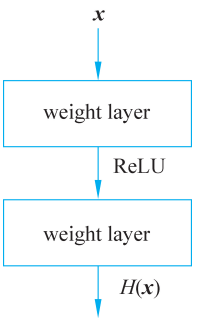


图 5-2　普通模块的结构

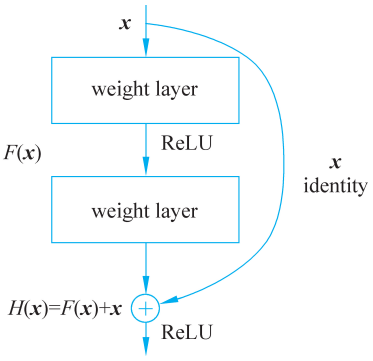


图 5-3　残差模块的结构

(4) $F(\boldsymbol{x})$与 $\boldsymbol{x}$ 是直接相加，而不是在某个维度拼接，所以要求 $F(\boldsymbol{x})$与 $\boldsymbol{x}$ 的形状相同。

(5) 残差模块的输出结果是在 $F(\boldsymbol{x})+\boldsymbol{x}$ 之后再加 ReLU 激活函数，而不是对 $F(\boldsymbol{x})$进行激活函数作用之后再加 $\boldsymbol{x}$，也就是 $F(\boldsymbol{x})+\boldsymbol{x}$ 在 ReLU 激活函数之前。

当没有从 $\boldsymbol{x}$ 到⊕的箭头时，残差模块就是一个普通的两层网络。残差模块中的网络可以是全连接层，也可以是卷积层。设第二层网络在激活函数之前的输出为 $H(\boldsymbol{x})$。如果在两层网络中，最优的输出就是输入 $\boldsymbol{x}$，那么对于没有捷径连接的网络，就需要将其优化成 $H(\boldsymbol{x})=\boldsymbol{x}$。对于有捷径连接的网络，即残差模块，如果最优输出是 $\boldsymbol{x}$，则只需要将其优化为

$$F(\boldsymbol{x})=H(\boldsymbol{x})-\boldsymbol{x}=0$$

显然，后者的优化比前者更简单，这也是残差的由来。

$F(\boldsymbol{x})$是与 $\boldsymbol{x}$ 求和前的网络映射，$H(\boldsymbol{x})$是从输入到与 $\boldsymbol{x}$ 求和后的网络映射。例如，如果将 5 映射到 5.1，那么引入残差前是 $F'(5)=5.1$，引入残差后是 $H(5)=5.1$，$F(5)=H(5)-5$，$F(5)=0.1$。这里的 F'和 F 都表示网络参数映射，引入残差后的映射对输出的变化更敏感。例如，输出从 5.1 变到 5.2，映射 F'的输出增加了 $1/51=2\%$，而对于残差结构输出从 5.1 到 5.2，映射 F 是从 0.1 到 0.2，增加了 100%。明显后者输出变化对权重的调整作用更大，所以效果更好。残差的思想都是去掉相同的主体部分，从而突出微小的变化。可以对多堆叠层采用残差学习。

残差模块可以解决深层神经网络准确率下降的问题。对于一个神经网络模型，如果该模型是最优的，那么训练就很容易将残差映射优化到 0，此时只剩下身份映射，那么无论怎么增加深度，理论上网络将一直处于最优状态。因为相当于后面所有增加的网络都将沿着身份映射(自身)进行信息传输，可以理解为最优网络后面的层数都是废掉的(不具备特征提取的能力)，实际上没起什么作用。这样，网络的性能也就不随着深度的增加而降低。

2. 捷径连接

在残差网络中有很多的旁路将输入直接连接到后面的层，这种结构也称为捷径连接或者跳过连接。捷径连接又分为实线连接和虚线连接。

(1) 实线连接

如果 F 和 $\boldsymbol{x}$ 维度相同，可以直接相加，不增加网络的参数以及计算复杂度，计算公式为

$$\boldsymbol{y}=F(\boldsymbol{x},\{\boldsymbol{W}_i\})+\boldsymbol{x}$$

在这种情况下的捷径连接的图形表示采用实线连接。

(2) 虚线连接

如果 $\boldsymbol{x}$ 和 F 的维度不同,需要先将 $\boldsymbol{x}$ 做一个变换,使特征矩阵形状相同,然后再相加。

虚线部分前后块的维度不一致,则体现在两方面上。

(1) 空间不一致

在跳接部分给输入的 $\boldsymbol{x}$ 加上线性映射 $\boldsymbol{W}$:

$$\boldsymbol{y}=\boldsymbol{F}(\boldsymbol{x},\{\boldsymbol{W}_i\})+\boldsymbol{x}\rightarrow\boldsymbol{y}=\boldsymbol{F}(\boldsymbol{x},\{\boldsymbol{W}_i\})+\boldsymbol{W}_s\boldsymbol{x}$$

(2) 深度不一致

如果深度不一致,则全 0 填充。

例 5-1 虚线连接。

跳接时加 1×1 卷积层升维,如图 5-4 所示,注意使用了虚线。实线的残差连接就是正常的直接相加,虚线表示在右侧边分支内增加了一个 1×1 卷积核的卷积操作。利用 1×1 卷积核的卷积操作可以升维或者降维,在这里使用升维,从 63 维升到 128 维,这样使得主分支与捷径连接的输出维数相同。

5.2.2 残差模块的类型

残差模块分为基本型和瓶颈型两种。基本型残差模块由两个 3×3 的卷积网络串接在一起组成,瓶颈型残差模块由 1×1、3×3、1×1 的三个卷积网络串接在一起组成,如图 5-5 所示。

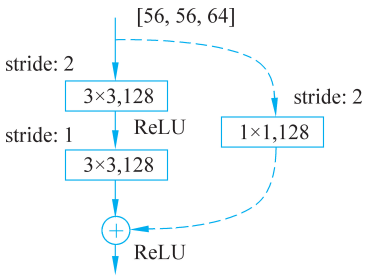
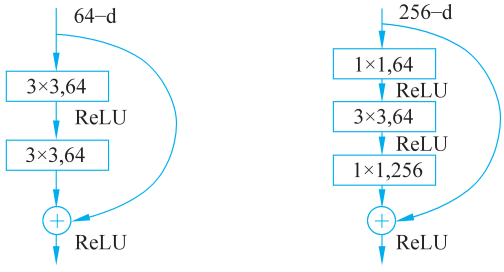


图 5-4 虚线连接



(a) 基本型残差模块　　(b) 瓶颈型残差模块

图 5-5 常用的两种残差模块

1. 基本型残差模块

(1) 3×3 卷积可以替代更大尺寸的卷积

在构建 CNN 时常使用 3×3 的卷积,而不是 5×5、7×7 等更大尺寸的卷积。这是由于在保证具有同样大小的输出和感受野前提下,用 3×3 卷积可以替代更大尺寸的卷积。两个 3×3 的卷积可以代替一个 5×5 的卷积;三个 3×3 的卷积可以代替一个 7×7 的卷积。所以 VGG 系列网络中全部使用了 3×3 的卷积。

假设图像大小为 $n\times n$,如果采用 5×5 卷积核的方案,stride=1,padding=0,其输出维度为 $(n-5)/1+1=n-4$。

采用 3×3 卷积的方案,同样图像大小为 $n\times n$,第 1 次 3×3 卷积后输出维度为

$$(n-3)/1+1=n-2$$

第 2 次 3×3 卷积后输出维度为

$$(n-2-3)/1+1=n-4$$

可以看出，采用一个 5×5 卷积核和两个 3×3 卷积核，它们卷积后的输出维度相同，输出的每一个像素的感受野也相同。这表明两个 3×3 的卷积可以代替一个 5×5 的卷积。

(2) 使用两个 3×3 卷积代替一个 5×5 卷积的优势分析

- 两个 3×3 卷积可以代替一个 5×5 卷积网络，导致层数增加，也提高了网络的非线性表达能力。
- 两个 3×3 卷积可以代替一个 5×5 卷积网络，使参数减少。两个 3×3 和一个 5×5 的参数比例为 3×3×2/(5×5)=0.72，同样的三个 3×3 和一个 7×7 的参数比例为 3×3×3/(7×7)=0.55，压缩将近一半，这是很大提升。

考虑这两点，残差网络中多数采用了两个 3×3 的卷积的结构构成基本型残差模块，如图 5-4(a)所示。

2. 瓶颈型残差模块

瓶颈型残差模块由 1×1、3×3、1×1 的三个卷积网络串接在一起组成。

1×1 卷积又称为网中网，在残差模块中，经常使用 1×1 卷积，如图 5-4(b)所示。两个 1×1 卷积和一个 3×3 卷积可构成瓶颈型残差模块。在残差网络中 1×1 卷积表面看起来无作用，但是其作用颇多，总结如下。

(1) 实现跨通道的特征整合与信息交互

如果当前层和下一层都只有一个通道，那么 1×1 卷积核确实没有什么作用。例如，输入 6×6×1 的矩阵，这里的 1×1 卷积形式为 1×1×1，卷积核中的元素为 3，经过 1×1 卷积，输出结果也是 6×6×1 的矩阵。但输出矩阵中的每个元素值是输入矩阵中每个元素值乘以 3 的结果，仅改变了原内容，如图 5-6 所示。

图 5-6　跨通道的特征整合

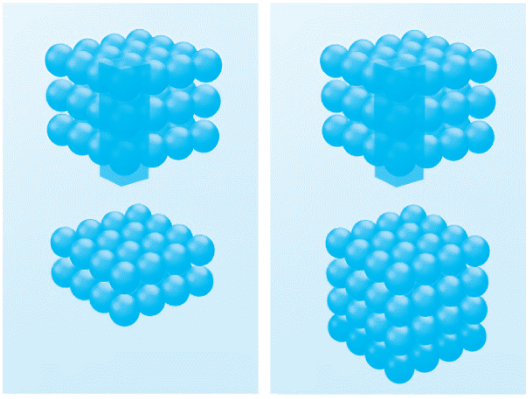
跨通道信息交互是通过通道的变换实现的。使用 1×1 卷积核实现降维和升维的操作其实就是通道间信息的线性组合变化，3×3×64 通道的卷积核后面添加一个 1×1×28 通道的卷积核，就变成了 3×3×28 通道的卷积核，原来的 64 个通道就可以理解为跨通道线性组合变成了 28 通道，只是在通道维度上做线性组合，$\boldsymbol{W}$ 和 H 上是共享权值的滑动窗口。

(2) 降维或升维

但是，如果它们分别为 m 层和 n 层，1×1 卷积可以起到一个跨通道聚合的作用，也可以起到降维(或者升维)的作用，改变参数量。

由于 1×1 并不改变高度和宽度，改变通道的直观结果就是可以将原本的数据量进行增加或者减少。改变的只是高度×宽度×通道中的通道这一维度的大小而已。降维和升维的

表示如图 5-7 所示。



(a) 降维　　(b) 升维

图 5-7　降维和升维的表示

使用 1×1 卷积核实现降维和升维的操作其实就是通道间信息的线性组合变化。

例 5-2　通道间的信息交互。

3×3×64 通道的卷积核后面添加一个 1×1×28 通道的卷积核，就变成了 3×3×28 通道的卷积核，原来的 64 个通道就可以理解为跨通道线性组合变成了 28 通道，这就是通道间的信息交互。

(3) 增加非线性特征

一个卷积核对应卷积后得到一个特征图，不同的卷积核具有不同的权重和偏置，卷积以后得到不同的特征图，提取了不同的特征。1×1 卷积核可以在保持特征图尺度不变的(即不损失分辨率)的前提下，利用后接的非线性激活函数，大幅增加非线性特性，同时网络也能做得越来越深。

在残差模块中，假设输入的特征图的维度是 $wh256$，并且最后要输出的也是 256 个特征图，则可以通过 1×1、3×3、1×1 的三个卷积网络串接完成。

5.2.3　残差模块的优势

1. 残差模块的主要作用

(1) 加了残差结构后，为输入 $\boldsymbol{x}$ 多一个选择的路径。如果神经网络学习到这层的参数是冗余的，它可以选择直接走这条“跳接”提供了曲线(快捷连接方式)，跳过这个冗余层，而不需要再去拟合参数使输出 $H(\boldsymbol{x})$ 等于 $\boldsymbol{x}$。

(2) 学习残差的计算量比学习输出等于输入要小。假设普通网络为 A，残差网络为 B，输入为 2，输出为 2(输入和输出一样是为了模拟冗余层需要恒等映射的情况)，那么普通网络就是 $A(2)=2$，而残差网络就是 $B(2)=F(2)+2=2$，显然残差网络中的 $F(2)=0$。网络中权重一般初始化为 0 附近的数，则 $F(2)$(经过权重矩阵)拟合 0 比拟合 $A(2)=2$ 更容易。

(3) ReLU 能够将负数激活为 0，而正数输入等于输出。这相当于过滤了负数的线性变化，让 $F(2)=0$ 变得更加容易。

(4) 残差网络可以表示成 $H(\boldsymbol{x})=F(\boldsymbol{x})+\boldsymbol{x}$，这就说明了在求输出 $H(\boldsymbol{x})$ 对输入 $\boldsymbol{x}$ 的导数(梯度)，也就是在反向传播时，$H(\boldsymbol{x})'=F(\boldsymbol{x})'+1$，残差结构的这个常数 1 也能保证在求

梯度时梯度不会消失。

2. 残差模块的优点

(1) 用恒等映射与残差相加，并没有增加模型的参数量，也没有增加计算复杂度。

(2) 增加残差模块后模型的收敛速度加快，即误差下降的梯度更大。

(3) 可以解决退化问题，至少不比没有加深网络差。

(4) 加了残差模块，网络就可以实现很深。

(5) 准确率也有了很大的提升。

5.3　ResNet 模型

残差神经网络(residual neural network，ResNet)由何恺明等在 2015 年提出，ResNet 的主要思想是在网络中增加了直连通道，即高速公路网络思路。此前的网络结构是将输入做一个非线性变换，而高速公路网络则允许保留之前网络层的一定比例的输出。ResNet 的思路和高速公路网络的思路也非常类似，允许原始输入信息直接传到后面的层中。这样某一层的神经网络可以不用学习整个输出，而是学习上一个网络输出的残差。

增加深度网络的深度后，会导致训练的难度增加。但是，通过残差的学习框架不仅可以简化网络训练，还适合在深度神经网络使用。与过去传统的方法比较，残差网络更易于优化，且在增加深度的同时，又可以获得准确性。

基于捷径连接的深度残差神经网络，其输入的一些部分将传递到下一层，因此，这些网络可以相当深，如 18 层、34 层、50 层、101 层、152 层的 ResNet-18、ResNet-34、ResNet-50、ResNet-101 和 ResNet-152 等模型，甚至层数达到 1202 层的极深层神经网络。

按照残差学习的基本思想，可以构造深度残差神经网络来解决模型的退化问题。ResNet 残差网络具有不同的网络层数，比较常用的是 50 层、101 层和 152 层。它们都由残差模块堆叠而成。

例 5-3　使用瓶颈型残差模块，可以减少参数数量和计算量。

将基本型残差模块应用到 ResNet34 和将瓶颈型残差模块应用到 dResNet50/101/152，其目的主要就是降低参数数量。基本型残差模块是两个 $3\times3\times256$ 的卷积，参数数量为 $3\times3\times256\times256\times2=1\ 179\ 648$，瓶颈型残差模块是第一个 $1\times1\times64$ 的卷积将 256 维通道降到 64 维，再通过 $3\times3\times64$ 卷积，最后通过 $1\times1\times256$ 卷积恢复，整体上用的参数数量为 $1\times1\times256\times64+3\times3\times64\times64+1\times1\times64\times256=69\ 632$，瓶颈型残差模块的参数数量比基本型残差模块减少了 94.1%，因此，瓶颈型残差模块可减少参数数量，从而减少计算量。基本型残差模块可以用于 34 层或者更浅的网络中；对于更深(如 101 层)的网络，则使用瓶颈型残差模块，可以减少参数数量和计算量。

5.3.1　ResNet 结构

传统的卷积网络或者全连接网络在信息传递时或多或少存在信息丢失、损耗等问题，同时会导致梯度消失或者梯度爆炸，进而使很深的网络无法训练。ResNet 在一定程度上解决了这个问题，它通过直接将输入信息绕道传到输出，保护信息的完整性，整个网络只需要学习输入、输出差别的部分，简化了学习目标，降低了学习难度。VGG19 和 ResNet 的比较如

图 5-8 所示。ResNet 与 VGG19 最大的区别在于有很多的旁路将输入直接连接到后面的层，这种结构也被称为捷径或者跳过连接。

ResNet 的结构使得网络具有学习恒等映射的能力，同时也具有学习其他映射的能力。因此 ResNet 的结构要优于传统的卷积网络结构。恒等映射是一个返回相同值的函数，该值用作其参数，也称为恒等关系或恒等转换。如果 f 是一个函数，则对于 $\boldsymbol{x}$ 的所有值，参数 $\boldsymbol{x}$ 的恒等关系表示为 $f(\boldsymbol{x})=\boldsymbol{x}$。

1. 在 VGG19 的基础上的修改

ResNet 是在 VGG19 的基础上进行了修改，并通过短路机制加入了构造的残差单元，VGG19、34 层 plain 和 34 层 ResNet 的比较如图 5-8 所示。ResNet 的主要变化如下。

(1) ResNet 直接使用 stride=2 的卷积，然后下采样。

(2) 用全局平均池化层替换了全连接层。

全局平均池化是指将特征图所有像素值相加求平均，得到一个数值，即用该数值表示对应特征图。其目的是替代全连接层，减少参数数量、计算量，防止过拟合。如图 5-9 所示，假设最终分成 10 类，则最后卷积层应该包含 10 个卷积核，即输出 10 个特征图，然后按照全局池化平均定义，分别对每个特征图累加所有像素值并求平均，最后得到 10 个数值，将这 10 个数值输入 Softmax 层中，得到 10 个概率值，即这张图片属于每个类别的概率值。

对整个网络从结构上做正则化防止过拟合，剔除了全连接层黑箱子操作的特征，直接赋予了每个通道实际的类别意义。

2. ResNet 的设计原则

ResNet 的设计原则：当特征图大小减小一半时，特征图的数量增加一倍，这保持了网络层的复杂度。从图 5-8 可以看到，ResNet 每两层间增加了短路机制，这就形成了残差学习，其中虚线表示特征图数量发生了改变。图中展示的 34 层 ResNet，还可以构建更深的网络，如表 5-1 所示。从表中可以看到，对于 18 层和 34 层的 ResNet，其进行的是两层间的残差学习。当网络更深时，其进行的是三层间的残差学习，三层卷积核分别是 1×1、3×3 和 1×1，隐藏层的特征图数量比较小，并且是输出特征图数量的 1/4。

5.3.2 ResNet 参数解析

表 5-1 给出了 5 种 ResNet，conv1、conv2_x、conv3_x、conv4_x 和 conv5_x。

每种 ResNet 按深度分为 18 层、34 层、50 层、101 层和 152 层。

可以看出，从 50 层之后，conv2_x、conv3_x、conv4_x 和 conv5_x 都采取 3×3 瓶颈模块，以减少计算量和参数数量。

1. 层数计算

以 101 层为例，说明层数计算。

首先经过 7×7×64 的卷积，共 1 层。

然后经过 3+4+23+3=33 个瓶颈模块，共 33×3=99 层。

最后经过全连接(fc)层进行分类，共 1 层。

共计 1+99+1=101 层，因此称为 101 层。

在计算层数时，仅包括卷积层和全连接层，不包括池化层等。