残差网络研究综述 *

郭玥秀，杨 伟†，刘 琦，王 玉

(河南大学 计算机与信息工程学院，河南 开封 475004)

摘 要：残差网络是深度学习领域中频繁采用的深度神经网络架构，其通过引入捷径连接有效地缓解了因为增加网络层数而产生的梯度消失或梯度爆炸问题，从而显著地提高了深度网络的训练速度和预测精度。自提出以来，残差网络获得了广泛关注，目前已经出现了多种改进变体。首先概述了残差网络的研究背景及意义；然后对残差单元和残差网络的框架进行了综述，随后从残差单元、网络框架和混合改进三方面阐述了残差网络的模型改进；最后总结了残差网络在一些领域的成功应用和未来可能的发展趋势。

关键词：残差网络；深度学习；神经网络；捷径连接；梯度消失；梯度爆炸

中图分类号：TP183 doi: 10.19734/j.issn.1001-3695.2018.12.0922

Survey of residual network

Guo Yuexiu, Yang Wei†, Liu Qi, Wang Yu

(*School of Computer & Information Engineering*, *Henan University*, *Kaifeng Henan* 475004, *China*)

Abstract: Residual network is a deep neural network architecture frequently used in the field of deep learning. It effectively mitigates the vanishing gradient or exploding gradient caused by increasing the number of network layers by introducing shortcut connections, thus significantly improving the training speed and prediction accuracy of deep network. Since the introduction, the residual network has gained wide attention, and various improvements have been proposed. Firstly, the research background and significance of the residual network are summarized. Then summarizes the residual unit and the framework of the residual network. The improved model of residual network is introduced from three aspects: residual unit, network framework and hybrid improvement. Finally, the successful applications of residual network in some fields and possible future development trends are summarized.

Key words: residual network; deep learning; neural network; shortcut connection; vanishing gradient; exploding gradient

0 引言

深度神经网络是由多个非线性处理层堆叠而成的机器学习模型。自从 AlexNet[1]在 2012 年以显著优势赢得 ILSVRC 比赛以来，深度神经网络已经成功地应用于图像识别、目标检测、语音识别、机器翻译、自动汽车驾驶、目标追踪和生物信息学等多个领域。特别地，越来越多的研究人员开始认识到网络深度的重要性[2]。VGGNet[3]通过把网络深度增加到 19 层，获得了 7.3%的 top-5 误差率。GoogleNet[4]则采用了 22 层的网络深度。此外，越来越深的卷积神经网络[3~10]在 ImageNet 或其他基准数据集上取得了更好的性能。更深层的网络导致更好的结果，多层特征可以通过网络深度来丰富其表达。

但是随着网络层数的增加，梯度消失或梯度爆炸[11~13]问题却出现了，这将导致深度神经网络在训练时难以收敛。然而随着归一化初始化[13~16]和中间归一化层[17]等的提出，这一问题在一定程度上得到了解决。特别地，针对极深神经网络难以训练问题，Srivastava 等人[18,19]提出了 Highway 网。该网络修改了每一层的激活函数，使前面一层的信息，有一定比例可以不经过矩阵乘法和非线性变换直接传输到下一层，仿佛一条信息高速公路，因此得名 Highway 网。Highway 网允许信息畅通无阻地流入更深层。在训练阶段，调整快捷连接参数以控制这些高速公路上允许的信息量。Highway 网的优点是使用随机梯度下降算法可以训练几百乃至上千层深的网络。然而 Highway 网却有两个明显的缺点，即训练速度慢且不能通过增加网络深度来显著地提升其性能。

当深度网络的收敛问题解决后，研究人员随后发现了网络退化问题。20 层以上的深度网络，随着网络层数的增加，分类性能却越来越差。特别地，50 层网络的测试误差率是 20 层网络的一倍。针对该问题，He 等人[20]在 Highway 网的基础上提出了残差网络。该网络采用捷径连接取代 Highway 网中的网关单元，可以保留全部的原始信息，并且减少了网络参数。残差网络的最大优点是不仅能够加速超深神经网络的训练，而且可以大幅提升深度网络的准确率。此外，残差网络在很大程度上避免了随着网络层数的增加而产生的梯度消失或梯度爆炸问题，这让训练极深的网络成为可能。残差网络是深度学习领域中的一个里程碑式突破。由于残差网络的良好性能，其在深度学习领域获得了广泛关注。目前，已经提出了多个改进变体。

本文综述了残差网络的最新研究进展，首先介绍了残差单元及其网络框架，然后论述了残差网络的模型改进及其应用，最后对本文进行了总结。

1 残差网络

残差网络的基本构建块是残差单元。本章首先对残差单元进行详细介绍，然后描述残差网络的框架。

1.1 残差单元

残差单元由卷积 Conv 层、批处理归一化 BN 层和非线

收稿日期：2018-12-26；**修回日期：**2019-03-07 **基金项目：**国家自然科学基金资助项目（61806074）

作者简介：郭玥秀（1994-），女（回族），重庆江北人，硕士研究生，主要研究方向为深度学习；杨伟（1983-），男（通信作者），河南信阳人，副教授，博士，主要研究方向为机器学习、深度学习（yang0sun@gmail.com）；刘琦（1996-），男，河南信阳人，硕士研究生，主要研究方向为深度学习；王玉（1993-），女，河南信阳人，硕士研究生，主要研究方向为深度学习.

性激活函数 ReLU 层堆叠而成。图 1 给出了一个残差单元的示意图。令第 l 个残差单元的输入为 x_l，则其输出可形式化为执行如下的数学计算：

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

其中：$F(x_l, W_l)$ 是残差函数；W_l 是该残差函数对应的权重参数；$f(\bullet)$ 是非线性激活函数 ReLU。

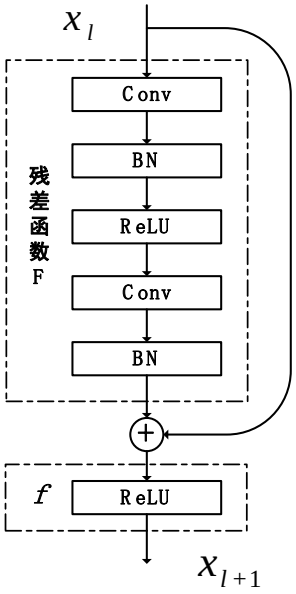


图 1　残差单元

Fig. 1　Residual unit

在式(1)中，x_l 和 $F(x_l, W_l)$ 的维度必须相同。如果不相同，当改变了输入/输出的通道，可以通过捷径连接执行一个线性映射 W_s 来匹配两者的维度。

$$x_{l+1} = f(W_s x_l + F(x_l, W_l)) \tag{2}$$

残差函数 F 的形式是灵活可变的，除了堆叠两层卷积层外，还可以堆叠三层卷积层。

1.2　残差网络的框架

残差网络的标准框架如图 2 所示。给定输入数据，首先残差网络将输入数据依次送入卷积层 Conv、非线性激活函数层 ReLU 和批处理归一化层 BN；然后处理的结果进一步送入到多个残差单元，再经过批处理归一化层 BN 和多个全连接层；最后得到输出结果。对于该框架中的参数，如卷积核的大小、卷积核的数量、执行卷积时的步幅等可以根据不同的学习任务进行调整，有时会在残差单元之后增加池化层进行降维。

2　残差网络的模型改进

自从残差网络被提出以来，已经出现了多个改进版本。概括起来，残差网络的改进方式可以粗略地分为基于残差单元的改进、基于网络框架的改进和混合改进三类。

2.1　基于残差单元的改进

对于图 1 中的残差单元，其最大的问题是函数 f 是非线性激活函数。然而当网络很深时，函数 f 会使得反向传播的梯度值趋近于零，不利于信息流在网络中的传播。当然，这也会增加网络的训练时间。为此，改进的残差单元都把函数 f 替换为恒等映射函数。在恒等映射函数下，残差单元的输出可以形式化为

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

给定第 l 个残差单元的输入，则较深的第 L 个残差单元的输出通过递归可以形式化为

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

假设目标损失函数为，从反向传播的链式法则可以得到

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

式(4)和(5)表明了在前向和反向阶段，信号都能够直接地从一个单元传递到其他任意一个单元。

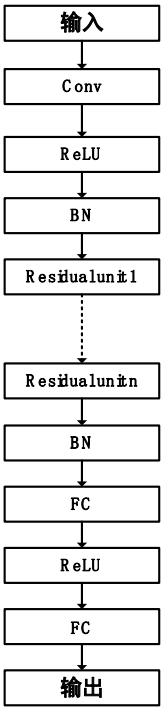


图 2　残差网络框架

Fig. 2　Framework of residual network

图 3 给出了改进残差单元的简化框架。基于残差单元的改进主要是通过修改框架中的残差函数进行的。对残差函数的修改主要表现在两方面：a)修改表示残差函数的网络层组合，代表的残差网络模型有“预激活”残差网络[21]、引入指数线性单元的残差网络[22]和宽残差网络[23]；b)把式(3)中的残差函数形式化为多个残差函数的和，每个残差函数仍通过多个网络层的堆叠进行表示，代表的残差网络模型有多重残差网络[24]、ResNeXt[25]、Shake-Shake 正则化的残差网络[26]。下面对其进行一一介绍。

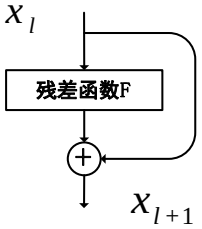


图 3　简化的残差单元

Fig. 3　Simplified residual unit

2.1.1　“预激活”残差网络[21]

“预激活”残差网络将残差函数中的激活函数 ReLU 和 BN 转移到卷积层之前，形成一种“预激活(pre-activation)”的方式。图 4 给出了该网络采用的残差函数。实验结果表明，预激活可以加强对模型的正则化，从而减少过拟合的影响。

2.1.2　引入指数线性单元的残差网络

为了缓解梯度消失问题，Shah 等人[22]使用指数线性单元(exponential linear unit ，ELU)[27]代替标准残差函数中 BN[17]层和 ReLU[28]层的组合。ReLU 函数输出的值均大于 0，当激活值的均值非 0 时，就会对下一层产生一个偏置值。如果激活值之间不能相互抵消(即均值非 0)，会导致下一层的激活单

元有偏移。如此叠加，单元越多时，偏移就会越大。相比 ReLU，ELU 可以取负值，减少了向前传播的变化和信息，这让单元激活均值可以更接近 0，类似于 BN 的效果但是只需要更低的计算复杂度。

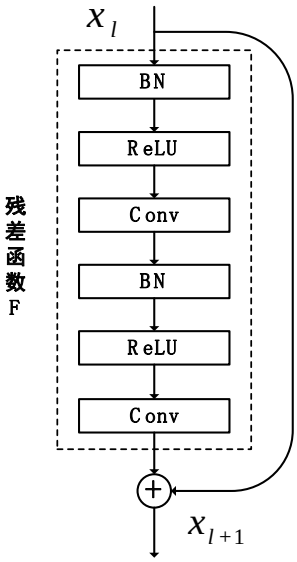


图 4　预激活”残差单元

Fig. 4　Pre-activation residual unit

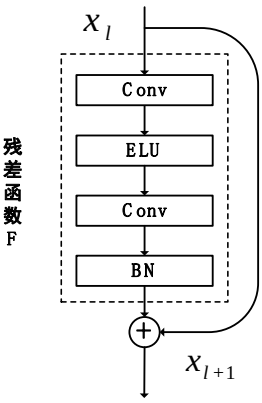


图 5　引入 ELU 的残差单元

Fig. 5　Residual unit with ELU

ELU 的公式为

$$f(x)=\begin{cases}x, x>0\\ \alpha(\exp(x)-1), x\le 0\end{cases} \tag{6}$$

其中：α 是一个可调整的参数，它控制着 ELU 负值部分在何时饱和。ELU 可将前面单元输入的激活值均值控制在 0，让激活函数的负值部分也可以携带信息并被使用。在降低计算复杂度的同时，还可以加快残差网络的学习速度。图 5 给出了引入指数线性单元后的残差函数。

2.1.3 宽残差网络

虽然通过增加残差网络的深度可以提高分类的精度，但有时为了少量的精度增加需要将网络层数翻倍。这不仅增加了计算成本，而且减少了特征的重用，也降低了网络的训练速度。为此，Sergey 等人[23]从“宽度”的角度入手提出了宽残差网络，即在减少残差网络深度的同时增加网络的宽度。这里的“宽度”涉及的是特征映射的通道数，在卷积层中是指增加卷积核的个数。宽残差网络的实验结果表明，宽残差网络的性能优于传统的窄而深的神经网络。一个简单的 16 层宽残差网络与 1 000 层的残差网络具有相同数量的参数和准确率，但训练速度要快几倍。此外，作者通过实验证实了在残差函数的卷积层之间引入 Dropout[29]，可以有效地防止过拟合。

2.1.4 多重残差网络

残差网络的一个缺点是每一个百分比的改进都需要显著地增加层数，这会导致网络计算和存储成本的增加[20]。为此，Abdi 等人[24]引入了多重残差网络(Multi-ResNet)。图 6 给出了其采用的残差单元。该网络可以在保持深度固定的同时，通过增加每个残差单元中残差函数的数量来增加网络的多样性，以提高分类精度。相对于 110 层的超深残差网络，Multi-ResNet 可看做是一个浅层多残差聚合网络。较浅的集合在优化期间对梯度更新有显著的贡献。特别地，通过对残差函数进行分组，把不同组的残差函数分配给不同的 GPU 进行计算，可以实现多 GPU 并行计算，显著提高网络的训练速度。

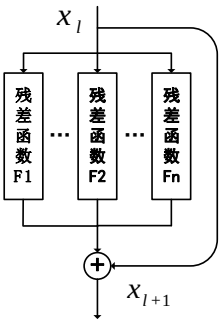


图 6　多重残差网络的残差单元

Fig. 6　Residual unit in Multi-resnet

2.1.5 ResNeXt 残差网络

为了提高预测的准确率而对残差网络进行加深或加宽时，超参数的数量会显著地增加并导致网络的计算开销增大。为此，Xie 等人[25]通过采用多分支的同构结构而提出了一种新的残差网络 ResNeXt。该网络引入了一个被称为“基数(cardinality)”的超参数——即独立路径的数量，从而提供了一种调整模型容量的新方式。图 7 给出了基数为 32 的 ResNeXt 构建块。对于残差函数中的每个卷积层，第一个是输入的通道数，中间是卷积核的大小，最后是输出的通道数。特别地，类似于标准的残差单元，求和后的结果会送到非线性激活函数 ReLU 中。从图中可以看出，ResNeXt 的残差函数是多个子残差函数的累加和。实际上，ResNeXt 同时借用了 VGGNet[3] 的堆叠思想和 Inception[4,17,30~34] 的 split-transform-merge 思想。由于拓扑结构相同，ResNeXt 可以采用组卷积进行高效地实现。ResNeXt 可以在不增加参数复杂度的前提下提高准确率，同时还减少了超参数的数量。实验表明，通过扩大“基数”值提高准确率比让网络加深或加宽来提高准确率更有效。

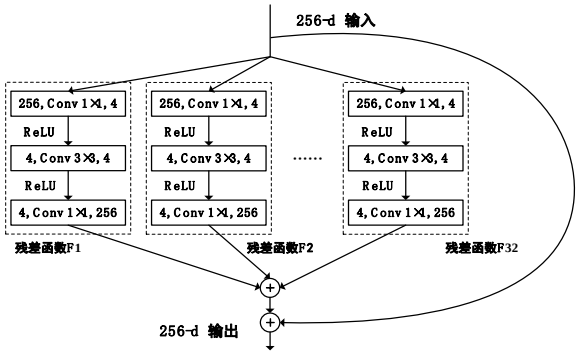


图 7　基数为 32 的 ResNeXt 构建块

Fig. 7　Building block of resnext with cardinality = 32

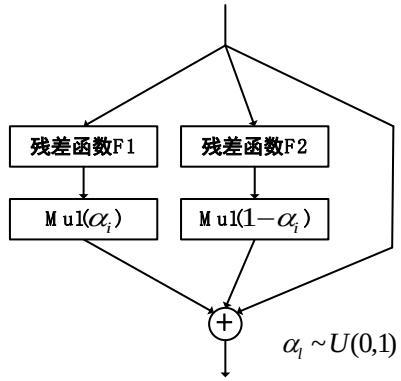
2.1.6 Shake-Shake 正则化残差网络

当训练数据集中的样本相对于学习参数较少时，残差网络容易过拟合。为了克服过拟合，许多正则化方法如权值衰

减、Dropout 正则化[29]、批处理归一化和早停法等都已经引入到残差网络。特别地，针对三分支网络的过拟合问题，Gastaldi[26]提出了 Shake-Shake 正则化残差网络。该网络通过采用随机仿射组合替换并行分支的标准求和来提高多分支网络的泛化能力。图 8 给出了 Shake-Shake 正则化残差网络的残差单元。对于训练时的前向传播，第一个残差单元的输出可以形式化为

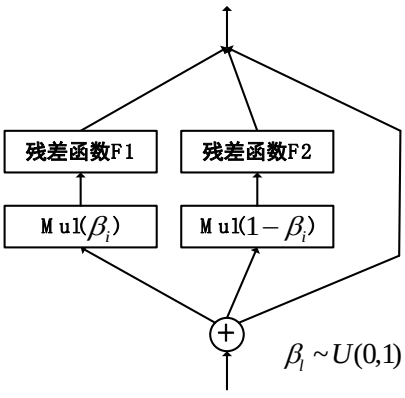
$$x_{l+1} = x_l + \alpha_l F_1(x_l, W_l^1) + (1-\alpha_l) F_2(x_l, W_l^2) \tag{7}$$

其中：α_l 是采样于均匀分布 U(0,1)的随机值。在每一次批处理中，α_l 的值都进行重新采样。在进行测试时，α_l 的值固定为其期望值 0.5。由于在训练过程中向梯度添加噪声有助于神经网络的训练和泛化[35,36]，所以 Shake-Shake 正则化残差网络进一步在反向传播时引入采样于均匀分布 U(0,1)的随机值 β_l 来增加梯度的多样性，如图 8(b)所示。在训练过程中，这些导致了前向信息流和反向信息流的随机混合，从而达到了更好的泛化效果。另外，文献[26]中的实验结果表明，在前向和反向传播时，为每一个通道采用不同的随机值能够进一步提高 Shake-Shake 正则化残差网络的泛化性能。



(a) 前向传播

(a) Forward propagation



(b) 反向传播

(b) Reverse propagation

图 8　Shake-Shake 正则化网络的残差单元

Fig. 8　Shake-Shake regularized residual unit

此外，考虑到 Shake-Shake 正则化只能应用于具有多分支的 ResNeXt 中，Yamada 等人[37]将随机深度(stochastic depth)思想[32]与 Shake-Shake 正则化进行结合进一步提出了 ShakeDrop 正则化网络。对于只有一个残差函数的残差单元，ShakeDrop 采用如式（8）计算第一个残差单元的输出。

$$x_{l+1} = \begin{cases} x_l + \alpha_l F(x_l, W_l) & \text{如果 } b_l = 0 \\ x_l + F(x_l, W_l) & \text{如果 } b_l = 1 \end{cases} \tag{8}$$

其中：$b_l \in \{0,1\}$ 是由指定概率 p_l 决定的伯努利随机变量；α_l 是采样于均匀分布 U(-1,1)的随机值。对于 $b_l = 0$ 时的反向传播，类似于 Shake-Shake 正则化，ShakeDrop 采用独立于 α_l 的随机变量 $\beta_l \sim U(0,1)$ 进行参数更新。特别地，ShakeDrop 不仅可以应用于 ResNeXt，还可以应用于 ResNet 和 PyramidNet。

2.2　基于网络框架的改进

原始的残差网络结构仍然存在一些缺陷，难以在非常深的网络上进行收敛，通过改进残差网络框架，可以提高残差网络的学习能力，并且更容易优化。本节将分别介绍 ROR[38]、深层“金字塔”残差网络[39]、加权残差网络[40]。

2.2.1 ROR

为了挖掘残差网络的优化能力，Zhang 等人[38]等人提出了一种新的方法——Residual networks of Residual networks（ROR）。ROR 放弃优化原始残差映射，转而优化残差映射中的残差映射。特别是 ROR 在原始残差网络上增加了层级智能快捷连接，然后构建了一个多级网络，如图 9 所示，以提高残差网络的学习能力。为了缓解过拟合，利用 drop-path 方法训练 ROR，并获得了明显的性能提升。在各种残差网络上建立 ROR，发现 ROR 不仅适用于原始残差网络，而且还适用于其他残差网络。ROR 显著提高了图像分类性能，可以改善各种残差网络。

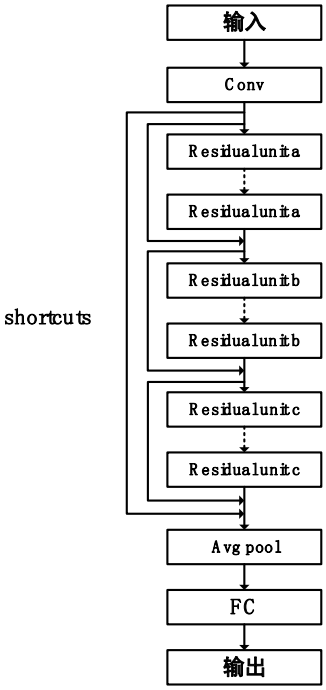


图 9　ROR 网络框架

Fig. 9　Framework of ROR

2.2.2 深层“金字塔”残差网络

Veit 等人[41]论证了残差网络可以看做是相对较浅网络的集合。他们的研究表明，从残差网络中删除一个单独的残差单元，即只保留一个快捷连接，不会对整体性能产生显著影响，从而证明删除残差单元等同于删除集成网络中的一些浅层网络。然而在普通网络如 VGG 网络中，删除任意一个网络层都会导致严重错误。

在残差网络中删除具有下采样功能残差单元中的构建块，会造成特征映射维度加倍，增大分类误差。但是当使用随机深度[32]训练残差网络时，删除前述的构建块不会降低分类性能。

随后，Han 等人[39]设计了一个深层“金字塔”残差网络(pyramidal residual network，PyramidNet)，如图 10 所示。该网络的宽度随深度的增加而逐渐增加，这类似于形状从顶部向下逐渐变宽的金字塔结构。

通过逐渐增加特征映射的通道数来代替在每个具有下采样功能的残差单元中急剧增加特征映射的通道数。另外，当增加特征映射维度时，设计的网络架构通过使用零填充(zero-padded)恒等映射捷径连接将普通网络和残留网络混合

使用，如图 11 所示。这种设计不会导致过度拟合问题，因为不存在其他参数，并且令人惊讶的是，与其他快捷方式相比，它显示出显著的泛化能力；并且使用新网络架构，删除具有下采样功能的残差单元不会降低性能。

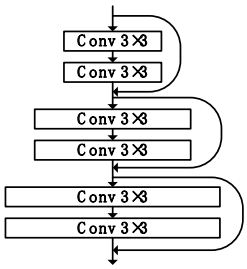


图 10　深层“金字塔”残差单元

Fig. 10　Pyramidal residual unit

2.2.3 加权残差网络

针对原始残差网络中 ReLU 和元素加法之间不兼容以及深度网络的初始化问题，Shen 等人[40]提出加权残差网络。该网络能够高效地组合来自不同层的残差，如图 12 所示。残差逐渐加入到 highway 信号中，使训练过程更可靠，即使没有“预激活”[21]策略，深度超过 1 000 层的网络也能快速收敛。此外，加权残差网络的计算量和 GPU 内存负担比原来的残差网络少得多。加权残差网络通过投影随机梯度下降进行优化。随着深度从 100 层增加到 1 000 层以上，加权残差网络模型在准确性和收敛性方面有很大的提高。

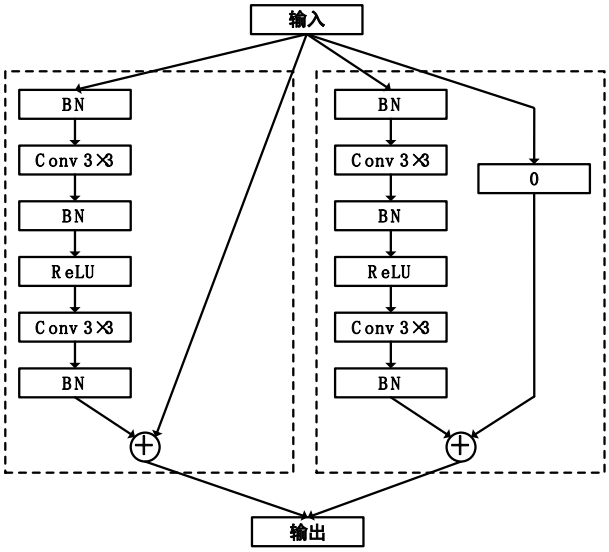


图 11　具有零填充恒等映射捷径连接的残差单元

Fig. 11　Residual unit with zero-padded identity mapping shortcut

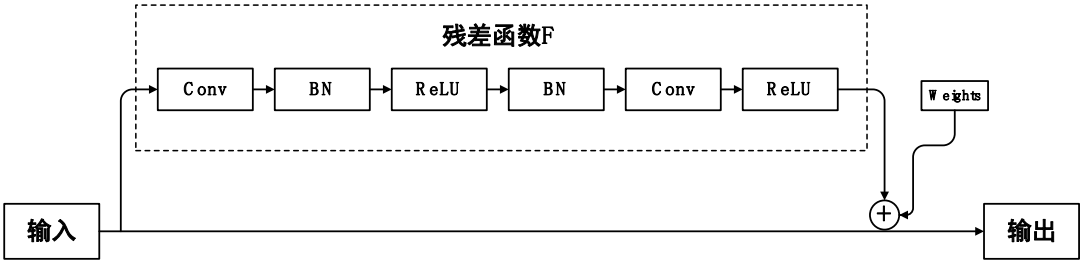


图 12　加权残差单元

Fig. 12　Weighted residual unit

2.3　混合改进

2.3.1 随机深度残差网络

随机深度残差网络(ResDrop)[32]在训练时使用伯努利随机变量随机禁用部分残差单元，而测试时使用全部深度的网络进行预测。训练期间，随机深度残差网络的深度会减小，进而会导致前向传播和反向传播的深度变短，所以其训练时间不会随网络的深度而线性增加。此外，训练期间网络深度的减少会增强早期层参数的梯度，这将使得 1 000 层以上的随机深度残差网络能够正常训练。特别地，但随机深度残差网络可视为不同深度网络的隐式集成，所以与恒定深度的深度残差网络相比不易过拟合。

此外，Yamada 等人[42]进一步把随机深度引入到“金字塔”残差网络框架中，提出了 PyramidSepDrop 网络模型，如图 13 所示。其中，含有交叉的圆表示随机深度的下降机制；F(x)是该机制的残差函数，由两部分组成，上部分用于增加通道，下部分的通道数与输入相同。

2.3.2 扩张残差网络

在卷积神经网络中使用池化和下采样会降低特征图的空间分辨率，使得丢失许多细节，从而影响模型对小型目标乃至目标之间关系的识别。这不仅限制了分类的准确率，也影响到把模型迁移到其他任务上的性能。Yu 等人[43]在原始残差网络[20]中加入扩张卷积来解决空间分辨下降问题，提出扩张残差网络(DRN)。扩张卷积可以增加神经元的感受野，使用扩张卷积替换残差网络[20]内部的下采样层，保持原有网络的感受野，并且还能提高图片空间分辨率。DRN 的目的是在不降低特征图分辨率与计算开销的情况下尽可能地增加卷积核的感受野。文献[20]中的每个 ResNet 模块有 5 组卷积，DRN 依次对后两组卷积进行修改。首先去除顶层的下采样，保持特征图的分辨率。原始残差网络第 4 和第 5 组卷积的空间分辨率分别下降 2 倍和 4 倍，因此对第 4 和第 5 组卷积依次使用 2 倍和 4 倍扩张率的扩张卷积。最后连接全局平均池化层和全卷积层得到分类的输出。经过扩张卷积操作特征图的尺度没有发生任何改变，并且感受野和原残差网络对应层一样。

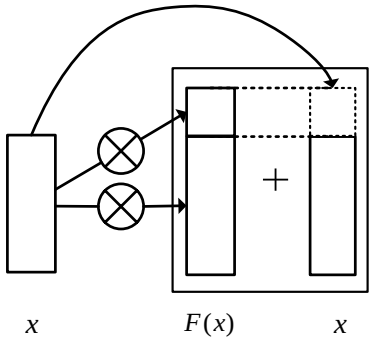


图 13　PyramidSepDrop 网络模型的残差块

Fig. 13　Residual block of pyramidsepdrop

DRN 可以直接由输入生成高分辨率的输出特征图，无须增加任何其他层、多余参数和再训练模型，但是扩张卷积会带来网格化现象，直接影响网络的性能。因此作者设计了消除网格化（degridding）的结构：在整个网络后面再加两个扩张率较小的卷积层，且去掉其中的捷径连接，从而使图像分类和目标位置检测效果都优于原始残差网络。

2.3.3 可逆残差网络

针对单个网络框架无法同时有效处理分类、密度估计、

生成任务的缺点，Behrmann 等人[44]基于残差网络是常微分方差的欧拉离散化思想提出了可逆残差网络。该方法仅需要改变标准残差网络的归一化机制，其可以利用深度学习框架中的已有实现。这种方法允许每个残差块的无约束架构，且每个残差块的 Lipschitz 常数只需要小于 1。此外，可逆残差网络定义了一个生成模型，通过最大似然算法可对其在无标签数据集上进行训练。在计算似然度时，可逆残差网络采用了近似的方法计算残差块的雅可比对数行列式。与 FFJORD[45]类似，可逆残差网络具有自由形式的雅可比行列式，这使得其能够学习比其他可逆模型使用的三角形映射更具表达性的变换。文献[46]提出了能够高效利用内存的深度可逆网络实现。实验结果表明，可逆残差网络的性能堪比当前最优的图像分类器和基于流的生成模型，它将通用架构应用在现实中又推进了一步。

3 残差网络的应用研究

残差网络由于其表征能力强，已被应用到计算机视觉、图像识别、行人检测等相关领域的模型中。对残差网络的应用研究主要从以下两方面展开：

a)模型应用。由于残差网络在 ILSVRC 2015 的成功，Google 开始借鉴残差网络的思想，重新评估 CNN 深度模型的设计，提出了 Inception-ResNet-v2[30]模型。证明在 Inception 网络中加入残差学习可以加快深度网络的训练速度，识别性能也显著提高[47]。文献[48]受到 ResNet 将输出与输入相加形成残差结构的启发，设计出一种将输出与输入并联，以实现每一层都能直接得到之前所有层的输出的密集卷积网络(dense convolutional network，DenseNet)。该网络可以缓解消失梯度问题，加强特征传播，鼓励特征重用，并大幅减少参数数量。DenseNet 提出后得到了广泛的应用，如语义分割[49]、语音识别[50]和图像分类[51]等。Chen 等人[52]受到递归神经网络(recurrent neural networks，RNN)[53]的启发，证明 ResNet 可以看做是 DenseNet[48]在跨层参数共享时的特例。深度残差网络通过残差路径隐含地重用这些特征，而 DenseNet 通过密集连接的路径不断探索新特征。结合 ResNet 和 DenseNet 两者的优点，提出了一个简单、高效和模块化的双路径网络结构(dual path networks，DPN)。这种新架构继承了 ResNet 和 DenseNet 的优点，通过双路径架构保持探索新特性的灵活性，实现有效的功能重用，提高参数效率，降低计算成本和内存消耗，具有很强的推广性。Zhang 等人[54]受到 ResNeXt 模型可有效降低超参数数量的启发，为移动端低功耗设备设计出一种更为高效的卷积神经网络——ShuffleNet，在大幅降低模型计算复杂度的同时仍然保持了较高的识别精度。文献[55]利用 101 层的残差网络设计出 IDW-CNN 模型，并且联合 VOC12 和 IDW 两个数据集共同训练该模型。在充分探索不同数据集的知识的同时，两个数据集的知识还可以相互转移，使语义图像分割性能大大提高。

b)领域应用。利用 ResNet 的残差学习模型和 VGG-19 的卷积核结构，张帆等人[56]设计出一种新的残差块，利用它可构建深度神经网络，并将其应用于脱机手写汉字的识别。裴颂文等人[57]成功设计网中网残差网络模型(NIN_ResNet)对表情图像做分类识别研究。在基于视频的行人重识别领域，文献[58]在 50 层残差网络的基础上，在分类训练时加入一个时域平均池化层、一个全连接层和一个分类层,有效地利用未标记数据，减少基于单标注样本视频行人重识别的注释工作量。在图像去噪领域，Zhang 等人[59]借鉴残差网络的思想，设计出较浅层的 CNN 网络，在不同噪声水平上训练，有效 f 实现了图像去噪。近年来，在图像超分辨率(super-resolution，SR)方面，非常深的卷积神经网络(CNN)取得了巨大的成功，并提供了分层特征。但是大多数基于 CNN 的深层次的 SR 模型并没有充分利用原始低分辨率(low-resolution，LR)图像的分层特征，所以性能较低。致力于建立更深且更简洁的网络，Tai 等人[60]提出了用于单幅图像超分辨率(single image super-resolution，SISR)的深度递归残差网络模型(deep recursive residual network，DRRN)。SISR 旨在将低分辨率图像(LR)恢复为高分辨率图像(HR)，可用于安全和监视成像[61]、医学成像[62]、卫星成像[63]等各种计算机视觉任务。DRRN 提出残差单元的递归学习，以保持模型紧凑，并且叠加多个递归块以学习 HR 与 LR 图像之间的残差图像。通过这种递归学习，DRRN 无须增加任何权重参数，只需增加深度就能提高准确性。Lim 等人[64]利用残差网络开发了一个增强型超深度超分辨率网络(EDSR)，性能超过了当前最先进的 SR 方法。该模型通过从常规 ResNet 架构中去除不必要的模块，使模型更紧凑。该模型还采用残差缩放技术在稳定训练程序的同时扩大模型尺寸，进一步提高了模型性能。但是 Zhang 等人[65]认为 EDSR 忽略了充分利用每个卷积层的信息，也忽略使用分层特征进行重构，因此提出了使用残差密集块(RDB)的残差密集网络，通过密集连通卷积层提取到丰富的局部特征，以充分利用原始 LR 图像的所有分层特征。

残差网络一直以来是深度学习领域的研究热点，但是仍存在一定问题，其网络模型的准确率是脆弱的，并且易受数据分布中微小自然变化的影响，鲁棒性有待进一步提高。

4 结束语

残差网络的研究取得了一定的成果，研究日益趋向成熟。因此，本文对残差网络的研究背景、意义进行了简要概述，重点从残差网络的基本原理、模型改进等方面对残差网络的最新研究进展进行了综述。此外，本文还探讨了残差网络在一些深度模型上的扩展应用。残差网络的成功应用仍将促进深度学习相关领域的进一步发展。

参考文献：

[1] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [C]// Proc of International Conference on Neural Information Processing Systems. [S.l.]:Curran Associates Inc, 2012: 1097-1105.

[2] Gates, G. The reduced nearest neighbor rule (corresp.) [J]. IEEE Trans on Information Theory, 1972, 18 (3): 431-433.

[3] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [C]//Proc of International Conference on Learning Representations. 2015.

[4] Szegedy C, Liu Wei, Jia Yangqing, *et al.* Going deeper with convolutions [C]//Proc of International Conference on Computer Vision and Pattern Recognition. 2015: 1-9.

[5] Zou W Y, Wang Xiaoyu, Sun Miao, *et al.* Generic object detection with dense neural patterns and regionlets [J]. Eprint Arxiv, 2014.

[6] Lin Min, Chen Qiang, Yan Shuicheng. Network in network [C]//Proc of International Conference on Learning Representations. 2014.

[7] Sermanet P, Eigen D, Zhang Xiang, *et al.* OverFeat: integrated recognition, localization and detection using convolutional networks [C]//Proc of International Conference on on Learning Representations. 2014.

[8] Romero A, Ballas N, Kahou S E, *et al.* FitNets: Hints for thin deep nets [J]. Computer Science, 2014.

[9] Lee C Y, Xie Saining, Gallagher P, *et al.* Deeply-supervised nets[C]//Proc of International Conference on Artificial Intelligence and Statistics. 2015: 562-570.

[10] Springenberg J T, Dosovitskiy A, Brox T, *et al.* Striving for simplicity: the all convolutional net [C]//Proc of International Conference on Learning Representations. 2014.

[11] Hochreiter S. Untersuchungen zu dynamischen neuronalen Netzen [C]// Proc of Master's Thesis, Institut Fur Informatik, Technische Universitat, Munchen. 1991.

[12] Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult [J]. IEEE Trans on Neural Networks, 1994, 5 (2): 157-166.

[13] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks [J]. Journal of Machine Learning Research, 2010, 9: 249-256.

[14] Saxe A M, McClelland J L, Ganguli S. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks [J]. Computer Science, 2013.

[15] He Kaiming, Zhang Xiangyu, Ren Shaoqing, *et al.* Delving deep into rectifiers: surpassing human-level performance on ImageNet classification [C]//Proc of IEEE International Conference on Computer Vision and Pattern Recognition. 2015: 1026-1034.

[16] Orr G B , M ü ller K R. Neural networks: tricks of the trade, this book is an outgrowth of a 1996 NIPS workshop [C]// Proc of Neural Networks: Tricks of the Trade, this book is an Outgrowth of a 1996 NIPS Workshop. London, UK: Springer-Verlag, 1998: 103-105.

[17] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift [C]// Proc of the 32nd International Conference on International Conference on Machine Learning. 2015: 448-456.

[18] Srivastava R K, Greff K, Schmidhuber J. Highway networks [J]. Computer Science, 2015.

[19] Srivastava R K, Greff K, Schmidhuber J. Training very deep networks [J]. Computer Science, 2015.

[20] He Kaiming, Zhang Xiangyu, Ren Shaoqing, *et al.* Deep residual learning for image recognition [C]// Proc of IEEE International Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[21] He Kaiming, Zhang Xiangyu, Ren Shaoqing, *et al.* Identity mappings in deep residual networks [C]// Proc of European Conference on Computer Vision. 2016: 630-645.

[22] Shah A, Kadam E, Shah H, *et al.* Deep residual networks with exponential linear unit [C]// Proc of International Symposium on Computer Vision and the Internet. New York: ACM Press, 2016: 59-65.

[23] Zagoruyko S, Komodakis N. Wide residual networks [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[24] Abdi M, Nahavandi S. Multi-residual networks: improving the speed and accuracy of residual networks[C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[25] Xie Saining, Girshick R, Dollar P, *et al.* Aggregated residual transformations for deep neural networks [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]:IEEE Computer Society, 2017: 5987-5995.

[26] Gastaldi X. Shake-Shake regularization [Z]. 2017.

[27] Clevert, Djork-Arné, Unterthiner T, *et al.* Fast and accurate deep network learning by exponential linear units (ELUs) [J]. Computer Science, 2015.

[28] Nair V, Hinton G E. Rectified linear units improve restricted Boltzmann machines [C]// Proc of International Conference on International Conference on Machine Learning. 2010: 807-814.

[29] Srivastava N, Hinton G, Krizhevsky A, *et al.* Dropout: a simple way to prevent neural networks from overfitting [J]. Journal of Machine Learning Research, 2014, 15 (1): 1929-1958.

[30] Szegedy C, Ioffe S, Vanhoucke V, *et al.* Inception-v4, Inception-ResNet and the impact of residual connections on learning[C]// Proc of IEEE Computer Vision and Pattern Recognition. 2016.

[31] Szegedy C, Vanhoucke V, Ioffe S, *et al.* Rethinking the Inception Architecture for Computer Vision [C]// Proc of IEEE Computer Vision and Pattern Recognition. 2016: 2818-2826.

[32] Huang Gao, Sun Yu, Liu Zhuang, *et al.* Deep networks with stochastic depth [C]// Proc of European Conference on Computer Vision. Berlin: Springer, 2016: 646-661.

[33] Larsson G, Maire M, Shakhnarovich G. FractalNet: ultra-deep neural networks without residuals [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[34] DeVries T, Taylor G W. Dataset augmentation in feature space [Z]. 2017.

[35] An Guozhong. The effects of adding noise during backpropagation training on a generalization performance [J]. Neural Computation, 1996, 8 (3): 643-674.

[36] Neelakantan A, Vilnis L, Le Q V, *et al.* Adding gradient noise improves learning for very deep networks [J]. Computer Science, 2015.

[37] Yamada Y, Iwamura M, Kise K. ShakeDrop regularization [Z]. 2018.

[38] Zhang Ke, Sun Miao, Yuan Xingfang, *et al.* Residual networks of residual networks: multilevel residual networks [J]. IEEE Trans on Circuits & Systems for Video Technology, 2018, 28(6): 1303-1314.

[39] Han D, Kim J, Kim J. Deep pyramidal residual networks [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 6307-6315.

[40] Shen Falong, Gan Rui, Zeng Gang. Weighted residuals for very deep networks [C]// Proc of International Conference on Systems and Informatics. Piscataway, NJ:IEEE Press, 2017: 936-941.

[41] Veit A, Wilber M, Belongie S. Residual networks behave like ensembles of relatively shallow networks [C]// Advances in Neural Information Processing Systems. 2016.

[42] Yamada Y, Iwamura M, Kise K. Deep pyramidal residual networks with separated stochastic depth [J]. 2016.

[43] Yu F, Koltun V, Funkhouser T A. Dilated residual networks [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2017: 636-644.

[44] Behrmann J, Grathwohl W, Chen R T Q, *et al.* Invertible residual networks [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[45] Grathwohl W, Chen R T Q, Bettencourt J, *et al.* FFJORD: free-form continuous dynamics for scalable reversible generative models [J]. arXiv: Learning, 2018.

[46] Sil C, Jonas T, Rashindra M. MemCNN: a Framework for developing memory efficient deep invertible networks [Z]. 2018.

[47] Russakovsky O, Deng Jia, Su Hao, *et al.* ImageNet large scale visual recognition challenge [J]. International Journal of Computer Vision, 2015, 115 (3): 211-252.

[48] Huang Gao, Liu Zhuang, Maaten L V D, *et al.* Densely connected convolutional networks [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 2261-2269.

[49] Krešo I, Oršić M, Bevandić P, *et al.* Robust semantic segmentation with ladder-densenet models [J]. ResearchGate, 2018 (6): 36.

[50] Park S, Jeong Y, Kim H S. Multi-resolution DenseNet based acoustic models for reverberant speech recognition [J]. Phonetics & Speech

Sciences, 2018, 10 (1): 33-38.

[51] Liu Wenqi, Zeng Kun. SparseNet: a sparse densenet for image classification [J]. ResearchGate. 2018 (4): 55.

[52] Chen Yunpeng, Li Jianan, Xiao Huaxin, *et al.* Dual path networks [Z]. 2017.

[53] Liao Qianli, Poggio T. Bridging the gaps between residual learning, recurrent neural networks and visual cortex [Z]. 2016.

[54] Zhang Xiangyu, Zhou Xinyu, Lin Mengxiao, *et al.* ShuffleNet: an extremely efficient convolutional neural network for mobile devices [C]// Proc of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2018: 6848-6856.

[55] Wang Guangrun, Luo Ping, Lin Liang, *et al.* Learning object interactions and descriptions for semantic image segmentation [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE Computer Society, 2017.

[56] 张帆, 张良, 刘星, 等. 基于深度残差网络的脱机手写汉字识别研究 [J]. 计算机测量与控制, 2017 (12): 259-262. (Zhang Fan, Zhang Liang, Liu Xing, *et al.* Recognition of off-line handwritten chinese character based on deep residual network [J]. Computer Measurement and Control, 2017 (12): 259-262.)

[57] 裴颂文, 杨保国, 顾春华. 网中网残差网络模型的表情图像识别研究 [J]. 小型微型计算机系统, 2018, 39 (12) :2681-2686. (Pei Songwen, Yang Baoguo, Gu Chunhua. Research on recognition of facial expression image based on NIN_ResNet model [J]. Mini-Micro Systems, 2018, 39 (12) :2681-2686.)

[58] Wu Yu, Lin Yutian, Dong Xuanyi, *et al.* Exploit the unknown gradually: one-shot video-based person Re-Identification by stepwise learning [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition . [S.l.]:IEEE Computer Society, 2018.

[59] Zhang Kai, Zuo Wangmeng, Chen Yunjin, *et al.* Beyond a Gaussian denoiser: residual learning of deep cnn for image denoising [J]. IEEE Trans on Image Processing, 2017, 26(7): 3142-3155.

[60] Tai Ying, Yang Jian, Liu Xiaoming. Image super-resolution via deep recursive residual network [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]:IEEE Computer Society, 2017: 2790-2798.

[61] Zou W W W, Yuen P C. Very low resolution face recognition problem [J]. IEEE Trans on Image Processing A Publication of the IEEE Signal Processing Society, 2012, 21 (1): 327.

[62] Shi Wenzhe, Caballero J, Ledig C, *et al.* Cardiac image super-resolution with global correspondence using multi-atlas PatchMatch [J]. Med Image Comput Comput Assist Interv, 2013, 16 (Pt 3): 9-16.

[63] Thornton M, Atkinson P. Subpixel mapping of rural land cover objects from fine spatial resolution satellite sensor imagery using super-resolution pixel-swapping [J]. International Journal of Remote Sensing, 2006, 27 (3): 473-491.

[64] Lim B, Son S, Kim H, *et al.* Enhanced deep residual networks for single image super-resolution [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition Workshops. [S.l.]:IEEE Computer Society, 2017: 1132-1140.

[65] Zhang Yulun, Tian Yapeng, Kong Yu, *et al.* Residual dense network for image super-resolution [Z]. 2018.