深度学习在遥感图像分类中的应用

方桂安，刘梦莎，刘玥，罗秋琳，马梓场，唐迅

摘 要—准确、高效的遥感图像分类是遥感图像解析的重要研究内容之一。近年来，随着机器学习技术的发展，深度神经网络日渐成为一种有效的遥感图像分类处理方法。本文分析了遥感图像分类目前存在的一些问题，并深入分析了典型的图像分类网络的模块结构原理及功能；然后根据遥感图像分类的研究现状和深度神经网络在遥感图像分类方向应用的研究现状，总结了深度神经网络在遥感图像分类技术应用中的未来发展趋势。

关键词—遥感图像，图像分类，深度学习

I. 概述

目标物体的分类与识别一直以来都是遥感图像解析的核心内容之一。如何在满足一定精度条件下对遥感图像进行分类信息提取，成为了遥感图像研究的关键问题。提高遥感图像的分类精度能够直接促进遥感技术的发展。遥感图像分类的主要目的是从遥感图像中获取地物信息，从而识别实际地物种类。其实质是将图像中的每个区域或象元点归为若干专题要素中的一种，或若干类别中的一类，并且完成图像数据从二维灰度空间到目标模式空间的转换。

随着人工智能在处理信息方面逐渐显现出来的优势，遥感图像分类技术趋于人工智能化，如人工神经网络、主动学习、支持向量机等。相较于普通人工神经网络，深度神经网络具有更多运算层级，在海量数据上应用统计学习的方法，从计算机视觉的角度提取遥感图像信息，能够极大地提高含有大量未知信息的遥感图像分类的精度。因此，深度神经网络日渐成为遥感图像分类研究中的热点。本文探讨和回顾遥感图像场景分类、目标分类等相关研究的现状，分析和总结了遥感图像传统分类算法研究存在的主要问题，并从解决遥感图像分类与识别的角度，深入分析了典型的图像分类网络的模块结构原理及功能。

方桂安，20354027，(e-mail: fanggan@mail2.sysu.edu.cn)。
刘梦莎，20354091，(e-mail: liumsh6@mail2.sysu.edu.cn)。
刘　玥，20354229，(e-mail: liuy2236@mail2.sysu.edu.cn)。
罗秋琳，20354095，(e-mail: luoqlin3@mail2.sysu.edu.cn)。
马梓场，20354103，(e-mail: mazy23@mail2.sysu.edu.cn)。
唐　迅，20354121，(e-mail: tangx66@mail2.sysu.edu.cn)。

II. 遥感图像分类

A. 原理简介

遥感图像是通过各种传感仪器发射的电磁波对远距离目标辐射后，目标反射的电磁波信息的成像，是由亮度特征构成的光谱空间。每种地物对不同波段的光的敏感程度不同，因此每种地物都有固定的光谱特征。但由于干扰的存在以及环境条件的不同 (如大气辐射、磁场变化、扫描仪视角、拍摄时间等)，光谱信息反映的地物特征不尽相同。遥感图像分类的任务就是通过对各类地物波谱特征进行分析来选择特征参数，反演推测目标地图的几何特征和物理特征，将特征空间按照类别划分成若干不相关的子空间，进而把影像内的像元划分到各子空间，从而实现分类。相对于普通图像，遥感图像有着自身的特点，主要包括:

1) **数据庞大**，地面上的每一个成像点都有对应的光谱信息，且遥感图像是周期采集的，具有连续性；
2) **不确定性**，遥感图像受到天气、光照等外界因素的影响，导致地物反射的光谱信息不尽相同，且遥感图像具有一定的局限性，存在同谱异物或同物异谱的现象；
3) **时效性**，遥感图像的获取周期短，具有很强的时效性；
4) **综合性**，遥感影像的处理需要参照其他地理信息。

B. 传统分类方法概述

遥感图像通过各波段像素值的大小来区分不同的地物。由于化学、物理等因素，不同地物在相同波段光的照射下的反射作用不同，因此得到的光谱信息也不同，这些光谱信息就是遥感图像分类的依据。为了从原始数据中抽出可供判别的统计量，对图像进行特征提取是遥感图像分类的基础，可以定量地提取出光谱特征、纹理特征和空间特征。

遥感图像分类方法分为监督分类和非监督分类。监督分类是指在部分已标注的遥感信息上的学习，建立分

类模型，预测未标记图像的分类。传统的监督分类方法有最大似然法、平行多面体法、线性判别法、马氏距离法和最小距离法。目前研究的较多的监督分类方法有人工神经网络、支持向量机和主动学习等。非监督分类是指在没有已标记信息的情况下，根据特征空间中的数据特点自建群分类，并根据样本的总体特征预测判断，对于特征相同的点聚类，没有结果判别的学习过程。常见的无监督学习有聚类分析、关联规则分析、迭代自组织数据分析技术和 K-均值算法等。

C. 现存问题

遥感图像信息含量大、地物种类多，不同的分类模型的性能也不尽相同。训练分类器是分类研究的关键部分。分类训练是在训练样本集上进行优化的过程，是一个机器学习过程。在传统的监督分类中，分类器从已标记分类信息的样本中进行训练学习，建立模型，对未标记的样本进行分类预测。随着互联网的高速发展，数据信息的共享度不断提高，从网络中获取大量的不含标记信息的样本已变得相对容易，而获取大量含有标记信息的样本仍较为困难。因此，如何通过对少量的含有标记信息的样本和大量的不含有标记信息的样本进行训练来提高机器学习性能，成为了当前机器学习中最受关注的问题之一。

针对遥感图像分类，典型的神经网络模型不能充分挖掘图像中该类别地物特征与周围地物特征之间的关联性，没有将上一地物特征对周边的影响及当前地物本身的特征和分布特性考虑到当前分类中，动态变化性较差。如何充分利用其像素的光谱特性和纹理特征构成的特征向量来构建不同的视图空间，从不同的角度充分挖掘有价值的未标记样本，并补充样本数据的数目，成为了研究的重点。

III. 分类网络

A. 数据处理流程

我们小组通过阅读相关文献，认为遥感图像分类与普通图像分类最大的区别在于不确定性（同谱异物或同物异谱）和综合性（需要参照地理信息）。故如果只是简单地进行图像增广，单源特征往往不能很好地反映出所有地物类别之间的差异，从而导致分类算法的泛化性能较差。

针对此问题，我们认为可以采取以下这种多源多特征融合 [1] 的方法。如图 1所示，对于每一个像素点，考

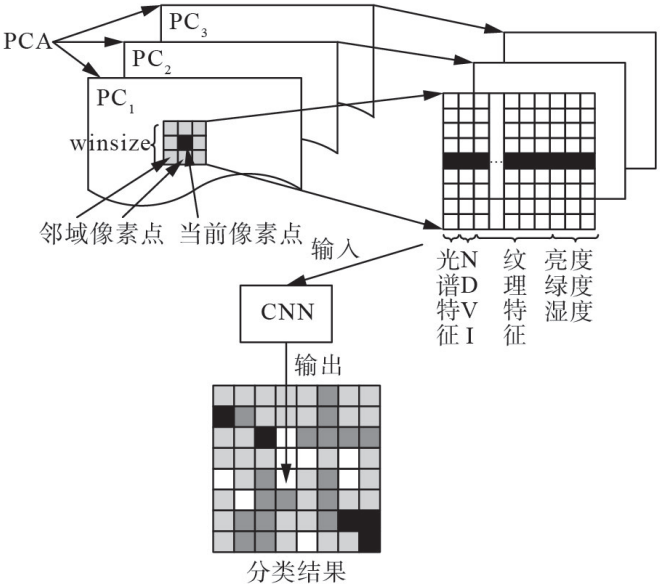

Fig. 1. **数据处理流程图**

虑了其周围大小为 winsize× winsize 的邻域像素，这有助于消除图像的斑点噪声。

首先利用 PCA 对原始数据进行变换，然后选择几乎包含所有波段 95% 以上信息的前 3 个主成分 (PCA_1 PCA_2 PCA_3) 作为变换后的原始图像。

接着提取训练样本对应的光谱值组成一维光谱特征向量 A，同时计算 NDVI（归一化植被指数），组成一维特征向量 B。

其次对每幅图像计算灰度共生矩阵 (gray level co-occurrence matrix，GLCM)，并基于 GLCM 提取均值、方差、熵、角二阶距、相关性、相异性、对比度和协同性共 8 种二阶概率统计的纹理滤波，按照提取顺序将其组成纹理特征矩阵 C 。

最后对图像进行 K-T 变换，提取亮度、绿度和湿度 3 个分量的数据组成特征矩阵 D。

按照图 1所描述的多源多特征融合方法，将 A、B、C、D 按照 [A B C D] 组成一个大小为 9×13×3 的特征融合矩阵，并将此矩阵输入分类网络中进行特征学习，最后进行分类处理。

B. 经典 CNN

CNN 的出现在图像分类领域取得了一系列的突破，并在大规模视觉任务上取得了优异的表现。深度 CNN (DCNN) 的巨大成功归功于其强大的特征学习能力。与传统的图像分类方法不同，基于 CNN 的分类方法是一个端到端的学习过程，只输入原始图像，在网络中进行

训练和预测过程，最后输出结果。这种方法放弃了人工提取特定图像特征的方法，打破了传统分类方法的瓶颈。这也是 CNN 用于图像分类的最大优势。本节主要介绍了基于 CNN 的图像分类模型，并按照时间顺序逐一介绍了具有代表性的经典模型。

1) LeNet: 1998 年，Lecun 等人建立了 LeNet-5 模型，用于对不同的人进行数字分类，并优于当时所有其他方法 [2]。这也是第一次将反向传播算法用于 CNN 的训练中。LeNet-5 模型是深度学习发展的基石，也是往后各种模型的灵感来源。

LeNet-5 网络有 7 层，包含大约 60k 个参数。如图 2所示，该网络分为两部分: 卷积区和全连接区。卷积区的基本单元是卷积层 (Conv) ，然后是最大池化层 (Pool) ，它是由卷积层和最大池化层的基本单元重复堆叠而成的。全连接区包含三个全连接层，每个层都有固定的神经元数量，依次为 120、84 和 10。这个模型使用 sigmoid 激活函数，在输出层使用 softmax 分类器。当卷积区的输出被传入全连接区时，全连接区的输入层将对小批次中的每个特征图进行展平。每个小批次中的向量长度为通道数 x 高度 x 宽度。

虽然 LeNet- 5 在早期的 MNIST 中能取得不错的成绩，但在更大的数据集上的表现并不令人满意。首先，神经网络计算复杂，在当时的硬件水平下，计算效率较低。其次，研究人员在参数初始化、优化算法等诸多领域没有大量深入研究，导致复杂神经网络的训练通常比较困难。在 LeNet-5 提出十多年后，神经网络一度被其他机器学习方法如支持向量机 (SVM) 超越。

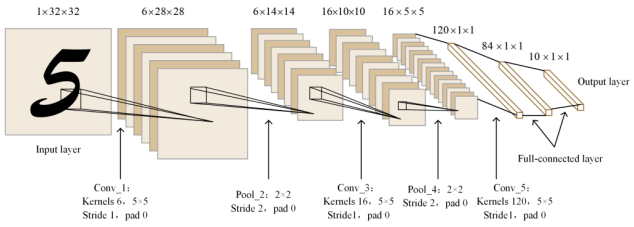

Figure 7. The architecture of the LeNet-5 network. The output shape is channel × height × width. Each convolutional layer uses size 5 × 5, padding 0, strides 1. Each pooling layer size 2 × 2 and strides 2.

Fig. 2. LeNet-5 网络

2) AlexNet: 2012 年，Krizhevsky 等人构建了 AlexNet [3]。这个网络以巨大的优势赢得了 ILSVR 2012 的比赛。它首次证明了学习到的特征可以超越人工设计的特征，从而一举打破了以往计算机视觉研究的状态。由于当时单个 GTX580 GPU 的能力有限，它采用了跨 GPU 的并行计算处理。

AlexNet 网络有 8 层，包含约 60M 的参数。它与 LeNet 的设计理念非常相似，但也有很大的区别。由图 3可以看出，AlexNet 包含 8 层转换，包括 5 层卷积和 2 层全连接隐藏层，以及 1 个全连接输出层。最后的全连接层给模型带来了大量的参数。ImageNet 中大多数图像的高度和宽度都比 MNIST 图像的高度和宽度大 10 倍以上，占据了更多的像素，所以需要在第一层中采用更大的卷积尺寸来提取物体特征。而 AlexNet 的改进之处如下。

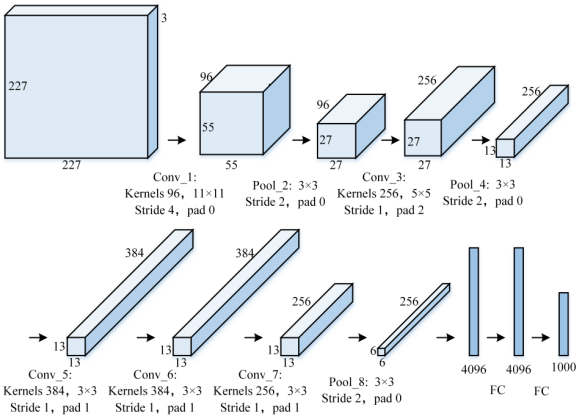

Fig. 3. AlexNet 的架构。第一层的卷积大小为 11 × 11，第二层缩减为 5 × 5，然后全部采用 3 × 3。$Conv_1, Conv_2, Conv_5$ 层之后是一个最大池化层，大小为 3 × 3，步长为 2。最后，有两个完全连接层为 4096，输出层为 1000 类。

(1) ReLU：激活函数从 sigmoid 改为 ReLU，它加速了模型的收敛并减少了梯度的消失。
(2) Dropout：该模型使用 dropout 来控制全连接层的模型复杂性，p=0.5，以缓解过拟合问题。
(3) 数据扩增：引入了大量的数据增强功能，如翻转，裁剪和颜色变化，以进一步扩大数据集，缓解过拟合问题。剔除法和数据扩增法被广泛用于后续的卷积神经网络。
(4) 重叠池化：相邻的池化层之间会有重叠的区域，这可以提高模型的准确性，缓解过度拟合。

3) VGGNet: 2014 年，Simonyan 等人提出了 VGG 模型 [4] 并获得了 ILSVR 2014 的亚军。该模型与 AlexNet 模型类似，也是采用卷积区后是全连接区的结构。VGG 模块的组成规则是连续使用几个相同的卷积层，然后是最大的池化层，卷积层保持输入高度和宽度不变，而池化层将其减半。VGG 网络有多种不同的层结构模型，图 4是 VGG-16。它包含 16 个权重层，网络串联了五个块，最后连接了两个 4096 的全连接层和一个 1000 分类的输出层。

虽然 AlexNet 的作者在卷积大小、输出通道数量、构造顺序等方面做了很多调整，但他们并没有为网络的构造提供常规的思路。VGGNet 给出的设计提供了这一思路，对 AlexNet 的改进如下。

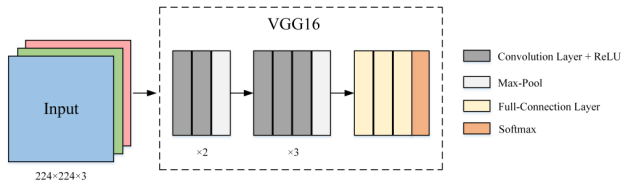

Fig. 4. VGG-16 网络架构。
卷积：size = 3 × 3, stride = 1, padding = 1.
池化：size = 3 × 3, stride = 2.

(1) 模块化网络。VGGNet 使用大量的基本模块来构建模型，这种想法已经成为 DCNN 的构建方法。
(2) 较小的卷积。在 VGGNet 上使用了大量的 3x3 卷积滤波器，与较大的卷积滤波器相比，可以保证网络的深度增加，在相同的感受野下，模型参数减少。
(3) 多尺度训练。它首先将输入图像缩放为不同大小的 $S \in (256, 512)$，然后随机裁剪为固定大小的 224x224，并将得到的多个窗口的数据一起训练。这个过程被看作是一种尺度抖动处理，可以达到数据增量的效果，防止模型过拟合。

4) Network in Network (NIN): 2014 年，Lin 等人提出了一个具有网中网结构的网络 NIN 模型 [5]。与传统卷积层中使用的线性滤波器加非线性激活函数不同，NIN 模型将 MLP 与卷积结合起来，使用更复杂的微神经网络结构代替了传统的卷积层。这种新型的层被称为”Mlpconv”，见图 5。

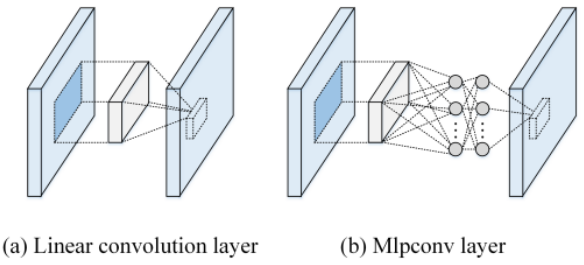

Fig. 5. 线性卷积层与 mlpconv 层的比较

NIN 网络与 LeNet、AlexNet 和 VGG 的区别在于将多个由卷积层和 MLP 组成的 Mlpconv 串联起来，建立一个深度网络。对于深度卷积神经网络来说，卷积层要实现良好的抽象表示，通常需要输入的数据是高度非线性的。滤波器是低级数据的生成线性模型（GLM），GLM 的抽象度很低。高层次的过滤器结合低层次的概念，生成高层次的抽象概念。而 MLP 具有很强的表达非线性函数的能力，用更有效的函数逼近器–MLP 代替 GLM 可以增强局部模型的抽象表达能力。作者认为在将每个局部模块组合成更高层次的概念之前，对其进行更好的抽象处理，有利于网络的构建，这就是 mlpconv 微网络。

NIN 是在 AlexNet 出现后不久提出的，它们的卷积层设置是相似的。但是在 NIN 中，这些不同的设计和贡献总结如下：

(1) Mlpconv：MLP 层相当于一个 1x1 卷积层。现在，它通常被用来调整通道和参数，也可以进行跨通道互动和信息整合。
(2) 全局平均池化（GAP）：全连接层不再用于输出分类，而是使用输出通道数量等于标签类别数量的微网络块，然后通过 GAP 层对每个通道中的所有元素进行平均，以获得分类置信度。

该模型之所以使用 GAP，是因为它比全连接层更有可解释性，更有意义。此外，全连接层由于参数太多，容易造成过拟合，而且它过于依赖 dropout 正则化。GAP 可以看作是一种结构化的正则化方法，用全连接层代替它可以大大减少模型的参数量，有效防止模型过拟合。

C. Inception 四部曲

2014 年，由 Christia Szegedy 等人提出的 GoogLeNet [6] 赢得了 2014 年 ILSVR 的冠军。该模型吸收了 NIN 的思想和 Arora 等人的理论工作，并引入了 Inception 模块的概念。在接下来的几年里，研究人员对 Inception 模块进行了多次改进，该模型的性能也得到了提高。

1) Inception V1: GoogLeNet 有 22 层，包括约 6M 的参数。该网络的基本模块是 Inception 模块（图 6a）。这个模块包含 4 个并行的分支。前三个分支使用不同大小的卷积层来提取不同空间大小下的信息。其中，1×1 的卷积可以减少通道的数量，压缩信息，从而降低模型的复杂度。最后一个分支的最大池化效应是降低分辨率，然后进行 1×1 的卷积来调整池化后的深度。总而言之，这个独特的设计提高了网络模型的宽度和对不同尺度甚至分辨率的适应性，实现了多尺度融合的效果。

GoogLeNet 模型类似于 VGGNet，其卷积部分也使用了模块化拼接。

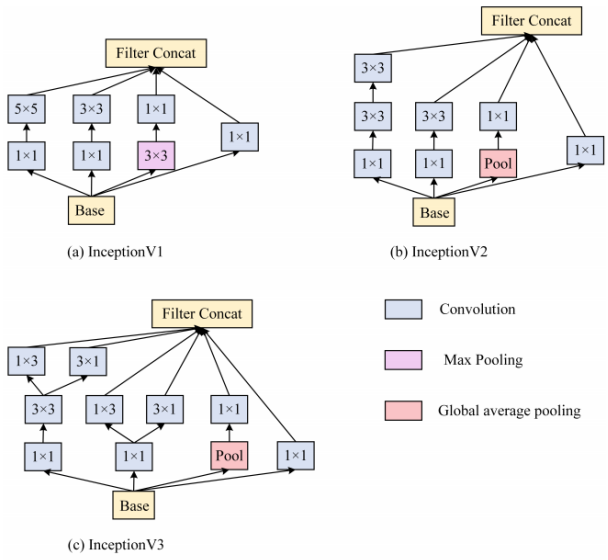

Fig. 6. Inception V1 到 V3 模块

提高网络性能最直接的方法是增加网络深度和网络宽度（每一层神经元的数量），但其缺点是随着网络规模、参数数量的增加，网络更容易过拟合，计算资源的使用将显著增加。作者认为，解决上述缺点的基本方法是将全连接层，甚至是卷积层转换为稀疏连接。首先，生物神经系统之间的联系也很稀少。其次，Arora 等人的主要研究结果表明，如果数据集的概率分布可以用一个大的、非常稀疏的深度神经网络来表示，那么可以通过分析最后一层激活情况的相关统计量，并将输出高度相关的神经元进行聚类，从而逐层构建最优网络拓扑。因此，Inception V1 所取得的进展如下：

- Inception 模块：虽然早期的传统神经网络使用随机稀疏连接，但计算机硬件在计算非均匀稀疏连接时效率低下。所提出的 Inception 模型不仅可以保持网络结构的稀疏性，而且还可以利用密集矩阵的高计算性能，从而有效提高模型的参数利用率。
- GAP：更换了全连接的图层，以减少参数。
- 辅助分类器：用于深层网络的辅助分类器是在训练过程中插入层与层之间的小 CNN，所产生的损失加到主网络损失中。

因此，GoogLeNet 的参数只有 AlexNet 的 1/12，但性能大大提高。

2) Inception V2: 与 V1 相比，V2 [7] 的改进如下：

(1) 更小的卷积。5×5 卷积被两个 3×3 卷积所取代。这也减少了计算时间，从而提高了计算速度，因为 5×5 卷积比 3×3 卷积成本要高 2.78 倍。

(2) 批量正则化 (BN)。BN 通过对每个 mini-batch 的输入进行归一化，使神经网络更快、更稳定。

在 CNN 中，BN 是通过一个归一化步骤来实现的，该步骤固定了每层输入的平均值和方差。理想情况下，归一化将在整个训练集上进行，但是为了将这一步与随机优化方法联合起来而去使用全局信息是不切实际的。因此，归一化在训练过程中被限制在每个小批量上。对于具有 d 维输入 $x = x^{(1)} \cdots x^{(d)}$ 的层，它将对每个维度的输入进行归一化，并使用 $B = x(i...m)$ 来表示整个训练集的大小为 m 的 mini-batch。BN 变换因此可以表示为：

- mini-batch 均值：$\mu_B = \frac{1}{m}\sum_{i=1}^{m} x_i$
- mini-batch 方差：$\sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_B)^2$
- 归一化：$\hat{x}_i = (x_i + \mu_B)/\sqrt{\sigma_B^2 + \epsilon}$
- 缩放和平移：$y_i = \gamma\hat{x}_i + \beta = BN_{\gamma,\beta}(x_i)$

γ, β 是优化过程中要学习的参数，ϵ 是一个为了数值稳定性而加到 mini-batch 方差的常数。

3) Inception V3: Inception V3 [8] 主要侧重于通过修改以前的 Inception 架构来减少计算能力的消耗。其主要改进如下：

(1) 因子化的卷积。这有助于提高计算效率，因为它减少了网络中所涉及的参数的数量。它还保持检查网络效率。这部分包含以下（2）和（3）。

(2) 较小的卷积。用更小的卷积代替更大的卷积，显然会导致更快的训练。

(3) 不对称卷积。3×3 卷积可以被 1×3 卷积和 3×1 卷积所取代。参数的数量减少了 33%。

(4) 网格大小减少。网格大小通常通过池化操作来实现。然而，为了解决计算成本的瓶颈问题，作者提出了一种更有效的技术。例如，在图 7中，320 个特征图是由步长为 2 的卷积完成的。通过最大池化得到了 320 个特征图。这两组特征图被连接成 640 个特征图，并进入 Inception 模块的下一个层次。

4) Inception V4: Inception V4 [9] 的主要目标是降低 Inception V3 模型的复杂性，该模型对每个 Inception 块做出统一选择。Inception 块包括 Inception 模块和 Reduction 模块，如图 8所示。图 9为 inception V4 的总体架构。图中所有没有标记为 “V” 的卷积都是相同的，这意味着它们的输出网格与输入的大小相匹配。标

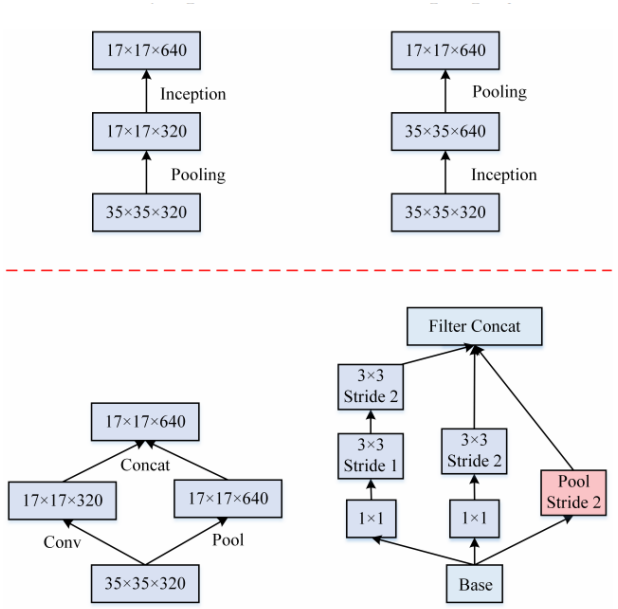

Fig. 7. 红线上有两种方法：左边的方法违反了 Inception V3 原则。右边的版本在计算成本上要高三倍。红线下的方法：一个高效的网格尺寸缩小模型既高效，又避免了 Inception V3 原则所建议的代表性瓶颈。

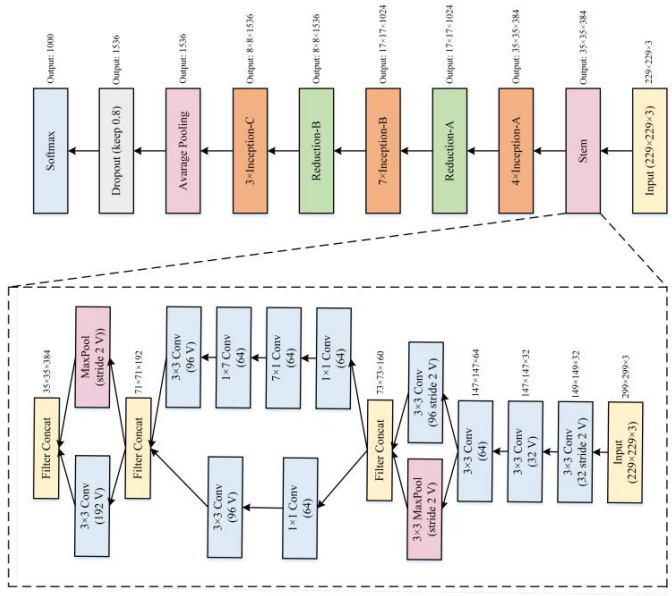

Fig. 9. Inception V4 的整体架构。图片右边是整体结构，图片左边是结构的 Stem 部分。

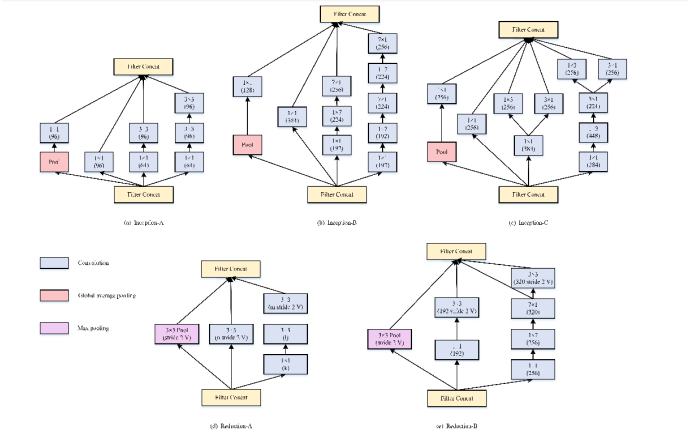

Fig. 8. Inception V4 块。它包括 Inception 模块和 Reduction 模块。

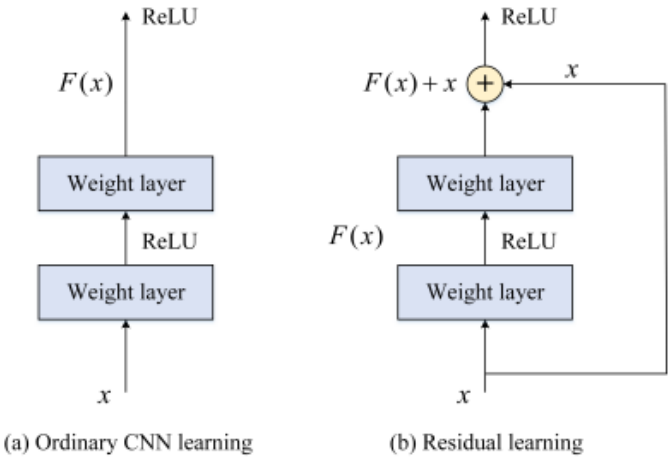

Fig. 10. 普通 CNN 学习和残差学习的比较

记有“V”的卷积是有效填充的，这意味着每个单元输入的 patch 完全包含在前一层中，输出激活映射的网格大小相应地减小。

D. 残差学习网络结构

1) **残差网络**: 2015 年，KaimingHe 等人提出的深度残差网络 ResNet [10] 获得了 ILSVR2015 的一等奖。回顾前面介绍的网络发展，网络深度的增加是一个共同的发展趋势，即增加网络的深度会提高网络的性能。然而，许多实验表明，在一定的深度范围内，单纯增加网络深度并不能有效地提高网络性能。另一个实验表明，20 层以内的网络层数增加带来了网络性能的提高，但如果超过 20 层的深度网络继续叠加网络层数，分类精度反而会下降。

对于这种现象，我们可能会盲目地把矛头指向梯度消失/爆炸的特征或过度拟合的问题。然而，有几十层的网络可以通过初始归一化和批量正则化在随机梯度下降的反向传播过程中轻松收敛。文章验证了网络退化不是由过拟合引起的。事实上，这是因为深度网络不能轻松地优化到预期性能——随机梯度下降 (SGD) 使优化变得困难。这种准确率不增反减的现象被称为”退化”。它已经严重影响了深度非线性网络的训练。训练残差连接在 ResNet 中是一种打破“退化”的方法，使深度神

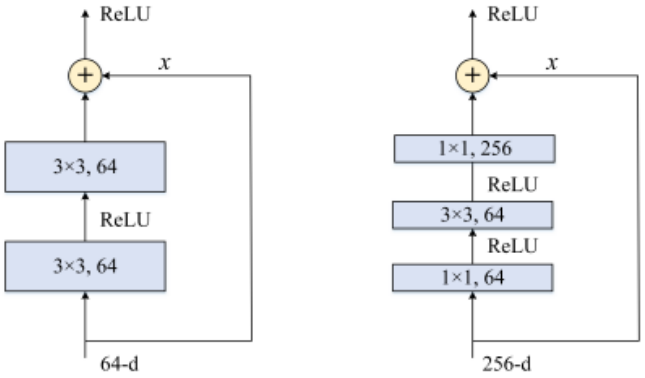

Fig. 11. 残差网络的两个模块

经网络达到高精确度。基于 VLAD 和 Fisher Vector 的残差矢量编码代表，以及捷径连接理论和实践的研究，残差网络使堆叠层的网络以最优状态继续积累多个直接映射层。这种残差连接在促进优化的同时可以增加网络的深度，精度也在不断提高。

残差网络的残差连接层，如图 10所示。$H(x)$ 是我们想要的理想映射，图 10的左边部分是普通的 CNN 学习，它需要直接拟合映射 $H(x)$。右边的残差学习是让残差块不直接学习目标映射，而是拟合一个与直接映射 $F(x) = H(x) - x$ 相关的残差映射。假设一定深度的网络趋于饱和，为了保证参数更新和梯度下一层的传播，只需要将 $F(x)$ 的权重和偏置更新为 0，然后直接映射 $H(x) - x$ 可以保证下一层的输入至少与前一层的输出相同。事实上，当理想映射 $H(x)$ 与直接映射非常接近时，残差映射也很容易捕捉到直接映射的细微波动。当然，非线性映射 $F(x)$ 比直接拟合更容易学习，这使得输入值能更快地通过跨层" 数据线" 向前传播。残差块包含两个具有相同通道数的 3×3 卷积层，每个卷积层后面都有一个 BN 和一个 ReLU 激活函数。另一个分支跳过卷积层将输入直接连接到最后一个 ReLU——如图 11(左) 中 ResNet-34 的构建块。当网络堆栈很深时，可以在 3×3 卷积层之后添加 1×1 卷积层，以控制通道的数量——如图 11 (右)，即 ResNet-50/101/152 的"瓶颈" 构建模块。

ResNet 可以说是站在了真正意义上的 DCNN 的最前沿。ResNet 的重要贡献如下：

1) 这种方法很容易优化，但当深度增加时，"普通" 网络 (简单堆叠层) 显示出更高的训练误差。
2) 它可以轻易地从大大增加的深度中获得准确性，产生比以前的网络更好的结果。

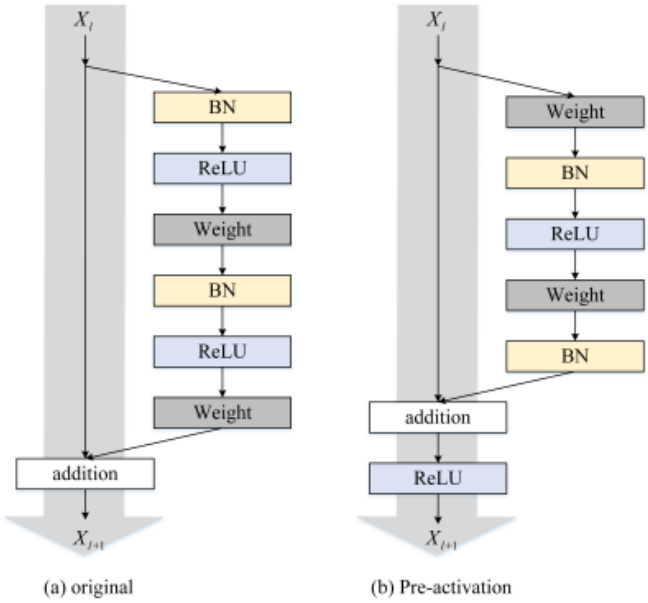

Fig. 12. (a) 原始的残余单元。(b) 具有完全预激活功能的残差单元

2) 残差网络的改进: 拥有数百层甚至更多的 CNN 确实非常有竞争力，但很多 DCNN 也存在训练难度大的挑战和过拟合的风险。在数据集有限的情况下，研究人员对 ResNet 的构建模块进行了如下改进。

1) **带有预激活功能的 ResNet**。He 等人提出了一种预激活结构，对 BN 和 ReLU 进行预激活以进一步提高网络性能。对 BN 和 ReLU 的布局进行了多次实验，得到了图 12（右）中性能最好的结构。它可以成功地训练具有 1000 层以上的 ResNet。同时，他们也证明了与其他捷径连接相比直接映射的重要性。
2) **随机的深度**。[23] 的作者指出，残差网络中存在许多层网络对输出结果的贡献很小。在网络训练过程中，采用了随机深度法，删除一些层可以大大缩短训练时间，有效提高 ResNet 的深度，甚至超过 1200 层。对 CIFAR-10/100 数据集的测试误差和训练时间仍有较好的改善。
3) **广义的残余网络 (WRN)**。随着残余网络的深度不断增加，越来越少的特征重用将使网络的训练非常缓慢。为了缓解这个问题，[24] 引入了一个宽落差块，拓宽了原始残差单元的权重层图 10(右)，在两个权重层之间增加了落差。与更深的 ResNet 相比，层数更少的 WRN 大大减少了训练时间，在 CIFAR&ImageNet 数据集上有更好的表现。
4) **ResNeXt**。虽然 Inception 和 ResNet 有很好的

性能，但这些模型只适用于一些数据集。由于涉及许多超参数和计算，使它们适应新的数据集不是一件小事。一个新的维度 “Cardinality C”(区块中的路径数) 被用来克服这个问题，实验证明，当我们增加容量时，增加 Cardinality C 比更深入或更广泛的网络效果更好。作者比较了图 13中三种数学计算的完全等价结构。实验结果表明，带有分组卷积的图 13c 块比其他两种形式更简洁、更快速,ResNeXt 使用这种结构作为一个基本块。

5) **扩张残差网络 (DRN)**。以解决下采样引起的特征图分辨率下降和特征信息丢失的问题。然而，简单地去除网络中的子采样步骤会降低感受野。因此，Yu 等人引入了扩张卷积，用来增加高层的感受野，并替换了残差网络内部下采样层的一个子集，补偿了因去除子采样而引起的感受野的减少。与参数量相同的 ResNet 相比，DRN 在图像分类中的准确度明显提高。

6) **其他模型**。Veit 等人将训练好的 ResNet 的一些层丢掉，并将其作为一个新的模型。仍有可观的性能。ResNet 中的 ResNet (RiR) 提出了一个深度双流架构，它包含了 ResNet 和标准 CNN，并且容易实现，没有计算开销。DropBlock 技术丢弃了被称为块的连续相关区域的特征，这是一种正则化，有助于避免数据科学专业人员面临的最常见问题，即过度拟合。Big Transfer (BiT) 提出了一种应用于 ResNet 的通用转移学习方法，该方法使用了最少的技巧，但在许多任务上获得了出色的性能。NFNet 提出了一种没有 BN 层的基于 ResNet 的结构，通过使用自适应梯度剪裁技术来实现惊人的训练速度和准确性。

3) 带有 Inception 的 ResNet: 从 2014 年到 2017 年，残差法和 Inception 法在图像分类任务中具有很强的统治力。研究人员将这两种结构结合起来，并增加了新的方法以达到更好的性能，这些经典的变体讨论如下。

1) **Inception-ResNet**。试图将 Inception 结构与 residual 结构相结合，并取得了良好的性能。它与 inceptionV4 来自同一篇论文，其组合是 Inception-ResNet-v1/v2，如图 14和 15所示。Inception-ResNet-V1 的计算成本与 Inception-V3 大致相同，它的训练速度更快，但达到的最终精度比 InceptionV3 略差。Inception-ResNet-V2 的计算成本与 Inception-V4 大致相同，它的训练速度比 InceptionV4 快得多，最终的准确率也比 InceptionV4 略高。

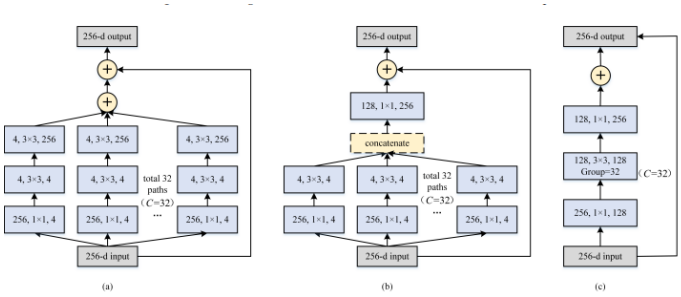

Fig. 13. (a-c)ResNeXt 的等效构建模块。

2) **Xception**。它是基于 InceptionV3 的设计思路。作者认为在 InceptionV3 中使用修改过的深度可分离卷积来代替卷积操作，通道之间的相关性和空间相关性应该被分开处理。有研究表明，使用可分离卷积可以减少 CNN 的大小和计算成本。但是 Xception 的修改旨在提高性能。Xception 在 ImageNet 上的准确度比 Inception-v3 略高，而参数则略微降低。[14] 的实验也表明，在 Xception 中加入类似于 ResNet 的残差连接机制可以大大加快训练时间，获得更高的准确率。

3) **PolyNet**。许多研究倾向于在图像分类中增加深度和宽度任务，以获得更高的性能。但是非常深的网络会有麻烦，那就是收益递减和训练难度增加。计算成本和内存需求的二次增长都是由不断扩大的网络引起的。这种方法探索了 Inception 和 ResNet 的结构多样性，这是在深度和宽度之外的一个新维度，从多项式的角度引入了一个更好的混合模型。

4) DenseNet: DCNN 经常面临梯度消失和退化的困境，且网络训练也是一个难题。上述改进的残差网络提出了一些解决方案，除此之外我们还必须提到优秀作品 DenseNet。它与 ResNet 和 Highway 网络有着相同的方向，传统卷积的连接是在每层和下一层之间，在 DenseNet 中，每一层都以前馈的方式连接到其他每一层。这样一来，每一层都可以直接接触到损失函数的梯度和原始输入信号，从而实现隐性的深度监督。密集块如图 16所示，每一层都从所有前面的层获得额外的输入，并将自己的特征图传递给所有后续层。可以表示为 $x = Hi([x_0, x_1, \cdots\cdots, X_{l-1}])$，残差连接为 $x_l = H_l(x_{l-1}) + x_{l1}$，它还设置增长率 k 表示通过一个层时增加的输入通道数量。[15]DenseNet-B 把

“$BN-ReLU-Conv(1\times1)-BN-ReLU-Conv(1\times1)$”称为”瓶颈层”，对 DenseNet 非常有效。而为了要求相同大小的特征图，DenseNet-C 使用“过渡层”——1×1 卷积将特征图的数量减少 $\theta\in(0,1)$，设置在不同的密集块之间以实现向下采样。瓶颈层和过渡层合称为 DenseNet-BC，它的性能是最好的。

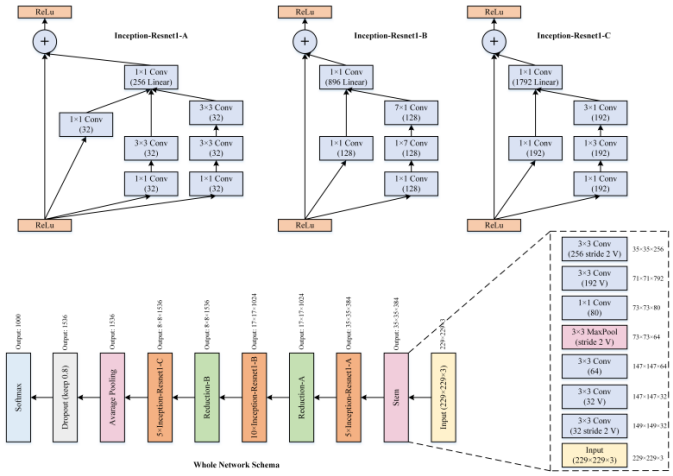

Fig. 14. Inception-ResNet-V1

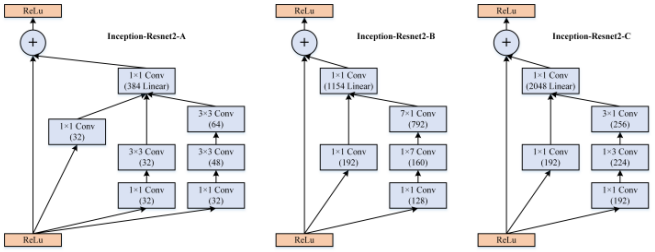

Fig. 15. Inception-ResNet-V2. 与 Inception-ResNet-V1 相比，某些层的参数数量有所增加

值得一提的是，Yan 等人提出的双路径网络 (DPN) 探索了残差学习和密集连接之间的关系，并结合了它们的优点。也可以说，通过 DPN，它们之间的数学表达是统一的。CondenseNet 主要通过分组卷积运算和训练时的剪枝来优化 DenseNet，以达到更高的计算效率和更少的参数。

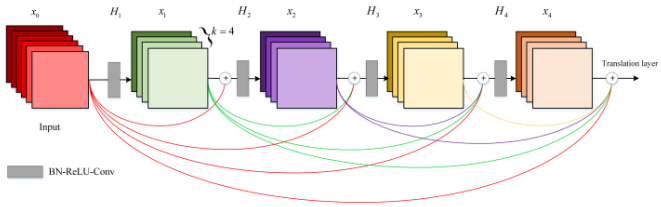

Fig. 16. 密集区块。5 层，增加率为 k=4

E. 注意力机制

注意力机制可以用人类的视觉机制来直观地解释。比如我们的视觉系统往往会关注图像中的部分信息进行辅助判断，而忽略不相关的信息。有一种模型基于这个想法，通过使用通道注意力或空间注意力来提高 CNN 模型的性能。它们可以看作是整个模型的小规模改进，可以移植到任何可行的模型中。

1) 残差注意力网络: 为了使网络能够学习对象的感知特征，Wang 等人 [16] 提出了残差注意力网络（RAN 或 Attention）将注意力机制整合到 CNN 中。RAN 的主要结构是由残差块堆叠而成，RAN 的整体架构如图 17所示。其中，三个超参数 p ，t ，r：p 是拆分为主干分支和掩码分支之前的预处理残差单元的数量；t 表示主干分支中的残差单元数；r 表示掩码分支中相邻池化层之间的残差单元数。

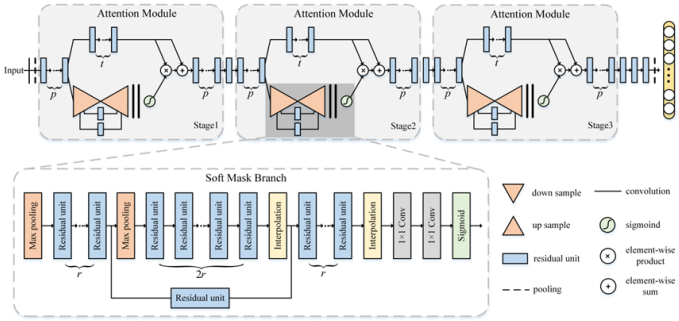

Fig. 17. Residual Attention Network 架构

Residual Attention Network 有两个分支：Trunk 分支是注意力模块中用于特征提取的上层分支，可以是 Pre-Activation ResNet 块或其他块，它的输入和输出分别是 x 和 $T(x)$。Mask 分支使用自下而上的 top-down 结构来学习相同大小的掩码 $M(x)$。注意力模块 H 的输出是：

$$H_{i,c}(x) = M_{i,c}(x) * T_{i,c}(x) \tag{1}$$

其中 i 跨越空间位置，c 是从 1 到 C 的通道索引。注意掩码可以在前向推理期间用作特征选择器，也可以在反向传播期间用作梯度更新滤波器。在 soft mask 分支中，输入特征的 mask 梯度为：

$$(\partial M(x,\theta)T(x,\theta))/\partial\phi = M(x,\theta)\partial T(x,\phi)/\partial\phi \tag{2}$$

其中 θ 是掩码分支参数，ϕ 是主干分支参数。然而，单纯堆叠注意力模块会导致性能下降。这是因为使用 mask 重复从 0 到 1 的点生成会降低深层特征的价值，而 soft

mask 可能会破坏主干分支的良好特性。一个更好的掩码构造为:

$$H_{i,c}(x) = (1 + M_{i,c}(x)) * F_{i,c}(x) \quad (3)$$

这被称为注意力残差学习。软掩码分支使用自下而上自上而下的全卷积结构。激活函数使用 softmax，它是每个通道和空间位置的简单 sigmoid。RAN 的积极作用如下:

1) 堆叠多注意力模块使 RAN 在识别嘈杂、复杂和杂乱的图像方面非常有效。
2) RAN 的分层组织使其能够根据每个特征图在层内的重要性自适应地为其分配权重。
3) 结合三个不同级别的注意力（空间、通道和混合）使模型能够使用这种能力来捕获这些不同级别的对象感知特征。

2) SENet: 2017 年 Wang 等人 [17] 提出的 Squeeze-and-Excitation Networks (SENet) 为 CNNs 引入了一个构建块，它在几乎没有计算成本的情况下改善了通道相互依赖性，它也是 ILSVR2017 在图像分类任务上的冠军。使用通道注意机制的 SE 模块可以添加到任何 baseline 架构中，以提高性能，而计算开销可以忽略不计。该架构的示意图如图 18所示。

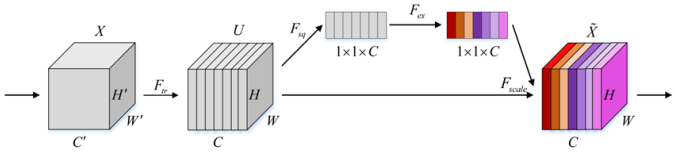

Fig. 18. Squeeze-and-Excitation (SE) 模块

通常，在创建输出特征图时，网络会平等地加权每个通道。SE 块就是通过添加注意力机制来自适应地对每个通道进行加权来改变这一点。首先，输入量 $X \in \mathbb{R}^{H' \times W' \times C'}$ 通过 F_{tr} 被映射到 $U \in \mathbb{R}^{H \times W \times C}$; 接着，通过 Squeeze 模块 F_{sq} 到 $1 \times 1 \times C$ 的向量来获得了对每个渠道的全局认知，接着可以通过 Excitation 模块 $F_{ex}: 1 \times 1 \times C \to 1 \times 1 \times C/r \to 1 \times 1 \times C$ 完全捕获通道的依赖关系。最后，这些 C 值可以用作原始特征图上的权重，并根据其重要性对每个通道的权值进行对应地调整。

3) BAM 和 CBAM: 我们通过 SENet 和 RAN 了解了特征图利用和注意机制的重要性。Wu 等人提出了引入通道注意模块和空间注意模块的瓶颈注意模块（BAM [18]）和卷积块注意模块（CBAM [19]）。下面将介绍这两种结构简单，思想相似的模型。

BAM: 如图 19所示，BAM 通过两条独立的路径 (通道注意力和空间注意力) 获取注意力图。对于给定的输入特征图 $F \in \mathbb{R}^{H \times W \times C}$，BAM 推断出 3D 注意力图 $M \in \mathbb{R}^{H \times W \times C}$。另外，计算得到的细化特征映射 F' 为 $F' = F + F * M(F)$。为了设计一个高效的模块，首先计算出的是通道关注度 $M_c(F) \in \mathbb{R}^{1 \times 1 \times C}$ 和空间注意力地图 $M_s(F) \in \mathbb{R}^{H \times W \times 1}$ 最后，将注意力图 $M(F)$ 与原始特征图 F 相乘得到 $F * M(F)$。

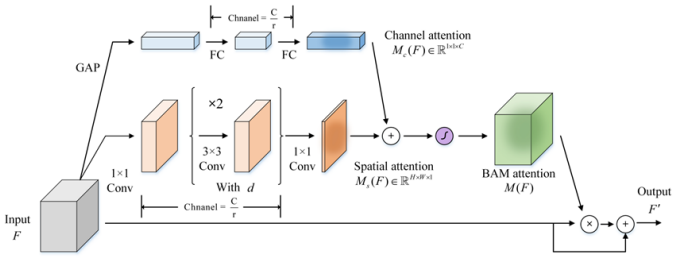

Fig. 19. BAM 模块架构

CBAM: 如图 20所示，CBAM 也通过两条独立的路径 (通道注意力和空间注意力) 获取注意力图。对于给定的输入特征图 $F \in \mathbb{R}^{H \times W \times C}$，CBAM 依次推断出一维通道注意力图 $M_c(F) \in \mathbb{R}^{1 \times 1 \times C}$ 和二维空间注意力地图 $M_s(F) \in \mathbb{R}^{H \times W \times 1}$。在乘法过程中，注意值被相应地广播——通道注意值沿着空间维度广播。通道细化的特征图 F' 和通道空间细化的特征图 F'' 总结为: $F' = M_c(F) * F$ 和 $F'' = M_S(F') * F'$

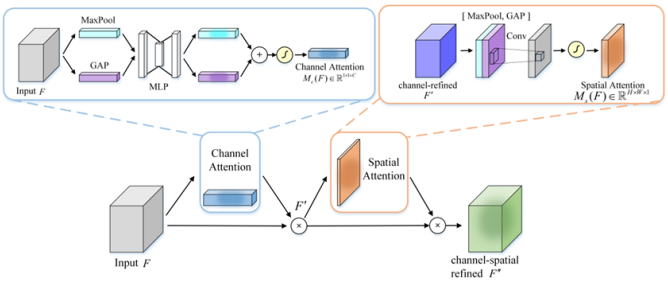

Fig. 20. CBAM 模块架构

简单来说，BAM 使用并联，CBAM 使用串联的通道注意力和空间注意力部分。在实际性能上，CBAM 对图像分类的效果略胜一筹。

4) GENet: 2018 年，Hu 等人 [20] 给 CNN 引入了特征上下文，其中包括两个操作: Gather 操作，ξG 和 Excite 操作，ξE。如图 21所示，对于给定的输入特征

图 $X \in \mathbb{R}^{H\times W\times C}$，聚集操作符 ξG：$X \in \mathbb{R}^{H\times W\times C} \rightarrow \hat{X} \in \mathbb{R}^{H'\times W'\times C}$ $(H' = [H/e], W' = [W/e])$，其中 e 是范围比。而激发操作符 ξE 使用最近邻插值进行调整：$\hat{X} \rightarrow \mathbb{R}^{H\times W\times C}$，然后在 sigmoid 之后，将 $\hat{X}$ 和原始输入 X 执行逐元素乘积运算:$\xi E(X, \hat{X}) = X * f(\hat{X})$。可以将此方法添加到 baseline 架构中，以通过稍微增加参数来获得更好的性能。

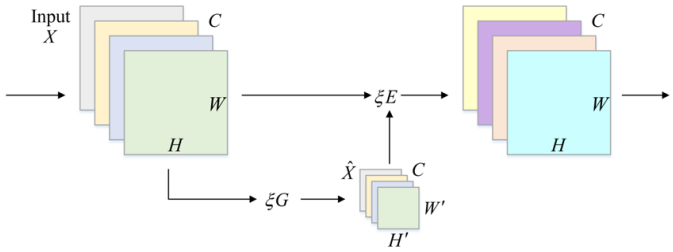

Fig. 21. Gather-Excite 模块架构

5) SKNet: 2018 年，Wang 等人 [21] 提出了选择性核网络 (SKNet)，它可以根据输入特征的多尺度自适应地调整感受野的大小。该网络主要分为三个操作：*Split*、*Fuse* 和 *Select*，如图 22所示。

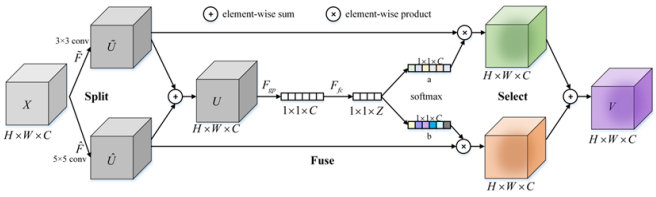

Fig. 22. 选择性内核卷积块

Split: 对于给定的输入特征图 $X \in \mathbb{R}^{H'\times W'\times C'}$，默认情况下执行两次转换：

$$\widetilde{F} : X \rightarrow \widetilde{U} \in \mathbb{R}^{H\times W\times C} \tag{4}$$

$$\hat{F} : X \rightarrow \hat{U} \in \mathbb{R}^{H\times W\times C} \tag{5}$$

$\widetilde{F}$ 和 $\hat{F}$ 由高效的分组/深度卷积、BN 和 ReLU 函数依次组成。

Fuse: 这两个分支通过一个基于元素的求和进行融合：$U = \tilde{U} + \hat{U}$ ，然后，使用 GAP 生成信道相关的信息:

$$F_{gp} : \mathbb{R}^{H\times W\times C} \rightarrow \mathbb{R}^{1\times 1\times C} \tag{6}$$

并且通过使用两个 FC 层来改变通道数:

$$F_{fc} : 1 \times 1 \times C \rightarrow 1 \times 1 \times Z \tag{7}$$

最后，将输出矩阵 a 和 b 输出。

Select: 矩阵 a 和 b，$\widetilde{U}$ 和 $\hat{U}$，并且最终的输出为

$$V = a * \widetilde{U} + b * \hat{U}, (a + b = 1) \tag{8}$$

其中 * 表示元素级积。

6) GSoP 网络: 2019 年，Gao 等人提出了全局二阶池化卷积网络（GSoP-Net），将二维 GAP 引入 CNN 的中间部分，它以协方差的形式体现了通道之间的关系。相关文献 [22]–[27] 也证明了将 GSoP 插入网络可以显着提高性能，并且一种 A^2-Net [28] 使用 GSoP 从输入信息的整个空间聚合和传播全局特征。该方法设计了一个简单有效的 GSoP 块，具有模块化程度高、内存占用少、计算复杂度低等优点。此外，它可以沿通道维度或位置维度捕获全局二阶统计信息，还可以方便地插入到现有的网络架构中，以更少的开销进一步提高其性能。

7) ECA-Net: 2019 年，Hu 等人提出了高效渠道注意力 (ECA) 模块，以解决 SE 块进行渠道降维的不合理性。文献表明，避免降维对于学习通道注意很重要，适当的跨通道交互可以在保持性能的同时显著降低模型复杂度。ECA-Net 利用一维卷积提出了一种不降维的局部跨通道交互策略（见图 23）。

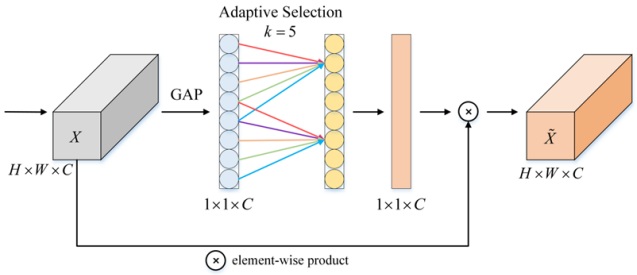

Fig. 23. 高效渠道注意力 (ECA) 模块

8) 协调注意力: 2021 年，Hou 等人 [29] 提出了高效移动网络的坐标注意 (CA) 模块，该模块使用两个一维 GAP 在两个空间方向 (位置信息) 捕获特征码，CA 块如图 24所示。对于给定的输入特征映射 $X \in \mathbb{R}^{H\times W\times C}$，首先，用 X GAP 和 Y GAP 分别对每个信道沿横坐标方向和纵坐标方向进行编码。其次，通过 Concat+Conv2d 生成坐标 attention，再通过 BN + 激活函数将其分解为两个具有空间方向特征的特征映射。最后，对原始输入 X 重新加权，r 为折减比。CA 块不仅捕获跨通道的信息，而且捕获方向敏感和位置敏感的信息，这有助于模型更准确地定位和识别感兴趣的对象。它是为移动网

络设计的轻量级模块，可以灵活地插入到经典的移动网络中，几乎没有计算开销，从而提高性能。

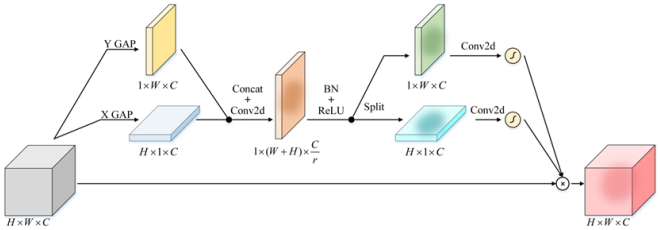

Fig. 24. 高效渠道注意力 (ECA) 模块

9) 其他注意模块和总结: SENet 是使用注意力机制进行图像分类的最成功的例子之一。不仅有上述 BAM、CBAM、GENet 、SKNet 、ECA-Net 和 CA [30] ，还有 GALA [31] 、AA-Net [32] 和 TA [33] 通过采用不同的空间注意机制或设计高级注意块来发展其想法。此外，GSoP-Net ，A 2Net、NLNet [34]、GCNet [35]、SCNet [36] 和 CCNet [37] 作为非局部/自注意力网络的典型示例，由于它们能够构建空间或通道，最近非常受欢迎。通过插入大型网络或移动网络，这些注意力网络在各种计算机视觉任务中非常有用。值得注意的是，NLNet 和 CCNet 不用于图像分类。

F. 更小或更有效的网络

现在有一种趋势，即建立更深层次、更复杂的网络，以实现更高的精度。然而，在许多现实世界的应用程序中，如手机、机器人技术和自动驾驶汽车，这些提高精度的进步并不一定会使网络在规模和速度方面更高效。在这里，我们将介绍更小、更有效的模型，这些模型致力于在一个计算量有限的平台上及时地执行识别任务。

1) SqueezeNet: 2016 年，F.N.Iandola 等人 [38] 提出了一种小型化的网络模型结构，即 SqueezeNet。该网络引入了 fire 模块（图 25），通过用 1×1 滤波器替换 3×3 滤波器，并减少通往 3×3 滤波器的输入通道的数量，从而减少了网络参数。它在网络中后期放置了向下采样，使卷积层有大的特征图。同时，该模型与 Deep Compression 相结合，以扩大模型的体积。与 AlexNet 相比，SqueezeNet 将参数的数量减少了近 50 倍，同时确保是没有精度损失的，模型体积被压缩到原来的 510 倍左右。之后，SqueezeNext 考虑了基于 SqueezeNet 的硬件。

2) MobileNet V1 到 V3: 2017 年，谷歌提出了 MobileNet V1 [39]，这是一个主要用于移动或嵌入式设

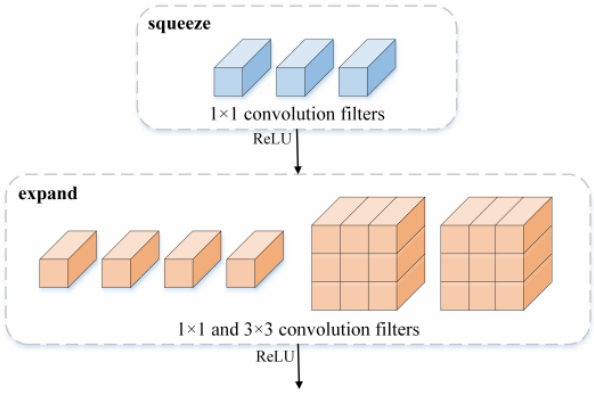

Fig. 25. SqueezeNet 架构中的 fire 模块

备的轻量级网络。该网络使用由深度卷积和点卷积 (1×1conv) 组成的深度可分卷积来代替标准卷积，如图 26所示。它可以大大减少计算成本和参数。同时，MobileNet V1 还提供了两个超参数（宽度乘法器 α 和分辨率乘法器 α）来有效地平衡计算和精度。

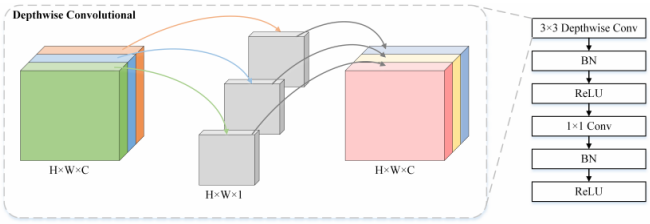

Fig. 26. 深度可分离的卷积

2018 年，MobileNet V2 [40] 引入了倒置残差和线性瓶颈来解决深度可分离卷积中大部分卷积核参数为零的问题。其原因是：特征信息在从高维空间映射到低维空间后很容易被 ReLU 破坏。MobileNet V2 的瓶颈残差块如图 32 所示。与标准残差块 1×1（压缩）→3×3→1×1（扩展）不同，倒置残余块为 1×1（压缩）→3×3→1×1（压缩）。这个块用线性变换替换了连接在 1×1 卷积后面的最后一个 ReLU，以避免信息丢失。

2019 年，基于之前的工作，MobileNet V3 [41] 采用了 SE 块和神经结构搜索 (NAS) 技术，实现了更高的效率和准确性。SE 块用于构建 channel-wise attention，采用块级搜索的平台感知 NAS 方法查找全局网络结构，然后采用层级搜索的 NetAdapt 方法对单个层进行顺序微调。该模型还使用 h-swish 激活函数，修改 swish 函数的 sigmoid 以提高准确度。

3) ShuffleNet V1 到 V2: 2017 年，由 Face++ 公司设计的 ShuffleNetV1 [42]，主要针对常见的移动平台（如

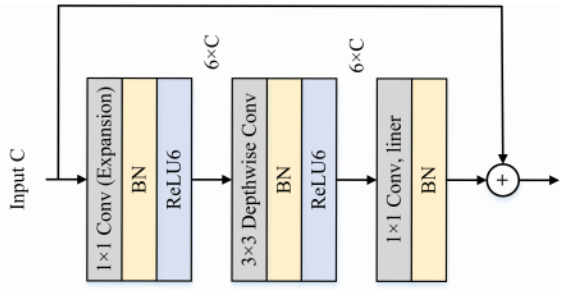

Fig. 27. MobileNet V2 的瓶颈残差块。C 为通道数，扩展比为 6。

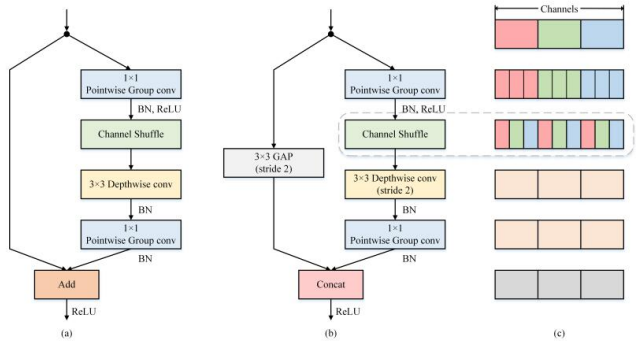

Fig. 28. (a) 步长 = 1 的 ShuffleNetV1 单元。(b) 步长 = 2 的 ShuffleNetV1 单元（下采样，2）。(c) 通道混洗操作

无人机、机器人和智能手机），采用了 Pointwise Group 卷积和 Channel shuffle 来改进残差块。前者是为了解决性能开销昂贵的点卷积导致通道数量有限的问题，以满足可能会严重损害准确度的复杂性约束。后者是为了解决组卷积阻塞通道组之间的信息流动并削弱表示的问题。ShuffleNet 单元（步长为 1）用逐点群卷积替换第一个 1x1 卷积，然后进行通道混洗操作（见图 28c），如图 28a 所示。ShuffleNet 单元（步长为 2）做了两个修改，即在“捷径”上增加了一个 3x3 GAP，并将通道连接替换为元素添加，如图 28b 所示。mobileNetV1 的多个版本已经试验了不同数量的卷积组数（g 组）和不同比例的滤波器数量（s）。

2018 年，ShuffleNetV2 [43] 对速度和精度有更高的要求。它不仅考虑了计算复杂性（每秒浮点运算，FLOPs），还考虑了其他因素，如内存访问成本（MAC）和平台特性。根据实验，作者给出了四条准则：

1）相等的通道宽度使 MAC 最小化；
2）过多的组卷积增加了 MAC；
3）网络碎片化降低了并行程度，这种碎片化的结构可能会降低效率，因为它对具有强大并行计算的设备不友好，如 GPU；

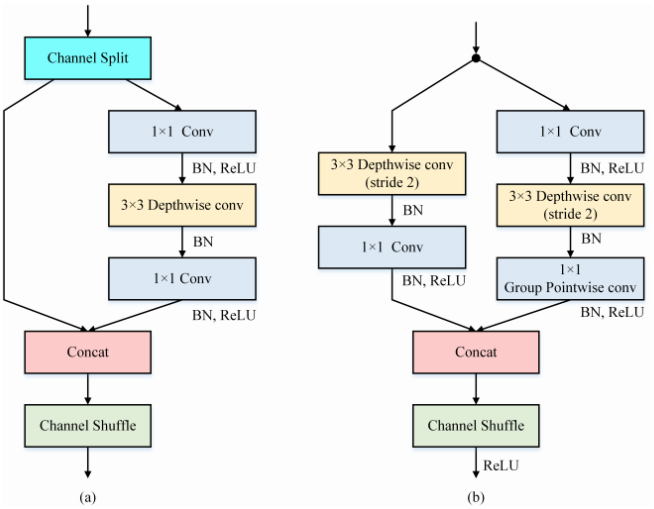

Fig. 29. (a)ShuffleNet V2 单元。(b) 步长 = 2 的 ShuffleNet V1 单元（下采样，2）

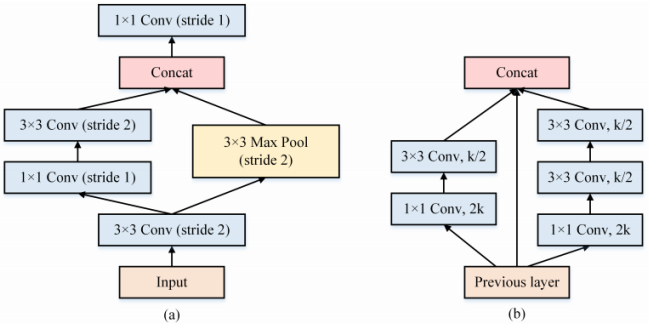

Fig. 30. (a) Stem 块的结构。(b) 双向致密层的结构。

4）元素方面的操作是不可忽略的，因为它也有很高的 MAC/FLOPs 比率。

ShuffleNetV2 单元避免了违反上述准则的情况，如图 29所示。

4) PeleeNet: 2018 年，PeleeNet [44] 在 DenseNet 的基础上做了改进，这是一个用于嵌入式平台的高效架构。它使用一个 steam 块（见图 30a）对输入图像进行首次下采样。原来的密集层被分成两路，以获得不同尺度的感受野，称为 2 路密集层（见图 30b）。它使输出通道的数量与过渡层的输入通道数量保持一致，因为 DenseNet 提出的压缩系数影响了特征表达。另一个改进是，瓶颈层的通道数根据输入形状而变化，而不是固定的 4 倍增长率，与 DenseNet 相比，可以节省 28.5% 的计算成本。最后，它使用后激活（Conv →BN→→ReLU）而不是前激活来提高实际速度。

5) MnasNet: 2018 年，Tan 等人的 MnasNet [45] 是一个自动化的移动神经架构搜索网络，用于使用强化学

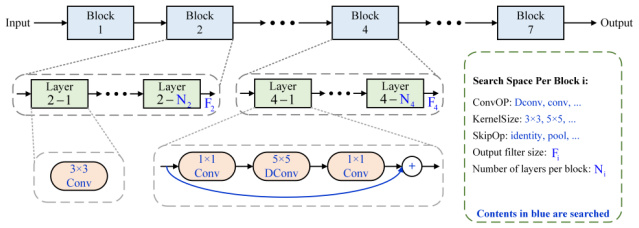

Fig. 31. 因素化分层搜索空间。卷积运算（ConvOP）：常规卷积（conv），深度卷积（dconv），以及具有不同扩展率的移动倒瓶颈卷积。卷积核大小（KernelSize）。跳过操作（SkipOp）：最大或平均池，身份残留跳过，或无跳过路径。

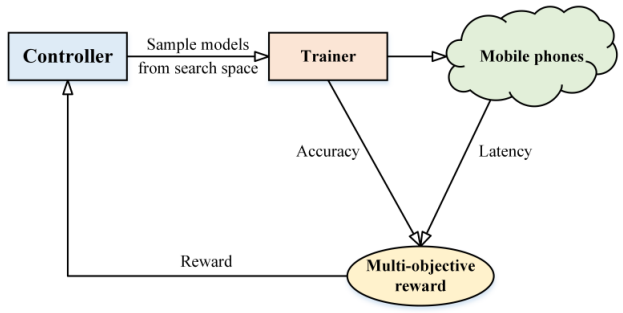

Fig. 32. 移动平台感知的 NAS 概述。

习（RL）来构建移动模型。它融合了 CNN 的基本精髓，从而在提高准确率和减少延迟之间取得了适当的平衡，在模型部署到移动设备上时描绘出高性能。MnasNet 明确地将速度信息纳入搜索算法的主要奖励功能，并通过在特定平台上执行模型来直接测量模型的速度。这个架构，一般来说，包括两个阶段如下。

分解层次搜索空间。该搜索空间支持整个网络中包含的不同层结构。CNN 模型被分解成不同的块，其中每个块都有一个独特的层结构。连接的选择要使输入和输出相互兼容，从而产生良好的结果以保持较高的准确率。图 31显示了 MnasNet 中搜索空间的示意图。

强化搜索算法。它采用了强化学习的方法，奖励最大化（多目标奖励），以实现两个主要目标（延迟和准确性）。在搜索空间中定义的每个 CNN 模型将被映射到一个由 RL 代理执行的行动序列中。控制器是一个 Recur-rent 神经网络（RNN），训练器训练模型并输出准确性。该模型被部署到移动电话上以估计延迟。准确度和延迟都被合并为一个多目标的奖励。该奖励被发送到 RNN，RNN 的参数被更新，以最大化总奖励。图 37 显示了移动平台感知 NAS 的概况。

6) EfficientNet V1 至 V2: 模型缩放的常规做法是任意增加 CNN 的深度或宽度，或者使用更大的输入图像分辨率进行训练和评估。相关文献 [46] 显示，网络深度和宽度之间存在一定的关系。虽然这些方法确实提高了准确性，但通常需要繁琐的人工调整，而且仍然经常产生次优的性能。2019 年，Tan 等人提出了 EfficientNetV1 [47]，它使用一个简单而高效的复合系数 Φ 来扩大 CNN，使网络的结构化程度更高。与上述任意缩放网络尺寸（如宽度 w、深度 d 和分辨率 r）的方法不同。这种模型缩放方法使用 Φ 来均匀地缩放网络 w、d 和 r，遵循：

$$\begin{aligned} \text{depth} &: d = \alpha^\phi \\ \text{width} &: w = \beta^\phi \\ \text{resolution} &: r = \gamma^\phi \\ \text{s.t } &\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2 \\ &\alpha \geq 1, \beta \geq 1, \gamma \geq 1 \end{aligned}$$

其中，α, β, γ 是可以通过小网格搜索确定的常数。ϕ 是一个用户指定的系数，控制模型扩展的资源多寡。最后，它通过利用多目标 NAS 开发一个 baseline 网络。主要的构建模块是移动倒置瓶颈 MBConv（见图 26和 31），SE 模块也会被添加。

2021 年，EfficientNetV2 [48] 提出了一个更小的模型和一个更快的训练方法。首先，EfficientNetV2 应用 FixRes 解决了 EfficientNetV1 的大图像尺寸导致大量内存占用的问题，使用较小的图像尺寸进行训练，但训练后不对任何层进行微调。相关资料也指出，使用较小的图像尺寸进行训练，准确度会稍好。其次，对于深度卷积往往不能充分利用设备加速计算的问题，EfficientNetV2 逐渐用 Fused-MBConv 取代了 MBConv，以更好地利用移动或服务器加速，即用单一的 3×3 卷积取代 MBConv 中的深度 3×3 卷积和扩展 1×1 卷积。然后用 NAS 自动搜索 MBConv 和 Fused-MBConv 的最佳组合。最后，这个网络提出了一个训练感知的 NAS 来搜索最佳组合。另一个重要的观点是 EfficientNetV2 使用了修正的渐进式学习，用不同的图像尺寸进行训练也会相应地改变正则化强度，以解决在训练过程中动态改变图像尺寸而导致的精度下降问题。

7) 其他前沿技术: 对于移动终端和嵌入式设备来说，手工设计的 CNN 模型已不再是时代的潮流。目前，更多的轻量级模型是与其他强大的算法相结合。不难看出，上面的一些介绍也包括混合的 CNN 架构设计。下面我们将简要介绍几种可与纯 CNN 结合的流行方法。

1) 在训练好的模型上，奇异值分解（SVD）通过压缩网络中全连接层的权重矩阵，可以达到模型压缩的效果；低秩滤波器使用两个 1× K 卷积而不是

一个 K × K 卷积来去除冗余，减少权重参数；网络修剪方法是舍弃网络中权重较低的连接，以降低网络的复杂性。量化通过牺牲算法的精度来减少每个权重所需的空间；神经网络的二进制化可以看作是一种极端的量化，它使用二进制表示网络权重，大大减少了模型的大小；深度压缩使用剪枝、量化和 Huffman 编码三个步骤来压缩原始模型，在不损失精度的情况下取得了惊人的压缩率。这种方法具有里程碑式的意义，引领了 CNN 模型小型化和加速研究的新热潮。

2) NAS 搜索，轻量级网络通常需要更小、更快、精度尽可能高。有太多的因素需要考虑，这对设计一个高效的模型是一个巨大的挑战。为了使架构设计过程自动化，首次引入了 RL 来搜索具有更高精度的高效架构。一个完全可配置的搜索空间可能会呈指数级增长，难以解决。因此，早期的架构搜索工作集中在元胞级结构上，并且同一单元在所有层中都被重复使用。相关文献探索了一个块级的层次结构搜索空间，允许在网络的不同分辨率块上有不同的层结构。为了减少搜索的计算成本，作者采用了基于梯度的优化的可微分架构搜索框架。着重于现有网络适应受限的移动平台，也有学者提出了更有效的自动网络简化算法。

3) 知识蒸馏（KD），KD 指的是模型压缩的概念。通过一步一步地教一个较小的网络，准确地使用一个已经训练好的较大的网络来做什么。这种训练设置有时被称为"teacher-student"，其中大网络是 teacher，小模型是 student。最后，student 网络可以达到与 teacher 网络相似的性能。

G. 其他高效方法

1) ViT: 基于自我注意机制的 transformer 在自然语言处理 (NLP) 中取得了巨大的成功。由于其强大的表示能力，研究人员正在研究如何将其应用于计算机视觉任务。以 ViT 为代表的纯 transformer 架构在图像分类任务中表现良好。最近，更多的 ViT 模型，例如 DeiT、PVT、TNT 和 Swin，已经被提出，以追求更强的性能。也有很多作品试图用卷积操作来增强纯 transformer 块或自注意层，例如 BoTNet、CeiT、CoAtNet、CvT。一些工作（如 DETR 方法）尝试将类似 CNN 的架构与 ViT 结合起来进行物体检测。当然，ViT 在更多的视觉任务、高/中/低级视觉和视频处理方面取得了成就，但是，我们在此不做过多介绍。现在，将 ViT 模型或混合方法应用于计算机视觉任务仍然存在巨大的挑战。在这个方向的遥感图像分类任务上，更多研究正在进行。

2) 自监督学习: 自监督已被应用于改善遥感图像分类中的 SOTA 模型。这种方法有三个主要步骤。

1) 在有标签的图像上训练一个 teacher 模型。
2) 使用 teacher 在未标记的图像上生成伪标签。
3) 结合有标签的图像和伪标签的图像来训练 student 模型。SSL 提出了一种基于 teacher-student 的半监督深度学习方法，以提高大型 CNN 的性为了使 student 有更强的学习能力。带噪声的 student 使 student 大于或至少等于 teacher，它为 student 添加噪声，如 RandAugment 数据增强、dropout 和随机深度。伪标签中的确认偏误问题会使 student 在伪标签不准确的情况下从不准确的数据中学习。元伪标签是通过 student 对标签数据集的表现反馈来不断调整的，这样 student 就可以从 teacher 那里学到更好的伪标签。

3) 迁移学习: DCNN 通常需要大量任务相关的数据和计算才能获得良好的性能。将这些 SOAT 网络应用于新任务可能会非常产生巨大的计算成本。迁移学习提供了一种解决方案，即网络在大型通用数据集上完成训练，然后使用其权值初始化后续任务。许多预先训练的模型，如 GoogLeNet 和 ResNet，已经在 ImageNet 等大型数据集上进行了训练，并用于遥感图像分类。该方法已被用作图像分类任务的常规解决方案。值得一提的是，BiT 提供了在许多任务中实现出色性能的通用方法。

4) 数据增强: CNN 经常面临由于有限的数据而导致的过拟合风险。传统的数据增强技术包括一系列方法，如翻转、颜色空间、裁剪、旋转、平移和噪声注入，这些方法可以改善训练数据集的属性和大小。不仅如此，它还具有显著提高 DL 模型泛化能力的潜力。自动数据增强有潜力解决传统数据增强方法的一些缺陷，通过使用学习到的数据增强策略训练 CNN 模型，可以显著提高半监督学习的准确性和模型鲁棒性。混合样本数据增强 (MSDA) 是将两个训练样本和它们的标签按照一定的比例随机混合，不仅可以减少一些困难样本的误识别，还可以提高模型的鲁棒性，使其在训练过程中更加稳定。

5) 其他优化方法: 实现更好的性能不仅仅是设计一个优秀的架构，基于训练的优化方法也很重要。除了上述

方法外，还有几种可靠的选择:

1) 优化器。优化器可以有效地最小化损失函数，以达到预期的性能，如 SGD，Adam，PMSProp 。锐度感知最小化 (SAM)，作为目前最好的解决方案，缓解了损失函数最小化和模型泛化能力之间的关系。
2) 正则化。BN 是大多数图像分类模型的一个关键组成部分，在训练集和测试集上都能达到更高的精度。更多的变体也扩展了这一思想，如层归一化和组归一化。但是最近的研究表明 BN 的一些重要缺陷会影响 CNN 的长期发展。NFNet 使用核心技术——自适应梯度剪切 (AGC)，在不进行归一化的情况下训练深度模型。

IV. 总结

A. 研究现状

综上所述，图像分类的研究蓬勃发展，但科研人员对于如何将前沿成果应用于遥感图像分类的研究仍然处于初步阶段。

对于复杂的遥感图像分类，传统的神经网络训练方法已经不能满足需求，使用深度神经网络为提高遥感图像分类精度开辟了一片新天地。通过将分类依据表达成向量，将少量含有标记信息的图像作为训练集，利用深度神经网络模型大量训练构建图像的数字化字典，对大量未标记的遥感图像数据分类。

对于包含海量信息的遥感图像，如何充分挖掘信息以更贴合遥感图像分类要求成为了研究的重点。遥感图像涵盖的纹理特征、光谱特征和空间特征都可以单独作为图像分类的依据，然而图像的这些特征信息在图像分类时尚未充分利用。对提取的这 3 类特征进行多角度充分利用，并将其共同作为分类依据，结合深度神经网络的训练模型，来提高遥感图像的分类精度，成为了该领域目前的研究热点。

B. 未来趋势

目前，遥感图像分类识别主要依赖于人工识别，遥感图像的数据主要来源于卫星影像，由于遥感图像的数据过于庞大，容易产生信息冗余，并且图像分辨率较低，容易导致不同信息相融合现象。由于遥感图像的本质特征，导致对遥感图像的分类较为困难，因此，如何准确分离出具有价值的信息，是遥感图像分类的重要任务。利用深度学习技术理论上是可以实现对遥感图像信息的分类，但是，由于技术上的不足，距离实际应用依然存在一定的差距。针对遥感图像的特殊性建立合适的深度学习模型，使用较好的优化算法，以实现识别率接近甚至超过人工识别，这是遥感图像分类识别的发展趋势。

参考文献

[1] 李亚飞, 董红斌. 基于卷积神经网络的遥感图像分类研究 [J]. 智能系统学报, 2018, 13(4):7.

[2] Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-Based Learning Applied to Document Recognition. Proc. IEEE 1998, 86,2278–2324.

[3] Krizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet Classification with Deep Convolutional Neural Networks. Neural Inf. Process.Syst. 2012, 25, 1097–1105.

[4] Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv 2014, arXiv:1409.1556.

[5] Lin, M.; Chen, Q.; Yan, S. Network In Network. arXiv 2013, arXiv:1312.4400.

[6] Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.;Rabinovich, A. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA,USA, 15 October 2015; pp. 1–9.

[7] Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv 2015, arXiv:1502.03167

[8] Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 12 December 2016; pp. 2818–2826.

[9] Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA; 2016.

[10] He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.

[11] Zagoruyko, S.; Komodakis, N. Wide Residual Networks. 2016, pp. 87.1–87.12. Available online: https://doi.org/10.5244/C.30.87 (accessed on 1 June 2021).

[12] Xie, S.; Girshick, R.; Dollar, P.; Tu, Z.; He, K. Aggregated Residual Transformations for Deep Neural Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE Computer Society: Los Alamitos, CA, USA, 2017; pp. 5987–5995.

[13] Yu, F.; Koltun, V.; Funkhouser, T. Dilated Residual Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 636–644.

[14] Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807.

[15] Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.

[16] Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; Tang, X. Residual Attention Network for Image Classification.arXiv2017, arXiv:1704.06904.

[17] Hu, J.; Shen, L.; Sun, G.; Albanie, S. Squeeze-and-Excitation Networks. InIEEE Transactions on Pattern Analysis and Machine Intelligence; IEEE: Piscataway, NJ, USA, 2019.

[18] Park, J.; Woo, S.; Lee, J.Y.; Kweon, I.S. BAM: Bottleneck Attention Module.arXiv2018, arXiv:1807.06514.

[19] Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module.arXiv2018, arXiv:1807.06521.

[20] Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Vedaldi, A.Gather-Excite: Exploiting Feature Context in Convolutional Neural Networks; NIPS' 18; Curran Associates Inc.: Red Hook, NY, USA, 2018; pp. 9423–9433.

[21] Li, X.; Wang, W.; Hu, X.; Yang, J. Selective Kernel Networks. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 510–519.

[22] Gao, Z.; Xie, J.; Wang, Q.; Li, P. Global Second-order Pooling Convolutional Networks. arXiv 2018, arXiv:1811.12006.

[23] Ionescu, C.; Vantzos, O.; Sminchisescu, C. Matrix Backpropagation for Deep Networks with Structured Layers. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 2965–2973.

[24] Lin, T.Y.; RoyChowdhury, A.; Maji, S. Bilinear CNNs for Fine-grained Visual Recognition.arXiv2015, arXiv:1504.07889.

[25] Cui, Y.; Zhou, F.; Wang, J.; Liu, X.; Lin, Y.; Belongie, S. Kernel Pooling for Convolutional Neural Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 3049–3058.

[26] Li, P.; Xie, J.; Wang, Q.; Gao, Z. Towards Faster Training of Global Covariance Pooling Networks by Iterative Matrix Square Root Normalization. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 947–955.

[27] Li, P.; Xie, J.; Wang, Q.; Zuo, W. Is Second-Order Information Helpful for Large-Scale Visual Recognition? In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2089–2097.

[28] Chen, Y.; Kalantidis, Y.; Li, J.; Yan, S.; Feng, J.A2-Nets: Double Attention Networks. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Red Hook, NY, USA, December 2018; NIPS' 18; Curran Associates Inc.: Red Hook, NY, USA, 2018; pp. 350–359.

[29] Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks.arXiv2019, arXiv:1910.03151.

[30] Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design.arXiv2021, arXiv:2103.02907.

[31] Linsley, D.; Shiebler, D.; Eberhardt, S.; Serre, T. Learning what and where to attend.arXiv2018, arXiv:1805.08819.

[32] Bello, I.; Zoph, B.; Le, Q.; Vaswani, A.; Shlens, J. Attention Augmented Convolutional Networks. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 3285–3294.

[33] Misra, D.; Nalamada, T.; Uppili Arasanipalai, A.; Hou, Q. Rotate to Attend: Convolutional Triplet Attention Module.arXiv2020, arXiv:2010.03045.

[34] Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local Neural Networks.arXiv2017, arXiv:1711.07971.

[35] Cao, Y.; Xu, J.; Lin, S.; Wei, F.; Hu, H. GCNet: Non-local Networks Meet Squeeze-Excitation Networks and Beyond.arXiv2019, arXiv:1904.11492.

[36] Liu, J.J.; Hou, Q.; Cheng, M.M.; Wang, C.; Feng, J. Improving Convolutional Networks with Self-Calibrated Convolutions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10093–10102.

[37] Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. CCNet: Criss-Cross Attention for Semantic Segmentation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 27 October–2 November 2019, Seoul, Korea; pp. 603–612.

[38] Iandola, F.N.; Han, S.; Moskewicz, M.W.; Ashraf, K.; Dally, W.J.; Keutzer, K. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size. arXiv 2016, arXiv:1602.07360.

[39] Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. arXiv 2017, arXiv:1704.04861.

[40] Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520

[41] Howard, A.; Pang, R.; Adam, H.; Le, Q.; Sandler, M.; Chen, B.; Wang, W.; Chen, L.C.; Tan, M.; Chu, G.; et al. Searching for MobileNetV3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 1314–1324.

[42] Zhang, X.; Zhou, X.; Lin, M.; Sun, J. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. arXiv 2017, arXiv:1707.01083.

[43] Ma, N.; Zhang, X.; Zheng, H.T.; Sun, J. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. In Computer Vision—ECCV 2018; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 122–138.

[44] Wang, R.J.; Li, X.; Ling, C.X. Pelee: A Real-Time Object Detection System on Mobile Devices. arXiv 2018, arXiv:1804.06882.

[45] Tan, M.; Chen, B.; Pang, R.; Vasudevan, V.; Sandler, M.; Howard, A.; Le, Q.V. MnasNet: Platform-Aware Neural Architecture Search for Mobile. arXiv 2018, arXiv:1807.11626.

[46] Raghu, M.; Poole, B.; Kleinberg, J.; Ganguli, S.; Sohl-Dickstein, J. On the Expressive Power of Deep Neural Networks. arXiv 2016, arXiv:1606.05336.

[47] Tan, M.; Le, Q.V. EffificientNet: Rethinking Model Scaling for Convolutional Neural Networks. arXiv 2019, arXiv:1905.11946

[48] Tan, M.; Le, Q.V. EfficientNetV2: Smaller Models and Faster Training. arXiv 2021, arXiv:2104.00298.