《人工智能导论》探究报告
图像分类

姓名：齐明杰　学号:2113997　AlexNet
姓名：高路博　学号:2111172　VGGNet
姓名：杨浩甫　学号:2113824　ResNet

目录

作业正文

1 问题描述

1.1 定义

图像分类问题，即根据各自在图像信息中所反映的不同特征，把不同类别的目标区分开来的图像处理方法。它利用计算机对图像进行定量分析，把图像或图像中的每个像元或区域划归为若干个类别中的某一种，以代替人的视觉判读。

1.2 分类

根据分类任务的目标不同，可以将该问题分为单标签图像分类和多标签图像分类两个子问题。对于单标签的图像分类问题，它可以分为跨物种语义级别的图像分类，子类细粒度图像分类，以及实例级图像分类三大类别。

1.3 评价指标

单标签图像分类问题：准确率 (Accuracy), 精确率（Precision）, 召回率（Recall）,F1-score, 混淆矩阵，ROC 曲线和 AUC 等；

多标签图像分类问题：平均准确率 (AP) 和平均准确率均值 (mAP)，汉明距离，1-错误率，覆盖率，排序损失等。

1.4 常用的图像分类数据集

MNIST 数据集，PASCAL，ImageNet 数据集等。

1.5 发展历史

图像分类任务从传统的方法到基于深度学习的方法，经历了几十年的发展。上个世纪 90 年代末本世纪初，SVM and K-nearest neighbors 方法被使用的比较多，以 SVM 为代表的方法，可以将 MNIST 分类错误率降低到了 0.56%，彼时仍然超过以神经网络为代表的方法，

本世纪的早期，虽然神经网络开始有复苏的迹象，但是受限于数据集的规模和硬件的发展，神经网络的训练和优化仍然是非常困难的。2009 年，ImageNet 数据集发布，发布早年，仍然是以 SVM 和 Boost 为代表的分类方法占据优势，直到

2012 年 AlexNet 的出现。AlexNet 是第一个真正意义上的深度网络，大幅提升了当时图像分类的准确度，在 2012 年 ImageNet 竞赛上以 15.3% 的 top-5 测试错误率夺得冠军，错误率比第二名低了 10.9

2014 年的冠亚军网络分别是 GoogLeNet 和 VGGNet，进一步加深网络深度与层数，自此，深度学习模型的分类准确率已经达到了人类的水平 (5% 10%)。2015 年，ResNet 获得了分类任务冠军。它以 3.57% 的错误率表现超过了人类的识别水平，并以 152 层的网络架构创造了新的模型记录。2017 年，也是 ILSVRC 图像分类比赛的最后一年，SeNet 获得了冠军。这个结构，仅仅使用了“特征重标定”的策略来对特征进行处理，通过学习获取每个特征通道的重要程度，根据重要性去降低或者提升相应的特征通道的权重。

2 论文摘要

第一篇 AlexNet 论文**首次将卷积神经网络 CNN 和深度学习 DL 相结合**，搭建出现代意义上第一个深度卷积神经网络模型。该模型创新性地使用 **ReLU 作为激活函数代替了传统的 Sigmoid 和 Tanh**；在多个 GPU 上进行模型的训练；使用 LRN 对局部的特征进行归一化；使用 dropout 技术减轻过拟合；采用重叠最大池化避免了平均池化的平均效应。该模型大幅提升了当时图像分类的准确度，其网络规模及输入规模远大于当时其他模型，刺激了沉寂多年的深度学习领域。

第二篇论文研究了 **CNN 的深度对图像识别的影响**，并搭建了一种具有较深层网络结构的 **vgg 网络**，从网络的深度入手，考察在参数总数基本不变的情况下，CNN 随着层数的增加，其效果的变化，探索了卷积神经网络深度对视觉识别精度的影响。作者为了使得增加网络深度可行，**采用了固定的卷积核大小 3x3**，使得模型深度变深，非线性次数变大，参数更少。但是该模型**训练时间过长，调参难度大**，而且需要的存储容量大，不利于部署。

Krizhevsky 与 Simonyan 的论文已经证明，神经网络的深度越深，其模拟效果越好，但一个严重的问题制约了神经网络的进一步发展。更大深度的网络的训练错误率，反而会比低层次网络的还要大，这被称为“退化”现象。**针对此问题，第三篇论文作者何恺明提出了残差网络的新模型**。残差模型的引入，**完美解决了“退化”问题**，并为构建更大深度的神经网络解除了桎梏。何恺明利用残差模型构建了 100 余层的神经网络，它的模拟效果达到了神经网络前所未有的新高度，**此外他还进一步探究 1000 层的神经网络，千层网络表现出良好的效果**。

3 核心内容

3.1 AlexNet: ImageNet Classification with Deep Convolutional Neural Networks

3.1.1 论文动机

这篇论文的动机是要解决物体识别任务中的挑战。物体识别是计算机视觉领域中的一个重要问题，它涉及将图像中的物体分类为不同的类别。为了解决这个问题，需要一个具有大量学习能力和先验知识的模型。卷积神经网络（CNNs）是一种具有这些特性的模型，它们可以通过改变深度和广度来控制其容量，并且对图像的本质做出了强有力且大多数正确的假设。

然而，在过去，CNNs 在处理高分辨率图像时仍然非常昂贵。幸运的是，随着 GPU 计算能力和 2D 卷积实现技术的提高，现在可以训练足够大且有趣的 CNNs 模型，并且最近出现了一些数据集（如 ImageNet），其中包含足够多标记示例以训练这些模型而不会出现过拟合问题。

因此，本文旨在训练一个最大化 CNNs 性能并减少训练时间的模型，并将其应用于 ImageNet 数据集上进行物体识别任务。通过使用更深、更广、更复杂的 CNNs 模型以及更多数据进行训练，我们希望能够提高物体识别的准确性，并且在未来可以将这些模型应用于更复杂的任务，如视频物体识别。

3.1.2 实验方法

这篇论文使用了卷积神经网络（CNNs）来解决物体识别问题。具体来说，作者使用了一个非常大的 CNNs 模型，并在 ImageNet 数据集上进行了训练和测试。该模型包含多个卷积层和池化层，以及全连接层和 softmax 分类器。

为了训练这个模型，作者使用了反向传播算法和随机梯度下降优化器。为了加速训练过程，作者还实现了高度优化的 GPU 实现，并使用多个 GPU 并行训练模型。此外，他们还采用了一种称为“dropout”的正则化方法来减少过拟合。

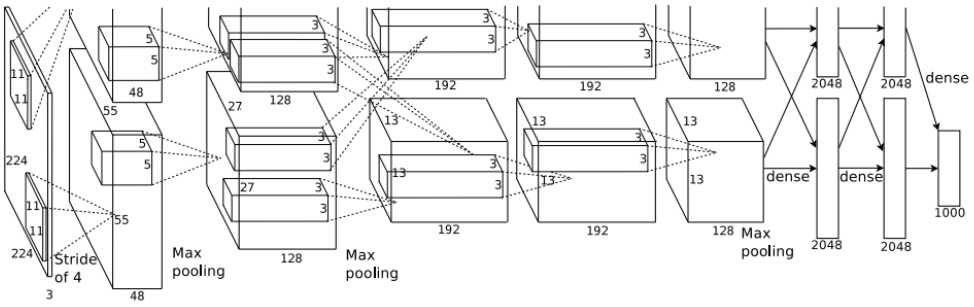

图 1: CNN 神经网络概念图

在训练过程中，该论文使用了以下方法或技巧：

- 非饱和神经元：研究人员使用了一种非饱和的激活函数（ReLU）来代替传统的 sigmoid 函数。这种激活函数比 sigmoid 函数更快，并且在训练过程中可以更好地避免梯度消失问题。
- GPU 实现卷积操作：为了加速训练，研究人员使用了高效的 GPU 实现卷积操作。
- 局部响应归一化（LRN）：研究人员使用了 LRN 来增强特征图之间的竞争关系。这种方法可以使得响应较大的神经元抑制周围神经元的响应，从而增强特征图之间的差异性。
- Dropout 正则化：为了减少过拟合，研究人员采用了一种称为 “dropout” 的正则化方法。在每次迭代时，随机删除一些神经元以减少过拟合。
- 随机梯度下降（SGD）算法：在训练过程中，研究人员使用了 SGD 算法，并将学习率设置得很小以避免震荡。
- 数据增强：为了增加数据量和减少过拟合，研究人员对训练数据进行了一些随机变换，如随机裁剪、水平翻转等。

在训练完成后，作者对该模型进行了广泛的测试，并与其他先进的物体识别算法进行比较。结果表明，该模型在 ImageNet 数据集上取得了迄今为止最好的结果，并且在多个子任务中都超过了其他算法。

总之，本文提出并实现了一个非常大、复杂且高效的 CNNs 模型，并将其应用于 ImageNet 数据集上进行物体识别任务。通过使用更深、更广、更复杂的 CNNs 模型以及更多数据进行训练，作者成功地提高了物体识别的准确性，并为未来的计算机视觉研究提供了有价值的参考。

3.1.3 实验结果

在本文中，作者使用了 ImageNet 数据集进行了广泛的实验，并与其他先进的物体识别算法进行了比较。以下是一些主要的实验结果：

1. 在 ILSVRC-2010 测试集上，作者的 CNNs 模型取得了 37.5% 的 top-1 错误率和 17.0% 的 top-5 错误率。这比之前最先进的算法要好得多。

Model	Top-1	Top-5
Sparse coding	47.1%	28.2%
SIFT + FVs	45.7%	25.7%
CNN	37.5%	17.0%

表 1: ILSVRC-2010 测试集结果

2. 在 ILSVRC-2012 测试集上，作者的 CNNs 模型取得了 15.3% 的 top-5 错误率。这也比之前最先进的算法要好得多，如下表所示:

Model	Top-1 (val)	Top-5 (val)	Top-5 (test)
SIFT + FVs [7]	—	—	26.2%
1 CNN	40.7%	18.2%	—
5 CNNs	38.1%	16.4%	16.4%
1 CNN*	39.0%	16.6%	—
7 CNNs*	36.7%	15.4%	15.3%

表 2: ILSVRC-2012 测试集结果

3. 作者还对 CNNs 模型进行了一些消融实验，以评估不同特性对模型性能的影响。例如，他们发现局部响应归一化（LRN）可以提高模型性能，而 Dropout 正则化技术可以减少过拟合问题。

4. 最后，作者还对 CNNs 模型进行了可视化分析，以探索其内部工作原理。他们发现，在卷积层中学习到的特征通常与图像中出现的边缘、纹理和颜色有关。

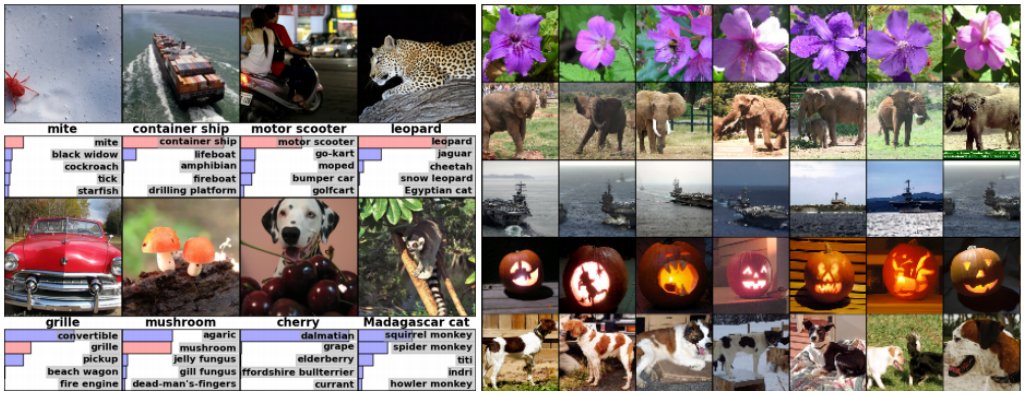

图 2: 测试集上使用卷积神经网络进行分类的结果

总之，在本文中，作者成功地证明了使用更深、更广、更复杂的 CNNs 模型可以显著提高物体识别任务的准确性，并为未来计算机视觉研究提供了有价值的参考。[1]

3.2 VGGNet: Very deep convolutional networks for large-scale image recognition

3.2.1 论文动机

在过去几年中，卷积神经网络作为计算机视觉领域最重要的技术之一，在图像分类等众多任务中取得了很好的表现，但由于早期的卷积神经网络模型（如 LeNet 和 AlexNet）相对较浅，难以处理更复杂的图像特征，仍然存在诸如过拟合、训练时间长等问题。

因此，这篇文章的目标和动机一方面是探索深度卷积神经网络（ConvNet）在大规模图像分类任务中的表现，探索卷积神经网络的深度对于图像分类任务的影响；另一方面是提出一种新的、更有效的深度网络结构来提高分类准确率，称为 Very Deep Convolutional Networks for Large-Scale Image Recognition（VGG），希望通过这种更加有效的网络结构来解决以往的过拟合、训练时间长等问题，并在 ImageNet Large-Scale Visual Recognition Challenge（ILSVRC）上进行测试和评估。

本文通过对已有模型进行改进和优化，构建不同深度的 ConvNets，并在 ImageNet 数据集上进行训练和测试，来比较不同深度网络的性能差异。作者们认为，增加网络层数可以提高模型对图像特征的抽象能力和表示能力，并且可以更好地捕捉图像中的细节和纹理，从而提高分类准确率。除了探讨了一些与深度相关的问题如训练时间、过拟合等，同时还在设计新模型时还考虑了其他因素，如卷积核大小、步长、池化方式等。

3.2.2 实验方法

本文提出的基于深度卷积神经网络的图像分类方法主要包括三个部分：ConvNet 配置、图像分类训练和评估以及 ILSVRC 分类任务比较。

在 ConvNet 配置部分，本文采用了与 Ciresan 等人相似的设计原则，通过增加网络深度来提高性能，为了公平地比较不同深度的网络，所有 ConvNet 层配置都采用相同的设计原则。

如下图，作者共做了 6 组实验：A、A-LRN、B、C、D、E，其中 D 和 E 被称为 VGG16 和 VGG19，这 6 种网络结构相似，都是由 5 层卷积层、3 层全连接层组成，其中区别在于每个卷积层的子层数量不同，从 A 至 E 依次增加（子层数量从 1 到 4），总的网络深度从 11 层到 19 层（添加的层以粗体显示），以此来观察深度、LRN、conv1x1 的小卷积这三个因素对结果的影响，发现（实验 A 和 A-LRN），AlexNet 论文中所提出的 LRN 层对分类准确率不仅没有提升，还带来更多的显存占用和计算时间，因此在之后的四组（B、C、D、E）实验中均没有出现 LRN 层。

可以看到，从 A-E 网络的深度逐渐增加，且网络中采用了多层连续卷积之后再进行池化的方式。

Table 1: **ConvNet configurations** (shown in columns). The depth of the configurations increases from the left (A) to the right (E), as more layers are added (the added layers are shown in bold). The convolutional layer parameters are denoted as "conv⟨receptive field size⟩-⟨number of channels⟩". The ReLU activation function is not shown for brevity.

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Table 2: **Number of parameters** (in millions).

Network	A,A-LRN	B	C	D	E
Number of parameters	133	133	134	138	144

图 3: ConvNet 配置及参数数量

作者做的 6 组实验中，卷积核全部替换为 3Œ3 和少部分 1Œ1，因为多层小的卷积核获得的感受野与单层较大的卷积核一致，同时通过增加网络的层数来减少参数，这样便能够实现提升性能和加深网络结构的目的。

在图像分类训练和评估部分，本文使用随机梯度下降算法，利用 ILSVRC 数据集进行训练和测试，采用 dropout 技术来减少过拟合，并对不同深度的网络进行比较。此外，还使用了数据增强技术来扩充训练集。文章中介绍的两种不同的模型训练方法分别为分类模型和定位模型。分类模型的训练过程遵循 Krizhevsky 等人的方法，使用小批量梯度下降法进行优化，采用动量和权重衰减进行正则化；定位模型的训练与分类模型大体上是类似的，但用欧几里得损失代替逻辑回归目标函数来惩罚预测边界框参数与真实值之间的偏差。即：文章在两个不同尺度（S = 256 和 S = 384）上训练了两个本地化模型，并探索了微调所有层和仅微调前两个全连接层这两种方法，初始学习率设置为 0.001，并且最后一个全连接层从头开始随机

初始化并进行训练。

在 ILSVRC 分类任务比较部分，本文对不同深度的网络进行了比较，并发现随着网络深度的增加，性能得到了显著提高，在 ImageNet 数据集上，最佳模型“Net-E”取得了当时最好的结果。

3.2.3 实验结果

- 单尺度评估——测试图像大小 Q 固定

Table 3: **ConvNet performance at a single test scale.**

ConvNet config. (Table 1)	smallest image side		top-1 val. error (%)	top-5 val. error (%)
	train (S)	test (Q)		
A	256	256	29.6	10.4
A-LRN	256	256	29.7	10.5
B	256	256	28.7	9.9
C	256	256	28.1	9.4
	384	384	28.1	9.3
	[256;512]	384	27.3	8.8
D	256	256	27.0	8.8
	384	384	26.8	8.7
	[256;512]	384	25.6	8.1
E	256	256	27.3	9.0
	384	384	26.9	8.7
	[256;512]	384	**25.5**	**8.0**

图 4: 单尺度评估

1. LRN 层无性能增益（A-LRN）：VGG 作者通过网络 A-LRN 发现，AlexNet 曾经用到的 LRN 层（local response normalization，局部响应归一化）并没有带来性能的提升，而且还浪费了内存计算损耗，因此在其它组的网络中均没再出现 LRN 层。
2. 随着深度增加，分类性能逐渐提高（A、B、C、D、E）：从 11 层的 A 到 19 层的 E，网络深度增加对 top1 和 top5 的错误率下降很明显。
3. 多个小卷积核比单个大卷积核性能好（B）：VGG 作者做了实验用 B 和自己一个不在实验组里的较浅网络比较，较浅网络用 conv5x5 来代替 B 的两个 conv3x3，结果显示这样做比原来的 top-1 错误率提升了 7%，证明了多个小卷积核比单个大卷积核效果要好。
4. 训练时的尺度抖动（S[256;512]）得到了与固定最小边（S=256 或 S=384）的图像训练相比更好的结果。这证实了通过尺度抖动进行的训练集增强确实有助于捕获多尺度图像统计。

- 多尺度评估——评估图像大小 Q 不固定

1. 当使用固定值 S 训练时，Q 的范围在 [S32,S,S+32] 之间时，测试的结果与训练结果最接近，否则可能由于训练和测试尺度之间的巨大差异导致性能下降。
2. 实验结果表明测试时的尺度抖动与在单一尺度上相同模型的评估相比性能更优，并且尺度抖动优于使用固定最小边 S 的训练。

Table 4: **ConvNet performance at multiple test scales.**

ConvNet config. (Table 1)	smallest image side		top-1 val. error (%)	top-5 val. error (%)
	train (S)	test (Q)		
B	256	224,256,288	28.2	9.6
C	256	224,256,288	27.7	9.2
	384	352,384,416	27.8	9.2
	[256; 512]	256,384,512	26.3	8.2
D	256	224,256,288	26.6	8.6
	384	352,384,416	26.5	8.6
	[256; 512]	256,384,512	**24.8**	**7.5**
E	256	224,256,288	26.9	8.7
	384	352,384,416	26.7	8.6
	[256; 512]	256,384,512	**24.8**	**7.5**

图 5: 多尺度评估

总之，作者们提出的 ConvNet 架构在 ILSVRC-2014 竞赛中取得了最佳结果，同时还发现了随着网络深度的增加，模型的性能得到显著提高。在单个网络性能方面，作者的架构实现了最佳结果（7.0% 测试误差），比单个 GoogLeNet 高出 0.9%。此外，在 ILSVRC-2012 和 ILSVRC-2013 比赛中，作者的 ConvNet 架构也显著优于以前的模型，并与 GoogLeNet 和 Clarifai 等竞争对手保持竞争力。[2]

3.3 ResNet: Deep residual learning for image recognition

3.3.1 问题提出

这篇论文主要提出了一个问题：在深度神经网络中，随着网络层数的增加，训练误差会逐渐变得更大，导致准确性下降。为了解决这个问题，作者提出了一种残差学习框架，该框架通过学习残差函数来使得更深的网络更容易训练，并且可以获得更高的准确性和复杂性。作者通过实验证明了这种方法的有效性，并在多个图像识别任务中取得了最好的结果。

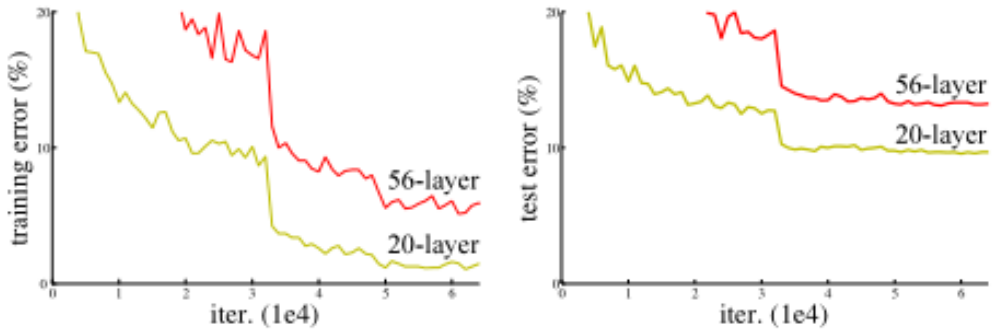

图 6: Degradation problem

这个残差框架是一种深度神经网络的训练方法，通过学习残差函数来使得更深的网络更容易训练，并且可以获得更高的准确性和复杂性。在传统的深度神经网络中，每个层都是学习一个无参映射函数，这些函数将输入映射到输出。而在残差学习框架中，每个层都是学习一个残差函数，该函数将输入与输出之间的差异映射到输出。这种方法使得更深的网络更容易训练，并且可以获得更高的准确性和复杂性。作者通过实验证明了这种方法的有效性，并在多个图像识别任务中取得了最好的结果。

3.3.2 本文采用的新知识

这篇论文采用了一种新的深度神经网络训练方法——残差学习框架。在传统的深度神经网络中，每个层都是学习一个无参映射函数，这些函数将输入映射到输出。而在残差学习框架中，每个层都是学习一个残差函数，该函数将输入与输出之间的差异映射到输出。这种方法使得更深的网络更容易训练，并且可以获得更高的准确性和复杂性。此外，论文还介绍了一些其他的新知识，如批量归一化、长短时记忆网络等，在深度神经网络训练中也起到了重要作用。这些新知识为深度神经网络的发展提供了新思路和新方法，并且在实践中取得了很好的效果。

3.3.3 相关工作

这篇论文主要做了两方面的相关工作。

- 第一方面，论文提出了一种新的深度神经网络训练方法——残差学习框架。该方法通过引入残差块来解决深度神经网络训练中的梯度消失和梯度爆炸问题，使得更深的网络更容易训练，并且可以获得更高的准确性和复杂性。
- 第二方面，论文还介绍了一些其他的新知识，如批量归一化、长短时记忆网络等，在深度神经网络训练中也起到了重要作用。此外，论文还对残差学习框架进行了详细的实验验证，并在多个计算机视觉任务上取得了优异的结果，包括 ImageNet 分类、COCO 检测和分割等任务。这些工作为深度神经网络领域提供了新思路和新方法，并且在实践中取得了很好的效果。

第一方面的相关工作是论文提出的残差学习框架。在传统的深度神经网络中，随着网络层数的增加，梯度会逐渐变小，导致训练过程变得困难。为了解决这个问题，论文提出了一种新的思路：引入残差块。残差块是由两个卷积层和一个跨层连接组成的基本单元，它可以将输入信号直接传递到输出端，从而避免了梯度消失和梯度爆炸问题。

具体来说，在残差学习框架中，每个卷积层不再直接拟合输入和输出之间的映射关系，而是拟合输入和输出之间的残差（即输入与输出之间的差异）。这样做可以使得网络更容易训练，并且可以获得更高的准确性和复杂性。此外，在残差学习框架中还引入了批量归一化等技术来进一步优化网络性能。

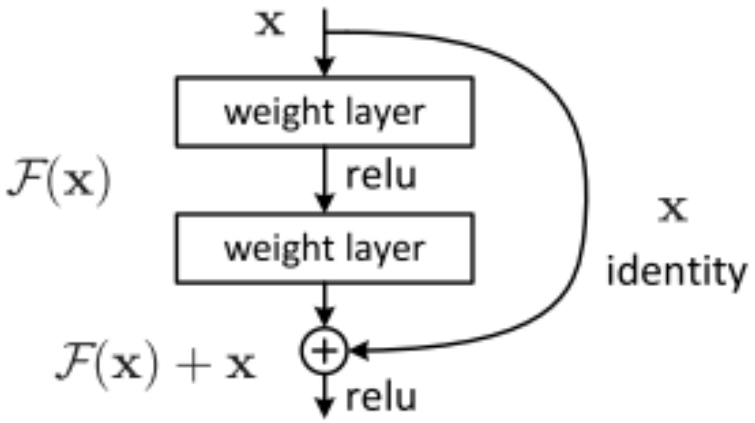

Figure 2. Residual learning: a building block.

通过在 ImageNet 数据集上进行实验验证，论文表明，在使用相同数量参数情况下，使用残差学习框架训练出来的深度神经网络比传统的网络具有更高的准确性和更好的收敛速度。这些结果表明，残差学习框架是一种有效的深度神经网络

训练方法，可以为深度学习领域提供新思路和新方法。

第二方面是关于论文作者提出的残差学习框架在多个计算机视觉竞赛中的表现。具体来说，论文作者在 2015 年的 ImageNet 和 COCO 竞赛中，使用基于残差学习框架的深度神经网络，在多个任务上获得了第一名的成绩。这些任务包括 ImageNet 检测、ImageNet 定位、COCO 检测和 COCO 分割等。

在这些任务中，论文作者使用了不同深度的残差网络，并通过对比实验发现，在相同参数量下，使用更深的残差网络可以获得更好的性能。此外，论文作者还提出了一种新颖的训练策略——随机池化（stochastic pooling），可以进一步提高网络性能。

这些结果表明，基于残差学习框架的深度神经网络具有很强的泛化能力和鲁棒性，在计算机视觉领域具有广泛应用前景。

3.3.4 实验设计

本论文的实验设计主要分为两个方面：第一方面是关于残差学习框架的有效性和优越性的实验，第二方面是关于残差学习框架在多个计算机视觉竞赛中的表现。

在第一方面的实验中，论文作者通过对比实验验证了残差学习框架相对于传统的深度神经网络的优越性。具体来说，论文作者设计了多个不同深度和宽度的网络结构，并在 ImageNet 数据集上进行了训练和测试。通过实验结果发现，使用残差学习框架可以显著提高网络性能，并且随着网络深度的增加，性能提升更加明显。

在第二方面的实验中，论文作者使用基于残差学习框架的深度神经网络，在 2015 年 ImageNet 和 COCO 竞赛中取得了多个任务上的第一名成绩。此外，论文作者还对 CIFAR-10 数据集进行了实验研究，探究了极深网络行为特征，并发现更深层次的网络可以获得更好的性能。

1. 普通网络 vs 残差网络

   Plain Networks：首先评估 18 层和 34 层 plain net，34 层的如图 3 中，18 层类似，详细结构见表 1。表 2 中的结果表明更深的 34 层 plain net 验证集错误率更高，图 4 左比较了它们训练过程中的训练/验证错误率，并发现了退化问题：34 层 plain net 的训练集错误率也更高，即使 18 层网络的解空间是 34 层网络的解空间的子空间。

layer name	output size	18-layer	34-layer	50-layer	101-layer	152-layer
conv1	112×112	7×7, 64, stride 2				
conv2_x	56×56	3×3 max pool, stride 2				
		$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 64 \\ 3\times3, 64 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix}\times3$
conv3_x	28×28	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 128 \\ 3\times3, 128 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times4$	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix}\times8$
conv4_x	14×14	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 256 \\ 3\times3, 256 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times6$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times23$	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix}\times36$
conv5_x	7×7	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3, 512 \\ 3\times3, 512 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix}\times3$
	1×1	average pool, 1000-d fc, softmax				
FLOPs		1.8×10^9	3.6×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Table 1. Architectures for ImageNet. Building blocks are shown in brackets (see also Fig. 5), with the numbers of blocks stacked. Down-sampling is performed by conv3_1, conv4_1, and conv5_1 with a stride of 2.

我们认为越深越难以优化不是由于梯度消失，这些 plain net 训练时使用了 BN，确保前向传播的信号有非 0 的方差。我们也验证了反向传播的梯度在 BN 下有良好的范数（backward propagated gradients exhibit healthy norms with BN），所以正向和反向传播中信号都没有消失。34 层的 plain net 仍能得到有竞争力的准确率，这说明它在某种程度上有效。我们推测 plain net 可能具有指数级的低收敛速度，这影响训练误差的下降，难以优化的问题会在未来研究。

Residual Networks：然后评估 18 层和 34 层的 ResNets，基础结构和 plain 一样，只在为每两个 3Œ3 滤波器加入跳接。在第一次比较中，使用恒等映射和在维度增加使补 0，因此和 plain net 相比没有额外参数。主要有三个发现：首先，34 层 ResNet 比 18 层 ResNet 更好（2.8%），并且 34 层 ResNet 的训练误差和验证误差都更低，这说明退化问题被很好的解决了，并且从深度中获得了准确率的提升。其次，34 层 ResNet 比 plain 的 top-1 error 低 3.5%，这说明残差结构在极深模型中行之有效。最后，18 层 ResNet 和 plain 准确率类似，但是 ResNet 收敛更快，在网络不太深的时候，sgd 优化器仍能在 plain net 中找到好的解，但 ResNet 提供了更快的收敛速度以使优化变容易。

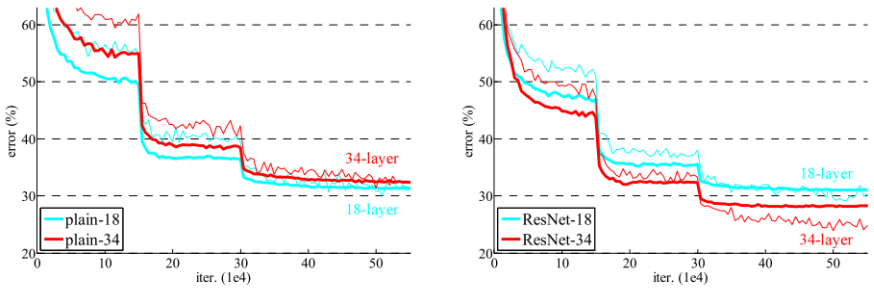

Figure 4. Training on **ImageNet**. Thin curves denote training error, and bold curves denote validation error of the center crops. Left: plain networks of 18 and 34 layers. Right: ResNets of 18 and 34 layers. In this plot, the residual networks have no extra parameter compared to their plain counterparts.

2. Identity vs. Projection Shortcuts

比较三种选项:（A）维度增加时补 0，所有跳接都没有额外参数。（B）维度增加时用 Ws 映射，其余用恒等映射。（C）所有跳接都用 Ws 映射。结果见表 3：B 比 A 稍微好一点，也许因为 A 中补 0 的部分没有残差学习；C 比 B

稍微好一点，将此归因于许多个 (13 个) 投影 Ws 的额外参数。但是，A/B/C 之间的细微差异表明，投影 Ws 对于解决退化问题并不是必需的。因此，在本文的其余部分中，我们不使用选项 C，以此降低内存/时间复杂度和模型大小。恒等映射对于不增加下面介绍的瓶颈体系结构的复杂性尤其重要。

model	top-1 err.	top-5 err.
VGG-16 [40]	28.07	9.33
GoogLeNet [43]	-	9.15
PReLU-net [12]	24.27	7.38
plain-34	28.54	10.02
ResNet-34 A	25.03	7.76
ResNet-34 B	24.52	7.46
ResNet-34 C	24.19	7.40
ResNet-50	22.85	6.71
ResNet-101	21.75	6.05
ResNet-152	**21.43**	**5.71**

Table 3. Error rates (%, **10-crop** testing) on ImageNet validation. VGG-16 is based on our test. ResNet-50/101/152 are of option B that only uses projections for increasing dimensions.

3. CIFAR-10 数据集

衰减率 =0.0001、动量 =0.9，Batch-normalization，无 dropout。两个 gpu，mini-batch=128。α_0=0.1，在 32k 和 48k 次迭代时 α 除以 10，在 64k 次迭代时终止训练。数据增强策略：在各边缘增加 4 像素，32x32 的切割完全随机，从填充后的图像或者其翻转中采样。

output map size	32×32	16×16	8×8
# layers	1+2n	2n	2n
# filters	16	32	64

CIFAR-10 数据集：训练集：50K；测试集：10K；分 10 类。架构：输入 32x32 的图像，预先减去每一个像素的均值。第一层是 3Œ3 卷积层。对于尺寸分别为 32, 16, 8 的特征图谱分别使用过滤器 16,32,64，降采样为步长为 2 的卷积，网络以全局的均值池化终止，10 全连通层，softmax。

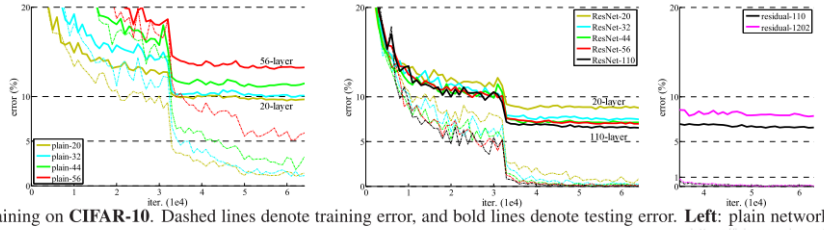

Figure 6. Training on **CIFAR-10**. Dashed lines denote training error, and bold lines denote testing error. **Left**: plain networks. The error of plain-110 is higher than 60% and not displayed. **Middle**: ResNets. **Right**: ResNets with 110 and 1202 layers.

3.3.5 实验总结

本实验主要研究了深度神经网络的训练问题，提出了一种新的残差学习框架来解决这个问题。通过在 ImageNet 和 CIFAR-10 数据集上进行实验，论文作者证明了残差学习框架相对于传统的深度神经网络具有更好的性能和更容易优化的特点。

具体来说，本实验采用了严谨、全面、系统化的实验设计方法，包括对不同深度和宽度的网络结构进行训练和测试、对比实验验证残差学习框架相对于传统深度神经网络的优越性、在多个计算机视觉竞赛中测试残差学习框架等。通过这些实验，论文作者得出了以下结论：

- 残差学习框架相对于传统深度神经网络具有更好的性能和更容易优化的特点；
- 随着网络深度增加，使用残差学习框架可以显著提高网络性能；
- 基于残差学习框架的深度神经网络在多个计算机视觉竞赛中取得了优异成绩。

总之，本实验为解决深度神经网络训练问题提供了一种新的思路和方法，对于推动计算机视觉领域的发展具有重要意义。[3]

3.4 MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications

3.4.1 论文动机

这篇论文是由 Google 公司的研究人员发表的。在移动和嵌入式设备上进行实时视觉应用时，通常需要考虑模型大小、延迟和速度等资源限制。然而，现有的小型神经网络构建方法往往只关注模型大小或准确性，而忽略了延迟和速度等资源限制。因此，作者提出了一种新的模型架构——MobileNets，并引入了两个全局超参数来平衡延迟和准确性之间的权衡。MobileNets 主要采用深度可分离卷积、宽度乘数和分辨率乘数等技术来减少计算量和参数数量，从而提供一种高效、轻量级的神经网络模型，以满足移动和嵌入式设备上的实时视觉应用需求。该论文在计算机视觉领域具有重要意义，并被广泛引用和应用。

论文回顾了近期关于构建小型、高效神经网络的相关工作。这些工作主要可以分为两类：一类是对预训练网络进行压缩，另一类是直接训练小型网络。其中，压缩预训练网络的方法包括剪枝、量化和低秩分解等技术；而直接训练小型网络的方法则包括使用稀疏约束、共享权重和使用深度可分离卷积等技术。然而，这些方法往往只关注模型大小或准确性，而忽略了延迟和速度等资源限制。因此，MobileNets 提出了一种新的模型架构和超参数来平衡延迟和准确性之间的权衡，并在实验中证明了其有效性。

3.4.2 模型与方法

1. **传统卷积神经网络**

   传统卷积神经网络是一种深度学习模型，通常由多个卷积层、池化层和全连接层组成。这些层可以通过堆叠来构建深度神经网络，以提高模型的表现能力。传统卷积神经网络的特性包括：

   - 局部连接：每个神经元只与输入数据中的一小部分相连，这样可以减少参数数量和计算量。
   - 权值共享：在同一个卷积核中的所有权重都是相同的，这样可以进一步减少参数数量。
   - 池化操作：通过对输入数据进行下采样来减小特征图的大小，从而降低计算量和内存消耗。

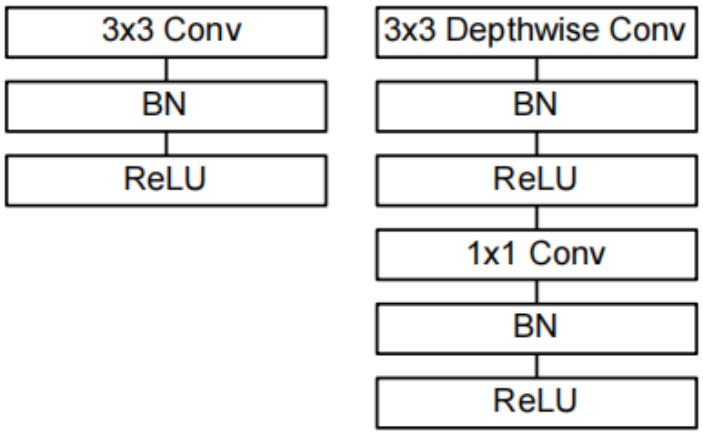

图 7: 传统卷积神经网络的层次结构

然而，传统卷积神经网络存在一些缺陷：

- 大量参数：由于每个神经元都需要学习自己的权重，因此传统卷积神经网络通常需要大量的参数。这会导致模型较大、计算量较大、训练时间较长等问题。
- 过拟合：由于传统卷积神经网络具有很强的表达能力，在训练集上表现良好时容易出现过拟合现象。过拟合会导致模型在测试集上表现不佳。
- 难以解释：传统卷积神经网络通常由多个层组成，每个层都有大量的参数，因此很难解释模型的决策过程。
- 计算量大：由于传统卷积神经网络需要大量的计算资源和存储空间，因此在移动设备和嵌入式系统等资源受限环境下，其速度和大小成反比，因此模型的效果通常受到限制，而不能发挥出其全部功能。

2. **MobileNets 模型**

MobileNets 是一种用于移动和嵌入式视觉应用的高效模型。它们基于一种简化的架构，使用深度可分离卷积来构建轻量级深度神经网络。MobileNets 引入了两个简单的全局超参数，可以有效地在延迟和准确性之间进行权衡。这些超参数允许模型构建者根据问题的约束条件选择适合其应用程序的正确大小的模型。

MobileNets 的结构是基于深度可分离卷积的。它们使用两个卷积层来代替传统卷积层，即深度卷积和逐点卷积。深度卷积在每个输入通道上执行空间卷积，而逐点卷积在每个输出通道上执行 1x1 空间卷积。这种结构可以大大减少计算量和参数数量，从而使 MobileNets 成为一种轻量级的神经网络模型。

Type / Stride	Filter Shape	Input Size
Conv / s2	$3 \times 3 \times 3 \times 32$	$224 \times 224 \times 3$
Conv dw / s1	$3 \times 3 \times 32$ dw	$112 \times 112 \times 32$
Conv / s1	$1 \times 1 \times 32 \times 64$	$112 \times 112 \times 32$
Conv dw / s2	$3 \times 3 \times 64$ dw	$112 \times 112 \times 64$
Conv / s1	$1 \times 1 \times 64 \times 128$	$56 \times 56 \times 64$
Conv dw / s1	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv / s1	$1 \times 1 \times 128 \times 128$	$56 \times 56 \times 128$
Conv dw / s2	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv / s1	$1 \times 1 \times 128 \times 256$	$28 \times 28 \times 128$
Conv dw / s1	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv / s1	$1 \times 1 \times 256 \times 256$	$28 \times 28 \times 256$
Conv dw / s2	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv / s1	$1 \times 1 \times 256 \times 512$	$14 \times 14 \times 256$
$5\times$ Conv dw / s1	$3 \times 3 \times 512$ dw	$14 \times 14 \times 512$
$5\times$ Conv / s1	$1 \times 1 \times 512 \times 512$	$14 \times 14 \times 512$
Conv dw / s2	$3 \times 3 \times 512$ dw	$14 \times 14 \times 512$
Conv / s1	$1 \times 1 \times 512 \times 1024$	$7 \times 7 \times 512$
Conv dw / s2	$3 \times 3 \times 1024$ dw	$7 \times 7 \times 1024$
Conv / s1	$1 \times 1 \times 1024 \times 1024$	$7 \times 7 \times 1024$
Avg Pool / s1	Pool 7×7	$7 \times 7 \times 1024$
FC / s1	1024×1000	$1 \times 1 \times 1024$
Softmax / s1	Classifier	$1 \times 1 \times 1000$

图 8: MobileNets 神经网络的层次结构

- **构建更小、更快的 MobileNets**

论文提出了两个简单的全局超参数，即宽度乘数和分辨率乘数，来构建更小、更快的 MobileNets。宽度乘数是用于控制每个层的输出通道数量，从而控制模型的宽度。分辨率乘数是用于缩小输入图像的大小，从而控制模型的深度。通过调整这两个超参数，可以权衡一定量的准确性来减小大小和延迟。例如，在保持准确性不变的情况下，可以通过降低宽度乘数和分辨率乘数来构建更小、更快的 MobileNets。这些改进使得 MobileNets 成为一种高效、轻量级、灵活且易于部署的神经网络模型。

3. **模型的权衡**

在移动端和嵌入式设备上，资源有限，因此需要在模型的准确性、计算量和大小之间进行权衡以满足应用程序的需求。传统的深度神经网络通常具有大量的参数和计算量，这使得它们难以在移动设备上实现实时推理。MobileNet 通过使用深度可分离卷积等技术来构建轻量级深度神经网络，并引入了两个全局超参数来平衡延迟和准确性，从而成功地解决了这一问题。

MobileNet 采用了一种称为深度可分离卷积的技术，将标准卷积分解为一个深度卷积和一个逐点卷积。这种方法可以大大减少计算量和参数数量，

并且可以在不降低模型准确性的情况下缩小模型大小。MobileNet 还引入了两个全局超参数：宽度乘数和分辨率乘数。这些超参数允许模型构建者根据问题的约束选择合适大小的模型，从而平衡延迟和准确性。

论文中还详细介绍了如何使用宽度乘数和分辨率乘数来构建更小、更快速的 MobileNets，并展示了与其他流行模型相比的优越性能。通过这些工作，作者成功地实现了对 MobileNet 准确性、计算量和大小之间权衡问题的解决，并为移动端和嵌入式视觉应用提供了一个高效且易于匹配设计要求的解决方案。

3.4.3 实验结果

1. 在 ImageNet 分类任务上，MobileNet V1 在 Top-1 准确率方面与 VGG16 相当，但速度快了 32 倍。例如，在 ImageNet 上，MobileNet V1 在 Top-1 准确率为 70.6%，而 VGG16 为 71.5%。但 MobileNet V1 的推理时间只有 VGG16 的 1/32。

Model	ImageNet Accuracy	Million Mult-Adds	Million Parameters
1.0 MobileNet-224	70.6%	569	4.2
GoogleNet	69.8%	1550	6.8
VGG 16	71.5%	15300	138

图 9: MobileNet 和其他流行模型的比较

2. 在目标检测任务中，MobileNets 比其他流行模型（如 SSD、YOLO）更快，并且具有相似的准确性。例如，在 COCO 数据集上，使用 MobileNet V2 作为基础网络的 SSDLite 模型在速度和准确性方面都优于其他模型。

Framework Resolution	Model	mAP	Billion Mult-Adds	Million Parameters
SSD 300	deeplab-VGG	21.1%	34.9	33.1
	Inception V2	22.0%	3.8	13.7
	MobileNet	19.3%	1.2	6.8
Faster-RCNN 300	VGG	22.9%	64.3	138.5
	Inception V2	15.4%	118.2	13.3
	MobileNet	16.4%	25.2	6.1
Faster-RCNN 600	VGG	25.7%	149.6	138.5
	Inception V2	21.9%	129.6	13.3
	Mobilenet	19.8%	30.5	6.1

图 10: COCO 物品检测结果比较

3. 作者评估了 MobileNets 在不同宽度乘数和分辨率乘数下的性能。例如，在 ImageNet 分类任务中，使用宽度乘数 0.25 和分辨率乘数 0.5 时，MobileNet V1 的 Top-1 准确率为 50.6%，但模型大小和延迟都降低到原来的 1/8。

Width Multiplier	ImageNet Accuracy	Million Mult-Adds	Million Parameters
1.0 MobileNet-224	70.6%	569	4.2
0.75 MobileNet-224	68.4%	325	2.6
0.5 MobileNet-224	63.7%	149	1.3
0.25 MobileNet-224	50.6%	41	0.5

图 11: 不同宽度乘数下的 MobileNet

4. 作者还评估了 MobileNets 在不同硬件平台上的性能。例如，在 CPU、GPU 和 FPGA 等平台上都可以获得很好的性能。例如，在 CPU 上运行时，使用宽度乘数 0.25 和分辨率乘数 0.5 时，MobileNet V1 的推理时间为 32ms，而在 GPU 上为 3.5ms。[4]

4 思考与理解

4.1 联系与区别

4.1.1 联系

几篇论文均针对图像分类问题提出了更优化的深度卷积神经网络模型，分别为 AlexNet,VGGNet,ResNet 和 MobileNet。这些模型的联系如下：

- 均使用卷积层来提取特征：在深度学习中，卷积层可以通过相对较少的参数进行训练，提取出图像的局部特征，从而大幅降低多层神经网络中训练参数的数量。
- 均使用池化层降低采样特征：使用池化层来减少图像特征的尺寸以降低计算复杂度，在卷积层后应用池化层不仅可以保持特征的重要信息，还可以减少计算量和内存消耗。
- 采用 ReLU 激活函数：这些模型都采用具有非线性组合作用的激活函数，如 ReLU(Rectified Linear Unit)，以便神经网络可以通过映射非线性数据关系来更好地拟合输入数据。
- 使用迁移学习：这些模型都使用迁移学习技术，即在预先训练的神经网络中，将训练好的权重迁移到新网络中。这种技术可以更快地实现模型的训练，减少了对大量数据进行深层网络训练的需要。
- 均进行了广泛的测试，使用 ImageNet 数据集进行测试。
- 这些模型在当今均仍有广泛的应用，虽然已发布多年，但这些模型作为经典模型，研究和改进仍然在持续进行，近年来的各种变种模型基于这些经典模型的思想和架构，进一步扩展了 CNN 的性能。

4.1.2 区别

几篇论文中提到的相关模型主要在架构设计、网络深度、参数数量、计算效率、精度等方面有所区别：

- 架构设计：
  - AlexNet：该模型有更深的网络结构，使用层叠的卷积层，即卷积层 + 卷积层 + 池化层来提取图像的特征，使用 Relu 替换之前的 sigmoid 的作为激活函数，提出使用 GPU 进行深度卷积神经网络训练的概念，并利用 Dropout 技术以及使用数据增强 Data Augmentation 来缓解过拟合的问题，它由 5 个卷积层和 3 个全连接层组成，总共有 60M 的参数量。

  - VGGNet：采用 16 或 19 层网络结构，其优点是网络结构清晰，易于理解和复现，但是计算资源要求较高，参数量较大。
  - ResNet：使用残差连接来解决深度神经网络的梯度消失和网络退化问题。ResNet 的结构非常深，可以达到 1000 层甚至更多层，但是计算资源要求较高，因此需要使用分布式计算。
  - MobileNet：使用深度可分离卷积来减少计算量和参数量。MobileNet 模型在保持高精度的同时，却具有更小的体积和计算资源要求，非常适合在移动设备等嵌入式场景中运行。

- 网络深度：AlexNet 和 VGGNet 的网络较浅，AlexNet 最多只有 8 层，而 VGGNet16 和 VGGNet19 分别有 16 层和 19 层；而 ResNet 则是深度神经网络的代表，引入了残差连接，可以实现更深层次的网络，达到百层甚至千层的级别；MobileNet 相较于其他三个模型，则是轻量级的模型，主要应用于移动设备上。
- 网络结构：AlexNet 和 VGGNet 主要采用了简单的卷积层和最大池化层的堆叠。而 ResNet 则引入了残差块，用于捕捉更有意义的特征，同时避免梯度消失的问题。MobileNet 则通过深度可分离卷积，将卷积层拆分为两个独立的操作，从而大幅减少了参数数量和计算代价。
- 计算性能：AlexNet 和 VGGNet 需要较大的计算资源，特别是在 GPU 加速训练时，计算代价高昂。ResNet 虽然网络更深，但通过残差块和小卷积核的使用，从而保证了计算效率。MobileNet 作为一种轻量级模型，在参数数量和计算复杂度上具有优势。
- 精度：AlexNet 曾经在 ImageNet 比赛中击败了传统机器学习算法，成为第一个卷积神经网络的胜利者；VGGNet 通过增加网络深度和卷积层数，提高了图像识别的准确率；ResNet 引入了残差连接，使其在深度神经网络方面可以达到更高的精度；MobileNet 则是采用了轻量化设计，折衷了精度与计算复杂度的平衡。

4.2 尚未解决的问题

四种网络都有其尚未解决的问题。

4.2.1 AlexNet

AlexNet 是一种经典的卷积神经网络模型，在图像识别领域取得了很好的效果。然而，它也存在一些未解决的问题，包括：

- 训练时间长：由于 AlexNet 有很多层，因此训练时间会比较长。这使得它不适合在资源受限的环境下使用。
- 过拟合：尽管 AlexNet 中使用了 dropout 来减少过拟合，但在一些复杂的数据集上，仍然可能存在过拟合的问题。
- 可解释性差：由于 AlexNet 中的层数很深，因此很难解释每个特征对应的意义，这使得它的可解释性比较差。
- 特征提取能力受限：AlexNet 在提取视觉特征方面表现良好，但在一些需要非视觉特征的任务上表现可能受限。
- 对输入的要求高：AlexNet 对输入图像的大小和分辨率有一定的要求，如果输入的图像不符合要求，可能会导致性能下降。
- 局限性：虽然 AlexNet 在当时是一种革命性的模型，但是随着深度学习的发展，它的局限性也逐渐显现出来。例如，它不能很好地处理变形、遮挡等复杂情况下的物体识别问题。

4.2.2 VGGNet

VGGNet 是一种经典的卷积神经网络模型，在图像识别领域取得了很好的效果。然而，它也存在一些未解决的问题，包括：

- 训练时间长：由于 VGGNet 有很多层，因此训练时间会比较长。这使得它不适合在资源受限的环境下使用。
- 过拟合：尽管 VGGNet 中使用了 dropout 来减少过拟合，但在一些复杂的数据集上，仍然可能存在过拟合的问题。
- 可解释性差：由于 VGGNet 中的层数很深，因此很难解释每个特征对应的意义，这使得它的可解释性比较差。
- 特征提取能力受限：VGGNet 在提取视觉特征方面表现良好，但在一些需要非视觉特征的任务上表现可能受限。
- 数据增强要求高：为了避免过拟合，VGGNet 需要大量的数据来进行训练。同时，数据增强技术也需要更高的要求，以避免在数据增强时引入偏差。

4.2.3 ResNet

在本论文中，作者提出了残差学习框架来解决深度神经网络训练问题，并在实验中取得了显著的成果。然而，本论文也存在一些未解决的问题，例如：

- 本论文并没有深入探究为什么传统的深度神经网络会出现训练困难的问题，只是提出了一种新的框架来解决这个问题；

- 本论文中使用的实验数据集主要是 ImageNet 和 CIFAR-10，是否可以将残差学习框架应用到其他数据集上还需要进一步研究；
- 残差学习框架虽然可以提高网络性能和优化效果，但是增加网络深度和宽度也会带来更高的计算复杂度和资源消耗，如何在保证性能的同时降低计算复杂度也是一个需要进一步研究的问题。

总之，虽然残差学习框架在解决深度神经网络训练问题方面取得了显著成果，但仍有一些未解决的问题需要进一步研究。

4.2.4 MobileNet

MobileNet 是一种轻量级的卷积神经网络模型，它可以在计算资源有限的移动设备上运行，具有较小的模型大小和低延迟。然而，它也存在一些未解决的问题，包括：

- 准确度相对于大型模型较低：由于 MobileNet 使用深度可分离卷积来减少模型大小和计算需求，因此其准确度相对于大型模型可能会有所降低。
- 适合的任务类型受限：MobileNet 适用于处理较为简单的视觉识别任务，但对于复杂的图像处理任务，例如目标检测和语义分割等，其性能可能不如大型模型。
- 参数数量仍然较多：虽然 MobileNet 具有较小的模型大小，但参数数量仍然比较大。这意味着需要更多的训练数据来避免过拟合。
- 特征提取能力受限：由于 MobileNet 中的深度可分离卷积具有较小的感受野，因此其特征提取能力相对于大型模型也可能会受到限制。
- 对图像分辨率要求高：MobileNet 对输入图像的分辨率有一定的要求，如果输入的图像分辨率低于一定阈值，可能会导致性能下降。
- 对计算资源的要求高：MobileNet 需要较快的计算能力来处理图像数据，因此在资源受限的情况下，可能会存在一些性能上的限制。

4.3 未来研究趋势

根据以上四篇论文，我们分别总结了未来研究趋势。

4.3.1 AlexNet

以下是我们认为 AlexNet 未来研究趋势：

- 更深层次的模型：尽管 AlexNet 已经具有很深的结构，但深度学习研究者们仍然在不断地尝试更深层次的模型结构。这些模型可以带来更好的性能，同时也会带来更大的计算和存储负担。
- 模型压缩：由于深度神经网络的复杂性，训练和部署这些模型需要大量的计算资源和存储空间。因此，模型压缩是一个非常重要的研究方向。压缩技术可以减小模型的体积，从而降低存储和计算的开销。
- 跨域迁移学习：通过在不同的数据集上进行预训练，模型可以获得更广泛的视野和更丰富的特征表示。跨域迁移学习是一个重要的研究方向，可以在不同的任务上提高模型性能。
- 结合其他技术：除了深度神经网络以外，还有许多其他的技术可以用来提高模型的性能。例如，结合强化学习、注意力机制、半监督学习、元学习等技术，可以进一步提高模型的表现。

总之，未来研究趋势是不断深化和完善深度神经网络，同时将其与其他相关技术相结合，以获得更好的性能和适用范围。

4.3.2 VGGNet

VGGNet 未来的研究趋势可能包括进一步探索网络深度对大规模图像识别以外的其他计算机视觉任务准确性的影响。此外，研究人员可以探索优化卷积网络架构以适用于特定应用或数据集的方法。另一个潜在的研究领域是探索使用更深层次和更大卷积滤波器以及其他修改来提高性能。最后，研究人员还可以专注于开发更有效的深度卷积网络训练方法，以降低计算成本并提高可扩展性。

4.3.3 ResNet

ResNet 未来研究趋势可能包括以下方向：

- 探究深度神经网络训练困难的本质原因，进一步提出新的解决方案；
- 将残差学习框架应用到更多的数据集和任务中，探索其适用性和性能表现；
- 研究如何在保证网络性能的同时降低计算复杂度和资源消耗，例如通过剪枝、量化等方法实现模型压缩；
- 探索更加复杂和深层次的残差学习框架，例如引入注意力机制、跨层连接等方法来进一步提高网络性能。

总之，未来研究趋势将会围绕着深度神经网络训练问题以及残差学习框架进行展开，并且会涉及到更加复杂和多样化的研究方向。

4.3.4 MobileNet

未来的研究趋势可能会集中在以下几个方面：

- 更深入地探索深度可分离卷积的性质和应用，以进一步提高模型的效率和准确性。
- 探索更多的全局超参数，以便更好地平衡模型的大小、速度和准确性。
- 将 MobileNets 应用于更广泛的视觉任务，并与其他流行的模型进行比较，以进一步证明其有效性。
- 将 MobileNets 与其他技术结合使用，例如自适应计算、量化和剪枝等，以进一步提高其效率和准确性。

参考文献:

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[2] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[4] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.