文章编号：2095-2163(2021)09-0113-07 中图分类号：TP391.4 文献标志码：A

基于深度迁移学习的饮食图像识别研究

王策仁，彭亚雄，陆安江
（贵州大学 大数据与信息工程学院，贵阳 550025）

摘 要： 卷积神经网络(CNN)应用于图像识别具有很大优势，但是需要足够深的网络和大量标签完善的数据集才能发挥其优越性。实际应用中，往往需要应对的是质量差和大小不一的数据集，且受硬件设备限制。为了提高图像识别效率和精度，提出一种基于深度卷积神经网络和迁移学习的识别算法。该算法首先对图像预处理和数据增强，后迁移大样本提取出的特征信息用于 CNN 特征提取，再接入微调网络对数据集再训练。实验结果显示，本文算法对饮食识别的精度和时间性能均有显著的提高，精确度最高可达 98%以上，精度提升最高可达 10%以上，时间性能提升幅度最高可达 110% 。

关键词： 深度学习；图像识别；卷积神经网络；迁移学习；微调网络；特征提取

Research on recognition of dietary images based on deep transfer learning

WANG Ceren, PENG Yaxiong, LU Anjiang
(Collage of Big Data and Information Engineering, Guizhou University, Guiyang 550025, China)

【Abstract】 Convolutional neural network (CNN) has tremendous advantages when applied to image recognition, but it requires a deep network and a large number of well-labeled datasets to exploit its superiority. In practical daily life, it often needs to deal with datasets with bad quality and inconsistent size and limited by hardware devices. In order to improve the efficiency and accuracy of image recognition, a recognition algorithm based on deep convolutional neural networks and transfer learning is proposed. The image is pre-processed and augmented first, then the feature information extracted from large samples is transferred for CNN feature extraction. Finally the fine-tuned network is accessed to target datasets. The experimental results indicate that the algorithm has a significant improvement in both accuracy and time performance for diet image recognition, with the accuracy up to more than 98%, precision improvement up to more than 10%, and time performance improvement up to more than 110%.

【Key words】 deep learning; image recognition; Convolutional Neural Networks (CNNs); transfer learning; fine-tuning networks; feature extraction

0 引 言

计算机视觉学科领域发展至今，图像分类问题一直是此领域的重要内容。随着互联网技术的飞速发展，网络上的图像无论从数据大小、数量基数和种类的丰富都呈现指数式的增长。近几年，机器学习在图像识别领域发展迅速，便于进行矩阵和浮点计算的 GPU 性能快速进步，而作为机器学习的重要分支，其可以通过大量数据和 GPU 硬件加速来增强深度学习快速在学习科研[1]、工业应用[2]、医学诊断[3]等领域得到广泛的运用。

深度学习拥有比较优秀的建模和分析能力，在图形分类中优势明显。为此需要两个前提条件：一是足够深且合适的网络模型，二是质量好且标注优秀的数据集提供训练[4]。此外，对实验的硬件要求很高。

饮食的种类丰富多样，又是人们日常生活的必要因素，随着国家、民族、地域的不同，其种类风格也大相径庭。由于食物图片的数量之多，采用传统的深度学习方法，对硬件和数据集的质量要求太高，花费的时间较多，并难以得到较好的结果。人们日常接触到的食物图像数据样本大多质量一般，要么是样本的数量较少，要么是样本中无效数据或标注较少[5]。但是，这些通过卷积神经网络提取出来的特征信息部分，具有不错的泛化能力，可通过对迁移学习合理运用增加模型性能。

1 神经网络算法分析

深度学习是机器学习领域的一个重要分支，2012 年 Hinton 等人[6]构建全连接网络和卷积神经网络结合的模型，并获得了当年 ImageNet 大规模视觉识别挑战大赛(ImageNet Large Scale Visual Recognition

基金项目：贵州省科技成果转化项目([2017]4856)。

作者简介：王策仁(1995-)，男，硕士研究生，主要研究方向：人工智能、图像识别；彭亚雄(1963-)，男，学士，副教授，主要研究方向：数字通信技术、音视频处理；陆安江(1978-)，男，博士，副教授，主要研究方向：嵌入式系统与集成技术、物联网安全、微传感技术。

收稿日期：2021-07-19

Competition, ILSVRC)[7]图像分类目标的冠军,同时文献[6]中首次使用了 GPU 来加速模型的训练,借以寻求更优的网络和更准确的算法,从此奠定了卷积神经网络在图像分类领域的领先。

本文采用在 ILSVRC 大赛图像分类项目中具有不同特性的 VGG16[8]、ResNet50[9]和 MobileNetV2[10]网络,用于特征提取,迁移其在 ImageNet(源域)上训练出的特征和权重知识到饮食图像分类(目标域)中[11],通过 Pre-training + Fine-Tuning(预训练+微调)方式再训练。

1.1 改进型 VGG16 神经网络

VGG16 网络模型是由牛津大学视觉几何组在 2014 开发的,整个网络共 16 层,由 13 层 conv 卷积层和 3 层 fc 全连接层共同组成(5 层 maxpool 池化层不计入层数)。最后通过 3 层全连接层输出1 000 个分类。

文献[8]中指出,网络的核心思路是通过堆叠对多个 3x3 卷积核来取代尺寸更大的卷积核,即通过感受野(receptive field)[12]的方式,由两个 3x3 卷积核堆叠取代 5x5 的卷积核;3 个 3x3 的取代 7x7 的卷积核,如图 1 所示。计算公式为:

$$N(i) = [N(i+1) - 1] \cdot S + K \tag{1}$$

其中:$N(i)$ 是第 i 层感受野;S 是步距;K 为卷积核大小。

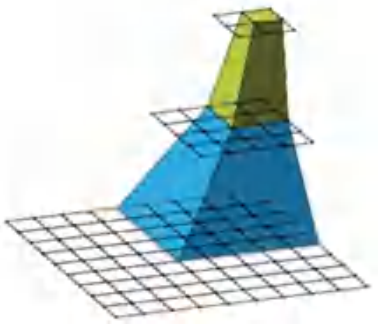

图 1 感受野

Fig. 1 Receptive Field

1.2 改进型 ResNet50 神经网络

ResNet50 在 2015 年由微软提出,其在结构上的优越性可以让网络突破1 000层组成超深的网络。得益于文献[9]中提出的残差(residual)结构,配合使用了 BN(Batch Normalization)[13]来加速训练,减少了依赖 dropout 等拟合的方法。

如图 2 所示,网络由多组残差模块构成。ResNet50 先通过一个 7x7,步距为 2 的卷积层,再经过 3x3,步距为 1 的最大池化层,后由 4 种不同的共 16 组残差模块组成。

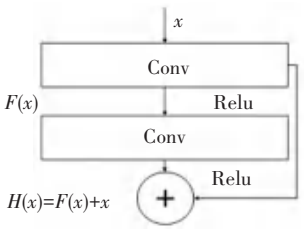

图 2 残差结构

Fig. 2 Architecture of the residual

其中,$F(x)$ 为正常通过卷积层的输出;$H(x)$ 为正常卷积层输出与捷径输出之和;*ReLu* 函数是较为主流的非线性激活函数。[14] 如图 3 所示,其公式定义为:

$$f(x) = \max(0, x) \tag{2}$$

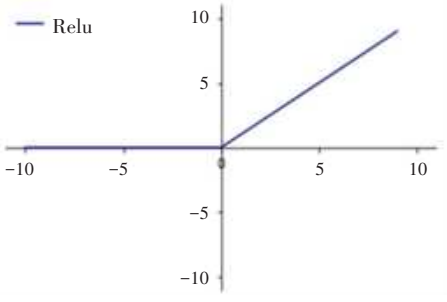

图 3 ReLu 函数图像

Fig. 3 Function of ReLu

文献[13]中提出的方式,可以使每个 *batch* 的数据满足均值为 0,方差为 1 的分布规律。公式如下:

batch 均值:

$$u_\beta = \frac{1}{m}\sum_{i=1}^{m} x_i \tag{3}$$

batch 的方差:

$$\sigma_\beta^{\ 2} = \frac{1}{m}\sum_{i-1}^{m}(x_i - u_\beta)^2 \tag{4}$$

标准化:

$$\hat{x}_i = \frac{x_i - u_\beta}{\sqrt{\sigma_\beta^2 - \varepsilon}} \tag{5}$$

最后得出:

$$y_i = \hat{\gamma x}_i + \beta N_{\gamma,\beta}(x_i) \tag{6}$$

其中,u、σ^2 由正向传播统计得到;γ、β 由反向传播训练得到;m 为一个 *batch* 总数。

1.3 改进型 MobileNetV2 神经网络

MobileNetV2 是由 Google 团队在 2018 年研发的网络模型,相比其它的卷积神经网络,沿用了文献[15]的 DW(Depthwise Convolution),卷积可以大量的减少网络的参数使用量和计算量,且仅牺牲小幅

的精确度，相比 VGG16 精确度减少了 0.9%，模型参数只有 VGG16 的 1/32[15]。

该网络首次提出 Inverted Residuals 倒残差结构（图 4）和采用 ReLu6 非线性激活函数（图 5）提高准确度，使模型变的更小。

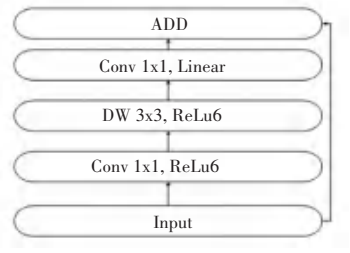

图 4　倒残差结构

Fig. 4　Inverted residuals structure

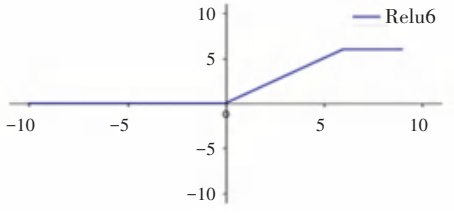

图 5　ReLu6 函数图像

Fig. 5　Function image of ReLu6

2　深度迁移网络与数据

本文以文献[8-10]提出的 3 种网络结构为基础，迁移 ImageNet 上的特征信息。首先载入不同的饮食数据集，对图像预处理和数据增强，然后通过对 CNN 网络接入分布式微调的方式进行训练。流程如图 6 所示。

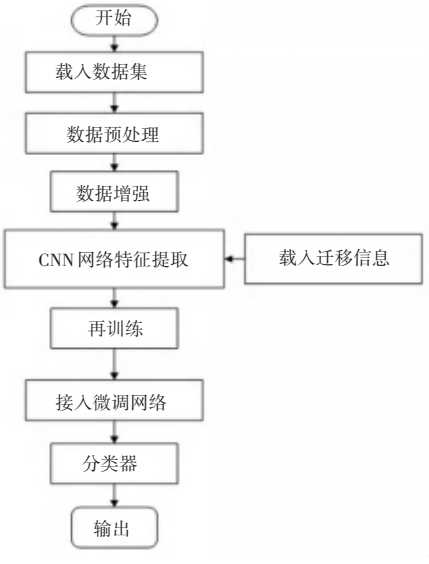

图 6　深度迁移学习加微调流程图

Fig. 6　Deep transfer learning and fine-tuning flow chart

2.1　数据预处理与分析

本文共使用 2 个数据集，数据来源于 kaggle 机器学习竞赛网站公开的饮食分类数据集。饮食 11 数据集中共12 000张图片，选取其中9 000张作为训练集，3 000张为验证集，数据集共 11 种分类，其中包括日常饮食中真实拍摄的食物，具有较好的代表性。饮食 101 数据集是一个涵盖范围更广的饮食数据集，共 101 类食物100 000张图，通过 python 编写的小程序对数据进行随机切分（比例控制在 8:2），即其中80 000张作为训练集，20 000张为测试集。

由于数据集中图片都是日常拍摄，其像素和尺寸不一，甚至有的图片的主体并不是目标物。所以，在训练之前需将图像进行预处理。首先进行统一的水平裁切处理（验证集不需进行裁切），将图片等比例填充成 280 x 280 的纯黑色图片，再还原成分辨率统一裁剪为 224 x 224 的彩色图片，并提高一定量的对比度和亮度，然后对数据集进行标准化和归一化处理后，再进行旋转变换、平移变化和随机组合，达到增强数据的效果。

2.2　深度迁移网络

本文采用特征迁移的方式进行迁移学习[16]，通过 ResNet50、VGG16 和 MoblieV2 网络对数据集进行特征提取，对最后几层全连接层失活，后接入分布式微调网络，通过微调网络进一步学习。

2.3　分布式微调网络

因本文所采用的数据集均是对特定类别的细化分类，通过原网络中学习的先验知识并不能完全的用于提取数据集的的特征，从而导致精确度下降，花费的训练时间还很多，所以需要对原网络全连接层进行失活，接入分布式微调网络，使用复杂度不一的微调网络，分布式的载入模型中，进而提高准确度，该方法即为分布式微调网络。

图 7~9 即为本文分布式微调网络结构。通过失活原 CNN 网络全连接层和分类器，再分布式加入具有不同特征的微调网络，对比不同形式的微调网络带来不同的结果进行分析。

图 9 是失活原网络全连接层后，直接通过一个 256 神经元的全连接线性层后，接入分类器；图 10 同样接入一个 256 神经元的全连接线性层后，进行激活和 Dropout[17]，Dropout 数值设为 0.5（即随机失活一半的神经元），最后接入分类器；图 11 则是接入1 024神经元的全连接线性层后，进行激活，再接入 512 神经元的全连接线性层后激活，逐级减半，最后接入 265 神经元的线性层后接入分类器。

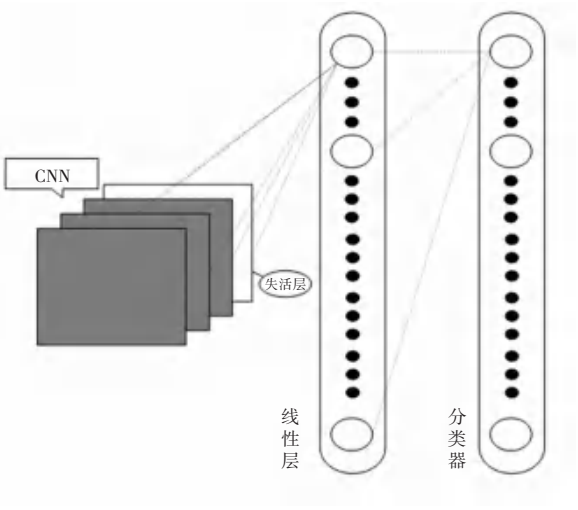

图 7 微调 1 网络结构

Fig. 7 Fine-Tuning 1

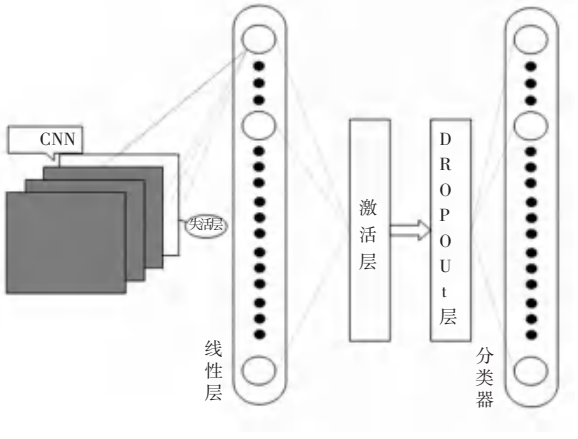

图 8 微调 2 网络结构

Fig. 8 Fine-Tuning 2

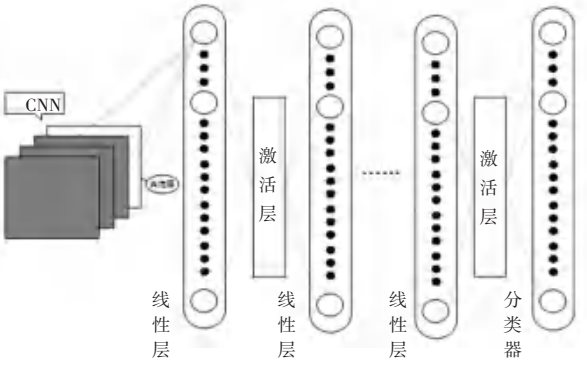

图 9 微调 3 网络结构

Fig. 9 Fine-Tuning 3

3 实验与结果分析

3.1 实验设置

本文实验基于深度学习 pytorch 库,实验模型和数据在 Windows 10 平台完成训练,计算机 CPU 为 intel i7 9750H,主频 2.6 GHz,最高睿频 4.5 GHz,GPU 使用 NVIDIA GTX 1660Ti,计算机内存为 16 GB。采用 CPU 和 GPU 并行计算,加速模型训练。

分别搭建 ResNet50、VGG16、MoblieNetV2 网络并载入该网络在 ImageNet 上训练得到的特征权重信息,对数据集进行特征提取。失活 CNN 网络中的全连接层后接入微调网络,对不同的微调网络训练结果记录并对比分析。

使用 torchvision 工具[18] 对数据集进行边缘裁剪、标准化和归一化等预处理后,对图片进行数据增强。

实验中各组 CNN 和微调网络组成的深度迁移学习网络模型中,Batchsize 均设置为 64。通过 CrossEntropy(交叉熵)[19] 来计算损失函数,公式如下:

$$H(p,q) = -\sum_{x}(p(x)\log q(x)) \tag{7}$$

其中,p 是真实值,q 是预测值,p 值经过 ReLu 函数激活后在 0~1 之间。同时实验使用 Adam 优化器来优化梯度下降过程。其可通过梯度的一阶二阶矩估计,自适应调节学习率。表达式如下:

$$m_t = u \cdot m_{t-1} + (1-u)\cdot g_t \tag{8}$$

$$n_t = v \cdot n_{t-1} + (1-v)\cdot g_t^2 \tag{9}$$

$$\hat{m}_t = \frac{m_t}{1-u_t} \tag{10}$$

$$\hat{n}_t = \frac{n_t}{1-v_t} \tag{11}$$

$$\theta_t = -\frac{\hat{m}_t}{\sqrt{\hat{n}_t + \varepsilon}}\cdot \eta \tag{12}$$

公式(8)、(9)分别是对梯度的一阶矩估计和二阶矩估计。其中,t 为梯度时刻;g 是梯度;u 和 v 是指数衰减率。而公式(10)、(11) 是对一阶、二阶矩估计的校正,近似对期望的无偏估计。公式(12) 则是对学习率 η 的动态约束。其中,ε 是一个极小值数,避免分母为 0。实验优化器默认学习率设置为 0.000 1。

3.2 结果分析

本实验基于 3 种 CNN 网络,通过迁移学习技术,迁移特征权重信息对数据特征提取后,接入分布式微调网络进行再学习和分类。经过上述实验设置,得到不同网络搭配的精确度和迭代时间。

图 10 ~ 12 分别对应 ResNet50、VGG16、MobileNetV2 分布式接入微调网络及传统 CNN 训练

的饮食 11 分类精确度对比图。

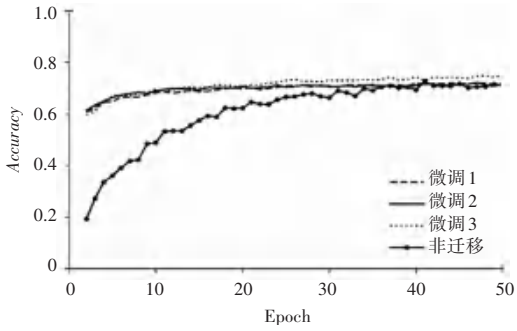

图 10　ResNet50 饮食 11 分类精确度对比图

Fig. 10　ResNet50 diet11 classification accuracy comparison chart

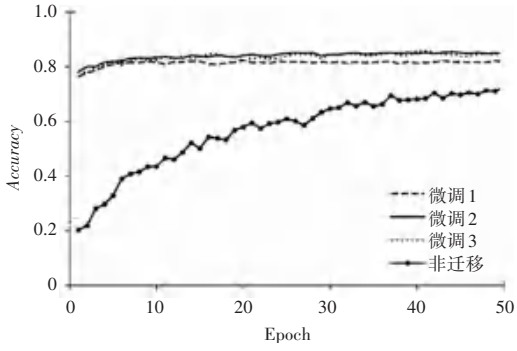

图 11　VGG16 饮食 11 分类精度对比图

Fig. 11　VGG16 diet11 classification accuracy comparison chart

从图中可以看出，分布式微调的 3 种方式在经过 2-4 次迭代后都可以达到精度较高且平稳上升的结果，同时训练出的精确度比传统 CNN 训练有很大提升。分步式微调的 3 种方式随着微调网络复杂度的变换，精确度也在提升。其中微调 3 方式效果最好。

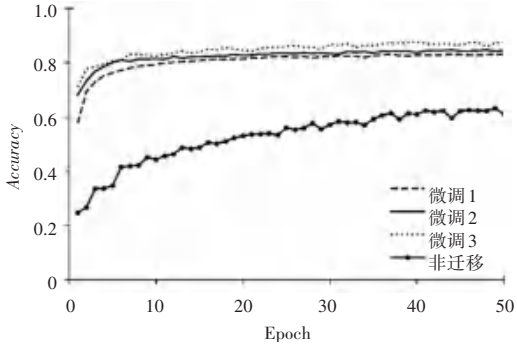

图 12　MobileNetV2 饮食 11 分类精度对比图

Fig. 12　MobileNetV2 diet11 classification accuracy comparison chart

图 13～15 则是对饮食 101 分类数据集的精确度对比图，其精确度对比结果同食物分类数据集的结果相似，迁移加分布式微调的可以快速且平稳得到一个优秀的效果且领先于传统 CNN 网络训练。对比饮食 11 分类，训练出的效果同样优势明显，且同样是微调 3 网络精确度更好。

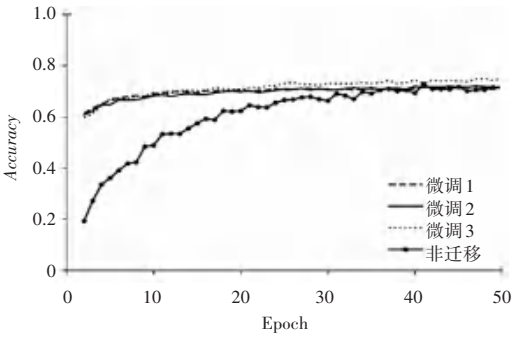

图 13　ResNet50 饮食 101 分类精度对比图

Fig. 13　ResNet50 diet101 classification accuracy comparison chart

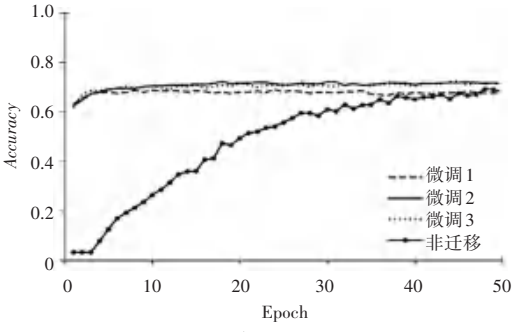

图 14　VGG16 饮食 101 分类精度对比图

Fig. 14　VGG16 diet101 classification accuracy comparison chart

通过上述精度对比图可以看出，迁移加微调 3 网络从精度上看，两个数据集都可以得到不错的精确度，远超原 CNN 训练的结果。ResNet50、VGG16、MobileNetV2 这 3 种网络在微调 3 模式下对比原网络，在 50 组迭代下饮食，11 分类分别提升 15%、13.5%和 24.4%，饮食 101 分类分别提 4.8%、3.5%和 5.1%。MoblieNetV2 提升最大，VGG16 提升较少。

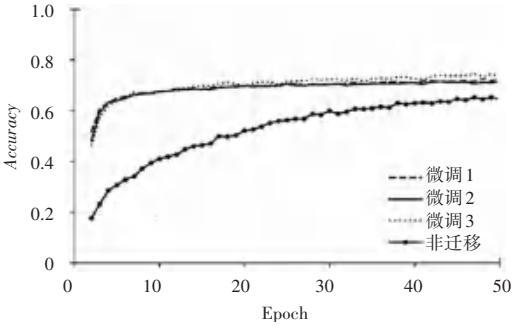

图 15　MoblieNetV2 饮食 101 分类精度对比图

Fig. 15　MoblieNetV2 diet101 classification accuracy comparison chart

对网络进行更深层次训练，精度对比结果见表 1。3 种网络进行微调 3 模式下训练 200 组后，对比原网络的精度，结果与 50 组迭代相似，ResNet50 的精确度最高，VGG16 次之，而在提升幅度上 ResNet50 提升最大，MoblieNetV2 次之。

表 1 200 组迭代精确度对比表

Tab. 1 Comparison of accuracy of 200 epoch of iterations %

算法	饮食 11 分类精确度	饮食 101 分类精确度
ResNet50 微调 3	98.2	86.7
Vgg16 微调 3	95.8	84.2
MobileNetV2 微调 3	93.5	81.9
ResNet50	78.2	78.1
Vgg16	77.3	77.7
MoblieNetV2	75.1	74.6

本文评价网络性能的参考因素:一是网络的精确度,二则是网络的训练时间(时间性能)。实验记录了每次迭代的平均训练时间作为参考,见表 2。在两个饮食分类数据集中,随着微调网络的复杂度增加,各个模式下花费的时间均有不同程度的增加,VGG16 的网络复杂度最高,故花费时间最多,ResNet50 次之,而以精简著称的 MoblieNetV2 时间最少。同样饮食 11 分类数据量和分类的复杂度都远小于饮食 101 分类,故饮食 101 分类的花费时间度远小于饮食 11 分类。

表 2 平均单位迭代时间

Tab. 2 Average iteration time s/epoch

算法	饮食 11 分类	饮食 101 分类
ResNet50 微调 1	136.12	274.79
ResNet50 微调 2	139.58	280.24
ResNet50 微调 3	148.89	330.68
ResNet50	196.23	523.77
Vgg16 微调 1	170.76	301.98
Vgg16 微调 2	174.26	306.17
Vgg16 微调 3	181.76	315.58
Vgg16	288.48	648.86
MoblieNetV2 微调 1	108.67	194.45
MoblieNetV2 微调 2	112.54	198.25
MoblieNetV2 微调 3	115.73	220.90
MoblieNetV2	140.95	320.09

3 种 CNN 网络搭配 3 种微调模式共 9 种模式,在花费时间上都远小于原 CNN 训练的结果。在饮食 11 分类中 3 种网络时间性能分别平均提升 25.5%、64.3%和 38.8%,在饮食 101 分类中提升更大,分别为 78.6%、110.7%和 56.9%。3 种 CNN 网络中,对硬件性能要求最高、使用参数量最多的 VGG16 网络的时间性能提升最大,而网络最为精简的 MoblieNetV2 所花费的时间最少。当然这和其精简的网络结构,更少的参数使用量和对设备性能的依赖性息息相关。

综合对比本实验的精确度和时间性能,在精确度层面,ResNet50 配合微调 3 模式得到的效果最好,200 组迭代后提升达到 10%和 7.6%,提升幅度最大;VGG16 网络的精度第二,提升幅度第三;MoblieNetV2 提升幅度第二,但精度最差。在对时间性能分析结果表明:VGG16 时间性能提升最大,平均时间性能提升在两个数据集分类中分别达到 64.3%和 110.7%,ResNet50 次之。可见,本实验采用的 CNN 复杂网络引入迁移学习预训练进行特征提取,后接入微调网络进行再训练,对复杂的网络、硬件设备要求越高的网络和数据复杂度越大的网络的提升就越大,无论是时间性能还是网络精度都有很好的效果。

学习配合分布式微调网络的策略可在极短的时间内获得优秀的结果。无论是精确度还是时间性能都远优于文献[8-10]CNN 网络的零基础训练。

4 结束语

本文选用迁移学习解决深度学习中的饮食图像分类问题,提出一种 CNN 网络配合迁移学习的方法,采用分布式微调的方式增加网络的性能。本文采用 3 种不同的 CNN 网络,配合分布式微调网络对数据集分类训练。实验结果表明,本文采用的方法在精确度和时间性能上都远强于原网络训练的结果。综合对比下,VGG16 配合微调 3 模式的性能最好,时间性能和精度都有不错的结果。下一步将建立更多的微调模式对不同的数据集进行训练,提高分类方法的性能和实用性。

参考文献

[1] 卢宏涛,张秦川.深度卷积神经网络在计算机视觉中的应用研究综述[J].数据采集与处理,2016,31(1):1-17.

[2] 田增垚,彭飞,孟庆东,等. 基于降噪循环神经网络的风电功率预测[J].微电子学与计算机,2021,38(3):27-32.

[3] 朱晓铭,陈林海,张帅,等. 基于卷积神经网络重建十二导联心电图[J].微电子学与计算机,2019,36(9):12-15.

[4] 吴正文. 卷积神经网络在图像分类中的应用研究[D]. 成都:电子科技大学,2015.

[5] 郑欣悦. 基于深度学习的少样本图像分类方法[D]. 北京:中国科学院大学(中国科学院国家空间科学中心),2019.

[6] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60(6): 84-90.

[7] DENG J, DONG W, SOCHER R, et al. Imagenet: A large-scale hierarchical image database [C]//2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009: 248-255.

(下转第 122 页)