文章编号：2095-2163(2022)07-0160-05 中图分类号：TP391 文献标志码：A

改进 MobileNetV3 的脱机手写汉字识别

程若然，周浩军，刘露露，贺 炎
（上海工程技术大学 电子电气工程学院，上海 201620）

摘 要：针对目前手写识别网络训练时间长、高资源消耗等问题，本文提出了一种脱机手写汉字识别网络模型。使用轻量级网络 MobileNetV3 作为主干网络，以减少网络参数量；针对汉字识别分类数庞大的特点，使用多尺度卷积核，提取更丰富的特征信息；针对形近字易产生识别错误的问题，使用注意力机制进行局部、全局特征提取并融合。实验结果表明，所提模型能在保持较少参数量的情况下，使其识别准确率有所提升。

关键词：脱机手写汉字识别；深度学习；MobileNetV3；特征融合；注意力机制

Offline handwritten Chinese character recognition based on improved MobileNetV3

CHENG Ruoran, ZHOU Haojun, LIU Lulu, HE Yan
(School of Electrical and Electronic Engineering ,Shanghai University of Engineering Science, Shanghai 201620, China)

【Abstract】 To solve the problem of long training time and high resource consumption of handwritten Chinese character recognition network, an offline handwritten Chinese character recognition network model is proposed in this paper. The lightweight network MobileNetV3 is used as the backbone network to reduce the number of network parameters. In view of the large number of Chinese character recognition classification, multi-scale convolution kernel is used to extract richer feature information. In view of the problem that Chinese characters with similar shapes are easy to be recognized incorrectly, an attention mechanism is adopted to extract local features and global features and fuse them. Experimental results show that the proposed model can improve the accuracy while keeping fewer parameters.

【Key words】 offline handwritten Chinese character recognition; deep learning; MobileNetV3; feature fusion; attention mechanism

0 引 言

文字是人类社会最重要的交流载体之一，随着互联网与人类社会的联系越发紧密，手写汉字识别在人们生活中起到越来越大的作用。手写文字识别主要分为联机手写汉字识别和脱机手写汉字识别。前者使用相关设备记录书写轨迹的各项数据，利用笔顺信息来进行文字识别；后者则使用图像采集设备获取手写文字图像，通过学习图像与汉字字符编码之间的映射来识别文字。

手写汉字识别从上世纪 80 年代起不断发展，传统方法逐渐形成“预处理、特征提取、分类”的流程来进行手写汉字识别[1]，并获得了不错的识别效果。但在实际应用中，更复杂的手写风格和识别模式使得文字识别率下降，用户难以获得最佳的性能体验。近年来，一些研究人员开始利用深度学习方法进行手写汉字识别。2013 年的 ICDAR 手写汉字识别竞赛[2]的优胜者，利用深度学习方法获得了远超传统方法的识别率，展现了深度学习在文字识别领域的极大潜能。

目前，基于深度学习的手写汉字识别方法存在训练时间长、高资源消耗的问题。为此，本文利用 MobileNetV3 作为主干网络进行脱机手写汉字识别，融合多尺度空间特征，提高了训练速度。

1 MobileNetV3 网络模型

MobileNetV3[3] 是 2019 年 Google 研发的 MobileNet 系列的新作。MobileNet 系列网络模型是为了能在移动端设备（如在手机上）运行而设计的轻量型网络，且继承了 V1 版本[4]的深度可分离卷积（Depthwise Separable Convolution）和 V2 版本[5]的逆残差（Inverted Residuals）和线性瓶颈（Linear Bottlenecks）结构。为了进一步提升分类准确率，V3 版本引入了 SE（Squeeze-and-excitation）结构，并对网络进一步剪枝以减少计算量，加快训练速度。

1.1 深度可分离卷积

如图 1 所示，标准卷积是每个卷积核与输入特征图的所有通道按位进行卷积计算，参数量为 $k \times k \times C_i \times C_o$、计算量是 $k \times k \times C_i \times C_o \times W \times H$。

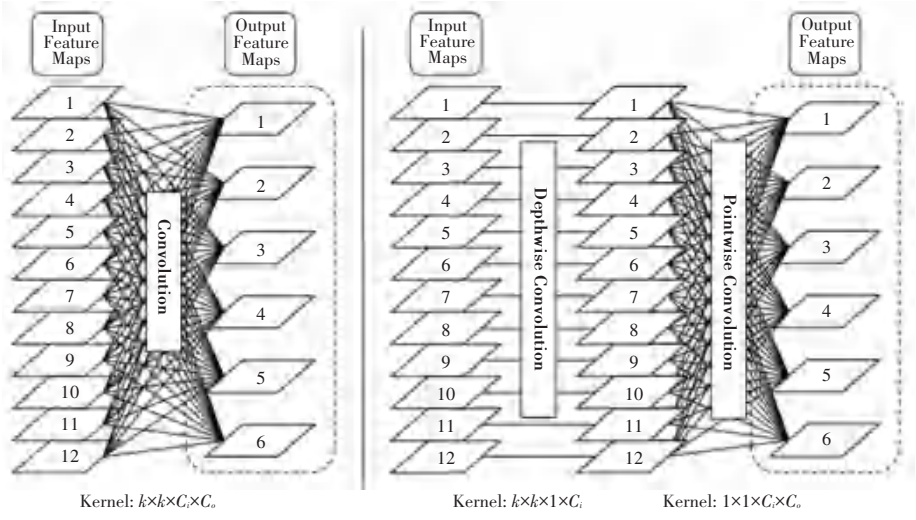
其中，k 是卷积核大小；C_i、C_o 是输入通道数（输

作者简介：程若然（1996-），女，硕士研究生，主要研究方向：计算机视觉。

收稿日期：2021-11-02

入特征图的通道数以及卷积核通道数）和输出通道数（即卷积核个数）；W 和 H 是输出特征图的宽和高。

深度可分离卷积分为深度卷积和点卷积两步完成：深度卷积是用 C_i 个大小为 $k \times k$、通道数为 1 的卷积核，对输入特征图的 C_i 个通道分别进行卷积计算，参数量为 $k \times k \times 1 \times C_i$、计算量为 $k \times k \times 1 \times C_i \times W \times H$；点卷积是使用 C_o 个通道大小为 1×1、通道数为 C_i 的卷积核，对深度卷积的输出进行标准卷积操作，参数量为 $1 \times 1 \times C_i \times C_o$，计算量为 $1 \times 1 \times C_i \times C_o \times W \times H$。因此，深度可分离卷积在参数量上减少为标准卷积的 $1/C_o + 1/(k \times k \times C_i)$，在计算量上减少为标准卷积的 $1/C_o + 1/(k \times k)$。当使用 3×3 大小的卷积核时，理论上深度可分离卷积的计算速度应是标准卷积的 8 ~ 9 倍，而精度只与标准卷积相差 1%。



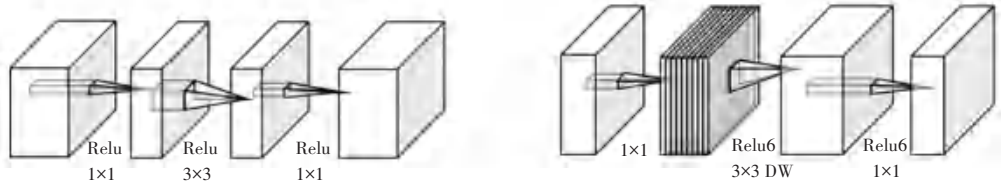
（a）标准卷积　　（b）深度可分离卷积

图 1　标准卷积与深度可分离卷积

Fig. 1　Standard convolution and depth separable convolution

1.2　带线性瓶颈的逆残差结构

逆残差结构是依据 ResNet[6] 的残差结构改进而得。残差结构先用 1×1 卷积核压缩输入特征图的通道数，再用 3×3 卷积核进行特征提取，最后用 1×1 卷积核扩张回原本的通道数，整体流程为“压缩-特征提取-扩张”，特征图通道数量先减小后增大。逆残差结构先用 1×1 卷积核扩张输入特征图的通道数，再用 3×3 卷积核进行深度卷积，最后用 1×1 卷积核压缩回原本的通道数，整体流程为“扩张-特征提取-压缩”，特征图通道数量先增大后减小。二者结构如图 2 所示。



（a）残差结构　　（b）倒残差结构

图 2　残差结构与倒残差结构

Fig. 2　Residual structure and inverted residual structure

由于低维分布嵌入到高维空间之后，再使用 RELU 激活函数由高维空间投影回低维空间，将会造成信息损失。针对该问题，MobileNet V2 引入了线性瓶颈，即将逆残差模块最后一层的 RELU 改为 *Linear* 激活函数。

1.3　MobileNetV3

MobileNet V3 在前代版本的基础上，首先调整了网络的输入输出层，减少了输入输出层的卷积操作，以降低计算机资源消耗，提升了 15% 的运算速度；其次去除了 *sigmoid* 函数的计算，改用计算消耗更小的 *h - swish* 函数；在逆残差模块中 3×3 深度可分离卷积之后引入 SE 模块[7]，结构如图 3 所示。

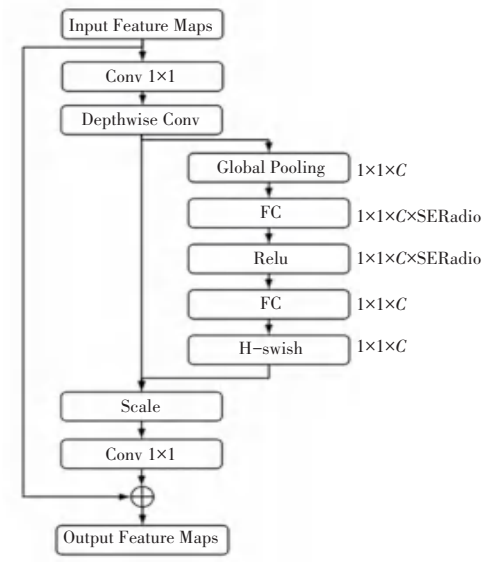


图 3 MobileNet V3-SE 模块

Fig. 3 MobileNet V3-SE module

MobileNet V3-SE 的基本实现过程为:先进行全局池化压缩(Squeeze)获得一个 1×1× C 的向量;然后经过两次“全连接层-激活(Excitation)”操作(为减少计算时间将第一次“全连接层-激活”操作的输出通道数压缩为原来的 1/4),输出 1×1× C 的向量;最后将得到的向量与深度可分离卷积的结果按位相乘,以调整每个通道的权值,从而提升网络精度。V3 总体网络结构的设计中,首先通过 NAS 算法,对网络结构进行搜索优化(如网络中 Block 的排列和结构),得到大体的网络构成,最后使用 NetAdapt 算法来确定每个 filter 的 channel 数量。

2 基于 MobileNetV3 的脱机手写汉字识别

本文研究目标是脱机手写汉字的识别,主要面临两个问题:

(1)汉字数量多,相当于数千级别的分类问题。如此大数量的分类网络,需要更丰富的特征信息。

(2)形近字的识别容易得到错误结果(如“已”和“已”)。

针对上述问题,本文以 MobileNetV3 为主干网络,设计了一种多尺度特征提取方案,并使用一种新的注意力机制[8]进行特征融合,改进后的 MobileNetV3 能够更好地适应脱机手写汉字识别任务。

2.1 改进的 MobileNetV3

本文在输入图像进行多次特征提取之后,加入一个特征提取模块来丰富特征信息。该模块包含两个支路:使用多尺度大小的卷积核来获取原始输入图像上不同范围大小的感受野,并利用注意力机制融合这两个分支的特征信息。优化后的网络模型如图 4 所示,其中虚线框所指模块即为本文的改进之处,MS-CAM 是多尺度通道注意力模块。

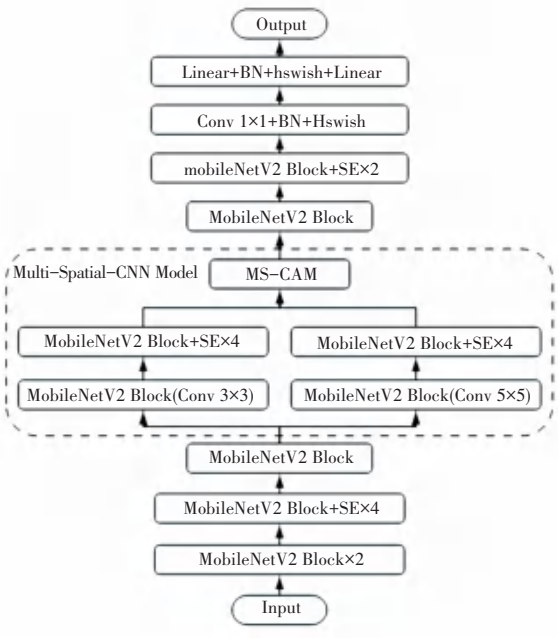


图 4 改进的 MobileNetV3

Fig. 4 Improved MobileNetV3

2.2 感受野

在卷积神经网络中,感受野是一个非常重要的概念,指的是每个网络层输出的特征图像上,每个神经元所能“看见”的原始输入图像上对应区域的范围大小。如图 5 所示,每一层的卷积核大小都为 3×3,卷积步长为 1,填充大小为 0。下一层特征图的每个神经元能看到上一层特征图 3×3 大小的区域,进而能看到再上一层 5×5 大小的区域。也就是说,越深的网络层越能看到原始输入图像上更多的内容。

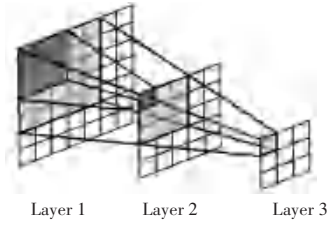


图 5 感受野示意图

Fig. 5 Receptive field diagram

网络中不同大小的感受野会带来不一样的性能表现,而感受野的大小则受到各参数的影响(如卷积核大小、卷积步长、填充大小等)。为了能够从原始输入图像获得更丰富的特征信息,本文网络使用 3×3 和 5×5 多尺度大小的卷积核,来获取原始输入图像上不同范围大小的感受野。

2.3　多尺度通道注意力特征融合

特征融合是来自不同层或分支特征的组合，是卷积神经网络中常见的操作内容，通常通过简单线性的操作（如：求和或拼接）来实现。文献[8]中认为，这样的特征融合方式并不是最佳选择，因此提出了一种新的基于注意力的特征融合方法。本文利用该方法中提出的多尺度通道注意力模块（MS-CAM），通过不同特征的通道注意力来赋予两个分支不同的权重，从而完成其融合，其结构如图 6 所示。

MS-CAM 除了进行池化，还有一个分支使用逐点卷积来进行特征提取，并将两个特征进行融合。

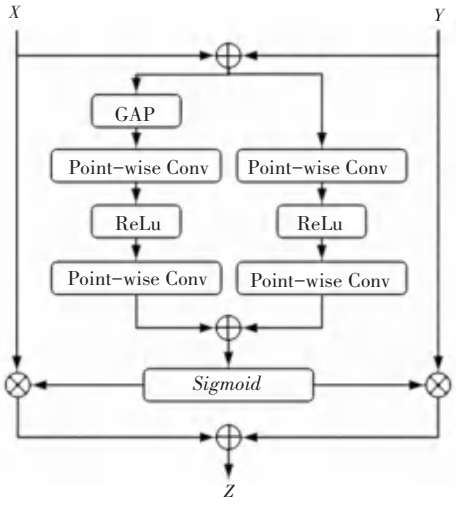


图 6　MS-CAM 结构图

Fig. 6　Structure of MS-CAM

前者更关注全局尺度上的大型对象，而后者更关注通道注意力不同尺度上下文的特征信息。SE 的计算方式非常容易丢失原始图像上的细节信息，而 MS-CAM利用多尺度特征提取方式，更好地捕获局部特征信息。

3　实验

3.1　数据集

本文使用 CASIA-HWDB[9] 脱机单字符数据库中的 HWDB 1.1 数据集进行实验。该数据集由中国科学院自动化研究所模式识别国家实验室建设。脱机单字符数据库包括 3 个字符集，其统计数据见表 1。其中，HWDB 1.0 数据集的 3 866 个汉字包含 GB2312-80 字符集 3 755 个一级汉字中的 3 740 个汉字；HWDB 1.1 数据集的 3 755 个汉字即为 GB2312-80 字符集一级汉字全集；HWDB 1.2 数据集的 3 319 个汉字与 GB2312-80 字符集一级汉字集不相交。

表 1　脱机单字符数据库

Tab. 1　Database of offline single character

数据集	汉字数	符号数	手写人数	汉字总样本数
HWDB 1.0	3 866	171	420	1 609 136
HWDB 1.1	3 755	171	300	1 121 749
HWDB 1.2	3 319	171	300	990 989

由于样本尺寸不一致，在输入网络之前均将其处理成 224×224 大小。训练、测试、验证集随机划分成 8∶1∶1。数据集内部分样本示例如图 7 所示。

图 7　部分样本示例

Fig. 7　Part of samples

3.2　评价标准

实验使用准确率（*Accuracy*）作为评价指标，当进行二分类时，其计算公式如下：

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

其中，*TP*、*TN*、*FP*、*FN* 属于混淆矩阵的概念，其含义见表 2。

表 2　混淆矩阵及其含义

Tab. 2　Confusion matrix

	类别为正（Positive）	类别为负（Negative）
预测为真（True）	真正（True Positive，TP） 将正类预测为正类的数量	真负（True Negative，TN） 将负类预测为负类的数量
预测为假（False）	假正（False Positive，FP） 将负类预测为正类的数量	假负（False Negative，FN） 将正类预测为负类的数量

准确率的含义就是被正确分类的样本数量与总样本数的比值。本文的目标任务是一个多分类(类别数 N = 3 755)问题,将二分类扩展为多分类,在计算第 i 个类别的准确率 ACC_i 时,应将第 i 类视为正类,其它类别视为负类,计算完所有类别的准确率后计算平均值。因此准确率的计算公式如下:

$$ACC = \frac{1}{N}\sum_{i=0}^{N-1}\frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_{ii}} \quad (2)$$

3.3 实验结果分析

本文所有的实验均在 Ubantu 18.04 系统上使用 CUDA 并行计算架构,并在 Cudnn 加速计算库的基础上搭建 PyTorch 框架,然后进行加速计算。实验所用显卡为 NVIDIA GEForce GTX3090(24 G),内存为 32.0 GB,CPU 为 Intel(R) Core(TM) i7-6950X CPU @3.00 GHZ。迭代次数 Epoch 为 100,优化器选择 Adam,优化参数选择默认。迭代学习率为 0.000 1,权值衰减率为 1e-5,批大小为 80。无预训练和其他前置任务,每个 Epoch 后进行一次训练和一次测试,测试时不更新参数。将结果最好的模型参数保存,最后在验证集上进行验证。

本文将 VGG19[10]、ResNet18[6]、MobileNetV2[5]、MobileNetV3[3]与改进网络进行对比实验。图 8 展示了 5 种网络使用同一数据集训练的准确率曲线,可以看出本文所提方法的准确率最高。

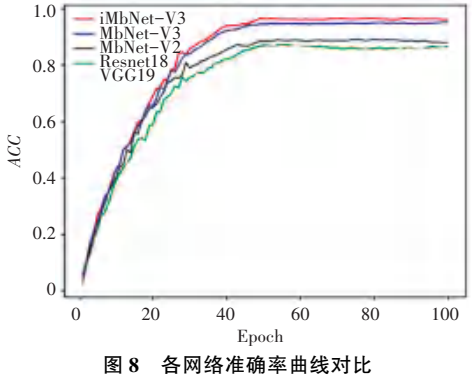


图 8 各网络准确率曲线对比

Fig. 8 Comparison of accuracy curves of all networks

表 3 显示了各网络模型的准确率和参数量对比结果,其中本文所提方法不但有最好的识别准确率,而且没有因为进行 3 755 类汉字的分类而增加过多的参数量。

表 3 各网络模型的对比结果

Tab. 3 Comparison results of various network models

网络模型	准确率 (ACC)	参数量
VGG19[10]	0.871 9	20.44 M
ResNet18[6]	0.873 2	33.16 M
MobileNetV2[5]	0.911 5	3.40 M
MobileNetV3[3]	0.960 0	5.40 M
Improved	0.966 8	5.86 M

4 结束语

本文针对目前手写识别网络训练时间长、高资源消耗的问题,提出了一种基于 MobileNetV3 的脱机手写汉字识别网络模型,在不降低识别率的基础上减少计算机资源消耗,加快训练速度。本文的主要改进工作:

(1)使用多尺度卷积核获取不同大小的感受野,丰富特征信息。

(2)采用多尺度通道注意力特征融合将多分支网络提取的特征进行全局和局部的特征融合,以提高网络性能。实验结果表明,本文提出的的改进网络获得了更好的识别结果。

参考文献

[1] 邓杰荣,梁森,曹昕妍,等. 基于深度学习的汉字识别方法研究综述[J]. 微纳电子与智能制造,2020,2(3):73-81.

[2] YIN F, WANG Q F, ZHANG X Y, et al. ICDAR 2013 Chinese handwriting recognition competition[C]//2013 12th international conference on document analysis and recognition. IEEE, 2013: 1464-1470.

[3] HOWARD A , SANDLER M, CHEN B, et al. Searching for MobileNetV3[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2020.

[4] HOWARD A G, ZHU M, CHEN B, et al.Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.

[5] SANDLER M, HOWARD A, ZHU M, et al. MobileNetV2: Inverted Residuals and Linear Bottlenecks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[6] Deep Residual Learning for Image Recognition[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016.

[7] JIE H, LI S, GANG S, et al. Squeeze-and-Excitation Networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, (99).

[8] DAI Y, GIESEKE F, OEHMCKE S, et al. Attentional feature fusion[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021: 3560-3569.

[9] LIU C L, YIN F, WANG D H, et al. CASIA online and offline Chinese handwriting databases[C]//2011 International Conference on Document Analysis and Recognition. IEEE, 2011: 37-41.

[10] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.