文章编号:1671-4598(2019)06-0022-04 DOI:10.16526/j.cnki.11—4762/tp.2019.06.006 中图分类号:TP3 文献标识码:A

基于轻量级卷积神经网络的实时缺陷检测方法研究

姚明海，杨 圳
（浙江工业大学 信息工程学院，杭州 310023）

摘要：应用机器视觉实现磁片表面缺陷的自动检测可以提高生产效率、降低生产成本；深度卷积神经网络具有高精度的分类性能，尤其在图像识别方面有显著的优点；但是目前提出的深度神经网络模型，由于参数量和计算量的巨大，在工业生产流水线上不能满足实时检测的需求；针对这个问题，基于深度可分离卷积和通道混洗，提出了一种轻量级高效低延时的卷积神经网络架构 MagnetNets；为了评估 MagnetNets 网络模型的性能，将 MagnetNets 网络模型与 MobileNets、ShuffleNet、Xception、MobileNetV2 在公开数据集 ImageNet 中做了对比实验；然后将 MagnetNets 网络模型应用在磁片缺陷检测系统中进行缺陷检测；实验结果表明，提出的网络架构显著地减少参数数量，具有良好的性能；同时在磁片缺陷检测系统中减少了延时，提高检测速度，缺陷检测识别率达到了 97.3%。

关键词：卷积神经网络；深度可分离卷积；通道混洗；缺陷检测

Research on Real—time Defect Detection Method Based on Lightweight Convolutional Neural Network

Yao Minghai，Yang Zhen
（College of Information Engineering，Zhejiang University of Technology，Hangzhou 310023，China）

Abstract：The application of machine vision to the automatic detection of surface defects on magnetic sheets can increase production efficiency and reduce production costs. Deep convolutional neural networks have high—precision classification performance，especially in image recognition. However，the deep neural network model proposed so far cannot meet the requirements of real—time detection in the industrial production line due to the huge amount of parameters and computation. To solve this problem，based on deep separable convolution and channel shuffling，we proposed a lightweight，high—efficiency and low—latency convolutional neural network architecture called MagnetNets. In order to evaluate the performance of the MagnetNets network model，we compared it with MobileNets，ShuffleNet，Xception，and MobileNetV2 in the public dataset ImageNet. And then the MagnetNets network model is applied to the defect detection system for magnetic defect detection. The experimental results show that the proposed network architecture significantly reduces the number of parameters and has good performance，At the same time，the delay is reduced and the detection speed is improved in the disk defect detection system and the defect detection recognition rate reaches 97.3%.

Keywords：convolutional neural network；deep separable convolution；channel shuffling；defect detection

0 引言

在传统的工业生产流水线上，对磁片的缺陷检测主要是以人工检测为主。随着智能信息化技术的发展，目前缺陷的自动检测方法主要分为两类。一种是使用传统的模式识别对表面缺陷进行检测。Yang 等[1]提出了一种利用平稳小波变换的新方法，用于在图像中自动检测各种光条件下的低对比度缺陷。Xie 等[2]提出了一种基于剪切变换的磁瓦图像缺陷提取方法。蒋红海等[3]将轮廓长度、相似度等作为特征向量，将支持向量机与凸凹缺陷相结合进分类检测。由于磁片表面缺陷不明显，纹理复杂和对比度低等难点，使用传统的模式识别方法对磁片表面缺陷进行检测存在通用性低和适应性不强等缺点。另一种则是基于卷积神经网路的方法来对缺陷图像进行识别和分类的。自 Krizhevsky 等[4]提出的 AlexNet 网络在 ImageNet 图像分类任务中取得了最好的成绩，开启了卷积神经网络的新纪元。为了进一步改善网络的性能，提高网络模型的分类检测的精度，使网络模型在实际中有更广泛的应用。研究者从网络的结构和应用等方面提出了 VGG[5]、GoogLeNet[6]、ResNet[7]、Xception[8]等一系列性能优良的网络。

在这些性能优良的卷积神经网络结构中，网络的深度越深和每层特征面数量越多，网络能够表示的特征空间也就越大，网络学习能力也越强[9]。虽然通过增加网络的深度和每层网络的特征面数量，使网络的性能得到了大幅度的提升。但是卷积神经网络的模型会随着网络深度和特征面数量变得越来越复杂，网络中的参数会大大增加，网络模型的计算量也会增加。因此会导致网络的实时性检测效率变低。

影响网络的实时分类检测效率主要是由于模型的存储大小和模型进行预测时的延时而引起的。对于模型的存储

收稿日期：2018-11-12； **修回日期**：2018-12-07。
基金项目：国家自然科学基金项目(61871350)。
作者简介：姚明海(1963-)，男，浙江省嘉善人，教授，博士生导师，主要从事模式识别、图像识别方向的研究。

而言，当网络的层数达到一定深度时，网络中需要保存的权值参数时巨大的，而保存大量的权值参数对设备的内存要求很高[10]。深度卷积神经网络模型当中会存在大量的参数冗余。这种冗余会对计算资源和存储资源造成巨大的浪费。因此减少网络模型的参数数量以及降低计算的复杂度是减少延时的核心所在。网络裁剪是通过寻找一种有效的评判手段来判断参数的重要性，将不重要的连接或者滤波器进行裁剪来减少网络的冗余。Song 等人[11]提出了一种几乎无损的网络裁剪压缩方法。Li 等人[12]提出了基于量级的裁剪方式，通过权值大小来作为滤波器的评价指标。核的稀疏化是在训练的过程中对权重的更新加以正则项进行引导，使其更加稀疏。Wen 等人[13]提出了一种能够学习一个稀疏结构的学习方式来降低计算消耗。虽然通过网络裁剪和核的稀疏化方法减少了网络参数，但是有些权重很小的参数会对模型的精度产生影响。而在磁片的实时缺陷检测系统中，不仅需要考虑延时性问题，还需要考虑检测精度的问题。因此设计出一种轻量级高效低延时的卷积神经网络架构 MagnetNets，使整个磁片缺陷检测系统具有高精度低延时的特点。

1　MagnetNets 网络架构

基于深度可分离卷积和通道混洗，本文设计了一种轻量级的卷积神经网络架构。在本部分首先介绍深度可分离卷积，然后介绍通道混洗，接着介绍由深度可分离卷积和通道混洗搭建而成的 MagnetNets 模块，最后介绍本论文提出的轻量级卷积神经网络的架构。

1.1　深度可分离卷积

深度可分离卷积是将标准的卷积方式因式分解为深度卷积和逐点卷积。深度卷积是将输入的特征图谱逐通道进行卷积，一个卷积核负责一个通道。通过深度卷积得到的每一个特征图谱，不能够包含输入特征图谱的所有信息。因此采用逐点卷积的方式将深度卷积输出的特征图谱再次进行多通道卷积，使信息能够尽可能的保留下来。通过深度可分离卷积可以在保证信息流通顺畅的情况下，减少模型的参数数量。深度可分离卷积与 Xception 中提出的 Inception 模块的极端形式是一致的。图 1 显示了如何将标准的卷积结构变成深度可分离卷积。

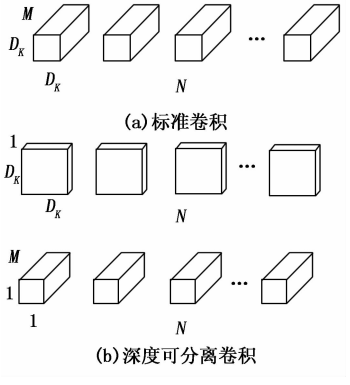

图 1　标准卷积和深度可分离卷积

1.2　通道混洗

通道混洗操作是在组卷积的操作上实施的。分组卷积的概念最早是在 AlexNet 中引入的。它通过将输入的多个特征图谱分成多个组数，然后对每个组分别进行卷积，随着将输入图谱的组数增加，所需要的网络参数量会大大减少，模型的计算速率也会随之增加。但是当多个分组卷积堆叠起来的时候，每个输出通道只能从有限输入通道获得信息，即一个组的输出只和这个组的输入有关，限制了模型的表达能力。在组卷积的基础上通过通道混洗操作，将输出的每一个组的特征图谱重新分配到每个组中，使每个组的输出都有来自其他组的上一层输入。图 2 显示了组卷积和组卷积与通道混洗的联合操作。

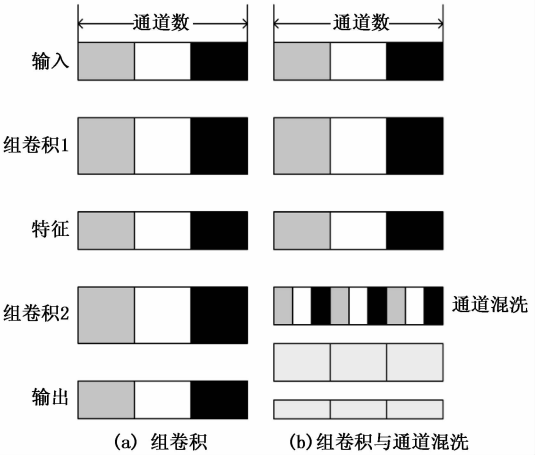

图 2　组卷积和组卷积与通道混洗

1.3　MagnetNets 模块

本论文所提出的网络结构主要受到 MobileNetV2[14] 网络模型的启发。通过组卷积和通道混洗替换 1x1 的逐点卷积，然后结合深度可分离反转残差卷积模块（如图 3 所示），构建了一种新的卷积神经网络的基本模块，我们将这种新提出的基本模块命名为 Magnet，如图 4 所示。深度可分离的反转残差卷积模块是将深度可分离卷积和残差网络的跳远连接结合起来所形成的一个网络模块。Magnet 模块主要使用了 3 个重要策略：1）延续 MobileNets[15] 中用大量 1x1 的卷积模块，同时采用组卷积和通道混洗相结合的方式取代 1x1 的卷积模块。由于 1x1 的卷积模块的参数数量只有 3x3 卷积模块参数数量的九分之一，同时将 1x1 的卷积进行分组，可以使参数数量随着组数的增加而减少，这种改进可以在保证精确度不降低的情况下减少网络模型的参数数量；2）反转残差。先通过一个 1x1 的卷积层把特征图谱的通道数扩张，然后在深度可分离卷积后再将通道数压缩回去。直接采用残差块中的跳远连接先进行压缩，会导致深度可分离卷积层提取到的特征数量减少，影响模型的精度。因此采用反转残差先将通道数扩张后再压缩；3）线性激活。经过反转残差以后，最后输出的特征图谱需要压缩。由于 Relu 函数的特性，对于负的输入，输出全为零。当对压缩后的特征图谱采用非线性的 Relu 函数激活，会损失已有

的特征，降低模型的表达。因而，采用线性函数进行激活。

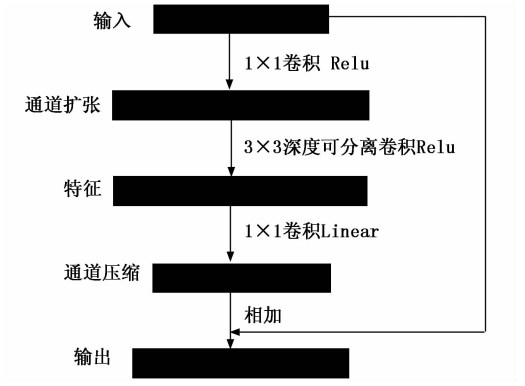

图 3 深度可分离反转残差卷积

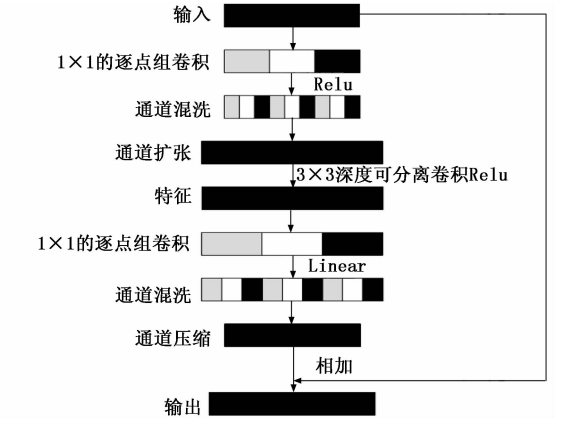

图 4 Magnet 模块

MobileNetV2 网络模型中采用大量的 1x1 的逐点卷积方法取代 3x3 的标准卷积，而在该模型中，1x1 的卷积方式占据大部分的计算量。本论文提出的方法通过组卷积和通道混洗替换 1x1 的逐点卷积，然后结合深度可分离反转残差卷积模块构成 Magnet 模块。它比使用深度可分离反转卷积模块构成的 MobileNetV2 更小，但是性能更加优越。

MagnetNet 模块将 $D_K \times D_K \times M$ 作为深度可分离卷积核的尺寸大小，一个 $D_I \times D_I \times S$ 大小的特征图作为输入，一个 $D_F \times D_F \times N$ 大小的特征图作为输出。D_K 是可分离卷积核的特征图谱大小，D_I 是输入特征图的图像尺寸，D_F 是输出特征图的图像尺寸，M 是深度可分离卷积核的个数，S 是输入特征图的通道数量，N 是输出特征图的通道数量。设定 L 为通道扩张的数量，G 为组卷积分的组数，中间输出的特征图和输入特征图的尺寸大小一样。

则 MobileNetV2 模块的计算量为：

$$C_{\text{MobileV2}} = 1 \times 1 \times S \times D_I \times D_I \times L + D_K \times D_K \times M \times D_I \times D_I + 1 \times 1 \times M \times D_F \times D_F \times N \tag{1}$$

MobileNetV2 模块的参数量为：

$$N_{\text{MobileV2}} = 1 \times 1 \times S \times L + D_K \times D_K \times M + 1 \times 1 \times M \times N \tag{2}$$

而 Magnet 模块的计算量为：

$$C_{\text{Magnet}} = \frac{1 \times 1 \times S}{G} \times \frac{D_I \times D_I}{G} \times L + D_K \times D_K \times M \times D_I \times D_I + \frac{1 \times 1 \times M}{G} \times \frac{D_F \times D_F}{G} \times N \tag{3}$$

Magnet 模块的参数量为：

$$N_{\text{Magnet}} = \frac{1 \times 1 \times S}{G} \times L + D_K \times D_K \times M + \frac{1 \times 1 \times M}{G} \times N \tag{4}$$

通过组卷积和通道混洗替换逐点卷积再和深度可分离反转残差卷积构建而成的 Magnet 模块的计算量与 MobileNetV2 的计算量比值为：

$$R_1 = \frac{C_{\text{Magnet}}}{C_{\text{MobileV2}}} = 1 - \frac{(G^2 - 1)(S \times L \times {D_I}^2 + M \times N \times {D_F}^2)}{G^2(S \times L \times {D_I}^2 + {D_K}^2 \times {D_I}^2 \times M + M \times N \times {D_F}^2)} \tag{5}$$

Magnet 模块的参数量与 MobileNetV2 的参数量比值为：

$$R_2 = \frac{N_{\text{Magnet}}}{N_{\text{MobileV2}}} = 1 - \frac{(G - 1)(S \times L + M \times N)}{G(S \times L + {D_K}^2 \times M + M \times N)} \tag{6}$$

当卷积层的输入以及中间层的输入特征图和输出特征图的尺寸大小一样时，并在经过通道混洗保证通道间信息流通的情况下，当组数 G 设置较大的时候，Magnet 模块的参数量和计算量是远小于 MobileNetV2 模块的。

1.4 MagnetNets 网络结构

MagnetNets 的总体结构是一系列的 Magnet 模块和一些普通的卷积层组合堆叠起来的。如图 5 显示了 MagnetNets 网络的整体架构以及该架构中的各种模块类型。MagnetNets 由一个输入尺寸为标准卷积层开始的，再与一些 Magnet 模块堆叠，然后通过一个不带任何参数平均池化。在最后一层卷积层的输出中通常会以一个全连接层作为输入，但是全连接层中的参数数量非常庞大，可能会导致过拟合，降低模型的表达效果。因此采用全局平均池化[16]来代替全连接层，同时搭载一个增加网络泛化能力的 Dropout 层，避免过拟合的发生。最后将 Dropout 层中的输出作为 Softmax 分类器的输入对图片进行分类和检测。

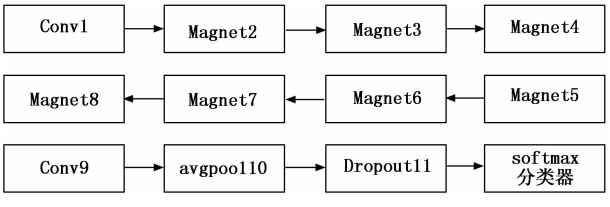

图 5 Magnets 网络架构

2 实验结果与分析

为了评估出本文提出的 MagnetNets 网络模型的性能，我们在 ImageNet 公用数据库的图像分类数据集上对 MobileNets、ShuffleNet[17]、Xception、MobileNetV2、Mag-

netNets 网络模型进行对比实验。然后将 MagnetNets 网络模型应用于磁片缺陷检测系统中对缺陷实时识别。实验结果表明 MagnetNets 网络模型具有很好的检测精度和泛化能力，并且在模型大小上面更加的轻量化。同时使整个磁片缺陷检测系统具有高精度低延时的特点，提高了检测的效率和精度。

2.1　模型评估

ImageNet 是一个计算机视觉系统识别项目，是目前世界上图像识别最大的数据库。本文在 ImageNet 的分类图片数据集上对 MobileNets、ShuffleNet、Xception、MobileNetV2、MagnetNets 网络模型进行训练和测试。使用 TensorFlow 框架对模型进行训练，并采用 Xavier 来初始化网络模型的参数，将 AdamOptimizer 作为优化器的优化算法。同时在每一层之后使用批量标准化（batch normalization），批处理大小（batch size）为 96，权重衰减（weight decay）为 0.00004。初始学习率（learning rate）设置为 0.045，学习率的衰减率（decay rate）为每代的 0.98。实验结果及与其他网络模型的对比的数据如表 1 所示。实验结果表明，在精确度方面，MagnetNets 网络模型能够达到 MobileNetV2 等网络模型的准确率，在模型的大小上面，MagnetNets 网络模型比其他的几种网络更加轻量化。

表 1　各网络模型的对比结果

CNN 架构	精确度/%	模型大小/MB
MobileNets	89.6	5.7
ShuffleNet	91.7	4.5
Xception	93.8	6.8
MobileNetV2	93.2	6.3
MagnetNets	93.5	2.9

2.2　缺陷检测

磁片缺陷检测系统主要由传送模块、视觉模块、分类检测模块、分拣模块 4 个部分组成，如图 6 所示。传送模块主要是在传送带上对生产打磨出来的磁片进行传输。当传送带上的磁片达到指定的视觉模块区域时，视觉模块对传送带上的磁片图像进行实时的采集。然后将采集到的图片送入由 MagnetNets 网络模型组成的分类检测模块中，来判断磁片是否具有缺陷，同时判断出缺陷是属于哪一种类型的。最后将缺陷检测的结果以信号的形式发送给分拣模块，分拣模块中的机械手根据信号将不同类型的缺陷分拣放置到不同的地方，从而完成整个磁片缺陷检测系统对磁片的分类。

图 6　磁片缺陷检测系统

磁片缺陷检测系统中的分类检测模块主要是使用磁片数据集对 MagnetNets 网络模型进行训练，网络模型训练时的 loss 曲线和训练精度如图 7 所示。其中磁片主要分为“正品”、“掉皮”、“开裂”、“缺角”四类磁片，如图 8 所示。

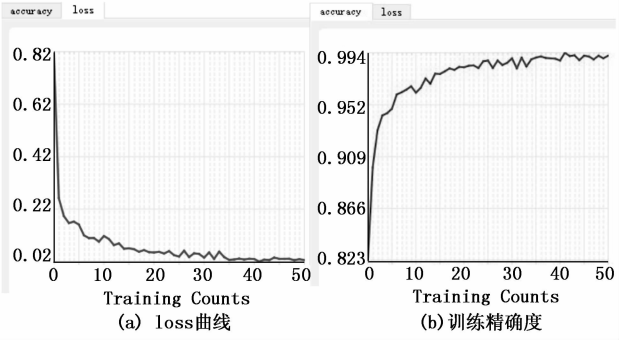

图 7　训练时的 loss 曲线和训练精确度

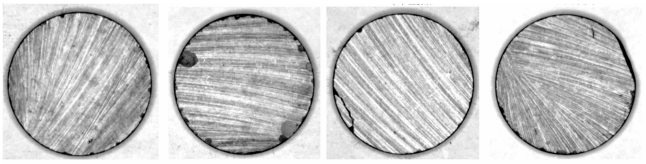

图 8　磁片示意图

在磁片的 4 种类型数据集中，“正品”和“掉皮”的数据集占据大部分，而“开裂”和“缺角”的数据集较少。为了防止在训练过程中，由于数据集不均衡而导致模型的性能下降，因此采用旋转和添加椒盐噪声的方式来扩展“开裂”和“缺角”的磁片数据集。

通过系统的实时性检测与评估，当磁片缺陷检测系统中加入 MagnetNets 网络模型后，检测过程中可以达到 30 ms/个，同时磁片的检测精度达到了 97.3%，提高了磁片缺陷检测系统的检测效率，节约了人力成本。

3　结束语

本文提出了一种基于轻量级卷积神经网络的实时缺陷检测方法。首先通过通道混洗和深度可分离卷积搭建出组成 MagnetNets 网络模型的 MagnetNet 模块，然后在公开数据集 ImageNet 上对 MagnetNets 网络模型进行性能和模型大小的评估。接着通过旋转和添加椒盐噪声的方式来扩展不平衡的磁片数据集，并放入所提出的轻量级卷积神经网络模型中进行训练，最后将训练好的轻量级卷积神经网络加入到磁片缺陷检测系统的分类检测模块中完成分类。通过系统的实时性分析，在提高检测速度的情况下仍然能够达到高精度的分类。但是由于不同形状的磁片需要重新对模型进行训练，该检测系统还不能针对于可变的磁片形状进行检测。因此下一步的研究方向放在具有自我发育机制的缺陷检测上面。

参考文献：

[1] Yang C, Liu P, Yin G, et al. Defect detection in magnetic tile images based on stationary wavelet transform. [J] NDT E Int. 2016, 83: 78-87.

（下转第 40 页）