

基于轻量级 CNN 与主动学习的工件疵病识别方法研究^①姚明海^② 杨 圳

(浙江工业大学信息工程学院 杭州 310023)

摘 要 应用图像识别技术实现工件疵病自动检测可以提高效率,降低人工成本。卷积神经网络(CNN)具有很强的特征提取能力,广泛应用于图像识别等领域。但是,已提出的网络模型普遍存在参数量和计算量巨大,以及海量工件数据集中、人工标注成本高等缺点,很难应用于工件疵病的实时自动识别。本文提出了一种基于轻量级 CNN 与主动学习的工件疵病识别方法,该方法通过深度可分离卷积和反转残差卷积构建一种轻量级卷积神经网络,并在识别过程中采用主动学习方法不断添加标注样本。实验表明,提出的方法识别精度达到 98.3%,并且能节省 18.8% 的人力标注成本。

关键词 卷积神经网络(CNN); 主动学习; 轻量级; 疵病识别

0 引 言

在制造业中,工件上的疵病主要通过人工检测进行识别。由于受个人精神状态和主观因素的影响,人工检测识别的效率和质量难以确保,传统的人工识别方法已无法满足生产需求。与传统的检测方法相比,基于卷积神经网络(convolutional neural network, CNN)的机器视觉是一种无接触、无损伤的自动检测技术,是实现设备自动化、智能化的有效手段,具有安全可靠、光谱响应范围宽、可在恶劣环境下长时间工作和生产效率高等突出优点^[1]。近年来,随着深度学习的发展,基于卷积神经网络的检测与识别模型已成功应用于人脸识别^[2,3]、图像检索^[4]、语音识别^[5,6]和自然语言处理^[7]等研究领域,在工业检测识别方面^[8]也得到广泛应用。

随着对网络性能的要求越来越高,使网络模型在实际中有更广泛的应用,许多学者通过对网络的结构进行改进提出了 VGG^[9]、GoogLeNet^[10]、ResNet^[11]等一系列性能优良的网络。这些网络模型通

过扩展网络的深度,增加每层网络的特征面,其识别精度得到了很大的提高^[12]。随着卷积神经网络的模型越来越复杂,网络中的参数大大增加,计算量也随之增加。在实际的工件疵病识别过程中,较大的网络虽然在识别精度上有了很大提升,但是降低了工件疵病实时检测的效率,同时网络的存储大小和网络参数的计算对硬件设备的要求很高,不具有普遍性。

训练卷积神经网络模型中巨大的网络参数需要大量数据集的支持。在使用传统的监督学习方法做分类时,通常是训练样本规模越大分类的效果越好。当训练样本的规模过于庞大时,训练所需的时间成本也会大幅增加。实际应用中发现,大量的数据集可以很容易地获取,但是标注样本的获取是比较困难的。而且数据标注的人工成本高,有些数据标注可能还需要领域内的专家才能完成,不可能实现快速或者低成本的数据标注。主动学习则可以通过标记大多数信息实例来以最低成本训练准确的预测模型^[13]。

为了解决高精度大网络模型实时检测效率低和

① 国家自然科学基金(61871350)资助项目。

② 男,1963年生,博士,教授;研究方向:模式识别和图像识别;联系人,E-mail: ymh@zjut.edu.cn
(收稿日期:2019-04-17)

获取标注数据集困难的问题,本文提出了一种基于轻量级卷积神经网络与主动学习的工件疵病识别方法。该方法首先在不损失精度的前提下对较大的网络模型进行优化,使网络模型更加轻量化。然后针对实际生产中大量数据获取比较容易但是人工标注成本较高的问题,采用一种改进的主动学习方法来解决。最后将该轻量级卷积神经网络和主动学习相结合应用在金手指工件上进行疵病识别。实验结果表明,在网络模型精度不损失的情况下,该方法减少了计算量和参数量,提高了检测效率,而且通过主动学习的方法节约了大量的人力标注成本。

1 模型压缩的可行性分析

卷积神经网络随着深度的增加,模型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。近年来,很多研究都集中在通过模型压缩来减少计算量和参数。目前对卷积神经网络进行压缩主要有 3 类:网络裁剪、核的稀疏化和轻量级的网络模型。

网络裁剪是针对具有冗余性的深度卷积神经网络模型,通过判断网络参数的重要性,更改网络的结构,将网络中的冗余部分裁剪掉,以达到压缩模型的目的。He 等人^[14]通过将网络模型的特征图谱裁剪前后的误差最小化,尽可能减少卷积核的通道数目,在影响指标最小的情况下,最大化地提升网络性能。核的稀疏化是在训练时对网络权重的更新加以正则项引导,使大部分不重要的权值为零。Guo 等人^[15]提出了一种动态的网络模型,在训练过程中通过一个可学习的掩码限制权重的更新,达到稀疏网络的目的。轻量级 CNN 是在模型训练之前就对网络结构进行优化,达到减少参数、降低计算量的目的。SqueezeNet^[16]利用瓶颈方法设计了一种达到 AlexNet^[17]相同精度水平的轻量级网络结构。MobileNets^[18]通过深度可分离卷积的方式来代替传统的卷积方式。这种卷积结构可以降低网络的复杂度,而不会造成明显的精度损失。ShuffleNet^[19]通过对通道数的重新排列,解决了进行卷积时信息流通不畅的问题。MobileNetV2^[20]通过引入残差网络中的

跳跃连接达到了在更小的模型参数下得到更高的分类精度和更快的实时性效果。

2 轻量级 CNN 模型

网络裁剪、核的稀疏化都是通过后处理的方式减小大型 CNN 模型的尺寸。然而,这种后处理的方式需要对专门的超参数进行优化。此外,随着网络架构的改变,模型需要额外的微调。本文基于深度可分离卷积和反转残差卷积设计的轻量级 CNN 是从模型的结构方面进行改进,所设计出来的网络模块可以在不损失精度的情况下降低模型的复杂度。

2.1 深度可分离卷积

在卷积神经网络模型中通常会采用标准卷积的方式来构建卷积层,如图 1 所示。标准的卷积方式会先在区域上提取特征,然后再考虑通道数,在减少网络参数的计算量方面有很大的局限性。假设输入特征图的大小为 $I \times I \times M$,卷积核为 $D \times D \times M$,输出特征图大小为 $O \times O \times N$,则标准卷积所需要的计算量如式(1)所示。

$$C_1 = D \times D \times M \times N \times O \times O \quad (1)$$

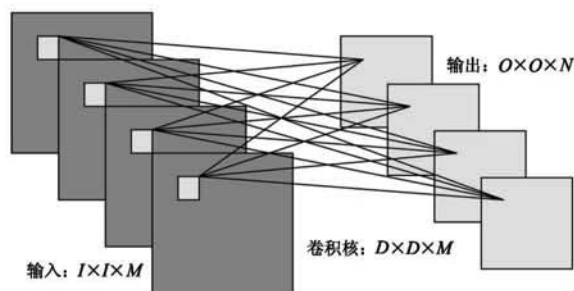


图 1 标准卷积

本文采用深度可分离卷积来构造卷积层,深度可分离卷积会同时考虑通道和区域的变化,在通道和区域上可分解为深度卷积和逐点卷积,如图 2 所示。深度卷积是将输入的特征图谱逐通道进行卷积,1 个卷积核负责 1 个通道;逐点卷积是将深度卷积输出的特征图谱进行多通道卷积。当逐点卷积中采用的卷积核为 $1 \times 1 \times N$ 时,深度可分离卷积所需要的计算量如式(2)所示。

$$C_2 = D \times D \times M \times O \times O + M \times N \times O \times O \quad (2)$$

在相同的输入和输出情况下,标准卷积和深度可分离卷积的计算量比值 R 如式(3)所示。通过计算量比值 R 可以得出在深度可分离卷积实现区域和通道分离的情况下,可以减少模型的计算量。

$$R = \frac{C_2}{C_1} = \frac{D \times D \times M \times O \times O + M \times N \times O \times O}{D \times D \times M \times N \times O \times O} = \frac{1}{N} + \frac{1}{D^2} \quad (3)$$

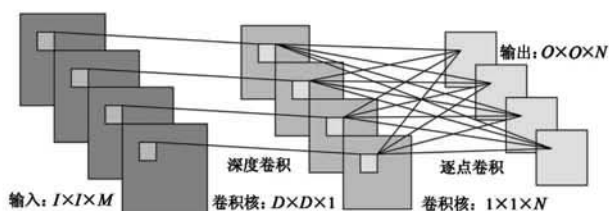


图2 深度可分离卷积

深度可分离卷积中大于 3×3 的卷积核都可以分解为连续的 3×3 的卷积层来降低网络模型的参数数量。因此,深度可分离卷积中通常采用的是 3×3 的卷积核。本文通过对 3×3 的卷积核进一步优化,将其在空间上分解为 2 个形状为 3×1 和 1×3 的连续卷积核,与原始的 3×3 的卷积核相比较计算量减少了 33%。然而将 3×3 卷积核在空间上分解为 2 个连续的卷积核会导致输入层向下传递的特征有所缺失,降低模型识别的精确度。同时大多数的模型在深度可分离卷积层都会采用步长为 2 的卷积操作取代池化操作,这会使特征的提取出现混叠现象。为了解决这些问题,通过在分解的 2 个非对称卷积层之间增加一种汇集策略取代采用步长为 2 的卷积操作,对输入信息进行提取和汇集。

2.2 反转残差卷积

当搭建层次较深的神经网络模型的时候,为了解决梯度消失和梯度爆炸的问题,通常会采用残差模块来搭建。残差模块中对信息的提取是对通道先进行压缩,然后提取特征,最后再对通道进行扩张。当采用 2.1 节中的深度可分离卷积进行卷积操作的时候,为了减少参数数量和计算量,会在深度卷积后面加上多个逐点卷积来压缩模型大小,使计算量都集

中在逐点卷积上。当深度可分离卷积中直接采用残差结构时,由于逐点卷积层已经压缩过网络了,再采用残差结构中先压缩后扩张的方式在压缩部分进一步压缩网络时,会导致最终提取到的特征信息减少。

本文提出一种反转残差结构,如图 3 所示。在经过逐点卷积层提取信息以后,通过增加 1×1 卷积核的个数与前一层进行卷积来扩张输出通道的数目,对扩张后的输出采用 3×3 的卷积核提取信息,最后通过减少 1×1 的卷积核个数将扩张后的通道数目压缩回去,同时保留残差结构中的跳跃连接,使前层的信息也能够直接向后层传递。

在残差模块中通常采用 Relu 函数来激活。由于 Relu 函数的特性,对于输入为负时,输出全为 0。在经过反转残差卷积模块以后,最后输出的特征图谱经过了压缩。当对压缩后的特征图谱采用非线性的 Relu 函数激活时,会损失已有的特征,降低模型的精确度。因此,在反转残差中采用线性函数进行激活。

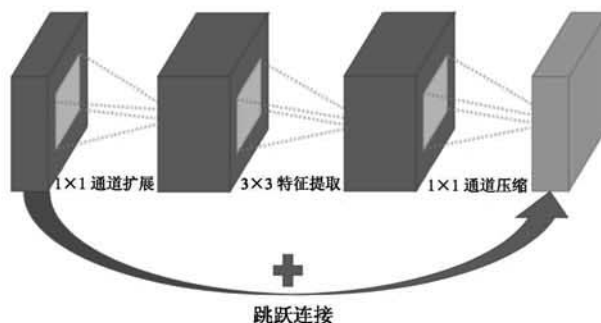


图3 反转残差卷积

2.3 网络模块结构

本文提出的轻量级卷积神经网络主要通过优化深度可分离卷积和反转残差卷积来构建一种名为空间可分离残差 (spatial separable residual, SSR) 卷积模块,如图 4 所示。在模块的初始阶段,通过第 1 次深度卷积操作后,采用反转残差中的逐点卷积扩张输出通道,再在扩张后的输出通道中将深度可分离卷积在空间上分解为 1 个 3×1 和 1 个 1×3 的深度可分离卷积,在 2 个非对称的深度可分离卷积之间采用最大池化的汇集策略来提取更多的信息,然后再采用 1×1 的逐点卷积将通道压缩,在压缩后的输出特征图谱中采用线性激活活尽可能地保留通道被压

缩后的信息。最后保留残差网络中的跳跃连接将输入信息传到神经网络的更深层。本文首先由 1 个初始的卷积层开始,然后连接 3 个空间可分离卷积模块,接着采用全局平均池化来汇集特征,最后搭载 1 个 softmax 分类器来搭建轻量级卷积神经网络 SSR-Net 模型。

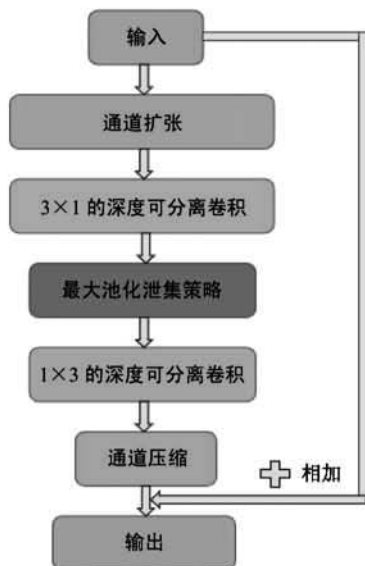


图 4 空间可分离残差模块

3 主动学习

在训练一个深度神经网络模型的时候,随着训练数据集的增加,网络模型的精度会不断提升。但是当训练数据集达到一个临界值时,精度基本不会产生变化,而这个临界值和数据样本本身有关。主动学习可以通过主动选择策略在未标注的样本集中挑选更具代表性的样本训练模型^[21,22]。主动学习可以使临界值更早地到来,提升了模型收敛的速度,同时节省了人力标注的成本。节省的人力标注成本可以通过式(4)来衡量。

$$S = \frac{N_1 - N_2}{N_1} \quad (4)$$

其中, N_1 代表随机选择所需要的样本数据集大小, N_2 代表主动学习所需要的样本数据集大小, S 代表可节省的人力标注成本。

一个主动学习算法 $A = (L, U_1, S, E, U_2)$ 可以通过 5 个部分来构建。其中 L 为学习模块; U_1 为一组已经标注好的训练数据集; S 为选择策略,用于

选取需要进行标注的样本; E 为执行者,通过某种选择策略对未标注样本进行标注; U_2 为整个未标注样本数据集。主动学习算法主要包含 2 个阶段:第 1 阶段为初始化阶段,利用已经标注好的样本集 U_1 去训练学习模块,获取网络学习模型的初代参数。第 2 阶段为循环更新阶段,执行者 E 根据选择策略 S 从未标注样本集 U_2 中选择部分样本进行标注,并加入到已标注好的训练样本集 U_1 中,重复训练网络模型,直到网络学习模型满足达到所需分类的精度为止。

主动学习中最重要就是“学习模块”和“选择策略”2 个部分。“学习模块”通过在训练集中不断增加新的标注样本,迭代训练网络模型,使网络模型达到最终的分精度。本文中选取第 2.3 节中提出的 SSR-Net 网络来作为主动学习的学习模块。“选择策略”是通过某种选择策略选择出更具代表性的样本,重新标注样本后,不断扩充标注样本集,加快学习模型的收敛速度。在常用的选择策略中,会选择“信息熵”这种单一的选择策略。信息熵是图像所包含信息量的一种度量,其值越大,则说明该图像包含的信息越多,不确定性越强,越难反映出图像的主要特征;反之则表明包含信息量越少,更容易反映出图像的主要特征。信息熵值的大小可以根据式(5)来计算。

$$H(y | x_i, \omega) = - \sum_{y \in N} P(y | x_i, \omega) \log P(y | x_i, \omega) \quad (5)$$

其中, y 代表分类的类别, $P(y | x_i, \omega)$ 表示在网络参数为 ω 时样本 x_i 属于 y 的概率, N 为所有可能分类类别的集合。

本文中采用“信息熵”和“多样性”2 种选择策略相结合来选取样本,使加入到标注样本集中的样本更加均匀。多样性策略就是通过比较未标注样本和已标注样本之间的相似度,将与已标注数据相似的样本作为简单样本,而那些不相似样本需要人工重新进行标注,提高整个数据集的多样性和准确性^[23]。

基于轻量级 CNN 和主动学习的算法流程图如图 5 所示,具体流程如下。

(1) 在所有样本中随机选择 m 个训练样本 U_1

$= (x_1, x_2, \dots, x_m)$ 进行人工标注来作为训练集。

(2) 通过 $U_1 = (x_1, x_2, \dots, x_m)$ 来训练第 2.3 节提出的 SSR-Net 网络模型, 得到网络参数 ω 。

(3) 在所有未标注的 n 个样本 $U_2 = (x_1, x_2, \dots, x_n)$ 中用式(5)计算出每个样本的信息熵 $H(y|x_i, \omega)$ 。

(4) 用式(6)计算出每个未标注样本 x_i 相对于标注样本 $U_1 = (x_1, x_2, \dots, x_m)$ 的相似度 $R(i, j)$ 。

$$R(i, j) = \sum_{y=1}^N \sum_{j=1}^n (P_y^i - P_y^j) \log \frac{P_y^i}{P_y^j} \quad (6)$$

P_y^i 表示已标注样本 x_i 在类别 y 中的概率, P_y^j 表示未标注样本 x_j 在类别 y 中的概率。

(5) 结合信息熵和相似度 2 种选择策略, 用式(7)计算 2 种策略的比值 $D(i, j)$, 对 $D(i, j)$ 进行降序排列, 根据 $D(i, j)$ 的大小在未标注的样本集中选择 k 个样本进行人工标注。

$$D(i, j) = \frac{H(y|x_i, \omega)}{R(i, j)} \quad (7)$$

(6) 将重新标注的 k 个样本加入 $U_1 = (x_1, x_2, \dots, x_m)$ 中, 组成新的已标注样本集, 重复执行步骤(2)、(3)、(4)、(5)、(6), 直到分类精度达到临界值为止。

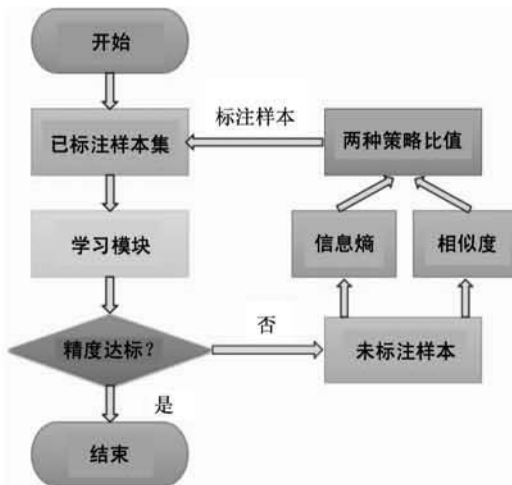


图5 算法流程图

4 实验与结果分析

本文采用深度可分离卷积和反转残差卷积搭建轻量级卷积神经网络, 并在训练网络的过程中结合主动学习来标注样本。为了验证模型的性能和通用

性, 在 Cifar-10 数据库中将该网络与轻量级网络 SqueezeNet、MobileNet、ShuffleNet 以及经典网络 AlexNet 进行比较。然后在结合主动学习标注的情况下, 将该轻量级卷积神经网络模型应用在金手指工件上, 对其进行疵病识别。

4.1 模型对比验证

本文的网络模型对比实验是在 Cifar-10 数据库中进行的。Cifar-10 数据库是由 5 万张彩色训练数据集和 1 万张彩色测试集组成的图像分类数据集。该数据集中图片的尺寸是 32×32 , 总共分为 10 个物体类别。实验中, 使用 TensorFlow 框架搭建网络模型, 采用 Adam 优化算法, 并以 Xavier 来初始化网络模型的参数。同时在每层后面使用批量标准化 (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 0.005, 学习率的衰减率 (decay rate) 为每代的 0.98。训练时的损失曲线如图 6 所示, 实验结果如表 1 所示。由表 1 可见, 本文提出的 SSR-Net 网络结构相对于其他几种网络模型不会出现精度高低与模型大小之间相互约束的情

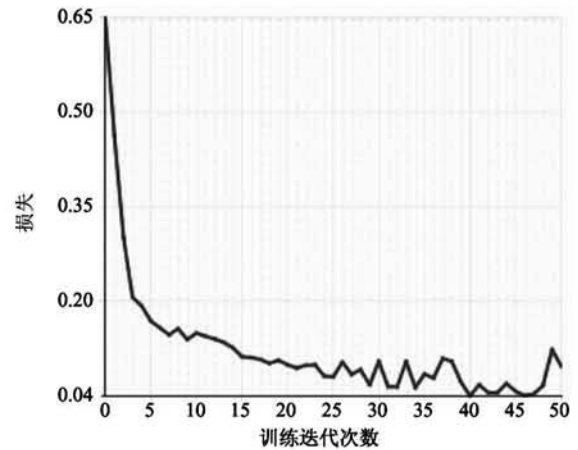


图6 损失曲线

表1 各种网络模型的识别精度和模型大小

网络模型	精确度 (%)	模型大小 (MB)
AlexNet	91.87	18.5
SqueezeNet	91.26	8.7
MobileNet	86.54	4.4
ShuffleNet	87.12	5.6
SSR-Net	94.45	3.4

况。在精确度上,该网络有更高的性能,同时在网络的模型规模上,该网络更加轻量化,更容易嵌入到移动平台上。

4.2 疵病识别

本文的实验对象为工业生产中采集到的金手指工件图像。在金手指的制造过程中,由于外部的一些因素会出现各种类型的疵病。所有的金手指数据集被分成了“正品”、“变色”、“刮痕”、“掉漆”4类,如图7所示。本文通过提出的 SSR-Net 轻量级卷积神经网络结合主动学习对金手指的疵病进行识别。首先选取 6 000 个已标注好的样本作为初始训练集放入 SSR-Net 网络中训练。当网络训练完以后,在没有标注的数据集中根据信息熵和多样性选择策略再次选择 2 000 个样本加入初始训练集中。重新对 SSR-Net 网络模型进行训练。针对再次训练好的网络,重复加入 5 次样本,训练出最终的模型参数。主动学习和随机挑选样本的对比如图8所示。采用本文的主动学习方法选择大约 13 000 个样本进行人工标定就能达到随机选择 16 000 个样本的准确率,可以使模型更快地收敛。同时,根据式(4)可以计算出节约了 18.8% 的人工标注成本。因此,该方法具有一定的工程应用意义。

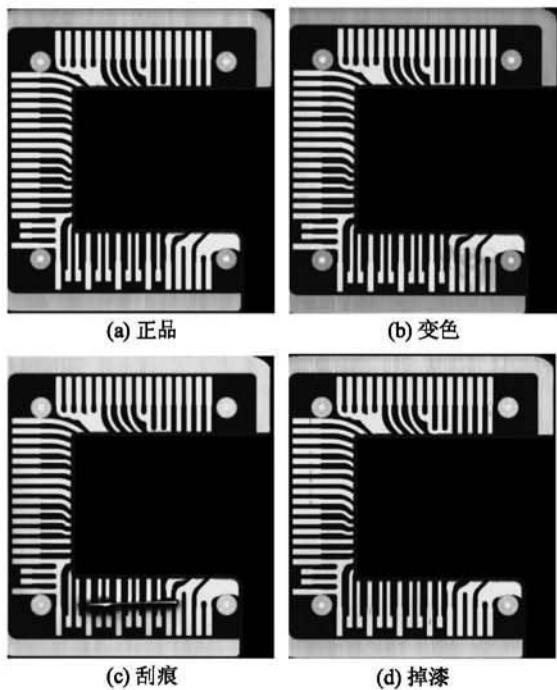


图7 4 种类型的金手指图像

对训练好的网络模型,分别选取 3 000 个样本作为验证集和测试集。各种金手指类型的识别结果如表2所示,疵病识别的准确率达到 98.3%。将 SSR-Net 网络模型嵌入到金手指工件生产流水线上的检测系统后,通过取多次在单位时间内检测到的金手指数目,然后计算平均值,得出金手指在流水线上的检测速度大约为 35 ms/个。

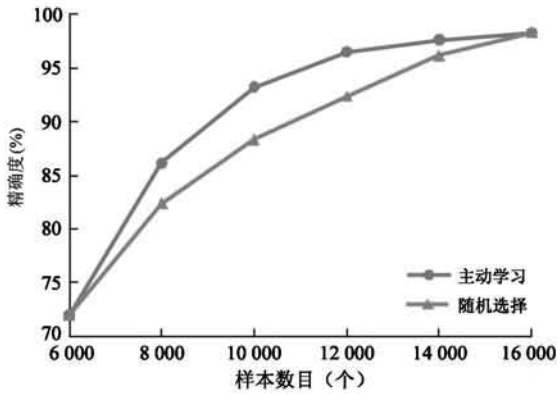


图8 随机挑选和主动学习的对比曲线

表2 各种金手指类型的识别结果

类型	待识别数目	已识别数目	识别率(%)
正品	1 478	1 478	100.0
变色	362	349	96.4
刮痕	436	427	97.9
掉漆	724	695	96.0
合计	3 000	2 949	98.3

5 结 论

本文提出了一种基于轻量级卷积神经网络和主动学习的工件疵病识别方法。通过深度可分离卷积和反转残差构建 SSR-Net 网络模型,并将 SSR-Net 与多种经典网络在 Cifar-10 数据库上进行对比验证。把金手指工件作为实验对象训练该网络模型,同时结合主动学习来标注样本。实验结果表明,本文提出的方法具有很好的泛化能力,网络模型更加轻量化,在标注样本上节约了大量的人力成本,使模型能够更早地达到最优。但是目前该方法针对不同的对象需要重新训练网络模型,下一步可以研究具有自我发育机制的轻量级疵病识别网络。

参考文献

- [1] 汤勃, 孔建益, 伍世虔. 机器视觉表面缺陷检测综述 [J]. 中国图象图形学报, 2017, 22(12):1640-1663
- [2] Zhang K, Zhang Z, Li Z, et al. Joint face detection and alignment using multitask cascaded convolutional networks [J]. *IEEE Signal Processing Letters*, 2016, 23(10): 1499-1503
- [3] Li H, Lin Z, Shen X, et al. A convolutional neural network cascade for face detection[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, USA, 2015: 1800-1804
- [4] 黎健成, 袁春, 宋友. 基于卷积神经网络的多标签图像自动标注[J]. 计算机科学, 2016, 43(7):41-45
- [5] Abdel-Hamid O, Mohamed A R, Jiang H, et al. Convolutional neural networks for speech recognition [J]. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2014, 22(10):1533-1545
- [6] Huang J T, Li J, Gong Y. An analysis of convolutional neural networks for speech recognition[C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, Australia, 2015: 4989-4993
- [7] Kim Y. Convolutional neural networks for sentence classification[EB/OL]. <https://arxiv.org>; Cornell University Library, 2014
- [8] 闫哲, 杜学丹, 曹森, 等. 基于深度学习的机器人抓取位置检测方法[J]. 高技术通讯, 2018, 28(1): 58-66
- [9] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [EB/OL]. <https://arxiv.org>; Cornell University Library, 2014
- [10] Szegedy C, Liu W, Sermanet P, et al. Going deeper with convolutions[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, USA, 2015: 1-9
- [11] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 770-778
- [12] 周飞燕, 金林鹏, 董军. 卷积神经网络研究综述[J]. 计算机学报, 2017, 40(6): 1229-1251
- [13] Fu Y, Zhu X, Li B. A survey on instance selection for active learning[J]. *Knowledge and Information Systems*, 2013, 35(2):249-283
- [14] He Y, Zhang X, Sun J. Channel pruning for accelerating very deep neural networks[C]//IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017: 1398-1406
- [15] Guo Y, Yao A, Chen Y. Dynamic network surgery for efficient DNNs[EB/OL]. <https://arxiv.org>; Cornell University Library, 2016
- [16] Iandola F N, Han S, Moskewicz M W, et al. SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size[EB/OL]. <https://arxiv.org>; Cornell University Library, 2016
- [17] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]//Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, Nevada, 2012: 1097-1105
- [18] Howard A G, Zhu M, Chen B, et al. Mobilenets: efficient convolutional neural networks for mobile vision applications[EB/OL]. <https://arxiv.org>; Cornell University Library, 2017
- [19] Zhang X, Zhou X, Lin M, et al. ShuffleNet: an extremely efficient convolutional neural network for mobile devices [EB/OL]. <https://arxiv.org>; Cornell University Library, 2017
- [20] Sandler M, Howard A, Zhu M, et al. Inverted residuals and linear bottlenecks: mobile networks for classification, detection and segmentation[EB/OL]. <https://arxiv.org>; Cornell University Library, 2018
- [21] Zhou Z, Shin J, Zhang L, et al. Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, USA, 2017: 4761-4772
- [22] Konyushkova K, Sznitman R, Fua P. Learning active learning from data[EB/OL]. <https://arxiv.org>; Cornell University Library, 2017
- [23] Chakraborty S, Balasubramanian V, Sun Q, et al. Active batch selection via convex relaxations with guaranteed solution bounds[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, 37(10):1945-1958

Research of workpiece defects recognition method based on lightweight CNN and active learning

Yao Minghai, Yang Zhen

(College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023)

Abstract

Automatic detection of workpiece defects through image recognition technology can improve efficiency and reduce labor cost. Because of its strong feature extraction ability, convolutional neural networks (CNN) are widely used in image recognition and other fields. However, the existing CNN models are difficult to be applied to automatic detection of workpiece defects in real time with the disadvantage of huge amount of parameters and computation, and the high cost of manual labeling in massive workpiece data sets. In this paper, a new approach for workpiece defect detection is proposed by combining lightweight CNN with active learning, which constructs a lightweight CNN by deep separable convolution and inverted residual convolution, and uses active learning method to add labeled samples continuously in the process of detection. Experiments show that the proposed method can achieve a high recognition accuracy of 98.3% and a labeling cost reduction of 18.8% simultaneously.

Key words: convolutional neural network (CNN), active learning, lightweight, defect recognition