# 基于对抗学习和知识蒸馏的神经网络压缩算法

刘金金[1,2,3],李清宝[1,2],李晓楠[1,2,3]
1.战略支援部队信息工程大学,郑州 450003
2.数学工程与先进计算国家重点实验室,郑州 450003
3.中原工学院 计算机学院,郑州 450007

**摘 要:**针对基于深度学习的人脸识别模型难以在嵌入式设备进行部署和实时性能差的问题,深入研究了现有的模型压缩和加速算法,提出了一种基于知识蒸馏和对抗学习的神经网络压缩算法。算法框架由三部分组成,预训练的大规模教师网络、轻量级的学生网络和辅助对抗学习的判别器。改进传统的知识蒸馏损失,增加指示函数,使学生网络只学习教师网络正确识别的分类概率;鉴于中间层特征图具有丰富的高维特征,引入对抗学习策略中的判别器,鉴别学生网络与教师网络在特征图层面的差异;为了进一步提高学生网络的泛化能力,使其能够应用于不同的机器视觉任务,在训练的后半部分教师网络和学生网络相互学习,交替更新,使学生网络能够探索自己的最优解空间。分别在CASIA WEBFACE和CelebA两个数据集上进行验证,实验结果表明知识蒸馏得到的小尺寸学生网络相较全监督训练的教师网络,识别准确率仅下降了1.5%左右。同时将本研究所提方法与面向特征图知识蒸馏算法和基于对抗学习训练的模型压缩算法进行对比,所提方法具有较高的人脸识别准确率。

**关键词:**知识蒸馏;对抗学习;互学习;模型压缩;人脸识别
**文献标志码:**A **中图分类号:**TP391 doi:10.3778/j.issn.1002-8331.2105-0295

## Neural Network Compression Algorithm Based on Adversarial Learning and Knowledge Distillation

LIU Jinjin[1,2,3], LI Qingbao[1,2], LI Xiaonan[1,2,3]
1.Information Engineering University, Zhengzhou 450003, China
2.State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou 450003, China
3.School of Computer Science, Zhongyuan University of Technology, Zhengzhou 450007, China

**Abstract:** In order to solve the problems that face recognition models based on deep learning are difficult to deploy in embedded devices and their real-time performance is poor, the existing model compression and acceleration algorithms are deeply studied, and a neural network compression algorithm based on knowledge distillation and adversarial learning is proposed. The algorithm framework consists of three parts: a large-scale teacher network for pre-training, a lightweight student network and a discriminator for adversarial learning. This paper improves the traditional knowledge distillation loss by adding an indicator function, so that the student network only learns the classification probability that the teacher network correctly identifies. Since the feature map of the middle layer has rich high-dimensional features, the discriminator in adversarial learning strategy is introduced to identify the difference between the student network and the teacher network in the feature map level. Furthermore, in order to improve the generalization ability of the student network and enable it to be applied to different machine vision tasks, the teacher network and the student network learn from each other in the latter part of the training and update alternately, so that the student network can explore its own optimal solution. The results are verified on CASIA WEBFACE and CelebA data sets respectively. The experimental results show that the recognition accuracy of the small size student network obtained by knowledge distillation is only about 1.5% lower than that of the teacher network with full supervision training. At the same time, the proposed method is compared with the feature map-oriented knowledge distillation algorithm and the model compression algorithm based on adversarial learning training, and the proposed method has a better performance.

**Key words:** knowledge distillation; adversarial learning; mutual learning; model compression; face recognition

**基金项目:**国家社会科学基金(15AGJ012)。

**作者简介:**刘金金(1988—),女,博士,研究方向为人工智能、机器视觉,E-mail:liujinjin0809@zzti.edu.cn;李清宝(1966—),男,博士研究生,教授,研究方向为信息安全、人工智能;李晓楠(1983—),女,博士,讲师,研究方向为图像处理。

**收稿日期:**2021-05-19 **修回日期:**2021-07-05 **文章编号:**1002-8331(2021)21-0180-08

深度神经网络在多种计算机视觉相关的任务中展现出了最优越的性能，例如图像分类[1]、工业视觉检测[2]、姿态估计[3]、行人再识别[4]和人脸识别[5]等。随着配套硬件设备的发展和对卷积神经网络认识的不断加深，研究表明越深的网络能够提取越抽象的语义信息，网络的表示能力越强。然而更深更宽的神经网络将难以收敛，并且会导致反向传播算法中的梯度消失[6-7]。残差网络 ResNet[1]和批量归一化（Batch Normalization，BN）[8]能够在一定程度上解决这一问题，但是具有大量参数的深度学习模型需要更大的存储空间和更强的运算单元，无法在移动终端上进行部署和实时推理，从而影响深度学习模型在实际应用中的落地和推广。例如公共区域的视频监控系统多部署在内存有限和计算能力较低的嵌入式设备上，无法实时准确地对视频帧中的多人进行身份识别和行为分析。本研究旨在改善深度学习网络在人脸识别系统中的部署和应用问题。

为了解决这一问题，研究人员采用多种技术压缩网络参数，主要包括量化或二值化、因子分解、网络剪枝和知识蒸馏[9]等。其中知识蒸馏的方法基于教师-学生的策略，旨在训练一个轻量级的学生网络，使学生网络模仿完备的大尺寸的教师网络输出的软目标，达到知识迁移的目的。相较于样本固有的一位有效信息标签，教师网络的输出能够提供“不正确”分类的相对概率，使分类概率具有更大的信息量。学生网络的权重更新是一个最小化知识蒸馏损失的过程，即最小化学生网络输出与教师网络输出、学生网络输出与真实标签间的差异。

虽然通过最小化知识蒸馏损失能够使学生网络模仿教师网络的输出，但是其性能仍有差距。主要有以下几个原因。首先学生网络只学习教师网络输出的分类概率分布，而忽略了包含丰富语义信息和空间相关性的中间特征图。一些现有方法直接对齐学生网络和教师网络的中间层表示，不能有效地转移潜在的空间相关性。其次由于教师网络和学生网络具有不同的拓扑结构，其最优解空间也存在差异。如果只采用蒸馏损失全程监督学生网络的训练过程，学生网络无法找到自己的最优解空间。此外，教师网络的预测不是完全正确的，在训练过程中，如果学生网络完全学习教师网络输出的软目标，会迁移错误的知识。

针对上述问题，本研究提出了一种针对分类概率和特征图两个层面的深度学习模型压缩算法。框架由三部分组成，分别为预训练得到的教师网络、小规模的学生网络和判别器。其中教师网络和学生网络可以为任意结构的卷积神经网络，判别器则为由多个全连接层构成的深度学习网络，并在训练过程中更新权重。为了减轻教师网络错误分类的影响，添加指示函数，优化知识蒸馏损失，使学生网络只学习正确的输出。此外由于人脸特征具有丰富的空间相关性，学生网络模仿教师网络提取的特征图是十分必要的。引入生成对抗网络[10]中的判别器，识别输入的特征图是“真”（教师网络）还是“假”（学生网络），使学生网络能够自动学习类间的相关性。为了使学生网络能够自主探索自己的最优解空间，在训练过程中打破教师网络和学生网络之间的单向转换通路，使其相互学习，交替更新。完整流程如图 1 所示，其中实线为网络的正向传播过程，虚线为目标函数的计算过程。

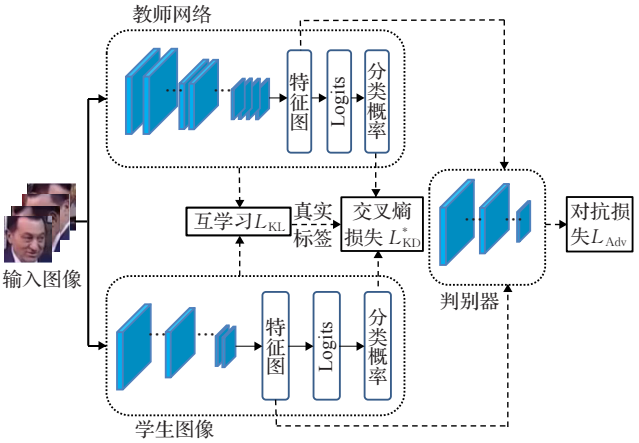


**图 1　对抗学习辅助下的知识蒸馏过程**

Fig.1　Process of knowledge distillation assisted by adversarial learning

本研究的贡献总结如下：

（1）改进经典的知识蒸馏损失，使其不仅能够学习教师网络输出的正确软目标，而且能够从中间层获取丰富的隐含知识。

（2）引入对抗学习中的判别器，鉴别教师网络和学生网络特征图的差异，进一步缩小大模型和容量有限的小模型最优解空间之间的差异。

（3）在训练过程中采用互学习的策略，使教师网络和学生网络学习对方的特征图，提升泛化能力。

（4）由于本研究针对人脸识别的应用，采用公开的人脸数据集训练模型，并与已有的理想算法进行比较，验证所提算法的有效性和先进性。

## 1　相关工作

神经网络的中存在很多冗余参数，文献[11]的研究表明，模型中有仅用 1%的深度卷积就能达到和原来网络相近的性能。神经网络压缩早在文献[12]的工作中就已为人所知，但最近由于现代深度模型的性能和计算需求的综合增长而受到了广泛关注。

### 1.1　卷积神经网络压缩算法

神经网络压缩的相关方法主要分为五大类：量化、剪枝、因子分解、精细模型设计和知识蒸馏等。

量化的方法是将网络的权重离散化，逐步将预先训练的全精度卷积网络转换为低精度卷积网络。基于这

一思想,Gong等人[13]使用$k$-means对权值进行聚类,然后进行量化。量化可以简化到二进制级别的−1和1,如XNOR-Net[14]和BinaryConnect[15],但后者不在参数更新期间量化,而是在前向和后向梯度传递的过程中二进制化权重。

剪枝则是根据一定规则剔除网络中链接的过程,根据粒度的粗细又可以分为面向链接的剪枝和面向卷积模板的剪枝。Han等人[16]将量化与剪枝相结合,进一步减少存储需求和网络计算。在HashedNet[17]中,网络连接被随机分组到哈希桶中,相同桶的连接共享权值。然而当使用卷积神经网络时,稀疏连接不一定会加速推理。出于这个原因,Li等人[18]裁剪完整的卷积模板,而不是单个的连接。因此剪枝后的神经网络仍然进行密集矩阵乘法,而不需要稀疏卷积库。

因子分解的方法旨在对卷积模板的矩阵进行低秩分解或找寻近似的低秩矩阵。其中,使用深度可分离卷积和点卷积的组合可以近似深度卷积模板,例如MobileNet[19]和ShuffleNet[20]。

相较于AlexNet[21]和VGG16[22]需要较多的计算资源,残差网络(Residual Network,ResNet)及其变体不仅减少了参数数量,同时保持(甚至提高)了性能。SqueezeNet[23]通过利用1×1卷积模板替换3×3卷积模板并减少3×3卷积模板的通道数量来进一步修剪参数。此外Inception[24]、Xception[25]、CondenseNet[26]和ResNeXt[27]也有效地设计更深更宽的网络,而不引入比AlexNet和VGG16更多的参数。Octave卷积[28]根据不同的频率将特征图进行因式分解,对不同频率的信息进行不同的存储和操作,以实现基于高低频率的轻量化存储方式。Octave卷积可用于ResNet、GoogLeNet等基线网络结构的优化,也可以对如MobileNet-v1&v2,ShuffleNet v1&v2等常规轻量化网络进行进一步地优化,能够有效减少深度神经网络对于存储空间的要求,实现轻量化。

除了人工设计轻量化深度神经网络以外,基于神经架构搜索的自动化模型设计的优势愈加凸显,如MnasNet[29]等。

### 1.2　知识蒸馏

知识蒸馏(Knowledge Distillation,KD)的目标是通过使用Softmax函数之前(Logits)或者之后的输出(分类概率),将知识从教师网络转移至学生网络。

为了使学生网络能够完成多个计算机视觉任务,文献[30]则是蒸馏多个教师网络,构成多分支的学生网络。文献[31]优化了教师-学生策略,弱化策略中两者的指导与学习关系,两个均从头训练,相互学习,并且采用循环训练策略同时训练多个网络。文献[32]中提出的教师-学生策略中,两种网络采用相同的架构,而使用不同分辨率的人脸图像,能在一定程度上解决实际应用中图像分辨率低的问题。学生网络的训练方法分为知识蒸馏和知识迁移两种,前者随机初始化学生网络参数,采用基于分类的交叉熵损失和特征向量间的欧氏距离更新网路,后者则是用教师网络的参数初始化学生网络,然后只采用交叉熵损失更新网络。

### 1.3　对抗学习

生成对抗网络通过对抗学习生成图像,使生成器能够模拟特定的特征分布空间。这一本质特性与知识蒸馏的目的存在交叉,可以将小容量的学生网络看作生成器,在给定相同输入图像的情况下,将学生输出映射到教师输出。

Belagiannis等人提出了一种基于对抗学习的网络压缩算法[33],去掉了KD损失中学生网络输出与真实标签的交叉熵损失,而是使用两者Logits间的L2范数,故不需要提前给出训练样本的真实标签。由于学生网络具有较小的容量,很难使其完全精确地模仿教师网络的软目标,增加对抗损失,使学生网络能够能快地收敛于教师网络的最优解空间。由于判别器过早达到平衡会使学生网络无法从教师网络学习到有效的梯度,引入对判别器的正则化,避免判别器支配后续的训练过程。

针对人脸识别任务的特性,本研究优化了现有的知识蒸馏方法,不只简单学习分类概率,同时考虑特征图间的知识迁移。为了使学生网络能够探索自己的最优解空间,加入判别损失这一更加宏观的标准,使学生网络在训练过程中具有更多的自主性。

## 2　算法分析

在视频监控系统的应用中,为了具有较好的人脸识别性能,多采用具有更深更宽结构的深度学习模型。相反,在一些现实场景中,为了满足资源有限的设备的需求,需要对已有模型进行剪枝或量化。为了解决这两个目标之间的权衡困难,本研究提出了一种对抗学习辅助下的知识蒸馏算法。

本研究主要从知识获取对象和知识蒸馏策略两个方面入手优化了现有算法,从以下三个方面进行详细阐述。

### 2.1　基于分类概率的知识蒸馏优化

知识蒸馏的基本思想是通过最小化教师网络和学生网络间的预测分布的差异,使学生网络近似于教师网络。神经网络通常通过使用Softmax输出层来产生分类概率,将计算出的每个类别的Logits转换为分类概率,如式(1)所示:

$$q_i=\frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \tag{1}$$

其中,$z_i$为Logits的第$i$个分量,$T$为温度参数,越高的温度会产生越软的类间分类概率。

知识蒸馏损失有两部分组成,一是分类概率间的交

叉熵,学生网络和教师网络使用相同的温度 $T$,二是学生网络的分类预测与真实标签间的交叉熵损失,温度为1,如式(2)所示:

$$L_{\mathrm{KD}}=\frac{1}{N}\sum_{i=1}^{N}[L_{\mathrm{CE}}(y_i,\sigma(z^{S_i}))+ T^2\times L_{\mathrm{CE}}(\sigma(z^{S_i}/T),\sigma(z^{T_i}/T))] \tag{2}$$

其中,$N$ 为小批量的尺寸,$L_{\mathrm{CE}}$ 代表交叉熵,也可以用相对熵,即 Kullback-Leibler 散度代替。$\sigma()$ 代表 Softmax 函数,$T$ 为蒸馏温度,$y_i$ 为样本 $i$ 的真实标签,$z^S\in\mathbb{R}^C$ 和 $z^T\in\mathbb{R}^C$ 分别为C类分类任务的学生网络和教师网络输出的Logits。

虽然训练初期教师网络比学生网络更准确,但教师仍然会有一些预测错误。当教师网络预测错误时,知识同样会转移到学生网络身上,这将会影响学生网络的表现。因此改进传统知识蒸馏的方法,忽略教师网络错误的预测分布,只把正确的预测分布传递给学生网络,具体目标函数如式(3)所示:

$$L^{*}_{\mathrm{KD}}=\frac{1}{N}\sum_{i=1}^{N}[L_{\mathrm{CE}}(y_i,\sigma(z^{S_i}))+ I(y_i,y_i^S)\times T^2\times L_{\mathrm{CE}}(\sigma(z^{S_i}/T),\sigma(z^{T_i}/T))] \tag{3}$$

其中,$I()$ 为指示函数,$y^S$ 为学生网络预测的标签。当教师网络能够正确预测输入样本的分类时,指示函数为1,学生网络同时学习样本标签和教师网络输出的软目标;教师网络无法正确分类时,指示函数为0,仅计算学生网络的分类情况和真实标签间的交叉熵。

## 2.2 对抗学习辅助下的特征图迁移

研究表明[34],位于较浅层面的卷积层会对边缘、角度、曲线等低级特征做出响应;下一级卷积层则能响应更复杂的特征,如圆和矩形。因此,当卷积操作逐渐加深,卷积层会提取出更复杂、高维的特征。另一方面,深度卷积特征也比浅层卷积特征更能代表网络的泛化能力。因此,选择教师网络Logits前的特征图作为学生网络学习的对象,分别记为 $f^T$ 和 $f^S$。

常用相似度计算方法有余弦距离、欧氏距离、马氏距离等。这些方法的目的均是使学生网络最大程度模仿教师网络输出的特征图。由于学生网络的容量小,它可能无法精确地再现某一特定的输出模态,并且实际中学生网络与教师网络具有不同的结构,没有必要精确地模拟一个教师网络的输出来达到良好的表现。因此本研究提出一个面向教师网络和学生网络的对抗学习机制。对抗训练策略缓解了人工设计损失函数的困难,已经在多个计算机视觉任务中显示出了优越性。

特征图学习机制由三部分组成,即教师网络、学生网络和判别器。教师网络和学生网络的输入为相同的人脸图像,将其输出的特征图作为判别器的输入,判别器鉴定其来自哪个网络。采用生成对抗网络中的对抗损失作为目标函数,如式(4)所示:

$$L_{Adv}=E_{f^T\sim p_{\mathrm{teacher}}}[\mathrm{lb}(D(f^T))]+ E_{f^S\sim p_{\mathrm{student}}}[\mathrm{lb}(1-D(f^S))] \tag{4}$$

在训练过程中,判别器的目的是最小化对抗损失,确保正确区分两个不同的分布;学生网络的目的则是使判别器无法区分其与教师网络的差异,以此构成对抗训练 $\min\limits_{S}\max\limits_{D}L_{\mathrm{Adv}}$。判别器和学生网络交替更新,直至判别器的识别准确率为1/2,此时网络收敛。

相较分类概率,高维特征图能够保留更多的特征,采用高维的特征图作为判别器的鉴定对象能够使判别器具有更强的鉴别能力,指导学生网络的更新,最小化与教师网络的差异。

## 2.3 学生网络和教师网络的深度互学习

相关研究表明[35],教师网络的影响不总是积极的。在网络训练的前期,知识蒸馏辅助学生网络的更新,但是随后会抑制学生网络的优化。实验结果表明,在某一时期,交叉熵损失会反向上升。此外模型蒸馏算法需要有提前预训练好的教师网络,且教师网络在学习过程中保持固定,仅对学生网络进行单向的知识传递,难以从学生网络的学习状态中得到反馈信息来对训练过程进行优化调整。

深度互学习[36]指即多个网络相互学习,每个网络在训练过程中不仅接受来自真值标记的监督,还参考同伴网络的学习经验来进一步提升泛化能力。真值标签提供的信息仅包含样本是否属于某一类,但缺少不同类别之间的联系,而网络输出的分类概率则能够在一定程度上恢复该信息,因此网络之间进行分类概率交叉学习可以传递学习到的数据分布特性,从而帮助网络改善泛化性能。其次,网络在训练过程中会参考同伴网络的经验来调整自己的学习过程,最终能够收敛到一个更平缓的极小值点,小的波动不会对网络的预测结果造成剧烈影响,从而具备更好的泛化性能。

本研究中采用学生网络和教师网络特征图间的Jensen-Shannon散度作为互学习的目标函数,如式(5)所示。相较于KL散度,JS散度是对称的,解决了KL散度非对称的问题。

$$L_{\mathrm{ML}}=JS(q^T\|q^S) \tag{5}$$

其中,$q^T$ 和 $q^S$ 分别为教师网络和学生网络的分类概率分布。联合基于对抗学习的特征图迁移和互学习方法,学生网络不仅能够模仿教师网络特征图的分布,同时能够保留自主学习的能力。

综上所述,用以训练学生网络的目标函数的完整形式为:

$$L_S=L^{*}_{\mathrm{KD}}+\alpha L_{\mathrm{Adv}}+\beta L_{\mathrm{ML}} \tag{6}$$

其中,$\alpha,\beta\in[0,1)$ 为超参数,用以平衡各部分间的权重。

## 3 实验结果与分析

### 3.1 实现细节

#### 3.1.1 数据集

本研究中所采用CASIA-WebFace[37]和CelebA[38]数据集作为训练样本集。CASIA-WebFace数据集中的样本来自IMDb网站,有10 575人的494 414张照片。CelebA数据集包含超过200 000的名人图片,每张图片有40个属性标注。该数据集中的图像具有较大的姿态变化和复杂背景。对所有图像进行归一化处理,尺寸统一为256×256。

#### 3.1.2 网络结构

由于残差网络在图像分类应用中具有最优的性能,所以本研究中采用ResNet-101作为教师网络。ResNet-101由两种残差块组成,一是Identity Block,输入和输出的维度相同,二是Conv Block,输入和输出的维度不同,用以改变特征向量的维度。教师网络预先在数据集上训练,在学生网络的训练过程中的互学习阶段微调。学生网络在实验中采用和教师网络同样采用残差结构的ResNet-18,在训练前随机分配链接权重。具体结构如表1所示。

**表1 教师网络和学生网络的具体结构**

Table 1 Detailed structure of teacher and student

| | ResNet-18 | ResNet-101 |
|---|---|---|
| conv1 | 7×7@64 | |
| | 3×3最大池化 | |
| conv2_x | $\begin{bmatrix}3\times3@64\\3\times3@64\end{bmatrix}\times2$ | $\begin{bmatrix}1\times1@64\\3\times3@64\\1\times1@256\end{bmatrix}\times3$ |
| conv3_x | $\begin{bmatrix}3\times3@128\\3\times3@128\end{bmatrix}\times2$ | $\begin{bmatrix}1\times1@128\\3\times3@128\\1\times1@512\end{bmatrix}\times4$ |
| conv4_x | $\begin{bmatrix}3\times3@256\\3\times3@256\end{bmatrix}\times2$ | $\begin{bmatrix}1\times1@256\\3\times3@256\\1\times1@1\,024\end{bmatrix}\times23$ |
| conv5_x | $\begin{bmatrix}3\times3@512\\3\times3@512\end{bmatrix}\times2$ | $\begin{bmatrix}1\times1@512\\3\times3@512\\1\times1@2\,048\end{bmatrix}\times3$ |
| 平均池化,全连接,Softmax | | |

判别器是模型中的重要组成部分,必须在简单性和网络容量之间取得平衡。判别器由三个全连接层组成(128fc-256fc-128fc),中间激活为非线性激活单元ReLU。输出则是由Sigmoid函数给出的二元预测,判定输入的特征图来自哪个网络。

#### 3.1.3 相关参数设置

在所有实验中,均采用随机梯度下降的算法。根据文献[39],动量设为0.9,权重衰减为0.000 1,小批量尺寸为128。对于两个训练集,均训练200轮次。初始学习率为0.1,分别在第80轮次和160轮次分别将学习率调整为0.01和0.001。

在前160轮中,只采用知识蒸馏损失和对抗损失指导学生网络的更新,使学生网络的分类错误率迅速下降;在后40轮中加入两个网络的互学习损失,对学生网络微调,具有更好的泛化能力。在多次实验中选取最优的超参数,使学生网络尽快收敛,故在上述两个训练阶段超参数分别取 $\alpha=0.5,\beta=0$ 和 $\alpha=0.5,\beta=0.05$ 。

### 3.2 算法流程

网络压缩分为两个阶段:首先,在学生网络训练的前半段,知识蒸馏损失和对抗损失指导学生网络和判别器的更新,在训练的后半段,教师网络和学生网络互学习,对模型进行微调,以提高泛化能力。算法伪码如算法1所示。

**算法1** 基于知识蒸馏和对抗学习的网络压缩算法

输入 预训练的教师网络;随机初始化的学生网络和判别器;训练数据集;学习率、小批量尺寸及 $\alpha,\beta$

输出 小尺寸的学生网络

1. for每个小批量
2. 输入样本 $x_i$ ,真实标签 $y_i$ ;
3. 将样本分别输入教师网络和学生网络,得到特征图 $f^T$ 和 $f^S$ ;
4. 将特征图通过激活函数,得到分类概率 $q^T$ 和 $q^S$ ;
5. 计算知识蒸馏损失 $L^*_{\mathrm{KD}}$ ;
6. 将特征图 $f^T$ 和 $f^S$ 输入判别器,计算对抗损失 $L_{\mathrm{Adv}}$ ;
7. 最大化对抗损失 $L_{\mathrm{Adv}}$ ,更新判别器;
8. if 训练轮数≤160
9. 计算 $L_S=L^*_{\mathrm{KD}}+\alpha L_{\mathrm{Adv}}$ ;
10. 使用反向传播算法更新学生网络的权重;
11. else
12. 计算互学习损失 $L_{\mathrm{ML}}$ ;
13. 计算 $L_S=L^*_{\mathrm{KD}}+\alpha L_{\mathrm{Adv}}+\beta L_{\mathrm{ML}}$ ;
14. 使用反向传播算法更新学生网络的权重;
15. 基于互学习损失 $L_{\mathrm{ML}}$ 微调教师网络;
16. end for

### 3.3 性能对比

本节中从两个方面定量地分析所提方法的有效性,首先比较全监督下训练的学生网络和不同温度下知识蒸馏得到的学生网络,如表2所示;然后在CASIA WEBFACE和CelebA数据集上训练其他知识蒸馏方法,

**表2 全监督训练与知识蒸馏性能对比**

Table 2 Comparison between fully supervised training and knowledge distillation

%

| 训练方法 | | 网络结构 | CASIA WEBFACE | CelebA |
|---|---|---|---|---|
| 全监督教师网络 | | ResNet-101 | 82.57 | 82.38 |
| 全监督学生网络 | | ResNet-18 | 79.62 | 79.35 |
| 本文知识蒸馏 | $T$=1 | ResNet-18 | 78.58 | 78.60 |
| | $T$=2 | ResNet-18 | 78.95 | 79.25 |
| | $T$=5 | ResNet-18 | 79.92 | 79.90 |
| | $T$=10 | ResNet-18 | 81.05 | 81.00 |

与所提方法比较,如表3所示。将数据集按照4:4:2的比例随机划分为训练集、测试集和验证集,表中准确率在验证集上获得,评价指标为Top-1准确率。

**表3 所提方法与其他知识蒸馏方法的性能对比**

Table 3 Comparison between proposed method and other knowledge distillation methods %

| 训练方法 | 网络结构 | CASIA WEBFACE | CelebA |
|---|---|---|---|
| 全监督教师网络 | ResNet-101 | 82.57 | 82.38 |
| 全监督学生网络 | ResNet-18 | 79.62 | 79.35 |
| KD | ResNet-18 | 78.43 | 78.52 |
| DML | ResNet-18 | 79.02 | 79.05 |
| FitNet | ResNet-18 | 78.43 | 78.61 |
| CD | ResNet-18 | 79.50 | 79.51 |
| KTAN | ResNet-18 | 79.56 | 79.45 |
| 本文知识蒸馏 | ResNet-18 | 81.05 | 81.00 |

表2中前两行分别为全监督训练方式下得到的教师网络和学生网络,网络结构分别固定为ResNet-101和ResNet-18。ResNet-101的参数量为$44.6\times10^6$,浮点运算量为1.8 GFLOPs,而ResNet-18的参数量仅为$11.2\times10^6$,浮点运算量为7.6 GFLOPs。同时利用本文所提出的知识蒸馏算法训练ResNet-18网络,温度$T\in\{1,2,5,10\}$。表2中还可以看出,由于网络深度不同,全监督训练下的学生网络在两种验证集上的准确率均低于教师网络,但是学生网络的参数数量仅为教师网络的1/4,模型尺寸和识别性能间有较好的平衡。采用知识蒸馏算法得到的学生网络在验证集上的性能略低于全监督训练的教师网络,但是随着温度$T$的增加,其性能超越了全监督训练的学生网络。由此可见,本研究所提出的知识蒸馏算法能够使小尺寸的网络学习到大规模网络的知识,有效地实现知识迁移。同时,温度$T$越高,输出的分类概率越平缓,得到的识别性能越好;但是随着温度的逐渐增加,性能提高趋于平缓,表明较高的温度会引起错误标签的概率增加,使学生网络更多地关注相关知识。因此在知识蒸馏方法中选择合适的温度$T$是比较重要的。

本研究中还将所提方法与其他知识蒸馏算法进行了比较,包括经典知识蒸馏算法KD、深度互学习算法DML、FitNet[40]、Channel Distillation[41]和KTAN[42]。FitNet将中间层的表示作为学生网络的学习对象,要求学生网络的中间层模仿老师网络特定的中间层的输出。CD中分别计算教师网络和学生网络每个通道的注意力信息,教师监督学生学习其注意力信息,已达到通道间传递信息的目的。KTAN除了采用分类概率间的交叉熵损失外,采用均方误差损失最小化学生网络和教师网络特征图之间的差异,同时引入判别器鉴定特征图的出处。

表3中的数据均在CASIA WEBFACE和CelebA数据集上获得,教师网络和学生网络的结构分别为ResNet-101和ResNet-18,训练参数如3.1.3小节所示。由表中数据可知,由于教师网络具有更深的结构,所以展现出最好的性能。五种引自文献的方法,除了CD和KTAN在CelebA数据集上的性能外,其余数据在Top-1性能中均低于全监督的学生网络。本文所提出的知识蒸馏方法相较其他知识蒸馏方法,学生网络不仅学习了教师网络正确的分类概率,同时利用对抗学习的方法使学生网络定性学习教师网络的特征图;在训练的后半段,教师网络和学生网络相互学习,促进学生网络探索自己的最优解空间,具有更好的泛化能力。在CASIA WEBFACE和CelebA两个数据集上,本研究的识别准确率均超过了全监督训练得到的学生网络,证明了所提方法的有效性和先进性。

### 3.4 消融实验

本节中,通过消融实验分别验证了式(6)中知识蒸馏损失,对抗损失和互学习损失的有效性。实验结果如表4所示。

**表4 基于不同损失函数的性能对比**

Table 3 Performance comparison based on different loss functions %

| 训练方法 | | CASIA WEBFACE | CelebA |
|---|---|---|---|
| 全监督学生网络 | | 79.62 | 79.35 |
| 本文知识蒸馏 | $L^*_{KD}$ | 78.56 | 78.64 |
| | $L^*_{KD}+L_{Adv}$ | 79.68 | 79.79 |
| | $L^*_{KD}+L_{ML}$ | 79.52 | 79.55 |
| | $L^*_{KD}+L_{Adv}+L_{ML}$ | 81.05 | 81.00 |

对比表3和表4中数据可以看出,改进后的知识蒸馏损失剔除了错误估计样本,使识别性能得到了小幅度提升。分别在知识蒸馏损失的基础上添加基于特征图的对抗损失和互学习损失后,性能的提成在1%左右,均超过了互学习DML算法、将中间层作为学习对象的FitNet和CD方法。对比KTAN,本研究方法摒弃了均方误差,使学生网络不必完全模拟教师网络的特征图,能够自主更新,性能得到了进一步的提升。

通过消融实验可知,本研究所采用的三种损失函数能够有效地完成从教师网络到学生网络的知识迁移,不仅包括分类概率知识,同时学习了特征图的分布,并且适度保留自主学习的空间。

## 4 结束语

深度学习方法越来越多地应用于计算机视觉任务中,并且在人脸识别任务中的准确率已经超越了人眼,但是仍然面对着训练时间长、泛化能力弱、部署困难等落地问题。本研究针对深度学习模型在嵌入式设备难以进行部署和实时性能差的问题,深入研究了现有的模型压缩和加速算法,提出了一种基于知识蒸馏和对抗学

习的神经网络压缩算法。算法框架由三部分组成,预训练得到的大规模教师网络,轻量级的学生网络和辅助对抗学习的判别器。学生网络通过知识蒸馏损失学习教师网络的分类概率,同时通过对抗损失模拟教师网络的特征图知识。鉴于教师网络和学生网络具有不同的最优解空间,在训练的后半段利用深度互学习理论,促使学生网络和教师网络相互学习,以促使学生网络探索自己的最优解。

针对人脸识别任务,采用CASIA WEBFACE和CelebA两个数据集作为训练集,通过消融实验验证了所提组合目标函数的有效性,同时与面向特征图知识蒸馏算法和基于对抗学习训练的模型压缩算法对比,实验数据表明,根据所提算法训练得到的学生网络具有较少的链接数,同时保证了较好的识别准确率。

## 参考文献:


[1] 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,June 27-30,2016.Los Alamitos:IEEE Computer Society,2016:770-778.

[2] 鲁统伟,徐子昕,闵锋.基于生成对抗网络的知识蒸馏数据增强[J/OL].计算机工程:1-13[2021-05-18].https://doi.org/10.19678/j.issn.1000-3428.0060395.
LU T W,XU Z X,MIN F.Knowledge distillation data augmentation based on generation adversarial network[J/OL].Computer Engineering,1-13[2021-05-18].https://doi.org/10.19678/j.issn.1000-3428.0060395.

[3] GÜLER R A,NEVEROVA N,KOKKINOS I.DensePose:dense human pose estimation in the wild[C]//IEEE Conference on Computer Vision and Pattern Recognition,Salt Lake City,June 18-22,2018.Los Alamitos:IEEE Computer Society,2018:7297-7306.

[4] LI W,ZHU X T,GONG S G.Person re-identification by deep joint learning of multi-loss classification[C]//International Joint Conference on Artificial Intelligence,Melbourne,August 19-25,2017.San Francisco:Morgan Kaufmann,2017:2194-2200.

[5] SUN Y,CHEN Y,WANG X,et al.Deep learning face representation by joint identification-verification[C]//Conference on Neural Information Processing Systems,Montreal,December 8-13,2014.Cambridge:MIT Press,2014:1988-1996.

[6] MACLAURIN D,DUVENAUD D,ADAMS R P.Early stopping is nonparametric variational inference[J].arXiv:1504.01344,2015.

[7] MAHSERECI M,BALLES L,LASSNER C,et al.Early stopping without a validation set[J].arXiv:1703.09580,2017.

[8] IOFFE S,SZEGEDY C.Batch normalization:accelerating deep network training by reducing internal covariate shift[C]//International Conference on Machine Learning,Lille,6-11 July,2015.New York:ACM,2015:448-456.

[9] HINTON G,VINYALS O,DEAN J.Distilling the knowledge in a neural network[J].arXiv:1503.02531,2015.

[10] GOODFELLOW I J,POUGET-ABADIE J,MIRZA M,et al.Generative adversarial networks[C]//Conference on Neural Information Processing Systems,Montreal,December 8-13,2014.Cambridge:MIT Press,2014:2672-2680.

[11] 毕鹏程,罗健欣,陈卫卫.轻量化卷积神经网络技术研究[J].计算机工程与应用,2019,55(16):25-35.
BI P C,LUO J X,CHEN W W.Research on lightweight convolutional neural network technology[J].Computer Engineering and Applications,2019,55(16):25-35.

[12] HASSIBI B,STORK D G.Second order derivatives for network pruning:optimal brain surgeon[C]//Conference on Neural Information Processing Systems,Denver,November 30-December 3,1992.Cambridge:MIT Press,1992:164-171.

[13] GONG Y C,LIU L,YANG M,et al.Compressing deep convolutional networks using vector quantization[J].arXiv:1412.6115,2014.

[14] RASTEGARI M,ORDONEZ V,REDMON J,et al.XNOR-net:imagenet classi-fication using binary convolutional neural networks[C]//European Conference on Computer Vision,Amsterdam,October 11-14,2016.Berlin:Springer,2016:525-542.

[15] COURBARIAUX M,BENGIO Y,DAVID J P.Binary-connect:training deep neural networks with binary weights during propaga-tions[C]//Conference on Neural Information Processing Systems,Montreal,December 7-12,2015.Cambridge:MIT Press,2015:3123-3131.

[16] HAN S,MAO H,DALLY W J.Deep compression:Compressing deep neural networks with pruning,trained quantization and human coding[C]//International Conference on Learning Representations,San Juan,May 2-4,2016.

[17] CHEN W L,WILSON J T,TYREE S,et al.Compressing neural networks with the hashing trick[C]//International Conference on Machine Learning,Lille,6-11 July 2015.New York:ACM,2015:2285-2294.

[18] LI H,KADAV A,DURDANOVIC I,et al.Pruning filters for efficient ConvNets[C]//International Conference on Learning Representations,Toulon,April 24-26,2017.

[19] HOWARD A G,ZHU M L,CHEN B,et al.MobileNets:efficient convolutional neural networks for mobile vision applications[J].arXiv:1704.04861,2017.

[20] ZHANG X Y,ZHOU X Y,LIN M X,et al.ShuffleNet:an extremely efficient convolutional neural network for mobile devices[C]//IEEE Conference on Computer Vision and Pattern Recognition,Salt Lake City,June 18-22,2018.

Los Alamitos:IEEE Computer Society,2018:6848-6856.
[21] KRIZHEVSKY A,SUTSKEVER I,HINTON G E.Imagenet classification with deep convolutional neural networks[C]//Conference on Neural Information Processing Systems,Lake Tahoe,December 3-6,2012.Cambridge:MIT Press,2012:1097-1105.
[22] SIMONYAN K,ZISSERMAN A.Very deep convolutional networks for large-scale image recognition[C]//International Conference on Learning Representations,San Diego,May 7-9,2015.
[23] IANDOLA F N,MOSKEWICZ M W,ASHRAF K,et al.SqueezeNet:Alexnet-level accuracy with 50x fewer parameters and<0.5 mb model size[J].arXiv:1602.07360,2016.
[24] SZEGEDY C,VANHOUCKE V,IOFFE S,et al.Rethinking the inception architecture for computer vision[C]//IEEE Conference on Computer Vision and Pattern Recognition,Las Vegas,June 27-30,2016.Los Alamitos:IEEE Computer Society,2016:2818-2826.
[25] CHOLLET F.Xception:deep learning with depthwise separable convolutions[C]//IEEE Conference on Computer Vision and Pattern Recognition,Honolulu,July 21-26,2017.Los Alamitos:IEEE Computer Society,2017:1800-1807.
[26] HUANG G,LIU S C,VAN DER MAATEN L,et al.Condensenet:an efficient densenet using learned group convolutions[C]//IEEE Conference on Computer Vision and Pattern Recognition,Salt Lake City,June 18-22,2018.Los Alamitos:IEEE Computer Society,2018:2752-2761.
[27] LEBEDEV V,GANIN Y,RAKHUBA M,et al.Speeding-up convolutional neural networks using fine-tuned CP-decomposition[C]//International Conference on Learning Representations,San Diego,May 7-9,2015.
[28] CHEN Y,FAN H,XU B,et al.Drop an octave:reducing spatial redundancy in convolutional neural networks with octave convolution[C]//Proceedings of the IEEE International Conference on Computer Vision,Seoul,October 27-November 2,2019.Piscataway:IEEE,2019:3435-3444.
[29] TAN M,CHEN B,PANG R,et al.Mnasnet:platform-aware neural architecture search for mobile[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,Los Angeles,June 16-20,2019.Piscataway:IEEE,2019:2820-2828.
[30] ASIF U,TANG J B,HARRER S.En-semble knowledge distillation for learning improved and efficient networks[C]//European Conference on Artificial Intelligence,Santiago de Compostela,29 August-8 September 2020.Amsterdam:IOS Press,2020:953-960.
[31] CHUNG I,PARK S,KIM J,et al.Feature-map-level online adversarial knowledge distillation[C]//International Conference on Learning Representations,Addis Ababa,April 26-30,2020:2006-2015.
[32] KARLEKAR J,FENG J S,WONG Z S,et al.Deep face recognition model compression via knowledge transfer and distillation[J].arXiv:1906.00619,2019.
[33] BELAGIANNIS V,FARSHAD A,GALASSO F.Adversarial network compression[C]//European Conference on Computer Vision,Munich,September 8-14,2018.Berlin:Springer,2018:431-449.
[34] ZEILER M D,FERGUS R.Visualizing and understanding convolutional networks[C]//European Conference on Computer Vision,Zurich,September 6-12,2014.Berlin:Springer,2014:818-833.
[35] CHO J H,HARIHARAN B.On the efficacy of knowledge distillation[C]//IEEE International Conference on Computer Vision,Seoul,October 27-November 2,2019.Piscataway:IEEE,2019:4793-4801.
[36] ZHANG Y,XIANG T,HOSPEDALES T M,et al.Deep mutual learning[C]//IEEE Conference on Computer Vision and Pattern Recognition,Salt Lake City,June 18-22,2018.Los Alamitos:IEEE Computer Society,2018:4320-4328.
[37] YI D,LEI Z,LIAO S C,et al.Learning face representation from scratch[J].arXiv:1411.7923,2014.
[38] LIU Z W,LUO P,WANG X G,et al.Deep learning face attributes in the wild[C]//IEEE International Conference on Computer Vision,Santiago,December 7-13,2015.Alamitos:IEEE Computer Society,2015:3730-3738.
[39] XU Z,HSU Y C,HUANG J W.Training shallow and thin networks for acceleration via knowledge distillation with conditional adversarial networks[C]//International Conference on Learning Representations,Vancouver,April 30-May 3,2018.
[40] ROMERO A,BALLAS N,KAHOU S E,et al.Fitnets:hints for thin deep nets[C]//International Conference on Learning Representations,San Diego,May 7-9,2015.
[41] ZHOU Z D,ZHUGE C R,GUAN X W,et al.Channel distillation:channel-wise attention for knowledge distillation[J].arXiv:2006.01683,2020.
[42] LIU P Y,LIU W,MA H D,et al.KTAN:knowledge transfer adversarial network[C]//International Joint Conference on Neural Networks,Glasgow,July 19-24,2020.Piscataway:IEEE,2020:1-7.