文章编号:1673-0291(2017)06-0027-07 DOI:10.11860/j.issn.1673-0291.2017.06.005

基于深度特征蒸馏的人脸识别

葛仕明[1,2],赵胜伟[1,2],刘文瑜[3],李晨钰[1,2]
(1.中国科学院信息工程研究所,北京 100095;2.中国科学院大学 网络空间安全学院,北京 100019;
3.北京大学 软件与微电子学院,北京 102600)

摘 要:在人脸识别系统中,深度学习由于强大的表征能力被广泛应用,但模型推理的高计算复杂度和特征表示的高维度分别降低了特征提取和特征检索的效率,阻碍了人脸识别系统的实际部署.为了克服这两个问题,本文提出一种基于深度特征蒸馏的人脸识别方法,该方法通过多任务学习实现大深度模型知识与领域相关数据信息的蒸馏,从而统一地压缩深度网络参数及特征维度.联合特征回归与人脸分类,以预训练的大网络为教师网络,指导小网络训练,将知识迁移得到轻量级的学生网络,实现了高效的特征提取.在 LFW 人脸识别数据集上进行了实验,学生模型在识别精度相比教师模型下降 3.7%的情况下,模型参数压缩到约 2×10^7、特征维度降到 128 维,相比教师模型分别获得了 7.1 倍的参数约减、32 倍的特征降维及 95.1%的推理复杂度下降,表明了方法的有效性和高效性.

关键词:深度学习;特征表示;知识蒸馏;模型压缩;人脸识别
中图分类号:TP183 **文献标志码**:A

Face recognition based on deep feature distillation

GE Shiming[1,2], *ZHAO Shengwei*[1,2], *LIU Wenyu*[3], *LI Chenyu*[1,2]
(1.Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100095, China;
2.School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100019, China;
3.School of Software & Mircoelectronics, Peking University, Beijing 102600, China)

Abstract: Deep learning has been widely used in face recognition system due to its powerful ability in feature representation. However, the high inferring complexity and feature representation reduce the efficiencies in feature extraction and retrieval respectively, which hinders the practical deployments of face recognition system. To address these issues, this paper proposes deep feature distillation in order to uniformly compress the deep network parameters and feature dimensions by distilling the knowledge from large teacher network and domain related data via multi-task deep learning. Combined feature regression and face classification, the method uses a pre-trained large depth network as a teacher network to guide the training of small network, which the knowledge transferred to the lightweight student network to achieve efficient feature extraction.

收稿日期:2017-11-01
基金项目:国家重点研发计划(2016YFC0801005);国家自然科学基金(61772513,61402463)
Foundation items: National Key Research and Development Plan (2016YFC0801005); National Natural Science Foundation of China (61772513, 61402463)
第一作者:葛仕明(1982—),男,广西宾阳人,副研究员,博士,博士生导师.研究方向为计算机视觉、深度学习和智能多媒体安全.
email:geshiming@iie.ac.cn.

引用格式:葛仕明,赵胜伟,刘文瑜,等.基于深度特征蒸馏的人脸识别[J].北京交通大学学报,2017,41(6):27-33.
GE Shiming, ZHAO Shengwei, LIU Wenyu, et al. Face recognition based on deep feature distillation[J]. Journal of Beijing Jiaotong University, 2017, 41(6):27-33.(in Chinese)

The experimental results on LFW benchmark show that in the condition of the student model recognition accuracy is reduced by 3.7% compared with the teacher model, the network has been compressed to about 2×10^7 in model size and 128 dimensional feature, which achieves the reductions of 7.1 times in model parameters, 32 times in feature dimension and 95.1% in inferring complexity. The results demonstrate the validity and efficiency of the proposed method.

Keywords: deep learning; feature representation; knowledge distillation; model compression; face recognition

随着数据采集和机器学习技术的进步,人脸识别技术取得了长足进展.对于实际部署的人脸识别系统来说,影响其性能的一个关键是高效的人脸特征表示.一方面,对于计算资源受限的设备来说,高效的人脸特征表示要求系统能够快速的提取出稳定的人脸特征;另一方面,对于云端部署来说,高效的人脸特征表示期望人脸的特征维度尽可能低,以实现快速的人脸比对检索.总之,人脸特征表示的鲁棒性、特征表示的计算复杂度及人脸特征维度,直接影响了人脸识别系统实际部署的精度和速度,因此高效的人脸特征表示在人脸识别系统中至关重要.

目前,深度学习由于其强大的表征能力,提取的特征相比传统方法手工构造的特征具有更强的鲁棒性,因此被广泛应用于人脸识别中.如 VGGFace[1] 采用 16 层的深度网络 VGGNet[2] 提取的 4 096 维人脸特征具有很强的辨识能力,在人脸识别标准数据集 LFW[3] 上能够达到 98.95%的识别精度.尽管如此,人脸识别模型提取的深度特征常采用高维特征以提升特征判别能力,模型需要上百万甚至上亿的模型参数,如 VGGFace 提取的人脸特征为 4 096 维,这导致人脸特征库存储空间过大、特征检索速度过慢,从而造成在存储和计算资源受限条件的前端设备上难以部署.

为了克服基于深度学习的人脸识别方法中模型大和特征维度高的问题,研究者分别从深度网络设计、海量训练数据的监督学习提出了一些解决方法.针对模型过大,一类方法是通过设计更小型的深度网络来解决,如 DeepID[4] 模型采用 7 层的小网络进行训练,然后集成 60 个小网络,尽管减少了模型参数,但是特征维度过高;另一类方法则是通过模型压缩方法实现,该类方法采用知识蒸馏或知识传递技术,将训练得到的大的教师模型进行压缩,尽管该类方法对教师模型的知识利用不充分,并且不能同时进行特征的压缩.针对特征维度高问题,FaceNet[5] 从上亿的海量人脸标注数据中进行模型的训练得到低维的人脸识别模型,但是对训练数据的要求很高,在面对有限训练数据的模型性能并不理想;另外有些方法则通过对深度特征进行降维来实现如 DeepID2[6] 等,这些方法不能实现端到端的训练,因此不是最优的.

传统基于交叉熵方法的训练困难性制约了小网络在人脸识别中的应用.Hinton[7] 提出的知识蒸馏方法提供了一种训练网络的新方式,即通过联合训练教师网络输出的概率分布向量联和训练集的人工标注对学生网络进行训练.通过在 MNIST 等数据集上的实验表明,学生网络能够取得超过教师网络的结果.对于输入的图像,教师网络输出的概率分布向量相比人工标注拥有更多的信息,多出的辅助信息可以加速学生网络的训练.学生网络联合学习概率分布向量和人工标注,在吸收教师网络知识的同时去除其中的错误,即对教师网络的知识进行蒸馏.受到知识蒸馏的启发,本文作者提出深度特征蒸馏方法,通过多任务损失函数的设计,联合特征回归和人脸分类,实现用已有教师模型和原始训练数据同时指导较小的学生网络的训练,在压缩模型体积的同时对特征进行蒸馏压缩,降低特征维度并提高特征的鲁棒性.教师模型为预先训练好的传统深度网络如 VGGFace 等;学生网络为轻量级的小网络如 Darknet、ResNet-34[8] 等,其模型参数较少,特征提取速度较快,同时获得的特征维度低,因此统一地实现了模型与特征的压缩.从而达到使用更小的模型和更少的计算进行更鲁棒地特征提取的目的.

本文的主要贡献包括:1)对当前基于深度学习的人脸识别模型进行了梳理与分析,揭示了当前模型实际部署的困难在于模型参数大从而计算复杂度高及特征维度高从而检索效率低;2)提出基于深度特征蒸馏的人脸识别方法,通过多任务学习蒸馏教师模型知识,统一实现深度网络与特征维度的压缩,解决深度模型实际部署问题;3)在 LFW 数据集上验证本文提出方法的有效性,揭示深度特征蒸馏方法相比传统训练方法的优势,并分析识别精度与特征维度之间的关系.

1 人脸识别相关研究工作

近年来,大规模人脸识别数据集的推动,使人脸

识别精度取得了极大提升.研究者提出了多种基于深度学习的人脸识别方法,如表 1 所示.

DeepFace[9]方法通过建立 3D 模型改进了人脸对齐的方法,在人脸数据集 SFC 上训练出一个 8 层卷积神经网络(Convolutional Neural Networks, CNNs)进行人脸特征提取,该模型具有超过 1.2 亿个参数,特征维度为 4 096 维,在 LFW 数据集上取得了 97.35%的识别精度.DeepID、DeepID 2 方法由文献[4,6]提出,DeepID 通过在人脸数据集 CelebFaces+[10]上训练一个 9 层卷积神经网络对约 10 000 个人做人脸分类,并取倒数第 2 层的输出作为人脸特征,接着用该特征训练 1 个联合贝叶斯模型用于人脸比对,最终模型大约有 1 700 万个参数,特征维度为 160 维,在 LFW 上取得了 97.45%的识别精度.DeepID 2 在卷积神经网络的训练中引入比对损失(contrastive loss)控制类内差异,大幅提升识别精度,该方法约有 1 000 万个参数,特征维度为 180 维,在 LFW 上取得了 99.15%的识别精度,超越人眼在该数据集上 97.53%的识别精度.

VGGFace[1]方法使用 VGG-16 进行2 622个人的分类,同时采用三联损失(triplet loss)进行训练,该模型有 1.38 亿个参数,特征维度为 4 096 维,在 LFW 上达到 98.95%的识别精度.FaceNet[5]方法使用三联损失来训练卷积神经网络,三联损失同时优化类间距离和类内距离,训练时分别挑选与输入人脸差异较大的相同个体人脸和差异较小的不同个体人脸组成一个批次(batch),让网络同时学习类间差别和类内共性,该模型有 1.4 亿个参数,特征维度为 128 维,在 LFW 上达到了 99.77%的识别精度.PSE[11]方法对现有公开数据集中的图像,从姿态,形状和表情 3 个方面合成新的人脸图像,极大地扩增了原有数据集的数据量,使用 VGG-19 作为网络结构,采取常规训练方法在扩容后的数据集上进行训练,最终模型约有 1.44 亿个参数,特征维度为 4 096维,在 LFW 上取得了 98.07%的识别精度.CNN-3DMM[12]方法采用卷积神经网络根据输入图像调节三维人脸模型的脸型和纹理参数,使用 ResNet-101 进行人脸识别,模型约有 3 000 万个参数,特征维度为 4 096 维,在 LFW 上达到了 92.35%的识别精度.GTNN[13]方法,使用基于张量的特征融合方式来融合深度人脸识别的特征和属性识别的特征,该方法约有 300 万个参数,在 LFW 上达到了 99.65%的识别精度.

表 1　基于深度学习人脸识别的几种方法

Tab.1　Face recognition methods based on deep learning

方法	训练数据规模×10^6	模型集成数目	模型层数	模型参数×10^6	识别精度/%
DeepFace	4.000	4	8	>120	97.35
DeepID	0.203	60	7	>17	97.45
DeepID2	0.203	25	7	10	99.15
VGGFace	2.600	1	16	138	98.95
FaceNet	260.000	1	22	140	99.60
PSE	0.500	1	19	144	98.07
CNN-3DMM	0.500	2	101	>30	92.35
GTNN	6.000	2	10	>3	99.65

基于深度学习的人脸识别模型参数需要上亿,特征维度较高,导致模型需要强大的计算资源,严格的计算硬件要求,在一定程度上制约了进一步发展和应用.为了解决这些问题,在保证识别精度的情况下,使用轻量级的小网络代替大网络,从而减少模型参数、特征维度及推理复杂度,成为人脸识别领域新的研究课题.

2　深度特征蒸馏方法

在知识蒸馏模型压缩方法[13]中,对于一个训练良好的深度网络(称为教师网络),其输出的特征通过简单的全连接神经网络便能够以很大概率正确预测出输入图像的类别,而人工标注只包括输入图像的类别信息,故教师网络输出特征所蕴含的信息基本包括了输入图像人工标注所蕴含的信息,即教师网络输出的特征相比输入图像的人工标注具有更多的信息量,附加的信息是一种辅助信息.学生网络为待训练的轻量级深度网络,其参数比教师网络少,前向传播比教师网络快,但同时带来了训练困难的问题.由于辅助信息的存在,学生网络从教师网络输出的特征中学习要比直接从原始数据中学习更加容易.基于这样的发现,本文提出深度特征蒸馏方法,通过损失函数的设计,用教师网络的特征指导监督学生网络的训练,同时联合人脸分类,对教师网络的

特征进行降维,从而得到更稳定且更高效的特征提取器.

2.1 深度特征蒸馏

深度特征蒸馏分为特征回归和人脸分类两部分:1)特征回归,学生网络直接从教师模型的特征中学习,从而将教师模型中的知识迁移到学生网络中,让学生网络获得教师网络的特征表示能力;2)人脸分类,学生网络把回归的 4 096 维特征通过全连接层压缩到低维,并接上 softmax 层进行人脸分类,从而实现对高维特征的二次加工.上述得到的低维特征融合了教师模型和训练数据中的知识,同时降低了特征的维度,因此直接取低维特征作为学生网络的输出,能够取得更好的效果.

如图 1 所示,深度特征蒸馏的目标是在教师网络 $\boldsymbol{G}$ 提取的特征指导下训练学生网络 $\boldsymbol{F}$.为了获得 $\boldsymbol{F}$,考虑一个图像 n 分类问题,训练集 $D=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_l,y_l)\}$,其中 $\boldsymbol{x}_i\in X\subset \mathbf{R}^{s\times s\times c}$,为 c 通道大小为 $s\times s$ 的图像数据,$y_i\in Y=\{1,2,\cdots,n\}$,为图像 $\boldsymbol{x}_i$ 的类别标签.对一个训练样本 $(\boldsymbol{x},y)\in D$,满足 $l_y=1,l_{i\neq y}=0$ 的 n 维向量 $\boldsymbol{l}=(l_1,\cdots,l_y,\cdots,l_n)$ 称为类别标签 y 的 one-hot 编码,深度网络 $\boldsymbol{F}(\boldsymbol{x},\boldsymbol{w})$ 最后一层输出一个 d 维向量 $\boldsymbol{p}$,传统训练方法通过最小化损失函数 $L_0(\boldsymbol{w};\boldsymbol{x},\boldsymbol{l})=H(\boldsymbol{p},\boldsymbol{l})$ 来优化模型参数 $\boldsymbol{w}$,其中 $H(\boldsymbol{p},\boldsymbol{l})$ 代表 $\boldsymbol{p}$ 和 $\boldsymbol{l}$ 的交叉熵.

本文提出的深度特征蒸馏方法改进了深度网络 $\boldsymbol{F}(\boldsymbol{x},\boldsymbol{w})$ 损失函数的设计,新的损失函数在原有交叉熵损失函数 $H(\boldsymbol{p},\boldsymbol{l})$ 的基础上,增加预先训练好的教师网络 $\boldsymbol{G}$ 的监督信号如下

$$K(\boldsymbol{w};\boldsymbol{x})=K(\boldsymbol{G}(\boldsymbol{x}',\boldsymbol{w}'),\hat{\boldsymbol{F}}(x,w))=\|\boldsymbol{G}(\boldsymbol{x}',\boldsymbol{w}')-\hat{F}(\boldsymbol{x},\boldsymbol{w})\|_2 \tag{1}$$

式中:教师网络 $\boldsymbol{G}$ 的输入 $\boldsymbol{x}'\subset \mathbf{R}^{s'\times s'\times c}$ 为学生网络输入图像 $\boldsymbol{x}$ 缩放前的图像,通常 $s\leqslant s'$,即学生网络输入比教师网络小;$\boldsymbol{w}'$ 为教师网络的参数,$\hat{\boldsymbol{F}}(\boldsymbol{x},\boldsymbol{w})$ 为学生网络中间层输出,其与教师网络输出 $\boldsymbol{G}(\boldsymbol{x}',\boldsymbol{w}')$ 具有相同的维度.最终损失函数表达式为

$$L(\boldsymbol{w};\boldsymbol{x},\boldsymbol{l})=H(\boldsymbol{p},\boldsymbol{l})+\alpha K(\boldsymbol{w};\boldsymbol{x})=H(\boldsymbol{F}(\boldsymbol{x},\boldsymbol{w}),\boldsymbol{l})+\alpha\|\boldsymbol{G}(\boldsymbol{x}',\boldsymbol{w}')-\hat{\boldsymbol{F}}(\boldsymbol{x},\boldsymbol{w})\|_2 \tag{2}$$

式中:α 为蒸馏强度;$\boldsymbol{p}$ 为学生网络 $\boldsymbol{F}(\boldsymbol{x},\boldsymbol{w})$ 最后一层特征向量;$\boldsymbol{l}$ 为输入 $\boldsymbol{x}$ 的标签向量.

通过反向传播(Back-Propagation,BP)算法迭代优化式(2),训练得到学生网络 $\boldsymbol{F}$,并最终取 $\boldsymbol{F}$ 中间层特征 $\hat{\boldsymbol{F}}(\boldsymbol{x},\boldsymbol{w})$ 和最后一层 $\boldsymbol{p}$ 之间的低维特征 $\tilde{\boldsymbol{F}}(\boldsymbol{x},\boldsymbol{w})$ 作为学生模型 $\boldsymbol{F}$ 的输出.

为达到模型压缩和加速的目的,学生网络 $\boldsymbol{F}(\boldsymbol{x},\boldsymbol{w})$ 通常选取一个轻量级的小网络,即 $\boldsymbol{w}$ 数目远小于 $\boldsymbol{w}'$,同时教师网络 $\boldsymbol{G}(\boldsymbol{x}',\boldsymbol{w}')$ 为训练良好的大网络,其模型参数 $\boldsymbol{w}'$ 包含大量知识.由于网络结构不同,教师网络 $\boldsymbol{G}$ 的知识无法直接通过参数学习迁移到学生网络 $\boldsymbol{F}$ 中,注意到 $\boldsymbol{G}$ 为预先训练好的大网络,其输出的特征 $\boldsymbol{F}_T=\boldsymbol{G}(\boldsymbol{x}',\boldsymbol{w}')$ 具有很多优良的性质,通过几个全连接层 $\boldsymbol{Q}$ 后得到 $\boldsymbol{Q}(\boldsymbol{F}_T)=\hat{\boldsymbol{G}}(\boldsymbol{x}',\boldsymbol{w}')$ 以很大概率正确预测出输入 $\boldsymbol{x}$ 的标签 $\boldsymbol{l}$,故数据对 $(\boldsymbol{x},\boldsymbol{F}_T)$ 相比原始数据 $(\boldsymbol{x},\boldsymbol{l})$ 具有更多

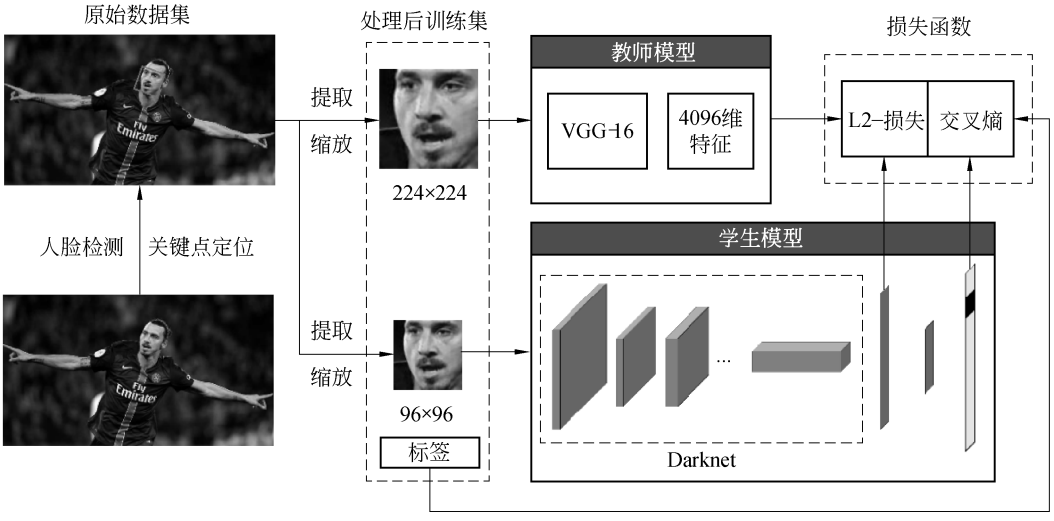

图 1 深度特征蒸馏人脸识别的框架

Fig.1 Frames of face recognition based on deep feature distillation

的知识,即 $G(D)$ 相比 D 具有更多的信息.让 $(\boldsymbol{x},\boldsymbol{F}_T)$ 通过监督信号式(1)监督学生网络 $\boldsymbol{F}$ 的训练,可以取得比直接从训练集 D 中训练更优的效果,同时让学生网络 $\boldsymbol{F}$ 接受分类信号的监督,监督信号式(1)和分类信号组成损失函数式(2)同时监督 $\boldsymbol{F}$ 的训练,最终 $\boldsymbol{F}$ 的中间层低维特征 $\widetilde{\boldsymbol{F}}(\boldsymbol{x},\boldsymbol{w})$ 融合特征 $\boldsymbol{F}_T$ 和类别信息 $\boldsymbol{l}$,实现了对深度特征 $\boldsymbol{F}_T$ 的蒸馏.

2.2　学生网络和教师网络的结构

如图 1 所示,教师网络 $\boldsymbol{G}$ 为预先训练好的 VGG-16 模型,其由 13 个卷积层和 3 个全连接层组成,输入为 224×224 像素的彩色图像,取 fc 6 层输出的 4 096 维向量作为人脸特征.学生网络 $\boldsymbol{F}$ 采用 Darknet[14],输入图像大小为 96×96×3 像素,学生网络结构见表 2.表 2 中 conv 表示卷积层,maxPool 表示最大池化层,argPool 表示平均池化层,fc 表示全连接层.可知学生网络 $\boldsymbol{F}$ 共有 15 个卷积层、4 个最大池化层、1 个平均池化层、1 个全连接层及 1 个 softmax 层,其中 softmax 层只用于训练网络,提取特征时直接将 fc 层的 d(d=128)维向量作为输出.

表 2　学生网络结构

Tab.2　Structure of student networking

层数	层类型	核大小	步长	通道数	参数量
1	conv	3×3	1	16	432
2	maxPool	2×2	2	16	0
3	conv	3×3	1	32	4608
4	maxPool	2×2	2	32	0
5	conv	1×1	1	16	512
6	conv	3×3	1	128	18.4×10^3
7	conv	1×1	1	16	2048
8	conv	3×3	1	128	18.4×10^3
9	maxPool	2×2	2	128	0
10	conv	1×1	1	32	4096
11	conv	3×3	1	256	73.7×10^3
12	conv	1×1	1	32	8192
13	conv	3×3	1	256	73.7×10^3
14	maxPool	2×2	2	256	0
15	conv	1×1	1	64	16.4×10^3
16	conv	3×3	1	512	0.3×10^6
17	conv	1×1	1	64	32.8×10^3
18	conv	3×3	1	512	0.3×10^6
19	conv	3×3	1	4096	18.8×10^6
20	avgPool	6×6	1	4096	0
21	fc	—	—	128	0.52×10^6
22	softmax	—	—	8419	1.08×10^6

2.3　学生模型和教师模型分析

2.3.1　模型参数

教师网络 $\boldsymbol{G}$ 的最后 1 个卷积层和第 1 个全连接层之间的参数数量为 $7\times7\times512\times4096\approx103\times10^6$,即约 1.03 亿个参数,而模型总共有约 1.3 亿个参数,这说明 $\boldsymbol{G}$ 的最后 1 个卷积层和第 1 个全连接层之间提供了该模型的主要参数来源.学生网络 $\boldsymbol{F}$ 使用平均池化层代替卷积层后紧跟的全连接层,大大缩减了模型参数,并且其在卷积层中大量使用 3×3 和 1×1 的小卷积核,进一步减少模型参数的数量,最终模型参数(不包含最后的 softmax 层)为 19.5×10^6,而 $\boldsymbol{G}$ 网络模型参数为 138×10^6,即学生网络 $\boldsymbol{F}$ 相比教师网络 $\boldsymbol{G}$ 模型参数压缩了 7.1 倍.

2.3.2　特征维度

教师网络 $\boldsymbol{G}$ 输出 4 096 维特征,而学生网络 $\boldsymbol{F}$ 输出 128 维特征,占 $\boldsymbol{G}$ 的 3.125%,大幅减少了人脸特征库的尺寸,使其能够存储相较之前 32 倍的数据.同时,在特征匹配中,两个特征之间的相似度通过计算余弦距离得到,在特征归一化之后,两个 4 096维向量计算余弦距离需计算4 096次乘法和 4 095次加法,而两个 128 维向量只需计算 128 次乘法和 127 次加法,节省了 97.875%的时间和计算量,大幅提高匹配速度.

2.3.3　推理复杂度

教师网络 $\boldsymbol{G}$ 前向传播进行一次特征提取需要计算 1.55×10^{10}次,而学生网络 $\boldsymbol{F}$ 进行一次特征提取只需计算 7.63×10^8 次,速度提升 20.3 倍,节省了 95.1%的时间和计算量.因此,大大缩减了人脸检索在提取特征时花费的大量时间和计算力,通过用学生网络 $\boldsymbol{F}$ 代替教师网络 $\boldsymbol{G}$,有效降低了这部分的时间占用.

3　深度特征蒸馏的实验

本文实验采用预先训练好的 VGGFace 模型作为教师网络,在人脸识别公开数据集 UMDFaces[15] 上利用提出的深度特征蒸馏方法对学生网络进行训练,并将训练好的学生网络模型在 LFW 上进行评估,与原模型 VGGFace 进行多方位的比较,以验证本文方法的有效性.

3.1　实验准备

1)采用 UMDFaces 用来训练学生网络.UMDFaces[15]数据集共包含 8 277 人的 367 888 张图片及其标注,其中每张图片至少包含一张人脸,标注为人脸的姓名、位置、姿态和性别及 21 个关键点,之后作者对数据集进行扩充及删减,最终公开的数据集中包含 8 419 人的 362 743 张图片,本文只使用其中的 362 700 张图片.

2)LFW 用来评估学生网络的性能.LFW 包含 5 749人的 13 233 张图片及其标注,每张图片标注一个人名,其中大约 1 680 个人包含两个以上的人

脸.在人脸识别评测中,LFW 提供了标准的 3 000 对相同人脸及 3 000 对不同人脸的图片,待评估模型对其中每一对人脸判断是否为同一人,计算判断的正确率,即判断错误次数除以总数,作为最终的识别精度.

按照 5:1 的比例把 UMDFaces 中的 362 700 张图片随机分为训练集和测试集:训练集有 302 250 张图片,用于学生网络的训练;测试集有 60 450 张图片,用于评估学生网络的分类效果及泛化性能.

如图 1 所示,对 UMDFaces 中的每张图片,使用人脸对齐方法结合标注的人脸区域进行人脸的对齐,提取并缩放到 224×224 像素和 96×96 像素两种尺寸,分别进行保存,保存完成后将其中 224×224 像素的人脸图像输入预先训练好的 VGGFace 模型,提取并存储模型 fc 6 层的 4 096 维特征,作为教师网络的特征输出,供后面实验使用.对 LFW 中的 6 000 对图片做同样的人脸对齐处理并输出 96×96 像素的人脸图像.

3.2 实验过程

3.2.1 训练学生网络

将 UMDFaces 训练集中的 96×96 像素人脸图像和标签及对应的 4 096 维特征作为学生网络的输入,其中人脸图像和标签产生分类损失 CL,人脸图像和特征产生回归损失 RL,总损失 TL 为分类损失 CL 与回归损失 CL 的加权和,训练时取式(2)中蒸馏强度 $\alpha=1$,则 $TL=CL+RL$.依次调整学生网络 fc 层输出特征维度为 128、256、512、1 024,产生 4 个网络,接着使用 GPU 版 Tensorflow(一种深度学习工具包)对这些网络进行训练.训练参数方面,统一批尺寸(batch size)为 256,即每 256 个数据对(图像,类别, 特征)组成一个批次(batch),初始学习率为 0.001,使用批归一化(Batch Normalization)层加速网络的训练并抑制过拟合的发生,梯度更新采用收敛速度较快的 Adam 优化算法.同时使用传统方法即只使用分类损失,并在其他参数均保持不变的情况下训练一个 128 维学生网络作为对照组.

3.2.2 评测学生网络

分别用训练完成的所有学生网络 $\boldsymbol{F}$ 对 LFW 中的 6 000 对人脸图片提取特征,用特征 $\boldsymbol{F}(\boldsymbol{p}_1)$ 和 $\boldsymbol{F}(\boldsymbol{p}_2)$ 之间的余弦相似度来度量人脸 $\boldsymbol{p}_1$ 和 $\boldsymbol{p}_2$ 的相似性,计算提取特征对的余弦相似度作为对应人脸对的识别分数,再设置阈值遍历 [0,1] 区间,对识别分数进行分割,计算相应识别精度.取最大的识别精度,并记录对应的最优分割阈值.

3.3 实验结果

3.3.1 训练结果

以下只对 128 维学生网络进行叙述,其中 256 维、512 维和 1 024 维学生网络训练过程与 128 维类似.

1)迭代 15 000 步后,学生网络在训练集上分类准确率达到 91.41%,测试集上达到 49.22%;无蒸馏对照组在训练集上分类准确率达到 85.94%,测试集上达到 27.73%.

2)迭代 60 000 步后,学生网络在训练集上分类准确率达到 99.22%,测试集上达到 79.69%;无蒸馏对照组在训练集上分类准确率达到 98.83%,测试集上达到 70.41%.

128 维学生网络的分类准确率和损失随迭代次数变化的曲线如图 2 所示.从图 2(a)中可以看出,迭代 60 000 步后,学生网络在训练集和测试集上的准确率和损失基本达到饱和;测试集上分类准确率整体低于训练集,损失整体高于训练集,这说明网络还是存在过拟合的情况.图 2(b)损失变化与图 2(d)极为相似,这里因为总损失中回归损失为主体部分,图 2(c)中训练集上分类损失在迭代40 000步左右后基本趋于 0.图 2(d)中训练集上回归损失迭代 40 000 步后始终在 10 附近徘徊,这说明通过回归学习教师网络特征是个更复杂的任务,具有更大的优化价值和优化空间,只使用分类损失的无蒸馏对照组在训练集和测试集上的分类准确率一直低于同时使用分类损失和回归损失的网络,也验证了这一点.

3.3.2 评测结果

如图 3 所示,128 维学生网络的识别精度可以达到 95.25%,相比教师网络 VGGFace 的 98.95% 只降低了 3.7%,这是在模型压缩了 7.1 倍、特征压缩了 32 倍的情况下取得的结果,说明本文提出的深度特征蒸馏方法是有效的.作为对照组的无蒸馏 128 维学生网络的识别精度只达到 92.48%,比使用蒸馏的 128 学生网络的 95.25%降低了 2.77%,这说明深度特征蒸馏方法比传统训练方法更有效.

不同特征维度的最终识别精度相差并不大,更高维的特征维度反而可能会造成识别精度的稍微下降.注意到特征维度越大,最优分割阈值越小,这说明高维特征的余弦相似度相对较小,相同人脸的相似度较低,这侧面反映了高维特征拥有更多信息干扰,使用深度特征蒸馏能够去部分干扰.

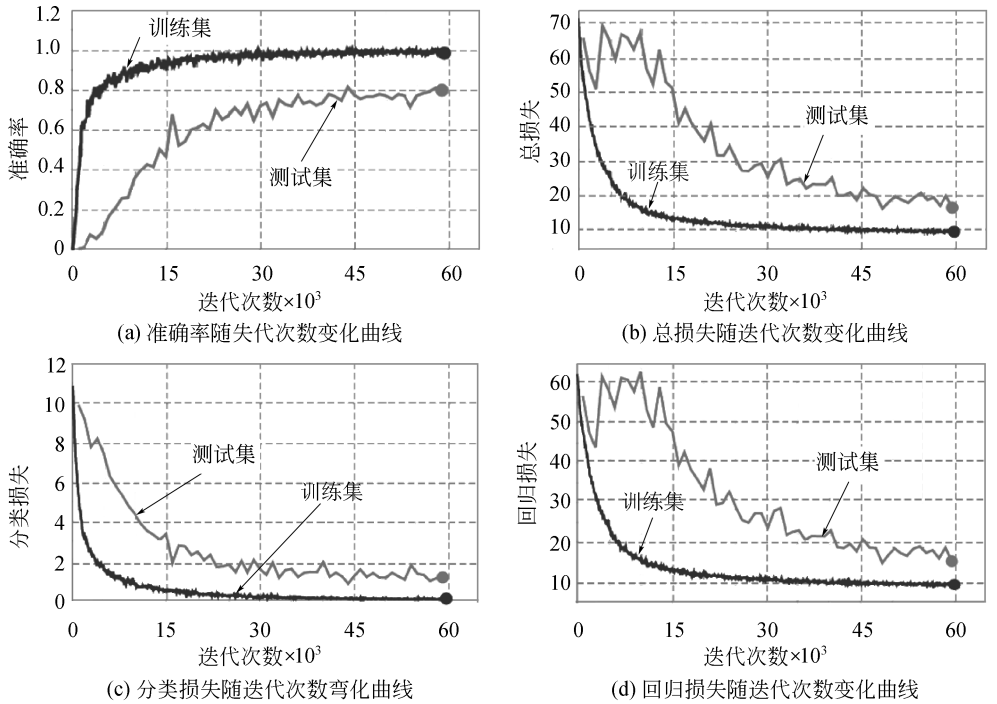

图 2　128 维学生网络训练结果

Fig.2　128 dimensional student network training results

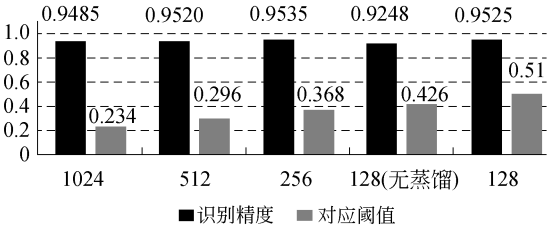

图 3　不同学生网络的评测结果

Fig.3　Results of different student networks

4　结论

1)本文作者提出深度特征蒸馏方法,以原有大模型作为教师网络,使用教师网络指导轻量级的学生网络的训练,通过联合回归与分类的多任务学习损失函数的设计,融合高维度深度特征回归及低分辨率人脸分类,能够在保证 95.25%识别精度的前提下,统一地完成深度模型与特征维度的压缩.

2)高维特征具有更多的信息干扰,使用深度特征蒸馏能够去除部分干扰,更注重发掘人脸的相似性,这在人脸检索等相似人脸发现任务中有着重要的作用.另外,使用深度特征蒸馏方法训练深度网络在人脸分类数据集 UMDFaces 上能够达到 79.69%的分类准确率,比传统训练方法的 70.41%取得了更高的分类准确率,深度特征蒸馏方法的训练方法更有效.

参考文献(References):

[1] PARKHI O M,VEDALDI A,ZISSERMAN A.Deep face recognition[C]//British Machine Vision Conference,2015,1(3):6－17.

[2] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10)[2017-09-01]. https://arxiv.org/abs/1409.1556.

[3] LEARNED-MILLER E,HUANG G B,ROYCHOWDHURY A,et al.Labeled faces in the wild: a survey[M].Springer International Publishing,2016:189－248.

[4] SUN Y,WANG X,TANG X.Deep learning face representation from predicting 10000 classes[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2014: 1891－1898.

[5] SCHROFF F,KALENICHENKO D,PHILBIN J.FaceNet: a unified embedding for face recognition and clustering[C]//IEEE Conference on Computer Vision and Pattern Recognition,2015:815－823.

[6] SUN Y,WANG X,TANG X.Deep learning face representation by joint identification-verification[C]//Advances in Neural Information Processing Systems,2014:1988－1996.

[7] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[EB/OL].(2015-03-09)[2017-09-01]. https://arxiv.org/abs/1503.02531.

[8] HE K,ZHANG X,REN S,et al.Deep residual learning for image recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition,2016:770－778.

(下转第 41 页)