DOI: 10.11992/tis.201906045
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20190828.1022.002.html

深度度量学习综述

刘冰[1,2], 李瑞麟[1,2], 封举富[1,2]
（1. 北京大学 信息科学技术学院, 北京 100871; 2. 北京大学 机器感知与智能教育部重点实验室, 北京 100871）

摘　要: 深度度量学习已成为近年来机器学习最具吸引力的研究领域之一, 如何有效的度量物体间的相似性成为问题的关键。现有的依赖成对或成三元组的损失函数, 由于正负样本可组合的数量极多, 因此一种合理的解决方案是仅对训练有意义的正负样本采样, 也称为“难例挖掘”。为减轻挖掘有意义样本时的计算复杂度, 代理损失设置了数量远远小于样本集合的代理点集。该综述按照时间顺序, 总结了深度度量学习领域比较有代表性的算法, 并探讨了其与 softmax 分类的联系, 发现两条看似平行的研究思路, 实则背后有着一致的思想。进而文章探索了许多致力于提升 softmax 判别性能的改进算法, 并将其引入到度量学习中, 从而进一步缩小类内距离、扩大类间距, 提高算法的判别性能。

关键词: 深度度量学习; 深度学习; 机器学习; 对比损失; 三元组损失; 代理损失; softmax 分类; 温度值
中图分类号: TP181　　**文献标志码:** A　　**文章编号:** 1673−4785(2019)06−1064−09

中文引用格式: 刘冰, 李瑞麟, 封举富. 深度度量学习综述 [J]. 智能系统学报, 2019, 14(6): 1064–1072.
英文引用格式: LIU Bing, LI Ruilin, FENG Jufu. A brief introduction to deep metric learning[J]. CAAI transactions on intelligent systems, 2019, 14(6): 1064–1072.

A brief introduction to deep metric learning

LIU Bing[1,2], LI Ruilin[1,2], FENG Jufu[1,2]
(1. School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China; 2. Key Laboratory of Machine Perception (MOE), Peking University, Beijing 100871, China)

Abstract: Recently, deep metric learning (DML) has become one of the most attractive research areas in machine learning. Learning an effective deep metric to measure the similarity between subjects is a key problem. As to existing loss functions that rely on pairwise or triplet-wise, as training data increases, and since the number of positive and negative samples that can be combined is extremely large, a reasonable solution is to sample only positive and negative samples that are meaningful for training, also known as Difficult Case Mining. To alleviate computational complexity of mining meaningful samples, the proxy loss chooses proxy sets that are much smaller than the sample sets. This review summarizes some algorithms representative of DML, according to the time order, and discusses their relationship with softmax classification. It was found that these two seemingly parallel research methods have a consistent idea behind them. This paper explores some improved algorithms that aim to improve the softmax discriminative performance, and introduces them into metric learning, so as to further reduce intra-class distance, expand inter-class distance, and, finally, improve the discriminant performance of the algorithm.

Keywords: deep metric learning; deep learning; machine learning; contrastive loss; triplet loss; proxy loss; softmax classification; temperature

在机器学习领域, 距离 (distance) 的概念从诞生之日起就一直有着广泛的应用。它提供了一种数据间相似性的度量, 即距离近的数据要尽可能相似, 距离远的数据要尽可能不同[1-2]。这种相似性学习的思想应用在分类问题上是著名的最近

收稿日期: 2019−06−24.　**网络出版日期:** 2019−08−28.
基金项目: 国家自然科学基金重点项目 (61333015).
通信作者: 封举富. E-mail: fjf@cis.pku.edu.cn.

邻 (nearest neighbors, NN)[3] 分类, 该方法将待测样本的类别分配为距其最近的训练样本的类别。这种最近邻分类思想催生了距离度量学习 (distance metric learning) 的产生[4]。

欧氏距离作为一种简洁有效的度量工具得到了度量学习算法的广泛青睐, 然而, 单一形式的距离度量无法普适所有实际问题。因此, 度量学习希望能够结合数据自身特点, 学习一种有效的度量方式, 用于求解目标问题。

早期度量学习算法的产生, 极大地改善了基于距离分类器的分类性能[5-6]、基于距离聚类的无监督问题[1] 以及特征降维[7] 的表现。而后, 随着深度学习[8-14] 的飞速发展, 结合深度神经网络在语义特征抽取、端到端训练优势的深度度量学习, 逐步进入人们的眼帘。

相比于经典度量学习, 深度度量学习可以对输入特征做非线性映射, 在计算机视觉领域得到了广泛的应用, 例如: 图像检索[15-16]、图像聚类[17]、迁移学习[18-19]、验证[20]、特征匹配[21]。除此之外, 对于一些极端分类[22-24] 任务 (类别数目很多, 但每类仅有几个样本), 深度度量学习仍有不错的表现。例如, 基于深度度量学习, FaceNet[25] 在 8 M 个个体、260 M 张图像的人脸识别任务中, 表现结果已经超越了人类水平。

标准的深度度量学习通过挖掘 2 个[26] 或 3 个[25] 正负样本对来约束类内距、扩大类间距。这为训练样本的采样带来了挑战: 由于训练样本数量极多, 因此只能挖掘有意义的样本参与训练。若负样本选取过难, 则易导致训练不稳定; 若选取过简单, 则易导致损失函数无梯度, 不利于模型的收敛。

代理损失[16] 的提出为每种类别分配了一个代理点, 由于代理点数量远远小于样本集合, 因而可以完整存储起来, 在训练过程中参与梯度回传, 从而为训练过程提供了全局的语义信息, 取得了更好的结果。

此外, 我们发现改进后的代理近邻损失与标准的分类任务有些相似: 一方面, 损失函数同时优化所有类别实现缩小类内距、扩大类间距; 另一方面, 如果我们移除了 softmax 线性变换的偏置项[27], 权重 $\boldsymbol{W}$ 的物理含义即为该类别的代理点。

标准分类任务结合 softmax 与交叉熵建立损失函数, 可以输出特征向量分别属于每一类的概率。然而 softmax 不具有较强的判别性, 因而很多算法提出温度值概念[27-29], 从特征梯度层面改进其性能, 具体细节我们将在后文展开综述。

度量学习起源于分类问题的最近邻思想, 经历了逐步演化最终至代理近邻损失函数。已有文献 [27] 证明移除偏置项、正则化输入特征 $\boldsymbol{x}$ 和权重 $\boldsymbol{W}$ 后的 softmax 分类任务可视为基于代理点的度量学习。考虑到代理近邻损失与 softmax 的相关性——softmax 的权值可视为该类别学到的代理点, 我们借鉴了带温度值的 softmax 分类思想, 将温度值引入代理损失, 从而进一步扩大类间距, 提高了度量学习的判别性能。至此, 我们将度量学习与分类这两条看似独立的分支建立了联系, 深入挖掘出二者背后统一的思想, 可谓"殊途同归"。

1　深度度量学习

在人脸识别、指纹识别等开集分类的任务中, 类别数往往很多而类内样本数比较少, 在这种情况下基于深度学习的分类方法常表现出一些局限性, 如缺少类内约束、分类器优化困难等。这些局限可以通过深度度量学习来解决: 深度度量学习通过特定的损失函数, 学习到样本到特征的映射 $f_\theta(\cdot)$。在该映射下, 样本特征间的度量 $d_{(i,j)}$(通常为欧式距离 $\left\|f_\theta(x_i)-f_\theta(x_j)\right\|_2$) 便可以反映样本间的相似程度: 类内样本对应的特征距离更近, 类间样本对应的特征距离更远。

1.1　对比损失

对比损失 (contrastive loss)[26,30] 是深度度量学习的开篇之作, 它首次将深度神经网络引入度量学习。在此之前, 经典度量学习最早应用于聚类问题[1], 如: 局部线性嵌入 (locally linear embedding, LLE)[31]、Hessian 局部线性嵌入 (Hessian LLE)[32]、主成分分析 (principal component analysis, PCA)[33] 等。它们通过定义样本 $\boldsymbol{x}$ 和样本 $\boldsymbol{y}$ 之间的马氏距离 $d(\boldsymbol{x},\boldsymbol{y})=(\boldsymbol{x}-\boldsymbol{y})^{\mathrm{T}}\boldsymbol{M}(\boldsymbol{x}-\boldsymbol{y})$, 约束相似样本马氏距离小, 不相似样本马氏距离大。其中 $\boldsymbol{M}$ 为马氏距离, 为 $d\times d$ 的半正定矩阵。相比于欧氏距离, 马氏距离考虑了特征之间的权重与相关性, 且凸问题易被优化, 因而得到了广泛应用。

然而, 传统方法主要存在两个弊端: 一是依赖于原始输入空间进行距离度量; 二是不能很好地映射与训练样本关系未知的新样本的函数。作者利用深度学习特征提取的优势, 将原始的输入空间映射到欧氏空间, 直接约束类内样本的特征尽可能接近而类间样本的特征足够远如式 (1):

$$l_{\text{contrast}}(\boldsymbol{X}_i,\boldsymbol{X}_j)=\boldsymbol{y}_{i,j}d_{i,j}^2+(1-\boldsymbol{y}_{i,j})[\alpha-d_{i,j}^2]_+ \tag{1}$$

其中, 若 $\boldsymbol{X}_i$ 与 $\boldsymbol{X}_j$ 类别编号相同则 $\boldsymbol{y}_{(i,j)}=1$, 否则 $\boldsymbol{y}_{(i,j)}=0$。$d_{(i,j)}$ 即为欧式距离, α 控制类间样本足够

远的程度。

1.2　三元组损失

对比损失仅仅只约束类内对的特征尽量近而类间对的特征尽量远，三元组损失 (triplet loss)[5,25] 在对比损失的基础上进一步考虑了类内对与类间对之间的相对关系：首先固定一个锚点样本 (anchor),希望包含该样本的类间对 (anchor-negative) 特征的距离能够比同样包含该样本的类内对 (anchor-positive) 特征的距离大一个间隔 (margin)，如式 (2)：

$$l_{\text{triplet}}(\boldsymbol{X}_a,\boldsymbol{X}_b,\boldsymbol{X}_n)=[d^2_{a,p}+m-d^2_{a,n}]_+ \tag{2}$$

式中 $\boldsymbol{X}_a$、$\boldsymbol{X}_p$、$\boldsymbol{X}_n$ 分别为锚点样本、与锚点样本同类的样本以及与锚点样本异类的样本；m 即为间隔。

然而，对于三元组的选取，采样策略是至关重要的：假设训练集样本数为 n，那么所有的三元组组合数为 $O(n^3)$，数量非常庞大。其中存在大量的平凡三元组，这些平凡三元组类间对的距离已经比类内对的距离大一个间隔，它们对应的损失为 0。简单的随机采样会导致模型收敛缓慢，特征不具有足够的判别性[14, 34-35]。因此一种合理的解决方案是仅挖掘对训练有意义的正负样本，也称为“难例挖掘”[25,36-39]。例如：HardNet[36] 旨在在一个训练 batch 中挖掘一些最难的三元组。然而如果每次都针对锚点样本挖掘最困难的类间样本，模型又很容易坍缩。因此，文献 [25] 提出了一种半难例 (semi-hard) 挖掘的方式：选择比类内样本距离远而又不足够远出间隔的类间样本来进行训练。

1.3　提升结构化损失

由于三元组损失一次采样 3 个样本，虽然能够同时考虑类内距、类间距以及二者的相对关系，但该损失没有充分利用训练时每个 batch 内的所有样本，因此文献 [18] 提出在一个 batch 内建立稠密的成对 (pair-wise) 的连接关系，具体实现是：对于每一个类内对，同时选择两个难例，一个距离 $\boldsymbol{x}^a$ 最近，一个距离 $\boldsymbol{x}^p$ 最近。提升结构化损失 (lifted structured loss)[18] 对应的损失函数为

$$\begin{aligned} J&=\frac{1}{2|\hat{P}|}\sum_{(i,j)\in\hat{P}}\max(0,J_{i,j})^2\\ J_{i,j}&=\max(\max_{(i,j)\in\hat{N}}\alpha-D_{i,k},\max_{(j,l)\in\hat{N}}\alpha-D_{j,l})+D_{i,j}\end{aligned} \tag{3}$$

式中：$\hat{P}$ 为 batch 内所有的正样本对集合；$\hat{N}$ 为 batch 内所有的负样本对集合。

这种设计结构性损失函数，以在一个训练 batch 中考虑所有可能的训练对或三元组并执行“软化的”难例挖掘在文献 [40] 中也得到了相似的应用。

1.4　多类 N 元组损失

Sohn 等[15] 将对比损失和三元组收敛比较慢的原因归结于训练时每次只挖掘一个负样本，缺少了与其他负样本交互过程。因此他们提出多类 N 元组损失 (multi-class N-pair loss)[15]：同类样本的距离不应每次只小于一组类间距，而应同时小于 n-1 组类间距离，从而实现类内对相似度显著高于所有类间对相似度。损失函数的设计借鉴了 (neighborhood component analysis, NCA)[6] 的表达形式，具体如式 (4) 所示：

$$\begin{aligned} l(\boldsymbol{X},\boldsymbol{y})=\frac{-1}{|\boldsymbol{P}|}\sum_{(i,j)\in P}\log\frac{\exp\{\boldsymbol{S}_{i,j}\}}{\exp\{\boldsymbol{S}_{i,j}\}+\sum\limits_{k:\boldsymbol{y}[k]\neq\boldsymbol{y}[i]}\exp\{\boldsymbol{S}_{i},\boldsymbol{k}\}}+\\ \frac{\lambda}{m}\sum_{i}^{m}\|f(\boldsymbol{X}_i)\|_2\end{aligned} \tag{4}$$

式中：i,j 表示同类样本；k 表示不同类样本；$\boldsymbol{P}$ 为一个 batch 内的所有正样本；m 为 batch size 的大小。另一方面，为了使分类面只与向量 $\boldsymbol{X}_i$ 的方向有关，与模长无关，作者对一个 batch 内的所有输入特征 $\boldsymbol{X}_i$ 利用 L_2 正则化。

1.5　成对聚类损失

由于三元组损失[25] 在锚点选取时的任意性，因此有些不满足类间距>类内距+间隔的样本，可能并没有被挖掘到，如图 1 所示。

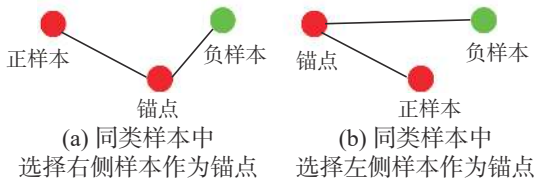


图 1　构建三元组时的两种不同方法[41]

Fig. 1　Two different cases when building triplets[41]

若样本以左侧方式组合，则负样本很易被检测到，从而距离得到优化；但若以右边方式设置锚点、正样本，则负样本由于满足约束，因而 loss 为 0，导致同类物体的距离没有被拉近，这一定程度上减缓了收敛的速度。这说明三元组损失对锚点的选取十分敏感，考虑到相似样本应该聚集成簇[42]，不同类样本应保持相对较远，因此他们[36] 提出成对聚类损失函数 (coupled clusters loss, CCL) 为同类样本估计了一个类内中心 $\boldsymbol{c}^p$：

$$\boldsymbol{c}^p=\frac{1}{N^p}\sum_{i}^{N^p}f(\boldsymbol{X}_i^p) \tag{5}$$

从而希望所有的正样本 $\boldsymbol{X}_i^p$ 到聚类中心 $\boldsymbol{c}^p$ 的距离加间隔 α 小于其他类间样本到聚类中心 $\boldsymbol{c}^p$ 的距离，对应的损失函数为

$$\begin{aligned} L(\boldsymbol{W},\boldsymbol{X}^p,\boldsymbol{X}^n)=\\ \sum_{i}^{N^p}\frac{1}{2}\max\{0,\|f(\boldsymbol{X}_i^p)-\boldsymbol{c}^p\|^2+\alpha-|f(\boldsymbol{X}_*^n)-\boldsymbol{c}^p\|^2\}\end{aligned} \tag{6}$$

式中: $\boldsymbol{X}_i^p$ 为正样本; X_*^n 为负样本; N^p 为同类正样本的数目; $\boldsymbol{c}^p$ 为正样本的聚类中心。

1.6 中心损失

处理开集识别问题的深度特征, 不仅需要具有可分离性 (separable), 还应具有判别性 (discriminative)。可判别性特征可以很好地容忍类内变化、分离类间变化, 进而可以应用在最近邻 (nearest neighbor, NN)[3] 和 k 近邻 (k-nearest neighbor, k-NN)[43] 算法中。然而, softmax loss 仅能约束特征具有可分离性、不具有判别性, 因此为 CNN 设计一个有效的损失函数是极为重要的。

中心损失 (center loss)[44] 结合了成对聚类损失 (CCL) 和 softmax loss 的优势, 用 CCL 约束类内, softmax 约束类间。具体做法是: 为每一类特征学习一个聚类中心, 随着训练的进行, 同步更新类内中心以及最小化特征与对应中心的距离。将聚类的 loss 与 softmax 联合训练, 利用超参平衡两个监督信号的权重。主观上, softmax 损失可以分离不同类别特征, center loss 可以使同类特征聚在一起中心点周围, 从而扩大类间距、缩小类内距, 学到了更具有判别性的深度特征。对应的损失函数为

$$L = L_S + L_C = \sum_{i=1}^{m} \log \frac{\exp\{\boldsymbol{W}_{y_i}^{\mathrm{T}} \boldsymbol{X}_i + \boldsymbol{b}_{y_i}\}}{\sum_{j=1}^{n} \exp\{\boldsymbol{W}_{y_i}^{\mathrm{T}} \boldsymbol{X}_i + \boldsymbol{b}_{y_i}\}} + \frac{\lambda}{2} \sum_{i}^{m} |\boldsymbol{X}_i - \boldsymbol{c}_{y_i}\|_2^2 \tag{7}$$

其中 $\boldsymbol{c}_{y_i}$ 表示第 y_i 类深度特征中心。

1.7 设备定位损失

Oh 等[45] 认为, 当前存在的大多数方法[15, 18, 25, 46] 通常只关注数据的局部区域 (如: 二元组、三元组或 n 元组), 并没有考虑到全局的结构信息, 因而降低了聚类和检索的表现。

作者指出, 一旦正样本对距离较远且二者之间被其他类别的负样本间隔开, 那么正样本对间相互吸引的梯度信号被负样本相互排斥的梯度信号所超过, 从而同类样本很难聚成一类, 而被错误地分开成了两类。因此, 他们提出一组聚类损失——设备定位损失 (facility location loss)[45] 来解决这个问题。

对于一组输入样本 $\boldsymbol{X} = \{\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_n\}$, 以 k-mediods[47] 思想为基础, 我们期望从中选取的聚类中心 $\{\boldsymbol{X}_S\}$(其中 $S \subset \{1, \cdots, n\}$) 满足在特征空间中与同类样本距离尽可能接近:

$$F(\boldsymbol{X}, \boldsymbol{S}; \theta) = -\sum_{i \in |\boldsymbol{X}|} \min_{j \in S} \|f(\boldsymbol{X}_i; \theta) - f(\boldsymbol{X}_j; \theta)\| \tag{8}$$

式 (8) 也被称为设备定位函数 (facility location function), 现已被广泛应用于数据求和[48-49] 与聚类。

由于最大化式 (8) 是一个 NP-hard 问题[50-51], 因此作者通过对子模块的贪婪求解, 找到了一个完备的优化下界, 复杂度为 $O(1-\frac{1}{e})$。具体方法是: 通过设计一个打分函数 $\tilde{F}$, 基于真实类别标签 y^*来评估聚类的好坏, 对应的公式为

$$\tilde{F}(\boldsymbol{X}, \boldsymbol{y}^*; \theta) = -\sum_{i \in |\boldsymbol{Y}|} \max_{j \in \{i: y^*[i]=k\}} F(\boldsymbol{X}_{\{i: y^*[i]=k\}}, \{j\}; \theta) \tag{9}$$

其中, $\{i : y^*[i] = k\}$ 表示当真实类别标签为第 k 类时, 对应的 $\{1, 2, \cdots, n\}$的元素构成的集合。

由于希望打分函数 $\tilde{F}$ 越大越好, 因此借鉴三元组损失的间隔思想, F 比 $\tilde{F}$ 相差一个间隔 $\Delta(\boldsymbol{y}, \boldsymbol{y}^*)$, 即

$$l(\boldsymbol{X}, \boldsymbol{y}^*) = [\max_{S \subset \{1, \cdots, n\}} \{F(\boldsymbol{X}, \boldsymbol{S}; \theta) + \gamma \Delta(\boldsymbol{y}, \boldsymbol{y}^*)\} - \tilde{F}(\boldsymbol{X}, \boldsymbol{y}^*; \theta)] \tag{10}$$

其中,

$$\begin{aligned} \boldsymbol{y}[i] &= \arg\min_j \|f(\boldsymbol{X}_i; \theta) - f(\boldsymbol{X}_{\{j \in S\}}; \theta)\| \\ \Delta(\boldsymbol{y}, \boldsymbol{y}^*) &= 1 - \mathrm{NMI}(\boldsymbol{y}, \boldsymbol{y}^*) \\ \mathrm{NMI}(\boldsymbol{y}_1, \boldsymbol{y}_2) &= \frac{\mathrm{MI}(\boldsymbol{y}_1, \boldsymbol{y}_2)}{\sqrt{H(\boldsymbol{y}_1) H(\boldsymbol{y}_2)}} \end{aligned} \tag{11}$$

其中, NMI 表示正则化互信息 (normalized mutual information, NMI)[52]。由于这种聚类方法在特征空间中有一个全局的感受野, 因此可以解决局部最优的问题。聚类的损失函数可以约束全局样本向类内中心靠拢、间隔项中的 NMI 矩阵可以使不同类别远离。

1.8 代理损失

为了克服三元组样本对采样困难的问题, 代理损失[16] 提出了一种用小规模的代理点来代替大规模的原始样本点的方法: 将原始样本用代理点来近似, 这样约束类内对和类间对的距离便可以转化为约束锚点样本与同类样本对应代理点和锚点与异类样本对应代理点的距离。随着训练的进行, 样本的特征与代理点都获得更新。

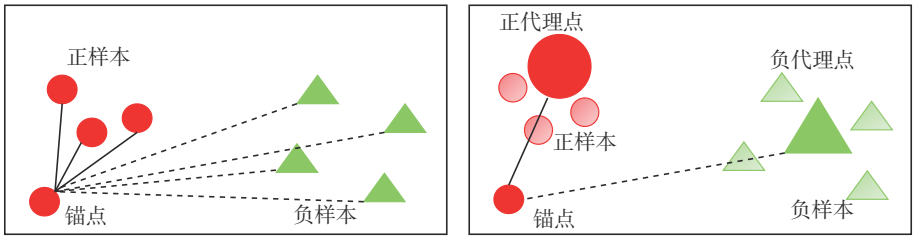
假设原始样本点和代理点的集合分别为 $\boldsymbol{X}$, $\boldsymbol{P}$, 且 $|\boldsymbol{X}|=n$, $|\boldsymbol{P}|=m$, $m \ll n$。有如下两种分配代理点的方式: 1) 动态分配策略: 选取距该样本最近的代理点作为代表该样本的代理点 (式 (12)); 2) 静态分配策略: 选取与样本类别数相同的代理点数目, 某一类样本被固定分配至对应该类别的代理点。

$$\boldsymbol{p}(x) = \arg\min_{p \in P} d^2(\boldsymbol{X}, \boldsymbol{p}) \tag{12}$$

代理损失借鉴了近邻成分分析 (neighborhood component analysis, NCA)[6] 的思路, 希望锚点样本与其同类代理点的距离尽可能近而与其异类代理点的距离尽可能远:

$$l^{\text{proxy}}(\boldsymbol{X},\boldsymbol{P})=-\log\left(\frac{\exp(-d^2(\boldsymbol{X},\boldsymbol{p}(x)))}{\sum_{z\in\boldsymbol{P}\setminus\{\boldsymbol{p}(x)\}}\exp(-d^2(\boldsymbol{X},z)}\right)\tag{13}$$

图 2 展示了三元组损失与代理损失在优化时的差别，代理点的设定使得“样本对”的数量大大减少：对于每一个锚点样本，图 (a) 中可以组成 12 个三元组，而图 (b) 中仅存在 2 个锚点–代理点对，样本挖掘的困难很大程度被克服了。



(a) 三元组损失　　(b) 代理损失

图 2　三元组损失 VS 代理损失示意图

Fig. 2　Triplet loss VS proxy loss

另一方面，作者也论证了代理损失与三元组损失的优化目标是一致的，通过三角不等式证明了代理损失是三元组损失的上界。

1.9　其他损失

除此之外，最近还有一些使用深度网络进行度量学习的工作。Hershey 等[17] 在二值化的真实标签和成对估计的亲和度矩阵之间的残差上使用了 Frobenius 范数；他们将此应用于语音谱信号聚类。然而，直接使用 Frobenius 范数是次优的，因为它忽略了亲和矩阵是正定的这一事实。为了克服这个问题，矩阵反向传播[53] 首先将真实和预测的亲和度矩阵投影到欧氏空间。然而，这种方法需要计算数据矩阵的特征值分解，具有数据样本量三次方的时间复杂度，因此对于大数据问题不适用。Ranked loss[54] 则是从秩的角度优化距离度量。

2　深度度量学习与 softmax 分类

利用深度神经网络的倒数第二层 (也叫瓶颈层) 特征，搭配 softmax 与交叉熵损失训练得到的分类器，同样适用于许多基于深度度量学习的应用[55]，例如诸如：物体识别[17, 51-58]、人脸验证[59-61]、手写数字识别[62] 等。然而，分类器训练与度量学习的目标实际是不同的[29]。前者旨在寻找最佳分类面，而后者旨在学习特征嵌入，使得相同类别的样本嵌入是紧凑的，而不同类别的样本嵌入是远离的。这促使我们研究度量学习和分类器训练之间的关系。

2.1　代理损失与 softmax 的关系

如果我们将代理近邻损失式 (13) 的分母中加入正样本项，则变为

$$l(\boldsymbol{X},\boldsymbol{P})=-\log\left(\frac{\exp(-d^2(\boldsymbol{X},\boldsymbol{p}(x)))}{\sum_{z\in\boldsymbol{P}}\exp(-d^2(\boldsymbol{X},z))}\right)\tag{14}$$

这样 log 函数内的式子可以看成是样本被分配到其对应代理点的概率，这里用 q 来表示概率即：

$$q(\boldsymbol{p}_i|\boldsymbol{X})=\frac{\exp(-d^2(\boldsymbol{X},\boldsymbol{p}_i))}{\sum_{j=1}^{m}\exp(-d^2(\boldsymbol{X},\boldsymbol{p}_j))}\tag{15}$$

这样式 (14) 可以看作上述后验概率结合交叉熵损失以及类别标签所得。

代理损失与 softmax 不同之处在于，softmax 将样本经过线性变换 $\boldsymbol{wx}+\boldsymbol{b}$ 之后进行归一化作为后验概率，而此处则是将样本与对应代理点的距离 $d^2(x, p_i)$ 进行归一化作为后验概率。如果我们将样本特征以及代理点的模长固定为常数 s，有：

$$d^2(\boldsymbol{X},\boldsymbol{p}_i)=\|\boldsymbol{X}\|_2^2+\|\boldsymbol{p}_i\|_2^2-2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{p}_i=2s^2-2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{p}_i\tag{16}$$

代入到式 (15) 中：

$$q(\boldsymbol{p}_i|\boldsymbol{X})=\frac{\exp(2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{p}_i-2s^2)}{\sum_{j=1}^{m}\exp(2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{p}_j-2s^2)}=\frac{\exp(2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{p}_i)}{\sum_{j=1}^{m}\exp(2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{p}_j)}\tag{17}$$

可以看作将线性变换参数的模长固定且去掉偏置项的 softmax，这与 Zhai 等[27] 的发现也是一致的。由此，我们在度量学习中的代理损失与 softmax 分类之间建立了联系。

2.2　温度放缩

softmax 损失函数对不同类别的特征有着较好的分离性，却不具有足够的判别性。因此，现阶段的工作提出了几种变体[44, 63-71] 来增强其判别性。最早在 2015 年，Hinton 为解决模型压缩[72]、知识迁移等问题，提出了温度值[28]的概念。他认为不同类别间的关系不应是非 0 即 1 的问题 (如：将猫误判为狗的损失直观上应该要比将猫误判为汽车的损失小)，因此，粗暴地使用 one-hot 编码丢失了类间和类内关联性的额外信息。由此作者提出带温度值的 softmax 函数，弱化不同类别之间的差异。损失函数：

$$L_{\text{soft}} = -\sum_{i=1}^{K} \boldsymbol{p}_i \log \boldsymbol{q}_i = -\sum_{i=1}^{K} \boldsymbol{p}_i \log \frac{\mathrm{e}^{z_i/T}}{\sum_j \mathrm{e}^{z_j/T}} \tag{18}$$

式中: z_i 为 logit, 即 $z_i = \boldsymbol{W}\boldsymbol{X} + b$; $\boldsymbol{p}_i$ 为软化后的类别标签; $\boldsymbol{q}_i$ 为压缩模型的输出。由式 (14) 可知, 当 T=1 时, 恢复到普通的 softmax 变换。文中令 T>1, 就得到了软化后的 softmax。这一思想在 Zhai 等[27] 的实验中得到了进一步验证。

文献 [27] 通过移除最后一层的偏置项, 并对权重与输入特征施加 L_2 正则, 从而完成了将任意分类网络向基于代理损失的度量学习转换的过程。考虑到在高维空间中, 单位球面上两个样本点之间的距离接近正态分布 $N\left(\sqrt{2}, \frac{1}{2\dim}\right)$, 其中 dim 表示特征维数[28]。为了使网络对类别差异变化更敏感, 他们引入温度值 σ 放缩余弦距离:

$$L_{\text{cls}}(\boldsymbol{X}, \boldsymbol{p}_y, \boldsymbol{p}_Z, \sigma) = -\log\left(\frac{\exp(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{p}_y/\sigma)}{\sum_{\boldsymbol{p}_z \in (\boldsymbol{p}_z \cup \boldsymbol{p}_y)} \exp(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{p}_z/\sigma)}\right) \tag{19}$$

作者在 CUB-200-2011[73] 数据集上, 探索了不同温度参数 σ 对实验结果的影响, 如表 1 所示。

表 1　不同温度值下 Recall@1 结果

Table 1　Recall@1 on CUB-200-2011 dataset across varying temperature

Tmp	0.005	0.01	0.03	0.05	0.1	0.5	1.0
R@1	19.3	38.8	57.0	61.3	61.6	54.8	50.5

由表 1 可知, 与 Hinton[28] 的思想一致, 当温度值 σ<1 时, 类别间差异放大, 学到的特征具有更强的判别性, 当温度值 σ=1 时, 判别性能急剧下降; 类似地, 温度值太低也会降低性能。然而, 针对这一现象, 本文作者并未做出合理的解释。

Zhang 等[29] 从样本特征梯度的角度分析了温度值如何影响特征分布。为方便起见, 作者令 $\alpha = 1/T$, 即:

$$q(m|x) = \log \frac{\mathrm{e}^{z_m/T}}{\sum_{j=1}^{M} \mathrm{e}^{z_j/T}} = \log \frac{\mathrm{e}^{\alpha z_m}}{\sum_{j=1}^{M} \mathrm{e}^{\alpha z_j}} \tag{20}$$

式中: m 为类别数, $m \in \{1, 2, \cdots, M\}$。$q(m|x)$ 为模型 softmax 预测概率输出。假设 $p(m|x)$ 为训练样本真实分布, 则可得到交叉熵损失函数:

$$L(\boldsymbol{X}, \alpha) = -\sum_{m=1}^{M} \log(q(m|\boldsymbol{X}, \alpha)) p(m|\boldsymbol{X}) \tag{21}$$

损失函数 $L(x, \alpha)$ 对 logit z_m 的梯度为

$$\frac{\partial L}{\partial z_m} = \alpha(q(m|\boldsymbol{X}) - p(m|\boldsymbol{X})) \tag{22}$$

进而得到对特征 $\boldsymbol{f}$ 的梯度:

$$\frac{\partial L}{\partial \boldsymbol{f}} = \sum_{m=1}^{M} \frac{\partial L}{\partial z_m} \frac{\partial z_m}{\partial \boldsymbol{f}} = \alpha \sum_{m=1}^{M} (q(m|\boldsymbol{X}) - p(m|\boldsymbol{X}))\boldsymbol{w}_m \tag{23}$$

由此, 作者观察到不同的 α 值为不同的样本 (难例、边界样本且被分对、内部易被分对样本) 分配不同大小的梯度, 从而改变了最终特征表达的分布。换句话说, 随着 α 的增大 (温度值 T 减小), 难例的梯度增大、而易分对样本梯度先增大后减小, 因而选择合适的 α 实际是在不同样本的梯度间做一种权衡。

3　结束语

本文综述了最近的一系列具有代表性的深度度量学习算法的文章, 并探讨了其与 softmax 分类的关系。深度度量学习最早源于对比损失, 由于未同时兼顾类内与类间的相对关系, 进而衍生出改进后的三元组损失。由于成对的二元组、三元组样本数量极多, 难例挖掘、半难例挖掘等采样策略针对正负样本采样问题起着关键作用。为减轻采样负担, 许多结构化的损失函数, 利用 batch 内更丰富的样本间结构关系来设计损失函数, 约束特征。还有一些基于聚类思想的损失函数, 如: 中心损失、代理损失等, 为每类样本学习一个代理点, 从而大大减少了类间样本数量, 使模型更易优化。

综述中, 我们发现搭配 softmax 与交叉熵损失训练得到的分类器, 同样适用于许多基于深度度量学习的任务, 这促使我们研究度量学习和分类器训练之间的关系。随着研究的深入, 我们发现代理损失与 softmax 分类这两条看似平行的研究思路, 实则背后有着一致的思想。针对 softmax 判别性不高的缺点, 许多算法引入温度值概念, 对原始的 softmax logit 作出改造, 并取得了很好的实验结果。在未来的研究中, 我们希望继续深入探索二者之间的关系。例如, 我们可以将 softmax 变体中间隔 margin 的概念引入代理近邻损失, 从而进一步缩小类内距离、扩展类间距。

参考文献:

[1] XING E P, NG A Y, JORDAN M I, et al. Distance metric learning, with application to clustering with side-information[C]//Proceedings of the 15th International Conference on Neural Information Processing Systems. Cambridge, USA, 2002: 521–528.

[2] LOWE D G. Similarity metric learning for a variable-kernel classifier[J]. Neural computation, 1995, 7(1): 72–85.

[3] COVER T M, HART P. Nearest neighbor pattern classific-

ation[J]. IEEE transactions on information theory, 1967, 13(1): 21–27.

[4] SUÁREZ J L, GARCÍA S, HERRERA F. A tutorial on distance metric learning: mathematical foundations, algorithms and software[J]. arXiv preprint arXiv: 1812.05944, 2018.

[5] WEINBERGER K Q, SAUL L K. Distance metric learning for large margin nearest neighbor classification[J]. Journal of machine learning research, 2009, 10: 207–244.

[6] GOLDBERGER J, ROWEIS S, HINTON G, et al. Neighbourhood components analysis[C]//Proceedings of the 17th International Conference on Neural Information Processing Systems. Vancouver, British Columbia, Canada, 2004: 513–520.

[7] VAN DER MAATEN L, POSTMA E, VAN DEN HERIK J. Dimensionality reduction: a comparative[J]. Journal of machine learning research, 2009, 10: 66–71.

[8] 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, USA, 2012: 1097–1105.

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

[10] SZEGEDY C, LIU Wei, JIA Yangqing, et al. Going deeper with convolutions[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 1–9.

[11] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770–778.

[12] HUANG Gao, LIU Zhuang, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 4700–4708.

[13] HU Jie, SHEN Li, SUN Gang. Squeeze-and-excitation networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 7132–7141.

[14] CHECHIK G, SHARMA V, SHALIT U, et al. Large scale online learning of image similarity through ranking [J]. Journal of machine learning research, 2010, 11: 1109–1135.

[15] SOHN K. Improved deep metric learning with multi-class n-pair loss objective[C]//Proceedings of the 39th Conference on Neural Information Processing Systems. Barcelona, Spain, 2016: 1857–1865.

[16] MOVSHOVITZ-ATTIAS Y, TOSHEV A, LEUNG T K, et al. No fuss distance metric learning using proxies[C]// Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 360–368.

[17] HERSHEY J R, CHEN Zhuo, LE ROUX J, et al. Deep clustering: Discriminative embeddings for segmentation and separation[C]//2016 IEEE International Conference on Acoustics, Speech and Signal Processing. Shanghai, China, 2016: 31–35.

[18] SONG H O, XIANG Yu, JEGELKA S, et al. Deep metric learning via lifted structured feature embedding[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 4004–4012.

[19] SENER O, SONG H O, SAXENA A, et al. Learning transferrable representations for unsupervised domain adaptation[C]//Proceedings of the 30th Conference on Neural Information Processing Systems. Barcelona, Spain, 2016: 2110–2118.

[20] BROMLEY J, GUYON I, LECUN Y, et al. Signature verification using a "siamese" time delay neural network [C]//Proceedings of the 6th International Conference on Neural Information Processing Systems. Denver, USA, 1993: 737–744.

[21] CHOY C B, GWAK J, SAVARESE S, et al. Universal correspondence network[C]//Proceedings of the 30th Conference on Neural Information Processing Systems. Barcelona, Spain, 2016: 2414–2422.

[22] PRABHU Y, VARMA M. FastXML: A fast, accurate and stable tree-classifier for extreme multi-label learning[C]// Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA, 2014: 263–272.

[23] YEN I E H, HUANG Xiangru, ZHONG Kai, et al. PD-sparse: a primal and dual sparse approach to extreme multiclass and multilabel classification[C]//Proceedings of the 33rd International Conference on International Conference on Machine Learning. New York, USA, 2016: 3069–3077.

[24] CHOROMANSKA A, AGARWAL A, LANGFORD J. Extreme multi class classification[C]//Neural Information Processing Systems Conference. Lake Tahoe, USA, 2013.

[25] SCHROFF F, KALENICHENKO D, PHILBIN J. FaceNet: A unified embedding for face recognition and clustering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 815–823.

[26] HADSELL R, CHOPRA S, LECUN Y. Dimensionality reduction by learning an invariant mapping[C]//2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York, USA, 2006, 2: 1735–1742.

[27] ZHAI A, WU Haoyu. Making classification competitive for deep metric learning[J]. arXiv preprint arXiv: 1811.

12649, 2018.
[28] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network[J]. arXiv preprint arXiv: 1503.02531, 2015.
[29] ZHANG Xu, YU F X, KARAMAN S, et al. Heated-up softmax embedding[J]. arXiv preprint arXiv: 1809.04157, 2018.
[30] CHOPRA S, HADSELL R, LECUN Y. Learning a similarity metric discriminatively, with application to face verification[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, USA, 2005: 539–546
[31] ROWEIS S T, SAUL L K. Nonlinear dimensionality reduction by locally linear embedding[J]. Science, 2000, 290(5500): 2323–2326.
[32] DONOHO D L, GRIMES C E. Hessian eigenmaps: Locally linear embedding techniques for high- dimensional data[J]. Proceedings of the national academy of sciences of the United States of America, 2003, 100(10): 5591–5596.
[33] JOLLIFFE I T. Principal component analysis[M]. Berlin: Springer, 2011.
[34] NOROUZI M, FLEET D J, SALAKHUTDINOV R. Hamming distance metric learning[C]//Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, USA, 2012: 1061–1069.
[35] CUI Yin, ZHOU Feng, LIN Yuanqing, et al. Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 1153–1162.
[36] MISHCHUK A, MISHKIN D, RADENOVIC F, et al. Working hard to know your neighbor's margins: Local descriptor learning loss[C]//Advances in Neural Information Processing Systems. Long Beach, USA, 2017: 4826–4837.
[37] HARWOOD B, KUMAR B G, CARNEIRO G, et al. Smart mining for deep metric learning[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 2821–2829.
[38] YUAN Yuhui, YANG Kuiyuan, ZHANG Chao. Hard-aware deeply cascaded embedding[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 814–823.
[39] WU Chaoyuan, MANMATHA R, SMOLA A J, et al. Sampling matters in deep embedding learning[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 2840–2848.
[40] USTINOVA E, LEMPITSKY V. Learning deep embeddings with histogram loss[C]//Proceedings of the 30th Conference on Neural Information Processing Systems. Barcelona, Spain, 2016: 4170–4178.
[41] LIU Hongye, TIAN Yonghong, WANG Yaowei, et al. Deep relative distance learning: Tell the difference between similar vehicles[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 2167–2175.
[42] LAW M T, URTASUN R, ZEMEL R S. Deep spectral clustering learning[C]//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia, 2017: 1985–1994.
[43] FUKUNAGA K, NARENDRA P M. A branch and bound algorithm for computing k-nearest neighbors[J]. IEEE transactions on computers, 1975, C-24(7): 750–753.
[44] WEN Yandong, ZHANG Kaipeng, LI Zhifeng, et al. A discriminative feature learning approach for deep face recognition[C]//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 499–515.
[45] SONG H O, JEGELKA S, RATHOD V, et al. Deep metric learning via facility location[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 5382–5390.
[46] BELL S, BALA K. Learning visual similarity for product design with convolutional neural networks[J]. ACM transactions on graphics (TOG), 2015, 34(4): 98.
[47] KAUFMAN L, ROUSSEEUW P J, DODGE Y. Clustering by Means of Medoids[M]//Dodge Y. Statistical Data Analysis Based on the L1-Norm and Related Methods. North-Holland: Elsevier, 1987.
[48] LIN Hui, BILMES J A. Learning mixtures of submodular shells with application to document summarization[C]// Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence. Catalina Island, USA, 2012: 479–490.
[49] TSCHIATSCHEK S, IYER R K, WEI Haochen, et al. Learning mixtures of submodular functions for image collection summarization[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada, 2014: 1413–1421.
[50] EMERSON A E. Handbook of theoretical computer science[M]. Amsterdam: Elsevier, 1990.
[51] KNUTH D E. Postscript about NP-hard problems[J]. ACM SIGACT news, 1974, 6(2): 15–16.
[52] MANNING C D, RAGHAVAN P, SCHÜTZE H. Introduction to information retrieval[M]. New York: Cambridge University Press, 2008.
[53] IONESCU C, VANTZOS O, SMINCHISESCU C. Training deep networks with structured layers by matrix backpropagation[J]. arXiv preprint arXiv: 1509.07838, 2015.
[54] WANG Xinshao, HUA Yang, KODIROV E, et al. Ranked list loss for deep metric learning[J]. arXiv preprint arXiv: 1903.03238, 2019.
[55] SHARIF RAZAVIAN A, AZIZPOUR H, SULLIVAN J, et al. CNN features off-the-shelf: an astounding baseline for recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Columbus, USA, 2014: 806–813.

[56] HINTON G E, SRIVASTAVA N, KRIZHEVSKY A, et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. arXiv preprint arXiv: 1207.0580, 2012.

[57] SERMANET P, EIGEN D, ZHANG Xiang, et al. OverFeat: Integrated recognition, localization and detection using convolutional networks[J]. arXiv preprint arXiv: 1312.6229, 2013.

[58] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification[C]//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 1026–1034.

[59] TAIGMAN Y, Yang MING, RANZATO M A, et al. DeepFace: Closing the gap to human-level performance in face verification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 1701–1708.

[60] SUN Yi, CHEN Yuheng, WANG Xiaogang, et al. Deep learning face representation by joint identification-verification[C]//Advances in Neural Information Processing Systems. Montreal, Quebec, Canada, 2014: 1988–1996.

[61] SUN Yi, WANG Xiaogang, TANG Xiaoou. Deeply learned face representations are sparse, selective, and robust[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 2892–2900.

[62] WAN Li, ZEILER M, ZHANG Sixin, et al. Regularization of neural networks using DropConnect[C]//Proceedings of the 30th International Conference on Machine Learning. Atlanta, GA, USA, 2013: 1058–1066.

[63] DENG Jiankang, ZHOU Yuxiang, ZAFEIRIOU S. Marginal loss for deep face recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, USA, 2017: 60–68.

[64] ZHANG Xiao, FANG Zhiyuan, WEN Yandong, et al. Range loss for deep face recognition with long-tailed training data[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 5409–5418.

[65] WANG Feng, CHENG Jian, LIU Weiyang, et al. Additive margin softmax for face verification[J]. IEEE signal processing letters, 2018, 25(7): 926–930.

[66] CHEN Binghui, DENG Weihong, DU Junping. Noisy softmax: Improving the generalization ability of DCNN via postponing the early softmax saturation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 5372–5381.

[67] WAN Weitao, ZHONG Yuanyi, LI Tianpeng, et al. Rethinking feature distribution for loss functions in image classification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 9117–9126.

[68] QI Xianbiao, ZHANG Lei. Face recognition via centralized coordinate learning[J]. arXiv preprint arXiv: 1801.05678, 2018.

[69] LIU Weiyang, WEN Yandong, YU Zhiding, et al. Sphereface: SphereFace: Deep hypersphere embedding for face recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 212–220.

[70] WANG Hao, WANG Yitong, ZHOU Zheng, et al. CosFace: Large margin cosine loss for deep face recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 5265–5274.

[71] LIU Weiyang, WEN Yandong, YU Zhiding, et al. Large-Margin Softmax Loss for Convolutional Neural Networks[C]//Proceedings of the 33rd International Conference on Machine Learning. New York, USA, 2016, 2(3): 7.

[72] BUCILUǍ C, CARUANA R, NICULESCU-MIZIL A. Model compression[C]//Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, USA, 2006: 535–541.

[73] WAH C, BRANSON S, WELINDER P, et al. The Caltech-UCSD Birds-200-2011 Dataset[R]. Computation & Neural Systems Technical Report, CNS-TR-2011-001, Pasadena, CA, USA: California Institute of Technology, 2011.

作者简介：

刘冰，女，1994 年生，博士研究生，主要研究方向为深度学习、计算机视觉和生物特征识别。

李瑞麟，男，1995 年生，硕士研究生，主要研究方向为深度学习、计算机视觉和生物特征识别。

封举富，男，1967 年生，教授，博士生导师，主要研究方向为图像处理、模式识别、机器学习和生物特征识别。主持和参与国家自然科学基金、"十一五"国家科技支撑计划课题、973 计划等项目多项。曾获中国高校科技二等奖、第一届亚洲计算机视觉国际会议优秀论文奖、北京大学安泰教师奖、北京大学大众电脑优秀奖、北京大学安泰项目奖等奖励多项。发表学术论文 300 余篇。