文章编号:2095-6134(2021)06-0841-11

基于原型学习改进的伪标签半监督学习算法*

杨雨龙[1],郭田德[1,2],韩丛英[1,2†]

(1 中国科学院大学数学科学学院, 北京 100049; 2 中国科学院大数据挖掘与知识管理重点实验室, 北京 100190)

(2021 年 4 月 15 日收稿; 2021 年 5 月 12 日收修改稿)

Yang Y L, Guo T D, Han C Y. Improving pseudo-labeling semi-supervised learning based on prototype learning[J]. Journal of University of Chinese Academy of Sciences, 2021, 38(6): 841-851.

摘 要 近年来,基于图像增广和一致性正则化的半监督学习(semi-supervised learning, SSL)方法被广泛应用并取得了很大的成功。然而,由于伪标签算法存在“认知偏误”问题,即模型的错误通过伪标签累积从而难以改正,因此很少有人关注基于伪标签(pseudo-labeling, PL)的半监督学习方法。提出一种特征图的原型图注意力特征修正模型(prototype attention layer, PAL):即在神经网络映射的特征空间上学习一个图注意力模型,将此模型应用于特征空间中,可以充分利用原型的信息来修正特征,将修正后的特征所产生的伪标签与原型分配产生的伪标签随机线性组合,从而得到新的伪标签。将这一模型应用到 2 种伪标签半监督学习框架上所得到的算法(prototype attention improved pseudo-labeling, PAIPL),在 CIFAR-10 和 CIFAR-100 的多个半监督分类问题上进行测试,分类准确率都得到了显著提升。特别地,将提出的修正模型应用于伪标签半监督学习 PLCB 框架时,又提出相互混合的监督技术,从而取得了更好的效果。还将提出的模型应用到其他多个伪标签半监督学习框架上,并在多个数据集上进行实验,验证了所提出的模型作为一个附加模块是普适且有效的。

关键词 半监督学习;伪标签;MixUp;图注意力模型;原型学习

中图分类号: TP391 **文献标志码**:A **doi**: 10.7523/j.issn.2095-6134.2021.06.015

Improving pseudo-labeling semi-supervised learning based on prototype learning

YANG Yulong[1], GUO Tiande[1,2], HAN Congying[1,2]

(1 *School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing* 100049, *China*;
2 *Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences, Beijing* 100190, *China*)

Abstract In recent years, semi-supervised learning (SSL) methods based on image augmentation and consistency regularization have been widely used and have achieved great success. However, little attention has been paid to pseudo-labeling (PL)-based semi-supervised learning methods because of the “confirmation bias” problem, i.e., errors in the model are accumulated by wrong pseudo-labels and thus difficult to be corrected. In this paper, we propose a feature refinement model based on the feature space graph. The model learns a graph attention model on the feature

* 国家自然科学基金(11731013,U19B2040,11991022)和中国科学院战略性先导科技专项(XDA27000000)资助

† 通信作者,E-mail: hancy@ucas.ac.cn

space mapped by the neural network. We apply this model to the feature space to make use of the information of the prototypes to refine the features. The pseudo-labels generated by the refined features are randomly and linearly combined with the pseudo-labels generated by the prototypes assignment to obtain new pseudo-labels. In this paper, we apply this module to two pseudo-labeling semi-supervised learning frameworks and achieve significant accuracy improvements in several CIFAR-10 and CIFAR-100 semi-supervised classification problems. In particular, we apply our feature refinement model to the pseudo-labeling semi-supervised learning framework PLCB and add the proposed mutual mix supervision techniques to achieve good results on this framework. By applying the proposed feature refinement module to several pseudo-labeling semi-supervised learning frameworks and conducting experiments on several datasets, the proposed algorithm is demonstrated to be universal and effective as an add-on module.

Keywords semi-supervised learning; pseudo-labeling; MixUp; graph attention model; prototype learning

深度神经网络被应用在计算机视觉和自然语言处理等许多领域,都取得了优秀效果。然而,训练一个深度神经网络需要数以百万计的标注样本和大量的计算资源,而对大量数据进行标注是很困难的。学术界已经研究了几种替代方案来缓解这一问题,如半监督学习(semi-supervised learning,SSL)、无监督学习和自监督学习。

半监督学习方法[1-2]是为解决海量无标签数据和高代价标注工作之间的矛盾而产生的。在半监督学习中,含有大量的未标注数据,只有小部分有标签数据。随着研究的深入,半监督学习算法在图像分类[3-4]、语义分割[5-6]、自然语言处理[7-8]等领域都取得了不错的结果。

本文主要研究基于伪标签(pseudo-labeling)的半监督学习图像分类算法。这类方法用输入图像在训练过程中的历史输出生成伪标签,并将其作为监督信号,然后以有监督的模式进行学习。

现有的伪标签方法存在“认知偏误(confirmation bias)”[9-10]的问题。“认知偏误”,也称为噪声积累,即模型的错误由于使用了自身提供的错误伪标签进行训练而得到加深。这种错误累积是由于伪标签方法仅使用单个样本自身的预测进行监督,一旦模型对样本预测错误,这一错误将被当作监督信号,而这个错误的监督信号无法通过与其他样本的比较得到修正。PLCB[11]通过使用 MixUp[12]引入成对图像的信息,在一定程度上缓解了认知偏误。然而,成对图像提供的流形信息有限,PLCB 依然会受到认知偏误的影响。

本文提出一种新的特征修正模型,即原型注意力模型,由于和神经网络结合,又称为原型注意力层(prototype attention layer, PAL)。假设在每一类样本的数据流形中存在 P 个具有代表性的点,即原型,所有样本都可找到某一原型与之相近。通过学习原型,得到数据流形的压缩表示,使得每个样本在训练时都能参考整个数据流形,从而缓解“认知偏误”。在特征空间中随机初始化 $C \times P$ 个向量作为原型,即每类有 P 个原型。将样本分类到某一原型的过程称为原型分配(prototypes assignment,PA)。把每个样本在当前迭代中的原型分配向量作为下一次迭代的伪标签,通过优化带正则项的交叉熵损失,来训练样本的原型。用学习到的原型合并样本特征,共同构建包括原型向量和样本特征的图,然后通过可学习的图注意力[13]模型来获得更好的特征。将 PAL 分别应用于 2 个伪标签半监督学习框架,得到 2 种使用原型学习改进的伪标签半监督学习算法(prototype attention improved pseudo-labeling, PAIPL):一种应用到软伪标签的自训练(self-training[14])框架,得到 PAIPL-S 算法;一种应用到伪标签的 PLCB 框架,得到 PAIPL-P。为了更好地使用伪标签,本文还提出相互混合的监督技巧,用于伪标签生成,从而使生成的伪标签既能在早期相对迅速地收敛,又具备了好的流形表示。PAIPL-S 和 PAIPL-P 算法在不增加图像预处理技术、而且训练使用小批量数据情形下,于 CIFAR-10 与 CIFAR-100 数据库上取得了很好的结果。本文提出的 PAIPL 算法架构如图 1 所示。

本文的贡献主要有以下 3 个方面:

1)克服了现有半监督学习方法对不同数据之间的关系信息利用不足的问题,提出一种基

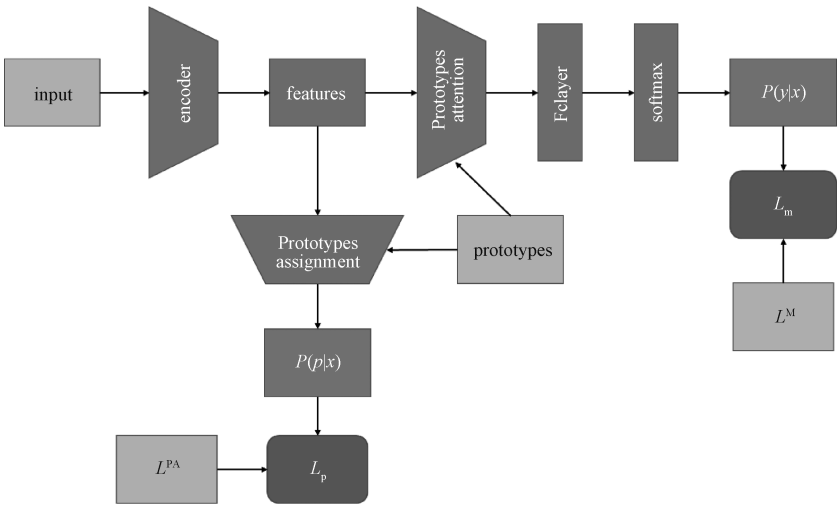


图 1　PAIPL 算法架构

Fig. 1　A schema of PAIPL

于原型的图注意力模型来生成特征:即通过训练学习原型,得到数据流形的一个压缩表示,通过图注意力模型,从原型中获得对样本分类有用的信息,合并原有特征,得到参考数据流形修正的特征;

2) 将原型注意力模型应用到 2 种伪标签半监督学习框架中,得到 2 种新的伪标签半监督学习算法:PAIPL-S 和 PAIPL-P。相对于没有加入原型注意力模型的基线方法,算法的准确率得到显著的提升;

3)提出一种相互混合监督的伪标签生成方法。传统的伪标签生成方法使用同一数据的历史输出作为伪标签,收敛速度快,但存在"认知偏误"问题。单纯基于流形信息的伪标签生成能通过邻域信息校正伪标签的错误,但可能出现过分平滑的现象。本文通过综合二者的优缺点,提出以二者的随机线性组合作为伪标签,使得模型既能获得前期的收敛速度,又能防止后期的过分平滑。同时,融合这 2 种方式的相互学习也能防止它们各自陷入自己的局部最优解。

1　相关工作

关于深度半监督学习的工作主要有 2 个分支,即一致性正则化和伪标签方法。在下面的讨论中,将深度神经网络(卷积神经网络)特征提取器记为 $f(\cdot)$,它将输入图像映射到一个高维特征向量 $f(\boldsymbol{x})$。分类器(全连接层后接 softmax 函数) $c(\cdot)$ 将特征向量作为输入,并输出分布向量 $p(\boldsymbol{y}\mid\boldsymbol{x})=c(f(\boldsymbol{x}))$。

一致性正则化方法对同一样本的不同数据增广进行预测,并最小化它们之间的差异。之前的研究在无标签数据上大多应用以下一致性正则化损失:

$$\| c(f(\mathrm{Aug}_1(\boldsymbol{x}))) - c(f(\mathrm{Aug}_2(\boldsymbol{x}))) \|_2^2 \quad (1)$$

其中:$\mathrm{Aug}_1(\cdot)$ 和 $\mathrm{Aug}_2(\cdot)$ 是 2 种不同的随机图像增广。π-model[15] 应用随机数据增广,要求模型对同一数据在 2 种不同数据增广下的预测结果相近。在此基础上,为保留更多的历史信息并稳定训练,文献[15]的 Temporal Ensemble 将历史预测的指数平均作为监督信号,最小化当前预测与历史平均预测的差异。另一种保留历史信息并稳定训练的算法是 Mean Teacher[10],以模型的参数的指数平均作为教师模型,用教师模型的预测来指导训练过程。然而,Mean Teacher 中教师模型会逐渐收敛到学生模型,使得一致性正则化损失的作用随着训练的进行而减小。Dual Student[16] 为解决这一问题,提出分别训练 2 个学生模型,以在样本点稳定的一个学生模型的预测作为在该点不稳定的另一个学生模型的监督信号。然而,这几种方法的数据增广只使用了常规的图像数据增广,多样性有限。虚拟对抗训练(virtual adversarial training, VAT)[17] 应用对抗训练来生成对抗样本,得到与传统图像增广不同的增广数据,并要求模型在对抗样本和原始样本上的预测相似。VAT 还使用熵最小化作为额外的正则化,使模型在未标记的数据上做出明确的预测。但对抗样本只集中在数据点的附近,且不能很好地覆盖数据流

形。最近,一些研究引入 MixUp 正则化作为训练信号,ICT[18] 和 MixMatch[19] 要求成对样本的线性插值的预测和其标签(或伪标签)的相应插值之间的一致性。ICT 应用 Mean Teacher 生成伪标签,而 MixMatch 则使用不同数据增广的预测的均值生成伪标签。在 MixMatch 中也应用了熵最小化。上述文献只建模成对样本,对数据流形的利用仍然不足。UDA[20] 专注于重度数据增广,通过元学习来选择数据增广方法,然后最小化原始样本和增广样本之间的预测差异来实现一致性正则化。但 UDA 的性能高度依赖于数据增广方法库的合理性和多样性。

伪标签方法通过对未标记的数据生成伪标签,再以有监督学习的形式训练。Lee[21] 直接将模型的预测作为伪标签。他们对模型进行预训练,在微调过程中使用伪标签。这种只考虑样本自身历史预测的伪标签算法受到认知偏误问题的严重影响。Self-training[14] 首先用有标签数据训练模型,然后用训练好的模型对无标签数据标注,将预测概率最大值大于某一阈值的数据加入有标签数据再次训练,反复如此直至再也不能向有标签数据集中添加数据。虽然通过逐步添加可信样本避免了使用明显错误的样本进行训练,但由于错误数据一经加入有标签数据集后就无法纠正,受到认知偏误的严重影响。还有一些研究考虑了生成伪标签的不确定性[22-23],使用 k 个最近邻点的距离作为不确定性的衡量标准,通过优化损失来缩小类内距离、扩大类间距离,但这样只能利用局部的流形信息。PLCB[11] 引入 MixUp,使模型能利用成对数据线性插值的信息。一些研究通过在 PLCB 中加入 dropout[24]、权重归一化[25]、类别分布对齐[26]、熵最小化[27],并在同一批次以固定比例加载有标签和无标签数据,在许多图像分类问题上都获得了显著的提升。然而,PLCB 只能使用成对图像的信息,对数据流形整体的利用不足。另有一些学者提出结合基于图的标签传播来获得更好的伪标签[28],此算法交替进行 2 个过程:1)用有标签数据和伪标签数据来训练模型;2)用从模型中得到的特征来构建最近邻图,并应用标签传播算法来调整伪标签。文献[28]虽然成功利用了整个数据流形,但这种方法需要对所有样本的特征建图,计算量过大,不适用于稍大的数据集。

2 基于原型学习改进的伪标签半监督学习算法

下面给出本文基于原型学习改进的伪标签半监督学习算法。半监督学习指的是从数据集 D 的 N 个训练样本集中学习一个模型 $f_\theta(\cdot)=c_\theta(h_\theta(\cdot))$,$D$ 被分割成有标签的数据集 $D_\mathrm{l}=\{(\boldsymbol{x}_i,\boldsymbol{y}_i)\}_{i=1}^{N_\mathrm{l}}$ 和无标签的数据集 $D_\mathrm{u}=\{\boldsymbol{u}_i\}_{i=1}^{N_\mathrm{u}}$。模型 $f_\theta(\cdot)$ 是一个卷积神经网络,特征提取器 h_θ 将输入样本映射为特征向量,分类器 c_θ 将特征向量映射为概率向量,分类器由全连接层和 softmax 组成。无标签样本 $\{\boldsymbol{u}_i\}_{i=1}^{N_\mathrm{u}}$ 的伪标签记作 $\{\tilde{\boldsymbol{y}}_i\}_{i=1}^{N_\mathrm{u}}$,是由上一次迭代中模型对未经增广的无标签样本预测得到的,对应数据集记作 $\tilde{D}_\mathrm{u}=\{(\boldsymbol{u}_i,\tilde{\boldsymbol{y}}_i)\}_{i=1}^{N_\mathrm{u}}$,用于训练的完整伪标签数据集记作 $\tilde{D}=D_\mathrm{l}+\tilde{D}_\mathrm{u}=\{(\boldsymbol{x}_i,\tilde{\boldsymbol{y}}_i)\}_{i=1}^{N}$。

与文献[11]类似,在训练时,每个批次中按照固定比例加载有标签数据,其余是无标签的数据,同时优化 2 个损失:主分支的损失 L_m 和用来学习原型的损失 L_p。

2.1 原型注意力层

MixMatch 和 PLCB 等方法只利用了成对数据的监督信号,而 PAIPL 在神经网络中增加了带原型注意力模型的隐藏层,以利用整个数据流形的信息。PAIPL 存储每个类的若干原型向量 $\boldsymbol{p}_i^j(1\leqslant i\leqslant P,1\leqslant j\leqslant C)$,并构造一个由训练样本特征和所有原型向量构成的图。利用图注意力网络收集所有原型的信息来改善样本的表示。

2.1.1 通过原型分配学习原型向量

为使用原型注意力模型,首先通过优化一个正则化的交叉熵损失来学习原型向量 $\boldsymbol{p}_i^j(1\leqslant i\leqslant P,1\leqslant j\leqslant C)$。获得样本的原型权重以及由原型权重得到伪标签的过程称为原型分配,如图 2。首先用特征提取器 $h_i=h_\theta(\boldsymbol{x}_i)$ 提取特征,将特征和原型映射到一个更高维的空间。与 NTXent 损失[29] 类似,将归一化的特征和原型做内积,并使用 softmax(·)得到原型分配向量。此外,PAIPL 添加分布对齐[26] 和熵最小化正则项[27],见式(6)和式(7)。PAIPL 还使用了一个损失来匹配原型分配生成的伪标签和主分支分类的伪标签。先对特征空间的样本特征和原型向量作升维:

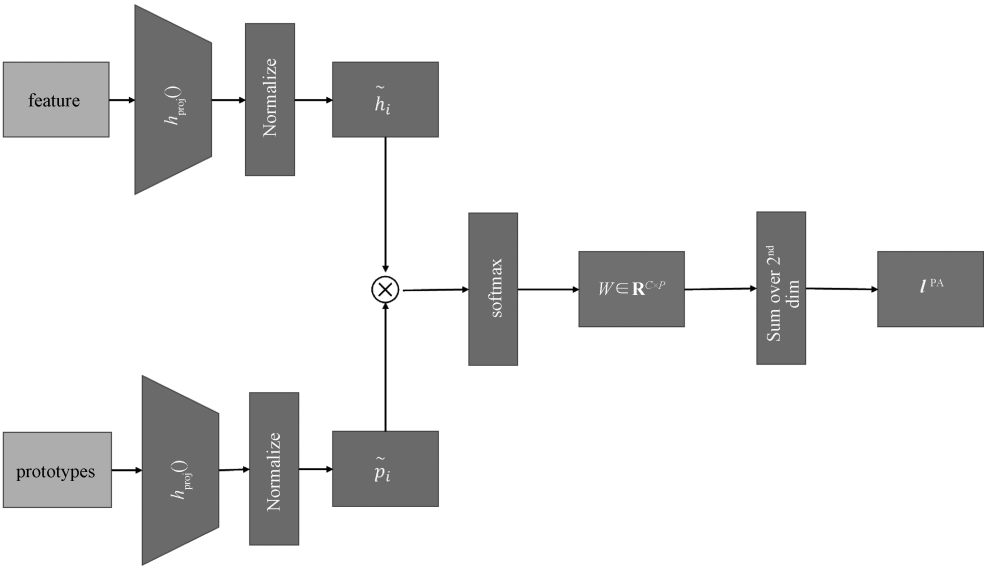


图 2　原型分配过程

Fig. 2　Prototypes assignment

$$\tilde{\boldsymbol{h}}_i = h_{\text{proj}}(\boldsymbol{h}_i), \tag{2}$$

$$\tilde{\boldsymbol{p}}_i = h_{\text{proj}}(\boldsymbol{p}_i), \tag{3}$$

其中:h_{proj} 是一个非线性函数,它将特征映射到一个更高维的空间。然后计算 $\tilde{\boldsymbol{h}}_i$ 和 $\tilde{\boldsymbol{p}}_j$ 之间的归一化余弦相似度:

$$w_{ij} = \text{softmax}\left(\frac{\tilde{\boldsymbol{h}}_i \cdot \tilde{\boldsymbol{p}}_j}{T \cdot |\tilde{\boldsymbol{h}}_i| \cdot |\tilde{\boldsymbol{p}}_j|}\right), \tag{4}$$

其中 T 是温度参数。使用交叉熵损失作为原型学习目标的主要损失

$$L_{\text{protos}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{P}\tilde{w}_{ij}^{\text{T}}\log(w_{ij}), \tag{5}$$

其中:$\tilde{w}_{ij}$ 是上一次迭代中样本 $\boldsymbol{x}_i$ 在原型 $\boldsymbol{p}_j$ 上分配的概率,作为本次迭代的伪标签。

在原型损失中添加 2 个正则项,类别分布对齐损失 R_{A}[26] 和熵最小化损失 R_{H}[27]。类别分布损失要求无标签样本中的原型分布与先验一致,即每个原型代表了数量均等的样本。熵最小化损失要求模型做出足够明确的判断,即每个样本归属于特定的某个原型。类别分布对齐损失 R_{A} 为

$$R_{\text{A}} = \sum_{i=1}^{P} p_i \log\left(\frac{p_i}{h_i}\right), \tag{6}$$

其中:p_i 是任意样本属于第 i 个原型的先验分布(假设为均匀分布),h_i 是模型在小批量数据上预测的经验分布 $\boldsymbol{h} = \sum_{i=1}^{B} f_\theta(\boldsymbol{x}_i)$。熵最小化损失 R_{H} 为

$$R_{\text{H}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} w_{ij}\log(w_{ij}), \tag{7}$$

其中:w_{ij} 是 $\boldsymbol{x}_i$ 被模型分配到原型 $\boldsymbol{P}_j$ 的概率。

本文还提出一个损失来匹配原型分配的伪标签和主分支产生的分类的伪标签

$$R_{\text{PM}} = -\frac{1}{N}\sum_{i=1}^{N}\tilde{\boldsymbol{y}}_i^{\text{T}}\log(\boldsymbol{y}_i^w), \tag{8}$$

其中 $\boldsymbol{y}_i^w = \left(\sum_{n=1}^{P} w_n, \cdots, \sum_{n=C\times P-P+1}^{C\times P} w_n\right)$ 是原型分配在第 i 类原型上的求和,即下一次迭代中原型分配产生的伪标签。$\tilde{\boldsymbol{y}}_i^{\text{T}}$ 是上一次迭代中原型分配在第 i 类原型上的求和,即本次迭代中原型分配产生的伪标签。

因此,总的原型损失为

$$L_{\text{p}} = L_{\text{protos}} + \lambda_{\text{A}}R_{\text{A}} + \lambda_{\text{H}}R_{\text{H}} + \lambda_{\text{PM}}R_{\text{PM}}. \tag{9}$$

2.1.2　原型注意力层构造

在学习到原型向量后,PAIPL 通过原型注意力层 f^{att},使用软注意力机制来调整输入图像的特征,如图 3。所有原型向量与输入图像的特征构成的图被用来改善输入图像的特征。非线性映射 h_{proj} 把特征向量 $\boldsymbol{f}_i$ 和原型向量 $\boldsymbol{p}_j$ 映射到另一个空间 $\tilde{\boldsymbol{f}}_i = h_{\text{proj}}(\boldsymbol{f}_i)$,$\tilde{\boldsymbol{p}}_j = h_{\text{proj}}(\boldsymbol{p}_j)$。PAIPL 先把 $\tilde{\boldsymbol{f}}_i$ 和 $\tilde{\boldsymbol{p}}_j$ 规范化到单位向量 $\boldsymbol{e}_i^x = \tilde{\boldsymbol{f}}_i / \|\tilde{\boldsymbol{f}}_i\|$,$\boldsymbol{e}_j^p = \tilde{\boldsymbol{p}}_j / \|\tilde{\boldsymbol{p}}_j\|$,再以特征作为查询值,原型作为键值,以下述方式计算注意力:

$$w_{ij} = \text{softmax}(a(\boldsymbol{e}_i^x, \boldsymbol{e}_j^p)), \tag{10}$$

其中:softmax(·)应用于所有原型的权重向量,$a(\cdot)$是用于计算未归一化注意力系数的线性映射。用原型向量的加权平均对所有原型的信息进行聚合:

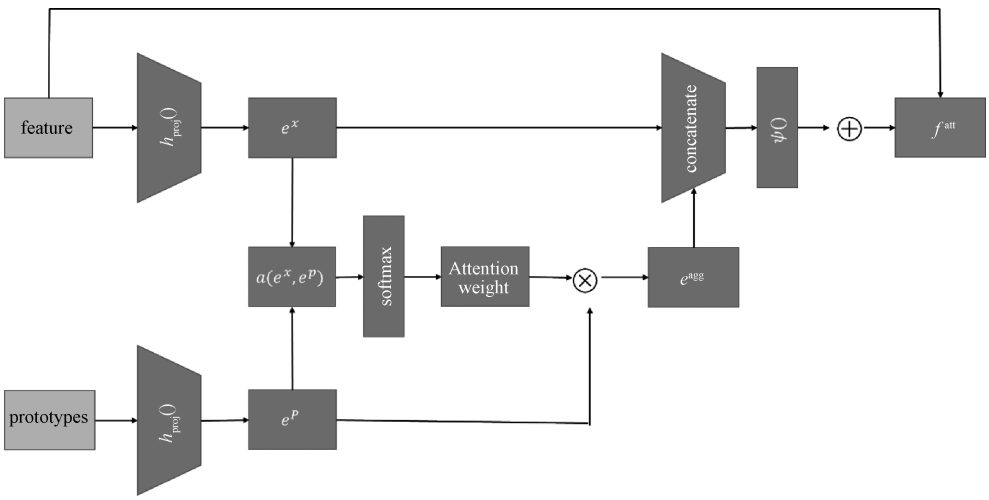


图 3 原型注意力层

Fig. 3 Prototypes attention layer

$$\boldsymbol{e}_i^{\text{agg}} = \sum_{j=1}^{P} w_{ij}\boldsymbol{e}_j^p. \tag{11}$$

将输入图像的嵌入 $\boldsymbol{e}^x$ 和聚合嵌入 $\boldsymbol{e}^{\text{agg}}$ 作为新的特征,投影回原来的低维特征空间:

$$\boldsymbol{f}_i^{\text{att}} = \boldsymbol{f}_i + \psi([\boldsymbol{e}_i^x, \boldsymbol{e}_i^{\text{agg}}]), \tag{12}$$

其中 $\psi(\cdot)$ 是一个 2 层非线性网络。使用了残差块来降低训练难度。

2.2 原型注意力层改进的半监督学习

PLCB 使用 MixUp 缓解“认知偏误”。MixUp 使用样本标签对 $((\boldsymbol{x}_p, \boldsymbol{y}_p), (\boldsymbol{x}_q, \boldsymbol{y}_q))$ 的凸组合训练神经网络:

$$\boldsymbol{x}^{\text{mix}} = \delta\boldsymbol{x}_p + (1-\delta)\boldsymbol{x}_q, \tag{13}$$

$$\tilde{\boldsymbol{y}}^{\text{mix}} = \delta\tilde{\boldsymbol{y}}_p + (1-\delta)\tilde{\boldsymbol{y}}_q, \tag{14}$$

其中 $\delta \sim B(\alpha, \alpha)$,即参数为 (α, α) 的 Beta 分布。于是交叉熵损失变为

$$L_{\text{mix}} = -\sum_{i=1}^{N} \delta[\tilde{\boldsymbol{y}}_{l,p}^{\text{T}} \log(f_\theta(\boldsymbol{x}_i^{\text{mix}}))] + (1-\delta)[\tilde{\boldsymbol{y}}_{l,q}^{\text{T}} \log(f_\theta(\boldsymbol{x}_i^{\text{mix}}))]. \tag{15}$$

MixUp 要求神经网络在训练样本对之间尽量近似于局部线性函数,从而实现决策边界的线性变化以实现更好的泛化。

PLCB 在 MixUp 损失的基础上同样增加了类别分布对齐损失 R_{A} 和熵最小化损失 R_{H}。值得注意的是,由于 PLCB 使用的数据经过了 MixUp 变换,R_{A} 和 R_{H} 中所用到的模型 $f_\theta(\boldsymbol{x}_i)$ 都是对线性组合后的样本 $\boldsymbol{x}_i$ 的预测。但在模型预测线性的假设下,2 个正则项的假设仍然成立。

故正则化 MixUp 的总损失函数为

$$L_{\text{m}} = L_{\text{mix}} + \lambda_{\text{A}} R_{\text{A}} + \lambda_{\text{H}} R_{\text{H}}, \tag{16}$$

其中 λ_{A} 和 λ_{H} 是超参数。

除了使用 MixUp 以外,PLCB 还应用了以下技巧:在每批次样本中同时加载有标签和无标签数据。每个批次中固定比例的样本是有标签的,而其余的是无标签的。在 PLCB 中,对于不同有标签数据数量的不同数据集,每个批次中有标签数据的比例是一个重要的超参数,对模型的结果有不小的影响。这是为了防止有标签数据的采样过度和采样不足。为了获得无噪声的伪标签,生成伪标签时不引入随机性,即不做图像增广且不使用 dropout,而训练时使用图像增广和 dropout。

将原型注意力层应用在 PLCB 上:1)在特征提取过程加入了原型注意力层,改变了特征提取器 $h_\theta(\cdot)$;2)加入了原型学习训练损失 L_{p}。总损失为

$$L = L_{\text{m}} + \lambda L_{\text{p}}, \tag{17}$$

其中 λ 为超参数。

使用软伪标签的自训练模型是经典的伪标签半监督算法之一。对于有标签数据,模型使用交叉熵损失进行有监督训练。而对于无标签数据,模型使用软伪标签,用交叉熵损失进行训练。软伪标签由模型上一次迭代的预测得到。

$$L_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N} \tilde{\boldsymbol{y}}_i^{\text{T}} \log(f_\theta(\boldsymbol{x}_i)), \tag{18}$$

其中 $\tilde{\boldsymbol{y}}_i$ 为样本 $\boldsymbol{x}_i$ 的伪标签。

在此基础上,同样增加了类别分布对齐损失 R_{A} 和熵最小化损失 R_{H}。

正则化软伪标签自训练的总损失函数为

$$L_{\text{m}} = L_{\text{CE}} + \lambda_{\text{A}} R_{\text{A}} + \lambda_{\text{H}} R_{\text{H}}, \tag{19}$$

其中 λ_{A} 和 λ_{H} 是超参数。

将原型注意力层应用在软伪标签的自训练框架上:1)在特征提取过程加入了原型注意力层,改变了特征提取器 $h_\theta(\cdot)$;2) 加入了原型学习训练损失 L_p。 总损失为:

$$L = L_m + \lambda L_p, \tag{20}$$

其中 λ 为超参数。

2.3　相互混合监督学习

为进一步提升模型性能,本文提出相互混合监督学习的技术。为了同时使用伪标签和相互学习[30]的方法,将 2 个分支的伪标签进行随机线性组合。一个分支是使用原型注意力层修正的传统神经网络,另一个分支是基于规范化余弦相似度的原型分配,详见 2.1.1。当得到 2.1.1 中的原型分配 w_{ij} 后,PAIPL 通过将每个原型向量分配到某个类中,将来自同一类的所有原型的概率相加,得到分类预测 $\boldsymbol{l}^{PA}$,如图 2 所示。采用线性预热(linear warm-up)的方式,每一个分支在一开始的时候都会注重于来自自己分支的监督,以达到训练初期较为迅速的收敛。在预热过程结束后,用 2 个分支的监督的随机线性组合对每个分支进行训练。具体来说:

$$\delta \sim U(0,1), \tag{21}$$

$$\alpha = \frac{i}{\text{warm}}\delta + \frac{\text{warm} - i}{\text{warm}}, \ (i \leqslant \text{warm}), \tag{22}$$

$$\alpha = \delta, \ (i > \text{warm}), \tag{23}$$

$$\boldsymbol{l}_i^{PA} = \alpha \boldsymbol{f}_{i-1}^{PA} + (1-\alpha)\boldsymbol{f}_i^{M}, \tag{24}$$

$$\boldsymbol{l}_i^{M} = \alpha \boldsymbol{f}_{i-1}^{M} + (1-\alpha)\boldsymbol{f}_i^{PA}, \tag{25}$$

其中:$\boldsymbol{f}_i^{PA}$ 和 $\boldsymbol{f}_i^{M}$ 是第 i 次迭代中原型分配分支和主分支的预测值。$\boldsymbol{l}_i^{PA}$ 和 $\boldsymbol{l}_i^{M}$ 是第 i 次迭代中用于训练原型分配分支和主分支的伪标签(第 $i-1$ 次迭代的预测)。warm 是预热过程的总迭代次数。

本文将加入原型注意力层和相互混合监督技术的伪标签半监督学习算法称为 PAIPL。将其在 PLCB 和软标签自训练上的应用分别称为 PAIPL-P 和 PAIPL-S。

3　实验

首先在几个标准的半监督学习基准上评估算法 PAIPL,包括各种不同有标签数据比例的 CIFAR-10 和 CIFAR-100。PAIPL 比原始框架 PLCB 表现更好,并且较性能优异的一致性算法 MixMatch 也有显著的提升。

3.1　数据集和训练细节

对 2 个常用的半监督学习数据集 CIFAR-10 和 CIFAR-100 进行实验。PAIPL 在不同数量的有标签数据下进行了测试。CIFAR-10 和 CIFAR-100 分别是 10 类和 100 类的自然图像数据集。CIFAR-10 包含 50 000 张训练图像和 10 000 张测试图像,大小为 32×32,平均分布在 10 个不相交的类上。CIFAR-100 包含 50 000 张训练图像和 10 000 张测试图像,大小为 32×32,均匀分布在 100 个不相交的类上。与文献[11]类似,我们为 CIFAR-10 和 CIFAR-100 都留出了 5 000 个样本作为验证集来调整超参数。而在与其他方法进行比较时,使用所有的 50 000 个训练样本。

实验使用不同比例的有标签数据。在 CIFAR-10 中,N_l = 250,500,1 000,4 000。 在 CIFAR-100 中,N_l = 500,1 000,4 000,$N_l + N_u$ = 50 000。N_l 表示标记样本的数量,N_u 表示未标记样本的数量。采用“13-CNN”[31]来提取特征,以便与前人的研究进行比较。

PAIPL 只使用非常简单的图像预处理:图像填充、颜色扰动、随机裁剪、水平翻转、图像归一化和高斯噪声。首先添加 2 个像素的边缘填充,并裁剪回原尺寸,得到 2 个像素的随机平移。然后进行颜色扰动,以增加数据的多样性。再以 0.5 的概率对图像进行水平翻转。用整个数据集的平均值和标准差对所有图像进行归一化。最后,加入均值为 0,标准差为 0.15 的高斯噪声。

使用随机动量梯度下降优化器训练模型,动量为 0.9,权重衰减为 10^{-4}。所有实验在对整个训练集进行训练之前都会进行预热。首先在有标签数据上预训练模型,只用 10 次迭代来获得后续训练的初始权重。训练了 400 次迭代,在 250 次和 350 次迭代时进行学习率衰减。

本文没有对正则化权重 λ_A 和 λ_H 进行大量的调参, 只按照文献[11]设置为 0.8 和 0.4。使用了 dropout,并在所有网络中使用权重归一化。

3.2　实验结果

首先展示在 CIFAR-10 和 CIFAR-100 上不同数量的有标签数据集的结果,见表 1。PAIPL-P 比目前最好的基于伪标签的方法 PLCB 有明显的改进,并且准确率比一致性方法 MixMatch 更高。

将 PAIPL-S 和 PAIPL-P 与它们对应的基线方法比较,结果显示,PAIPL-S 和 PAIPL-P 相较于软伪标签自训练和 PLCB 都有明显提升。这说

表 1 与其他方法的精度比较

Table 1 Accuracy comparison with previous methods

	methods	CIFAR-10			CIFAR-100	
		500	1 000	4 000	4 000	10 000
Supervised	Supervised *	56.02	64.89	80.47	45.26	58.39
	Supervised (M) *	62.49	71.14	84.02	47.55	60.97
Consistency	π-model[15]	–	–	87.64	–	60.81
	Mean Teacher[10]	72.55	80.96	88.59	54.64	63.92
	Dual Student[16]	–	85.83	91.11	–	67.23
	ICT[18]	–	84.52	92.71	–	–
	MixMatch[19]	90.35	92.25	93.76	–	–
Pseudo- Labeling	LP[28]	67.60	77.98	87.31	53.80	61.57
	LP+MT[28]	75.98	83.07	89.39	56.27	64.08
	Soft Self-training *	64.92	78.05	89.10	53.40	62.03
	PAIPL-S(ours) *	68.10	79.96	90.31	54.98	64.30
	PLCB[11]	91.20	93.15	94.05	62.45	67.85
	PAIPL-P(ours) *	92.89	93.38	94.18	65.04	71.74

注:M 表示使用 MixUp 正则化, * 表示本文的实验结果。

明 PAIPL 对 2 种伪标签半监督学习框架都是有效的。

我们将 PAIPL-P 与在 CIFAR-10 和 CIFAR-100 中使用 13-CNN[31] 架构的其他方法进行比较,尽管只使用小的批次以及基本的数据预处理和简单的预热策略,PAIPL-P 仍然取得了优异的结果。ICT 和 MixMatch 通过引入 MixUp,缓解了半监督学习中的认知偏误。PLCB 不仅引入了 MixUp,还加入了类别分布对齐和熵最小化等许多额外技术。PAIPL 引入了更复杂的流形信息,从而获得更好的性能提升。这表明本文提出的算法有效地缓解了伪标签学习的认知偏误问题。

3.3 消融实验

通过消融实验(ablation study),测试 PAIPL 的不同模块的效果。表 2 展示了 PAIPL-P 在 CIFAR-100 上的使用 4 000 和 10 000 个有标签数据的实验。由于 PAIPL-P 是基于 PLCB 改进的,本文将实验结果与 PLCB 进行比较。在 PLCB 的基础上,增加了 2 个模块:原型注意力层(PAL)和相互混合监督(MM)。

表 2 消融实验结果

Table 2 Results of ablation study

methods	CIFAR-100	
	4 000	10 000
Supervised *	45.26	58.39
Supervised(M) *	47.55	60.97
PLCB[11]	62.45	67.85
PLCB+PAL *	64.42	70.78
PLCB+MM *	63.08	68.72
PAIPL-P *	65.04	71.74

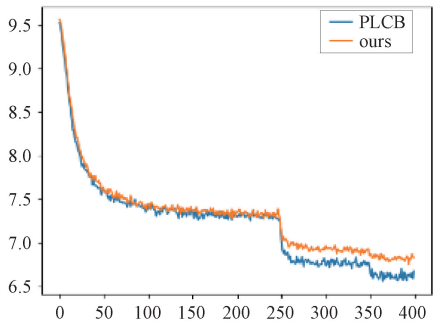
注:在有 4 000 和 10 000 个有标签样本的 CIFAR-100 上进行实验, * 表示本文的实验结果。

由于提供的数据非常有限,有监督学习(Supervised)只能得到较低的准确率。在有监督学习中加入 MixUp 后(Supervised(M)),由于 MixUp 的正则化要求边界更平滑,结果得到了改善。PLCB 加入了无标签数据,形成半监督学习框架,与有监督方法相比,准确率有较大的提高。本文通过增加原型注意力层(PLCB+PAL),可以获得比 PLCB 更高的准确率。这是由于原型注意力层提供了更复杂的流形信息,而不仅仅是成对数据的信息。加入相互混合监督(PLCB+MM),结合了 2 个分支的优点,比 PLCB 准确率有所提升,但相对原型注意力层带来的提升效果较弱。将 2 种结构同时加入后(PAIPL-P)得到了最好的结果。

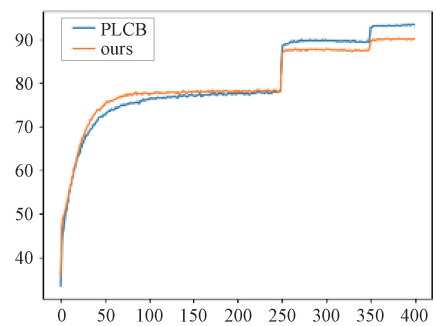
4 实验结果分析

首先展示 PAIPL 的有效性。图 4 比较了 PLCB 和 PAIPL-P 精度曲线和损失曲线。PLCB 的训练损失持续下降,测试精度在学习率第 2 次下降前持续上升,而在学习率第 2 次下降后,测试精度反而下降了,而 PAIPL-P 的精度不断提高。这说明在训练后期 PLCB 出现了过拟合问题,这意味着 PAIPL-P 提供了比 PLCB 更充分的正则化。

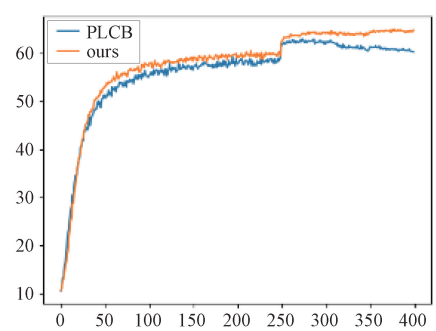
然后展示 PAIPL-P 的 t-SNE[32] 降维可视化结果。在 500 个有标签样本的 CIFAR-10 上训练模型,并将所有测试样本映射到特征空间。特征



(a) 训练损失曲线



(b) 训练精度曲线



(c) 测试精度曲线

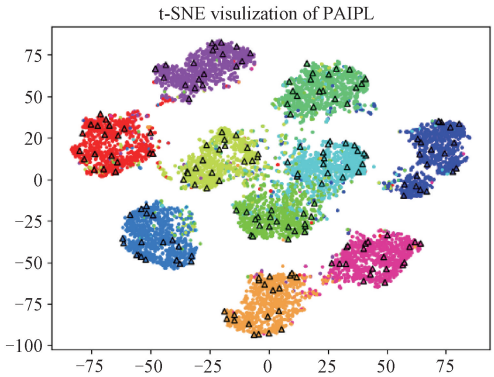
实验在 4 000 个有标签数据的 CIFAR-100 上进行。

图 4　PLCB 和 PAIPL 的曲线对比

Fig. 4　Comparison curves between PLCB and PAIPL (ours)

向量和原型向量都用 t-SNE 映射到二维空间。图 5 显示了 t-SNE 的可视化结果。所有的原型都位于它们所属的真实聚类中,并且较为均匀的分布在整个聚类区域。这表示学习的原型能用很小的成本,较好地表达大部分的数据流形。PAIPL 利用学习到的原型捕捉到数据流形的全局压缩信息,而基于 MixUp 的方法,如 MixMatch 和 PLCB,只能使用成对数据信息。

图 6 展示了飞机这个大类中,不同特征空间中原型附近的样本图像。首先用带 500 个有标签数据的 CIFAR-10 上训练模型,并将所有测试样本映射到特征空间。每个原型的 2 个近邻样本被挑选出来,每个原型周围的图像非常相似,而不同原型周围的图像,虽然来自同一类,看起来差异更



不同颜色表示不同的类别,三角形表示学到的原型。

图 5　测试样本特征的 t-SNE 可视化结果

Fig. 5　Visualization result of t-SNE of test sample features

大。这表示 PAIPL 学到的原型可以看作是子聚类的中心,原型注意力层可以看作是细粒度分类。这对于更复杂的数据集更加重要,所以 PAIPL 在 CIFAR-100 的改进比在 CIFAR-10 上的改进更加明显。

(a) 第1原型近邻样本　(b) 第2原型近邻样本

(c) 第3原型近邻样本　(d) 第4原型近邻样本

图 6　飞机图像部分原型近邻样本

Fig. 6　Images near different prototypes of airplane class

最后讨论相互混合监督学习的作用机理。PAIPL 中有 2 个分支,主分支是加入原型注意力层的传统的前馈神经网络,在训练早期,该分支会在伪标签的监督下快速收敛。然而,随着训练的进行,它将受到认知偏误的影响。另一个分支是原型分配,在训练早期,由于原型尚未充分训练,该分支使用质量较差的原型进行预测,结果较差。随着训练的进行,原型会得到更好的训练,该分支会变得更强。但训练后期若只使用原型分配分支又会出现预测过于平滑的现象。所以本文使用线性预热来获得伪标签。在训练初期,每个分支更倾向于受到来自该分支的监督。而在预热过程结束后,用 2 个分支预测的随机线性组合对每个分支进行训练。

5　结论

本文提出一种新型的特征修正模型 PAL。这种特征修正模型可以广泛应用在伪标签半监督学

习框架中,并与相互混合监督结合,得到基于原型学习改进的伪标签半监督学习算法。PAIPL 包含 2 部分:1)用于改善特征的可学习的原型注意力层;2)用于结合修正特征伪标签和原型分配伪标签的相互混合监督。本文将 PAIPL 算法应用到 2 种不同的伪标签半监督学习框架上,软伪标签的自训练框架和伪标签的 PLCB 框架,得到 2 种新的伪标签半监督学习算法 PAIPL-S 和 PAIPL-B。实验结果显示 PAIPL-P 优于最新的伪标签方法和一致性正则化方法。根据本文的研究可以看出伪标签方法可以和一致性训练方法一样,在半监督学习中起到重要作用。未来的工作可以使用更大批量的数据和更强的图像预处理来获得更好的效果,也可以考虑将自监督学习的成果移植到半监督学习中。

参考文献

[1] Ouali Y, Hudelot C, Tami M. An overview of deep semi-supervised learning[EB/OL]. arXiv: 2006.05278. (2020-07-06)[2021-04-15]. https://arxiv.org/abs/2006.05278.

[2] Oliver A, Odena A, Raffel C, et al. Realistic evaluation of deep semi-supervised learning algorithms[EB/OL]. arXiv: 1804.09170. (2019-06-17)[2021-04-15]. https://arxiv.org/abs/1804.09170.

[3] Chen T, Kornblith S, Swersky K, et al. Big self-supervised models are strong semi-supervised learners[EB/OL]. arXiv: 2006.10029. (2020-10-26)[2021-04-15]. https://arxiv.org/abs/2006.10029.

[4] Xie Q Z, Luong M T, Hovy E, et al. Self-training with noisy student improves ImageNet classification[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 13-19, 2020, Seattle, WA, USA. IEEE, 2020: 10684-10695.

[5] Ibrahim M S, Vahdat A, Ranjbar M, et al. Semi-supervised semantic image segmentation with self-correcting networks[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 13-19, 2020, Seattle, WA, USA. IEEE, 2020: 12712-12722.

[6] Ouali Y, Hudelot C, Tami M. Semi-supervised semantic segmentation with cross-consistency training[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 13-19, 2020, Seattle, WA, USA. IEEE, 2020: 12671-12681.

[7] He J X, Gu J T, Shen J J, et al. Revisiting self-training for neural sequence generation[EB/OL]. arXiv: 1909.13788. (2020-10-18)[2021-04-15]. https://arxiv.org/abs/1909.13788.

[8] Chen L X, Ruan W T, Liu X Y, et al. SeqVAT: virtual adversarial training for semi-supervised sequence labeling[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 8801-8811.

[9] Li Y T, Liu L, Tan R T. Decoupled certainty-driven consistency loss for semi-supervised learning[EB/OL]. arXiv: 1901.05657. (2020-07-31)[2021-04-15]. https://arxiv.org/abs/1901.05657.

[10] Tarvainen A, Valpola H. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results[EB/OL]. arXiv: 1703.01780. (2018-04-16)[2021-04-15]. https://arxiv.org/abs/1703.01780.

[11] Arazo E, Ortego D, Albert P, et al. Pseudo-labeling and confirmation bias in deep semi-supervised learning[C]//2020 International Joint Conference on Neural Networks (IJCNN). July 19-24, 2020, Glasgow, UK. IEEE, 2020: 1-8.

[12] Zhang H, Cisse M, Dauphin Y N. Mixup: beyond empirical risk minimization[EB/OL]. arXiv: 1710.09412. (2018-04-27)[2021-04-15]. https://arxiv.org/abs/1710.09412.

[13] Veličković P, Cucurull G, Casanova A, et al. Graph attention networks[EB/OL]. arXiv: 1710.10903. (2018-02-04)[2021-04-15]. https://arxiv.org/abs/1710.10903.

[14] Riloff E, Wiebe J. Learning extraction patterns for subjective expressions[C]//Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing. July 11-12, 2003, Sapporo, Japan. Association for Computational Linguistics, 2003: 105-112.

[15] Laine S, Aila T. Temporal ensembling for semi-supervised learning[EB/OL]. arXiv: 1610.02242. (2017-03-15)[2021-04-15]. https://arxiv.org/abs/1610.02242.

[16] Ke Z H, Wang D Y, Yan Q, et al. Dual student: breaking the limits of the teacher in semi-supervised learning[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). October 27-November 2, 2019, Seoul, South Korea. IEEE, 2019: 6728-6736.

[17] Miyato T, Maeda S I, Koyama M, et al. Virtual adversarial training: a regularization method for supervised and semi-supervised learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(8): 1979-1993.

[18] Verma V, Lamb A, Kannala J, et al. Interpolation consistency training for semi-supervised learning[C]//Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence. August 10-16, 2019. Macao, China. California: International Joint Conferences on Artificial Intelligence Organization, 2019: 3635-3641.

[19] Berthelot D, Carlini N, Goodfellow I J, et al. MixMatch: a holistic approach to semi-supervised learning[EB/OL]. arXiv: 1905.02249. (2019-10-23)[2021-04-15]. https://arxiv.org/abs/1905.02249.

[20] Xie Q Z, Dai Z H, Hovy E, et al. Unsupervised data

augmentation for consistency training [EB/OL]. arXiv: 1904.12848. (2020-11-5) [2021-04-15]. https://arxiv.org/abs/1904.12848.

[21] Lee D H. Pseudo-label: the simple and efficient semi-supervised learning method for deep neural networks [EB/OL]. (2013-07) [2021-04-15]. https://www.researchgate.net/publication/280581078_Pseudo-Label_The_Simple_and_Efficient_Semi-Supervised_Learning_Method_for_Deep_Neural_Networks.

[22] Zhuang C X, Zhai A, Yamins D. Local aggregation for unsupervised learning of visual embeddings[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). October 27-November 2, 2019, Seoul, South Korea. IEEE, 2019: 6001-6011.

[23] Kuo C W, Ma C Y, Huang J B, et al. Manifold graph with learned prototypes for semi-supervised image classification [EB/OL]. arXiv: 1906.05202. (2019-06-13) [2021-04-15]. https://arxiv.org/abs/1906.05202.

[24] Srivastava N, Hinton G E, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research, 2014, 15: 1929-1958.

[25] Salimans T, Kingma D P. Weight normalization: a simple reparameterization to accelerate training of deep neural networks [EB/OL]. arXiv: 1602.07868 (2016-06-04) [2021-04-15]. https://arxiv.org/abs/1602.07868.

[26] Tanaka D, Ikami D, Yamasaki T, et al. Joint optimization framework for learning with noisy labels[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. June 18-23, 2018, Salt Lake City, UT, USA. IEEE, 2018: 5552-5560.

[27] Grandvalet Y, Bengio Y. Semi-supervised learning by entropy minimization [C]// NIPS'04: Proceedings of the 17th International Conference on Neural Information Processing Systems. December 13-18, 2004, Vancouver, British Columbia, Canada. MIT Press, 2004: 529-536.

[28] Iscen A, Tolias G, Avrithis Y, et al. Label propagation for deep semi-supervised learning [C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). June 15-20, 2019, Long Beach, CA, USA. IEEE, 2019: 5065-5074.

[29] Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations [EB/OL]. arXiv: 2002.05709. (2020-07-01) [2021-04-15]. https://arxiv.org/abs/2002.05709.

[30] Zhang Y, Xiang T, Hospedales T M, et al. Deep mutual learning[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. June 18-23, 2018, Salt Lake City, UT, USA. IEEE, 2018: 4320-4328.

[31] Athiwaratkun B, Finzi M, Izmailov P, et al. There are many consistent explanations of unlabeled data: why you should average [EB/OL]. arXiv: 1806.05594. (2019-02-21) [2021-04-15]. https://arxiv.org/abs/1806.05594.

[32] Van der Maaten L, Hinton G. Visualizing data using t-SNE [J]. Journal of Machine Learning Research, 2008, 9(86): 2579-2605.