基于伪标签的可防御稳定网络

刘佳美,孙 涵,林 磊
(南京航空航天大学 计算机科学与技术学院/人工智能学院,江苏 南京 211106)

摘 要:针对域自适应问题中无法较好地同时提升模型迁移能力和防御攻击能力导致其在目标域中不稳定且易受攻击的问题,提出了一种基于伪标签的可防御稳定网络。在条件域对抗网络的框架下,首先通过高斯混合模型对经过预训练输出的源域特征和目标域特征进行共同聚类,得到基于类别概率的软伪标签来引入更为可靠的目标域信息,以拉近两域之间的距离;接着将源域和目标域数据输入学生网络和教师网络,教师网络参数根据历史上学生网络参数通过指数移动平均方法迭代更新,通过约束特征的类内一致性以减轻错误的伪标签带来的不利影响;与此同时,采用主动防御的思想,在训练中增加源域的对抗样本,使模型学习到更鲁棒的特征,提高其在目标域数据对对抗攻击的防御能力。在 Office-31 数据集上的实验结果表明,所提出的基于伪标签的可防御稳定网络能够有效提高模型的迁移能力和防御能力,从两个不同的方面提高了网络的鲁棒性。

关键词:域自适应;聚类算法;伪标签;平均教师模型;主动防御
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2022)06-0034-05
doi:10.3969/j.issn.1673-629X.2022.06.006

Pseudo-label Based Defensible Stable Network

LIU Jia-mei, SUN Han, LIN Lei
(School of Computer Science and Technology/Artificial Intelligence, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China)

Abstract: Aiming at the problem that the migration ability and defense ability of the model cannot be improved at the same time in the domain adaptation, which makes it unstable and vulnerable to attack in the target domain, a defensible and stable network based on pseudo label is proposed. Under the framework of conditional domain adaptation network, firstly, the pre-trained source domain features and target domain features are co-clustered by Gaussian mixture model, and soft pseudo-label based on category probability is obtained to introduce more reliable target domain information, so as to shorten the distance between the two domains. Then the data are input into the student network and teacher network. According to the student network parameters, the teacher network parameters are updated by the exponential moving average to reduce the adverse effects of false labels. At the same time, using the idea of active defense, the model can learn more robust features and improve the defense ability of the data in the target domain against the attack by adding the source domain adversarial samples in the training. Experimental results on Office-31 show that the proposed algorithm can effectively improve the migration ability and defense ability of the model, and improve the robustness of the network from two different aspects.

Key words: domain adaptation; clustering algorithm; pseudo label; mean teacher model; active defense

0 引 言

随着算力和数据的急剧增加,如今的计算机视觉任务都已经获得了较好的效果。然而大多数任务仅局限于同一分布的数据,当更换数据集时常常需要重新训练模型,且对于新任务的标签标记耗时费力,常常难以获得。由此,域自适应问题应运而生。域自适应中通常包含一个有标签数据的源域和一个无标签数据的目标域,两域之间数据分布相关但不同[1]。域自适应旨在克服域偏移[2],将在源域中学习的知识迁移到目标域能够有较好的预测。而为了解决目标域数据没有标签的问题,Cao 等人[3]引入了半监督学习中伪标签的思想,采用基于原型聚类进行两域对齐;Nie[4]考虑了聚类对齐与数据平衡之间的关系;Dai 等人[5]则是使用基于密度聚类后通过使用对比损失解决特征表示

收稿日期:2021-07-10 **修回日期**:2021-11-11
基金项目:中央高校基本科研业务费专项资金资助项目(NZ2019009)
作者简介:刘佳美(1999-),女,硕士研究生,CCF 会员(H6893G),研究方向为计算机视觉;孙 涵,博士,副教授,CCF 会员(33361M),研究方向为图像处理、计算机视觉。

的不连续问题。但是由于伪标签的生成常常伴随着噪声,过度自信的伪标签与真实值不符时反而对模型学习带来更多的困难。为了解决这个问题,Zou 等人[6]从正则化的角度出发,将伪标签作为交替优化的连续潜变量联合优化,通过网络自训练实现标签正则化和模型正则化。另一方面,可以抵御对抗攻击的域自适应方法也在探索中,Zhang 等人[7]提出了通过代理损失最小化将干净数据上的模型精度和鲁棒性分离为两个损失项来训练更鲁棒的模型。

受文献[6]启发,笔者从标签和模型两方面减轻伪标签可能带来的负面影响,通过基于概率的聚类对每个样本生成该样本属于不同类别的概率,实现聚类时就生成软伪标签,减少标签正则化的过程。再者,通过平均教师模型,让教师网络集成学生网络的历史参数,再通过一致性损失达到模型正则化的目的,有利于输出结果更加稳定。最后,通过训练时增加对抗样本,在保证模型的域自适应能力有所提升的同时,更大限度地提高其防御攻击能力。

该文贡献如下:

(1)提出了基于伪标签的可防御稳定网络(pseudo-label based defensible stable network,PDSN),在经典域对抗网络框架下增加了基于概率的伪标签生成模块,并通过平均教师模型减轻错误的伪标签带来的噪声,使结果更稳定。

(2)结合抵御对抗攻击的方法,增加训练时的对抗样本,在不大幅减少网络迁移能力的情况下,实现稳定能力和防御能力均有所提升的可靠网络。

(3)在 Office-31 数据集上进行了实验,与其他域自适应方法结果进行对比分析,证明了该方法的有效性。

1　相关工作

1.1　对抗判别网络

对抗判别算法主要利用对抗训练来学习域不变特征和判别能力,通过判别器判别源域或者是目标域,借此使目标域边缘特征分布与源域对齐。2016 年 Ben-David 等[8]开创性地提出在神经网络之上,使用对抗训练的方法学习域不变的特征和判别能力。在此思想之上,域对抗神经网络算法(DANN)[9]利用标准反向传播和随机梯度下降的单前馈网络实现域对抗训练。在此之后,陆续提出了更多的域对抗训练思想的方法,如对抗判别域自适应方法(ADDA)[10]反向标记 GAN 损失,将优化器分离成两个,用于生成器和鉴别器。

条件域对抗网络(CDAN)[11]则是考虑了怎样对齐两域之间的条件分布,利用多线性映射对特征和类别进行联合域自适应,引入熵作为权重系数调节极大极小优化方法。其损失由分类损失和对抗损失两个部分组成。分类损失值即源域图像本身拥有的真实值和经过网络产生的预测值之间的交叉熵损失。对抗损失即对抗网络判别样本为源域图像还是目标域图像的二分类交叉熵损失。以下为 CDAN 定义的极大极小问题。

$$\min_{G}\max_{D}\varepsilon(G)-\lambda\varepsilon(D,G) \tag{1}$$

$$\varepsilon(G)=\mathbb{E}_{(x_i^s,y_i^s)\sim D_s}\mathcal{L}(G(x_i^s),y_i^s) \tag{2}$$

$$\varepsilon(D,G)=-\mathbb{E}_{x_i^s\sim D_s}\log[D(T(h_i^s))]-\mathbb{E}_{x_j^t\sim D_t}\log[1-D(T(h_j^t))] \tag{3}$$

式中,λ 为权衡超参数,$\mathcal{L}(\cdot,\cdot)$ 为交叉熵损失,G 为源域分类器,D 指域鉴别器与类别联合变量 h 的多线性映射。CDAN 通过以上极大极小策略能达到对齐两域条件分布的目的。

1.2　对抗攻击与主动防御

自 Szegedy 等人[12]提出深度神经网络容易被对抗样本欺骗产生错误的预测,越来越多的对抗攻击和主动防御的问题被人们关注[13-14]。在图像分类中,对抗攻击的经典方法有 L-BFGS 方法和快速梯度攻击方法等等[15]。在对抗攻击的方法中,通常开始扰动前,设定一定的扰动步长,迭代数次达到对原图像产生足够的干扰的目的。对抗攻击的测试方法可分为黑盒测试和白盒测试两种。黑盒测试因为对模型的未知,测试时主要依据的方法是输入和输出数据之间的关系,白盒测试则是从模型本身进行测试[16]。与对抗攻击相对应的主动防御的方法[17]主要有修改训练过程或样本、修改网络、使用附加网络三种方式。而该文就是通过在训练的过程中不断增加源域的对抗样本,以提升网络的鲁棒性。

2　基于伪标签的可防御稳定网络

2.1　总体结构

如图 1 所示,提出的网络结构由三个部分组成。

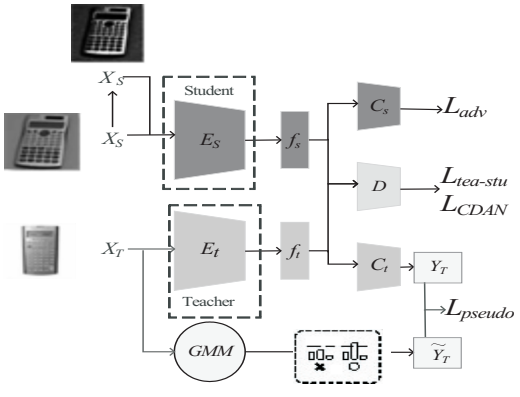

图 1　PDSN 结构

首先采用基于概率的聚类方法对源域和目标域的特征进行聚类,通过比较样本在不同类别上的概率值

对伪标签值进行选择,得到软伪标签,减少错误的硬伪标签带来的噪声影响。再通过最小化软伪标签 $y_t^{'}$ 和目标输出 $\hat{y}_t$ 之间的距离损失 $\mathcal{L}_{\text{pseudo}}$ 减少结果差异;其次将数据经过平均教师模型,让学生网络正常训练,教师网络通过对学生网络历史参数集成后更新优化,再通过对损失 $\mathcal{L}_{\text{tea-stu}}$ 的最小化让学生网络学习到较为稳定的分类结果;最后通过主动防御模块最小化生成的源域的对抗样本和正常图像之间的距离,增强网络的鲁棒性。其总体损失为 CDAN 框架下的原损失与以上三部分损失之和:

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{CDAN}} + \mathcal{L}_{\text{pseudo}} + \mathcal{L}_{\text{tea-stu}} + \beta\mathcal{L}_{\text{adv}} \tag{4}$$

2.2 软伪标签生成

基于距离的无监督聚类方法通常根据距每个样本距离中心点的距离进行类别划分,基于密度的方法将密度相连的点进行集合形成簇的聚类。不同于以上将一个样本直接划分为某个确定的类别,仿佛“非此即彼”这样硬聚类的方法。混合高斯模型(GMM)的聚类方法属于软聚类,认为每个样本可以属于多个类,与类协同训练有异曲同工之妙。将数据通过 GMM 聚类时,每个样本都会计算其属于某个类别的概率值向量 x^K,其中 K 为目标域类别,计算公式为:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \sigma_k) \tag{5}$$

其中,π_k 是混合系数,$\mathcal{N}(\cdot)$ 代表其每个类别的高斯混合模型参数。该文将源域数据和目标域数据经特征提取器后将特征输入 GMM 聚类,在向量 x^K 中选择概率值最大且大于 0.5 的类别作为伪标签。

通过最小化目标域样本聚类后产生的伪标签 $\tilde{y}_i^t$ 与模型预测值之间的平均绝对误差作为损失。

$$\mathcal{L}_{\text{pseudo}} = \frac{1}{N_0} \sum_{i=1}^{N_t} w(x_i^t)\, \mathcal{J}(C(F(x_i^t))), \tilde{y}_i^t) \tag{6}$$

式中,$\mathcal{J}(\cdot,\cdot)$ 是平均绝对误差 MAE,$w(x_i^t)$ 表示经过 GMM 聚类后生成该类伪标签的概率值,当聚类概率小于 0.5 时,将 $w(x_i^t)$ 置为零,即代表丢弃置信度较低的伪标签,使置信度较高的伪标签与模型预测结果进行损失计算。

2.3 伪标签噪声减弱

为了进一步减少错误的伪标签可能带来的负迁移,考虑通过让网络本身学习到更稳定的网络参数,为此使用了平均教师模型[18]。该模型使用两个模型结构相同的网络进行训练,分别为学生网络和教师网络。学生网络参数根据梯度下降法更新得到;教师网络参数根据历史上学生网络参数通过指数移动平均方法(EMA)加权迭代并不断反向传播更新,让教师网络的预测训练学生网络,即通过最小化学生网络的预测结果和教师网络的预测结果之间的 L2 损失:

$$\mathcal{L}_{\text{tea-stu}} = \min \sum_{i=1}^{n} (C_{\text{tea}}(F_{\text{tea}}(x_i)) - C_{\text{stu}}(F_{\text{stu}}(x_i)))^2 \tag{7}$$

式中,$F(\cdot)$ 为网络的特征提取部分,$C(\cdot)$ 为网络的分类器。两者之间相互促进,形成良好的循环,达到使网络输出更稳定,减弱伪标签噪声的目的。

2.4 主动防御

该文在训练过程中对源域数据进行对抗样本的生成,将其同样加入训练数据集[7],得到其预测结果。能在一定程度上抵御对抗攻击,使网络更加鲁棒。在基于伪标签的可防御稳定网络中,将原图像的预测分布 $p(x_i^s)$ 和对抗样本的预测结果 $q(\hat{x}_i^s)$ 通过计算 KL 散度来表示二者之间的拟合程度,增加对抗样本损失:

$$\mathcal{L}_k = D_{\text{KL}}(p \mid\mid q) = \sum_{i=1}^{n} q(\hat{x}_i^s) \log \frac{p(x_i^s)}{q(\hat{x}_i^s)} \tag{8}$$

上述方法的目的是能够使分类决策边界同时远离对抗样本和正常样本,减小因为扰动带来的不确定的结果,鼓励网络变得更加可防御。这样能够促使网络被成功攻击的概率降低,提高网络的安全性和鲁棒性。

3 实验结果与分析

3.1 实验数据

主要在 Office-31 数据集下展开实验。Office-31 数据集[19]是域自适应问题中的基准数据集,其中共包含 3 个子数据集,每个子数据集中都拥有自行车、键盘、背包等 31 个类别的物体。如图 2 所示,不同子数据集的拍摄背景和物体都各有差异。

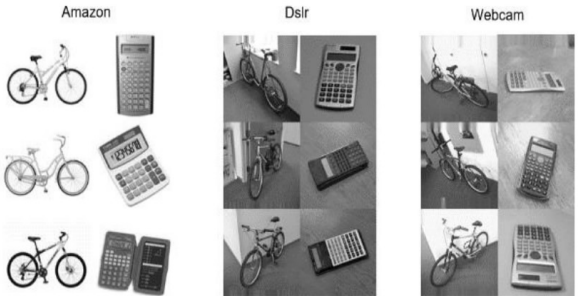

图 2 Office-31 数据集示例

Amazon 域中的图片下载自亚马逊网站,背景干净,物体拍摄视角通常为正面,画质清晰,分辨率为 300 ×300,图片数量最多;DSLR 图片域由不同摄影设置的数码单反相机拍摄,其中的物品均放置在现实生活场景中,背景复杂且拍摄视角多变,不再是单一的正面视角,分辨率多为 1 000 ×1 000,图片数量最少;Webcam 域由网络摄像头拍摄,其图片同样背景复杂,画质相比 DSLR 数据集较为模糊,分辨率在 600 ×600 左右。在三个两两组合的六对迁移域 A→W,D→W,W→D,A→D,D→A 和 W→A 中,困难程度各有不同。

3.2　实验设置

将该方法与 Source－only 方法、DAN、DANN、CDAN 进行实验对比。Source-only 使用 Resnet-50 作为骨干网络,在训练的过程中仅使用交叉熵损失,目标域数据不参与训练,并且将其训练模型不做任何修改在目标域数据集中进行测试。DAN 是典型的基于差异的方法,使用了 8 层的 AlexNet 网络,而在分类层之前增加了多个适配层,并且使用了最大平均差异(MK-MMD)的多个核变量的总和,计算在源域上的分类损失和适配层的差异损失。DANN 是传统与对抗网络,利用标准反向传播和随机梯度下降的单前馈网络实现域对抗训练。

实验中的超参数设置主要为,骨干网络使用 Resnet-50,epochs=6,每个 epoch 迭代 iters=5 000,统一设置 batch-size 为 12,学习率为 0.001。此外主动防御模块扰动为 0.3,步长设置为 0.01。实验中调整参数扰动步数 pertub-steps=5,10,20,损失函数中 β =0,0.3,0.5,0.8,1,关于生成源域对抗图像的 distance,与文献[7]方法一致选择 L2 或者 INF。

3.3　评价标准

通过对目标域正常样本的测试的准确率的计算来评估对模型的迁移能力:

$$\text{Accuracy} = \frac{\sum_{i=1}^{K} \text{TP}_i}{\sum_{i=1}^{K} (\text{TP}_i + \text{FN}_i)} \tag{9}$$

对于算法鲁棒性的定量分析,采用 PGD 方法(project gradient descent)进行白盒测试,评估指标为在攻击测试时的准确率。

3.4　结果分析

3.4.1　算法结果对比

从表 1 中可以看出,PDSN 与其他方法相比在 6 个域组合中都获得了明显的提升,相较于 CDAN 平均提升 1.67%,在 D→A 和 W→A 两个域组合上更为明显。这是因为软伪标签的生成和平均教师模型方法让大量的 A 域的数据获得了平滑而正确的伪标签,原本因为图像本身导致的域偏移以及图像数量的不平衡问题得到明显的改善。

表 1　在 Office-31 上不同算法的结果准确度　　%

方法	A→D	D→A	W→A	W→D	A→W	D→W	平均
source-only	81.12	65.01	64.32	99.19	77.23	96.13	80.5
DAN	85.95	66.85	64.23	100	86.78	98.63	83.74
DANN	83.33	72.74	71.12	100	90.41	97.66	85.86
CDAN	90.80	73.73	72.61	100	92.62	98.67	88.07
PDSN(Ours)	91.57	75.65	77.21	100	93.58	99.25	89.54

3.4.2　消融实验与参数对比实验

表 2 中,“S”表示使用软伪标签方法和平均教师方法,“D”表示加入了主动防御的对抗样本损失。可以验证在 PDSN 方法中,伪标签和平均教师模型是改进方法中提升准确率的主要原因。而为了提高鲁棒性进行主动防御加入了对抗样本的损失,对于大部分域组合的准确率有一定程度的影响,却能够提高网络的抗对抗性。

表 2　改进技术结果准确率对比　　%

S	D	A→D	D→A	W→A	W→D	A→W	D→W	平均
×	×	90.80	73.73	72.61	100	92.62	98.67	88.07
×	√	90.78	74.65	75.97	100	89.94	99.25	88.93
√	×	91.76	76.89	77.17	100	92.96	99.25	89.67
√	√	91.57	75.65	77.21	100	93.58	99.25	89.54

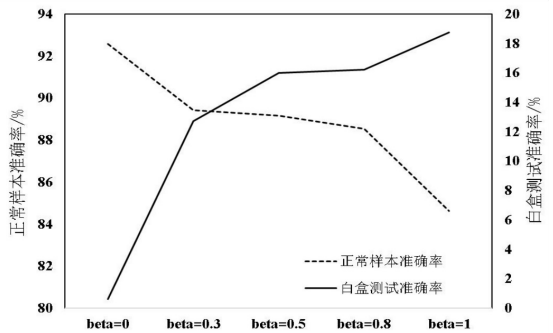

图 3　不同 beta 取值实验结果

图 3 代表了权衡超参数 β 对网络迁移能力和防御能力的影响,实验为在 A→W 域组合中进行。

此外,对不同的 distance 和 beta 参数做了更多的实验,测试了不同的取值,结果如表 3 所示。

表 3　不同 distance 选择实验

Distance	Beta	正常样本准确率/%	白盒测试准确率/%
L2	0.3	93.58	8.46
INF	0.3	89.43	12.75
L2	0.5	91.45	9.75

续表 3

Distance	Beta	正常样本准确率/%	白盒测试准确率/%
INF	0.5	89.18	16.00
L2	0.8	91.45	10.93
INF	0.8	88.55	16.25
L2	1	90.94	14.65
INF	1	84.65	18.75

4 结束语

探讨了怎样更好地提高域自适应模型的迁移能力和防御能力,提出了基于伪标签的可防御网络,采用软伪标签和平均教师模型使网络输出更加稳定可靠,利用主动防御方法有效地减轻了对抗样本攻击的影响。在 Office-31 数据集上进行了对比实验、消融实验和参数对比实验,验证了此网络提高鲁棒能力的有效性。此外,该方法由于源域对抗样本生成部分存在训练时间长的问题,未来研究中,将尝试通过对源域数据随机采样来生成对抗样本等方法解决这一问题。

参考文献:

[1] MORI S. Domain adaptation in natural language processing [J]. The Japanese Society for Artificial Intelligence,2012,27(4):365-372.

[2] TORRALBA A, EFROS A A. Unbiased look at dataset bias [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Providence, RI: IEEE, 2011: 1521-1528.

[3] 曹俊年. 基于深度学习的图像域适应问题的研究[D]. 成都:电子科技大学,2020.

[4] 聂玲俐. 基于聚类与平衡的无监督领域适应方法研究[D]. 南京:南京邮电大学,2020.

[5] DAI Z, WANG G, ZHU S, et al. Cluster contrast for unsupervised person re-identification[J]. arXiv:2103.11568,2021.

[6] ZOU Y, YU Z, KUMAR B, et al. Domain adaptation for semantic segmentation via class-balanced self-training[C]//Proceedings of the European conference on computer vision. Germany: Springer, 2018: 289-305.

[7] ZHANG H, YU Y, JIAO J, et al. Theoretically principled trade-off between robustness and accuracy[C]//Proceedings of the international conference on machine learning. New York: ACM, 2019: 7472-7482.

[8] BEN-DAVID S, BLITZER J, CRAMMER K, et al. A theory of learning from different domains[J]. Machine Learning, 2010, 79(1): 151-175.

[9] GANIN Y, USTINOVA E, AJAKAN H, et al. Domain-adversarial training of neural networks[J]. The Journal of Machine Learning Research, 2016, 17(1): 2096-2030.

[10] TZENG E, HOFFMAN J, SAENKO K, et al. Adversarial discriminative domain adaptation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu, HI, USA: IEEE, 2017: 2962-2971.

[11] LONG M, CAO Z, WANG J, et al. Conditional adversarial domain adaptation[C]//Proceedings of the 32nd international conference on neural information processing systems. Montréal, Canada: [s. n.], 2018: 1647-1657.

[12] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[J]. arXiv:1312.6199, 2013.

[13] 陈晋音,张敦杰,黄国瀚,等. 面向图神经网络的对抗攻击与防御综述[J]. 网络与信息安全学报,2021,7(3):1-28.

[14] 李明慧,江沛佩,王 骞,等. 针对深度学习模型的对抗性攻击与防御[J]. 计算机研究与发展,2021,58(5):909-926.

[15] 王 超,魏祥麟,田 青,等. 基于特征梯度的调制识别深度网络对抗攻击方法[J]. 计算机科学,2021,48(7):25-32.

[16] 魏运清. 基于 RNN 结构深度学习系统的白盒自动化测试方法的研究[D]. 济南:山东大学,2019.

[17] 侯 勇,郑钰炜. 深度学习中的对抗防御算法研究[J]. 滁州学院学报,2021,23(2):10-20.

[18] TARVAINEN A, VALPOLA H. Weight-averaged, consistency targets improve semi-supervised deep learning results[J]. arXiv:1703.01780, 2017.

[19] SAENKO K, KULIS B, FRITZ M, et al. Adapting visual category models to new domains[C]//Proceedings of the European conference on computer vision. Germany: Springer, 2010: 213-226.