doi:10.3772/j.issn.1002-0470.2020.07.005

基于鉴别模型和对抗损失的无监督域自适应方法①

赵文仓② 袁立镇 徐长凯

(青岛科技大学自动化与电子工程学院 青岛 266061)

摘 要 对于许多任务而言,收集注释良好的图像数据集来训练深度学习算法成本过高且耗时,而仅在渲染图像训练的模型通常无法推广到真实图像。针对上述问题,无监督域自适应算法试图在2个域之间映射一些表示或提取域不变的特征,将2个域映射到共同的特征空间。本文结合源域的有标签数据和目标域的无标签数据,提出了基于生成对抗网络(GAN)架构的无监督域自适应方法。方法使用鉴别模型,无需权重共享、对抗损失和辅助分类任务,以无监督的方式学习从一个域到另一个域的变换。对抗鉴别的无监督域自适应方法能有效减少训练域和测试域分布之间的差异,减轻域移位的有害影响,并显著地提高识别率。实验结果证明对抗鉴别方法比其他域自适应方法更有效且更简单,扩充样本的同时提高了网络的泛化性能。

关键词 深度学习;无监督;域自适应;生成对抗网络(GAN);辅助分类任务

0 引 言

深度前馈架构为计算机视觉及其他领域的各种任务带来了深刻的先进技术。只有当有大量标记的训练数据可用时,才会出现这些性能上的飞跃。深度卷积网络在大规模数据集上训练时,可以学习各种任务和视觉领域中通用的表示[1]。然而,由于数据集偏差或域移位[2]的现象,在大型数据集上与这些表示一起训练的识别模型不能很好地推广到新的数据集和任务[3]。

上述问题的解决方案是无监督域自适应方法。域自适应方法试图减轻域移位的有害影响。最近的域自适应方法学习深度神经变换,将2个域映射到共同的特征空间。这通常通过优化表示以最小化域移位的一些度量来实现,例如最大平均差异(maximum mean discrepancy, MMD)[4]或相关距离[5]。另一种方法是从源表示中重建目标域[6]。在机器翻译中,丁亮等人[7]将 Bi-LSTM 用于构建自动编码器,有效翻译系统的性能。曾远柔等人[8]通过优化非线性映射函数来对齐子空间和目标子空间,用界标无人管理域自适应法来实现。Ganin 等人[9]引入梯度反转层,将梯度乘以小的负数,以训练特征提取器使域分类器不能区分源域和目标域。Tzeng 等人[10]考察了用于半监督域自适应的类似设置。该方法不是采用梯度反转层以直接最大化域分类器的损失,而是最大化域混淆以"最大程度地混淆"域分类器。当域分类器在二进制标签上输出均匀分布时,它是"最大混淆的",这表明域分类器不能确定输入图像的学习特征表示是来自源域还是目标域,通过加入软标签损失,用来保持源域和目标域各类之间相对分布的一致性。

虽然这些方法已经取得了良好的进展,但它们仍然不能与仅在目标领域进行训练的纯监督方法相提并论。生成对抗网络(generative adversarial network, GAN)[11]优于其他生成方法的优点是其在训练期间不需要复杂的采样或推理,对抗性方法寻求

① 国家自然科学基金(61171131),山东省重点研发计划(2013YD01033)和国家留学基金委项目(201608370049)资助。

② 男,1973年生,博士,教授;研究方向:模式识别与智能系统;联系人,E-mail: zhao_coinslab@outlook.com

(收稿日期:2019-07-19)

通过关于域鉴别符的对抗性目标来最小化近似域差异距离。针对上述问题,本文提出了一种基于鉴别模型和对抗损失的无监督域适合方法,该方法在MNIST、MNIST-M 和 SVHN 数字数据集上实现了最先进的视觉自适应结果。为了更好地验证对抗鉴别方法,本文将该方法在较复杂的 2 组遥感影像数据集上进行适应。对抗鉴别方法与现有方法相比具有的优势为与特定任务的体系结构分离,跨标签空间的泛化以及训练稳定等。

1 生成对抗网络

根据生成对抗网络对抗训练生成逼真图像的思想,本文提出了基于对抗网络的域自适应框架,如图 1所示。首先使用源域中的标签学习鉴别表示,然后使用通过域-对抗性损失学习的非对称映射将目标数据映射到同一空间的单独编码。以无监督的方式学习鉴别表示,运用无权重共享、对抗性损失以及辅助分类任务。

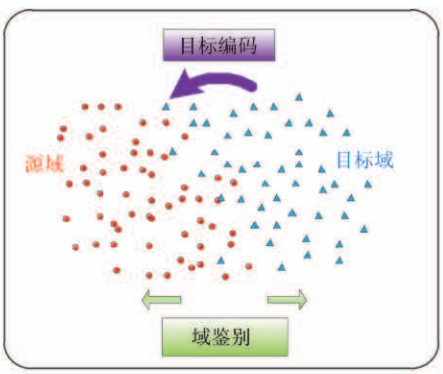


图 1 结合鉴别模型的无监督域自适应方法

1.1 GAN 架构

使用 Goodfellow 等人的符号,定义了 2 个网络之间的极小极大博弈所使用的值函数 $V(G, D)$:

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log D(x)] + E_{z\sim p_z(z)}[\log(1 - D(G(z)))] \tag{1}$$

其中,$x \sim p_{data}(x)$ 从实数据分布中抽取样本,$z \sim p_z(z)$ 从输入噪声中抽取样本,$D(x;\theta_d)$ 是鉴别器,$G(z;\theta_g)$ 是生成器。如式(1)所示,目标是找到参数 θ_d,其最大化正确区分真样本 x 和假样本 $G(z)$ 的对数概率,同时找到最小化对数概率 $1 - D(G(z))$ 的参数 θ_g。表达式 $D(G(z))$ 表示生成的数据 $G(z)$ 被鉴别为真的概率。如果鉴别器正确地对假输入进行分类,则 $D(G(z)) = 0$。目标是使 $D(G(z))$ 越大越好,即以假乱真。所以使数值 $1 - D(G(z))$ 最小化:当 $D(G(z)) = 1$ 时,或鉴别器将生成器的输出错误分类为实际样本时,会发生这种情况。因此,鉴别器的任务是学习正确地将输入分类为真实或假的,而生成器试图欺骗鉴别器以认为其生成的输出是真实的,二者形成对抗关系。对抗能更好地学习,而对抗学习的关键就是如何表示和优化对抗性损失。

1.2 对抗性损失

对于未标记的目标域,策略是通过最小化源和目标特征分布之间的差异来指导特征学习[10,12,13]。为此目的,有几种方法使用最大平均差异损失,计算 2 个域均值之间差异的范数。除了源上的常规分类损失之外,深度域混淆(deep domain confusion, DDC)[14]方法使用 MMD 来学习既具有鉴别性又具有域不变性的表示。相比之下,相关对齐(correlation alignment, CORAL)[15]方法提出匹配 2 个分布的均值和协方差。

域自适应的目标是从源数据分布中学习在不同但相关的目标数据分布上的良好性能模型。而生成对抗网络的思想是通过对抗训练生成与真实图像逼真的图像。对抗性学习方法是训练健壮的深度网络的有前景的方法,并且可以跨不同领域生成复杂样本。

本文的对抗性损失定义为固定 G 的参数不变,优化 D 的参数,即 $\max V(D, G)$,等价于 $\min[-V(D,G)]$。因此 D 的损失函数等价为

$$J^{(D)}(\theta^D,\theta^G) = -E_{x\sim p_{data}(x)}[\log D(x)] - E_{\tilde{x}\sim p_g}[\log(1 - D(\tilde{x}))] \tag{2}$$

鉴别器认为来自真实数据样本的标签为 1 而来自生成样本的标签为 0。因此,其优化过程是类似于 Sigmoid 的二分类,即 Sigmoid 的交叉熵。

在固定鉴别器参数不变的情况下,生成器的代价函数可表述为

$$J^{(G)} = \max_D V(G, D) = E_{x\sim p_{data}(x)}[\log D_G^*(x)]$$

$$+ E_{z\sim p_z(z)}[\log(1 - D_G^*(G(z)))]$$
$$= E_{x\sim p_{data}(x)}[\log D_G^*(x)] + E_{\tilde{x}\sim p_g}[\log(1 - D_G^*(x))]$$
$$= E_{x\sim p_{data}(x)}[\log\frac{p_{data}(x)}{p_{data}(x) + p_g(x)}] + E_{x\sim p_g}[\log\frac{p_g(x)}{p_{data}(x) + p_g(x)}] \quad (3)$$

当 $p_g = p_{data}$ 时，生成器的损失为

$$J^{(G|D^*)} = E_{x\sim p_{data}(x)}[\log\frac{1}{2}] + E_{x\sim p_g}[\log\frac{1}{2}] = \log\frac{1}{2} + \log\frac{1}{2} = -\log 4 \quad (4)$$

引入 JS 散度(Jensen-Shannon divergence)，生成器的代价函数等价为

$$J^{(G)} = E_{x\sim p_{data}(x)}[\log\frac{2p_{data}(x)}{p_{data}(x) + p_g(x)}] + E_{x\sim p_g(x)}[\log\frac{2p_g(x)}{p_{data}(x) + p_g(x)}] - \log(4)$$
$$= KL(p_{data} \| \frac{p_{data} + p_g}{2}) + KL(p_g \| \frac{p_{data} + p_g}{2}) - \log(4)$$
$$= -\log(4) + 2 \times JSD(p_{data} \| p_g) \quad (5)$$

由于 JS 散度具有非负性，当两者分布相等时，其散度为 0。因此，$D(x)$ 训练得越好，$G(z)$ 就越接近最优，则生成器的损失越接近于生成样本分布和真实样本分布的 JS 散度。

用交替迭代的方法优化参数，其优化流程如下。

初始化：采用批随机梯度下降进行训练，超参数 $k = 1$；批大小 Batchsize $= m$；

for number of training iterations **do**

　for k steps **do**

　　抽样出 m 个噪声 $p_z(z)$ 样本 $\{z^{(1)}, z^{(2)}, z^{(3)} \cdots z^{(m)}\}$

　　抽样出 m 个数据 $p_x(x)$ 样本 $\{x^{(1)}, x^{(2)}, x^{(3)} \cdots x^{(m)}\}$

　　计算鉴别器的代价函数：

$$J^{(D)} = \frac{1}{m}\sum_{i=1}^{m}[-\log D(x^{(i)}) - \log(1 - D(G(z^{(i)})))]$$

　　通过 Adam 梯度下降算法更新鉴别器参数：

$$\theta_d = Adam(\nabla\theta_d(J^{(D)}), \theta_d)$$

　end for

　抽样出 m 个噪声 $p_z(z)$ 的样本 $\{z^{(1)}, z^{(2)}, z^{(3)} \cdots z^{(m)}\}$

　计算生成器的代价函数：

$$J^{(G)} = \frac{1}{m}\sum_{i=1}^{m}[\log(1 - D(G(z^{(i)})))]$$

　通过 Adam 梯度下降算法更新生成器的参数：

$$\theta_g = Adam(\nabla\theta_g(J^{(G)}), \theta_g)$$

end for

2　对抗鉴别的无监督域自适应方法

2.1　对抗性无监督域自适应

基于鉴别模型和对抗损失的无监督适应方法的一般框架如图 2 所示。在无监督领域自适应中，假设源图像 X_s，从源域分布 $p_s(x, y)$ 绘制的标签 Y_s，以及服从目标分布 $p_t(x, y)$ 的目标图像 X_t，没有标签。目的是学习目标表示即目标特征映射 F_t 和分类器 C_t，它可以在测试时将目标图像正确地分类为

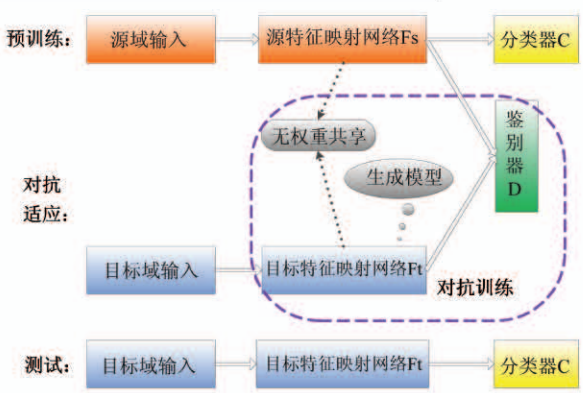


图 2　本文方法的框架

N 类别中的一个。由于目标域无标签,不能对目标进行直接监督学习,先域自适应学习源特征映射 F_s 以及源分类器 C_s,然后再学习使该模型适应于目标域。

最小化源域映射后的特征空间 $F_s(X_s)$ 和目标域映射后的特征空间 $F_t(X_t)$ 之间的距离。由于源域有标签,可以学习源域的特征映射 F_s 和源域的分类器 C_s 来分类:

$$\min_{F_s,C} L_{cs}(X_s,Y_s) = E_{(x_s,y_s)\sim(X_s,Y_s)} - \sum_{n=1}^{N} 1_{[n=y_s]} \log C(F_s(x_s)) \tag{6}$$

把 F_s 和 C_s 迁移到目标域。为使实验结果更为显著,将源域分类器 C_s 直接作为目标分类器 C_t,即设置 $C = C_s = C_t$。因此,只需要学习 F_t,为了获得 F_t,需要优化分类器 D,借鉴第 1 节 GAN 网络的思想,优化 D 的目标函数即域分类器损失为

$$L_{ad}^{D}(X_s,F_s,X_t,F_t) = -E_{x_s\sim X_s}[\log D(F_s(x_s))] - E_{x_t\sim X_t}[\log(1-D(F_t(x_t)))] \tag{7}$$

其中,$F_t(x_t)$ 可以视为输入的噪声,希望分类器 D 能尽量分类出得到的特征是来自于哪一个域,因此,$D(F_t(x_t))$ 理想的结果是 0,$D(F_s(x_s))$ 理想结果是 1,最小化 L_{ad}^{D} 可以得到当前的域分类器 D。希望 D 尽可能区分不出二者,训练 X_t,即进行对抗性优化,先固定 F_s、F_t 优化 D,再固定 D 优化 F_s、F_t。

在有监督的情况学习源域特征映射 F_s,使用源域的标签 x_l 通过潜在的空间鉴别损失进行监督训练得到最终的源识别最佳表示。因为目标域未标记,需要对这些特征映射的特定参数化。当源域的特征映射参数 F_s 被确定,就要确定目标域的特征映射参数 F_t,目标特征映射几乎总是在特定功能层方面与源匹配,只使用源初始化目标映射参数,而不共享二者权重,需要源映射和目标映射之间约束 $\psi(F_s,F_t)$。在最小化 $F_s(X_s)$ 和 $F_t(X_t)$ 之间的距离的同时,也要保证 F_t 的识别性能,将这个约束定义为逐层的约束,即在分层表示中,每个层参数表示为 F_s^l 或 F_t^l,对于一组等效层 $\{l_1, l_2, \cdots, l_n\}$,约束表示为

$$\psi(F_s, F_t) = \{\psi_{li}(F_s^{li}, F_t^{li})\}_{i\in\{1,\cdots,n\}} \tag{8}$$

并且用它最普遍的约束,即源域的分层和目标域的分层完全一致:

$$\psi_{li}(F_s^{li}, F_t^{li}) = \psi_{li}(F_s^{li}) = \psi_{li}(F_t^{li}) \tag{9}$$

确定了 F_t 的参数后,利用对抗性损失来学习实际的特征映射。在训练 GAN 时,分别优化 2 个独立的目标,即生成器和鉴别器。鉴别器的对抗损失 L_{ad}^{D} 不变,特征映射的损失变为

$$L_{ad}^{F}(X_s, X_t, D) = E_{x_t\sim X_t}[\log(1-D(F_t(x_t)))] \tag{10}$$

这个目标函数与极大极小损失有相同的定点属性,但其针对目标特征映射 $F_t(x_t)$ 拥有更强的梯度。这种方式是将源特征映射 F_s 和目标特征映射 F_t 独立开来,并且仅仅去学习目标特征映射 F_t,因为源特征映射 F_s 可以通过直接训练得到。这模拟了 GAN,其中真实图像的分布保持固定,生成器 G 生成的分布来匹配真实图像的分布。

在生成器试图拟合 1 个不变的分布的时候,对抗损失是一个标准的选择方案。但是,在 2 个分布都发生变化的情况下,当 F_t 收敛到最优的时候此目标将会震荡,鉴别器的变化会导致预测的符号发生反转 。为确保 F_s 和 F_t 之间的独立性并且避免震荡的出现,采用使用交叉熵损失函数对统一分布训练特征映射:

$$L_{ce}^{F}(X_s,X_t,D) = -\sum_{m\in\{s,t\}} E_{x_m\sim X_m}[\frac{1}{2}\log D(F_m(x_m)) + \frac{1}{2}\log(1-D(F_m(x_m)))] \tag{11}$$

2.2 辅助分类任务

在域自适应应用场景中,源域样本中往往包含有目标域中不存在的类别样本。为了能够充分利用到源域样本,本文引入辅助分类任务,其思想源自多任务学习。结合辅助的任务学习共同的特征表示,这样最大限度地丰富训练样本,增强学习到特征的泛化性能,而且有效增大类间距离和减小类内距离,有利于提高分类精度。

辅助损失函数定义为

$$L_{sup}(x_{sup}, y_{sup}, \theta_f, \theta_{sup}) = -\sum_{k} 1[y_{sup} = k]\ln p_k \tag{12}$$

式中，x_{sup}、y_{sup}分别为辅助样本数据和标签，θ_{sup}为辅助分类线性分类器的网络参数，分类器最后输出的单元数为辅助分类的类别数；1 为指示数，即当 $y_{sup}=k$ 成立时取值 1，否则为 0；p_k 为 softmax 层的输出，即 $p_k=\text{softmax}[\theta_{sup}^{\mathrm{T}}f(x_{sup};\theta_f)]$。辅助损失函数是类别的损失，所以要求它的损失越小越好。

2.3　算法流程

本文方法的参数更新流程如表 1 所示。

图 3　数字数据集适应示例

表 1　算法流程

输入：源域数据 $S=\{(x_s^i,y_s^i)\}_{i=1}^{ns}$，目标域数据 $T=\{(x_t^i)\}_{i=1}^{nt}$，初始化参数 θ_f、θ_{sup}、θ_{main}、θ_d，随机初始化系数 α 和对抗更新系数 β，迭代次数 $i=0$，初始学习率 $\varphi(0)$，最大迭代次数 max_step，批大小 Batchsize $=m$；

While $i<$ max_step：

输入样本：$\{(x^i,y^i)\}_{i=1}^{m/4}$，$\{(x_{sup}^i,y_{sup}^i)\}_{i=1}^{m/4}$，$\{(x_t^i)\}_{i=1}^{m/2}$

$$f_{ad}=conv(x^i,\theta_f);\ f_{sup}=conv(x_{sup}^i,\theta_f);\ f_t=conv(x_t^i,\theta_f)$$

$$\nabla\theta_{ad}=\frac{\partial L_{ad}(f_{ad},\theta_{ad})}{\partial\theta_{ad}};\ \nabla\theta_{sup}=\frac{\partial L_{sup}(f_{sup},\theta_{sup})}{\partial\theta_{sup}};\ \nabla\theta_d=\frac{\partial L_d(f_{ad},\theta_d)}{\partial\theta_d}+\frac{\partial L_d(f_t,\theta_d)}{\partial\theta_d}$$

$$\nabla f_{ad}=\frac{\partial L_{ad}(f_{ad},\theta_{ad})}{\partial f_{ad}}+\beta\cdot\frac{\partial L_d(f_{ad},\theta_d)}{\partial f_{ad}};\ \nabla f_{sup}=\alpha\cdot\frac{\partial L_{sup}(f_{sup},\theta_{sup})}{\partial f_{sup}};$$

$$\nabla f_t=\beta\cdot\frac{\partial L_d(f_t,\theta_d)}{\partial f_t}$$

$$\nabla\theta_f=\nabla f_{ad}\cdot\frac{\partial conv(x^i,\theta_f)}{\partial\theta_f}+\nabla f_{sup}\cdot\frac{\partial conv(x_{sup}^i,\theta_f)}{\partial\theta_f}+\nabla f_t\cdot\frac{\partial conv(x_t^i,\theta_f)}{\partial\theta_f}$$

更新参数：

$$\theta_{ad}=\theta_{ad}-\phi(i)\cdot\nabla\theta_{main},\ \theta_{sup}=\theta_{sup}-\phi(i)\cdot\nabla\theta_{sup},\ \theta_d=\theta_d-\phi(i)\cdot\nabla\theta_d,\ \theta_f=\theta_f-\phi(i)\cdot\nabla\theta_f$$

$$i=i+1$$

End while

输出：θ_f，θ_{ad}，并预测标签 $\hat{y}_t^i=C(x_t^i,\theta_f,\theta_{ad})$

3　实　验

3.1　MNIST、MNIST-M、SVHN 数字数据集适应

本研究在 MNIST[16]、MNIST-M[17] 和 SVHN[18] 数字数据集之间的无监督域自适应调整任务中验证了本文方法，这些数据集都由 10 个数字（0～9）类组成，数据集示例见图 3。所有的实验都在无监督的设置中进行，其中目标域中的标签被隐藏，主要考虑在 2 个方向上进行适应，即 MNIST 到 MNIST-M，SVHN 到 MNIST。

（1）从 MNIST 到 MNIST-M。MNIST 数据集的数字图像作为源域，MNIST-M 数据集的数字图像作为目标域。MNIST-M 数据集是针对无监督域自适应提出的 MNIST 的变体。它的图像是通过每个 MNIST 数字为二进制掩码和它的背景图像反相创建的。背景图像是随机从伯克利分割数据集中（BSDS200）[19] 均匀采样。实验遵循文献[17]中建立的训练协议，从 MNIST 采样 2 000 个图像，从 MNIST-M 采样 1 000 个图像。

（2）从 SVHN 到 MNIST。在 2 个不同的域上测试本文方法。SVHN 为街景门牌号数据集，包含着现实世界的复杂因素。对 SVHN 的训练具有挑战性，适应比较困难。在训练的前期，分类错误仍然很高。由于 SVHN 更加多样化，因此预计在 SVHN 上

训练的模型将更加通用并且可以在 MNIST 数据集上合理地执行。

对于上述实验,使用简单修改的 LeNet 架构在 tensorflow[20] 中实现。对抗性鉴别器由 3 个完全连接层组成,前 2 层具有 500 个隐藏单元,第 3 层是最终鉴别器输出。每个 500 单元层使用 ReLU 激活功能。优化使用 Adam 优化器[21] 进行 10 000 次迭代,学习率为 0.002,$\beta1$ 为 0.5,$\beta2$ 为 0.99,批量大小为 256 个图像,即源域与目标域各 128 个。所有训练图像都转换为灰度,并重新缩放为 28 ×28 像素。

实验结果如图 4 和表 2 所示。根据图表可以明显看出,本文方法在“MNIST 到 MNIST-M”数据集上实现了比以前方法更好的结果,而且曲线上升趋势良好,紧追“只有目标域”的表现。此外,与其他方法相比,该方法在具有挑战性的从 SVHN 到 MNIST 适应任务上展现出令人信服的结果,也表明本文方法有可能推广到其他各种设置。

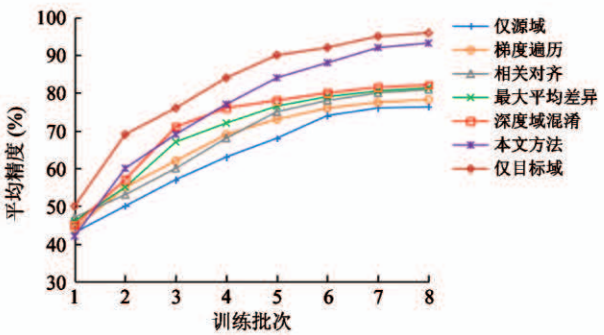


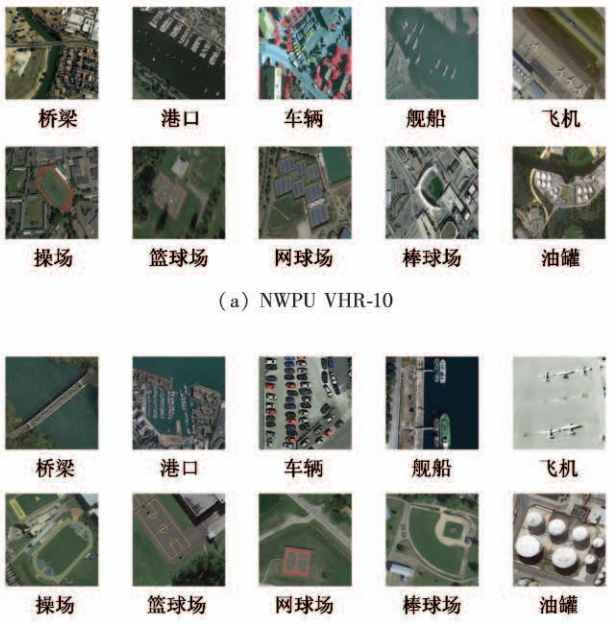
图 4　各方法的精度随训练批次的变化

表 2　数字数据集的分类精度

方法	MNIST→MNIST-M	SVHN→MNIST
仅源域	0.763	0.614
梯度遍历	0.782	0.751
相关对齐	0.809	0.587
最大平均差异	0.813	0.776
深度域混淆	0.821	0.704
本文方法	0.932	0.794
仅目标域	0.959	0.915

3.2　遥感影像数据集适应

为了更好地验证本文方法,将该方法在 2 组遥感影像数据集上适应,示例图像如图 5 所示。



(a) NWPU VHR-10

(b) NWPU-XUAN10

图 5　遥感数据集各类示例

NWPU VHR-10 数据集是公开的 10 个对象类地理空间物体检测数据集,这 10 类物体分别是飞机、舰船、油罐、棒球场、网球场、篮球场、操场、港口、桥梁和车辆。该数据集包含 800 个非常高分辨率(VHR)的遥感影像。对图像进行人工切割尺寸为 256 ×256,并人工分类标注。

NWPU-RESISC45 数据集含有 45 类场景的遥感影像,每类影像都包含有 700 张图片,尺寸均为 256 ×256。选出与 NWPU VHR-10 重叠的 10 个类每类随机选用 100 张,共 1 000 张影像,命名为 NWPU-XUAN10。

该实验网络的各个参数,如卷积核大小、步长和卷积层的层数如图6所示。特征训练层使用了预训

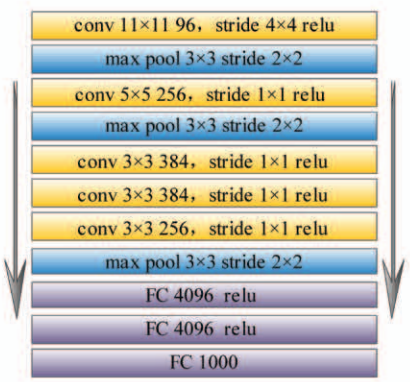


图 6　本文方法的网络结构

练的 Alexnet 网络架构，对抗性鉴别器由 3 个完全连接层组成，前 2 层具有 4 096 个隐藏单元，第 3 层是对抗性鉴别器输出。除输出外，这些层使用 ReLU 激活功能。然后，使用与数字实验中相同的超参数训练，再进行 10 000 次迭代。

从 NWPU VHR-10 到 NWPU-XUAN10 的分类精度与批次关系以及最终结果如图 7 和表 3 所示。同时进行“仅源域”和本文方法监督目标模型的混淆矩阵到深度适应实验，并将 NWPU VHR-10 数据集的混淆矩阵列于图 8。

从表 3 可以看出，本文方法在精度上实现了更好的结果，优于其他方法。在图 7 中，本文方法逐渐赶超最优的域分离网络方法，并且还有上升的趋势。图 8 中，本文方法表现均衡，对于容易混淆的篮球场、操场和网球场这 3 类场景的辨识度也有了一定的提高。由此表明在域自适应中对抗网络和辅助任务可以很好地学习到域不变特征，并提高网络的泛化能力与分类精度。

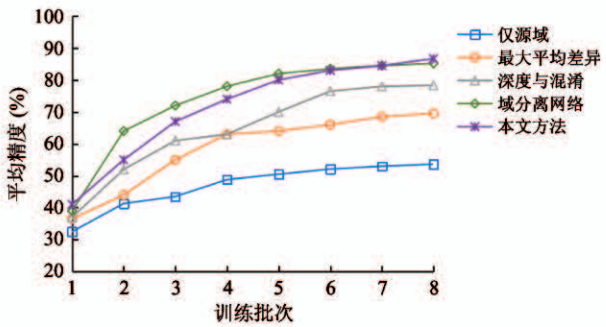
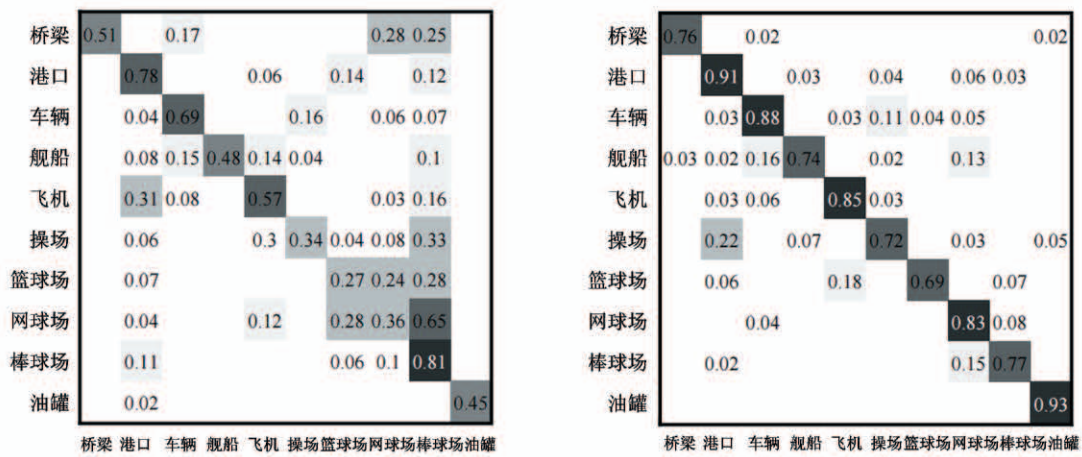


图 7　各方法的精度随训练批次的变化

表 3　遥感数据集的分类精度

方法	平均分类精度
仅源域	0.536
最大平均差异	0.695
深度域混淆	0.783
域分离网络	0.852
本文方法	0.867



(a) 仅源域　(b) 本文方法

图 8　NWPU VHR-10 数据集混淆矩阵

4　结　论

本文提出了一种基于鉴别模型和对抗学习目标的无监督域自适应方法，域自适应网络结合鉴别模型，无需权重共享、对抗性损失和辅助分类任务，并建立了基于深度卷积神经网络的分类框架，使源特征映射网络与目标特征映射网络形成对抗的关系，引入辅助分类任务，扩充训练样本。这种对抗鉴别的无监督域适应方法在数字数据集上实现了比以前方法更佳的结果，并在具有挑战性的从 SVHN 到 MNIST 适应任务上展现出良好的结果，也表明本文方法有可能推广到其他各种设置。最后在遥感数据集上的实验表明，对抗网络和辅助任务可以很好地学习到域不变特征，并提高网络的泛化能力与分类精度。

参考文献

[1] Donahue J, Jia Y Q, Vinyals O, et al. DeCAF: a deep convolutional activation feature for generic visual recognition[C]//International Conference on Machine Learning, Beijing, China, 2014: 647-655

[2] Gretton A, Smola A, Huang J Y, et al. Covariate Shift by Kernel Mean Matching[M]. In: Dataset Shift in Machine Learning, Cambridge: MIT press, 2009: 131-160

[3] Torralba A, Efros A A. Unbiased look at dataset bias[C]//2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, USA, 2011: 1521-1528

[4] Long M S, Cao Y, Wang J M, et al. Learning transferable features with deep adaptation networks[C]//International Conference on Machine Learning (ICML), Lille, France, 2015: 97-105

[5] Sun B C, Feng J S, Saenko K. Return of frustratingly easy domain adaptation[J]. *arXiv*:1511.05547, 2015

[6] Ghifary M, Kleijn W B, Zhang M J, et al. Deep reconstruction-classification networks for unsupervised domain adaptation[C]//European Conference on Computer Vision, Amsterdam, The Netherlands, 2016: 597-613

[7] 丁亮, 何彦青. 融合领域知识与深度学习的机器翻译领域自适应研究[J]. 情报科学, 2017, 35(10): 125-132

[8] 曾远柔, 王红霞. 机器视觉中基于界标的无人管理域自适应算法研究[J]. 高技术通讯, 2018, 28(7): 614-619

[9] Ganin Y, Lempitsky V. Unsupervised domain adaptation by backpropagation[C]//Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 2015: 1180-1189

[10] Tzeng E, Hoffman J, Darrell T, et al. Simultaneous deep transfer across domains and tasks[C]//2015 IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 4068-4076

[11] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]//Advances in Neural Information Processing Systems 27, Montreal, Canada, 2014: 2672-2680

[12] Liu M Y, Tuzel O. Coupled generative adversarial networks[C]//Advances in Neural Information Processing Systems 29, Barcelona, Spain, 2016: 469-477

[13] Sener O, Song H, Saxena A, et al. Learning transferrable representations for unsupervised domain adaptation[C]//Advances in Neural Information Processing Systems 29, Barcelona, Spain, 2016: 2110-2118

[14] Tzeng E, Hoffman J, Zhang N, et al. Deep domain confusion: maximizing for domain invariance[EB/OL]. https://arxiv.org/abs/1412.3474: Cornell University, 2014

[15] Sun B C, Saenko K. Deep CORAL: correlation alignment for deep domain adaptation[EB/OL]. https://arxiv.org/abs/1607.01719: Cornell University, 2016

[16] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. *Proceedings of the Institute of Electrical and Electronics Engineers*, 1998, 86(11): 2278-2324

[17] Ganin Y, Ustinova E, Ajakan H, et al. Domain-adversarial training of neural networks[J]. *Journal of Machine Learning Research*, 2016, 17(59): 1-35

[18] Netzer Y, Wang T, Coates A, et al. Reading digits in natural images with unsupervised feature learning[C]//NIPS Workshop on Deep Learning and Unsupervised Feature Learning, Granada, Spain, 2011: 12-17

[19] Arbelaez P, Maire M, Fowlkes C, et al. Contour detection and hierarchical image segmentation[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011, 33(5): 898-916

[20] Abadi M, Agarwal A, Barham P, et al. Tensorflow: large-scale machine learning on heterogeneous distributed systems[EB/OL]. https://arxiv.org/abs/1603.04467v1: Cornell University, 2016

[21] Kingma D, Ba J. Adam: a method for stochastic optimization[EB/OL]. https://arxiv.org/abs/1412.6980: Cornell University, 2014

Unsupervised domain adaptation method based on discriminative model and adversarial loss

Zhao Wencang, Yuan Lizhen, Xu Changkai

(College of Automation and Electronic Engineering, Qingdao University of Science & Technology, Qingdao 266061)

Abstract

Collecting well-annotated image datasets to train deep learning algorithms is prohibitively expensive and time consuming for many tasks, and models trained purely on rendered images often fail to be generalized to real images. Regarding the issues above, the unsupervised domain adaptive algorithm attempts to map some features that represent or extract domain invariance between two domains, mapping the two domains to a common feature space. Considering the labeled data of the source domain and the unlabeled data of the target domain, an unsupervised domain adaptation method based on generative adversarial network (GAN) architecture is proposed, which uses the discriminant model, without weight sharing, adversarial loss and auxiliary classification tasks, and learns the transformation from one domain to another in an unsupervised manner. The unsupervised domain adaptive method for adversarial discrimination can effectively reduce the difference between the training and test domain distributions, mitigating the harmful effects of domain shifts, and significantly improve the recognition rate. The experimental results prove that the adversarial discriminant method is more effective and simpler than other domain adaptive methods, expanding the sample and improving the generalization performance of the network.

Key words: deep learning, unsupervised, domain adaptation, generative adversarial network (GAN), auxiliary classification task