复现文档

Adversarial Feature Augmentation for Unsupervised Domain Adaptation

论文主要内容

论文题目为，无监督领域适应的对抗性特征增强。无监督指利用源域中已有的知识（标签信息）去学习目标域的样本的类别，对抗学习指特征增强。主要内容为以下几点

1. 使用 GAN对数据的特征空间进行增强，而不是对数据本身进行增强，为训练域不变特征提取器提供训练数据。
2. 无监督领域自适应(Unsupervised Domain Adaptation)，训练一个域不变(domain-invariant) 的特征提取器，其可以在对未给出label的目标数据集工作，为对目标域分类提供基础。
3. 训练结果通过对未给出label的目标数据集分类的准确性体现。

总结：使用极大极小值博弈（DI）强制模型的域不变特性，并且使用特征增强为其提供更多数据，增强模型。

数据集

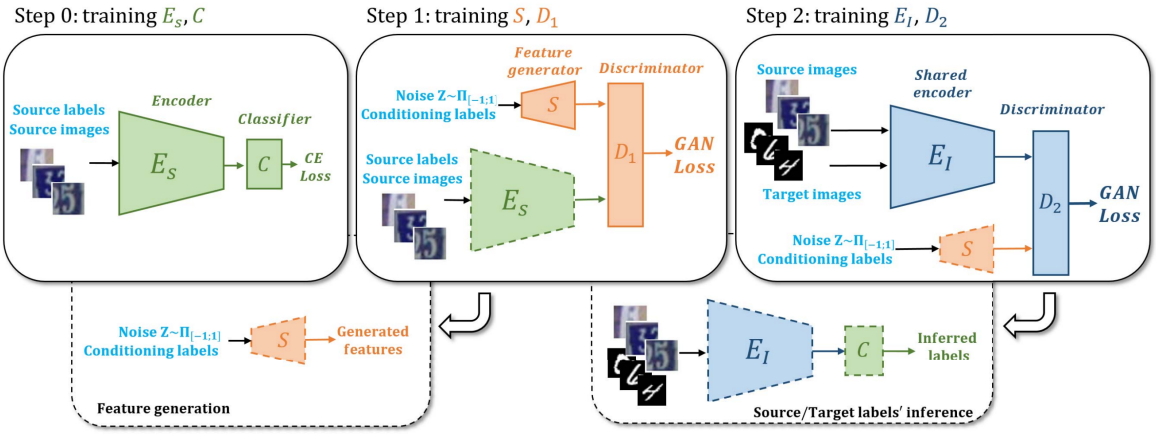
mnist 目标数据

svhn 源数据

模型衡量

模型对于未给出lable的目标数据的分类准确率

模型结构



步骤总览

- Step 0
  - 利用有标签的源域数据，训练特征提取器E_s 和一个分类器C
  - E_s为一个参考的特征空间
  - C为一个表现良好的参考分类器，用于Step2
- Step 1
  - 使用E_s提取出的特征，通过GAN(采用CGAN框架)，训练S和D_1
  - S的作用为特征增强，为Step 2提供特征数据
  - D_1的作用为区分特征由S产生还是由源域数据通过E_s提取而来
- Step 2
  - 使用S生成的特征和源域以及目标域的图片，通过GAN，训练E_I和D_2
  - E_I为一个具有域不变特性的鲁棒性更好的特征提取器，其将源域和目标域图片中提取的特征和特征生成器生成的特征对齐，并使用S进一步增强效果
  - 最后将与E_I和Step 0 中的分类器C结合，达到给目标域分类的目的

模型训练具体步骤

Step 0

E_s表示一个ConvNet(卷积神经网络) 特征提取器，C 表示一个完全连接的 softmax 层。优化目标如下

$$\min_{\theta_{E_s},\theta_C} \ell_0 = \mathbb{E}_{(x_i,y_i)\sim(X_s,Y_s)} H(C \circ E_s(x_i), y_i)$$

其中θ_{E_s}和θ_C分别表示E_s和C的参数，X_s，Y分别是源样本（x_i）和源标签（y_i）的分布，H表示softmax交叉熵函数。

Step 1

经过训练的S可以生成与源特征相似的特征样本。 利用 CGAN 框架，定义了以下极大极小博弈：

$$\min_{\theta_S}\max_{\theta_{D_1}} \ell_1 = \mathbb{E}_{(z,y_i)\sim(p_z(z),Y_s)} \left\| D_1(S(z||y_i)||y_i) - 1 \right\|^2 + \mathbb{E}_{(x_i,y_i)\sim(X_s,Y_s)} \left\| |D_1(E_s(x_i)||y_i) \right\|^2$$

其中θ_{E_s} 和θ_{D_1} 分别表示S和D_1的参数，$p_z(z)$是抽取噪声样本的分布1，$||$表示级联操作。

特征增强只需要使用S，其将噪声向量和one-hot标签代码的串联作为输入，特征向量作为输出，见模型结构图中左边的虚线框

$$F(z|y) = S(z||y)$$

其中$z \sim p_z(z)$和F是属于与y相关的类标签的特征向量

Step 2

域不变的编码器E_I通过极大极小博弈训练，在用Step 0 训练得到的权重初始化后 (E_I和E_s具有相同的架构) 开始训练。

$$\min_{\theta_{E_I}}\max_{\theta_{D_2}} \ell_2 = \mathbb{E}_{x_i\sim X_s\cup X_t} \left\| D_2(E_I(x_i)) - 1 \right\|^2 + \mathbb{E}_{(z,y_i)\sim(p_z(z),Y_s)} \left\| D_2(S(z||y_i)) \right\|^2$$

其中θ_{E_I} 和θ_{D_2} 分别表示E_I和D_2的参数。 由于模型 E_I 是使用源域和目标域进行训练的，因此特征提取器域不变。它将源样本和目标样本映射到公共特征空间中，其中的特征与通过 S 生成的特征无法区分。 由于后者经过训练以产生与源特征无法区分的特征，特征提取器 E_I 可以与Step 0 中的 C 分类器结合并用于分类，达到分类目标域的目的。

$$\tilde{y}_i = C \circ E_I(x_i)$$

其中 x_i 是来自源或目标数据分布的通用图像，而$\tilde{y}_i$ 是预测的label，见模型结构图中右边的虚线框

取得成果

1. **特征增强取得的优势**

蓝色曲线为使用了特征增强后对于目标域的分类准确率，橙色曲线为未使用特征增强的分类准确率，绿色曲线未采用 LS-ADDA 的分类准确率。明显蓝色准确率最高。

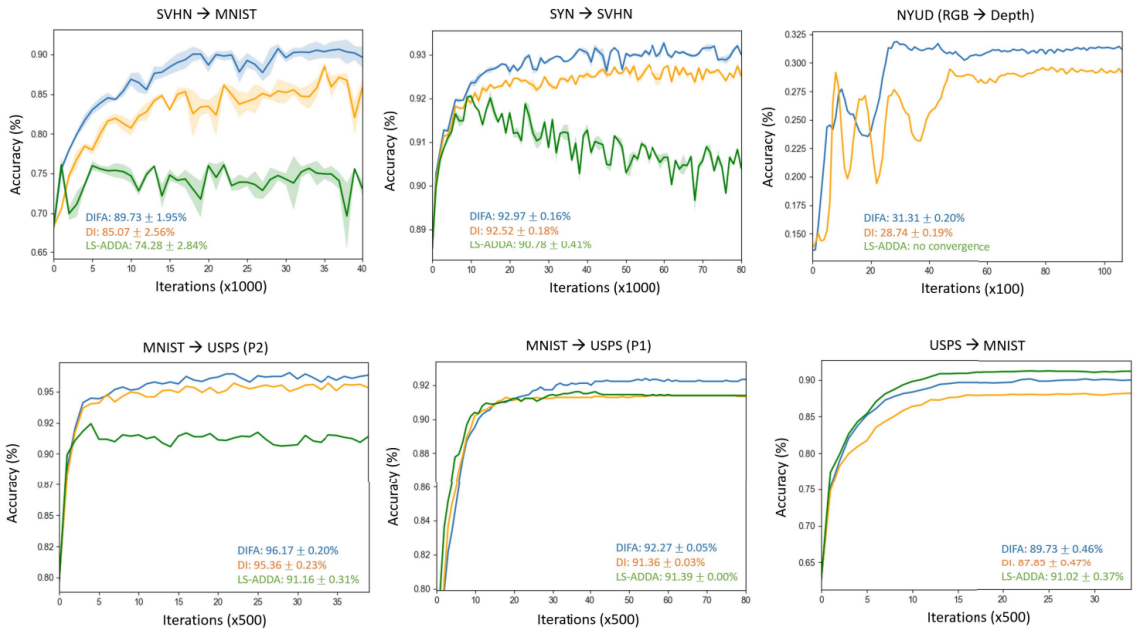


Figure 3. Accuracies on target samples evaluated throughout the training of the feature extractors of *LS-ADDA* (*green*), *DI* (*orange*) and DIFA (*blue*). Inference was performed by combining the feature extractor being learned with C of Step 0, Section 3.1. In the NYUD experiment the green curve is missing due to non-convergence of *LS-ADDA*. SVHN → MNIST and SYN → SVHN plots were obtained averaging over three different runs; confidence bands are portrayed.

2. **与其他模型比较的分类准确率**

Ours(DI)未使用特征增强，Ours(DIFA)使用特征增强，两者对比其他模型基本都具有优势。

Table 2. Comparison of our method with competing algorithms. The row *LS-ADDA* lists results obtained by our implementation of Least Squares ADDA. The row *Ours (DI)* refers to our approach in which only domain-invariance is imposed. The row *Ours (DIFA)* refers to our full proposed method, which includes feature augmentation. (*) DTN [32] and UNIT [17] use extra SVHN data (531, 131 images). (**) Protocols P1 and P2 are mixed in the results section of Bousmalis et al. [1]. Convergence not reached is indicated as *no conv.*

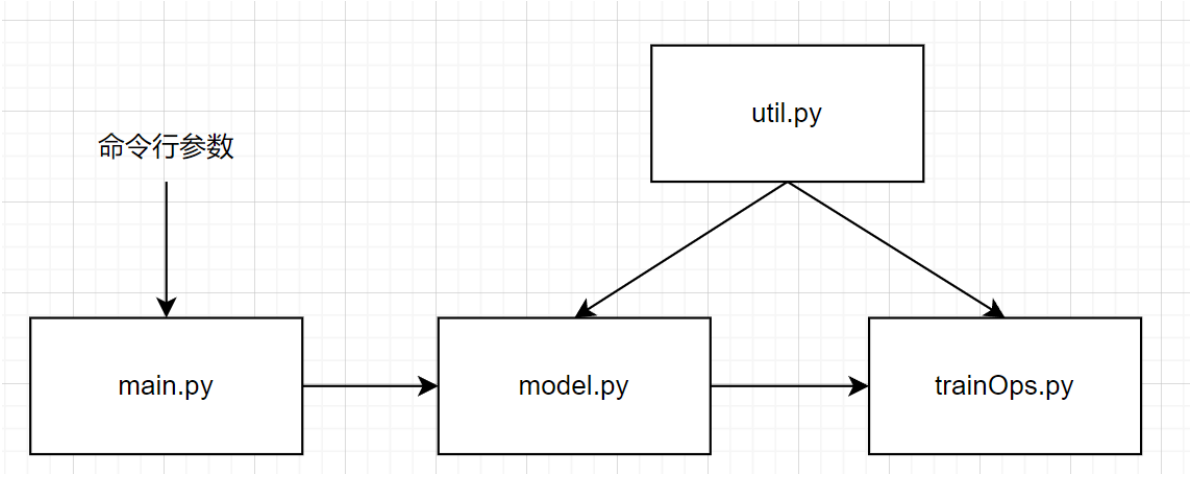
	SVHN→MNIST	MNIST→USPS$_{P1}$	MNIST→USPS$_{P2}$	USPS→MNIST	SYN→SVHN	NYUD
Source	0.682	0.723	0.797	0.627	0.885	0.139
DANN [8, 9]	0.739	0.771 ± 0.018 [34]	-	0.730 ± 0.020 [34]	0.911	-
DDC [34]	0.681 ± 0.003	0.791 ± 0.005	-	0.665 ± 0.033	-	-
DSN [2]	0.827	-	-	-	0.912	-
ADDA [34]	0.760 ± 0.018	0.894 ± 0.002	-	**0.901 ± 0.008**	-	0.211
Tri [28]	0.862	-	-	-	**0.931**	-
DTN [32]	0.844*	-	-	-	-	-
PixelDA** [1]	-	-	0.959	-	-	-
UNIT [17]	**0.905***	-	0.960	-	-	-
CoGANs [18]	no conv. [34]	0.912 ± 0.008	0.957 [17]	0.891 ± 0.008	-	-
LS-ADDA	0.743 ± 0.028	0.914 ± 0.000	0.912 ± 0.003	**0.910 ± 0.004**	0.908 ± 0.004	no conv.
Ours (DI)	0.851 ± 0.026	0.914 ± 0.000	0.954 ± 0.002	0.879 ± 0.005	0.925 ± 0.002	0.287 ± 0.002
Ours (DIFA)	**0.897 ± 0.020**	**0.923 ± 0.001**	0.962 ± 0.002	0.897 ± 0.005	**0.930 ± 0.002**	**0.313 ± 0.002**
Target	0.992	0.999	0.999	0.975	0.913	0.468 [34]

论文的创新部分

1. 首次使用GAN进行特征增强，其他模型大多对原图像的增强。
2. 开创了一种新的增强思路，即增强特征，不增强目标分布图像。在增强特征取得的最终结果比依赖生成目标图像来处理无监督域适应任务的方法表现得更好
3. 将无监督领域自适应与GAN所增强的数据相结合，训练出一个域不变的特征提取器。

代码结构

文件名	作用
mian.py	处理命令行运行时的参数
model.py	构建模型
trainOps.py	训练模型
utils.py	工具



具体代码见github

运行见视频演示