

用于失配隐写分析的对抗子领域自适应网络

章蕾, 王宏霞

(四川大学网络空间安全学院, 四川 成都 610065)

摘要: 当训练集数据和测试集数据来自不同的载体源时, 即在载体源失配的条件下, 通常会使一个表现优异的隐写分析器检测准确率下降。在实际应用中, 隐写分析人员往往需要处理从互联网上采集的图像。然而, 与训练集数据相比, 这些可疑图像很可能具有完全不同的捕获和处理历史, 导致隐写分析模型可能出现不同程度的检测性能下降, 这也是隐写分析工具在现实应用中很难成功部署的原因。为了提高基于深度学习的隐写分析方法的实际应用价值, 对测试样本信息加以利用, 使用领域自适应方法来解决载体源失配问题, 将训练集数据作为源领域, 将测试集数据作为目标领域, 通过最小化源领域与目标领域之间的特征分布差异来提高隐写分析器在目标领域的检测性能, 提出了一种对抗子领域自适应网络(ASAN, adversarial subdomain adaptation network)。一方面从生成特征的角度出发, 要求隐写分析模型生成的源领域特征和目标领域特征尽可能相似, 使判别器分辨不出特征来自哪一个领域; 另一方面从减小域间特征分布差异的角度出发, 采用子领域自适应方法来减少相关子领域分布的非期望变化, 有效地扩大了载体与载密样本之间的距离, 有利于分类精度的提高。通过在多个数据集上对3种隐写算法进行检测, 证实了所提方法可以有效地提升模型在数据集失配和算法失配时的检测准确率, 减少了失配问题带给模型的负面影响。

关键词: 图像隐写分析; 载体源失配; 对抗学习; 领域自适应

中图分类号: TP309

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2022028

Adversarial subdomain adaptation network for mismatched steganalysis

ZHANG Lei, WANG Hongxia

School of Cyber Science and Engineering, Sichuan University, Chengdu 610065, China

Abstract: Once data in the training and test sets come from different cover sources, that is, under the condition of cover source mismatch, it usually makes the detection accuracy rate of an outstanding steganalysis model to be

收稿日期: 2022-02-15; 修回日期: 2022-05-10

通信作者: 王宏霞, hxwang@scu.edu.cn

基金项目: 国家自然科学基金(61972269); 四川省科技计划(2022YFG0320); 河南省网络空间态势感知重点实验室开放课题(HNTS2022003)

Foundation Items: The National Natural Science Foundation of China (61972269), Sichuan Science and Technology Program (2022YFG0320), Open Foundation of Henan Key Laboratory of Cyberspace Situation Awareness (HNTS2022003)

引用格式: 章蕾, 王宏霞. 用于失配隐写分析的对抗子领域自适应网络[J]. 网络与信息安全学报, 2022, 8(3): 76-86.

Citation Format: ZHANG L, WANG H X. Adversarial subdomain adaptation network for mismatched steganalysis[J]. Chinese Journal of Network and Information Security, 2022, 8(3): 76-86.

reduced. In practical applications, the analyzers need to process images collected from the Internet. However, compared with the training set data, these suspicious images are likely to have completely different capture and processing histories, which may lead to the degradation of steganalysis model. It is also why steganalysis tools are difficult to deploy successfully in the real-world applications. To improve the practical application value of steganalysis methods based on deep learning, test sample information is utilized and domain adaptation method is used to solve the problem of cover source mismatch. Regarding the training set data as the source domain and test set data as the target domain, the detection performance of steganalysis models in the target domain is enhanced by minimizing the discrepancy between the feature distribution of source domain and target domain. ASAN (adversarial subdomain adaptation network) was proposed from the perspective of feature generation on the one hand. The source domain features and target domain features generated by the steganalysis model were required to be as similar as possible, so that the discriminator cannot distinguish which domain the features came from. On the other hand, to reduce the difference of feature distribution between domains, the subdomain adaptation method was adopted to reduce the unexpected change of the distribution of related subdomains. The distance between the cover and stego samples was enlarged effectively to improve the classification accuracy. After testing three steganography algorithms on multiple datasets, it is confirmed that the proposed method can effectively improve the detection accuracy rate of the model in the case of dataset mismatch and algorithm mismatch and it can also reduce the negative impact of the mismatch problem of the model.

Keywords: image steganalysis, cover source mismatch, adversarial learning, domain adaptation

0 引言

图像隐写是一种利用数字图像的视觉冗余来嵌入秘密信息的隐蔽通信手段，隐写的安全性是进行隐蔽通信的双方需要考虑的问题，即不能被第三方发现隐写行为的存在。最近几年兴起的基于最小化失真的隐写技术通过设计嵌入失真函数，以此衡量图像每个元素的修改对隐写安全性的影响，并通过最小化总嵌入失真实现隐写的高安全性。这种根据图像的内容自适应地将秘密信息嵌入图像纹理丰富或边缘区域的隐写算法又被称为自适应隐写算法，具有代表性的自适应隐写算法有 S-UNIWARD^[1]、WOW^[2]和 HUGO^[3]。

图像隐写分析的目的是检测隐写信号，以区分载体图像与载密图像。传统的隐写分析方法根据统计矩、图像校准、邻域相关性等来构造隐写分析特征进而进行分类。然而，随着自适应隐写算法的出现，秘密信息被嵌入难以建模的纹理复杂区域，给隐写分析带来了很大的挑战。为了提高隐写分析的准确率，设计算法时需要考虑越来越复杂的图像统计特性，甚至所设计的隐写分析特征需要达到几万维^[4-6]。卷积神经网络在图像隐写分析中表现出色，基于深度学习的隐写分析技

术得到广泛关注^[7-12]。这些基于深度学习的隐写分析器往往能够达到比传统的基于手工设计特征的隐写分析方法更高的检测准确率。

然而，基于深度学习的方法十分依赖训练集的数据和标签，如果训练集的图像内容不够丰富、数量不够大则模型很容易过拟合，且在模型训练时需要将载体图像和对应的载密图像成对地送入模型中，这为训练集的收集增加了难度。为了提升模型的泛化性能，研究者在训练模型的时候往往会通过随机裁剪、旋转等操作进行数据增强^[7-10]来扩充数据集。以往的深度学习隐写分析研究假设训练数据和测试数据服从相同的概率分布。然而测试集的图像与训练集的图像差距较大时，模型在测试集上的表现往往不如在训练集上的表现。训练集和测试集存在差异的因素有很多，如图像来自不同的数据集，经过不同的图像处理过程，不同的隐写算法以及不同的嵌入率等。以上所述就是隐写分析中的载体源失配（CSM, cover source mismatch）问题。现实的应用场景中有各种各样的载体源失配问题，因为测试集的图像与训练集的图像总会有差别，研究者很难获取数据集、隐写算法、嵌入率等重要信息。这也是基于深度学习的隐写分析模型很难在实际应用场景中部署的

原因。为了提升隐写分析模型在载体源失配情况下的检测准确率,本文提出了一种对抗子领域自适应网络(ASAN, adversarial subdomain adaptation network)以减小训练集和测试集之间的差异。

1 相关工作

近年来,已经有一些工作研究如何减轻载体源失配现象对隐写分析器的负面影响,总体上分为两类。一类是子空间迁移学习方法,这类方法试图寻找一个投影将特征投影到一个公共子空间来减小两个领域之间的分布差异。Jia 等^[13]试图找到一个投影矩阵将源领域和目标领域数据投影到一个公共的特征子空间,在重构矩阵上联合低秩约束和稀疏表示来保留局部和全局数据结构。Xue 等^[14]提出了一种基于子空间学习的无监督隐写分析方法,该方法通过对重构系数矩阵的低秩和稀疏约束来保持数据的全局和局部结构以获得新的特征表示,使训练数据和测试数据的特征分布接近。针对 JPEG 重压缩, Yang 等^[15]构造多分类器来检测测试图像的重压缩马尔可夫特性,然后将隐写分析特征转移到一个新的特征子空间。Feng 等^[16]提出了一种基于贡献的特征迁移算法,试图通过评估样本特征和维度特征的贡献来学习两种变换以迁移训练集特征。另一类方法是通过构造领域自适应分类器,直接将领域自适应准则作为正则化项整合到损失函数中。Yang 等^[17]改进了 ARTL (adaptation regularization transfer learning)^[18]算法,在图拉普拉斯正则化中加入了条件分布,将联合分布自适应和几何结构作为正则化项整合到一个标准监督分类器中,以构建领域自适应分类器。在空域中,针对数据集载体源失配问题, Zhang 等^[19]同时考虑了输入图像和预测标签的联合分布。当一个在纹理复杂度低的数据集上训练的模型检测纹理复杂度高的测试集时,隐写分析器的准确率会大幅下降, Zhang 等^[19]提出的 J-Net 可以在一定限度上有效地缓解这种下降。

除上述两类方法, Hu 等^[20]提出了一种解决 CSM 问题的新思路,通过度量图像的纹理复杂度,然后在一幅图像中寻找最接近有效范围的块来代替整幅图像进而提高隐写分析器的准确率。

Zhang 等^[21]通过研究特征距离与特征运动模式的相似度,为每个测试样本选择了专用的训练集训练分类器来解决 CSM。对于嵌入率失配,一个通用的思路^[12,22-23]是先在低嵌入率下训练模型,然后通过渐近地用低嵌入率的数据对模型进行微调,最后得到一个可以检测低嵌入率隐写信号的模型。现在的隐写分析器很多是基于深度学习的,然而以往很多研究^[13-17,20-21]针对的隐写分析特征是基于手工设计的,如 PEV^[24]、Rich Model^[4]、DCTR^[25]、CC-PEV^[26]等,很少对深度隐写分析网络的载体源失配问题进行研究。因此,本文以基于深度学习的隐写分析模型为研究对象,探究缓解载体源失配问题的方法。

隐写分析中的 CSM 问题与迁移学习中的领域自适应问题相似。领域自适应关注于解决特征空间和类别空间一致仅特征分布不一致的问题^[27], CSM 问题也是如此。一些领域自适应方法会设计并最小化一个度量准则来减小域间差异。最大均值差异(MMD, maximum mean discrepancy)^[28]是领域自适应中应用最为广泛的度量,它将两个变量映射到再生核希尔伯特空间(RKHS, reproducing Kernel Hilbert space),然后测量两种分布在 RKHS 中的距离。在 MMD 的基础上,研究者提出了许多领域自适应方法。单核方法,如深度领域混淆^[29];多核方法,如深度自适应网络^[30]、联合自适应网络^[31]、深度子领域自适应网络^[32]等。Zhu 等^[32]提出的非对抗的深度子领域自适应网络通过最小化局部最大均值差异(LMMD, local maximum mean discrepancy)以对齐相关子领域的分布。而另一些领域自适应方法则采用对抗的方式学习一种既能用来进行分类又不能使判别器区分是哪个领域的表示。领域对抗神经网络(DANN, domain-adversarial neural network)^[33]最先在领域自适应中加入了对抗机制,通过与领域判别器进行对抗,得到域间通用的特征。动态对抗自适应网络(DAAN, dynamic adversarial adaption network)^[34]分别为边缘分布和条件分布设计判别器,并动态和定量地评估了这两种分布对抗学习的相对重要性。

这两类领域自适应方法是从两个不同的角度进行考虑的,仅使用对抗学习的领域自适应方法不

足以将域间特征分布的距离最小化，而仅使用最小化度量的方法又缺失了对域无关特征的学习。因此，针对这些局限性，本文提出了用于失配隐写分析的对抗子领域自适应网络。一方面，从减小域间差异的角度出发，对齐相关子领域的特征分布，扩大类间距离，缩小类内距离；另一方面，通过构造一个领域判别器与隐写分析模型进行对抗学习，使模型在两个领域上生成的特征更为相近。

2 对抗子领域自适应网络

2.1 符号表示

图像隐写分析中的载体源失配问题表示如下：源领域用 $D_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ 表示，其中包含 n_s 个有标签样本， y_i 是一个二维的标签向量，它的第一个分量对应载体图像 (cover) 的概率，第二个分量对应载密图像 (stego) 的概率。当 y_i 与源领域上的样本相关时， $y_i = [1, 0]$ 或 $[0, 1]$ 。另外，当 y_i 与目标领域中的样本相关时， y_i 拥有一个软标签，所有分量的总和等于 1。类似地， $D_t = \{x_j^t\}_{j=1}^{n_t}$ 表示目标领域，包含 n_t 个无标签样本。 D_s 和 D_t 服从概率分布 p 和 q ，且 $p \neq q$ 。 y_d 表示领域的标签。本文的主要目的是使模型在 D_s 和 D_t 生成的特征尽可能相似，同时减小不同领域中相关

子领域的特征分布差异，学习可迁移表示。

2.2 网络结构

本文提出的对抗子领域自适应网络结构如图 1 所示。蓝色的数据流代表源领域图像，黄色的数据流代表目标领域图像。同时包含源领域图像和目标领域图像的批数据作为模型输入，预训练隐写分析模型共有 L 层。图像经过模型的第 l ($l \in L$) 层后提取到的特征 f^l 会同时进入分类分支 (绿色部分) 和判别分支 (橙色部分)。在分类分支中，分类器给出对一批源领域图像的预测值 $\hat{Y}_s$ ，通过计算 $\hat{Y}_s$ 与真实标签 Y_s 之间的交叉熵损失得到分类损失。分类器给出对一批目标领域数据的预测值 $\hat{Y}_t$ ，与源领域特征 f^{sl} 、目标领域特征 f^{tl} 、真实标签 Y_s 一起用于计算局部最大均值差异损失。在判别分支中，特征会首先经过一个梯度反转层 (GRL, gradient reversal layer) [33] 以简化对抗训练过程。紧接着，领域判别器会对特征的领域标签进行预测，将预测值 $\hat{Y}_d$ 与真实领域标签值 Y_d 计算交叉熵损失作为对抗损失。通过同时最小化这 3 项损失进行对抗训练，特征提取器提取的特征会更加相似，特征的分布差异也会更小，从而缓解了载体源失配问题对隐写分析器的负面影响。

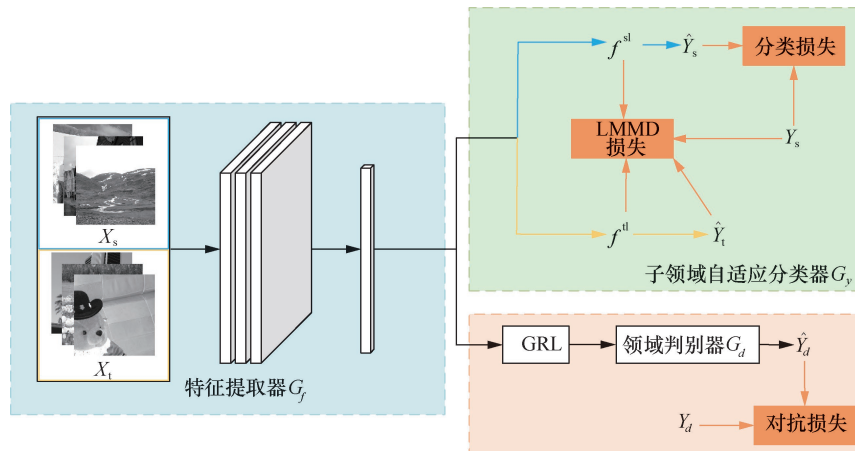


图1 对抗子领域自适应网络结构
Figure 1 Structure of adversarial subdomain adaptation network

2.3 对抗学习

本文提出的对抗子领域自适应网络主要由特征提取器 G_f 、子领域自适应分类器 G_y 和领域判别器 G_d 组成，主要是为了提高预训练隐写分析模

型在失配情况下的检测准确率。将现有隐写分析模型去掉分类层即可作为特征提取器 G_f ， G_f 对输入的图像进行特征提取，提取到的特征是由子领域自适应分类器 G_y 和领域判别器 G_d 共享的。

G_f 学习的目标是:提取后的特征尽可能使分类器对样本进行正确分类,同时使判别器无法区分样本是来自源领域还是目标领域。子领域自适应分类器 G_y 的网络结构与原隐写分析模型的分层相同,根据源领域特征 f^{sl} 和目标领域特征 f^{tl} 分别给出分类结果。在分类的同时进行子领域自适应,通过最小化 LMMD 损失来减小域间相关子领域特征分布的差异。领域判别器 G_d 根据输入的特征给出预测领域的标签,即判断该特征来自哪一个领域。而 G_f 作为与 G_d 对抗的另一方,则希望其生成的特征足够相似以欺骗领域判别器 G_d ,这就使 G_f 寻找源领域和目标领域共有的特征。 G_y 和 G_d 的双重约束,可使 G_f 生成的特征尽可能相似,同时特征分布的差异尽可能地小。

本文的领域分类网络是由梯度反转层和领域判别器组成的。采用梯度反转层可以简化对抗训练过程。梯度反转层实现的原理是:在前向传播过程中保持输入特征 f^l 不变,在反向传播过程中将梯度乘以一个负的定值 μ ,实现梯度自动取反以达到对抗训练的目的。采用两个伪函数来分别表示梯度反转层在正向传播和反向传播中的作用。

前向传播:

$$R_\mu(f^l) = f^l \quad (1)$$

反向传播:

$$\frac{dR_\mu}{dx} = -\mu I \quad (2)$$

其中, I 是一个单位矩阵。

2.4 子领域自适应

主流的领域自适应方法主要学习全局领域偏移,即对齐全局源和目标分布,这种全局对齐的方法会使分类器忽略很多细节,如同一个领域内属于不同类数据之间的关系,不同领域间属于同类数据之间的关系,导致分类器并不能很好地描述两个不同类之间的界限。本文根据标签类别将源领域和目标领域分别划分为两个子领域。源领域使用真实标签进行划分,而目标领域数据使用网络预测的概率分布。本文通过计算局部最大均值差异^[32]来衡量属于同一类的两

个子领域之间的特征分布差异。LMMD 的计算如下。

$$d_H(p, q) \triangleq E_c \left\| E_{p^{(c)}} [\phi(x^s)] - E_{q^{(c)}} [\phi(x^t)] \right\|_H^2 \quad (3)$$

其中, H 是被赋予特征核 k 的再生核希尔伯特空间, $E[\cdot]$ 为数学期望。 $p^{(c)}$ 和 $q^{(c)}$ 分别是源领域和目标领域中属于类别 c 的子领域特征分布。 $\phi(x^s)$ 和 $\phi(x^t)$ 分别表示将源领域和目标领域中的样本映射到再生核希尔伯特空间的特征映射,核 $k(x^s, x^t) = \langle \phi(x^s), \phi(x^t) \rangle$ 是映射 $\phi(x^s)$ 和 $\phi(x^t)$ 的内积。通过最小化局部最大均值差异,同一类别的相关子领域的特征分布更相近。具体地, LMMD 损失的计算方式为

$$L_l(p, q) = \frac{1}{C} \sum_{c=1}^C \left[\sum_{i=1}^{n_s} \sum_{j=1}^{n_t} w_i^{sc} w_j^{sc} k(f_i^{sl}, f_j^{sl}) + \sum_{i=1}^{n_t} \sum_{j=1}^{n_t} w_i^{tc} w_j^{tc} k(f_i^{tl}, f_j^{tl}) - 2 \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} w_i^{sc} w_j^{tc} k(f_i^{sl}, f_j^{tl}) \right] \quad (4)$$

其中, $k(f_i^l, f_j^l)$ 表示源领域特征和目标领域特征的内积。 w_i^c 是样本 x_i 属于类 c 的权重,计算如式(5)所示。

$$w_i^c = \frac{y_{ic}}{\sum_{(x_j, y_j) \in D} y_{jc}} \quad (5)$$

其中, y_{ic} 是标签向量 y_i 的第 c 个值。

2.5 损失函数

在本文的训练过程中一共有 3 个目标:一是实现源领域图像分类误差的最小化;二是减小相关子领域特征分布之间的差异;三是使领域分类误差的最大化,寻找域无关特征。因此可以将总的损失函数写为

$$L(\theta_f, \theta_y, \theta_d) = \frac{1}{n_s} \sum_{i=1}^{n_s} L_y(G_y(G_f(x_i^s)), y_i^s) + \lambda \sum_{l \in L} L_l(p, q) - \frac{\omega}{n_s + n_t} \sum_{i=1}^{n_s + n_t} L_d(G_d(G_f(x_i)), d_i) \quad (6)$$

其中, λ 和 ω 是权衡参数, L_y, L_l, L_d 则分别代表分类损失、LMMD 损失和对抗损失。 $\theta_f, \theta_y, \theta_d$ 分别代表 G_f, G_y 和 G_d 的参数。 y_i^s 代表源领域图像的标签, d_i 则代表图像的领域标签 (源领域的图像 $d_i = 0$, 目标领域的图像 $d_i = 1$)。训练收敛后, 参数 $\hat{\theta}_f, \hat{\theta}_y, \hat{\theta}_d$ 将得到式(6)的一个鞍点。

$$(\hat{\theta}_f, \hat{\theta}_y) = \operatorname{argmin}_{\theta_f, \theta_y} L(\theta_f, \theta_y, \theta_d) \quad (7)$$

$$(\hat{\theta}_d) = \operatorname{argmax}_{\theta_d} L(\theta_f, \theta_y, \theta_d) \quad (8)$$

通过采用 GRL 简化对抗训练过程, 损失函数可以改写为

$$\begin{aligned} \tilde{L}(\theta_f, \theta_y, \theta_d) = & \frac{1}{n_s} \sum_{i=1}^{n_s} L_y(G_y(G_f(x_i^s)), y_i^s) + \\ & \lambda \sum_{l \in L} L_l(p, q) + \frac{\omega}{n_s + n_t} \sum_{i=1}^{n_s + n_t} L_d(G_d(R_\mu(G_f(x_i))), d_i) \end{aligned} \quad (9)$$

在训练完成之后, 可以用标签预测器 $y(x) = G_y(G_f(x))$ 对目标领域的图像进行分类。

3 实验结果及分析

3.1 实验设置

本实验首先使用 3 种隐写算法 S-UNIWARD、WOW、HUGO 在 BOSSbase 1.01 数据集^[35]训练 SRNet 模型^[9]; 然后将训练好的 SRNet 模型作为本文载体源失配的研究对象。因为 SRNet 是一个端到端的深度学习网络, 没有引入过多的手工设计成分, 并且其检测效果在空域中达到前沿水平, 所以选择该模型。本文探究了数据集失配和隐写算法失配这两种载体源失配问题。将本文方法与 SRM^[4]、SRNet^[9]、J-Net^[19]进行比较。J-Net 也用于解决基于深度学习的隐写分析模型面临的载体源失配问题, 其使用联合最大均值差异 (JMMD, joint maximum mean discrepancy)^[31]来最小化数据集之间的差异。本实验的评价指标为平均总精度 P_A , $P_A = \frac{1}{2} \operatorname{ave}(p_c + p_s)$, 文献[17]也采用了平均总精度作为评价指标。其中 p_c 是载体图像的分类准确率, p_s 是载密图像的分类准确率。在失配实验中, 在训练集随机选取 500 对图像 (一幅载

体图像和对应的载密图像为一对) 作为源领域数据, 在测试集上随机选取 500 对图像作为目标领域图像来完成迁移任务。与需要大量标签数据进行有监督训练的隐写分析方法不同, 本文将待测图像加以利用, 对预训练隐写分析模型进行领域自适应。其他失配隐写分析的工作^[13-14, 17, 20]也使用了 500 对图像参与训练。

本文的整体隐写分析过程包括 3 个阶段, 分别是预训练阶段、领域自适应阶段和测试阶段。在预训练阶段, 隐写分析模型在源领域所属数据集上进行有监督的训练。预训练完成后, 将预训练隐写分析模型去掉分类层作为本文的特征提取器 G_f 。通过构造领域判别器来实现对抗子领域自适应, 减小特征分布的差异。在本文实验中, 领域判别器 G_d 由 3 个网络层组成以对特征的领域标签进行预测。SRNet 的特征维数为 512, 隐藏层维数设置为 512, G_d 的网络结构如表 1 所示, 其中 Linear 表示线性层, BN1d 表示一维批标准化层, ReLU 和 Sigmoid 为激活函数。子领域自适应分类器 G_y 的网络结构与原隐写分析模型的分层相同。在测试阶段, 将目标领域图像送入自适应后的网络中, 标签预测器 $y(x) = G_y(G_f(x))$ 从而给出图像的预测结果。

表 1 领域判别器 G_d 的网络结构
Table 1 Network structure of the domain discriminator G_d

网络层	结构	输出尺寸
输入		1×512
第 1 层	Linear+BN1d+ReLU	1×512
第 2 层	Linear+BN1d+ReLU	1×512
第 3 层	Linear+Sigmoid	1

在预训练阶段, 本文采用 256×256 大小的 BOSSbase 1.01 图像进行训练。本文使用 PIL 库中的 `resize` 函数对原 512×512 大小的图像进行下采样调整其大小为 256×256; 接着采用 S-UNIWARD、WOW、HUGO 3 种隐写算法对原始载体图像嵌入秘密信息生成载密图像。数据集的划分为: 5 000 对训练集, 1 000 对验证集和 4 000 对测试集。在预训练过程中没有使用数据增强。本文 0.2 bit/pixel 的 SRNet 模型是通过 0.4 bit/pixel 的 SRNet 模型进行 100 个 epoch 的再训练得到的。对于 SRM, 使

用 BOSSbase1.01 数据集前 5 000 对图像的特征训练分类器。在进行失配实验时,提取目标领域 500 对图像的空域富模型 (SRM, spatial rich model) 特征,利用训练好的分类器对其进行分类。

模型的预训练以及自适应过程均是在 Pytorch 框架下进行的。对于所有任务,本文使用动量为 0.9 的小批量随机梯度下降 (SGD, stochastic gradient descent) 和文献[32]中的学习速率衰减策略。学习率初始化为 0.01, 并按照 $\eta_\varepsilon = \eta_0 / (1 + \varepsilon)^\beta$ 更新, ε 是训练过程中取值范围为 0 到 1 的线性变化, $\eta_0 = 0.01, \alpha = 10, \beta = 0.75$ 。通过这种策略,在经过 200 个 epoch 后,学习率会降为 0.001 7。本文没有固定权衡因子 λ , 而是通过一个渐进的过程逐步将它从 0 改变到 1, 即 $\lambda_\varepsilon = 2 / \exp(-10\varepsilon) - 1$, 这样做是为了抑制在训练初期模型不稳定的输出。对抗损失权重参数 $\omega (\omega = 10)$ 在整个实验中是固定的。训练批大小设为 16。

3.2 数据集失配实验

数据集失配实验从 BOSSbase 1.01 数据集中随机选取 500 幅并嵌入秘密信息得到 500 对大小为 256×256 的图像作为源领域数据, 分别从数据集 UCID^[36]、DIV 2K^[37]、Flickr25k^[38] 上随机选取 500 对图像作为目标领域数据。

表 2 和表 3 分别给出了嵌入率为 0.4 bit/pixel 和 0.2 bit/pixel 时数据集失配的实验结果。当在 BOSSbase1.01 数据集上训练的模型检测来自 UCID.v2、DIV2K 和 Flickr 25K 的数据时, 即数据集失配时, 基于手工特征的传统隐写分析方法 SRM 受数据集失配的影响比 SRNet 更大, 其检测准确率下降更多。SRM 的特征提取过程不是特定于训练集数据的, 但分类器的训练却是特定于训练集的。载体源之间的失配也会导致这种传统隐写分析方法的检测准确率下降。对于 SRNet, 大部分情况下模型的准确率会下降, 准确率上升的特例来自 Flickr 25K 数据集, 这可能是对其进行的图像处理使其纹理复杂度较低导致的。

从表 2 和表 3 的实验结果可以看出, 无论是在纹理复杂度高还是低的数据集上, 本文提出的 ASAN 方法都可以有效地减轻数据集失配对基于深度学习的隐写分析器的影响。在嵌入率为 0.4 bit/pixel 的情况下, 本文方法将隐写分析器的准确率提

高 3.8%~11.5%, 平均提高 8.3%; 嵌入率为 0.2 bit/pixel 时, 本文方法将隐写分析器的准确率提高 10.1%~17.2%, 平均提高 12.8%。嵌入率为 0.2 bit/pixel 时的提升效果优于 0.4 bit/pixel, 两种嵌入率下的检测准确率均高于 SRM、SRNet、J-Net。

表 2 数据集失配时准确率对比 (嵌入率为 0.4 bit/pixel)
Table 2 Detection accuracy comparison in dataset mismatch(0.4 bit/pixel)

数据集	隐写算法	SRM	SRNet	J-Net	ASAN
UCID v.2	SUNI	56.9%	81.6%	83.0%	91.5%
	WOW	55.9%	81.0%	82.3%	91.9%
	HUGO	59.2%	78.8%	81.7%	88.7%
DIV2K	SUNI	56.0%	83.5%	84.5%	93.4%
	WOW	55.0%	80.5%	81.4%	92.0%
	HUGO	57.0%	80.1%	81.7%	90.7%
Flickr 25K	SUNI	50.5%	87.4%	87.9%	91.6%
	WOW	51.5%	84.3%	86.5%	88.1%
	HUGO	50.6%	88.4%	91.0%	95.0%
平均值		54.7%	82.8%	84.4%	91.4%

表 3 数据集失配时准确率对比 (嵌入率为 0.2 bit/pixel)
Table 3 Detection accuracy comparison in dataset mismatch(0.2 bit/pixel)

数据集	隐写算法	SRM	SRNet	J-Net	ASAN
UCIDv.2	SUNI	51.5%	70.3%	74.8%	80.4%
	WOW	52.1%	68.2%	67.8%	84.0%
	HUGO	53.1%	70.6%	70.7%	81.5%
DIV2K	SUNI	52.0%	71.6%	75.9%	83.1%
	WOW	51.8%	69.9%	70.4%	84.5%
	HUGO	53.0%	70.2%	70.9%	81.1%
Flickr 25K	SUNI	50.0%	76.8%	79.6%	89.2%
	WOW	50.5%	71.0%	74.4%	88.2%
	HUGO	50.0%	78.3%	81.0%	90.5%
平均值		51.6%	71.9%	73.9%	84.7%

3.3 隐写算法失配实验

隐写算法失配实验将已知隐写算法 (用来训练 SRM 和 SRNet 模型的算法) 进行嵌入得到的 500 个图像对作为源领域数据, 将其他隐写算法进行嵌入得到的 500 个图像对作为目标领域数据。实验数据集为 BOSSbase1.01。在隐写算法失配时, SRM 仍受失配影响, 其检测准确率下降, 但下降幅度低于数据集失配。可见相比隐写算法

失配, SRM 可能对数据集之间的不一致更为敏感。而对于 SRNet, 大部分隐写算法失配情况下模型的准确率会下降, 准确率上升的特例出现在测试算法为 S-UNIWARD 时。

表 4 和表 5 分别给出了嵌入率为 0.4 bit/pixel 和 0.2 bit/pixel 时隐写算法失配的实验结果。从实验结果可以看出本文提出的 ASAN 方法可以有效地减轻隐写算法失配对基于深度学习的隐写分析器的影响。嵌入率为 0.4 bit/pixel 时, 本文方法将隐写分析器的准确率提高 4.7%~10.2%, 平均提高 7.2%; 嵌入率为 0.2 bit/pixel 时, 本文方法将隐写分析器的准确率提高 2.5%~11.2%, 平均提高 6.8%, 检测准确率均高于 SRM、SRNet、J-Net。

表 4 隐写算法失配时准确率对比 (嵌入率为 0.4 bit/pixel)
Table 4 Detection accuracy comparison in steganographic algorithm mismatch (0.4 bit/pixel)

训练集	测试集	SRM	SRNet	J-Net	ASAN
SUNI	WOW	67.4%	73.1%	80.0%	82.2%
	HUGO	69.3%	74.6%	78.6%	79.9%
WOW	SUNI	71.9%	87.8%	88.2%	92.5%
	HUGO	67.7%	84.0%	84.9%	89.1%
HUGO	SUNI	67.6%	85.2%	85.3%	94.3%
	WOW	62.9%	82.2%	83.1%	92.4%
平均值		67.8%	81.2%	83.4%	88.4%

表 5 隐写算法失配时准确率对比 (嵌入率为 0.2 bit/pixel)
Table 5 Detection accuracy comparison in steganographic algorithm mismatch (0.2 bit/pixel)

训练集	测试集	SRM	SRNet	J-Net	ASAN
SUNI	WOW	57.0%	67.3%	68.7%	72.1%
	HUGO	57.9%	68.9%	69.3%	73.4%
WOW	SUNI	59.8%	74.9%	74.7%	85.9%
	HUGO	56.8%	72.1%	72.2%	83.3%
HUGO	SUNI	57.6%	73.2%	74.8%	75.7%
	WOW	54.5%	70.7%	71.4%	77.4%
平均值		57.3%	71.2%	71.9%	78.0%

3.4 t-SNE 可视化

为了进一步评估 ASAN 的性能, 本文使用 t-SNE 降维技术^[39]对数据集失配时网络学习到的特征分布进行可视化, 如图 2 所示。来自源领域的样本用黄色数值表示, 来自目标领域的样本用蓝色数值表示, 两个领域各 1 000 个样本点。数值 0 代表

载体图像, 数值 1 代表载密图像。同时为了更直观地表示, 本文将载体图像用更浅的黄色或蓝色表示。该部分实验将数据集从 BOSSbase1.01 迁移至 UCIDv.2, 检测 S-UNIWARD 算法的实验结果, 嵌入率为 0.4 bit/pixel。

图 2(a)为进行领域自适应之前 SRNet 提取的特征分布图, 可以观察到, 在源领域 (黄色) 上, 载体图像特征 (浅黄色的 0) 和载密图像特征 (深黄色的 1) 之间存在聚类现象, 很少有样本混淆到另一个类别的特征分布中, 此时模型的检测准确率较高。然而, 在目标领域 (蓝色) 上, 由于载体源失配问题的发生, 特征的分布发生变化, 模型检测准确率降低。图 2(b)和图 2(c)分别为采用 J-Net^[19]和本文方法 ASAN 进行领域自适应之后的特征分布图。可以看到本文提出的方法可以更好地将源领域和目标领域进行对齐, 扩大类间距离, 减小类内距离, 从而提升模型的准确率。结合 t-SNE 特征分布图进行分析, 载体源失配会导致预训练隐写分析模型提取的特征分布发生变化, 从而使检测准确率下降。通过充分利用待测图像的信息, J-Net 和 ASAN 可以将源领域数据和目标领域数据的特征分布进行对齐以提升检测准确率。而 J-Net 全局地将特征分布进行了对齐, 虽然对齐之后全局分布大体相似但具体到两个类别中, 分类器不能充分描述分类边界。ASAN 方法通过结合子领域自适应和对抗学习两种策略, 局部地对齐属于同一类别的相关子领域, 同时使源领域和目标领域的特征更加相似, 特征分布更紧密, 达到了更好的分类效果。

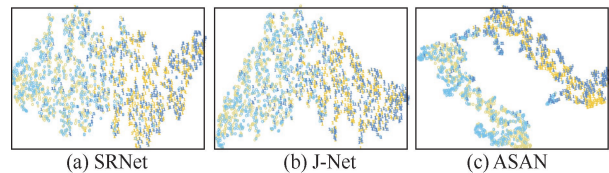


图 2 t-SNE 对比

Figure2 Comparison result of t-SNE

3.5 消融实验

消融实验探究了子领域自适应和对抗学习对最终分类效果的不同影响。选取的载体源失配场景为数据集失配, 实验设置与 3.2 节一样, 表 6 与表 7 分别为嵌入率为 0.4 bit/pixel 和 0.2 bit/pixel 时的消融实验结果。表中 ADV 列表示只采用对抗学

习, DSAN 列表示只采用子领域自适应, 而 ASAN 列则表示同时采用这两种策略进行领域自适应的结果。由表 6 和表 7 中实验结果可以看出子领域自适应和对抗学习两种方法都可以提升模型的准确率, 二者效果接近, 而将二者结合起来可以更好地进行领域自适应, 从而提升模型的准确率。

为了更直观地观察这种差异, 本实验采用 UCIDv.2 数据集, 检测 S-UNIWARD 算法时的准确率曲线如图 3 所示。其中图 3(a)是嵌入率为 0.4 bit/pixel 时的准确率曲线, 图 3(b)是嵌入率为 0.2 bit/pixel 时的准确率曲线。可以观察到, 在训练过程中 ASAN 的提升高于其他两种方法, 这种优势在嵌入率为 0.4 bit/pixel 时更为明显。

表 6 消融实验准确率对比 (嵌入率为 0.4 bit/pixel)
Table 6 Detection accuracy comparison of ablation experiments (0.4 bit/pixel)

数据集	隐写算法	ADV	DSAN	ASAN
UCID v.2	SUNI	85.5%	85.1%	91.5%
	WOW	87.7%	86.0%	91.9%
	HUGO	84.4%	84.6%	88.7%
DIV2K	SUNI	86.7%	86.7%	93.4%
	WOW	85.4%	85.6%	92.0%
	HUGO	85.5%	84.9%	90.7%
Flickr 25K	SUNI	90.1%	92.7%	91.6%
	WOW	87.3%	89.9%	88.1%
	HUGO	93.2%	93.8%	95.0%
平均值		87.3%	87.7%	91.4%

表 7 消融实验准确率对比 (嵌入率为 0.2 bit/pixel)
Table 7 Detection accuracy comparison of ablation experiments (0.2 bit/pixel)

数据集	隐写算法	ADV	DSAN	ASAN
UCID v.2	SUNI	75.0%	76.9%	80.4%
	WOW	72.7%	77.3%	84.0%
	HUGO	72.9%	76.4%	81.5%
DIV2K	SUNI	76.1%	80.7%	83.1%
	WOW	73.1%	79.5%	84.5%
	HUGO	75.0%	77.0%	81.1%
Flickr 25K	SUNI	84.7%	84.9%	89.2%
	WOW	75.3%	78.5%	88.2%
	HUGO	86.4%	85.4%	90.5%
平均值		76.8%	79.6%	84.7%

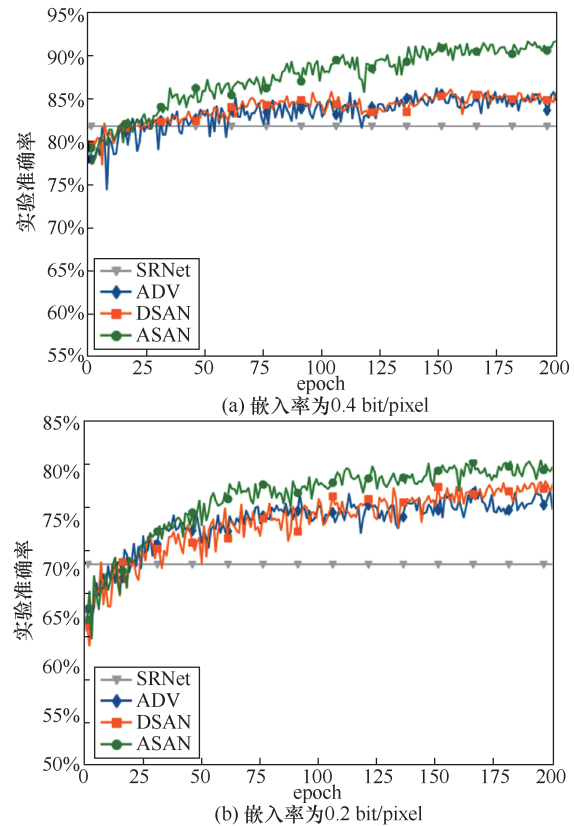


图 3 消融实验准确率曲线
Figure 3 Accuracy curves of ablation experiments

4 结束语

为了减轻载体源失配问题对基于深度学习的隐写分析模型的负面影响, 本文提出了对抗子领域自适应网络 ASAN, 从学习域无关特征和减小子领域特征分布差异这两个角度, 将子领域自适应和对抗学习这两种领域自适应方法结合, 提高了隐写分析模型在载体源失配情况下的准确率。通过在多个公开数据集以及多种嵌入率上进行了大量的实验, 结果证明本文所提出的方法可以在数据集失配以及隐写算法失配时都提供较好的检测性能。在未来的工作中, 将继续研究 JPEG 域的失配隐写分析问题, 进一步提升基于深度学习的隐写分析方法的实际应用价值。

参考文献:

- [1] HOLUB V, FRIDRICH J, DENEMARK T. Universal distortion function for steganography in an arbitrary domain[J]. EURASIP Journal on Information Security, 2014, (1): 1-13.
- [2] HOLUB V, FRIDRICH J. Designing steganographic distortion using directional filters[C]//Proceedings of 2012 IEEE International

- Workshop on Information Forensics and Security (WIFS). 2012: 234-239.
- [3] PEVNÝ T, FILLER T, BAS P. Using high-dimensional image models to perform highly undetectable steganography [C]//Proceedings of International Workshop on Information Hiding. Springer. 2010: 161-177.
 - [4] FRIDRICH J, KODOVSKY J. Rich models for steganalysis of digital images[J]. IEEE Transactions on Information Forensics and Security, 2012, 7(3): 868-882.
 - [5] SONG X, LIU F, YANG C, et al. Steganalysis of adaptive JPEG steganography using 2D Gabor filters[C]//Proceedings of the 3rd ACM Workshop on Information Hiding and Multimedia Security. 2015: 15-23.
 - [6] FENG G, ZHANG X, REN Y, et al. Diversity-based cascade filters for JPEG steganalysis[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 30(2): 376-386.
 - [7] YE J, NI J, YI Y. Deep learning hierarchical representations for image steganalysis[J]. IEEE Transactions on Information Forensics and Security, 2017, 12(11): 2545-2557.
 - [8] YEDROUDJ M, COMBY F, CHAUMONT M. Yedroudj-net: an efficient CNN for spatial steganalysis[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018: 2092-2096.
 - [9] BOROUMAND M, CHEN M, FRIDRICH J. Deep residual network for steganalysis of digital images[J]. IEEE Transactions on Information Forensics and Security, 2019, 14(5): 1181-1193.
 - [10] ZHANG R, ZHU F, LIU J, et al. Depth-wise separable convolutions and multi-level pooling for an efficient spatial CNN-based steganalysis[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 1138-1150.
 - [11] YOU W, ZHANG H, ZHAO X. A siamese CNN for image steganalysis[J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 291-306.
 - [12] 沈军, 廖鑫, 秦拯, 等. 基于卷积神经网络的低嵌入率空域隐写分析[J]. 软件学报, 2021, 32(9): 2901-2915.
SHEN J, LIAO X, QIN Z, et al. Spatial steganalysis of low embedding rate based on convolutional neural network[J]. Journal of Software, 2021, 32(9): 2901-2915.
 - [13] JIA J, ZHAI L, REN W, et al. Transferable heterogeneous feature subspace learning for JPEG mismatched steganalysis[J]. Pattern Recognition, 2020, 100: 107105.
 - [14] XUE Y, YANG L, WEN J, et al. A subspace learning-based method for JPEG mismatched steganalysis[J]. Multimedia Tools and Applications, 2019, 78(7): 8151-8166.
 - [15] YANG Y, KONG X, FENG C. Double-compressed JPEG images steganalysis with transferring feature[J]. Multimedia Tools and Applications, 2018, 77(14): 17993-18005.
 - [16] FENG C, KONG X W, LI M, et al. Contribution-based feature transfer for JPEG mismatched steganalysis[C]//Proceedings of 2017 IEEE International Conference on Image Processing (ICIP). 2017: 500-504.
 - [17] YANG Y, KONG X, WANG B, et al. Steganalysis on Internet images via domain adaptive classifier[J]. Neurocomputing, 2019, 351: 205-216.
 - [18] LONG M, WANG J, DING G, et al. Adaptation regularization: a general framework for transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2013, 26(5): 1076-1089.
 - [19] ZHANG X, KONG X, WANG P, et al. Cover-source mismatch in deep spatial steganalysis[C]//Proceedings of 18th Workshop on Digital Watermarking. 2019: 71-83.
 - [20] HU D, MA Z, FAN Y, et al. Study on the interaction between the cover source mismatch and texture complexity in steganalysis[J]. Multimedia Tools and Applications, 2019, 78(6): 7643-7666.
 - [21] 张逸为, 张卫明, 俞能海. 针对特定测试样本的隐写分析方法[J]. 软件学报, 2018, 29(4): 987-1001.
ZHANG Y W, ZHANG W M, YU N H. Specific testing sample steganalysis[J]. Journal of Software, 2018, 29(4): 987-1001.
 - [22] OZCAN S, MUSTACOGU A F. Transfer learning effects on image steganalysis with pre-trained deep residual neural network model[C]//Proceedings of 2018 IEEE International Conference on Big Data (Big Data). 2018: 2280-2287.
 - [23] EL BEJ R, SAIDI M, HERMASSI H, et al. An improved CNN steganalysis architecture based on “Catalyst Kernels” and transfer learning[C]//Proceedings of International Conference on Digital Economy. 2018: 119-128.
 - [24] PEVNY T, FRIDRICH J. Merging Markov and DCT features for multi-class JPEG steganalysis[C]//Proceedings of Security, Steganography, and Watermarking of Multimedia Contents IX. 2007: 650503.
 - [25] HOLUB V, FRIDRICH J. Low-complexity features for JPEG steganalysis using undecimated DCT[J]. IEEE Transactions on Information Forensics and Security, 2015, 10(2): 219-228.
 - [26] KODOVSKÝ J, FRIDRICH J. Calibration revisited[C]//Proceedings of the 11th ACM Workshop on Multimedia and Security. 2009: 63-74.
 - [27] PAN S J, YANG Q. A survey on transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 22(10): 1345-1359.
 - [28] BORGWARDT K M, GRETTON A, RASCH M J, et al. Integrating structured biological data by kernel maximum mean discrepancy[J]. Bioinformatics, 2006, 22(14): 49-57.
 - [29] TZENG E, HOFFMAN J, ZHANG N, et al. Deep domain confusion: Maximizing for domain invariance[J]. arXiv preprint arXiv:1412.3474, 2014.
 - [30] LONG M, CAO Y, WANG J, et al. Learning transferable features with deep adaptation networks[C]//Proceedings of International Conference on Machine Learning. 2015: 97-105.
 - [31] LONG M, ZHU H, WANG J, et al. Deep transfer learning with joint adaptation networks[C]//Proceedings of International Conference on Machine Learning. 2017: 2208-2217.

- [32] ZHU Y, ZHUANG F, WANG J, et al. Deep subdomain adaptation network for image classification[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 32(4): 1713-1722.
- [33] 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.
- [34] YU C, WANG J, CHEN Y, et al. Transfer learning with dynamic adversarial adaptation network[C]//Proceedings of 2019 IEEE International Conference on Data Mining (ICDM). 2019: 778-786.
- [35] BAS P, FILLER T, PEVNÝ T. "Break our steganographic system": the ins and outs of organizing BOSS[C]//Proceedings of International Workshop on Information Hiding. 2011: 59-70.
- [36] SCHAEFER G, STICH M. UCID: an uncompressed color image database[C]//Proceedings of Storage and Retrieval Methods and Applications for Multimedia. 2004: 472-480.
- [37] AGUSTSSON E, TIMOFTE R. Ntire 2017 challenge on single image super-resolution: Dataset and study[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2017: 126-135.
- [38] HUISKES M J, LEW M S. The MIR flickr retrieval evaluation[C]//Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval. 2008: 39-43.
- [39] VAN DER MAATEN L, HINTON G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9(11): 2579-2605.

[作者简介]



章蕾(1998-),女,吉林长春人,四川大学硕士生,主要研究方向为图像隐写分析与深度学习。



王宏霞(1973-),女,河北赵县人,四川大学教授、博士生导师,主要研究方向为多媒体信息安全、信息隐藏与数字水印、数字取证、智能信息处理。