基于域适应的电磁泄漏还原图像中文文本识别

吕志强[1,2], 于 超[1,2], 李海洋[1,2], 张 宁[1,2]

[1] 中国科学院信息工程研究所第四研究室 北京 中国 100093
[2] 中国科学院大学网络空间安全学院 北京 中国 100093

摘要 计算机显示系统会在信息的传输和显示过程中产生电磁泄漏, 利用 TEMPEST 技术(Transient Electromagnetic Pulse Emanation Surveillance Technology), 可以很容易地将辐射的电磁信息截获,在通过电磁泄漏途径获取的视频图像中, 图像中的文字往往含有十分重要的信息, 也是我们更为关注的内容, 因此对于从电磁泄漏途径得到的图像, 其文字区域的识别是一项至关重要的工作,然而通过接收机接收的电磁泄漏的视频信号信噪比很低, 然而通过接收机接收的电磁泄漏的视频信号信噪比很低, 这使得还原的图像难以进行有效的文本识别。现有的针对低信噪比中文文本图像的文字识别工作非常少。在本文中, 我们提出了一种基于域适应思想的 CRNN(Convolutional Recurrent Neural Network)文字识别模型。该模型用电磁泄漏环境下采集的无标注文本图像作为目标域数据, 正常的带标注文本图像作为源域数据, 将卷积神经网络(Convolutional Neural Network, CNN)结合上域判别模块(Domain Discrimination Module,DDM), 然后采用半监督学习的训练方式使得卷积神经网络最终提取的特征层是带随机噪声的目标域数据集和正常的源域数据集的公共特征, 由于是两者的公共特征, 因此也就最小化各种随机噪声带来的影响, 并且可以最大化地利用这些鲁棒的公共特征来进行后续的字符分类。提升了真实噪声环境条件下的文字识别准确率。本文模型在电磁泄漏还原实景下的公开数据集 RCTW-17、CASIA-10 上进行了测试, 评价指标为精确率(Precision)和归一化平均编辑距离(Normalized Average Edit Distance, NAED), 相比于主流的识别模型, 基于域适应的 CRNN 对于电磁泄漏还原的文本图像的精确率和归一化平均编辑距离有了明显的提升。

关键词 电磁泄漏, 文本识别, 域适应, 半监督学习, 神经网络
中图法分类号 TP183/TP309.2 DOI 号 10.19363/J.cnki.cn10-1380/tn.2026.03.16

Chinese Text Recognition in Electromagnetic Emission Reconstructed Images Based on Domain Adaptive

LV Zhiqiang[1,2], YU Chao[1,2], LI Haiyang[1,2], ZHANG Ning[1,2]

[1] The 4th Laboratory, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100093, China

Abstract Electromagnetic emission exists in the process of information transmission and display in computer display system. Using TEMPEST technology (Transient Electrical Pulse Analysis Surveillance Technology), radiated electromagnetic information can be easily intercepted. In video images obtained through electromagnetic leakage, the text in the image often contains very important information, which is also the focus of our attention. Therefore, for images obtained through electromagnetic leakage, the recognition of the text area is a crucial task. However, the signal-to-noise ratio of the emitted video signal received by the receiver is very low, and it makes the restored image difficult for effective text recognition. There are few text recognition methods for Chinese text images with low signal-to-noise ratio. In this paper, We propose a CRNN (Convolutional Recurrent Neural Network) text recognition model based on domain adaptation, which uses the unlabeled text images collected in the electromagnetic emission environment as the target domain data, and uses the normal labeled text images as the source domain data. The model combines the Convolutional Neural Network (CNN) with the Domain Discrimination Module(DDM), and then then the semi supervised learning training method is adopted to make the final feature layer extracted by the convolutional neural network be the common features of the target domain dataset with random noise and the normal source domain dataset. As they are common features of both, the impact of various random noise is minimized, and these robust common features can be maximized for subsequent character classification. which improves the accuracy of text recognition in images emitted from target computer. This model was tested on publicly available datasets RCTW-17 and CASIA-10k in the context of electromagnetic leakage restoration, and the evaluation indicators were Precision and Normalized Average Edit Distance (NAED). Compared with mainstream recognition models, The domain adaptation based CRNN has significantly improved the accuracy and normalized average editing distance of text images restored by electromagnetic leakage.

通讯作者: 吕志强, 博士, 副研究员, Email: lvzhiqiang@iie.ac.cn。
本课题得到国家重点研发计划(No.2018YFF01014303)资助。
收稿日期: 2020-12-24; 修改日期: 2021-03-03; 定稿日期: 2023-08-13

Key words electromagnetic emission; text recognition; domain adaptive; few-shot learning; neural network

1 背景

当今的信息化时代, 计算机等信息技术设备被广泛应用于各个领域。虽然大部分的信息技术设备都有相应的安全机制, 例如防火墙、身份加密系统等, 但是在这些信息技术设备的运转过程中会产生大量的电磁辐射, 这些电磁辐射携带着计算机等信息技术设备的重要敏感信息, 利用 TEMPEST 技术[1] (Transient Electromagnetic Pulse Emanation Surveillance Technology), 可以很容易地将辐射的电磁信息截获, 获取其正在处理的文字和图形信息, 给信息安全带来了很大的威胁[2-6]。

TEMPEST 是针对由电子信息设备泄密电磁发射产生的电磁信息安全问题的研究, 主要研究信息设备屏幕显示系统泄密发射和抑制理论。TEMPEST 技术主要包括了对电磁泄漏信号中所携带的敏感信息进行分析研究、测试接收、截获还原以及泄漏防护的一系列技术。由于技术的敏感性和复杂性, 早期的TEMPEST研究基本由政府和军方垄断, 随着信息和通信技术的迅速发展与广泛应用, 现在 TEMPEST 已经成为电磁兼容领域的重要组成部分。计算机是现代社会中不可缺少的信息设备, 由于视频信号的重要性, 计算机视频信息泄密电磁发射一直是 TEMPEST 领域最重要的研究热点。在 1985 年, 荷兰学者 Van Eck 第一次展示了用电视机接收系统接收计算机显示器的电磁辐射, Van Eck 分析了计算机显示器的信息泄漏安全问题, 并且制作了一个低成本的基于电视机改造的截获接收机, 在几百米范围内成功窃取了目标计算机的显示图像[2]。Smulders 于 1990 年对 RS-232 总线的信息泄漏机理进行了分析, 并给出了窃取实验结果[6]。1998 年, Markus 等人在实验室中利用 Data Safe/ESL Model 400 TEMPEST 信息泄漏监测器实现了对文本图像的截获还原, 并分析了视频辐射信息不同频率成分的信息泄漏效应[7]。2002 年 Markus 利用光电倍增管接收了彩色显示器电子扫描视频信号的电磁辐射, 并用接收机对文字视频泄漏信息进行了再现[8]。2011 年, 日本大阪大学的 Hidenori Sekiguchi 和 Shinji Seto 从信息技术设备的可接收距离的角度评估了可以利用信息技术设备的电磁辐射来重建视频信息泄漏的威胁, 给出了不同频率范围最远接收距离的理论估值[9]。

在通过电磁泄漏途径获取的视频图像中, 图像中的文字往往含有十分重要的信息, 也是我们更为关注的内容, 因此对于从电磁泄漏途径得到的图像, 其文字区域的识别是一项至关重要的工作。传统的文字识别技术主要有模板匹配方法[10]和基于字符特征的方法[11-15], 模板匹配方法是将图像进行预处理后与模板图像进行相关运算得到匹配值, 匹配值高的即认为是该字符, 其中较为代表性的工作为 Yokobayashi 和 Wakahara 在 CMY 颜色平面上根据字符颜色和宽度进行局部自适应二值化并以最大直方图宽度对字符和背景进行分割, 然后使用归一化的互相关作为匹配度量, 检测目标图像和图像模板的匹配程度[10], 这种文字识别的方法的实现较为简单, 对于背景单一, 字符显示完整的图像有较高的识别率, 但是对于背景稍微复杂的文本图像, 其识别能力大大下降。基于字符特征的方法是通过统计字符的特征, 然后根据这些特征的相似度利用分类器进行分类判别, de Campos 等人针对字符的特点设计了六种字符局部特征子—形状上下文, 几何模糊, 缩放不变特征变换, 旋转图, 滤波器最大响应和区域描述子, 然后再把这六种特征送入分类器进行分类[11], 这种方法虽然较于模板匹配方法的鲁棒性更好一些, 但是这种方法最大的缺点就是人们需要花费大量时间做特征的设计, 这是一件相当费工夫的事情, 且人工设计的特征较为单一, 此类单一的特征在字体变化, 模糊或背景干扰时泛化能力也很差, 并不能满足业界需求。

随着近年来深度学习技术的快速发展, 针对传统文本识别解决方案的不足, 学界业界纷纷基于深度学习来训练文本识别模型, 其性能也较于传统文本识别的解决方案有了很大提升。Wang 等人提出了第一个端到端的文本识别模型, 该模型基于卷积神经网络首先通过 MSER 算法提取候选字符区域, 再利用后续的卷积神经网络和分类器对提取的字符候选区域进行分类[16]。Jaderberg 等人提出了一种新的卷积神经网络架构, 允许字符检测与分类共享卷积特征[17]。Bai 等人提出了称为卷积递归神经网络(CRNN)的模型, 因为它是 CNN 和 RNN (Recurrent Neural Network)的组合, 因此对于文字这种序列式的对象, CRNN 具有优于传统的卷积神经网络模型的独特优势, 实现了十分优秀的性能指标, 是近年来极具代表性和被广泛应用的文本识别模型[18]。Yin 等人在滑动窗口上的字符分类器输出使用基于连接器时间分类(Connectionist Temporal Classification, CTC)[19,20]的算法进行解码, 避免了基于 RNN 的模型训练时导致的梯度消失/爆炸现象, 可以简单有效地训练模型[21]。Liu 等人在文字特征提取网络中引

入了面向不规则文字的空间注意力机制, 提升了对于不规则文字的识别精度[22]。Qiao 等人提出了一个语义增强的编解码框架, 它通过预测全局的语义信息去引导解码过程, 并利用一个预训练语言模型的词嵌入来对预测的语义信息进行监督, 提升了模型对于低质量图像的识别精度[23]。

虽然近年来基于深度学习的文本识别技术相较于传统的文本识别技术在性能上有了很大的提升, 但是目前大部分的工作都集中在提升高信噪比英文数据集的对不规则字体及复杂背景的泛化能力和识别精度上, 针对低信噪比中文数据集的文本识别的工作很少, 且目前对于低信噪比图像的去噪技术[24-26]也需要人工生成大量带噪声的数据集来训练相应的模型, 而合成的带噪声图像与真实电磁泄漏环境中的带噪声图像存在一定差异, 因此泛化效果并不很好。我们通过电磁泄漏获得的图像大部分都是较为规则的长序列字体但是却包含大量的随机噪声, 其随机噪声以采集期间由于不良照明或高温引起的高斯噪声和图像传输期间产生的瑞利噪声为主, 由于输入的文本图片带有这些随机噪声, 因此通过常规文本识别模型的卷积神经网络提取的特征也带有随机噪声的特征, 将这些特征送入后续的字符分类网络后会极大地影响字符的判别, 导致识别文本的精度过低、编辑距离较大。综上所述针对电磁泄漏还原的低信噪比图像的中文文本识别是一项具有重要实际意义的任务。

针对上述问题, 本文提出了基于域适应的 CRNN 模型。本文主要创新点和贡献如下。

(1) 本文提出了一种将局部文字特征和全局文字特征都进行域对齐的模型。本模型训练时采用半监督学习的方式, 将训练的样本数据划分为带标注的源域和无标注的目标域, 其中源域数据为高信噪比文本图像, 目标域数据为电磁泄漏还原的文本图像, 在卷积神经网络的浅层和最后一层分别结合上带有梯度反转层(Gradient Reversal Layer, GRL)的局部域判别模块和全局域判别模块, 从而在局部和全局层面上对目标域数据集和源域数据集的字符特征分布进行对齐, 使得卷积神经网络最终提取的特征层是带随机噪声的目标域数据集和正常的源域数据集的公共特征, 由于是两者的公共特征, 因此也就最小化各种随机噪声带来的影响, 并且可以最大化地利用这些鲁棒的公共特征来进行后续的字符分类。

(2) 本文在全局域判别模块中引入了 Focal Loss 函数。由于中文文本图像具有不同的字体扭曲失真、字迹和真实背景的不同组合, 因此在全局图像级别将源域数据和带目标域数据的整个分布彼此完全匹配既十分困难也没有必要。根据这一特点, 我们在全局域判别模块引入了 Focal Loss 函数, 增加了全局特征分布相似的源域数据和目标域数据对齐损失的权重, 使得模型的训练效果更好, 降低了训练难度, 防止梯度爆炸。

(3) 针对公开的标准中文数据集, 通过对比实验证明, 基于域适应的 CRNN 模型对于电磁泄漏还原的文本图像相比于其他主流识别模型的识别准确率更高。

本文共分为 4 节, 第 1 节对文章的研究背景、研究意义进行了阐述, 同时介绍了国内外在文本识别研究方面的现状; 第 2 节介绍基于域适应 CRNN 文本识别算法设计; 第 3 节介绍实验所需数据的构建细节以及实验验证与结果分析部分; 第 4 节是对全文的总结。

2 基于域适应 CRNN 文本识别算法设计

域适应 CRNN 网络结构如图 1 所示。原有的 CRNN 网络结构包含两个重要的模块: 文字特征学习模块, LSTM 上下文序列预测模块。基于域适应的 CRNN 网络结构将文字特征学习模块划分成低层次的特征学习模块和高层次的特征学习模块, 并分别在这两个模块后面添加了梯度反转层和对应的局部特征域判别模块和全局特征域判别模块。目标是尽量训练出更完美的域判别模块, 使其能准确地判别文本图像来自哪种环境, 同时尽量使共享的文字特征学习模块提取的特征更好地欺骗域判别器, 最

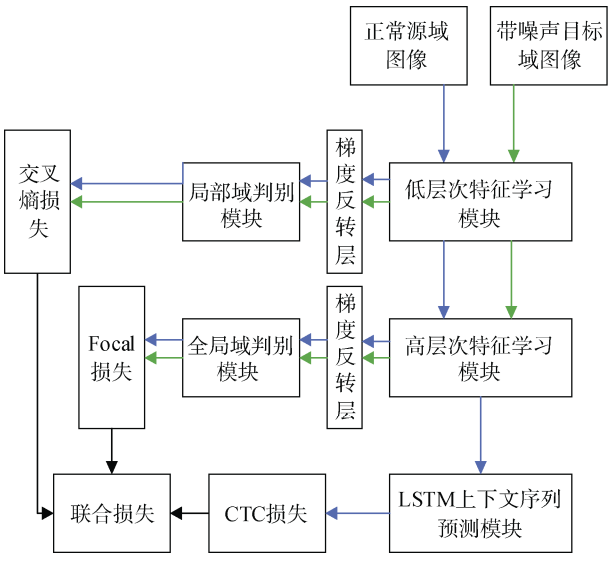


图 1 域适应 CRNN 网络结构图

Figure1 Domain adaptive CRNN framework

终使得域判别模块与文字特征学习模块在迭代过程中形成对抗, 从而使特征提取更偏向域无关的方向, 集中于要识别的目标文字本身。该框架在训练时输入源域和目标域样本图像, 有标注的源域文本图像通过文字特征学习模块、域适应模块和 LSTM 上下文序列预测模块来修正文字识别网络和解决域适应问题, 无标注的目标域文本图像仅通过文字特征学习模块和域适应模块来解决域适应问题, 最终将识别损失(对于源域样本)和域适应损失(对于所有样本)最小化, 使得两个不同域上的数据在高维特征空间上的数据分布相似, 以最大化地提取与噪声无关的公有特征。

2.1 域适应结构的理论分析

定义 1. 迁移学习(Transfer Learning)。给定一个有标签的源域 $D_s=\{x_i^s,y_i^s\}_{i=1}^{n_s}$ (其中 x_i^s 为源域样本, y_i^s 为源域样本标签, i 为源域样本索引, n_s 为源域样本数量)和一个无标签的目标域 $D_t=\{x_j^t\}_{j=1}^{n_t}$ (其中 x_j^t 为目标域样本, j 为目标域样本索引, n_t 为源域样本数量), 源域和目标域数据分布不同, 即 $P(x_s)\neq P(x_t)$, 迁移学习的任务就是通过有标签的源域数据 D_s 学习一个机器学习网络 $f: x \rightarrow y$ 来预测目标域数据 D_t 的标签 y_t[27]。

定义 2. 域适应(Domain Adaptation)。给定一个有标签的源域 $D_s=\{x_i^s,y_i^s\}_{i=1}^{n_s}$ 和一个无标签的目标域 $D_t=\{x_j^t\}_{j=1}^{n_t}$, 假设源域和目标域的特征空间和标签空间相同, 并且边缘分布概率也相同 $P(y_s|x_s)=P(y_t|x_t)$, 但是这两个域的边缘分布不同, 即 $P(x_s)\neq P(x_t)$。域适应算法的任务就是, 利用有标签的源域数据 D_s 学习一个机器学习网络 $f: x \rightarrow y$ 来预测目标域数据 D_t 的标签 y_t[27]。

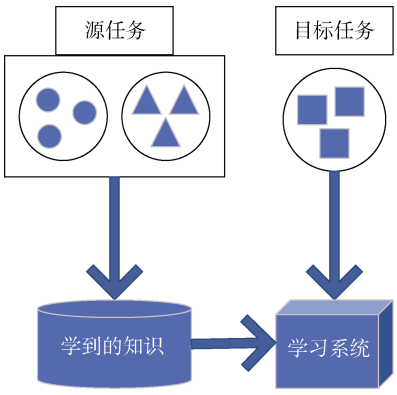


图 2 迁移学习的学习过程

Figure 2 Learning process of transfer learning

域适应和迁移学习都是从一个或多个源领域中通过训练该模型, 得出有用的知识并将其用在新的目标任务上(未标记的同一类有相似特征的物品或者未标记的不同类物品), 其本质是知识的迁移再利用。与迁移学习相比, 域适应还需要保证源域和目标域的特征空间和标签空间相同, 并且边缘分布概率也相同, 可以说域适应是一种约束条件更为严格的迁移学习。

文本识别问题是要给出图像文本区域的各字符类别, 我们用 C 表示文本区域的类别信息, F 表示输入的文本图片, 我们可以把文字识别问题视作学习并求解后验概率 $P(C|F)$的问题。由于带有随机噪声的目标域和正常的源域数据分布不同, 因此两个领域之间存在域偏移问题, 即带有随机噪声的目标域的联合分布 $P_t(C,F)$和正常的源域的联合分布 $P_s(C,F)$ 并不相同。根据贝叶斯公式有

$$P(C,F)=P(C|F)P(F)$$

我们假设目标域的条件概率 $P_t(C|F)$和源域的条件概率 $P_s(C|F)$相同, 即给定一张图像, 不管图像属于哪个域, 文字识别的结果应该是相同的, 由贝叶斯公式来看, 目标域与源域产生域偏移问题的主要原因是带噪声的文本图像和不带噪声的文本图像的边缘分布不相同, 即 $P_t(F)\neq P_s(F)$, 因此相除两者之间域差异使得其联合分布接近的主要方法就是要使得两者的边缘分布尽可能相同。总而言之, 我们希望不管是来自带有随机噪声的目标域文本图片还是来自正常的源域文本图片, 文字识别网络都可以进行正确的文本识别, 我们的模型应该尽可能使从不同域的图片映射到高维空间向量特征且服从相同分布, 将目标域和源域的图像在高维特征空间对齐, 即 $P_t(F)=P_s(F)$。

对于卷积神经网络而言, 越深的卷积层学习到的图像特征越抽象。以图 3 为例, 低层次特征往往是泛化的、易于表达的, 如纹理、颜色、边缘、棱角等, 高层次特征往往是复杂的、抽象的, 是具有辨别性的关键特征。对于卷积神经网络的各层输出来说, 浅层(靠近输入)往往能提取到上述的低层次特征, 深层(靠近输出)往往能提取到上述的高层次特征, 最后使用后续特定任务的网络分析最后的高层次特征图并给出预测结果[28]。

随着卷积神经网络的加深, 网络的特征图越来越多, 其所包含的信息也越来越复杂, 只对齐最终的特征图不能保证域适应结果完善, 因此, 为了使带随机噪声的目标域和正常的源域文本图像更好地在高维特征空间对齐, 我们设计了两个域判别模块, 第一个域判别模块对应着低层次的特征学习模块,

主要是用于对齐文本图像的低层次中文字符局部特征，这些低层次中文字符特征有：“横线”“竖线”“斜线”“横折”“左半圆”“右半圆”等，第二个域判别模块对应着高层次的特征学习模块，主要用于对齐文本图像的高层次中文字符全局特征，这些高层次的特征包含更多的文本语义信息，是更为抽象的特征。

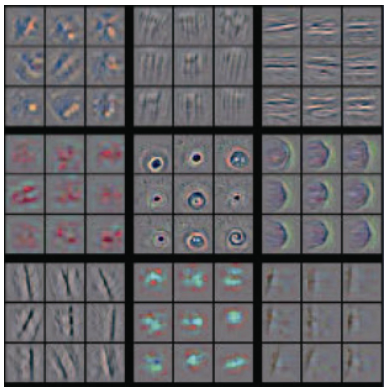


图 3　卷积神经网络学习的低层次特征

Figure 3　Low-level features learned by convolutional neural networks

2.2　全局域适应模块

全局域适应模块由高层次特征学习模块、梯度反转层和全局域判别模块构成。高层次的特征学习模块由二维卷积神经网络组成，能够在低层次特征学习模块学到的局部文字特征的基础上进行全局语义特征的学习，学到的高层次特征图将会被送入全局域判别模块和后续的 LSTM 上下文序列预测模块，其具体参数如表 1 所示。

表 1　高层次特征学习模块参数

Table1　Configurations of High-level feature learning

模块	类型
输入	经低层次特征学习模块学得的局部特征图，尺寸:Batch×256×8×32
	Conv, in:256,out:256,k:3×3,ReLU
	Conv, in:256,out:256,k:3×3,BatchNorm,ReLU
	MaxPool,k:2×2,s:2×1,p:1×1
	Conv, in:256,out:512,k:3×3,BatchNorm,ReLU
	Conv, in:512,out:512,k:3×3,BatchNorm,ReLU
	MaxPool,k:2×2,s:2×1,p:1×1
	Conv, in:512,out:512,k:2×2,ReLU
输出	经高层次特征学习模块学得的全局特征图，尺寸:Batch×512×1×33

定义 3. 梯度反转层。如图 4 所示，梯度反转层可以将神经网络反向传播的误差乘以一个负系数，使得该层之后的网络与该层之前的网络训练目标相反，以实现对抗训练的目的。

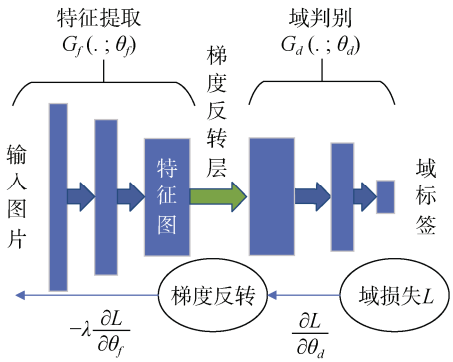


图 4　梯度反转层原理图

Figure 4　Schematic diagram of gradient reversal layer

全局域判别模块是接在高层次特征学习模块后的梯度反转层上的，其作用是对齐高层次的全局特征，该模块主要由三个卷积层构成，其参数如表 2 所示。

表 2　全局域判别模块参数

Table 2　Configurations of global domain adaptation

模块	类型
输入	经梯度反转层的全局特征图，尺寸:Batch×512×1×33
	Conv, in:512,out:128,k:3×3,BatchNorm,ReLU,Dropout
	Conv, in:512,out:128,k:3×3,BatchNorm,ReLU,Dropout
	Conv, in:128,out:128,k:3×3,BatchNorm,ReLU,Dropout
	Average Pooling:1×33,Fully connected:128×2, Softmax
输出	全局域判别模块预测的标签，尺寸:Batch×2

由于中文的文本图像具有不同的字体扭曲失真、字体笔迹和真实背景的不同组合，因此在全局图像级别将带有随机噪声的目标域文本图像和正常的源域文本图像和整个分布彼此完全匹配是非常困难且没有意义的。因此，我们设计的全局域判别模块是一种“弱”对齐模块，该模块的对抗对齐损失集中强调全局特征相似的源域和目标域对齐，而不强调整体不相似的样本对齐。全局特征相似指的是虽然目标域的文本图像却带有随机噪声而源域文本图像不带有随机噪声，但是两者的字迹、字体的扭曲和失真以及其背景等图像级别的特征差异是较小的，图 5 给出相应的文本图像示例，第一张和第二张是全局特征相似的目标域和源域文本图像，两者的字迹、背景以及文字的扭曲程度等全局特征比较接近，易于对齐并提取公共特征，而第一张和第三张图片的这些全局图像级别特征差异较大，难以对其进行对齐。

如图 6 所示，在特征空间中，容易区分的目标域样本距离源域样本非常远，而难以区分的目标域样本距离源域样本很近，即特征相似的目标域样本和源域样本在域判别阶段是难以区分的。由于我们要

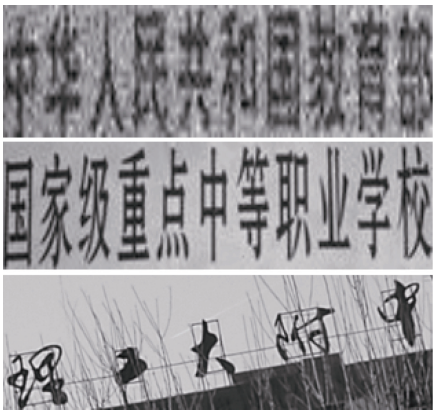


图 5 不同图像级别特征差异的源域和目标域文本图像

Figure 5 Source and target text images with different image levels feature differences

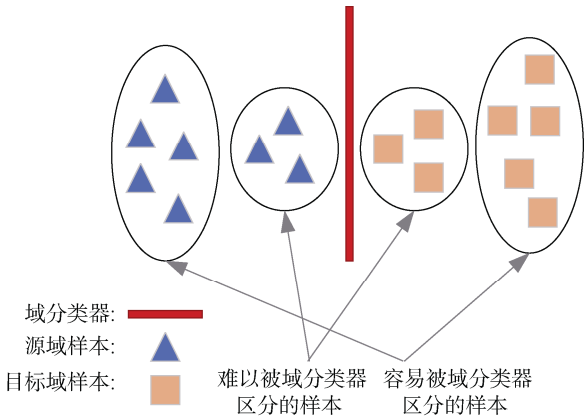


图 6 不同特征相似度的源域和目标域样本

Figure 6 Source domain and target domain samples with different feature similarity

将对抗对齐损失集中在全局特征相似的源域和目标域样本，而尽量忽略全局特征差异较大的样本，因此我们不能用常规的交叉熵损失函数来作为全局域对齐的损失函数，因为常规的交叉熵损失对于很容易被分类的样本也会产生不容忽视的损失值。在这里我们引入一种可以最大程度忽略容易被分类的样本而强调难以被分类的样本的损失函数—Focal Loss[29]，Focal Loss 被提出来主要是为了解决目标检测中检测框正样本与负样本的比例严重不均衡问题。因为在目标检测问题中，神经网络选出的框有成千上万个，然而这么多框中仅有极少的框为正样本目标框，其他大部分都为负样本背景框，这是不利于模型对具有类别信息的正样本的学习的，其次，难以分类的样本很少，易分类样本的数量多，这就导致易分类样本产生的梯度会主导模型的学习过程，会削弱难分类样本的作用，降低目标检测网络对难分类样本的判别能力。

Focal Loss 是在原来的交叉熵损失函数上进行的改进，交叉熵损失函数 $\mathrm{CE}(p_t)$ 如下:

$$\mathrm{CE}(p_t) = -\log(p_t)$$

p_t 定义为

$$p_t = \begin{cases} p, & y = 1 \\ 1-p, & y = 0 \end{cases}$$

其中 y 为样本标签，p 为网络预测的输出概率，而 Focal Loss 在交叉熵损失函数的基础上进行加权，让模型注重学习难以进行学习的样本，Focal Loss 的损失函数 $\mathrm{FL}(p_t)$ 如下:

$$\mathrm{FL}(p_t) = -(1-p_t)^{\alpha}\log(p_t)$$

这里的 $(1-p_t)^{\alpha}$ 称为调制系数，其中 $\alpha \geq 0$，不同的 α 对应着不同的损失曲线，图 7 和图 8 分别为当样本为正和样本为负的情况下，不同 α 对应的 Focal Loss 损失曲线，当 $\alpha = 0$ 时，Focal Loss 函数即为交叉熵损失函数。对于正类样本而言，被预测为正样本概率越大的样本就越易于被分类，由图 7 的 Focal Loss 函数曲线可以观察到，随着 α 的数值增大，这些易于被分类样本造成的损失越小。同理对于负类样本而言，被预测为正样本概率越小的样本就越易于被分类，由图 8 的 Focal Loss 函数曲线也可以观察到，随着 α 的数值增大，这些易于被分类样本造成的损失也越小。因此我们可以得出结论，不论正样本还是负样本，使用 Focal Loss 作为损失函数可以减少易于被分类样本的损失值，从而达到训练模型时减少易于被分类样本的权重，使模型在训练时更专注于难分类的样本的目的。

由于 Focal Loss 函数可以将损失集中在难以被分类的样本中，因此我们可以利用这一特性，应用 Focal Loss 函数在全局域判别模块上，以使得其对抗对齐损失集中强调全局特征相似的源域文本图像和目标域文本图像。在这里我们令源域的文本图像为

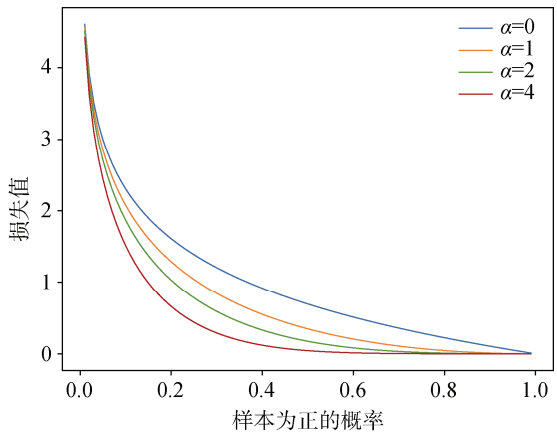


图 7 样本为正时的 Focal Loss 曲线

Figure 7 Focal Loss curve of positive samples

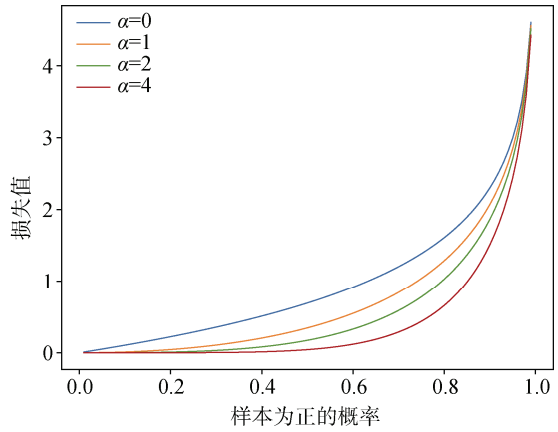


图 8　样本为负时的 Focal Loss 曲线

Figure 8　Focal Loss curve of negative samples

负样本, 其域标签为 0, 目标域的文本图像为正样本, 其域标签为 1, 此外我们令通过全局域判别模块得到的域标签 P 为

$$P = D_g(F_g(x_i))$$

其中, x_i 表示输入网络的图片, F_g 表示高层次的特征提取模块, D_g 表示全局域判别模块。此时整个全局域判别模块的损失定义如下:

$$L_{\text{globals}} = -\frac{1}{n_s}\sum_{n=1}^{n_s}\alpha(P_s)^{\gamma}\log(1-P_s)$$

$$L_{\text{globalt}} = -\frac{1}{n_t}\sum_{n=1}^{n_t}\beta(1-P_t)^{\gamma}\log(P_t)$$

$$L_{\text{global}} = L_{\text{globals}} + L_{\text{globalt}}$$

其中, P_s 与 P_t 分别为源域文本图像和目标域文本图像通过全局域判别模块预测的域标签, n_s 和 n_t 分别为源域文本图像和目标域文本图像的样本数量, γ 指数取 4, α 取 0.25, β 取 0.1, L_{globals} 为源域文本图像样本的损失, L_{globalt} 为目标域文本图像样本的损失, L_{global} 为总全局域判别模块的损失。

2.3　局部域适应模块

局部域判别模块由低层次特征学习模块、梯度反转层和局部域判别模块构成, 低层次特征学习模块由二维卷积神经网络构成, 用以学习层次较低的文字局部特征, 学到的局部特征图将会被送入高层次特征学习模块和局部域判别模块, 其具体参数如表 3 所示。

对于中文的文本图像来说, 由于其低层次的特征往往是如: “横”“竖”“撇”“点”“折”等这些易于泛化的特征, 它们不像全局特征那样组合得复杂多变, 因此对于带噪声的目标域文本图像和正常的源域文本图像而言, 两者的低层次局部特征是比较容易匹配对齐的, 所以我们对于从低层次特征学习模块学到的低层次特征使用最小平方损失函数进行“强”对齐训练, 即对于提取的低层次特征图的每一个像素都进行域判别, 对齐低层次特征图中的每一个局部感受野。如图 9 所示, 我们设计的局部域判别模块针对于学到的低层次特征图, 使用卷积核尺寸为 1×1 的全卷积模块 Conv1×1 Moudle 将维度为 256 的低层次特征图逐渐提取为维度为 1 且宽度和高度不变的特征图, 然后用 Sigmod 函数对最终提取的特征图的每个像素都进行域分类。全卷积模块 Conv1×1 Moudle 可以接受任意尺寸的输入特征图, 同时保留了原始输入特征图中的空间信息从抽象的特征中恢复出每个像素所属的类别, 即从图像级别的分类进一步延伸到像素级别的分类。局部域判别模块具体的参数如表 4 所示。

综上所述, 对于局部域判别模块预测的像素级域标签, 我们使用最小平方损失函数来进行训练,

表 3　低层次特征学习模块参数

Table 3　Configurations of low-level feature learning

模块	类型
输入	输入的原始图片, 尺寸:Batch×1×32×128
	Conv, in:1,out:64,k:3×3,ReLU
	MaxPool,k:2×2,s:2×2
	Conv, in:64,out:64,k:3×3,BatchNorm,ReLU
	MaxPool,k:2×2,s:2×2
	Conv, in:64,out:128,k:3×3,ReLU
	Conv, in:128,out:128,k:3×3,BatchNorm,ReLU
输出	经低层次特征学习模块学得的局部特征图, 尺寸:Batch×256×8×32

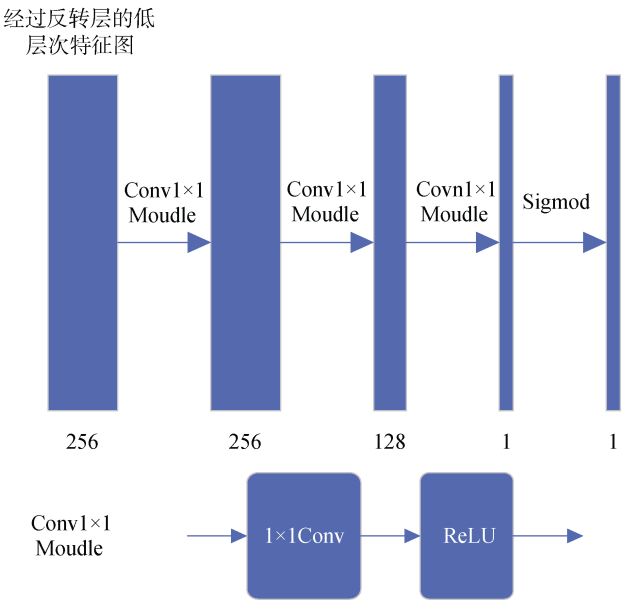


图 9　局部域判别模块架构

Figure 9　The framework of local domain adaptation

表 4 局部域判别模块参数

Table 4 Configurations of local domain adaptation

模块	类型
输入	经梯度反转层的局部特征图, 尺寸:Batch×256×8×32
	Conv, in:256,out:256,k:1×1,ReLU
	Conv, in:256,out:128,k:1×1,ReLU
	Conv, in:128,out:1,k:1×1,ReLU
	Sigmod
输出	局部域判别模块预测的像素级域标签, 尺寸:Batch×1×8×32

在这里我们同样令源域的文本图像为负样本, 其域标签为高和宽与输入的局部特征图相同的 0 像素值域标签图, 目标域的文本图像为正样本, 其域标签为高和宽与输入的局部特征图相同的 1 像素值域标签图, 此外我们还令通过局部域判别模块得到的像素级域标签 R_{hw} 为

$$R_{hw} = D_l(F_l(x_i))_{hw}$$

其中 x_i 表示输入网络的图片, h 和 w 表示预测的像素级域标签在整个预测的域标签图上的坐标, F_l 表示低层次的特征提取模块, D_l 表示局部域判别模块。此时整个局部域判别模块的损失定义如下:

$$L_{\text{locals}} = -\frac{1}{n_s HW}\sum_{i=1}^{n_s}\sum_{h=1}^{H}\sum_{w=1}^{W}\alpha(R_{shw})^2$$

$$L_{\text{localt}} = -\frac{1}{n_s HW}\sum_{i=1}^{n_t}\sum_{h=1}^{H}\sum_{w=1}^{W}\beta(1-R_{thw})^2$$

$$L_{\text{local}} = L_{\text{locals}} + L_{\text{localt}}$$

其中, R_{shw} 与 R_{thw} 分别为源域文本图像和目标域文本图像通过局部域判别模块预测的像素级域标签, H 和 W 分别为预测的域标签图的高度和宽度, n_s 和 n_t 分别为源域文本图像和目标域文本图像的样本数量, α 取 0.25, β 取 0.1, L_{locals} 为源域文本图像样本的损失, L_{localt} 为目标域文本图像样本的损失, L_{local} 为总局部域判别模块的损失。

2.4 序列语义信息标注模块

输入的文本图像在经过被域判别模块作用的特征提取模块后, 得到了富含语义信息的鲁棒特征图, 我们对此特征图进行序列化, 然后使用双向的 LSTM[30]对特征序列进行预测, 对序列中的每个特征向量进行学习, 并输出预测标签分布, 然后使用 CTC 转录[31], 把从双向 LSTM 获取的一系列标签分布转换成最终的标签序列。

由于 LSTM 是不能直接接受卷积神经网络提取的特征图作为输入的, 因此我们需要对提取到的特征图进行特征图序列化的调整。如图 10 所示, 共生成 33 个特征向量序列, 每一个特征向量在特征图(C=512,H=1,W=33)上按列从左到右生成, 每一列包含 512 维特征, 这意味着第 i 个特征向量是所有的特征图第 i 列像素的连接, 这 33 个特征向量就构成一个带有语义特征的向量序列 S=(S_1,S_2,S_3,...,S_{33}), 这些特征向量序列就作为 LSTM 循环的输入, 每个特征向量作为 LSTM 在一个时间步(time_step)的输入。由于卷积层, 最大池化层和激活函数都是在局部区域上执行, 因此它们是平移不变的, 如图 11 所示, 特征图的每列(即一个特征向量)对应于原始图像的一个矩形区域(称为感受野), 并且这些矩形区域与特征图上从左到右的相应列具有相同的顺序。

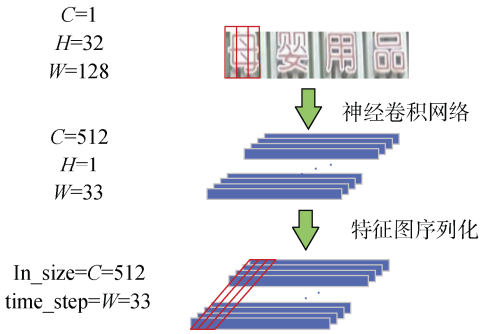


图 10 特征图序列化示意图

Figure 10 Feature map serialization

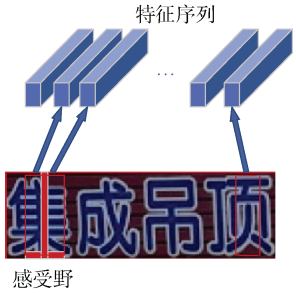


图 11 特征图向量序列及其感受野

Figure 11 Feature map vector sequence and its receptive field

2.4.1 长短期记忆网络

在特征图序列化后, 我们使用双向的长短期记忆网络(Long Short-Term Memory, LSTM)对提取的特征序列结合序列的前向信息和后向信息进行预测, 并且设置其隐状态维数为 256。

LSTM 是一种特殊的时间循环神经网络, 是为了解决一般的循环神经网络(Recurrent Neural Network, RNN)存在的长期依赖问题而专门设计出来的, LSTM 通过门控状态来控制序列信息的传输状态, 保留任务需要长时间记忆的序列信息, 丢弃不必要的序列信息, 而不像普通的 RNN 那样仅有一种序列信息叠加的方式, 因此 LSTM 对很多需要“长期记

忆”的任务来说, 尤其好用[32,33]。对于文字识别的任务来说, 当前时刻的输出不仅和此序列之前的特征序列有关, 还与此序列之后的特征序列密切相关, 预测文本图像中的文字, 特别是被噪声污染严重的文字, 不仅需要根据前面的文字来判断, 还需要考虑它后面的文字内容, 这样才能真正做到基于上下文判断。因此我们将两个 LSTM 上下叠加在一起, 如图 12 所示, 一个正向的 LSTM, 利用上文语义序列的信息, 一个逆向的 LSTM, 利用下文语义序列的信息, 这样双向 LSTM 就能够同时利用上下文信息, 预测的序列最终由这两个 LSTM 共同决定, 会比单向的 LSTM 最终的预测更加准确。

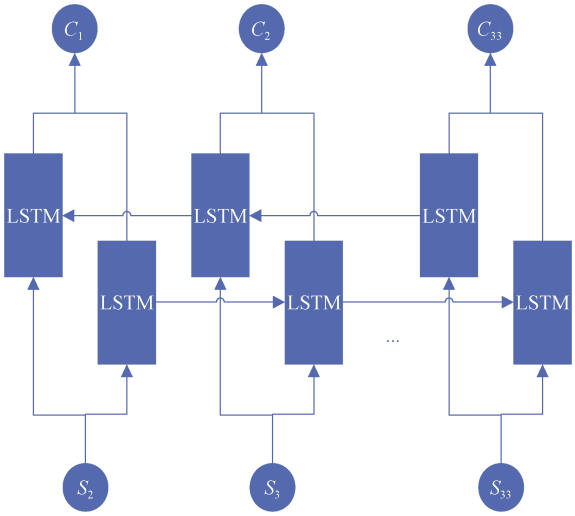


图 12 双向 LSTM 结构图

Figure 12 Bilateral LSTM

输入双向 LSTM 的每一个特征向量都对应文本图像中的一个小矩形区域, 如图 13 所示, 双向 LSTM 的目标就是预测这个小矩形区域为哪个字符, 即根据输入的特征向量 x^{T}, 输出一个所有字符的概率分布 y^{T}, 这是一个长度为字符类别数 n 的向量, 所有的字符概率分布构成后验概率矩阵作为 CTC 层的输入。

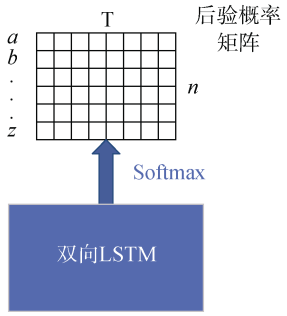


图 13 双向 LSTM 输出的后验概率矩阵

Figure 13 The posterior probability matrix of the bidirectional LSTM output

2.4.2 CTC 转录

CTC 转录是将双向 LSTM 对特征向量所预测的后验概率矩阵转换成对应的标签序列, CTC 算法并不要求输入输出是严格对齐的, 因此它非常适合训练不定长文字的识别。

由于 LSTM 进行时序分类的过程中, 不可避免地会出现字符信息冗余的现象, 即同一个目标字符连续被特征向量标注两次及以上, 因此我们需要一种序列合并机制, 比如我们要识别图 14 的文本, 经双向 LSTM 的预测后有 5 个时间步的特征向量, 理想情况下 t_1、t_2 和 t_3 都被预测为“中”, 而 t_4 和 t_5 都被预测为“国”, 我们将连续重复的字符进行合并, 最终得到的结果为“中国”。这是一种最简单的合并情况, 对于稍微复杂一点的文本, 例如: “好好好好学习”, 用这种方法合并得到的结果是“好学习”, 这显然是不准确的。CTC 算法针对这种字符信息冗余现象引入了一种 blank 机制, 我们以“—”符号代替 blank, C 代表 C 为所有字符类别集合, 经过双向 LSTM 和 Softmax 输出的序列类别为 C’=C∪{—}, 转录时对字符序列先删除连续重复的字符, 再删除所有“—”字符, 例如输出的预测字符序列为“好好—好学习”, 则最后被 CTC 转录为“好好学习”。

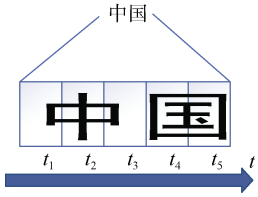


图 14 字符信息冗余示意图

Figure 14 Character information redundancy

CTC 的转录方式是一对多的, 对于不同的后验概率矩阵, 其转录结果可能相同, 例如“好好—好学习”和“好—好好学习”对应的标签都是“好好学习”。基于这种一对多的转录方式, CTC 会尝试所有转录路径, 然后将同一文本标签 l 的每种路径概率相加, 最后选择相加后概率最大的作为其最终标签。在训练时, 对于给定双向 LSTM 的输入 x 的情况下, 输出为目标文本标签 l 的概率为

$$p(l\,|\,x)=\sum_{\pi\in B}p(\pi\,|\,x)$$

其中 B 代表转录后是文本 l 的所有路径集合, π 则是其中的一条路径, 其中, 对于任意一条路径 π 有

$$p(\pi\,|\,x)=\prod_{t=1}^{T}y_{\pi_t}^{t}$$

$y_{\pi_t}^{t}$ 为双向 LSTM 的输出, T 是特征序列长度, 下

标 t 表示 π 路径的每一个时刻。训练的目标就是要优化网络参数, 使得目标文本标签对应的 $p(l|x)$ 最大化, 因此 CTC 的损失函数 L_{ctc} 定义如下:

$$L_{\text{ctc}} = -\sum_{(x,l)\in S} \ln p(l|x)$$

其中 S 为源域文本图像的训练集, 目标域的文本图像无标注, 不参与 CTC 损失。

当训练好文本识别的模型后, 我们输入 x, 希望能得到对应的目标文本标签 l^*, 即我们希望输出 l^*的条件概率最高:

$$l^* = \arg\max_{l} p(l|x)$$

CTC 一般使用贪婪算法[34]或者 Beam search 算法[35]以寻找概率最大路径。本模型在寻找概率最大路径时使用贪婪算法, 即认为每一个时刻最大概率的类别组成的路径即为最优路径:

$$\beta(x) \approx \gamma(\pi^{'})$$

其中, $\pi^{'} = \arg\max_{\pi} p(\pi|x)$

2.5 网络总损失

因使用 CRNN 作为基本文字识别框架结构, 所以本文网络的总损失包括 CTC 损失和域适应部分的损失, 将总损失表示为 L:

$$L = L_{\text{ctc}} + L_{\text{global}} + L_{\text{local}}$$

其中, L_{ctc} 代表源域文本图像的 CTC 损失; L_{global} 代表源域和目标域之间的高层维度的全局域判别损失; L_{local} 代表对齐源域和目标域之间的低层维度的局部域判别损失。

3 实验与分析

3.1 数据集

本文采用合成数据集方法 SynthText[36]构建中文数据集进行模型预训练。采用标准中文数据集 RCTW-17 和 CASIA-10K 构建域适应微调训练阶段的源域和目标域数据集, 其中从 RCTW-17 和 CASIA- 10K 的训练集随机抽取 20%作为源域, 为了使训练的模型适用于电磁泄漏还原的真实场景, 从电磁泄漏还原的真实场景下采集 10%的 RCTW-17 和 CASIA-10K 训练集作为目标域。验证时, 从电磁泄漏还原的真实场景下采集 RCTW-17 和 CASIA-10K 测试集的 10%作为本实验的验证集。

3.1.1 标准数据集

RCTW-17[37]是阅读中文文字图像的比赛 ICDAR2017 Competition on Reading Chinese Text in the Wild 的中文自然场景文本数据集, 包含了建筑、标志牌、条幅、商场、墙壁等带有文字的图像和手机上带有文字的图像截图等。该数据集共 12263 张图像, 其中 8034 张作为训练集, 4229 张作为测试集, 使用四边形框标注文本行, 绝大多数字体为楷书, 极少数艺术字, 几乎没有手写字体, 绝大多数背景和文字非常清晰。在评估阶段, 需要进行行级预测, 其样本如图 15 所示。

图 15 RCTW-17 样本图像
Figure 15 Samples from RCTW-17

CASIA-10K[38]是中国科学院自动化所 PAL 团队在《Multi-Oriented and Multi-Lingual Scene Text Detection With Direct Regression》中提出的中文自然场景文本数据集。该数据集包含各种场景下的 10000 张图像, 其中 7000 张图像用于训练, 而 3000 张图像用于测试。对于每个文本行, 标注了四边形的 8 个坐标。在评估阶段, 需要行级预测, 其样本如图 16 所示。

图 16 CASIA-10K 样本图像
Figure 16 Samples from CASIA-10K

3.1.2 电磁泄漏实景数据

在电磁泄漏还原的真实场景中, 由于设备中

信号放大原件和信号探测设备的原子热运动和物体热辐射以及信号在电磁信道的传输, 我们通过接收设备接收并还原的图像包含大量的随机噪声, 其中虽然高斯噪声和瑞利噪声占绝大部分, 但其他的随机噪声我们也不能完全忽略, 因为这个原因我们人工合成的低信噪比图像很难去拟合在真实电磁泄漏环境中还原的低信噪比图像, 用人工合成的低信噪比图像去训练的文字识别模型对于真实电磁泄漏环境的低信噪比图像的泛化效果并不能够达到很好的预期。

为了使训练的模型在真实的电磁泄漏环境的文本图像上的泛化效果更好, 我们从真实的场景下采集 RCTW-17 和 CASIA-10K 训练集的 10%作为目标域。在域适应微调训练阶段采用半监督训练方式, 即仅源域的文本图像需要标注而带噪声的目标域文本图像无需标注, 在电磁泄漏还原的真实场景中, 我们仅需要不断采集 RCTW-17 和 CASIA-10k 测试集的文本图像即可, 无需依次再按照标记文件进行标注。为了最大限度使训练的模型适用于电磁泄漏还原的真实场景, 在采集文本图像时, 我们将电磁泄漏信号接收天线依次放置在 0.5 m、1 m、2 m、4 m 的位置处, 每处分别按均分比例采集训练集文本图像和测试集文本图像作为目标域训练集和测试集, 不同距离的样本图像如图 17 所示, 从上到下的图像采集距离依次增大。

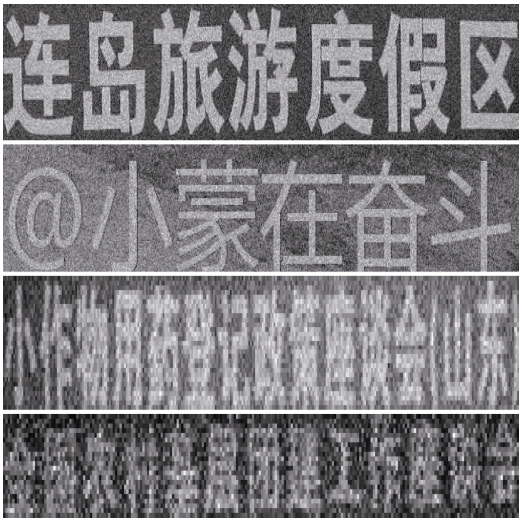


图 17　电磁泄漏还原的样本图像

Figure 17　Sample images emitted from target computer

3.2　实验验证

3.2.1　模型训练细节

基于域适应 CRNN 模型是基于 Pytorch1.1 框架使用 Python 语言实现的, 实验环境如下: 操作系统为 Ubuntu 16.04.5 LTS, 操作环境为 Docker, CPU 型为 Intel(R) Core(TM) i7-7820X CPU @ 3.60GHz, GPU 型号为 GeForce RTX 2080 Ti, 内存大小为 32GB。训练时, 模型首先在用 SynthText 合成的 80 万张中文文本图像上进行预训练, 待模型收敛后再进行域适应微调训练直到再次收敛。收敛判断标准为精确率在连续 5 个训练周期内增长不超过 0.01%。预训练阶段共耗时约 44 小时, 适应微调训练共耗时 6.3 小时, 为了加速训练, 训练过程中所有样本均缩放到 32×128, 反向传播过程使用 Adam 优化器[39], 学习率设置为 0.001。模型中的所有层选择“Xavier”初始化[40], 批大小为 128。

3.2.2　模型训练结果

图 18 为域适应 CRNN 训练阶段的 CTC 损失曲线, 由图可知大概在 20000 次迭代后模型就已收敛。域适应部分的 loss: 全局域适应损失 dloss_t 和局部域适应损失 dloss_t_p 的最终目的是区分不出样本来自哪个域, 即分对分错的可能性理论上各占 50%, 所以稳定时 loss 曲线在一个恒定范围内上下波动。如图 19 和图 20 所示, 可以看出 1000 次以后幅值趋于稳定, 然后在 0.1 到 0.6 范围内稳定地波动, 证明对抗训练已经收敛。

实验结果使用精确率和编辑距离对模型进行评估。

定义 4. 精确率。指模型正确判别的样本在全部测试样本中所占的比例。

定义 5. 编辑距离。对于字符串 s1, s2, 他们的编辑距离是将 s1 转换成 s2 所需的最小编辑操作数。编辑操作包括插入字符, 删除字符和替换字符。

实验时将本文模型与其他具有代表性的主流文字识别模型进行横向对比, 包括 Bai[18]、Yin[21]、Liu[22]

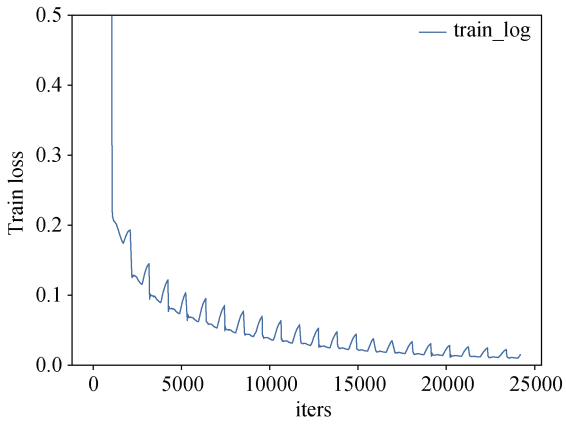


图 18　CTC 训练损失曲线

Figure 18　CTC training loss

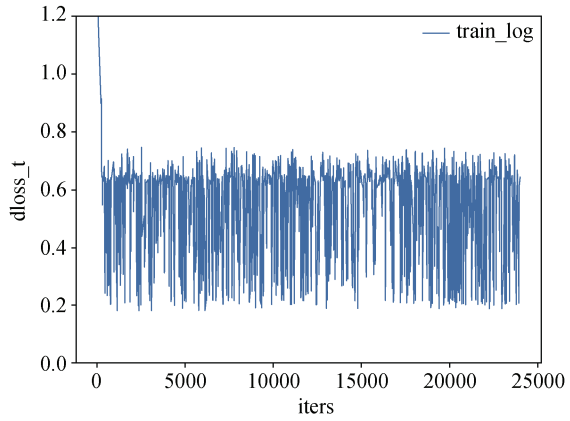


图 19　全局域判别训练损失曲线

Figure 19　Global domain adaptation training loss

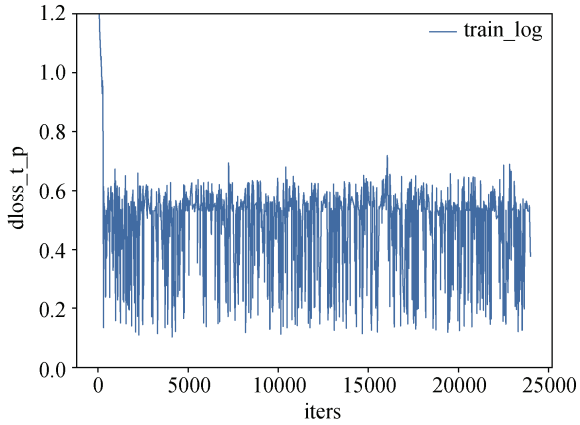


图 20　局部域判别训练损失曲线

Figure 20　Local domain adaptation training loss

和 Qiao[23]的模型。作为对照组的 Bai 、Yin、Liu 和 Qiao 的模型均采用与本模型一致的训练数据，评价指标为精确率(Precision)和归一化平均编辑距离(Normalized Average Edit Distance, NAED)。测试时，使用从电磁泄漏还原的真实场景下采集的 RCTW-17 和 CASIA-10k 的测试集。表 5 和表 6 展示了不同模型测试结果的精确率和归一化平均编辑距离。

由表 5 和表 6 可以看出，对于两种实际电磁泄漏场景下的图像集，相比于 Bai、Yin、Liu 和 Qiao 的模型，基于域适应的 CRNN 在 Precision/NAED 方面分别提升了 14.5%/0.13 和 13.0%/0.14、11.4%/0.07 和 9.7%/0.10、9.7%/0.10 和 7.8%/0.08、3.9%/0.02 和 4.0%/0.02。由于 Bai、Yin 和 Liu 这三者的文字识别模型没有特殊的特征强化或者特征抑制结构，因此并不能在文本图像被污染的情况下有效地学习到文字特征，这就导致了 Bai、Yin 和 Liu 的模型性能较差。而 Qiao 的模型加入了可以学习语义信息的语义模型，使用了可以有效强化文字特征的语义增强编解码框架，因此其对于被噪声污染的文本图像的识别效果较好。基于域适应的 CRNN 在低层次的局部特征和高层次的全局特征上都进行了域对齐，使得整个特征提取模块可以最大限度地提取只与文字相关的公共特征而尽量弱化噪声等非公共特征，因此基于域适应的 CRNN 效果最好。

表 5　不同模型在 CASIA-10k 数据集上的结果对比

Table 5　Different models' results comparison on CASIA-10k dataset

	Precision(%)/NAED
Bai et al. [18]	28.5/0.54
Yin et al. [21]	31.6/0.48
Liu et al. [22]	33.3/0.51
Qiao et al. [23]	39.1/0.43
Ours	43.0/0.41

表 6　不同模型在 RCTW-17 数据集上的结果对比

Table 6　Different models' results comparison on RCTW-17 dataset

	Precision(%)/NAED
Bai et al. [18]	31.1/0.53
Yin et al. [21]	34.4/0.49
Liu et al. [22]	36.3/0.47
Qiao et al. [23]	40.1/0.41
Ours	44.1/0.39

为了进一步分析基于域适应的 CRNN 各结构的有效性，分析域适应各结构对文字识别精度和归一化编辑距离的影响，我们分别进行了独立的对比实验。域适应 CRNN 模型存在四种情况，如表 7 和表 8 所示，其中 L 代表使用局部域判别模块，G 代表使用全局域判别模块，F 代表使用 Focal Loss 函数作为全局域判别模块的损失函数。仅使用局部域判别模块的情况下，在 CASIA-10k 和 RCTW-17 上的 Precision/NAED 分别为 41.7%/0.44 和 42.5%/0.42；仅使用全局域判别模块且不使用 Focal Loss 函数时，在 CASIA-10k 和 RCTW-17 上的 Precision/NAED 分别为 39.2%/0.51 和 39.9%/0.48；仅使用全局域判别模块且使用 Focal Loss 函数时，在 CASIA-10k 和 RCTW-17 上的 Precision/NAED 分别为 40.9%/0.46 和 41.4%/0.44；当局部域判别模块和全局域判别模块结合使用且 Focal Loss 函数作为全局域判别模块的损失函数时，在 CASIA-10k 和 RCTW-17 上的 Precision/NAED 分别为 43.0%/0.41 和 44.1%/0.39。上述实验数据表明：对于源域文本图像与带噪声的目标域文本图像，它们的低层次的局部级特征差异是最主要的，对齐它们之间的局部特征对于文字识别模型性

表 7 不同域适应结构在 CASIA-10k 数据集上的结果对比

Table 7 Different domain adaptation structures' results comparison on CASIA-10k dataset

Method	L	G	F	Precision(%)/NAED
基于域适应的 CRNN	√			41.7/0.44
		√		39.2/0.51
		√	√	40.9/0.46
	√	√	√	43.0/0.41

表 8 不同域适应结构在 RCTW-17 数据集上的结果对比

Table 8 Different domain adaptation structures' results comparison on RCTW-17 dataset

Method	L	G	F	Precision(%)/NAED
基于域适应的 CRNN	√			42.5/0.42
		√		39.9/0.48
		√	√	41.4/0.44
	√	√	√	44.1/0.39

能的提升更大; 使用 Focal Loss 函数作为全局域判别模块的损失函数时, 可以使全局域判别模块的损失集中在整体特征相似的源域和目标域数据, 从而降低对齐难度, 能有效提升模型性能; 当局部域判别模块和全局域判别模块共同作用时, 效果最佳。

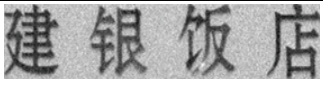
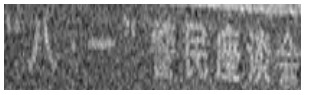
为了更直观地进行模型对比, 本文用表 9 展示了基于域适应的 CRNN 模型与其他 4 种被测模型针对电磁泄漏还原得到的文本图像的识别情况。由表中的识别结果可见, 对于前几张被噪声污染较小且文本字数较少的图像, 这 5 种模型基本都能做到正确地识别, 然而对于后面两张被噪声污染较大且文本字数较多的图像, 本文模型的文字识别正确率相较于前四种模型要更高。对于文本图像中的标点符号, 五种模型都存在漏检的现象, 这是因为标点的尺寸较小, 受噪声的影响更大, 难以有效地提取它的特征。

4 结论

通过电磁泄漏还原技术得到的图像中, 其文字往往含有十分重要的信息, 但是这种通过电磁泄漏方式得到的图像往往伴有不同程度的随机噪声, 主流的文本识别模型难以对其进行有效的识别。本文针对电磁泄漏还原图像的特点提出了一种基于域适应的 CRNN 文字识别模型。该模型能够在不进行常规去噪等预处理的情况下直接对带噪中文文本图像实现无分割识别。在电磁泄漏还原下环境下的公开数据集 RCTW-17 和 CASIA-10K 上的测试结果表明, 相比于其他主流识别模型, 基于域适应的 CRNN 在电磁泄漏还原图像中的中文识别率有了明显的提升。

表 9 不同模型对电磁泄漏还原图像的识别结果对比

Table 9 Comparison of recognition results between different methods

电磁泄漏还原图像	Bai et al. [18]	Yin et al. [21]	Liu et al. [22]	Qiao et al. [23]	Ours
 GT: 建银饭店	建银饭店	建银饭店	建银饭店	建银饭店	建银饭店
 GT: @小蒙在奋斗	@小**菜**在奋斗	@小蒙在奋斗	@小蒙在**面**斗	@小蒙在奋斗	@小蒙在奋斗
 GT: 洋地村 13.5km	洋地村 35km	洋地村 135km	洋地村 135km	洋地村 135km	洋地村 13.5km
 GT:“八·一”警民座谈会	**人**一**薰**民座**微**会	八一**擎**民**圆姿**会	**人**一**罄**民**庭爽**会	**人**一警民座**淡**会	八一”**制**民座谈会
 GT:全国农村基层团建工作座谈会	全**团**农村**瓦里困**工作座**微**会	**笙困**农村基层团**延**工作**参睿**会	**金园**农村基层团**便**工作座**说**会	全国农村基层团**键**工作座**议**会	全国农村基层团建工作座谈会

在下一步的研究工作中, 我们将研究如何将全局域判别和局部域判别得到的信息有效地利用起来, 从而更有效地提取源域和目标域的公共字符特征, 进一步提升模型的性能。

参考文献

[1] Highland H J. Electromagnetic eavesdropping machines for christmas?[J]. *Computers & Security*, 1988, 7(4): 341-341.

[2] Van Eck W. Electromagnetic radiation from video display units: An eavesdropping risk?[J]. *Computers & Security*, 1985, 4(4): 269-286.

[3] Elibol F, Sarac U, Erer I. Realistic eavesdropping attacks on computer displays with low-cost and mobile receiver system[C]. *IEEE 20th European Signal Processing Conference*, 2012: 1767-1771.

[4] TOSAKA T, YAMANAKA Y, FUKUNAGA K, et al. Method for Determining Whether or Not Information is Contained in Electromagnetic Disturbance Radiated From a PC Display[J]. *IEEE Transactions on Electromagnetic Compatibility*, 2011, 53(2): 318-324.

[5] Highland HaroldJoseph. Tempest over leaking computers[J]. *Computers & Securit*y, 1987, 6(6):457-458.

[6] Smulders P. The threat of information theft by reception of electromagnetic radiation from RS-232 cables[J]. *Computers & Security*, 1990, 9(1): 53-58.

[7] Kuhn M G, Anderson R J. Soft tempest: Hidden data transmission using electromagnetic emanations[C]. *Proceedings of the 4th International Workshop on Information Hiding*, 1998: 124-142.

[8] Kuhn M G. Optical time-domain eavesdropping risks of CRT displays[C]. *Proceedings 2002 IEEE Symposium on Security and Privacy*, 2002: 3-18.

[9] Sekiguchi H , Seto S . Estimation of receivable distance for radiated disturbance containing information signal from information technology equipment[C]. *IEEE International Symposium on Electromagnetic Compatibility*, 2011: 942-945.

[10] Yokobayashi M, Wakahara T. Segmentation and recognition of characters in scene images using selective binarization in color space and GAT correlation[C]. *International Conference on Document Analysis and Recognition*, 2005: 167-171.

[11] Campos T E, Babu B R, Varma M. Character recognition in natural images[C]. *VISAPP*, 2009: 05-08.

[12] Weinman J, LearnedMiller E, Hanson A R. Scene text recognition using similarity and a lexicon with sparse belief propagation[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, 31(10): 1733-1746.

[13] Mishra A, Alahari K, Jawahar C. Top-down and bottom-up cues for scene text recognition[C]. *IEEE Conference on Computer Vision and Pattern Recognition*, 2012: 2687-2694.

[14] Novikova T, Barinova O, Kohli P, et al. Large-lexicon attribute-consistent text recognition in natural images[C]. *European Conference on Computer Vision*, 2012: 752-765.

[15] Newell A J, Griffin L D. Multiscale Histogram of Oriented Gradient Descriptors for Robust Character Recognition[C]. *Proceedings of the International Conference on Document Analysis and Recognition*, 2011: 1085-1089.

[16] Wang T, Wu D, Coates A, et al. End-to-end text recognition with convolutional neural networks[C]. *Proceedings of the International Conference on Pattern Recognition*, 2012: 3304-3308.

[17] Jaderberg M, Vedaldi A, Zisserman A. Deep features for text spotting[C]. *Proceedings of the 13th European Conference on Computer Vision*, 2014: 512-528.

[18] Shi B, Bai X, Yao C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition[J]. *IEEE Transactions on Pattern Analysis and Machine intelligence*, 2017, 39(11): 2298-2304.

[19] Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition[J]. *IEEE Signal Processing Magazine*, 2012, 29(6): 2012.

[20] Krizhevsky A , Sutskever I , Hinton G . ImageNet Classification with Deep Convolutional Neural Networks[C]. *Proceedings of the 25th International Conference on Neural Information Processing Systems*, 2012, 1097-1105.

[21] Yin F, Wu Y C, Zhang XY, et al. Scene text recognition with sliding convolutional character models[EB/OL]. 2017: ArXiv Preprint ArXiv: 1709.01727.

[22] Liu W , Chen C , Wong K Y , et al. STAR-Net: A SpaTial Attention Residue Network for Scene Text Recognition[C]. *British Machine Vision Conference*, 2016: 19-22.

[23] Qiao Z , Zhou Y , Yang D , et al. SEED: Semantics Enhanced Encoder-Decoder Framework for Scene Text Recognition[EB/OL]. 2020: ArXiv Preprint ArXiv:1109.01354.

[24] Zhang K , Zuo W , Chen Y , et al. Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising[J]. *IEEE Transactions on Image Processing*, 2016, 26(7):3142-3155.

[25] Tao L , Zhu C , Xiang G , et al. Llcnn: A convolutional neural network for low-light image enhancement[C]. *IEEE Visual Communications and Image Processing*, 2017: 1-4.

[26] Guo Z , Sun Y , Jian M , et al. Deep Residual Network with Sparse Feedback for Image Restoration[J]. *Applied Sciences*, 2018, 8(12).

[27] Shi Honglei. Research on Domain Adaptation Algorithm Based on Single Source and Multi-source[D]. East China Normal University, 2015.
(时红垒. 基于单源及多源的域适应算法研究[D]. 华东师范大学, 2015.)

[28] Zeiler M D, Fergus R. Visualizing and Understanding Convolutional Networks[C]. *European Conference on Computer Vision*, 2014: 818-833.

[29] Lin T Y, Goyal P, Girshick R, et al. Focal Loss for Dense Object Detection[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2017, 42(2): 2999-3007.

[30] Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult[J]. *IEEE Transactions on Neural Networks*, 1994, 5(2):157-166.

[31] Graves A, Mohamed A R, Hinton G. Speech recognition with deep recurrent neural networks[C]. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013: 6645-6649.

[32] A. G F, Schmidhuber J, Cummins F. Learning to Forget: Continual Prediction with LSTM[M]. Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale, 1999: 73.

[33] Hochreiter S, Schmidhuber J. Long short-term memory[J]. *Neural Computation*, 1997, 9(8): 1735-1780.

[34] Graves A, Fernández S, Gomez F, et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks[C]. *Proceedings of the 23rd International Confe-*

rence on Machine Learning, 2006: 369-376.

[35] Markus Freitag, Yaser AlOnaizan. Beam searchstrategies for neural machine translation[EB/OL]. 2017: ArXiv Preprint ArXiv: 1702.01806.

[36] Gupta A, Vedaldi A, Zisserman A. Synthetic Data for Text Localisation in Natural Images[C]. *IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 2315-2324.

[37] Shi B, Yao C, Liao M, et al. ICDAR2017 competition on reading chinese text in the wild[C]. *Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition*, 2017: 1429-1434.

[38] He W, Zhang X Y, Yin F, et al. Multi-oriented and multi-lingual scene text detection with direct regression[J]. *IEEE Transactions on Image Processing*, 2018, 27(11): 5406-5419.

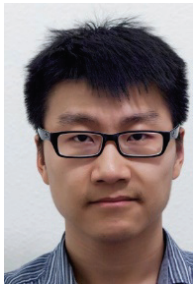
[39] Kingma D P, Ba J. Adam: A method for stochastic optimization[EB/OL]. 2014: ArXiv Preprint ArXiv:1412.6980.

[40] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks[C]. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010: 249-256.

吕志强 于 2007 年在哈尔滨工业大学微电子学与固体电子学专业获得博士学位。现任中国科学院信息工程研究所副研究员。研究领域为信号处理及系统实现。研究兴趣包括: 高噪声图像处理与文本识别。Email: lvzhiqiang@iie.ac.cn

于超 2019 年在中国民航大学电子信息工程专业获得学士学位。现在中国科学院大学网络空间安全专业攻读硕士学位。研究领域为网络空间安全。研究兴趣包括：图像处理、文本识别。Email：yuchao@iie.ac.cn

李海洋 于 2012 年在中央民族大学通信工程专业获得学士学位。现在中国科学院大学网络空间安全学院通信与信息系统专业攻读硕士学位。研究领域为嵌入式硬件安全、电磁信息安全。研究兴趣包括: 物理安全检测、电磁信息安全。Email: lihaiyang@iie.ac.cn



张宁 于 2013 年在哥伦比亚大学电气工程专业获得硕士学位。现任中国科学院信息工程研究所工程师。研究领域为网络空间安全。研究兴趣包括硬件安全等。Email: zhangning@iie.ac.cn