软件学报 ISSN 1000-9825, CODEN RUXUEW
2025,36(9):4187−4206 [doi: 10.13328/j.cnki.jos.007264] [CSTR: 32375.14.jos.007264]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于掩码信息熵迁移的场景文本检测知识蒸馏*

陈建炜[1], 沈英龙[1], 杨 帆[1,2], 赖永炫[3]

[1](厦门大学 自动化系, 福建 厦门 361005)
[2](厦门大学深圳研究院, 广东 深圳 518057)
[3](厦门大学 软件工程系, 福建 厦门 361005)
通信作者: 杨帆, E-mail: yang@xmu.edu.cn

摘 要: 自然场景文本检测的主流方法大多使用复杂且层数较多的网络来提升检测精度, 需要较大的计算量和存储空间, 难以部署到计算资源有限的嵌入式设备上. 知识蒸馏可通过引入与教师网络相关的软目标信息, 辅助训练轻量级的学生网络, 实现模型压缩. 然而, 现有的知识蒸馏方法主要为图像分类任务而设计, 提取教师网络输出的软化概率分布作为知识, 其携带的信息量与类别数目高度相关, 当应用于文本检测的二分类任务时会存在信息量不足的问题. 为此, 针对场景文本检测问题, 定义一种新的信息熵知识, 并以此为基础提出基于掩码信息熵迁移的知识蒸馏方法 (mask entropy transfer, MaskET). MaskET 在传统蒸馏方法的基础上引入信息熵知识, 以增加迁移到学生网络的信息量; 同时, 为了消除图像中背景信息的干扰, MaskET 通过添加掩码的方法, 仅提取文本区域的信息熵知识. 在 ICDAR 2013、ICDAR 2015、TD500、TD-TR、Total-Text 和 CASIA-10K 这 6 个公开标准数据集上的实验表明, MaskET 方法优于基线模型和其他知识蒸馏方法. 例如, MaskET 在 CASIA-10K 数据集上将基于 MobileNetV3 的 DBNet 的 $F1$ 得分从 65.3% 提高到 67.2%.

关键词: 自然场景; 文本检测; 知识蒸馏; 信息熵
中图法分类号: TP18

中文引用格式: 陈建炜, 沈英龙, 杨帆, 赖永炫. 基于掩码信息熵迁移的场景文本检测知识蒸馏. 软件学报, 2025, 36(9): 4187–4206. http://www.jos.org.cn/1000-9825/7264.htm

英文引用格式: Chen JW, Shen YL, Yang F, Lai YX. Knowledge Distillation for Scene Text Detection via Mask Information Entropy Transfer. Ruan Jian Xue Bao/Journal of Software, 2025, 36(9): 4187–4206 (in Chinese). http://www.jos.org.cn/1000-9825/7264.htm

Knowledge Distillation for Scene Text Detection via Mask Information Entropy Transfer

CHEN Jian-Wei[1], SHEN Ying-Long[1], YANG Fan[1,2], LAI Yong-Xuan[3]

[1](Department of Automation, Xiamen University, Xiamen 361005, China)
[2](Shenzhen Research Institute of Xiamen University, Shenzhen 518057, China)
[3](Department of Software Engineering, Xiamen University, Xiamen 361005, China)

Abstract: Mainstream methods for scene text detection often use complex networks with plenty of layers to improve detection accuracy, which requires high computational costs and large storage space, thus making them difficult to deploy on embedded devices with limited computing resources. Knowledge distillation assists in training lightweight student networks by introducing soft target information related to teacher networks, thus achieving model compression. However, existing knowledge distillation methods are mostly designed for image classification and extract the soft probability distributions from teacher networks as knowledge. The amount of information carried by such methods is highly correlated with the number of categories, resulting in insufficient information when directly applied to the binary classification task in text detection. To address the problem of scene text detection, this study introduces a novel concept of information

* 基金项目: 国家自然科学基金面上项目 (62173282); 厦门市自然科学基金面上项目 (3502Z20227180)
收稿时间: 2023-03-28; 修改时间: 2024-01-11, 2024-07-18; 采用时间: 2024-08-02; jos 在线出版时间: 2025-01-15
CNKI 网络首发时间: 2025-01-15

entropy and proposes a knowledge distillation method based on mask entropy transfer (MaskET). MaskET combines information entropy with traditional knowledge distillation methods to increase the amount of information transferred to student networks. Moreover, to eliminate the interference of background information in images, MaskET only extracts the knowledge within the text area by adding mask operations. Experiments conducted on six public benchmark datasets, namely ICDAR 2013, ICDAR 2015, TD500, TD-TR, Total-Text and CASIA-10K, show that MaskET outperforms the baseline model and other knowledge distillation methods. For example, MaskET improves the *F*1 score of MobileNetV3-based DBNet from 65.3% to 67.2% on the CASIA-10K dataset.

Key words: natural scene; text detection; knowledge distillation (KD); information entropy

在自然环境中, 场景文本无处不在, 常见于高速公路上的路标、街边的广告牌以及各种产品的包装信息中. 获取自然场景图像中的文本信息有助于自动导航和定位、智能安防、图像检索等领域的研究.

随着深度全卷积网络[1]在图像分割方面取得重大进展[2], 图像分割已成为主流文本检测方法的基础框架. 全卷积网络能够输出逐像素分类结果, 因此更有利于检测出弯曲、多方向等复杂的场景文本. 然而, 该类方法以庞大的模型参数来换取检测精度的提升, 例如在 ICDAR 2015 数据集上排名前列的 TextFuseNet[3], 其参数量超过百万, 存储内存更是高达 820 MB, 难以应用于计算资源有限的场景. 解决这个问题的一个有效方法便是知识蒸馏 (knowledge distillation, KD)[4].

知识蒸馏通过将预先训练好的大型网络 (教师网络) 的知识转移到小型网络 (学生网络), 从而提高学生网络的性能表现. 教师网络只在训练学生网络时使用, 最终实际部署只需满足学生网络的计算资源要求, 经过蒸馏训练过后的学生网络参数量远小于教师网络, 精度也比原先没有蒸馏训练的学生网络高, 在性能和复杂度上达到了较好的平衡[5].

目前, 大多数知识蒸馏方法是为图像分类设计, 而专门针对文本检测模型开发的知识蒸馏方法较少. 由于基于分割网络的文本检测模型的像素级分类与图像分类有相同之处, 目前业界普遍的做法[6,7]是将用于分类的蒸馏方法直接套用到基于分割的检测模型上. 然而本文发现, 当图像分类的蒸馏方法[4,6,8]直接应用在文本检测任务时, 并不总能表现出理想的效果, 甚至有时会使得学生网络精度下降. 基于分类的知识蒸馏方法一般以迁移教师网络输出的软目标 (软化概率分布) 知识为主, 对于文本检测这种像素级二分类任务而言, 其概率分布包含的泛化信息严重不足, 难以有效指导学生网络的训练.

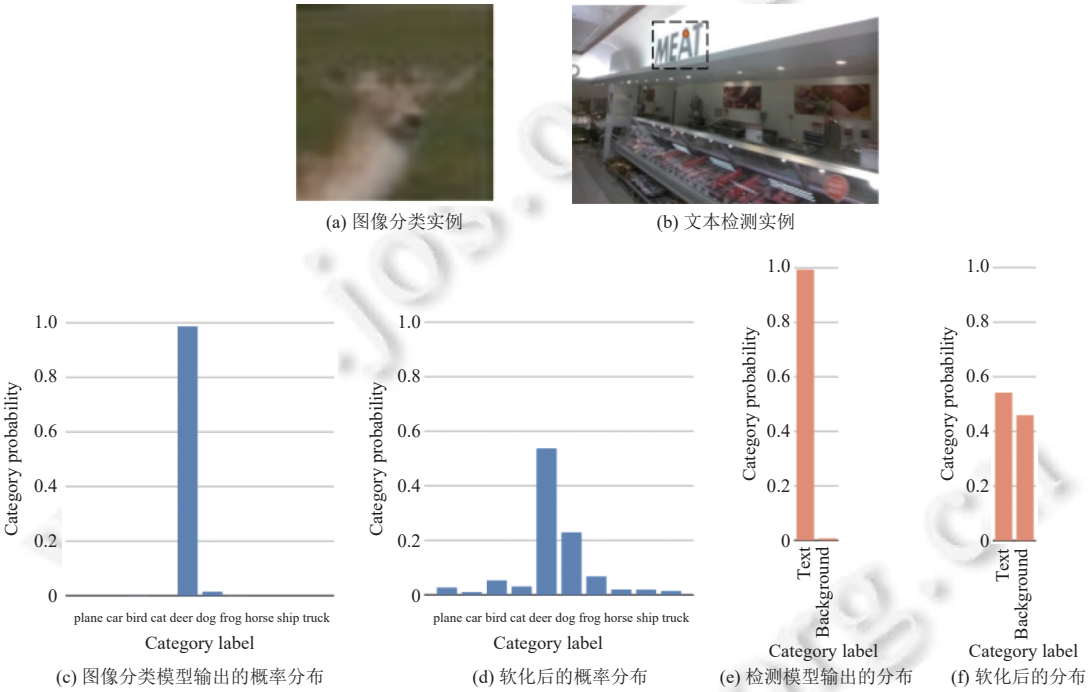
软目标知识的实质是教师网络输出的 logit 经带有温度超参数 T (T 为大于 1 的正数) 的 Softmax 函数“软化”后得到的类概率分布 $\left(p_i = \exp(z_i/T)\sum_j \exp(z_j/T)\right)$, 软化的平滑程度与温度 T 有关, T 越大, 分布越平滑. 软化的分布蕴含教师网络丰富的泛化信息, 即类别之间相似性[4]. 学生网络通过接受教师网络软目标的监督, 可以有效学习到不同类别间的相似信息, 从而比仅通过硬目标 (one-hot 标签) 训练的效果好. 然而, 学生网络学到的泛化信息量与类别数有关, 类别数越多, 软化后的概率分布就携带越多的信息, 因此软目标更适合图像多分类问题. 以 CIFAR-10[9] 图像分类问题为例, 训练的模型 (ResNet101[10]) 输出的类概率分布和软化后 (温度 T=5) 的分布如图 1 所示, 在图 1(d) 软目标分布中, 鹿的概率最高, 其次是狗, 并且误判成狗的可能性远大于车等其他 8 个类别, 说明狗比车更像鹿, 而车与鹿有较大差距. 实际上, 图 1(a) 中的鹿和狗的确有相似之处, 而默认温度 T 的分布 (图 1(c)) 中, 模型分配给错误类别的概率接近 0, 难以直接反映出类别间的关联性.

自然场景文本检测任务可以看作是对图像上的每一个像素点做非文本即背景的二分类问题, 对于类别数较少的任务, 软化后的分布所能得到的类间相似信息便会减少, 甚至无法获得有效信息. 在图 1(b) 文本区域 (图中虚线矩形框表示) 中, 圆点所示的像素点的概率分布经过温度 T 的软化操作后, 其文本概率从接近 1 (图 1(e)) 降低到靠近 0.5 (图 1(f)), 相当于原本属于文本区域的点逐渐变成边缘点 (概率值 0.5 的像素点). 在图 1(b) 中, 圆点标注的像素点属于文本, 与虚线矩形框的边缘点相距较远, 两者之间并没有明显的相似性, 由此可见, 对于文本检测问题, 软目标提炼的泛化信息量不如图像分类问题.

在知识蒸馏中, 教师网络的知识对学生网络的学习起着至关重要的作用. 对于文本检测任务而言, 学生网络未能从教师网络的软目标知识中获得足够多的泛化信息, 因此在后续实验中, 和不使用蒸馏训练的 baseline 网络相比较, 软目标蒸馏方法提升有限. 软目标的泛化信息量不足, 自然而然的想法是加入更多的知识, 以提供更多有用

的信息来指导学生网络的训练, 例如 FitNets[8]在软目标知识的基础上, 额外增加教师网络中间特征图知识.

为此, 本文提出了文本检测模型的信息熵知识, 其灵感来源于对文本检测模型输出的分割图的混乱程度与其泛化能力强弱之间的分析. 本文在第 3 节通过可视化分析, 说明信息熵知识具有放大检测模型对文本边缘关注的现实含义. 学生网络相较于教师网络泛化能力差, 主要表现在学生网络预测的文本边缘不如教师网络完整, 而教师网络的信息熵知识反映了模型对文本边缘的关注, 将其作学生网络额外的“文本边缘”的监督, 促使学生网络预测的文本框更加完整地包围文本边缘, 从而实现了提升模型泛化性能的目的. 在第 4 节, 本文将信息熵知识作为教师网络中的迁移对象, 提出了基于掩码信息熵迁移的知识蒸馏方法 (mask entropy transfer, MaskET), 充分利用文本检测分割图的信息熵知识, 促进学生网络学习教师网络提炼的文本边缘信息. 为了避免背景噪声的干扰, 本文使用掩码 (mask) 操作提取出文本区域的信息熵, 有效提高了学生网络在复杂背景下正确检测的能力. 在 6 个自然场景文本检测数据集上的实验表明, MaskET 能显著提升基线文本检测模型的召回率和 $F1$ 得分, 且优于其他蒸馏方法. 例如, MaskET 在 ICDAR 2015[11] 数据集上将基于 MobileNetV3 的 DBNet 的 $F1$ 得分从 78.7% 提高到 79.8%.



(a) 图像分类实例 (b) 文本检测实例

(c) 图像分类模型输出的概率分布 (d) 软化后的概率分布 (e) 检测模型输出的分布 (f) 软化后的分布

图 1 图像分类与文本检测软目标对比

1 相关工作

1.1 基于深度学习的场景文本检测

基于深度学习的文本检测方法与计算机视觉技术发展息息相关, 随着目标检测、实例分割和图像分割等技术取得重大突破, 文本检测的研究思路得到了扩展. Liao 等人[12]提出可微分阈值算法, 自适应预测每一个像素点的二值化阈值, 简化分割网络的后处理步骤. Lyu 等人[13]修改实例分割网络 Mask R-CNN[14]的 mask 分支, 以支持文本实例分割和字符级分割. Zhou 等人[15]使用单个卷积网络直接预测图像文本行的方向、角度、四边形坐标以及分割图, 消除中间步骤, 极大提高检测速度. Tian 等人[16]提出使用固定宽度的预选框替换原 Faster R-CNN[17]中不同尺度的预选框, 通过将文本行切割成密集细小的预选框, 提高小文本的检测精度. Liao 等人[18]修改单阶段目标检测 SSD[19]的卷积核及默认框的尺度比例以适应文本行长条形的特点. Qin 等人[20]提出 Mask R-CNN[14]的掩码分支应该学习预测整个实例的形状, 而不是预测每个像素分类为文本的概率, 从而有效解决了基于实例分割的文本

检测方法难以区分同个建议框内多个文本的问题. Dai 等人[21]在实例分割网络 deep snake[22]生成的分割图上, 根据文本的中心和大小生成初步建议框, 再将建议框的轮廓逐步回归到文本的轮廓, 缓解了以往图像分割的检测方法对像素级噪声敏感的问题. Liao 等人[23]提出可微二值化 (DB) 模块, 该模块将二值化过程集成到分割网络中, 以产生更准确的结果, 同时提出一种高效的自适应尺度融合 (ASF) 模块, 通过自适应融合不同尺度的特征来提高尺度鲁棒性.

基于图像分割或者实例分割的检测方法能进行准确的像素级预测, 更适合检测任意形状的文本, 因而成为基于深度学习的检测方法的主流. 该类文本检测模型大多使用复杂且层数较多的网络来提升检测精度, 需要较大的计算量和存储空间. 然而, 该类模型较少关注如何将其轻量化, 从而更好部署在资源有限的平台上.

1.2 知识蒸馏及其在文本检测的应用

如何将性能显著但参数量过大的深度网络部署在资源有限的设备上是模型压缩的主要研究问题. 知识蒸馏旨在通过迁移教师网络知识来获得一个精度较高而规模较小的学生网络, 是目前模型压缩的主流解决方案.

知识蒸馏的核心思想是构造一种包含教师网络和学生网络的训练框架, 从预先训练好的教师网络中提炼出知识, 将知识作为训练学生网络的额外监督信息来实现知识的迁移. Hinton 等人[4]将软化后的 logit 称为软目标知识, 软目标包含不同类别间的相似性信息. Romero 等人[8]引入卷积神经网络的中间层输出的特征图, 作为提示信息 (hint) 引导学生网络模仿教师网络的特征表达. He 等人[24]提出让学生网络学习经过编码器压缩的教师网络的知识. Liu 等人[25]将教师网络的特征图和通过对抗式学习获取的信息作为迁移到学生网络的结构化知识. SD[26]使用最深层的分类器作为教师网络, 通过辅助网络指导前面每个残差块的学习, 实现无需教师网络的自蒸馏. SAD[27]将相邻层网络的注意力图作为浅层网络的蒸馏训练目标. 为了解决用于目标检测的蒸馏方法不能有效提取教师网络定位知识的问题, Zheng 等人[28]提出了定位蒸馏方法 (localization distillation), 通过有选择性地提取特定区域的语义和定位信息, 将教师网络的定位知识有效地迁移到了学生网络. Zhou 等人[29]提出一种通用的跨模态知识 (UniDistill), 将不同模态的知识用统一的鸟瞰视角特征表示, 以实现自动驾驶中的雷达点云和相机图像的跨模态知识蒸馏. Zagoruyko 等人[30]将教师网络的中间层特征图按通道计算统计量构建二维的注意力图, 再将注意力从教师网络迁移到学生网络上; 其提出的注意力图和 Romero 等人[8]提出的知识类似, 是一种基于中间特征图的知识, 而非网络输出层. 陈建炜等人[31]使用辅助网络将信息熵知识从学生网络的深层往浅层迁移, 实现无需教师网络的自蒸馏; 本文和其不同在于, 提出了掩码操作来提取文本区域的信息熵知识, 且知识来自预先训练好的教师网络, 而非网络自身.

目前, 知识蒸馏的应用以计算机视觉任务为主, 并且绝大多数蒸馏方法是为图像分类而设计, 为文本检测领域设计的知识蒸馏方法较少. 现有的研究[5−7]侧重于将图像分类的蒸馏方法直接迁移到文本检测任务上, 并取得了一定的成效. Yang 等人[32]提出了一种快速文本检测方法, 该检测框架包括一个轻量级的学生网络和一个复杂的教师网络, 借鉴 FitNets[8]在图像分类上的成功经验, 提取教师网络的中间特征图来指导学生网络的训练, 在准确性和运行效率间取得了更好的平衡. 为了解决基于分割的文本检测方法难以分离过于靠近的文本以及后处理耗时的问题, Yang 等人[33]受实例分割网络 YOLACT[34]的启发, 提出了实时检测网络, 并模仿快照蒸馏方法[35], 从先前的迭代模型中提取有用的信息来监督学生网络, 以提高其检测精度. 百度飞桨团队[7]在 PP-OCRv2 的开源 OCR 系统上引入了图像分类领域广泛使用的深度互学习蒸馏方法[36], 其中两个学生网络在训练过程中相互学习对方输出的分割图, 同时增加一个教师网络来指导两个学生网络训练, 以此获得更强大的文本检测模型.

1.3 信息熵在知识蒸馏中的应用

信息熵在知识蒸馏中常应用于原始软目标蒸馏方法的损失函数, 软目标蒸馏方法通过最小化学生网络的类概率分布与教师网络软化后的分布之间的相对熵 (KL 散度) 来迁移教师网络的知识. 相对熵可以衡量两个分布之间的差异, 因此可以用作软目标的损失函数. 除此之外, 信息熵作为系统的不确定性或者混乱程度的度量, 在变分信息蒸馏[37]、无数据蒸馏[38]、多教师知识蒸馏[39]等蒸馏方法中也有所应用. Ahn 等人[37]提出变分信息蒸馏 (variational information distillation, VID), VID 将知识迁移过程看作是教师网络和学生网络中间层知识的互信息最大化的过程, 当教师-学生网络之间的互信息最大时, 可以认为学生网络已经完全学到教师网络的知识. Chen 等人[38]

指出, 类别平衡的数据集的分布近似于均匀分布且具有最大的信息熵值, 并以此为基础提出了一个新的无数据知识蒸馏框架 DAFL (data-free learning), 可通过最大化数据集的信息熵, 来促使生成器能够产生类别平衡的合成数据. Kwon 等人[39]提出自适应知识蒸馏 (adaptive knowledge distillation, AKD), 根据教师网络软目标的熵值大小来决定它的重要性. 其背后原理是信息熵能够衡量教师网络对分类结果的不确定程度, 熵值越大, 其不确定性越高, 分配的权重理应更小.

2 适用于文本检测的信息熵知识

知识蒸馏通过将教师网络 (网络结构复杂的大网络) 的泛化知识迁移到学生网络 (网络层数少的小网络) 来提高学生网络的泛化能力, 最为关键的地方在于如何定义具有教师网络泛化信息的“知识”. 绝大多数知识蒸馏方法是为图像分类设计, 而专门针对文本检测模型设计的知识蒸馏方法较少. 基于分割的检测方法能够对图像上每一个像素点做文本或非文本的二分类, 与图像分类有相通之处, 因此分类的知识蒸馏方法可以直接迁移到文本检测任务上. 然而, 本文发现有些在图像分类上广泛使用的蒸馏方法[4,8]在较大数据集上普遍效果不佳, 其原因是这些蒸馏方法以图像分类的方式提取文本检测模型的知识, 缺乏足够的泛化信息. 针对软目标知识存在泛化信息不足的问题, 本文定义了一种更具有泛化性的信息熵知识, 和教师网络的软目标知识同时使用, 有效提升学生网络的精度.

信息熵知识的灵感来源于教师网络和学生网络输出的分割图 (segmentation map) 的视觉混乱程度的对比, 分割图上每一个像素点的值代表属于文本的概率值. 教师网络泛化能力强, 预测更加准确[40], 因此误判的情况少, 表现在一些模糊的噪声区域 (图 2(a) 和 (b) 中红色虚线标识区域) 中文本的概率接近 0, 而学生网络的分割图在一些噪声区域有较高的概率值, 整体看起来更加混乱. 在知识蒸馏的概念中知识是体现教师网络泛化能力强的信息, 而分割图的混乱程度可以反映预测能力的强弱, 直接与泛化能力相关联. 因此, 可以尝试从分割图混乱程度中挖掘泛化信息[40], 本文将分割图看作是由图上所有像素点组成的一个系统, 每个像素点的概率分布满足文本或非文本的 0-1 分布, 视觉上看到的混乱实质是整个系统的不确定性, 而信息熵可以表示系统的不确定性. 本文进一步从信息熵角度分析输出的分割图, 发现信息熵蕴含了模型对文本边缘的关注信息.

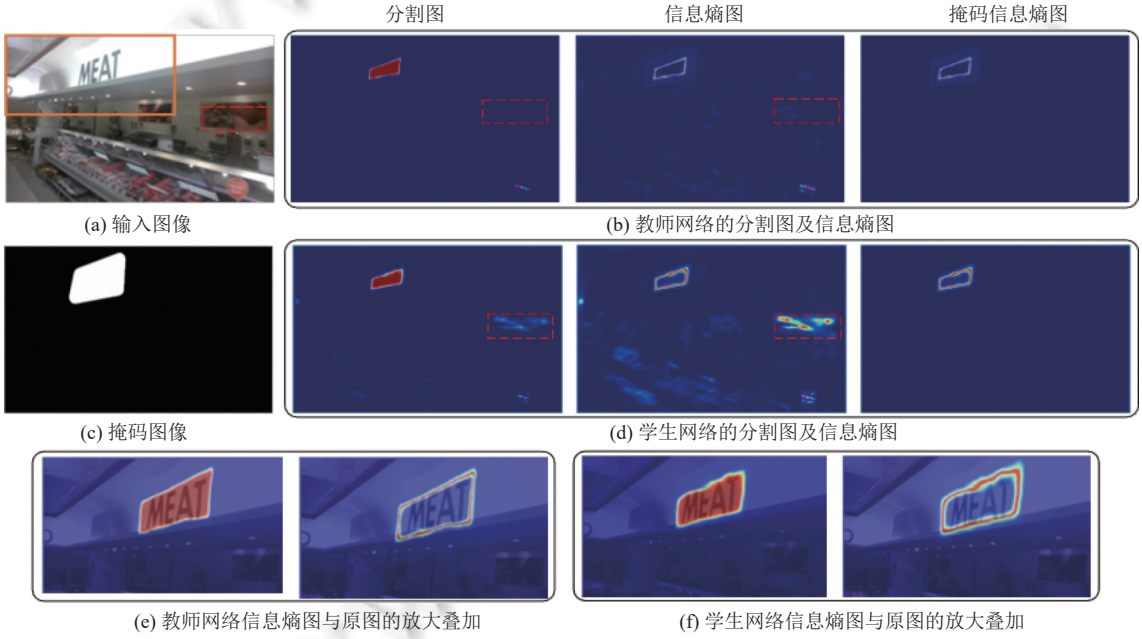


图 2 DBNet 的分割图和信息熵图可视化

为了更加直观分析信息熵知识, 本文首先将 DBNet[12]预测的每一个像素点的概率值转换为信息熵, 再用不同颜色来标识信息熵值的高低, 得到如图 2(b) 和 (d) 所示的信息熵图, 可以看出, 文本边缘的区域呈红色, 而文本中

心区域熵值低 (包裹的蓝色区域). 由于 DBNet 预测的结果是按比例缩小的, 因此本文将文本区域 (图 2(a) 橙色矩形区域) 对应的信息熵图进行放大, 同时将原图叠加到信息熵图上, 以探究信息熵图和输入图像中的文本框的联系. 观察图 2(e) 和 (f) 可发现, 相较于分割图, 信息熵图放大了模型对边缘的注意力, 并且教师网络的信息熵图形成的边缘比学生网络的熵值图不仅更加接近矩形, 而且更完整地包围文本边缘, 如图 2(f) 学生网络的信息熵图对字母“E”和“T”的边缘都有所缺失. 另外, 学生网络和教师网络两者的信息熵图混乱程度不同, 学生网络在一些噪声区域 (图 2(d) 红色虚线矩形标注) 也有高熵值.

教师网络和学生网络两者之间信息熵图的差别体现了他们检测性能和泛化能力上的差距, 本文通过对学生网络施加教师网络的信息熵监督, 促进学生网络学习教师网络关于文本边缘的信息, 进而提升学生网络检测文本边缘的能力. 同时, 教师网络的信息熵图也存在对噪声区域 (非目标文本区域) 的信息熵值, 本文通过一个掩码 (图 2(c)) 的操作提炼文本区域的信息熵, 进一步提高学生网络在复杂易混淆的背景下正确检测的能力, 并称这种知识迁移方式为掩码信息熵迁移 (MaskET).

本文将信息熵迁移用于文本检测模型, 与第 2.3 节中提到的变分信息蒸馏 (VID)[37]、无数据知识蒸馏 (DAFL)[38]、多教师知识蒸馏 (AKD)[39]等引入信息熵的知识蒸馏相比较, 其特点在于本文提出的信息熵知识具有实际含义, 信息熵图是监督学生网络训练的知识, 反映了模型对边缘的注意力, 而其他蒸馏方法中的信息熵并没有相对应的具体含义. 例如 VID 利用信息熵推导教师网络和学生网络中间层特征图之间的互信息, 实际上迁移的是教师网络的中间层知识; DAFL 和 AKD 则是用信息熵的值越大, 其混乱程度越高的性质来约束教师网络的输出. 总而言之, VID 等蒸馏方法中的信息熵仅有数学含义, 并没有类似关注文本边缘这样的实际含义, 而本文提出的 MaskET 蒸馏方法充分利用信息熵知识能放大教师网络关于文本边缘注意力的特性, 从而有效提升学生网络的性能表现.

3 MaskET 方法

3.1 框架介绍

图 3 展示本节提出的知识蒸馏框架, 其核心是通过教师网络向学生网络迁移知识来提高学生网络的泛化能力.

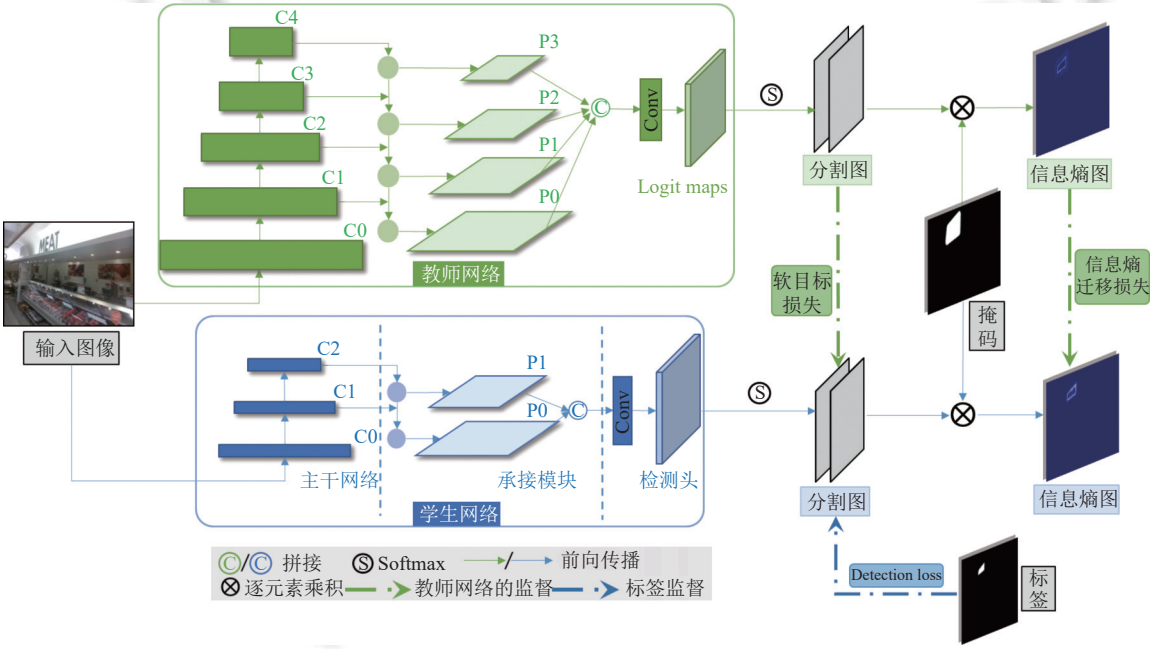


图 3 基于掩码信息熵迁移的知识蒸馏学习框架

具体而言, MaskET 训练框架包含两个参数规模不同的文本检测模型, 其中规模较大的作为教师网络, 规模较小的模型则作为学生网络. 教师网络预先进行训练并因为其更深的网络结构表现出良好的效果, 而学生网络通过接受来自教师网络的知识而提高检测精度. 训练图像输入到训练好的教师网络和学生网络, 经过前向传播, 各自输出分割图 (segmentation map). 学生网络的监督信息来自教师网络的知识和标签, 其中教师网络传递给学生网络的知识由两部分构成: (1) 来自教师网络的分割图, 本文使用逐像素点的概率值[25]作为学生网络的软目标; (2) 来自教师网络的掩码信息熵图, 其保留了图像中文本区域的关键信息.

3.2 掩码信息熵迁移

掩码信息熵迁移方法 (MaskET) 利用教师网络向学生网络迁移其信息熵知识来提高学生网络的性能, 加入掩码的作用是保留教师网络分割图文本周围区域的信息熵, 从而抑制其他可能的噪声区域的信息熵, 使传递给学生网络的信息熵知识更精炼.

由第 2 节可知, 信息熵图放大了模型对场景文本边缘的注意力, 并且由于教师网络和学生网络性能的差距, 教师网络的信息熵图形成的边缘比学生网络更加接近矩形, 而且更完整地包围文本边缘. 因此, 向学生网络迁移教师网络的信息熵知识, 有助于提高学生网络对场景文本边缘的注意力, 从而提高学生网络的检测效果.

如图 4 所示, 在训练过程中, 教师网络和学生网络输出尺寸相同但精度不同的分割图, 之后两张分割图被转化为信息熵图并分别加入掩码, 得到仅包含文本区域信息熵知识的熵值图. 学生网络通过接受来自教师网络信息熵图的监督进行训练, 具体表现在对训练过程添加信息熵迁移的损失. 这一完整的过程即掩码信息熵迁移 (MaskET).

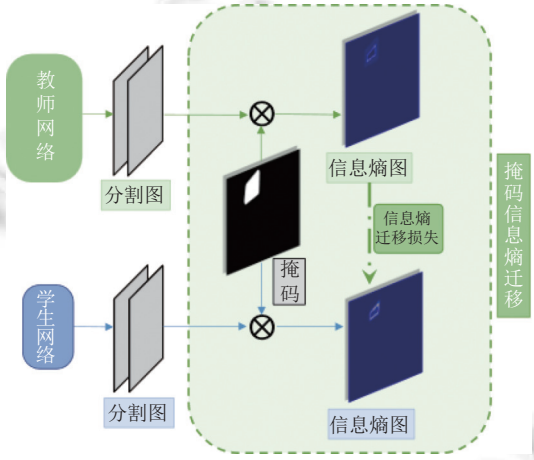


图 4 掩码信息熵迁移的示意图

3.3 文本检测模型

本文使用图像分割作为教师网络和学生网络的检测框架, 其结构按照功能划分为主干网络、承接模块、检测头.

● 主干网络 (backbone): 整个文本检测模型中参数量最大的部分, 作为核心主干负责提取多层次的图像特征, 一般以主干网络大小来区别学生网络和教师网络.

● 承接模块 (neck): 起承上启下的作用, 如图 3 所示, 特征金字塔 (feature pyramid network, FPN)[41]作为承接模块, 自上而下融合从高层到低层的特征图, 并将拼接的特征图输入到检测头网络中.

● 检测头 (detection head): 主要目的是使用上采样或者转置卷积把特征图放大到和输入图像同样大小, 然后对放大后的图像上的每一个像素点做二分类, 本文将 Softmax 归一化操作之前的特征图记为 logit map, 经过归一化的称为分割图 (segmentation map).

模型的检测损失函数 L_{det} 计算如下:

$$L_{\text{det}} = \alpha L_s + L_o \tag{1}$$

其中, L_s 表示分割图的分类损失, 一般使用交叉熵, L_o 表示非分割图的损失构成, 例如 DBNet 对阈值图 (threshold

map) 的可微二值化损失[12], EAST 对几何图 (geometry map) 的形状损失[15], 表示损失的权重系数, 按照相应论文[12,15]的最优参数设置.

3.4 损失函数

本文提出的 MaskET 蒸馏方法将教师网络文本区域的信息熵知识迁移到学生网络, 其中信息熵由文本检测模型的分割图计算得来, MaskET 将信息熵图与掩码相乘, 从而仅保留文本区域的信息熵, 消除背景信息的影响.

给定输入维度为 $H\times W\times 3$ 的图像 I, 其中 H 代表图像高, W 代表图像宽, 3 表示图像通道类型为 RGB, 文本检测模型对输入图像 I 预测出图像上每一个像素点分类成文本或者背景的概率值, 即检测头输出的分割图形状为 $H\times W\times 2$, 数字 2 表示分割网络输出两个通道, 一个通道输出图像上每一个像素点属于文本的概率值, 另一个通道表示像素点属于背景 (非文本区域) 的概率.

根据香农熵的定义, 计算分割图中每一个像素点对应的熵值:

$$E^{(h,w)}=-(P^{(h,w,0)}\log_2(P^{(h,w,0)})+P^{(h,w,1)}\log_2(P^{(h,w,1)})) \tag{2}$$

其中, $E^{(h,w)}$ 表示分割图中位于 (h,w) 像素点的信息熵值, 分别用 $E_T^{(h,w)}$ 和 $E_S^{(h,w)}$ 表示教师网络和学生网络的信息熵. P 表示检测头输出的分割图, 分别用 P_T 和 P_S 表示教师网络和学生网络的输出.

为了提取文本区域的信息熵, 本文使用与输入图像同样大小的二进制掩码 (mask, M):

$$M^{(h,w)}=\begin{cases}1, & \text{if } (h,w)\in t\\ 0, & \text{Otherwise}\end{cases} \tag{3}$$

其中, t 表示文本框, h 和 w 分别代表像素点的纵坐标和横坐标, (h, w) 的像素点如果处于文本区域, 那么 $M^{(h,w)}=1$, 否则为 0. 训练时所需要的掩码直接来源于文本检测数据集的标签, 不需要额外的获取代价. 因为 MaskET 方法的适用对象是自然场景文本检测, 该类任务的数据集的标签代表了图片中的文本区域, 且在训练时已转换为统一大小的矩阵, 可以直接作为掩码使用, 因此获取掩码不需要额外的代价.

如图 2 所示, 教师网络的信息熵图能够提取到更好的边缘轮廓信息, 因此本章提出的掩码信息熵迁移的知识蒸馏方法旨在将教师网络的信息熵知识迁移到学生网络. 为最小化学生网络和教师网络两者文本区域上每一个像素点信息熵的差, 本文使用带有掩码的 L_1 损失来计算其信息熵迁移损失:

$$L_{\text{et}}=\frac{1}{N}\sum_{(h,w)}M^{(h,w)}\left\|E_T^{(h,w)}-E_S^{(h,w)}\right\|_1 \tag{4}$$

其中, $N=\sum_{h,w}M^{(h,w)}$, 表示文本区域的像素点个数. 掩码的操作在熵值图上进行, 熵值图的通道数为 1, 掩码和学生、教师网络的熵值图差进行点乘, 传递信息量的代价为 O(n). 而其他蒸馏方法大多数需要对齐学生、教师网络的特征图, 特征图的通道数较高, 所需要的代价将远大于 O(n). 例如, 结构化知识蒸馏 (structured knowledge distillation, SKD)[25]需要对齐师生网络的中间层特征图的成对相似性, 需要的代价为 O(n^2).

与传统图像分类软目标的知识蒸馏方法相似, 本文增加逐像素点分类[25]蒸馏损失, 希望学生网络不仅学习到教师网络局部边缘的信息熵知识, 同时也能模仿教师网络分割图上的概率分布, 用 KL 散度最小化教师网络的分割图和学生网络的分割图上每一个像素点的分布:

$$L_{\text{pi}}=\frac{1}{W\times H}\sum_{(h,w)}\sum_{i=1}^{c}P_T^{(h,w,i)}\log\left(\frac{P_T^{(h,w,i)}}{P_S^{(h,w,i)}}\right) \tag{5}$$

其中, $P_T^{(h,w,i)}$ 和 $P_S^{(h,w,i)}$ 分别表示教师网络和学生网络的分割图上位于 (h, w) 的第 i 个类别的概率值, c 为总类别数, 文本检测任务中固定为 2.

按照原始蒸馏方法的理论, 引入温度超参数 τ 来控制教师网络分割图上每点概率分布的平滑程度, 即:

$$P^{(h,w,i)}=\frac{\exp(q^{(h,w,i)}/\tau)}{\sum_i\exp(q^{(h,w,i)}/\tau)} \tag{6}$$

其中, $q^{(h,w,i)}$ 表示 logit map 上位于 (h, w) 的第 i 个类别的值. 和原始知识蒸馏方法要求 τ 是大于 1 的正数不同, 在

MaskET 方法中 $\tau = 1$, 其原因从第 5.4.2 节中探究温度超参数的实验结果中分析而来.

训练学生网络的最终损失函数 L 包含文本检测损失 L_{det}、信息熵迁移损失 L_{et} 和逐像素点分类损失 L_{pi}:

$$L = L_{\text{det}} + \lambda_1 L_{\text{et}} + \lambda_2 L_{\text{pi}} \tag{7}$$

其中, λ_1 是 L_{et} 权重参数, 设置为 6, λ_2 是 L_{pi} 权重参数, 固定为 10. 参数值的设置由第 5.4.3 节中探究损失权重的实验得来.

3.5 训练方法

如算法 1 所示, MaskET 方法是两阶段训练方法. 第 1 阶段训练教师网络, 利用给定的数据集, 最小化检测损失 L_{det}, 目的是获得精度较高的大网络. 在阶段 2 训练学生网络的过程中, 固定教师网络参数, 在教师网络的掩码信息熵和逐像素点概率值, 以及标签的监督下训练一个以轻量级卷积神经网络为主干的学生网络.

算法 1. MaskET 训练算法.

输入: 训练数据集 D_{train}; 教师网络 θ_T; 学生网络 θ_S;
输出: 学生网络的最优参数 θ_S^*.

阶段 1: 训练教师网络
初始化教师网络参数
FOR *epoch* = 1 **TO** epochs **DO** //训练 1 200 轮
  FOREACH *minibatch* **IN** D_{train} **DO**
  教师网络前向传播;
  计算检测损失 L_{det};
  使用 $\nabla_{\theta_T} L_{\text{det}}$ 更新 θ_T;
  END
END
阶段 2: 训练学生网络
固定教师网络参数;
初始化学生网络参数
FOR *epoch* = 1 **TO** epochs **DO** //训练 1 200 轮
  FOREACH *minibatch* **IN** D_{train} **DO**
  教师网络前向传播;
  学生网络前向传播;
  计算学生网络和教师网络的信息熵 E_T 和 E_S;
  生成教师、学生网络的软目标 $P_T^{(h,w,i)}$ 和 $P_S^{(h,w,i)}$;
  计算学生网络的总训练损失 L;
  使用 $\nabla_{\theta_S} L$ 更新 θ_S;
  END
END

4 实 验

4.1 数据集与评价指标

4.1.1 数据集

本文实验中使用了 6 个标准自然场景文本检测数据集.

● ICDAR 2013[42]: 来源于 ICDAR 2013 竞赛挑战 2, 包含 229 张训练集图像和 233 张测试集图像. 该数据集文本排布特点为水平, 提供字符级和单词级标注, 以矩形标注.

● ICDAR 2015[11]: 来源于 ICDAR 2015 竞赛挑战 4, 是以矩形边界框标注的英文文本数据集, 其中训练图片有 1 000 张, 500 张用作测试图片. ICDAR 2015 使用谷歌眼镜随机采集街景图片, 因此数据集特点是具有任意方向的文本, 像素值低.

● TD500[43]: 以矩形文本框标注的中英文混合数据集, 其中有 300 张图片用于训练集, 测试数据集 200 张, 包含水平以及倾斜的自然场景文本. MSRA-TD500 使用数码相机采集来自办公室或者室外街道的图像, 分辨率从 1296×864 到 1920×1280 之间.

● TD-TR: 本文参考相关文献[12,15]中的做法, 将 400 张 HUST-TR400[44]数据集图像和 TD500 的训练数据集合并, 因此 TD-TR 数据集的训练集有 700 张, 测试集与 TD500 一致.

● Total-Text[45]: 多边形标注的弯曲文本数据集, 共 1 555 张图片, 其中训练集有 1 255 张, 测试集 300 张. Total-Text 数据集特点是方向多样性, 包含水平方向和多方向、多种弯曲样式, 如圆形、波浪形等.

● CASIA-10K[46]: 该数据集包含 10 000 张图像, 其中训练集图像有 7 000 张, 其余 3 000 张图像用于测试. 数据集采集自中文场景, 每条文本行标注其四边形的 8 个坐标值.

4.1.2 评价指标

本文使用精确率 (precision, P)、召回率 (recall, R) 和精确率召回率的调和平均 ($F1$-score, $F1$) 来评价文本检测方法的性能. 根据检测的矩形框 D_i 和标签 G_j 之间的交并比 IoU (intersection over union) 统计出正确检测的矩形框集合 T_p, 计算公式如下:

$$IoU = \frac{area(G_j \cap D_i)}{area(G_j \cup D_i)} \tag{8}$$

其中, $area(G_j \cap D_i)$ 和 $area(G_j \cup D_i)$ 分别表示 G_j 和 D_i 的交集和并集区域面积.

进而可求得精确率 P、召回率 R 和 $F1$ 如下:

$$P = \frac{|T_P|}{|D|} \tag{9}$$

$$R = \frac{|T_P|}{|G|} \tag{10}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{11}$$

4.2 实验设置

教师网络和学生网络规模大小的差别体现在主干网络, 其中教师网络使用参数量较大的 ResNet50[10]作为主干网络, 而学生网络采用轻量级网络 MobileNetV3[47]. 本文选择广泛使用的 DBNet 和 EAST 作为学生网络的检测头, 分别记为 MV3-DB 和 MV3-EAST 以区别不同架构的学生网络. 由于教师网络已经提前训练好, 此处仅讨论训练学生网络时的相关设置. 模型训练的优化器为 SGD, 其中初始学习率为 0.007, 优化器动量设置为 0.9, 权重衰减设置为 0.000 1, 并且使用多项式学习率调整策略. 总共训练 1 200 epoch, 一个 batch size 为 8. 训练图像经过随机翻转, 旋转 (−10, 10), 最后统一裁剪至 640×640. 如果没有特别说明, 图像的数据增强采用随机旋转和随机剪裁. 为了保证不超出显存, 训练 EAST 时训练图像统一缩放至 512×512, 训练 DB 时, 缩放至 640×640.

所有实现深度学习框架 PyTorch, 在 Windows 10 平台上使用单张 1080Ti 显卡训练. 实验环境如表 1 所示.

表 1 环境配置

项目	配置
处理器	Intel(R) Core(TM) i7-7800X CPU @ 3.50 GHz
内存	DDR4 2133 MHz 96.0 GB
显卡	NVIDIA GTX 1080Ti
操作系统	Windows 10 (19042.1110)

4.3 对比实验

4.3.1 与其他知识蒸馏方法对比

在本节实验中, 学生网络 (baseline 模型) 的主干网络采用 MobileNetV3, 教师网络则是将主干网络替换为 ResNet50, 在 ICDAR 2013 等 6 个数据集上使用可微二值化 (DB) 作为文本检测分割头 (MV3-DB), 由于 EAST 复现的精度在 TD500 和 TD-TR 数据集上与原论文差距过大和不支持 Total-Text 这类弯曲文本数据集, 因此仅在 ICDAR 2013、ICDAR 2015 和 CASIA-10K 数据集上进行实验. 本文将 MaskET 和目前学术界广泛使用的 8 种蒸馏方法进行对比, 比较其在测试集上的精度. 这 8 种蒸馏方法在文本检测模型上的应用设置如下.

- 软目蒸馏方法 (soft target, ST)[4]: Hinton 等人提到的最原始蒸馏方法, 迁移教师网络输出的温度超参数大于 1 的软化概率分布知识;
- 中间层知识蒸馏方法 (FitNets)[8]: 不仅利用教师网络的软目标知识, 还用到了中间层的特征图知识;
- 知识适配蒸馏 (knowledge adaptation, KA)[24]: 预先训练的自编码器压缩教师网络的中间特征图知识, 学生网络再通过适配器来接收教师网络的知识;
- 结构化知识蒸馏 (structured knowledge distillation, SKD)[25]: 使用 KL 散度对齐教师-学生网络分割图上的每一个像素点概率和教师网络中间层特征图上成对相似性;
- 变分信息蒸馏 (variational information distillation, VID)[37]: 使用变分分布转换教师网络-学生网络特征金字塔合并后的中间特征图;
- 自蒸馏 (self-distillation, SD)[26]: 在特征金字塔的每一层设置辅助分类器, 从最深层分类器的提炼出软目标和特征图, 迁移到每个辅助分类器;
- 自我注意力蒸馏 (self attention distillation, SAD)[27]: 将特征金字塔的深层部分的注意图当作蒸馏的知识, 浅层的网络模仿更深层网络的注意力图, 例如 P1 层的注意力图的蒸馏目标是 P2 层.
- 通道式知识蒸馏 (channel-wise distillation, CD)[48]: 通过带有温度超参数的 Softmax 函数将教师-学生网络中间特征图和 logit 图的激活值转化为软化概率分布, 并利用 KL 散度对齐.
- 类内特征差异蒸馏 (intra-class feature variation distillation, IFVD)[49]: 通过 IFV 模块构造每张图片内同一类像素点之间的差异图 (intra-class feature variation map), 并使学生网络模仿教师网络的 IFV 图进行训练.

实验结果如表 2 所示 (最高 $F1$ 值加粗显示), 相较于直接训练的 baseline 模型 (学生网络), MaskET 在各大数据集上均能显著提升学生网络 MV3-DB 的 $F1$, 并且优于其他知识蒸馏方法. 在 ICDAR 2013 数据集上 $F1$ 达到 76.1%, 比 baseline 高出 2.3 个百分点, 远超过其他知识蒸馏方法; 其他 5 个数据集也都有 1 个百分点左右的提升. 同样如表 3 所示 (最高 $F1$ 值加粗显示), 以 MV3-EAST 作为学生网络, 在 ICDAR 2013、ICDAR 2015 和 CASIA-10K 数据集上, MaskET 方法均有可观的性能提升, $F1$ 分别有 1.4%、0.9% 和 0.7% 的提高. 表 2、表 3 的实验结果为对应论文中的图像分类蒸馏方法在 DB 和 EAST 文本检测模型上复现的结果.

对比表 2 和表 3, MaskET 在 MV3-EAST 上的提升效果相对较小. 在 CASIA-10K 数据集上, MV3-DB 高出 baseline 有 1.9 个百分点, 但 MV3-EAST 仅有 0.7 个点的提升. 由于学生网络 MV3-EAST 和教师网络 ResNet50-DB 两者分割图的分辨率不一致, 为了能够计算两者的信息熵差, 教师网络的分割图通过池化的操作, 缩小至与 MV3-EAST 分割图相同的尺寸, 因此计算信息熵图时丢失了部分信息, 而 MV3-DB 不需要缩放分割图, 因而效果更好.

表 2 和表 3 的实验结果也反映 MaskET 虽然能提升 $F1$ 值和召回率, 但精确率也有所下降, 第 4.4 节消融实验中的第 4.4.3 节也证明了该结论. 其原因在于信息熵损失函数仅考虑学生网络和教师网络预测一致的正样本损失, 未惩罚学生网络误检下的信息熵损失, 导致学生网络检测框数量增多, 进而导致 TP 样本和 FP 样本增大. 根据评价指标的计算公式 (9)–公式 (11), 可知检测框数量$|D|$增多, $|T_P|$自然会增加, 召回率也随之提高, 同时$|T_P|$提高的幅度没有$|D|$大, 因此精确率降低. 但是, 绝大多数情况下 MaskET 会大幅提升 baseline 模型的召回率, 虽然精确率有所下降, 但综合评价指标 $F1$ 值是提高的.

从表 2 和表 3 的结果可以看出, 在不同规模大小的数据集上其他蒸馏方法难有一致的性能提升. 在小规模数

据集 ICDAR 2013、TD500 和 TD-TR 上, ST、KA、FitNets 以及 SKD 都能不同程度上提升学生网络的 $F1$ 指标, 然而在大一些的数据集如 ICDAR 2015、Total-Text 和 CASIA-10K 上, 绝大多数蒸馏方法难以有效提高学生网络的性能表现, 甚至出现性能下降, 不如直接训练的 baseline, 例如图像分类任务中最常用的软目标知识蒸馏方法. 在 CASIA-10K 这类难度更大的数据集上, 教师网络和学生网络之间学习能力差距更加明显, 导致知识迁移的效率降低[50], 而 MaskET 在 CASIA-10K 有较大的提升, 这说明通过迁移教师网络的信息熵知识可以一定程度下缓解这个问题.

表 2 MV3-DB 在不同数据集上的知识蒸馏实验结果 (%)

方法	ICDAR 2013			TD500			TD-TR			ICDAR 2015			Total-Text			CASIA-10K		
	P	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1
baseline	83.7	66.0	73.8	78.7	71.4	74.9	83.6	74.4	78.7	87.1	71.8	78.7	87.2	66.9	75.7	88.1	51.9	65.3
ST[4]	82.5	65.8	73.2	77.0	73.0	74.9	84.6	73.5	78.7	85.4	72.2	78.2	87.4	65.3	74.8	88.8	49.4	63.5
KA[24]	82.5	66.8	73.8	79.5	71.3	75.2	86.3	72.5	78.8	85.0	73.3	78.7	85.9	66.8	75.2	87.8	51.4	64.8
FitNets[8]	84.7	65.4	73.8	78.6	73.3	75.8	85.3	74.0	79.2	85.3	73.3	78.8	87.4	67.5	76.2	88.0	52.3	65.6
SKD[25]	82.4	68.8	75.0	81.2	70.6	75.5	84.8	74.5	79.3	87.4	71.6	78.7	87.4	67.0	75.9	88.6	51.6	65.2
VID[37]	81.9	69.7	75.3	81.4	71.5	76.1	86.6	72.0	78.6	86.4	72.6	78.9	86.5	66.7	75.3	87.4	50.9	64.3
SD[26]	83.5	67.8	74.8	79.4	72.2	75.6	85.0	74.0	79.1	85.1	73.0	78.6	87.0	67.6	76.1	87.1	52.0	65.1
SAD[27]	82.8	66.7	73.9	78.7	72.3	75.4	87.3	72.0	78.9	86.7	72.7	79.1	86.5	67.1	75.6	88.4	50.7	64.4
CD[48]	83.0	67.2	74.3	79.0	72.3	75.5	83.4	73.2	78.0	86.0	72.9	78.9	86.6	66.5	75.2	87.9	50.5	64.1
IFVD[49]	83.3	68.8	75.4	80.0	70.1	74.7	85.3	73.7	79.1	86.8	71.7	78.5	86.1	67.0	75.4	87.3	51.2	64.6
MaskET	83.7	69.8	**76.1**	78.0	74.4	**76.2**	84.9	74.4	**79.3**	84.3	75.8	**79.8**	84.6	70.5	**76.9**	86.4	55.0	**67.2**

表 3 MV3-EAST 在不同数据集上的知识蒸馏实验结果 (%)

方法	ICDAR 2013			ICDAR 2015			CASIA-10K		
	P	*R*	*F*1	*P*	*R*	*F*1	*P*	*R*	*F*1
baseline	81.7	64.4	72.0	80.9	75.4	78.0	66.1	64.9	65.5
ST[4]	77.8	64.9	70.8	80.9	75.1	77.9	64.7	65.1	64.9
KA[24]	78.6	64.0	70.5	78.2	76.4	77.3	67.7	63.0	65.3
FitNets[8]	82.4	65.8	73.2	78.0	77.8	77.9	65.4	64.2	64.8
SKD[25]	79.5	66.3	72.3	81.9	75.6	78.6	66.6	64.7	65.6
VID[37]	80.8	64.8	71.9	81.4	75.3	78.2	65.4	63.9	64.6
SD[26]	80.2	63.8	71.1	79.6	74.7	77.1	66.2	63.5	64.8
SAD[27]	81.4	65.6	72.6	80.2	76.5	78.3	65.7	64.1	64.9
CD[48]	81.7	65.2	72.6	79.0	77.3	78.2	66.7	64.0	65.3
IFVD[49]	81.6	65.5	72.7	80.4	76.0	78.1	68.2	62.4	65.2
MaskET	82.2	66.3	**73.4**	81.0	76.9	**78.9**	69.8	62.9	**66.2**

综上所述, 本文提出的 MaskET 训练方法在多个数据集上均能有效提升学生网络 MV3-DB 和 MV3-EAST 的性能.

4.3.2 显著性检验分析

为了检验 MaskET 方法的提升效果是否在统计上显著, 本文使用 Wilcoxon 秩和检验进行检验. 具体来说, 我们基于 ICDAR 2013 数据集, 使用每一种蒸馏方法对 MV3-DB 进行训练, 并重复 10 次实验, 统计得到的精确率 P、召回率 R 和 $F1$. 首先, 比较 MaskET 与对比方法在对应指标上的平均值, 若 MaskET 的平均值小于对比方法, 则不进行检验; 若 MaskET 的平均值大于对比方法, 则进一步检验这种差异是否显著. 以 MaskET 与 IFVD 在精确率 P 指标上的检验为例, 检验的原始假设为 $H_0:\mu^P_{\text{MaskET}}=\mu^P_{\text{IFVD}}$, 备选假设为 $H_0:\mu^P_{\text{MaskET}}\neq\mu^P_{\text{IFVD}}$, 其中 μ^P 代表了对应蒸馏方法在精确率 P 指标上的平均值.

表 4 展示了 MaskET 与对比方法在精确率 P、召回率 R 和 $F1$ 上检验结果. 结果为–1, 代表 MaskET 方法在该指标上的均值低于对比方法; 结果为 1, 代表在 95% 的置信度下认为 MaskET 方法在对应指标上显著优于对比方法; 检验结果为 0, 代表 MaskET 方法在对应指标上并不显著优于对比方法. 此外, 我们也在括号中报告了检验得到的 p 值.

从表 4 可以看出, MaskET 在仅精确率 P 上不显著优于 FitNets、SD 及 IFVD; 仅在召回率 R 上不显著优于 VID. 除此之外, MaskET 在精确率 P 和召回率 R 上都显著优于对比方法, 且在 $F1$ 上显著优于所有对比方法.

表 4 MaskET 与其他蒸馏方法 Wilcoxon 秩和检验结果 h (p)

指标	ST	KA	FitNets	SKD	VID	SD	SAD	CD	IFVD
P	**1** (0.021)	**1** (0.002)	−1	**1** (0.002)	**1** (0.005)	0 (0.381)	**1** (0.011)	**1** (0.020)	0 (0.240)
R	**1** (0.001)	**1** (0.001)	**1** (0.001)	**1** (0.025)	0 (0.42)	**1** (0.002)	**1** (0.001)	**1** (0.028)	**1** (0.022)
$F1$	**1** (0.001)	**1** (0.001)	**1** (0.017)	**1** (0.011)	**1** (0.021)	**1** (0.019)	**1** (0.002)	**1** (0.021)	**1** (0.034)

4.3.3 不同知识蒸馏损失项对 MaskET 的影响

本文提出的 MaskET 方法的蒸馏损失是由逐像素点分类蒸馏损失 L_{pi} 和掩码信息熵迁移损失 L_{et} 组成, 因此本部分实验在 TD500 和 ICDAR 2015 数据集上探究不同知识蒸馏损失和信息熵图的掩码对 MaskET 的性能影响, 其中 baseline 是 MV3-DB, 教师网络是 ResNet50-DB.

从表 5 所展示的实验结果可以得出以下结论.

表 5 不同知识蒸馏损失项对 MaskET 的影响 (%)

损失项	TD500			ICDAR 2015		
	P	R	$F1$	P	R	$F1$
baseline	78.7	71.4	74.9	87.1	71.8	78.7
L_{pi}	81.7	69.9	75.3	86.2	72.6	78.8
L_{et} (w/o mask)	80.8	70.3	75.2	85.5	72.5	78.5
L_{et} (w/ mask)	77.0	74.0	75.5	84.2	74.7	79.2
L_{et} (w/o mask)+ L_{pi}	83.3	69.7	75.9	85.2	73.1	78.7
L_{et} (w/ mask)+ L_{pi}	78.0	74.4	**76.2**	84.3	75.8	**79.8**

注: w/o 是without的缩写, w/ 是with的缩写

(1) 单独使用某一项损失项, 掩码信息熵迁移损失 (L_{et} (w/ mask)) 提升较为明显, 而不使用掩码的信息熵迁移损失 (L_{et} (w/o mask)) 效果较差.

(2) 同样增加逐像素点分类蒸馏损失 (L_{pi}) 的情况下, 使用掩码的信息熵迁移损失 (L_{et} (w/ mask)) 训练的学生网络获得了最高的检测精度, $F1$ 相较于 baseline 在 TD500 和 ICDAR 2015 上分别提高了 1.3% 和 1.1%, 而不使用掩码的蒸馏损失 (L_{et} (w/o mask)+ L_{pi}) 效果略差. 使用掩码提升效果更为明显的原因是该操作使教师网络仅保留了文本区域的信息熵知识, 避免了背景噪声的影响.

这些结果表明, 掩码信息熵迁移是逐像素点分类蒸馏的有效补充, 掩码信息熵知识反映了教师网络的在文本区域边缘的信息, 让学生网络获得更多的泛化信息, 提升了网络的性能.

4.3.4 温度超参数对 MaskET 的影响

MaskET 中信息熵图的计算和逐像素点分类蒸馏都依赖于教师网络的分割图, 按照文献 [4,25] 的多分类蒸馏思路, 分割图是 logit map 经过带有温度超参数 τ 的 Softmax 归一化“软化”得到, 为了评估温度对 MaskET 的影响, 本节比较 MV3-DB 在不同温度 τ 的测试精度, 实验数据集为 ICDAR 2015, 温度分别设置为 1、2、10、20、30、50.

从表 6 实验结果可以看出, $\tau = 1$ 时 MaskET 效果最好, 将学生网络从 78.7% 提高到 79.8%. 然而随着温度的升高, 学生网络的 $F1$ 逐渐下降, 当 τ 升至 50, 相较于 baseline, $F1$ 值下降了 1.2 个百分点. 从图 5 以看出, $\tau = 1$ 时, 教师网络的信息熵图蕴含更丰富的知识, 能够明显分辨出边缘, 但是随着 τ 升高, 与图 1 中所展示的那样教师网络输出的概率分布软化效果越加明显, 其分布逐渐接近均匀分布, 这使得信息熵值逐渐接近最大值, 丢失了有效的边

缘信息, 因而学生网络训练时难以获得足够的知识.

表 6 不同温度 (τ) 下 MaskET 的实验结果 (%)

方法	P	R	$F1$
baseline	87.1	71.8	78.7
τ=0.5	82.8	76.2	79.4
τ=1	84.3	75.8	**79.8**
τ=1.5	83.0	76.8	79.7
τ=2	87.7	71.7	78.9
τ=10	85.5	72.2	78.3
τ=20	89.2	69.4	78.0
τ=30	88.7	69.0	77.6
τ=50	89.5	68.4	77.5

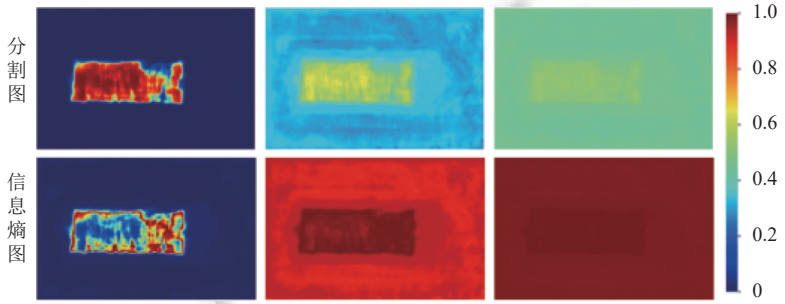


图 5 不同温度下教师网络的分割图与信息熵图

4.3.5 损失权重超参数对 MaskET 的影响

训练学生网络的最终损失函数 L 包含文本检测损失 L_{det}、信息熵迁移损失 L_{et} 和逐像素点分类损失 L_{pi}, 其中 L_{et} 的权重参数为 λ_1, L_{pi} 的权重参数为 λ_2. 为了评估损失权重超参数 λ_1 和 λ_2 对 MaskET 的影响, 本节比较 MV3-DB 在不同损失权重下的测试精度, 实验数据集为 ICDAR 2015. 在 λ_1 设置为 6 时, λ_2 分别设置为 2、4、6、8、10、12; 在 λ_2 设置为 10 时, λ_1 分别设置为 2、4、6、8、10、12.

从图 6 可以看出, 提高信息熵迁移损失的权重 λ_1 可以有效提高召回率, 但在一定程度上会降低精确率; 提高逐像素点分类损失的权重 λ_2 可以有效提高精准率, 但召回率会降低. 当 λ_1 取 6、λ_2 取 10 的时候, $F1$ 值分别达到最大, 在精准率和召回率之间达到了最好的平衡. 因篇幅所限, 其他数据集上的实验结果不再展示, 但其所给出的最优超参数也非常接近, 因此本文实验中统一将 λ_1 设置为 6, λ_2 设置为 10.

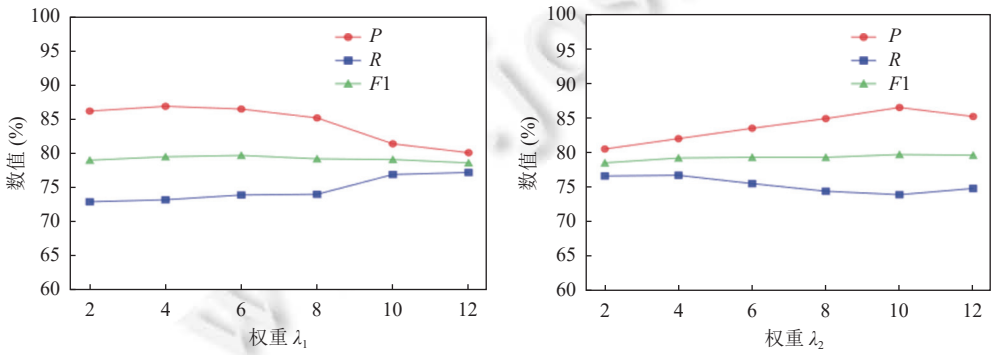


图 6 损失权重 λ_1、λ_2 对 MaskET 的影响

4.4 模型复杂度分析

深度学习方法在实际应用时, 除了考虑算法的精度, 还要考虑部署环境, 尤其是对实时性要求高的场合, 因此本文从参数量 (parameters)、存储大小 (storage size) 来评估模型占用系统内存大小; 从每秒浮点操作次数 (floating-point operations per second, FLOPs)、每秒传输帧数 (frames per second, FPS) 评价模型的运算量及其前向速度.

表 7 是教师网络 ResNet50-DB 和学生网络 MV3-DB 的复杂度对比结果, 其中实验环境与第 5.2 节一致, 在 ICDAR 2015 数据集上测试模型推理速度. 可以看到, 学生网络的存储大小 (storage size) 仅为 3.70 MB, 是教师网络的 1/25, GPU 运算量 (FLOPs) 只有 ResNet50-DB 的 1/15, 推理速度 (FPS) 更是 ResNet50-DB 的 3 倍. MaskET 蒸馏训练的学生网络不仅检测精度高, 模型占用的系统内存小, 前向推理速度也较快, 因此较为适合部署在资源受限的嵌入式设备上.

表 7 学生网络与教师网络的模型复杂度对比

模型	FLOPs	Parameters	FPS	Storage size (MB)
Teacher (ResNet50-DB)	1 841 959 296	24 640 226	13	94.9
MaskET (MV3-DB)	116 313 742	914 488	39	3.70

表 8 是学生网络 MV3-DB 和现有文本检测算法 DBNet[12]和 TextFuseNet[3]的复杂度对比结果, 可以看出, 现有的文本检测模型大多使用复杂且层数较多的网络来提升检测精度, 这类模型往往需要较大的计算量和存储空间. 例如, 基于 ResNet50 的 DBNet 的存储大小 (storage size) 达到了 110.36 MB. 与其相比, 本文的学生网络的存储大小 (storage size) 仅为 3.70 MB, 在模型大小和模型精度之间达到了很好的平衡, 从而可以更好地部署在资源有限的平台上.

表 8 学生网络与现有算法模型复杂度对比

模型	FLOPs	Parameters	FPS	Storage size (MB)
MaskET (MV3-DB)	116 313 742	914 488	39	3.70
DBNet (ResNet50)	1 862 274 048	24 785 218	15	110.36
DBNet (ResNet18)	1 279 331 328	12 269 378	32	52.77
TextFuseNet (ResNet101)	2 447 861 928	43 777 346	4.1	872.5

4.5 可视化分析

4.5.1 特征图可视化

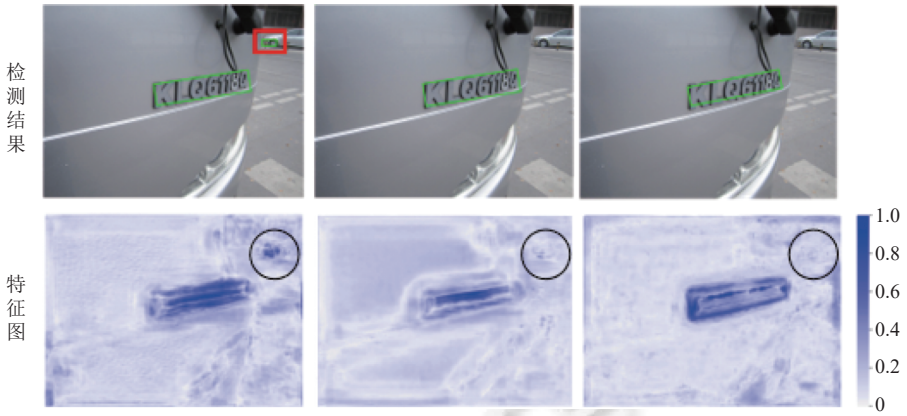
为了更直观地分析 MaskET 提升学生网络性能的内部机理, 本文选取 baseline (MV3-DB)、MaskET 和 ResNet50-DB, 将它们的融合特征 (图 3 中特征金字塔拼接后的特征图) 进行可视化. 图 7 特征图可视化结果中颜色深浅反映网络关注程度, 图像中的区域颜色越深, 说明网络对此处的激活响应更高, 即对该位置的关注度越高, 蓝色越淡表示激活值越低, 可以理解为网络对该位置的像素区域注意力低.

从后文图 7 预测结果中可以看到, 学生网络存在误检的情况, 即图中红色矩形标注的区域, 误把另一辆车的车辆当作文本, 而教师网络 ResNet50-DB 和经过 MaskET 训练的模型则没有. 学生网络误检的可能原因是网络在非目标区域的地方也有较高的关注度, 对照特征可视化结果, 网络在“车轮”区域 (图中黑色椭圆标注) 的颜色深度不亚于文本区域, 因此最终网络输出结果时难以分辨出噪声, 造成误判. 对比 baseline, 教师网络和 MaskET 学到的特征图不仅背景噪声处激活值低得多, 而且仅在文本区域有较高的响应值. 这说明引入 MaskET 知识蒸馏方法后, 通过带有掩码操作的信息熵引导学生网络去学习教师网络的文本位置的边缘信息, 忽略背景的信息, 使得学生网络对目标区域有更高的关注度, 抑制对噪声区域的注意力, 有效避免了噪声信息影响到模型的特征学习, 进而提升学生网络的检测性能.

4.5.2 与基线模型的检测结果对比

除了定量分析 MaskET 的性能提升, 本文从实际检测效果上定性比较 MaskET 和基线模型. 图 8 展示 MV3-DB

在不同数据集上检测结果的可视化图像, 图像中第 2 行表示基线模型的检测结果, 第 2 行表示 MaskET, 矩形框表示检测出来的文本框, 椭圆表示两者的差异之处.



(a) Baseline (b) MaskET (c) ResNet50-DB

图 7 特征图可视化结果

(a) ICDAR 2013 (b) TD500 (c) TD-TR

(d) ICDAR 2015 (e) Total-Text (f) CASIA-10K

图 8 基线模型 MV3-DB 与 MaskET 的检测结果对比

从结果中可以看出, 经过 MaskET 知识蒸馏训练得到的学生网络的检测效果较好. 对比基线模型, MaskET 能够加强学生网络对小目标的检测能力, 例如 baseline 模型没有检测到图 8(a) 中字母“F”和“T”, 图 8(d) 中天花板上

的出口标志以及图 8(f) 中门牌上的英文. 同时可以发现基线模型存在错检的情况, 例如图 8(b), baseline 模型错把背景的一部分当作文本, 而 MaskET 判别能力更强, 不会将背景检测为文本. 如图 8(c) 所示, MaskET 训练的学生网络检测的文本框均能更加完整包裹住文本的边缘, 可见学生网络通过学习蕴含教师网络关于文本边缘注意力的信息熵知识, 有效提升边缘检测能力.

5 有效性威胁

5.1 内部有效性威胁

内部有效性威胁主要来源于对文本检测模型、MaskET 和其对比方法的实现过程, 以及在实验分析中计算相关指标的过程. 为了减少这方面的威胁, 我们在复现本文检测模型时, 使用了相关论文的开源代码来实现, 并保证模型参数严格遵循原论文中的最佳参数; 在实现 MaskET 方法时, 我们尽可能地使用 Python 中现有成熟框架的封装代码来实现; 在实现对比蒸馏方法时, 我们参考了相对应论文的开源代码, 并将其迁移至我们的项目环境中. 此外, 在实验分析时, 尽管相关指标的计算并不复杂, 我们仍参考了相关工作的开源项目, 将其计算评价指标的代码迁移至我们的环境中, 以实现评价指标的计算.

5.2 外部有效性威胁

外部有效性威胁主要来源于实验数据集以及预训练模型的参数文件. 为了评价所提方法的效果, 我们在 6 个标准自然场景文本检测数据集上进行了验证, 尽管这些数据集是业界广泛使用的 benchmark 数据集, 但可能存在对特定领域或场景的偏见, 我们仍然需要考虑模型在其他数据集上的泛化性能. 特别是实际应用时, 如果真实的数据包含不同风格的文本类型、场景或质量, 模型可能无法良好地适应, 从而影响其在实际应用中的效果. 此外, 在训练模型前, 我们加载了业界广泛采用的预训练权重文件, 此权重文件基于 ImageNet 数据集进行分类训练而得到. 尽管这一做法对于模型的训练效果有所提升, 在实际应用时我们仍需考虑领域适应的问题, 预训练权重文件的数据集是否适配实际应用场景中的文本检测任务是需要考虑的因素.

5.3 结构有效性威胁

结构有效性威胁主要来源于实验中的超参数选择及实验设计. 为了减少这方面的威胁, 我们在实现文本检测模型、对比蒸馏方法时时遵循了原始论文中的最佳参数推荐; 对于温度超参数及损失权重超参数, 我们进行了敏感性分析实验, 并选择实验效果最佳的超参数作为我们的实验设置. 在实验设计方面, 我们从统计上分析了所提方法与对比方法的显著性差异, 设计了消融实验以验证我们方法中各个模块的有效性, 并进行了可视化实验以便从直觉上理解我们的方法. 但由于模型的单次训练时间较长, 我们仅在单个数据集上进行了上述实验, 因此超参数的选择及其他实验结果可能存在偏向性, 在未来的工作中, 我们将继续研究这些超参数及各个模块对方法性能的影响.

6 结 论

针对图像分类的软目标知识存在泛化信息不足的问题, 本文定义了一种全新的更具泛化性的信息熵知识. 在此基础上, 本文进一步提出了基于掩码信息熵迁移的知识蒸馏方法 (MaskET), 通过掩码的操作, 仅提取教师网络关于文本框区域的信息熵知识, 用于指导学生网络训练. 在 6 个公共标准数据集的实验表明, MaskET 能有效提高基线模型的 $F1$, 并且与其他蒸馏方法对比, 取得最好的效果. 未来我们将进一步探索把信息熵作为知识融入自蒸馏的框架, 以省去知识蒸馏预先训练教师网络带来的额外训练成本.

References:

[1] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proc. of the 2015 IEEE Conf. on Computer Vision and Pattern Recognition. Boston: IEEE, 2015. 3431–3440. [doi: 10.1109/CVPR.2015.7298965]

[2] Yuan YH, Chen XL, Wang JD. Object-contextual representations for semantic segmentation. In: Proc. of the 61st European Conf. on Computer Vision. Glasgow: Springer, 2020. 173–190. [doi: 10.1007/978-3-030-58539-6_11]

[3] Ye J, Chen Z, Liu JH, Du B. TextFuseNet: Scene text detection with richer fused features. In: Proc. of the 29th Int'l Joint Conf. on Artificial Intelligence. Yokohama: IJCAI, 2020. 516–522.

[4] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv:1503.02531, 2015.

[5] Gao H, Tian YL, Xu FY, Zhong S. Survey of deep learning model compression and acceleration. Ruan Jian Xue Bao/Journal of Software, 2021, 32(1): 68–92 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6096.htm [doi: 10.13328/j.cnki.jos.006096]

[6] Li ZH, Xu PF, Chang XJ, Yang LY, Zhang YY, Yao LN, Chen XJ. When object detection meets knowledge distillation: A survey. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2023, 45(8): 10555–10579. [doi: 10.1109/TPAMI.2023.3257546]

[7] Du YN, Li CX, Guo RY, Cui C, Liu WW, Zhou J, Lu B, Yang YH, Liu QW, Hu XG, Yu DH, Ma YJ. PP-OCRv2: Bag of tricks for ultra lightweight OCR system. arXiv:2109.03144, 2021.

[8] Romero A, Ballas N, Kahou SE, Chassang A, Gatta C, Bengio Y. FitNets: Hints for thin deep nets. arXiv:1412.6550, 2014.

[9] Krizhevsky A. Learning multiple layers of features from tiny images [MS. Thesis]. Toronto: University of Toronto, 2009.

[10] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Proc. of the 2016 IEEE Conf. on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778. [doi: 10.1109/CVPR.2016.90]

[11] Karatzas D, Gomez-Bigorda L, Nicolaou A, Ghosh S, Bagdanov A, Iwamura M, Matas J, Neumann L, Chandrasekhar VR, Lu SJ, Shafait F, Uchida S, Valveny E. ICDAR 2015 competition on robust reading. In: Proc. of the 13th Int'l Conf. on Document Analysis and Recognition. Tunis: IEEE, 2015. 1156–1160. [doi: 10.1109/ICDAR.2015.7333942]

[12] Liao MH, Wan ZY, Yao C, Chen K, Bai X. Real-time scene text detection with differentiable binarization. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI, 2020. 11474–11481. [doi: 10.1609/aaai.v34i07.6812]

[13] Lyu PY, Liao MH, Yao C, Wu WH, Bai X. Mask TextSpotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. In: Proc. of the 15th European Conf. on Computer Vision. Munich: Springer, 2018. 71–88. [doi: 10.1007/978-3-030-01264-9_5]

[14] He KM, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision. Venice: IEEE, 2017. 2980–2988. [doi: 10.1109/ICCV.2017.322]

[15] Zhou XY, Yao C, Wen H, Wang YZ, Zhou SC, He WR, Liang JJ. EAST: An efficient and accurate scene text detector. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 2642–2651. [doi: 10.1109/CVPR.2017.283]

[16] Tian Z, Huang WL, He T, He P, Qiao Y. Detecting text in natural image with connectionist text proposal network. In: Proc. of the 14th European Conf. on Computer Vision. Amsterdam: Springer, 2016. 56–72. [doi: 10.1007/978-3-319-46484-8_4]

[17] Ren SQ, He KM, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. In: Proc. of the 28th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2015. 91–99.

[18] Liao MH, Shi BG, Bai X, Wang XG, Liu WY. Textboxes: A fast text detector with a single deep neural network. In: Proc. of the 31st AAAI Conf. on Artificial Intelligence. San Francisco: AAAI, 2017. 4161–4167. [doi: 10.1609/aaai.v31i1.11196]

[19] Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC. SSD: Single shot multibox detector. In: Proc. of the 14th European Conf. on Computer Vision. Amsterdam: Springer, 2016. 21–37. [doi: 10.1007/978-3-319-46448-0_2]

[20] Qin XG, Zhou Y, Guo YH, Wu DY, Tian ZH, Jiang N, Wang HB, Wang WP. Mask is all you need: Rethinking mask R-CNN for dense and arbitrary-shaped scene text detection. In: Proc. of the 29th ACM Int'l Conf. on Multimedia. Association for Computing Machinery, 2021. 414–423. [doi: 10.1145/3474085.3475178]

[21] Dai P, Zhang S, Zhang H, *et al*. Progressive contour regression for arbitrary-shape scene text detection. In: Proc. of the 2021 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 7389–7398. [doi: 10.1109/CVPR46437.2021.00731]

[22] Peng SD, Jiang W, Pi HJ, Li XL, Bao HJ, Zhou XW. Deep snake for real-time instance segmentation. In: Proc. of the 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020. 8530–8539. [doi: 10.1109/CVPR42600.2020.00856]

[23] Liao MH, Zou ZS, Wan ZY, Yao C, Bai X. Real-time scene text detection with differentiable binarization and adaptive scale fusion. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2023, 45(1): 919–931. [doi: 10.1109/TPAMI.2022.3155612]

[24] He T, Shen CH, Tian Z, Gong D, Sun CM, Yan YL. Knowledge adaptation for efficient semantic segmentation. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 578–587. [doi: 10.1109/CVPR.2019.00067]

[25] Liu YF, Chen K, Liu C, Qin ZC, Luo ZB, Wang JD. Structured knowledge distillation for semantic segmentation. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 2599–2608. [doi: 10.1109/CVPR.2019.00271]

[26] Zhang LF, Song JB, Gao AN, Chen JW, Bao CL, Ma KS. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 3712–3721. [doi: 10.1109/ICCV.2019.00381]

[27] Hou YN, Ma Z, Liu CX, Loy CC. Learning lightweight lane detection CNNs by self attention distillation. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 1013–1021. [doi: 10.1109/ICCV.2019.00110]

[28] Zheng ZH, Ye RG, Hou QB, Ren DW, Wang P, Zuo WM, Cheng MM. Localization distillation for object detection. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2023, 45(8): 10070–10083. [doi: 10.1109/TPAMI.2023.3248583]

[29] Zhou SC, Liu WZ, Hu C, Zhou SC, Ma C. UniDistill: A universal cross-modality knowledge distillation framework for 3D object detection in bird's-eye view. In: Proc. of the 2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Vancouver: IEEE, 2023. 5116–5125. [doi: 10.1109/CVPR52729.2023.00495]

[30] Zagoruyko S, Komodakis N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv:1612.03928, 2016.

[31] Chen JW, Yang F, Lai YX. A self-distillation approach via entropy transfer for scene text detection. Acta Automatica Sinica, 2023, 49(11): 1−12 (in Chinese with English abstract). [doi: 10.16383/j.aas.c210598]

[32] Yang P, Zhang F, Yang G. A fast scene text detector using knowledge distillation. IEEE Access, 2019, 7: 22588–22598. [doi: 10.1109/ACCESS.2019.2895330]

[33] Yang P, Yang G, Gong X, Wu PP, Han Xu, Wu JS. Instance segmentation network with self-distillation for scene text detection. IEEE Access, 2020, 8: 45825–45836.[doi: 10.1109/ACCESS.2020.2978225]

[34] Bolya D, Zhou C, Xiao FY, Lee YJ. YOLACT: Real-time instance segmentation. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 9156–9165. [doi: 10.1109/ICCV.2019.00925]

[35] Yang CL, Xie LX, Su C, Yuille AL. Snapshot distillation: Teacher-student optimization in one generation. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 2854–2863. [doi: 10.1109/CVPR.2019.00297]

[36] Zhang Y, Xiang T, Hospedales TM, Lu HC. Deep mutual learning. In: Proc. of the 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 4320–4328. [doi: 10.1109/CVPR.2018.00454]

[37] Ahn S, Hu SX, Damianou A, Lawrence ND, Dai ZW. Variational information distillation for knowledge transfer. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 9155–9163. [doi: 10.1109/CVPR.2019.00938]

[38] Chen HT, Wang YH, Xu C, Yang ZH, Liu CJ, Shi BX, Xu CJ, Xu C, Tian Q. Data-free learning of student networks. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 3513–3521. [doi: 10.1109/ICCV.2019.00361]

[39] Kwon K, Na H, Lee H, Kim NS. Adaptive knowledge distillation based on entropy. In: Proc. of the 2020 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing. Barcelona: IEEE, 2020. 7409–7413. [doi: 10.1109/ICASSP40776.2020.9054698]

[40] Vu TH, Jain H, Bucher M, Cord M, Pérez P. ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: Proc. of the 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 2512–2521. [doi: 10.1109/CVPR.2019.00262]

[41] Lin TY, Dollár P, Girshick R, He KM, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 936–944. [doi: 10.1109/CVPR.2017.106]

[42] Karatzas D, Shafait F, Uchida S, Iwamura M, Bigorda LGI, Mestre SR, Mas J, Mota DF, Almazàn JA, de las Heras LP. ICDAR 2013 robust reading competition. In: Proc. of the 12th Int'l Conf. on Document Analysis and Recognition. Washington: IEEE, 2013. 1484–1493. [doi: 10.1109/ICDAR.2013.221]

[43] Yao C, Bai X, Liu WY, Ma Y, Tu ZW. Detecting texts of arbitrary orientations in natural images. In: Proc. of the 2012 IEEE Conf. on Computer Vision and Pattern Recognition. Providence: IEEE, 2012. 1083–1090. [doi: 10.1109/CVPR.2012.6247787]

[44] Yao C, Bai X, Liu WY. A unified framework for multioriented text detection and recognition. IEEE Trans. on Image Processing, 2014, 23(11): 4737–4749. [doi: 10.1109/TIP.2014.2353813]

[45] Ch'ng CK, Chan CS. Total-Text: A comprehensive dataset for scene text detection and recognition. In: Proc. of the 14th IAPR Int'l Conf. on Document Analysis and Recognition. Kyoto: IEEE, 2017. 935–942. [doi: 10.1109/ICDAR.2017.157]

[46] He WH, Zhang XY, Yin F, Liu CL. Multi-oriented and multi-lingual scene text detection with direct regression. IEEE Trans. on Image Processing, 2018, 27(11): 5406–5419. [doi: 10.1109/TIP.2018.2855399]

[47] Howard A, Sandler M, Chen B, Wang WJ, Chen LC, Tan MX, Chu G, Vasudevan V, Zhu YK, Pang RM, Adam H, Le Q. Searching for MobileNetV3. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 1314–1324. [doi: 10.1109/ICCV.2019.00140]

[48] Shu CY, Liu YF, Gao JF, Yan Z, Shen CH. Channel-wise knowledge distillation for dense prediction. In: Proc. of the 2021 IEEE/CVF Int'l Conf. on Computer Vision. Montreal: IEEE, 2021. 5291–5300. [doi: 10.1109/ICCV48922.2021.00526]

[49] Wang YK, Zhou W, Jiang T, Bai X, Xu YC. Intra-class feature variation distillation for semantic segmentation. In: Proc. of the 16th European Conf. on Computer Vision. Glasgow: Springer, 2020. 346–362. [doi: 10.1007/978-3-030-58571-6_21]

[50] Cho JH, Hariharan B. On the efficacy of knowledge distillation. In: Proc. of the 2019 IEEE/CVF Int'l Conf. on Computer Vision. Seoul: IEEE, 2019. 4793–4801. [doi: 10.1109/ICCV.2019.00489]

附中文参考文献:

[5] 高晗, 田育龙, 许封元, 仲盛. 深度学习模型压缩与加速综述. 软件学报, 2021, 32(1): 68–92. http://www.jos.org.cn/1000-9825/6096.htm [doi: 10.13328/j.cnki.jos.006096]

[31] 陈建炜, 杨帆, 赖永炫. 一种基于信息熵迁移的文本检测模型自蒸馏方法. 自动化学报, 2023, 49(11): 1−12. [doi: 10.16383/j.aas.c210598]

陈建炜(1996－), 男, 硕士生, 主要研究领域为文本检测, 知识蒸馏.

杨帆(1982－), 男, 博士, 副教授, 主要研究领域为机器学习, 数据挖掘.

沈英龙(2000－), 男, 硕士生, 主要研究领域为机器学习, 异常检测.

赖永炫(1982－), 男, 博士, 教授, CCF 专业会员, 主要研究领域为大数据分析和管理, 边缘计算.