自动化学报
Acta Automatica Sinica
ISSN 0254-4156,CN 11-2109/TP

《自动化学报》网络首发论文

题目：一种基于信息熵迁移的文本检测模型自蒸馏方法
作者：陈建炜，杨帆，赖永炫
DOI：10.16383/j.aas.c210598
收稿日期：2021-06-29
网络首发日期：2022-06-09
引用格式：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[J/OL]．自动化学报. https://doi.org/10.16383/j.aas.c210598

一种基于信息熵迁移的文本检测模型自蒸馏方法

陈建炜 [1] 杨 帆 [1, 2] 赖永炫 [3]

摘 要 前沿的自然场景文本检测方法大多基于全卷积语义分割网络 (Fully convolutional network, FCN), 利用分割网络输出的像素级分类结果, 有效检测任意形状的文本. 该类方法的主要缺陷是模型规模大, 前向推理时间过长, 占用较大内存, 这在实际应用中限制了其部署. 为达到模型容量和效率的折衷, 本文提出一种面向文本检测模型的知识蒸馏方法. 知识蒸馏在图像分类任务中作为模型压缩和提升模型精度的技巧而广泛使用, 然而知识蒸馏需要提前训练庞大的教师网络, 会增加训练成本, 并且由于师生网络的学习能力差异, 造成知识迁移效率下降; 同时, 传统的知识蒸馏方法多用于图像分类任务, 当应用到文本检测模型时无法取得令人满意的性能. 本文提出了一种基于信息熵迁移的自蒸馏训练方法 (Self-distillation via entropy transfer, SDET), 将文本检测网络的深层网络输出的分割图的信息熵作为待迁移的知识, 通过一个辅助网络将其直接反馈给浅层网络, 从而提升模型对文本框边缘的关注; SDET 同时利用信息熵和标签信息监督浅层网络的训练, 最终提升网络的性能. 与依赖教师网络的知识蒸馏不同, SDET 仅在训练阶段增加一个辅助网络, 以微小的额外训练代价实现无需教师网络的自蒸馏. 在 TD500、TD-TR、ICDAR2013、ICDAR2015、Total-Text 和 CASIA-10K 六个自然场景文本检测的标准数据集上的实验表明 SDET 能显著提升基线文本检测网络的召回率和 F1 得分, 且优于其它蒸馏方法.

关键词 自然场景, 文本检测, 知识蒸馏, 自蒸馏, 信息熵

引用格式 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, 2022, **48**(x): 1−13

DOI 10.16383/j.aas.c210598

A Self-Distillation Approach via Entropy Transfer for Scene Text Detection

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

Abstract Most of the state-of-the-art text detection methods in natural scenes are based on full convolutional network (FCN), which can effectively detect arbitrary shape text by using the pixel level classification results from the segmentation network. The main defects of these methods, i.e. large size of the networks, time-consuming of forward reasoning and large memory occupation, hinders their deployment in practical applications. In order to achieve a tradeoff between model capacity and efficiency, in this paper we propose a knowledge distillation method for text detection model. Knowledge distillation is widely used in image classification tasks as an approach of model compression and for improving model accuracy. However, traditional knowledge distillation usually needs to pretrain a large teacher network, which increases the training cost and may reduce the efficacy of knowledge transfer due to the difference of learning ability between teacher and student. Meanwhile, existing knowledge distillation methods, which are mostly used for image classification, cannot achieve satisfactory performance when applied to text detection models. In this paper, we propose a novel self-distillation method, Self-distillation via entropy transfer (SDET), which takes the information entropy of the segmentation map output by the deep layers of the text detection network as the knowledge to be transferred, and feeds it directly back into the shallow layers through an auxiliary network, so as to enhance the model's attention on the edge of textbox. In sumary, SDET simultaneously uses information entropy and label information to supervise the training of shallow layers, and ultimately improves the performance of the network. Different from traditional knowledge distillation which relies on teacher network, SDET utilizes an auxiliary network in the training stage and realizes self-distillation at a small extra training cost. Experiments on TD500, TD-TR, ICDAR2013, ICDAR2015, Total-Text and CASIA-10K datasets demonstrate that SDET can significantly improve the recall rate and F1 score of baseline text detection networks, and outperforms other

收稿日期 2021-06-29 录用日期 2022-02-10
Manuscript received June 29, 2021; accepted February 10, 2022
科技创新 2030— "新一代人工智能"重大项目 (2021ZD0112600), 国家自然科学基金委员会面上项目 (62173282, 61872154), 广东省自然科学基金 (2021A1515011578), 深圳市基础研究专项面上项目 (JCYJ20190809161603551) 资助
Supported by the National Key R & D Program of China (2021ZD0112600), National Natural Science Foundation of China (62173282, 61872154), Natural Science Foundation of Guangdong Province (2021A1515011578), the Shenzhen Fundamental Research Program (JCYJ20190809161603551)

本文责任编委
Recommended by Associate Editor

1. 厦门大学航空航天学院自动化系 厦门 361005 2. 厦门大学深圳研究院 深圳 518057 3. 厦门大学信息学院软件工程系 厦门 361005
1. Department of Automation, School of Aerospace Engineering, Xiamen University, Xiamen, 361005 2. Shenzhen Research Institute, Xiamen University, Shenzhen, 518057 3. Department of Software Engineering, School of Informatics, Xiamen University, Xiamen, 361005

distillation methods.

Key words Natural scene, text detection, knowledge distillation, self-distillation, information entropy

Citation Chen Jian-Wei, Yang Fan, Lai Yong-Xuan. A self-distillation approach via entropy transfer for scene text detection. *Acta Automatica Sinica*, 2022, **48**(x): 1−13

近年来, 自然场景文本理解广泛应用于自动导航和定位、手机拍照识别和智能安防等, 吸引了大批计算机视觉研究人员的关注. 文本检测作为场景文本理解中重要的一步, 直接影响后续文本识别的准确率. 随着深度全卷积网络[1] 在语义分割方面取得重大进展[2], 越来越多场景文本检测方法采用语义分割作为基本检测框架, 如掩码文本检测器 (Mask textspotter[3]) 修改实例分割网络掩码区域卷积神经网络 (Mask region convolutional neural network, Mask R-CNN[4]) 的掩码 (Mask) 分支以实现更加准确的字符分割. 得益于全卷积网络对图像上每一个像素点的分类能力, 基于分割的文本检测模型更有利于检测出弯曲、多方向等复杂场景文本. 然而为了提高检测精度, 该类模型往往规模庞大, 例如在多个数据集上取得最高性能的文本聚合网络 (TextFuseNet[5]) 使用 101 层的深度残差网络 (101-layer residual net, ResNet-101[6]) 提取图像的多级特征, 这导致前向推理需要花费更多时间, 且占据较大存储空间, 不利于部署在计算资源有限或者有实时性要求的场景, 例如智能手机、智能眼镜、无人驾驶汽车等. 为了减小模型规模同时保持较高检测精度, 研究者目前采取的一种主流方法是知识蒸馏 (Knowledge distillation, KD[7]). 由于其思路简单直接且在实践中被证明有效, 知识蒸馏不仅常用于模型压缩, 也被广泛应用于提升小规模网络的性能.

知识蒸馏也被称为 “师生学习”(Teacher-student learning, TSL), 主要思想是将一个较大规模的教师网络的知识迁移给一个紧凑的学生网络. 经典的知识蒸馏方法[7] 将教师网络预测类别的概率分布作为训练学生网络的软目标 (Soft target, ST), 通过带有 “温度”超参数的 Softmax 函数来控制软目标的平滑程度, 最后在软目标和硬目标 (如独热 (One-hot) 标签) 的同时监督下, 学生网络泛化能力得到提升. 知识蒸馏在图像分类任务上[7−10] 已经得到了广泛而成功的应用, 但是当我们**将传统基于师生网络的知识蒸馏方法应用到自然场景文本检测模型**上时, 发现存在以下三个问题:

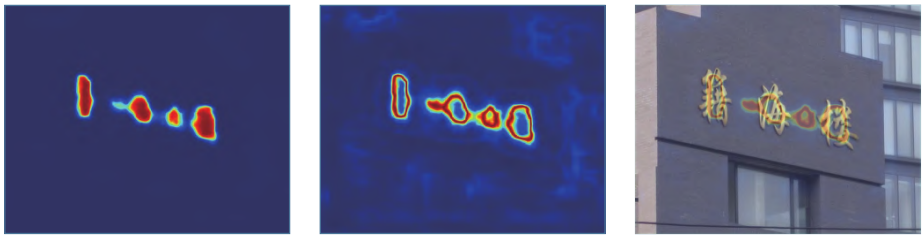
1) 学生网络常常不能通过对教师网络的学习达到理想精度, 例如在数据集 ICDAR 2015[11] 和 Total-Text[12] 上. 传统知识蒸馏方法存在 “教学效率”的问题[13]: 随着数据集的增大, 学生和教师网络之间学习能力的差异越来越显著, 这导致教师网络的知识难以被学生网络充分吸收, 因此在较大数据集上传统的知识蒸馏方法普遍效果不佳.

2) 传统的知识蒸馏方法是两阶段进行的, 必须提前训练教师模型, 再把知识迁移到学生模型. 为获得性能优越的教师网络 (通常规模较大), 需要花费大量时间来训练和调参.

3) 已有的文本检测网络的知识蒸馏研究[14] 仅仅是将现有的图像分类中的知识蒸馏方法直接应用到文本检测模型中, 没有考虑文本检测模型自身输出信息的特点.

不同于图像分类, 文本检测模型更关注文本边缘的像素点信息. 以基于分割的文本检测网络作为研究对象, 该类检测模型都会输出对每一个像素点属于文本的概率值. 从信息熵角度分析分割模型输出的分割图 (Segmentation map, SM), 概率值的高低反映模型的置信度. 对抗熵最小化的语义分割领域适应方法 (Adversarial entropy minimization, AdvEnt[15]) 提到在源域上训练的语义分割模型输出的分割图置信度高, 熵值低, 而对目标域的图像预测不准确, 输出高熵值. 除了领域差 (Domain gap) 造成信息熵值的差异, 对于基于分割的文本检测网络而言, 其中心和边缘同样存在显著的信息熵差. 如图 1(a) 模型仅对文本中心附近区域 (红色区域) 有较高的概率预测值, 而边缘区域概率值低. 本文将模型预测的每一个像素点的概率值转换为信息熵, 则边缘区域的信息熵高, 如图 1(b) 信息熵图所示外围红色区域, 而中心区域熵值低 (包裹的蓝色区域). 图 1(c) 将信息熵图和原图叠加, 可发现熵值图能有效放大模型对边缘的注意力, 因此本文认为, 分割图的信息熵作为蒸馏知识, 能更有效地提升网络检测文本边缘的能力.

综上, 本文针对文本检测网络提出了一种基于信息熵迁移的自蒸馏训练方法 (Self-distillation via entropy transfer, SDET), 克服传统师生网络必须提前训练教师网络的不足并且充分利用文本检测结果的信息熵. SDET 从深监督[16] 和自我注意力蒸馏[17] 获得灵感: 对于一个文本检测模型的网络结构而言, 网络深层的分类器由于抽取到更加抽象的语义特征, 因此预测的结果比浅层更加确定; 而浅层获得的特征细节虽然更丰富, 但是预测的准确性不如深层分类器, 两者信息熵存在差异. 因此 SDET



(a) 分割图　(b) 信息熵图　(c) 叠加图

(a) Segmentation map　(b) Entropy map　(c) Overlay map

图 1　DB 文本检测网络的分割图和信息熵图可视化

Fig. 1　Segmentation map and entropy map visualization of DB text detection network

让网络深层通过信息熵引导网络浅层的训练以达到知识迁移的目的. 具体来说, SDET 通过在网络的浅层部分连接一个辅助分类器 (Auxiliary network), 将网络深层的信息熵作为网络浅层的训练目标. 从师生学习的角度来看, 深层可被视为教师模型, 而浅层则看作学生模型, 因此 SDET 是一种自蒸馏方法. 需要注意的是, 引入的辅助分类器仅存在于训练阶段, 使用时可删除辅助分类器, 因此并不影响文本检测模型的推理速度.

本文的贡献可归纳为以下三点: 1) 本文将自蒸馏方法应用于文本检测模型, 首次提出一种基于信息熵的自蒸馏方法 (SDET). SDET 以网络深层的信息熵来监督网络浅层的学习, 通过促进浅层网络学习文本框边缘信息来提升网络的精度, 从而避免了训练一个大规模的教师网络; 2) 与传统的知识蒸馏方法相比, SDET 不仅节约了教师网络的训练代价, 而且能更有效地提升网络精度; 值得注意的是, SDET 无需精细地调参, 在 ICDAR 2013、TD500、TD-TR、Total-Text、ICDAR2015 和 CASIA-10K 六个标准数据集上的对比实验表明, 使用默认参数的 SDET 显著优于其他六种知识蒸馏方法, 这进一步增加了其使用的方便性; 3) 多个标准数据集上的实验进一步表明 SDET 可适用于不同架构和不同规模的文本检测网络, 同时也优于深监督方法.

1　相关工作

1.1　基于深度学习的文本检测

基于深度学习的自然场景文本检测[18] 大体可分为两类: 基于边界框回归和基于图像分割.

基于边界框回归的方法受目标检测框架的启发, 利用目标检测算法 (如更快速区域卷积神经网络 (Faster region convolutional neural network, Faster R-CNN[19])、单发多盒检测器 (Single shot multibox detector, SSD[20]) 等) 产生候选文本框, 经过非极大值抑制 (Non-maximum suppression, NMS) 后处理 (Post-process) 获得最终文本实例. Liao 等[21] 提出端到端识别的文本盒 (TextBoxes) 算法, 将 SSD 中的默认框统一设置成长条形, 取消正方形的边框, 其卷积核由 3×3 替换成 1×5, 以适应文本行的特点; 为检测出不同大小的文本框, 与 SSD 类似地引入多尺度训练. Tian 等[22] 认为文本检测和目标检测的不同点在于文本行大都是水平而且连续, 因此提出基于连接的文本建议网络 (Connectionist text proposal network, CTPN) 算法, 在 Faster R-CNN 的基础上将文本行分割成宽度固定的小建议框, 以提高检测精度. Zhou 等[23] 提出一种快速准确的文本检测器 (Efficient and accurate scene text detector, EAST), 采用'U'形的全卷积网络[24], 自上而下合并特征图, 训练目标由分割图的类别平衡交叉熵损失和几何形状损失, 同时调节分类损失和几何损失的权重参数. 例如标注形式为四边形和旋转角的, 则几何损失采用交并比 (Intersection over union, IOU) 损失, EAST 消除了以往文本检测的区域建议等步骤, 提高检测速度.

基于图像分割的方法是目前主流的文本检测方法. 该方法通过全卷积网络[1] 结构对图像的每一个像素做分类, 更有利于检测出复杂背景下的任意形状文本. 如 Liao 等[25] 除了在分割图上和二值图使用二元交叉熵损失外, 巧妙地使用可微二值化 (Differentiable binarization, DB) 的方法解决文本检测后处理阈值难以选择的问题, 即添加阈值图的 L1 损失, 其中三者损失函数的权重系数依次为 1、1 和 10, 由此简化文本检测的后处理, 进一步提高文本检测的精度和速度. Wang 等[26] 提出渐进式尺度扩展网络 (Progressive scale expansion network, PSENet), 通过从最小核逐渐扩展到最大尺寸的文本示例, 有效解决基于分割的算法不能分离相邻的、过于接近的文本的问题; Ye 等[5] 使用 Mask R-CNN 提取字符和单词级别的特征, 并额外引入一

个语义分割分支以获取图像全局特征, 再通过多路径融合网络合并字符级、单词级和全局级特征, 产生更准确的文本检测结果; Wang 等[27] 在轻量级主干网络上级联多个特征金字塔增强模块 (Feature Pyramid Enhancement Module, FPEM), 使得不同层次的特征更具有判别力, 并使用特征融合模块 (Feature fusion module, FFM) 汇聚不同层次的特征形成最终的特征用于预测文本区域. Xu 等为检测不规则的场景文本, 提出文本场 (TextFiled[28]) 的文本检测方法, 在图像分割的基础上引入了方向场 (Direction field) 的概念, 其中场的方向表示像素点的相对位置, 长度代表像素点为文本的概率, 有效检测弯曲文本.

1.2 知识蒸馏与自蒸馏

知识蒸馏最早是由 Hinton[7] 提出, 用来从大网络 (教师网络) 迁移知识到小网络 (学生网络), 以提高学生网络的学习能力. 早期的知识蒸馏[7] 经过软化的全连接层 (Fully connected layer, FC) 输出的 logits 当作教师网络的知识, 定义该类知识为软目标. Romero 等[9] 扩展了知识蒸馏的形式, 认为迁移中间特征图同样有利于学生网络的学习. Zagoruyko[10] 通过让学生网络模仿教师网络中间特征图的注意力图来提高学生网络的性能, 其中注意力图编码了教师网络中间层特征图的信息, 因而比直接迁移中间特征图有更好的效果. He[29] 为了解决学生网络和教师网络迁移特征的不一致, 使用预先训练的自编码器将教师网络的特征输出到潜在空间, 经过压缩的特征更容易让学生网络学习. Liu 等[30] 充分考虑语义分割任务中图像上的每一个像素点与周围像素的关联性或者结构性, 提出结构化知识蒸馏 (Structured knowledge distillation, SKD), 其中结构化知识包括教师网络特征图的相似性和通过对抗式学习策略获得的更高层次的结构信息. Wang[31] 等提出类内特征变化蒸馏 (Intra-class feature variation distillation, IFVD), 以每一个像素特征到其类别中心的相似性表征类内特征变化 (Intra-class feature variation, IFV), 替代结构化知识蒸馏 (SKD) 的逐像素点的成对相似性, 更有利于学生网络模仿教师网络的特征变化.

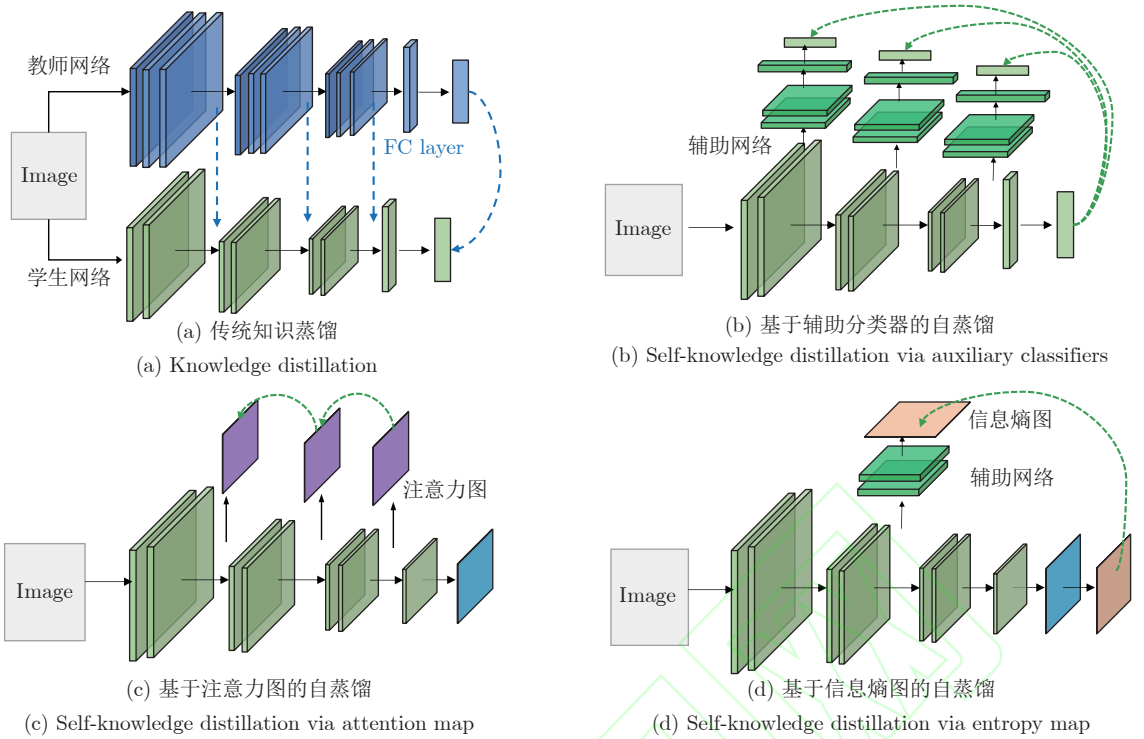
最近的研究[32] 已经将知识蒸馏扩展到自蒸馏. 自蒸馏是让模型学习自身的知识, 其学生网络和教师网络其实是同一个网络, 其最大的好处是避免训练一个规模较大的教师网络. 例如 Zhang 等[32] 提出先将卷积神经网络按照深度划分成几个浅层部分, 每个浅层都设置一个分类器, 在训练阶段从最深层分类器的提炼出软目标和特征图, 迁移到每个浅层分类器, 按照知识蒸馏的概念[7] 可以将最深层分类器视为教师模型, 浅层分类器作为学生模型. Hou[17] 提出了自注意力蒸馏方法, 认为模型中提取的注意力图会编码丰富的上下文信息, 经过逐层蒸馏 (即浅层的网络模仿更深层网络的注意力图), 增强了车道线检测模型的表示学习.

本文首次提出将基于信息熵的自蒸馏用于文本检测模型. 图 2 展示了本文所提出的 SDET 方法与其它主要知识蒸馏方法的框架: 图 2(a) 是传统的学生 − 教师网络框架的知识蒸馏, 图 2(b) 表示使用辅助分类器实现图像分类网络的自蒸馏[32], 图 2(c) 通过提炼自我注意图, 实现车道线分割网络的自蒸馏[17], 图 2(d) 展示了本文提出的以信息熵为迁移目标的文本检测自蒸馏方法 SDET.

图 2(a) 中的方法以转移师生网络的软目标和特征图匹配为基础, 必须预训练一个精度高的教师网络 (通常规模较大), 而其它三种自蒸馏方法仅靠网络自身来提炼知识, 省去了教师网络的构建和训练. SDET 和图 2(b) 所示方法[32] 类似, 都是基于辅助分类器实现自蒸馏, 不同之处在于图 2(b) 方法[32] 中包含了四个辅助分类器, 每一个辅助分类器的训练目标由三部分构成, 分别是最深层分类器的软目标损失、图像标签的交叉熵损失和最深层分类器的中间特征图 L2 损失. 而与其相比, SDET 只需一个辅助分类器, 监督信息只包含最深层分类器的信息熵, 也只需要一个超参数用来平衡原始模型的检测损失和转移信息熵的损失 (后续实验表明该超参数可设为 1 即可取得满意性能). SDET 和图 2(c) 所示的基于注意力的自蒸馏方法 (Self attention distillation, SAD[17]) 的相似点在于它们都使用网络深层信息监督网络浅层的学习, 不同之处在于 SAD 需要在相邻层间构造多层注意力图, 而 SDET 只关注深层的转移分割图的信息熵, 并使用了一个额外的辅助分类器.

2 基于信息熵迁移的自蒸馏方法

图 3 展示了本文所设计的 SDET 方法的整体框架. 可分为两个部分: 1) 基于语义分割网络的文本检测基线模型 (Baseline model), 其特点是模型输出对每一个像素点是否为文本的二分类结果; SDET 将基线模型输出的分割图转换成信息熵图来监督自蒸馏模块和浅层网络. 2) 自蒸馏模块 (Self-distillation module), 实质上是一个辅助分类器网络 (Auxiliary network); 在蒸馏训练中辅助网络输出的概率图将转换成信息熵图, 而在检测阶段时可



(a) 传统知识蒸馏
(a) Knowledge distillation

(b) 基于辅助分类器的自蒸馏
(b) Self-knowledge distillation via auxiliary classifiers

(c) 基于注意力图的自蒸馏
(c) Self-knowledge distillation via attention map

(d) 基于信息熵图的自蒸馏
(d) Self-knowledge distillation via entropy map

图 2　不同知识蒸馏方法对比
Fig. 2　Comparison of different knowledge distillation methods

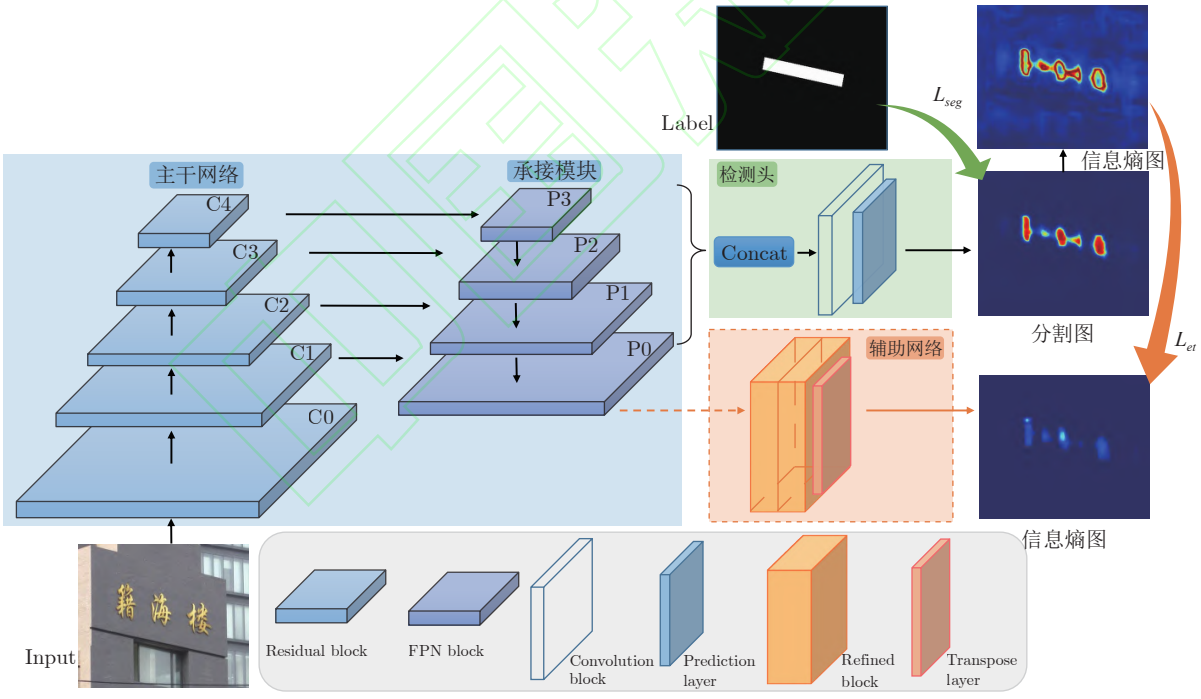


图 3　SDET 训练框架
Fig. 3　SDET training framework

移除自蒸馏模块, 如图 3 虚线框表示. 如前所述, 可以认为深层网络是教师网络, 辅助网络和浅层网络构成一个学生网络, 它们之间传递的知识是文本边缘信息. 由于提炼的知识来自网络自身, 因此将该方法称为信息熵迁移的自蒸馏方法 (Self-distillation via entropy transfer, SDET).

2.1 基线模型

基于语义分割的文本检测框架可分成三个部分:

1) 主干网络 (Backbone), 常用移动端高效卷积神经网络 (MobileNet[33]) 或者残差网络 (Residual net, ResNet[6]) 等图像分类卷积网络, 负责抽取图像特征;

2) 承接模块 (Neck), 常用特征金字塔 (Feature pyramid networks, FPN[34]), 聚合不同层次的特征;

3) 检测头 (Detection head, DH), 主要作用是预测图像上每一个像素点属于文本的概率.

一个基于分割的文本检测网络把一张 H × W× 3 (高为H, 宽为W 的 RGB 三通道) 的图片 $\boldsymbol{I}$ 作为输入, 经过主干网络的特征抽取, 得到不同层次 (C0 ∼ C4 层) 的特征; 特征金字塔自上而下聚合 C1 到 C4 层的特征, 输出融合低层和高层信息的多层 (P0 ∼ P3 层) 特征, 将这些特征拼接成特征图 $\boldsymbol{M}$, 输入到检测端网络, 最终计算得到尺寸为H× W × 2 (通道数 2 表示输出为 "文本"和 "背景"的二分类结果) 的分割图 $\boldsymbol{P}$. 其检测头的损失函数 (Detection head loss, L_{dh}) 为:

$$L_{\mathrm{dh}} = L_s + \lambda \times L_o \tag{1}$$

其中 L_s 表示图像上每一个像素点分类损失, L_o 表示其它部分的损失, 如文献 [25] 中采用可微二值损失, 文献 [23] 中采用几何损失, 在此不再赘述. λ 为平衡两者之间的超参数.

2.2 自蒸馏模块

自蒸馏模块仅在训练阶段使用, 推理阶段完全丢弃, 不会影响文本检测. 正如图 3 中展示, 把自蒸馏模块加入文本检测模型是简单和直接的, 只需要把特征金字塔输出的结果输入到辅助分类器, 后续实验表明特征金字塔从何处连接辅助分类器取决于对应位置特征图大小.

自蒸馏方法与深监督网络 (Deeply-supervised nets, DSN[16]) 类似, 同样在主干网络的某一分支引入辅助网络, 但和深监督不同之处在于, 其辅助网络的监督信号仅仅来自于网络后半部分的信息熵而不是图像的标签. 深监督广泛运用于图像分类[16]、语义分割网络[2] 等, 它通过训练额外的辅助分类器来提高网络泛化性能和加快网络收敛, 但其辅助网络的结构并没有统一的设计方法. 因此自蒸馏的重点是设计合适的辅助网络. 实验中发现, 结构不合适的辅助网络蒸馏效果欠佳, 因而本文提出通过各种精炼卷积块 (Refined block[35]) 构造适合主干网络的辅助网络, 不断提炼和组合输入的特征图, 以期获得令人满意的蒸馏效果. 图 4 给出三种辅助网络, 它们适应于不同网络规模的主干网络和不同架构的文本检测分割头, 在消融实验中我们将具体分析辅助网络对自蒸馏的影响.

将辅助网络的输入特征图记为 $\boldsymbol{FM}$ (Feature map), 网络的输出是分割图, 记为 $\boldsymbol{SM}$ (Segmentation map), 网络中核心部分 (图 4 中阴影) 记为 RF(Refine function), 则辅助网络可统一表达为:

$$\boldsymbol{f}_I = Conv(\boldsymbol{FM}) \tag{2}$$

$$\boldsymbol{f}_R = RF(\boldsymbol{f}_I) \tag{3}$$

$$\boldsymbol{SM} = Upsample(\sigma(Conv(\boldsymbol{f}_R))) \tag{4}$$

式 (2) − (4) 表示, 从特征金字塔输出的特征 $\boldsymbol{FM}$ 经过简单卷积过滤抽取, 得到 RF 的输入特征 (Input features, $\boldsymbol{f}_I$); 核心模块 RF 对 $\boldsymbol{f}_I$ 进一步组合

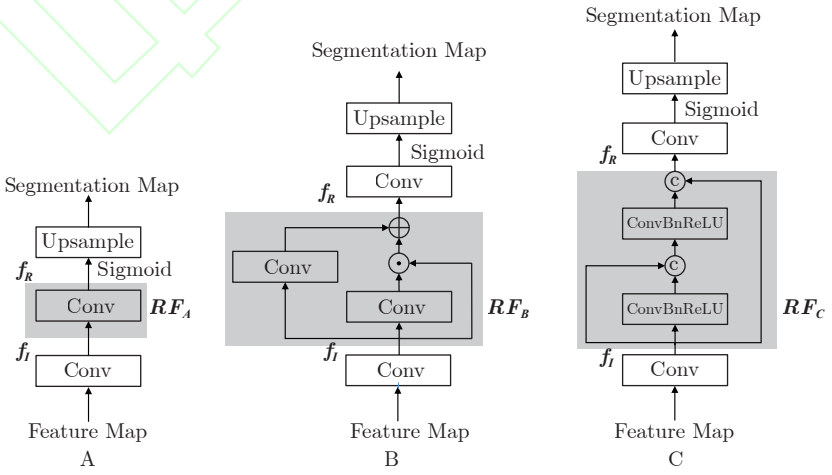


图 4 辅助网络的三种结构形式

Fig. 4 The three types of auxiliary networks

特征得到细化特征 (Refine features, $\boldsymbol{f}_R$), 经过卷积运算将 $\boldsymbol{f}_R$ 通道数降为 2, 使用 Sigmoid 函数输出概率, 最后插值放大 (Upsample) 到原图大小, 得到模型对输入图片的文本和背景的预测结果. 不同辅助网络的主要区别体现在 RF 上, RF 负责将特征 $\boldsymbol{f}_I$ 提炼成更加精细的特征 $\boldsymbol{f}_R$, 如图 4 所示, 按照网络的复杂程度可划分为三种类型:

A 型 RF 模块是直接使用 3×3 卷积,

$$RF_A(\boldsymbol{f}_I) = Conv(\boldsymbol{f}_I) \tag{5}$$

B 型 RF 计算公式来自于文献 [35], 先使用 3×3 卷积压缩输入特征 $\boldsymbol{f}_I$ 的通道数, 然后将压缩后的特征图与 $\boldsymbol{f}_I$ 做乘积, 将乘积的值与先前另一分支上 3×3 卷积的结果求和, 得到 $\boldsymbol{f}_R$.

$$RF_B(\boldsymbol{f}_I) = (Conv_1(\boldsymbol{f}_I)\odot\boldsymbol{f}_I)\oplus Conv_2(\boldsymbol{f}_I) \tag{6}$$

C 型 RF 的设计灵感来源于文献 [24], 其核心思想是自上而下逐级融合拼接不同特征层次的特征; 与文献 [24] 不同, 式 (7) 中低层特征 f_I 总是参与拼接运算 (concat), 并且提炼特征的过程加入了批归一化层 (Batch normalization, BN) 和 ReLU 激活函数.

$$\begin{aligned} RF_C(\boldsymbol{f}_I) = concat(\boldsymbol{f}_I,\ & ConvBnReLU_1 \\ & (concat(\boldsymbol{f}_I,\ ConvBnReLU_2(\boldsymbol{f}_I)))) \end{aligned} \tag{7}$$

2.3　损失函数

我们将文本检测模型的主干网络的检测头记为 d, 输出的分割图记为 $\boldsymbol{P}_d$, 把自蒸馏模块的辅助分类器记为 a, 输出的分割图记为 $\boldsymbol{P}_a$. 根据香农熵的定义, 某一个位于坐标 (h, w) 的像素点对应的信息熵可根据其属于文本的概率 $P^{(h,w,0)}$ 和属于背景的概率 $P^{(h,w,1)}$ 定义为

$$\begin{aligned} E^{(h,w)} = & -(P^{(h,w,0)}\times\log_2(P^{(h,w,0)}) \\ & + P^{(h,w,1)}\times\log_2(P^{(h,w,1)})), \end{aligned} \tag{8}$$

则深层网络和辅助网络的信息熵图分别用 $E_d^{(h,w)}$ 和 $E_a^{(h,w)}$ 表示.

为了鼓励辅助网络输出分割图的信息熵与检测头的分割图的信息熵一致, SDET 最小化其信息熵迁移损失 (Entropy transfer loss, L_{et}), 即最小化下式:

$$L_{et} = \frac{1}{\mathrm{H}\times\mathrm{W}}\sum_{h,w}\left|E_d^{(h,w)} - E_a^{(h,w)}\right| \tag{9}$$

因此, 训练总损失 (Loss, L) 包括文本检测的损失 L_{dh} 和自蒸馏损失 L_{et}:

$$L = L_{\mathrm{dh}} + \gamma\times L_{et} \tag{10}$$

其中 γ 是平衡文本检测和自蒸馏的超参数, 在本文所有实验中, 我们将 γ 固定设为 1, 即可取得满意的效果, 无需额外调参.

2.4　训练方法

如算法 1 所示, 训练时, 基线模型和自蒸馏模块中的辅助网络同时优化更新. 输入批量图像数据, 分别经过基线模型和辅助网络, 各自预测出图像上的每一个像素点的概率值, 按照式 (8) 将其转化为信息熵, 通过最小化式 (10), 同时训练基线模型和辅助网络. 训练终止条件是模型迭代次数达到预设的次数. 测试阶段断开辅助网络与特征融合网络 FPN 的连接, 仅评测基线模型的检测头输出的结果.

算法 1 SDET 训练流程

Algorithm 1: Self-distillation via entropy transfer

Input: D_{train}: Training Data;
$d(\cdot;\theta_d)$: Text Detection Model;
$a(\cdot;\theta_a)$: Auxiliary Network

Output: the optimized parameters θ_d^*, θ_a^*

Initialize
$\theta_d \leftarrow$ kaiming_normal (θ_d);
$\theta_a \leftarrow$ kaiming_normal (θ_a);

for *epoch* $= 1$ **to** *epochs* **do**
  foreach *minibatch* B in D_{train} **do**
    $P_d,\ FM \leftarrow d(B;\theta_d)$;
    $P_a \leftarrow a(B;\theta_a)$;
    Calculate entropy map E_d and E_a;
    Compute self-distillation loss and total loss L;
    Update θ_d using $\nabla_{\theta_d}L$;
    Update θ_a using $\nabla_{\theta_a}L$;
  end
end

3　实验

3.1　数据集

本文实验中使用了文本检测研究常用的 6 个标准数据集:

ICDAR2013[36]: 总共有 462 张图片, 训练集图片有 229 张, 其余 233 张用于测试集. 文字都是英文的且水平对齐, 提供字符级和单词级标注.

TD500[37]: 有 300 张图片用于训练集, 其余 200 张是测试集. TD500 数据集具有任意方向的矩形文本框, 包含中英文, 以行为单位标注.

TD-TR: 我们参考相关文献 [23, 25] 中的做法, 将 HUST-TR400[38] 数据集里 400 张图片添加到

TD500 的训练数据集, 因此 TD-TR 数据集有 700 张训练集, 测试集 200 张.

ICDAR 2015[11]: 总共有 1500 张图片, 分成训练集 1000 张和 500 张测试集. 由于使用谷歌眼镜拍摄的街边图片, 因此图像模糊, 分辨率仅为 720×1280.

Total-Text[12]: Total-Text 具有任意方向的不同形状文本, 包括水平的矩形文本和弯曲的文本形状等. 其中训练集 1255 张, 测试集 300 张, 标注单位是单词.

CASIA-10K[39]: 数据集共有 10000 张图像, 其中 7000 张图像用于训练, 3000 张图像用于测试. 数据集采集自中文场景, 每个文本行标注其四个顶点坐标.

3.2 评价指标

根据 ICDAR 2015 评价方法[11], 我们用信息检索领域的精确率 (Precision, P)、召回率 (Recall, R) 和 F1 得分 (F1-score, F) 来综合评估文本检测算法的性能. 计算 P 和 R 依赖于交并比 (Intersection over union, IoU). IoU 由第 i 个检测的矩形框 D_i 和第 j 个标签 (Ground-truth, GT) G_j 的之间交集和并集的比值定义; 如果 $IoU \geq 0.5$, 该检测结果是正确的. 具体定义 IoU 计算式为

$$IoU = \frac{area(G_j \cap D_i)}{area(G_j \cup D_i)} \tag{11}$$

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. 根据检测结果的 IoU 可以统计出正确检测的矩形框集合 $\boldsymbol{T}_p$, 则精确率 P、召回率 R 和 F 可定义如下,

$$P = \frac{|\boldsymbol{T}_P|}{|\boldsymbol{D}|} \tag{12}$$

$$R = \frac{|\boldsymbol{T}_P|}{|\boldsymbol{G}|} \tag{13}$$

$$F = \frac{2 \times P \times R}{P + R} \tag{14}$$

3.3 实验设置

实验目的是评价本文所提出的自蒸馏方法, 以及和其它蒸馏方法进行对比, 因此对于所有使用到的基线模型, 包括 λ 在内的所有超参数均按其原文献推荐的最优超参数设置, 以使其性能达到最优. 在此基础上加入知识蒸馏, 探索其是否能够进一步提升基线模型的性能. 对于自蒸馏模块, 主要超参数为 γ. 如 2.3 所述, 本文中简单设定 γ =1, 后续实验表明即可取得满意性能而无需精细调参.

在消融实验中, 我们使用主干网络为 MobileNetV3 的 EAST 模型来分析自蒸馏算法的影响元素. 除此之外的其它对比实验均采用可微二值化分割头的 DB 网络模型, 分别采用 MobilieNetV3 和 ResNet50 作为主干网络 (Backbone). 如果没有特别说明, 图像的数据增强采用随机旋转 (−10°, 10°) 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. 优化器采用随机梯度下降 (Stochastic gradient descent, SGD), 并且使用多项式学习率调整策略, 训练主干网络是 MobileNetV3, 批大小 (Batch size) 设置为 8, 训练 1200 轮, 而训练 ResNet50 时, 批大小设置为 4, 同样训练 1200 轮. 所有实现基于 Pytorch 平台, 在单张 1080Ti 显卡上训练.

3.4 消融实验

使用自蒸馏方法需要考虑如何设计合适的辅助分类器以及在特征金字塔的哪个特征层次连接分类器. 首先, 为探索不同辅助网络设计对自蒸馏的影响, 我们比较了图 4 中 3 种辅助分类器设计 (用-A, -B, -C 表示) 对 SDET 的影响. ICDAR 2013 和 ICDAR 2015 数据上的实验结果如表 1 中所示, 其中 MV3-EAST/DB 表示主干网络采用 MoblieNetV3, 分割头使用 EAST 或者 DB 的文本检测模型, Method 列中的 SDET-A/B/C 表示使用何种辅助分类器. 实验结果表明, 对同一个基线模型, 采用不同的辅助分类器的 SDET 性能存在差异, 例如对于 MV3-EAST, 简单 A 型抑制了 SDET 的作用, 而稍复杂的 B 型和 C 型都能不同程度上提升 baseline 的 *F1-score*; 不同模型对辅助分类器有所偏好, 如 MV3-DB 更适合用 A 型, 而不适应对 MV3-EAST 有较大提升的 B 型, 这可能是因为不同的模型对特征抽取组合不同. 总的来说, C 型辅助分类器较为鲁棒, 均能有效提升 MV3-EAST 和 MV3-DB 基线模型的 F1-score. 其它数据上的实验结果基本一致.

其次, 从主干网络提取的特征往往需要经过特征金字塔这类特征融合模块, 它们融合高层抽象的语义信息和底层的细节信息, 再输出不同层次的特征, 如 P0 ∼ P3. 因此可以连接辅助分类器的位置共有 4 个, 我们用 MV3-EAST 作为 baseline 模型, 在 ICDAR 2015 数据集上考虑不同的位置对 SDET 的影响. 实验结果如表 2, 其中 SDET-Px 表示将辅助网络连接在 Px 位置上. 观察结果, 可以得出在 P2 和 P3 位置放置辅助分类器有利于 SDET, P0 和 P1 则会抑制 SDET 的训练. 其它数据上的实

表 1　不同辅助分类器对 SDET 的影响 (%)
Table 1　The impact of different auxiliary classifiers on SDET(%)

Model	Method	ICDAR 2013			ICDAR 2015		
		P	R	F	P	R	F
MV3-EAST	baseline	81.7	64.4	72.0	80.9	75.4	78.0
	SDET-A	78.8	65.9	71.8	78.8	76.3	77.5
	SDET-B	84.4	66.5	**74.4**	81.3	77.0	**79.1**
	SDET-C	81.4	67.4	73.7	78.9	77.7	78.3
MV3-DB	baseline	83.7	66.0	73.8	87.1	71.8	78.7
	SDET-A	84.1	68.8	**75.7**	86.5	73.9	**79.7**
	SDET-B	81.1	67.3	73.6	87.8	71.7	78.9
	SDET-C	84.9	67.9	75.4	87.8	73.0	**79.7**

表 2　不同特征金字塔位置对 SDET-B 的影响 (%)
Table 2　The impact of different feature pyramid positions on SDET-B(%)

Method	Feature map size	P	R	F
baseline		80.9	75.4	78.0
SDET-P0	16 × 16	79.1	75.8	77.4
SDET-P1	32 × 32	79.5	76.5	78.0
SDET-P2	64 × 64	80.7	77.4	79.0
SDET-P3	128 × 128	81.3	77.0	**79.1**

验表现是一致的. 可能原因是 P2 和 P3 的特征图尺寸较为合适, 保留了足够多的信息, 而 P0 和 P1 的特征图缺乏检测需要的底层细节信息, 因此效果略差. 因而, 我们可根据不同的主干网络抽取特征的能力不同来选择相应的金字塔位置, 实验中, 我们发现对于轻量级网络, 如 MobileNetV3, 可以选择 P3 或者 P2 位置, 而对于主干网络为 ResNet50 这样的大网络, 将辅助分类器连接到 P1 层.

3.5　和主流蒸馏方法对比

我们将 SDET 和目前主流的六种蒸馏方法在 ICDAR2013 等数据集上进行对比, 比较其在测试集上的精度. 这六种蒸馏方法分成两大类, 分别是传统的学生 − 教师框架的知识蒸馏方法, 即 ST[7]、FitNets[9]、KA[29] 和 SKD[30], 另一类则是近年来流行的自蒸馏方法 SD[32] 和 SAD[17]. 其中, ST 表示转移教师网络输出的软化概率值; FitNets 通过 L2 范数最小化教师 − 学生网络的中间特征图, 特征图通道不一致时, 使用 1×1 卷积转化; KA 使用卷积编码器作为特征适配器来实现教师网络与学生网络特征之间的适配; SKD 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, 其次匹配特征图对应的相似性矩阵; SD 在特征金字塔的每一层连接辅助分类器, 浅层分类器训练目标包括标签信息和最深层分类器的软目标; SAD 以特征金字塔的深层部分的注意图当作浅层的蒸馏目标, 例如 P1 层的注意力图的蒸馏目标是 P2 层.

实验中, 学生模型 (baseline 模型) 的主干网络采用 MobileNetV3, 分别使用可微二值化和 EAST 作为最后的文本检测分割头, 教师网络则是将主干网络替换为 ResNet-50.

从表 3 和表 4 可以明显看出本文所提出的自蒸馏方法 SDET 在不同规模的数据集下均能提高基线 DB 和 EAST 模型的 F1 综合指标, 并取得了最佳表现 (**加粗字体为该数据上 8 种方法中的最高 F-score**). 尤其是在 ICDAR 2013 数据集上, 相较于 baseline 的学生网络, 经过自蒸馏训练的 DB 模型在精确率、召回率和 F1 分别有 0.4%、2.8% 和 1.9% 的提升, 同样 SDET 有效提升 EAST 模型的 F1, 从 72.0% 提高到 74.4%. 图 5(见下页顶部) 通过 3 个真实图像上的文本检测效果直观展示了 SDET 对基线模型 (学生网络) 的性能提升, 其中第 (a) 行的图像中用方框展示了文本所在位置 (即真实标签 Ground truth), 第 (b) 行中的方框展示了基线模型对三幅图像的检测结果, 用圆框凸显和真

表 3　MV3-DB 在不同数据集上的知识蒸馏实验结果 (%)
Table 3　Experimental results of knowledge distillation of MV3-DB on different datasets (%)

Method	ICDAR 2013			TD500			TD-TR			ICDAR 2015			Total-Text			CASIA-10K		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
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[7]	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[29]	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[9]	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[30]	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
SD[32]	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[17]	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
ours	84.1	68.8	**75.7**	80.6	72.2	**76.2**	85.6	74.6	**79.7**	86.5	73.9	**79.7**	87.5	68.4	**76.8**	87.4	53.4	**66.3**

表 4 MV3-EAST 在不同数据集上的知识蒸馏实验结果 (%)

Table 4 Experimental results of knowledge distillation of MV3-EAST on different datasets (%)

Method	ICDAR 2013			ICDAR 2015			CASIA-10K		
	P	R	F	P	R	F	P	R	F
baseline	81.7	64.4	72.0	80.9	75.4	78.0	66.1	64.9	65.5
ST[7]	77.8	64.9	70.8	80.9	75.1	77.9	64.7	65.1	64.9
KA[29]	78.6	64.0	70.5	78.2	76.4	77.3	67.7	63.0	65.3
FitNets[9]	82.4	65.8	73.2	78.0	77.8	77.9	65.4	64.2	64.8
SKD[30]	79.5	66.3	72.3	81.9	75.6	78.6	66.6	64.7	65.6
SD[32]	80.2	63.8	71.1	79.6	74.7	77.1	66.2	63.5	64.8
SAD[17]	81.4	65.6	72.6	80.2	76.5	78.3	65.7	64.1	64.9
ours	84.4	66.5	**74.4**	81.3	77.0	**79.1**	70.8	63.0	**66.7**

实框有显著差异之处. 第 (c) 行中的方框展示 SDET 训练后的模型检测结果. 从图 5 中可以发现, (b) 中基线模型的预测结果存在检测边缘漏判、不完全或误判的情况, 而自蒸馏训练的网络检测出的结果相对精确和鲁棒, 仅将最后一幅图中的茶杯手柄误测为字母'D'.

图 5 SDET 与基线模型的检测结果对比: (a) 真实标签; (b) 基线模型检测结果; (c)SDET 训练后的模型检测结果

Fig. 5 Comparison of detection results between SDET and baseline models: (a)ground-truth; (b)detection results of baseline models; (c) detection results of models trained with SDET

对比表 3 其它蒸馏方法在不同规模数据集上的实验结果, 可以观察对于 MV3-DB 学生网络, 其它蒸馏方法难以有一致性的稳定提高. 例如在小数据集 TD 500 上, 传统蒸馏方法能在不同程度上提升学生模型的 F1 指标, 然而在大一些的数据集上如 ICDAR 2015、Total-Text 和 CASIA-10K 上, 绝大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, 如图像分类任务中常用的 ST 方法. 根据文献 [13] 中关于知识蒸馏效率的研究, 可能的原因是教师网络和学生网络之间的学习能力差距随着训练数据集的增大而增大, 尤其是在 CASIA-10K 这类难度更大的数据集上, 学习能力差距更加明显, 导致传递知识的效率降低. 观察表 4 展示的对于 MV3-EAST 学生网络的结果也会发现除了 SKD、SAD 和本文所提 SDET 方法外, 其它蒸馏方法也缺乏一致性的性能提升, 而 SDET 提升最为显著.

总的来说, 本文提出的自蒸馏方法 SDET 在没有训练一个教师网络的情况下, 在多个数据集上效果都超出其它蒸馏方法, 且无需额外调参. 由于训练一个合适的教师网络不仅需要较大内存, 还需要耗费大量时来调参, 相对比下自蒸馏框架可大大节约内存和时间两方面的需求, 还能带来令人满意的性能提升, 因而具有很大的优势.

3.6 和深监督方法对比

SDET 方法和深监督网络 (Deeply supervised nets, DSN[16]) 相似, 两者都需要外接辅助分类器, 主要的不同点在于 SDET 辅助分类器的监督信号来自深层分类器预测结果的信息熵, 而 DSN 则来自标签信息. 为了对比两者的效果, 我们在数据集 ICDAR2013、TD500、TD-TR、ICDAR 2015、Total-Text 和 CASIA-10K 上分别用 SDET 和 DSN 两种方式来训练主干网络为 MobileNetV3、分割头是 DB 的文本检测基线模型. 从表 5(见下页顶部, **加粗字体为该数据上三种方法中的最高 F-score**) 中可以发现 SDET 在各数据集上都能取得更好的性能.

在 SDET 中, 浅层分类器的学习目标由标签改成深层分类器预测结果的信息熵, 这种方式能提高

表 5 DSN 与 SDET 在不同数据集上的对比 (%)
Table 5 Comparison of DSN and SDET on different datasets (%)

Method	ICDAR 2013			TD500			TD-TR			ICDAR 2015			Total-Text			CASIA-10K		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
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
DSN[16]	84.4	68.0	75.3	79.7	71.5	75.4	86.4	72.2	78.7	85.8	73.4	79.1	86.1	67.9	75.9	87.9	52.3	65.6
ours	84.1	68.8	**75.7**	80.6	72.2	**76.2**	85.6	74.6	**79.7**	86.5	73.9	**79.7**	87.5	68.4	**76.8**	87.4	53.4	**66.3**

表 6 SDET 在不同数据集上提升 RES50-DB 的效果 (%)
Table 6 The effect of SDET on improving RES50-DB on different datasets (%)

Method	ICDAR 2013			TD500			TD-TR			ICDAR 2015			Total-Text			CASIA-10K		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
baseline	86.3	72.9	79.0	84.1	75.9	79.8	87.3	80.4	83.7	90.3	80.1	84.9	87.7	79.4	83.3	90.1	64.7	75.3
ours	82.7	77.2	**79.9**	79.9	81.5	**80.7**	87.2	83.0	**85.0**	90.3	82.1	**86.0**	87.4	81.8	**84.5**	86.0	68.7	**76.4**

网络性能的一个可能原因是信息熵具备更多的信息量. 从图 1 可以看出信息熵放大了模型对边缘的注意力; 另一方面, 相较于 DSN 中固定不变的标签信息, SDET 深层分类器的信息熵随着训练迭代不断地动态调整, 浅层分类器也可随之动态地学习, 一个直观的理解是其学习过程从易到难, 逐步提高难度.

3.7 推广到大网络

传统的知识蒸馏方法常用在较小的学生网络上, 如果希望应用到较大的学生网络上, 则必须训练一个比学生网络大得多的教师网络. 例如训练 Backbone 为 ResNet50 的网络, 可能会需要训练 ResNet101 当作教师网络. 而自蒸馏仅靠传递自身的知识, 无需训练庞大的教师网络, 优势显著. 我们用主干网络为 ResNet50 的 DB 模型当作基线模型, 在 ICDAR2013 等 6 个数据集上运用 SDET 进行自蒸馏, 其主干网络直接加载 Pytorch 上预训练的 ResNet50, 未使用可变卷积, 算法测试时, 输入图像统一为 736×736, 结果如表 6 所示.

从表 6 可以观察到 SDET 能有效提升 ResNet50 的性能表现, 在六个数据集上 F1-score 均有提升 (**加粗显示**), 在数据集 TD-TR、ICDAR 2015、Total-Text 和 CASIA-10K 上均有超过 1 个百分点的提高. 对比基线模型可以发现, 精确率 P 并没有改善, F1-score 的提升来自于 SDET 显著地提升了模型的召回率 R, 在各个数据集分别提高了 4.3%、5.6%、2.6%、2%、2.4% 和 4.0%. 根据式 (12) 和 (14) 可知召回率提升反映了有效检出 $|\boldsymbol{T}_P|$ 值增大, 可能原因是浅层分类器经过来自深层分类器的信息熵监督训练, 促进网络学习边缘的知识, 从而检测边缘更加准确, 使得 IoU 普遍增大, $|\boldsymbol{T}_P|$ 也随之提高.

4 结论

本文提出了一种基于信息熵迁移的自蒸馏训练方法 (SDET), 用于自然场景文本检测模型. SDET 无需提前训练教师网络, 仅在训练阶段添加一个辅助网络来传递信息熵的知识以提高文本检测模型的性能, 能够在很大程度上节约内存和训练时间. 在六个流行的标准数据集上的对比实验表明 SDET 无需精细的调参过程即可提升不同规模大小的基线模型 (如 MV3-DB 和 ResNet50-DB), 比已有的知识蒸馏方法和深监督方法更具有优势. SDET 的不足之处在于不能用于仅有边界框回归的文本检测算法, 例如 CTPN[22], 因为该类网络没有输出对每个像素点的概率预测, 因而不能计算信息熵. 其次, 本文仅设计了 3 种较简单的辅助网络, 而不同的文本检测网络可能需要不同的辅助网络, 未来我们将探索神经网络结构搜索与 SDET 的结合, 通过自动调整辅助网络的结构以寻找最优的辅助网络.

References

1 Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE, 2015. 3431−3440

2 Yuan Y H, Chen X L, Wang J D. Object-contextual representations for semantic segmentation. arXiv: 1909.11065, 2019.

3 Lyu P Y, Liao M H, Yao C, Wu W H, Bai X. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. In: Proceedings of the European Conference on Computer Vision. Munich, Germany: Springer, 2018. 67−83

4 He K M, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 2961−2969

5 Ye J, Chen Z, Liu J H, Du B. TextFuseNet: Scene text detection with richer fused features. In: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence. Yokohama, Japan: ijcai. org, 2020. 5165−22

6 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, Nevada, USA: IEEE, 2016. 770−778

7 Hinton G E, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv: 1503.02531, 2015.

8 Lai Xuan, Qu Yan-Yun, Xie Yuan, Pei Yu-Long. Topology-guided adversarial deep mutual learning for knowledge distillation. *Acta Automatica Sinica*, 2021, **x**(x): 1−9
(赖轩, 曲延云, 谢源, 裴玉龙. 基于拓扑一致性对抗互学习的知识蒸馏. 自动化学报, 2021, **x**(x): 1−9)

9 Romero A, Ballas N, Kahou S E, Chassang A, Gatta C, Bengio Y. Fitnets: Hints for thin deep nets. arXiv: 1412.6550, 2014.

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

11 Karatzas D, Gomez-Bigorda L, Nicolaou A, Ghosh S, Bagdanov A, Iwamura M, et al. ICDAR 2015 competition on robust reading. In: In: Proceedings of the 13th International Conference on Document Analysis and Recognition. Nancy, France: IEEE, 2015. 1156−1160

12 Ch'ng C K, Chan C S. Total-text: A comprehensive dataset for scene text detection and recognition. In: Proceedings of the 14th International Conference on Document Analysis and Recognition. Kyoto, Japan: IEEE, 2017. 935−942

13 Cho J H, Hariharan B. On the efficacy of knowledge distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, South Korea: IEEE, 2019. 4794−4802

14 Yang P, Yang G W, Gong X, Wu P P, Han X, Wu J S, Chen C S. Instance segmentation network with self-distillation for scene text detection. *IEEE Access*, 2020, **8**: 45825−45836

15 Vu T-H, Jain H, Bucher M, Cord M, Pérez P. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE, 2019. 2517−2526

16 Lee C Y, Xie S N, Gallagher P, Zhang Z Y, Tu Z W. Deeply-supervised nets. In: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics. San Diego, California, USA: PMLR, 2015. 562−570

17 Hou Y N, Ma Z, Liu C X, Loy C C. Learning lightweight lane detection cnns by self attention distillation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Seoul, South Korea: IEEE, 2019. 1013−1021

18 Wang Run-Min, Sang Nong, Ding Ding, Chen Jie, Ye Qi-Xiang, Gao Chang-Xin and Liu Li. Text detection in natural scene image: a survey. *Acta Automatica Sinica*, 2018, **44**(12): 2113−2141
(王润民, 桑农, 丁丁, 陈杰, 叶齐祥, 高常鑫, 刘丽. 自然场景图像中的文本检测综述. 自动化学报, 2018, **44**(12): 2113−2141)

19 Ren S Q, He K M, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. arXiv: 1506.01497, 2015.

20 Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y, Berg A C. Ssd: Single shot multibox detector. In: Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016. 21−37.

21 Liao M H, Shi B G, Bai X, Wang X G, Liu W Y. Textboxes: A fast text detector with a single deep neural network. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. San Francisco, California, USA: AAAI, 2017. 4161−4167

22 Tian Z, Huang W L, He T, He P, Qiao Y. Detecting text in natural image with connectionist text proposal network. In: Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands: Springer, 2016. 56−72

23 Zhou X Yu, Yao C, Wen H, Wang Y Z, Zhou S C, He W R, et al. East: an efficient and accurate scene text detector. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 2017. 5551−5560

24 Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Proceedings of the 2015 Medical Image Computing and ComputerAssisted Intervention. Munich, Germany: Springer, 2015. 234−241

25 Liao M H, Wan Z Y, Yao C, Chen K, Bai X. Real-time scene text detection with differentiable binarization. In: Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence. New York, USA: AAAI, 2020. 11474−11481

26 Wang W H, Xie E Z, Li X, Hou W B, Lu T, Yu G, Shao S. Shape robust text detection with progressive scale expansion network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE, 2019. 9336−9345

27 Wang W H, Xie E Z, Song X G, Zang Y H, Wang W J, Lu T, et al. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, South Korea: IEEE, 2019. 8440−8449

28 Xu Y C, Wang Y K, Zhou W, Wang Y P, Yang Z B, Bai X. Textfield: Learning a deep direction field for irregular scene text detection. *IEEE Transactions on Image Processing*, 2019, **28**(11): 5566−5579

29 He T, Shen C H, Tian Z, Gong D, Sun C M, Yan Y L. Knowledge adaptation for efficient semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE, 2019. 578−587

30 Liu Y F, Chen K, Liu C, Qin Z C, Luo Z B, Wang J D. Structured knowledge distillation for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE, 2019. 2604−2613

31 Wang Y K, Zhou W, Jiang T, Bai X, Xu Y C. Intra-class feature variation distillation for semantic segmentation. In: Proceedings of the European Conference on Computer Vision. Glasgow, UK: Springer, 2020. 346−362

32 Zhang L F, Song J B, Gao A, Chen J W, Bao C L, Ma K S. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, South Korea: IEEE, 2019. 3713−3722

33 Howard A, Sandler M, Chu G, Chen L C, Chen B, Tan M X, Wang W J, Zhu Y K, Pang R M, Vasudevan V. Searching for mobilenetv3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, South Korea: IEEE, 2019. 1314−1324

34 Lin T Y, Dollár P, Girshick R, He K M, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. Honolulu, HI, USA: IEEE, 2017. 2117−2125

35 Chen Z Y, Xu Q Q, Cong R M, Huang Q M. Global context-aware progressive aggregation network for salient object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA: AAAI, 2020. 10599−10606

36 Karatzas D, Shafait F, Uchida S, Iwamura M, I Bigorda L G, Mestre S R, et al. ICDAR 2013 robust reading competition. In: Proceedings of the 12th International Conference on Document Analysis and Recognition. Washington DC, USA: IEEE, 2013. 1484−1493

37 Yao C, Bai X, Liu W Y, Ma Y, Tu Z W. Detecting texts of ar-

bitrary orientations in natural images. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI, USA: IEEE, 2012. 1083−1090

38 Xue C H, Lu S J, Zhan F N. Accurate scene text detection through border semantics awareness and bootstrapping. In: Proceedings of the European conference on computer vision. Munich, Germany: IEEE, 2018. 355−372

39 He W H, Zhang X Y, Yin F, Liu C L. Multi-oriented and multilingual scene text detection with direct regression. *IEEE Transactions on Image Processing*, 2018, **27**(11): 5406−5419

陈建炜 厦门大学航空航天学院自动化系硕士研究生. 主要研究方向为计算机视觉, 图像处理.
E-mail: jianweichen@stu.xmu.edu.cn
(**CHEN Jian-Wei** Master student in the Department of Automation, School of Aerospace Engineering, Xiamen University. His research interest covers computer vision and image processing.)

杨 帆 厦门大学自动化系副教授. 主要研究方向为机器学习, 数据挖掘和生物信息学. 本文通信作者.
E-mail: yang@xmu.edu.cn
(**YANG Fan** Associate professor in the Department of Automation at Xiamen University. His research interest covers machine learning, data mining, and bioinformatics. Corresponding author of this paper.)

赖永炫 厦门大学信息学院软件工程系教授. 主要研究领域为大数据分析和管理, 智能交通, 深度学习, 车载网络.
E-mail: laiyx@xmu.edu.cn
(**LAI Yong-Xuan** Professor in the Software Engineering Department, School of Informatics, Xiamen University, China. His research interest covers big data management and analysis, intelligent transportation systems, deep learning, and vehicular networks.)