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, 利用像素级分类结果有效检测任意形状的文本, 其主要缺点是模型大、推理时间长、内存占用高, 这在实际应用中限制了其部署. 提出一种基于信息熵迁移的自蒸馏训练方法 (Self-distillation via entropy transfer, SDET), 利用文本检测网络深层网络输出的分割图 (Segmentation map, SM) 信息熵作为待迁移知识, 通过辅助网络将信息熵反馈给浅层网络. 与依赖教师网络的知识蒸馏 (Knowledge distillation, KD) 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 (Self-distillation, SD). 在多个自然场景文本检测的标准数据集上的实验结果表明, SDET 在基线文本检测网络的召回率和 F1 得分上, 能显著优于其他蒸馏方法.

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, 2024, **50**(11): 2128−2139

DOI 10.16383/j.aas.c210598 **CSTR** 32138.14.j.aas.c210598

Self-distillation 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, 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 forward reasoning and large memory occupation, hinder their deployment in practical applications. In this paper, we propose self-distillation via entropy transfer (SDET), which takes the information entropy of the segmentation map (SM) 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. Different from traditional knowledge distillation (KD) which relies on teacher network, SDET utilizes an auxiliary network in the training stage and realizes self-distillation (SD) at a small extra training cost. Experiments conducted on multiple standard datasets for natural scene text detection demonstrate that SDET significantly improves the recall rate and F1 score of the baseline text detection networks, and outperforms other distillation methods.

Key words Natural scene, text detection, knowledge distillation (KD), self-distillation (SD), information entropy

Citation Chen Jian-Wei, Yang Fan, Lai Yong-Xuan. Self-distillation via entropy transfer for scene text detection. *Acta Automatica Sinica*, 2024, **50**(11): 2128−2139

近年来, 自然场景文本理解广泛应用于自动导航与定位、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的重要一步, 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[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[5] 使用 101 层的深度残差网络[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

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 National Key Research and Development Program of China (2021ZD0112600), National Natural Science Foundation of China (62173282, 61872154), Natural Science Foundation of Guangdong Province (2021A1515011578), and Shenzhen Fundamental Research Program (JCYJ20190809161603551)
本文责任编委 金连文
Recommended by Associate Editor JIN Lian-Wen
1. 厦门大学航空航天学院 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 厦门 361005
1. School of Aerospace Engineering, Xiamen University, Xiamen 361005 2. Shenzhen Research Institute, Xiamen University, Shenzhen 518057 3. School of Informatics, Xiamen University, Xiamen 361005

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, 研究者们目前采取的一种主流方法是知识蒸馏 (Knowledge distillation, KD)[7]. 由于其思路简单和直接, 在实践中被证明是有效的. 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", 主要思想是将一个较大规模的教师网络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, 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] 已经获得了广泛而成功的应用, 但当将传统基于学生−教师网络的知识蒸馏方法应用到自然场景文本检测模型上时, 尚存在以下 3 个问题:

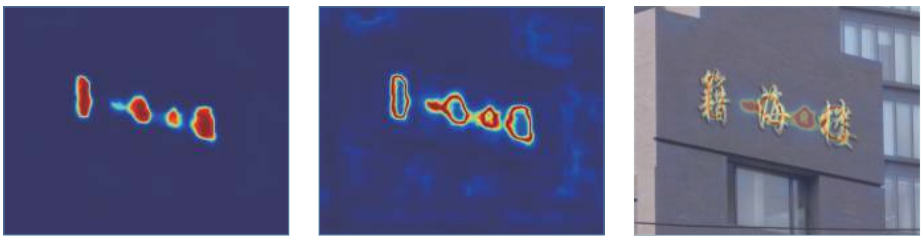
1) 学生网络常常不能通过对教师网络的学习达到理想精度, 例如在 ICDAR2015[11] 和 Total-text[12] 数据集上, 传统知识蒸馏方法存在"教学效率"问题[13], 随着数据集的增大, 学生和教师网络之间学习能力的差异越来越显著, 这导致教师网络的知识难以被学生网络充分吸收. 因此, 在较大数据集上, 传统知识蒸馏方法普遍效果不佳.

2) 传统知识蒸馏方法分两阶段进行, 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), 需要花费大量时间进行训练和调整参数.

3) 已有的文本检测网络的知识蒸馏研究[14] 仅将现有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. 在对抗熵最小化的语义分割领域适应方法[15] 中, 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, 但对目标域的图像预测不准确, 输出高熵值. 除了领域差造成信息熵值的差异, 对基于分割的文本检测网络, 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) 为信息熵图和原图叠加. 可以看出, 熵值图能有效放大模型对边缘的注意力, 因此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), 克服了传统学生−教师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, 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 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. 具体地, SDET 通过在网络的浅层部分连接一个辅助分类器, 将网络深层的信息熵作为网络浅层的训练目标. 从师生学习的角度看, 深层可被视为教师模型, 浅层则看作为学生模型, 因此 SDET 是一种自蒸馏方法 (Self-distillation, SD). 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



(a) 分割图　(b) 信息熵图　(c) 叠加图

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

图 1　可微二值化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 differentiable binarization text detection network

检测模型的推理速度.

本文的主要贡献有以下 3 点: 1) 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 无需精细地调整参数, 在 ICDAR2013、TD500、TD-TR、Total-text、ICDAR2015 和 CASIA-10K 六个标准数据集上的对比实验结果表明, 使用默认参数的 SDET 性能显著优于其他 6 种知识蒸馏方法. 3) 在多个标准数据集上的实验结果进一步表明, SDET 可适用于不同架构和不同规模的文本检测网络, 同时性能也优于深监督方法.

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] 大体可分为基于边界框回归和基于图像分割两类.

1) 基于边界框回归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 multi-box detector, SSD)[20] 等产生候选文本框, 经过非极大值抑制后处理获得最终文本实例. Liao 等[21] 提出端到端识别的文本盒算法, 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 替换为 1×5, 以适应文本行特点; 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”形全卷积网络[24], 自上而下合并特征图, 训练目标为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.

2) 基于图像分割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] 除了在分割图和二值图上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 L_1 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, 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或过于接近的文本问题; 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, 使得不同层次的特征更具有判别力, 并使用特征融合模块汇聚不同层次特征, 形成最终的特征, 用于预测文本区域. Xu 等[28] 为检测不规则的场景文本, 提出文本场的文本检测方法, 在图像分割的基础上引入了方向场概念, 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 等[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), 以提高小网络的学习能力. 早期的知识蒸馏[7] 经过软化的全连接层 (Fully connected layer, FC) 输出值作为教师网络知识, 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 等[10] 通过让学生网络模仿教师网络中间特征图的注意力图, 以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效果. He 等[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 等[31] 提出类内特征变化蒸馏, 以每个像素特征到其类别中心的

相似性表征类内特征变化, 替代结构化知识蒸馏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.

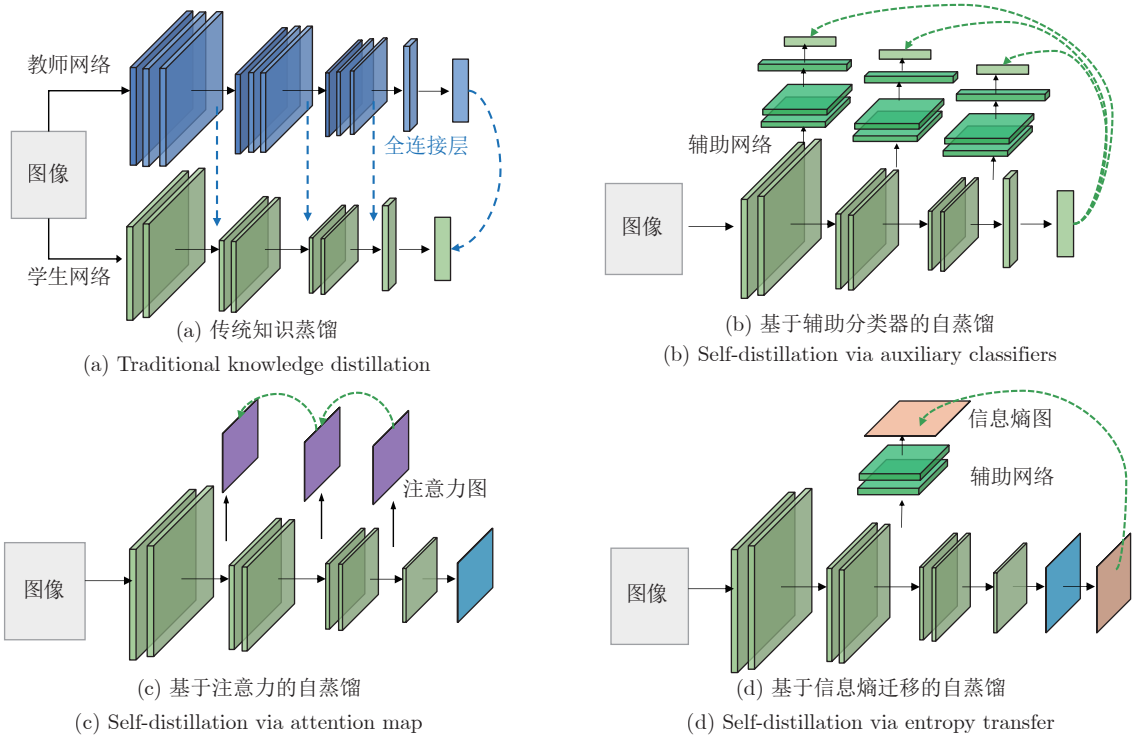
文献 [32] 已经将知识蒸馏扩展到自蒸馏. 自蒸馏是让模型学习自身的知识, 即学生网络和教师网络是同一个网络, 其最大好处是避免训练一个规模较大的教师网络. 例如 Zhang 等[32] 提出先将卷积神经网络按照深度划分为几个浅层, 每个浅层都设置一个分类器, 在训练阶段, 从最深层分类器提炼出软目标和特征图, 迁移到每个浅层分类器, 按照知识蒸馏概念[7] 可以将最深层分类器视为教师模型, 浅层分类器作为学生模型. Hou 等[17] 提出自注意力蒸馏方法, 认为模型中提取的注意力图会编码丰富的上下文信息, 经过逐层蒸馏即浅层的网络模仿更深层网络的注意力图, 增强了模型的表示学习能力.

本文首次提出将基于信息熵自蒸馏用于文本检测模型. 图 2 展示了本文提出的 SDET 方法与其他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) 是使用辅助分类器实现图像分类网络的自蒸馏; 图 2(c) 通过提炼自我注意力图, 实现车道线分割网络的自蒸馏; 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) 方法以转移学生–教师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), 而其他三种自蒸馏方法仅靠网络自身提炼知识, 省去了教师网络的构建和训练. SDET 与图 2(b) 方法类似, 都是基于辅助分类器实现自蒸馏, 不同之处在于图 2(b) 方法中包含了 4 个辅助分类器, 每个辅助分类器的训练目标由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 L_2 损失 3 个部分构成. 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, 只需要使用一个超参数, 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设为 1 即可取得满意性能). SDET 与图 2(c) 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 展示了本文设计的 SDET 方法的训练框架, 可分为以下 2 个部分: 1) 基于语义分割网络的文本检测基线模型. 特点是模型输出对每个像素点



(a) 传统知识蒸馏
(a) Traditional knowledge distillation

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

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

(d) 基于信息熵迁移的自蒸馏
(d) Self-distillation via entropy transfer

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

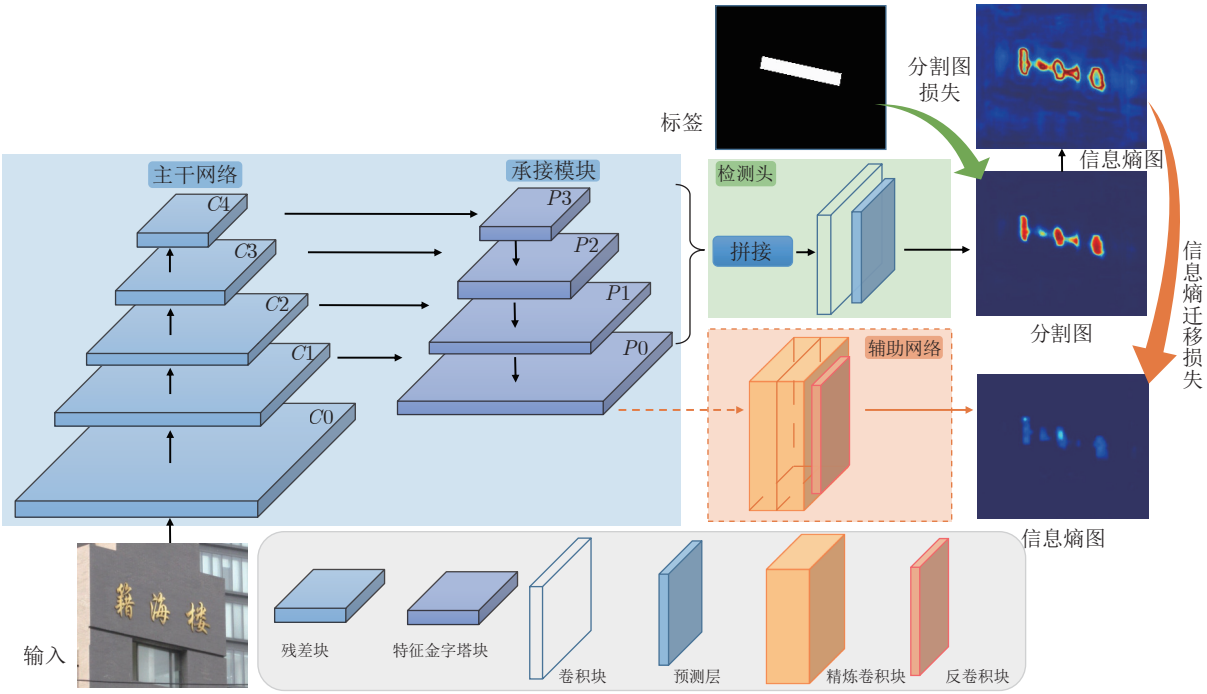


图 3 SDET 训练框架

Fig. 3 SDET training framework

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, 以监督自蒸馏模块和浅层网络. 2) 自蒸馏模块实质上是一个辅助分类器网络. 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, 可移除自蒸馏模块, 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, 因此称该方法为信息熵迁移的自蒸馏方法.

2.1 基线模型

基于语义分割的文本检测框架可分为以下 3 个部分:

1) 主干网络. 常用移动端高效卷积神经网络[33]或残差网络 (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. 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. 主要作用是预测图像上每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 $H \times W \times 3$ (高为 H 、宽为 W 的 RGB 三通道) 图片 I 作为输入, 经过主干网络的特征抽取, 得到不同层次 ($C0 \sim C4$) 的特征; 特征金字塔整合 $C1$ 到 $C4$ 层的特征, 输出融合低层和高层信息的多层 ($P0 \sim P3$) 特征, 将这些特征拼接成特征图 M, 输入到检测端网络, 最终计算得到尺寸为 $H \times W \times 2$ (通道数 2 表示输出为“文本”和“背景”的二分类结果) 的分割图 P. 其检测头的损失函数 L_{dh} 为:

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

式中, L_s 表示图像上每个像素点分类损失, L_o 表示其他部分的损失, 如文献 [25] 采用可微二值损失, 文献 [23] 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. 如图 3 所示, 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来自网络后半部分的信息熵, 而不是图像的标签. 深监督广泛应用于图像分类[16]、语义分割网络[2] 等领域, 它通过训练额外的辅助分类器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[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 给出辅助网络的 3 种结构形式, 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, 在第 3.4 节具体分析辅助网络对自蒸馏的影响.

辅助网络的输入特征图记为 FM, 网络输出的分割图记为 SM, 网络中核心部分 (图 4 中阴影部分) 记为 RF, 则辅助网络可统一表达为:

$$f_I = \mathrm{Conv}(FM) \tag{2}$$

$$f_R = RF(f_I) \tag{3}$$

$$SM = \mathrm{Upsample}(\sigma(\mathrm{Conv}(f_R))) \tag{4}$$

式 (2) ~ (4) 表示从特征金字塔输出的特征 FM 经过简单卷积过滤抽取, 得到 RF 的输入特征 f_I; 核心模块 RF 对 f_I 进一步组合特征得到细化特征 f_R, 经过卷积运算将 f_R 通道数降为 2, 使用 Sigmoid 函数输出概率 σ, 最后插值放大到原图大小, 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 f_I 提炼成更加精细的特征 f_R. 如图 4 所示, 按照网络的复杂程度可划分为 3 种类型:

A 型 RF 模型是直接使用 3×3 卷积:

$$RF_A(f_I) = \mathrm{Conv}(f_I) \tag{5}$$

B 型 RF 模型先使用 3×3 卷积压缩输入特征 f_I 的通道数, 然后将压缩后的特征图与 f_I 相乘, 再将乘积值与先前另一个分支上 3×3 卷积结果求和, 得到 f_R[35]:

$$RF_B(f_I) = (\mathrm{Conv}_1(f_I)\odot f_I)\oplus \mathrm{Conv}_2(f_I) \tag{6}$$

C 型 RF 模型核心思想是自上而下逐级融合拼接不同特征层次的特征[24]. 与文献 [24] 不同的是, 式 (7) 中低层特征 f_I 总是参与拼接运算, 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:

$$RF_C(f_I) = \mathrm{concat}(f_I, \mathrm{ConvBNReLU}_1(\mathrm{concat}(f_I, \mathrm{ConvBNReLU}_2(f_I)))) \tag{7}$$

2.3　损失函数

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 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 P_a. 根据香农熵定义, 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:

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

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 L_{et}, 即最小化下式:

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

因此, 训练包括文本检测损失 L_{dh} 和自蒸馏损失 L_{et} 的总损失 L:

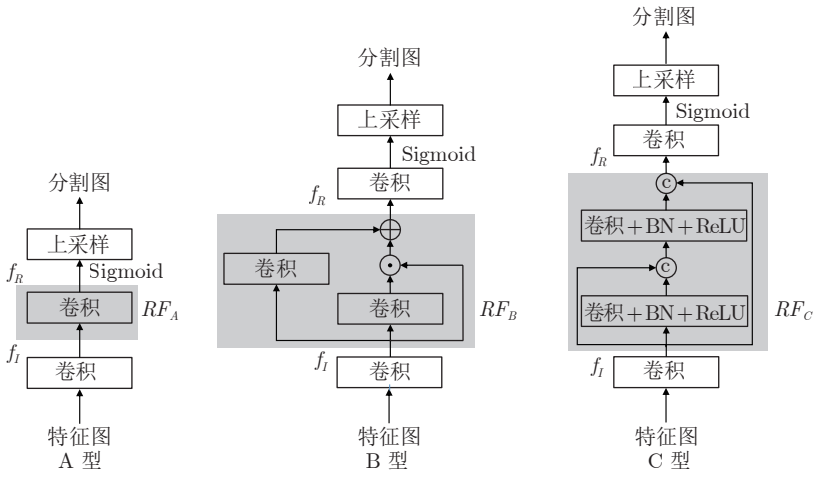


图 4　辅助网络的 3 种结构形式

Fig. 4　The three types of auxiliary networks

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

式中, γ 是平衡文本检测和自蒸馏的超参数. 在本文实验中, γ 设置为 1 即可取得满意效果, 无需额外调整参数.

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

输入. 训练数据集 D_{train}、文本检测模型 $d(\cdot;\theta_d)$、辅助网络 $a(\cdot;\theta_a)$.

输出. 文本检测模型、辅助网络的最优参数 θ_d^* 和 θ_a^*.

1) 初始化. 初始化检测模型参数 θ_d、辅助网络参数 θ_a;
2) for $epoch = 1$ to $epochs$ do;
3) for each $minibatch$ B in D_{train} do;
4) 检测模型前向传播;
5) 辅助网络前向传播;
6) 计算深层网络、辅助网络的信息熵 E_d 和 E_a;
7) 计算自蒸馏损失 L_{et} 和总损失 L;
8) 使用 $\nabla_{\theta_d} L$, 更新 θ_d;
9) 使用 $\nabla_{\theta_a} L$, 更新 θ_a;
11) end;
12) end.

3 实验

3.1 数据集

本文实验使用了文本检测研究常用的 6 个标准数据集:

1) ICDAR2013[36] 数据集共有 462 张图片, 其中 229 张图片用于训练集, 其余 233 张图片用于测试集. 该数据集文字都是英文且水平对齐, 提供字符级和单词级标注.

2) TD500[37] 数据集共有 500 张图片, 其中 300 张图片用于训练集, 其余 200 张图片用于测试集. 该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.

3) TD-TR 数据集. 参考文献 [23, 25], 将 HUST-TR400[38] 数据集中 400 张图片添加到 TD500 训练数据集中, 形成 TD-TR 数据集. TD-TR 数据集共有 900 张图片, 其中 700 张图片用于训练集, 其余 200 张图片用于测试集.

4) ICDAR2015[11] 数据集共有 1 500 张图片, 其中 1 000 张图片用于训练集, 其余 500 张图片用于测试集. 该数据集由于是使用谷歌眼镜拍摄的街边图片, 因此图像模糊, 分辨率仅为 $720 \times 1\,280$ 像素.

5) Total-text[12] 数据集共有 1 555 张图片, 其中 1 255 张图片用于训练集, 其余 300 张图片用于测试集. 该数据集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. 标注单位是单词.

6) CASIA-10K[39] 数据集共有 10 000 张图片, 其中 7 000 张图片用于训练集, 其余 3 000 张图片用于测试. 该数据集采集自中文场景, 每个文本行标注其 4 个顶点坐标.

3.2 评价指标

根据 ICDAR2015 评价方法[11], 使用信息检索领域的精确率 P、召回率 R 和 F1 得分 F, 综合评估文本算法的性能. 计算 P 和 R 依赖于交并比 IoU. IoU 由第 i 个检测的矩形框 D_i 和第 j 个标签 G_j 间的交集/并集比值定义, 如果 IoU ≥ 0.5, 该检测结果正确. 定义 IoU 表达式为:

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

式中, $\text{area}(G_j \cap D_i)$ 和 $\text{area}(G_j \cup D_i)$ 分别表示 G_j 和 D_i 的交集/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 T_p, 则精确率、召回率和 F1 得分定义如下:

$$P = \frac{|T_p|}{|D|} \tag{12}$$

$$R = \frac{|T_p|}{|G|} \tag{13}$$

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

3.3 实验设置

本文实验目的是评价本文提出的自蒸馏方法与其他蒸馏方法的性能对比, 因此所有基线模型, 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. 如第 2.3 节所述, 本文设定 $\gamma = 1$ 即可取得满意性能, 而无需

精细调整参数.

消融实验使用 Pytorch 平台, 在单张 1080Ti 显卡上训练. 主干网络使用 MobileNetV3 的 EAST 模型分析自蒸馏算法的影响元素. 其他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. 一般地, 图像的数据增强采用随机旋转 (−10°, 10°) 或随机剪裁. 为了保证不超出显存, 在训练 EAST 时, 训练图像统一缩放至 512 × 512 像素; 在训练 DB 时, 统一缩放至 640 × 640 像素. 优化器采用随机梯度下降, 并且使用多项式学习率调整策略, 在训练主干网络是 MobileNetV3 时, 批大小设置为 8, 训练 1 200 轮; 在训练 ResNet50 时, 批大小设置为 4, 同样训练 1 200 轮.

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, 比较了图 4 中 3 种辅助分类器 (即 A 型、B 型、C 型) 对 SDET 的影响. 在 ICDAR2013 和 ICDAR2015 数据集上的实验结果如表 1 所示, 其中 MV3-EAST、MV3-DB 分别表示主干网络采用 MoblieNetV3 和分割头使用 EAST、分割头使用 DB 的文本检测模型. 实验结果表明, 对同一个基线模型, 采用不同辅助分类器的 SDET 性能存在差异. 例如, 对 MV3-EAST, 简单的 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基线的 F1 得分; 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. 总之, C 型辅助分类器较具有鲁棒性, 均能有效提升 MV3-EAST 和 MV3-DB 基线模型的 F1 得分. 其他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 $P0 \sim P3$. 因此可以连接辅助分类器的位置共有 4 个, 用 MV3-EAST 作为基线模型, 在 ICDAR2015 数据集上, 不同特征金字塔位置对 B 型的影响见表 2, 其中 $P0 \sim P3$ 分别表示将辅助网络连接在 0 ~ 3 位置上. 由表 2 可以看出, 在 $P2$ 和 $P3$ 位置放置辅助分类器, 有利于 SDET 的训练; 在 $P0$ 和 $P1$ 位置放置辅助分类器, 则会抑制 SDET 的训练. 其他位置上的实验表现一致, 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, 可根据不同主干网络抽取特征的能力不同, 选择相应的金字塔位置. 同时可以看出, 轻量级网络 (如 MobileNetV3) 可以选择 $P3$ 或 $P2$ 位置; 对于主干网络为 ResNet50 的大网络, 将辅助分类器连接到 $P1$ 层效果较好.

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

模型	方法	ICDAR2013			ICDAR2015		
		P	*R*	*F*	*P*	*R*	*F*
MV3-EAST	基线	81.7	64.4	72.0	80.9	75.4	78.0
	A 型	78.8	65.9	71.8	78.8	76.3	77.5
	B 型	84.4	66.5	**74.4**	81.3	77.0	**79.1**
	C 型	81.4	67.4	73.7	78.9	77.7	78.3
MV3-DB	基线	83.7	66.0	73.8	87.1	71.8	78.7
	A 型	84.1	68.8	**75.7**	86.5	73.9	**79.7**
	B 型	81.1	67.3	73.6	87.8	71.7	78.9
	C 型	84.9	67.9	75.4	87.8	73.0	**79.7**

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

方法	特征图尺寸 (像素)	*P*	*R*	*F*
基线	—	80.9	75.4	78.0
*P*0	16 × 16	79.1	75.8	77.4
*P*1	32 × 32	79.5	76.5	78.0
*P*2	64 × 64	80.7	77.4	79.0
*P*3	128 × 128	81.3	77.0	**79.1**

3.5　与主流蒸馏方法的对比

将 SDET 和目前主流的 6 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 6 种蒸馏方法分为以下 2 类: 1) 传统的学生−教师框架的知识蒸馏方法 (即学生−教师蒸馏法 (Student-teacher, ST))[7]、中间层特征蒸馏法 (FitNets)[9]、知识适配法 (Knowledge adaptation, KA)[29] 和 SKD[30]. 其中, ST 表示转移教师网络输出的软化概率值; FitNets 通过范数最小化教师−学生网络的中间特征图, 当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实现教师网络与学生网络特征间的适配; SKD 使用 KL (Kullback-Leibler) 散度对齐教师−学生网络分割图上的每个像素点概率, 然后匹配特征图对应的相似性矩阵. 2) 近年流行的 SD[32] 和 SAD[17]. 其中, 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力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, 学生网络的主干网络采用 MobileNet-V3, 分别使用可微二值化和 EAST 作为最后的文本检测分割头, 教师网络将主干网络替换为 ResNet50.

由表 3 和表 4 可以看出, 本文提出的自蒸馏方法 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 (加粗数字为最高 F1 得分). 尤其是在 ICDAR-2013 数据集上, 相较于基线的学生网络, 经过自蒸馏训练的 DB 模型在精确率、召回率和 F1 得分上分别有 0.4%、2.8% 和 1.9% 的提升, 同样 SDET 有效提升 EAST 模型的 F1 得分, 从 72.0% 提高到 74.4%. 图 5 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, 其中图 5(a) 中方框为文本所在位置 (即真实标签), 图 5(b) 方框为基线模型对 3 幅图像的检测结果, 用圆框凸显与真实标签有显著差异, 图 5(c) 方框为 SDET 训练后的模型检测结果. 由图 5 可以看出, 图 5(b) 中基线模型的预测结果存在检测边缘漏判、不完全或误判情况, 而自蒸馏训练的网络检测出的结果具有相对较高的精确性和鲁棒性, 仅将图中茶杯手柄误测为字母 D.

同时, 由表 3 还可以看出, 在 MV3-DB 学生网络, 其他蒸馏方法难以有一致性的稳定提高. 例如在小数据集 TD500 上, 传统蒸馏方法能在不同程度上提升学生网络的 F1 指标, 但在大一些的数据集 (如 ICDAR2015、Total-text 和 CASIA-10K) 上, 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 (如图像分类任务中常用的 ST 方法). 原因可能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[13]. 由表 4 可以看出, 除了 SKD、SAD 和本文 SDET 方法外, 其他蒸馏方法缺乏一致性的性能提升, 其中 SDET 提升最为显著.

综上所述, 本文提出的自蒸馏方法 SDET 在没有训练一个教师网络情况下, 在多个数据集上, 效

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 (%)

方法	ICDAR2013			TD500			TD-TR			ICDAR2015			Total-text			CASIA-10K		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
基线	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	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	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	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	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	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	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
本文方法	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 (%)

方法	ICDAR2013			ICDAR2015			CASIA-10K		
	P	R	F	P	R	F	P	R	F
基线	81.7	64.4	72.0	80.9	75.4	78.0	66.1	64.9	65.5
ST	77.8	64.9	70.8	80.9	75.1	77.9	64.7	65.1	64.9
KA	78.6	64.0	70.5	78.2	76.4	77.3	67.7	63.0	65.3
FitNets	82.4	65.8	73.2	78.0	77.8	77.9	65.4	64.2	64.8
SKD	79.5	66.3	72.3	81.9	75.6	78.6	66.6	64.7	65.6
SD	80.2	63.8	71.1	79.6	74.7	77.1	66.2	63.5	64.8
SAD	81.4	65.6	72.6	80.2	76.5	78.3	65.7	64.1	64.9
本文方法	84.4	66.5	**74.4**	81.3	77.0	**79.1**	70.8	63.0	**66.7**

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)

果都超出其他蒸馏方法且无需额外调整参数. 由于训练一个合适的教师网络不仅需要较大内存, 还需要耗费大量时间调整参数, 自蒸馏框架可大大节约内存和时间, 还能带来令人满意的性能提升, 因而具有很大优势.

3.6　SDET 与 DSN 方法对比

SDET 方法与 DSN 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. 在 ICDAR2013、TD500、TD-TR、ICDAR2015、Total-text 和 CASIA-10K 数据集上分别用 SDET 和 DSN 两种方式训练主干网络为 MobileNetV3、分割头是 DB 的文本检测基线模型, 实验结果见表 5. 由表 5 可以看出, SDET 在各数据集上都能取得更好的性能, 加粗数字为 3 种方法的最高 F1 得分.

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, 这种方式能提高网络性能的原因是: 1) 信息熵具备更多的信息量. 由图 1 可以看出, 信息熵放大了模型对边缘的注意力. 2) 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, 其学习过程从易到难, 逐步提高难度.

3.7　推广到大网络

传统知识蒸馏方法常用在较小的学生网络上, 如果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知识, 无需训练庞大的教师网络, 其优势更加显著. 用主干网络为 ResNet50 的 DB 模型 (ResNet50-DB) 当作基线模型, 在 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 像素, 实验结果如表 6 所示.

由表 6 可以看出, SDET 能有效提升 ResNet50 的性能表现, 在 6 个数据集上 F1 得分均有提升 (数字加粗显示), 在数据集 TD-TR、ICDAR2015、Total-text 和 CASIA-10K 上均有超过 1% 的提高. 对比基线模型可以发现, 精确率并没有改善, F1 得分的提升是由于 SDET 显著地提升了模型的召回率, 在 6 个数据集上分别提高了 4.3%、5.6%、2.6%、2.0%、2.4% 和 4.0%. 由式 (12) 和式 (14) 可知, 召回率提升反映了有效检出 $|T_P|$ 值增大, 可能原因是

表 5　SDET 与 DSN 在不同数据集上的对比 (%)

Table 5　Comparison of SDET and DSN on different datasets (%)

方法	ICDAR2013			TD500			TD-TR			ICDAR2015			Total-text			CASIA-10K		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
基线	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	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
本文方法	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 ResNet50-DB 的效果 (%)

Table 6　The effect of SDET on improving ResNet50-DB on different datasets (%)

方法	ICDAR2013			TD500			TD-TR			ICDAR2015			Total-text			CASIA-10K		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
基线	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
本文方法	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**

浅层分类器经过来自深层分类器的信息熵监督训练, 促进网络学习边缘知识, 从而检测边缘更加准确, 使得 IoU 普遍增大, $|T_P|$ 值也随之提高.

4 结束语

本文提出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传递信息熵知识, 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 6 个标准数据集上的对比实验结果表明, SDET 无需精细地调整参数过程, 即可提升不同规模大小的基线模型 (如 MV3-DB、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 (如 CTPN), 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. 本文存在的不足是仅设计了 3 种简单的辅助网络, 而不同的文本检测网络需要不同的辅助网络. 未来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, USA: IEEE, 2015. 3431−3440

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

3 Lv 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 29th International Joint Conference on Artificial Intelligence. Yokohama, Japan: 2020. 516−522

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, USA: IEEE, 2016. 770−778

7 Hinton G E, Vinyals O, Dean J. Distilling the knowledge in a neural network. arXiv preprint 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*, 2023, **49**(1): 102−110
(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, 2023, **49**(1): 102−110)

9 Romero A, Ballas N, Kahou S E, Chassang A, Gatta C, Bengio Y. FitNets: Hints for thin deep nets. arXiv preprint 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 preprint arXiv: 1612.03928, 2016.

11 Karatzas D, Gomez-Bigorda L, Nicolaou A, Ghosh S, Bagdanov A, Iwamura M, et al. ICDAR2015 competition on robust reading. In: Proceedings of the 13th International Conference on Document Analysis and Recognition. Nancy, France: IEEE, 2015. 1156−1160

12 Chng C K, Chan C S. Total-text: A comprehensive data-set 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, et al. 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, 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 18th International Conference on Artificial Intelligence and Statistics. San Diego, 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, et al. 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, 等. 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 preprint arXiv: 1506.01497, 2015.

20 Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y, et al. SSD: Single shot multi-box detector. In: Proceedings of the European Conference on Computer Vision. Amsterdam, Netherlands: 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 31st AAAI Conference on Artificial Intelligence. San Francisco, 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, Netherlands: Springer, 2016. 56−72

23 Zhou X Y, 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, USA: IEEE, 2017. 5551−5560

24 Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Proceedings of the Medical Image Computing and Computer Assisted 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 34th 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, et al. Shape robust text detection with progressive scale expansion network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 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, 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, 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, et al. 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, 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. ICDAR2013 robust reading competition. In: Proceedings of the 12th International Conference on Document Analysis and Recognition. Washington DC, USA: IEEE, 2013. 1484–1493

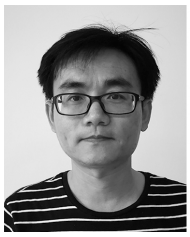
37 Yao C, Bai X, Liu W Y, Ma Y, Tu Z W. Detecting texts of arbitrary orientations in natural images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA: IEEE, 2012. 1083–1090

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 multi-lingual 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 at the School of Aerospace Engineering, Xiamen University. His research interest covers computer vision and image processing.)



杨 帆 厦门大学航空航天学院副教授. 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 at the School of Aerospace Engineering, Xiamen University. His research interest covers machine learning, data mining, and bio-informatics. Corresponding author of this paper.)

赖永炫 厦门大学信息学院教授. 主要研究方向为大数据分析和管理, 智能交通系统, 深度学习和车载网络. E-mail: laiyx@xmu.edu.cn

(**LAI Yong-Xuan** Professor at the School of Informatics, Xiamen University. His research interest covers big data analysis and management, intelligent transportation systems, deep learning, and vehicular networks.)