doi:10.6041/j.issn.1000-1298.2020.08.037

基于语义分割的食品标签文本检测

田　萱[1,2]　王子亚[1]　王建新[1,2]

(1.北京林业大学信息学院,北京 100083;2.国家林业草原林业智能信息处理工程技术研究中心,北京 100083)

摘要:食品包装上的标签文本含有生产日期、营养成分、生产厂家等食品相关信息,这些不仅为消费者购买食品提供了重要依据,也有助于食品监督抽检机构发现潜在的食品安全问题。食品标签文本检测是食品标签自动识别的前提,有助于降低人工录入成本、提高数据处理效率。基于食品包装图像构建数据集,提出了一种基于语义分割的距离场模型,以检测食品标签。该模型包含像素分类和距离场回归两类任务,其中像素分类任务分割处理图像中的文本区域,距离场回归任务预测文本区域内的像素点到该区域边界的归一化距离。为提升模型的检测性能,在回归预测模块中通过增加注意力模块优化模型结构,并针对距离场回归任务损失值过小、影响模型训练优化问题对其损失函数进行了改进。消融实验结果表明,增加注意力模块和损失函数的改进使得模型的准确率分别提高了 4.39、3.80 个百分点,有效提高了检测准确率。食品包装图像数据集的对比实验表明,采用本文模型检测食品标签文本具有较好的性能,其召回率、准确率分别达到 87.61%、76.50%。

关键词:食品标签;语义分割;文本检测;文本识别

中图分类号:TP391.1　　**文献标识码**:A　　**文章编号**:1000-1298(2020)08-0336-08　　**OSID**:

Text Detection of Food Labels Based on Semantic Segmentation

TIAN Xuan[1,2]　WANG Ziya[1]　WANG Jianxin[1,2]

(1. *School of Information Science and Technology*, *Beijing Forestry University*, *Beijing* 100083, *China*
2. *Engineering Research Center for Forestry-oriented Intelligent Information Processing of National Forestry and Grassland Administration*, *Beijing* 100083, *China*)

Abstract: The label texts on food package include some information like production date, nutrition facts and production corporation etc. The information provides important foundation for consumers to buy food. It also can help the food supervision and inspection administrations to discover the potential problems of food safety. Food label detection is the groundwork of food label recognition. It can help to decrease the heavy workload of manual inputting and advance efficiency of data processing. The dataset of food label was constructed firstly, and then a semantic segmentation based distance field model (DFM) was proposed. In DFM two tasks were included: pixel classification and distance field regression. The pixel classification task was used to segment the text from background regions, and the distance field regression task was used to predict the normalized distance from the pixel located in the text region to the boundary of text region. For effectively using the correlation of two tasks, an attention module was added into DFM to optimize the model structure. In addition, the loss function was improved to resolve the loss value of the distance field regression as it was too small to train smoothly. The results of ablation experiment showed that the accuracy of the proposed model was increased by 4.39 percentage points and 3.80 percentage points respectively according to the improvement of attention module and loss function. The comparative experiments of different model methods showed that DFM had good performance in detecting the text of food labels, and the recall rate and precision were 87.61% and 76.50%, respectively.

Key words: food labels; semantic segmentation; text detection; text recognition

收稿日期:2020-05-13　修回日期:2020-06-03

基金项目:国家重点研发计划项目(2018YFC1603305、2018YFC1603302)

作者简介:田萱(1976—),女,副教授,主要从事智能信息处理与文本挖掘研究,E-mail:tianxuan@bjfu.edu.cn

通信作者:王建新(1972—),男,教授,主要从事数据挖掘与大数据处理研究,E-mail:wangjx@bjfu.edu.cn

0　引言

食品标签是印刷于食品封装容器上的文字，主要包括产品名称、配料、生产商、产地和营养成分等用于描述该食品相关信息的内容。消费者根据该信息查询并判断该食品是否符合自身需求，食品监督抽检机构根据该信息进行食品安全相关的数据挖掘和分析[1-2]。作为食品标签识别的必要步骤，食品标签的文本检测（即定位其边界）对于食品安全监督和保障非常重要。

文本检测是图像信息分析和提取过程中的重要技术，其关键在于如何区分和定位复杂的文本区域和背景区域。传统的图像文本检测主要利用人工设计特征检测图像中的文本，主要包括基于连通域分析和基于滑动窗口两类检测方法。其中，笔画宽度变换（Stroke width transformation，SWT）[3]和最大稳定极值区域（Maximally stable extremal regions，MSER）[4]是经典的基于连通域分析的方法，分别利用图像中文本的笔画宽度和像素特征获取文本区域。基于滑动检测窗的方法采用了自顶向下策略检测文本，该类方法[5-6]采用不同尺寸窗口对整幅自然场景图像进行扫描，然后提取窗口区域的人工设计特征[7-8]，结合已训练的分类器判断图像中窗口位置是否为文本。

食品标签检测属于复杂场景下的文本检测[9]，不同的食品包装包含不同尺寸、颜色、形状、对比度的标签文本，并且食品标签的背景相较于印刷文档更加复杂，传统的文本检测方法[10-13]受到人工设计特征的局限性，难以满足复杂的食品标签文本检测需求。目前，深度学习技术已经广泛应用于目标检测领域[14]，并且对文本检测技术的创新具有重要的推动作用，能够有效避免人工设计特征存在的局限性。其中，基于区域建议的方法[15-16]由于需要预设固定形状的候选框，使其对不规则文本的检测受到限制。语义分割技术[17]能够从图像中分割任意形状目标区域，经典的语义分割算法包括 FCN[18]、SegNet[19]、DeepLab[20]和 PSPNet[21]等，目前已经出现众多基于语义分割的自然场景文本检测算法[22-25]，但这些方法检测食品标签的准确率有待提升。本文借鉴经典语义分割算法的思路，提出基于语义分割的距离场模型，以检测食品标签文本，使计算机能够有效分割图像中的文本，实现食品包装上文本内容的自动定位。

1　数据集构建

鉴于目前没有公开可用的食品包装图像数据集，本文通过自采数据集形式进行模型训练和实验评估。通过手机拍摄，从不同颜色以及透明的袋、盒、瓶和罐状包装表面采集得到食品包装图像共计 509 幅，这些图像涵盖了茶叶类、肉制品、奶制品、调味品、食用油、干果类和酒类等主要食品类型。图像中包含不同的食品标签文本，文本框共计 13 205 个，其中矩形文本框 11 880 个，弯曲不规则文本框1 325 个。但本文在研究过程中利用由若干个坐标点顺时针连接构成的多边形表示文本区域，因此并不对规则矩形和弯曲标签文本进行特别区分。

1.1　数据集划分

采用自助采样方法划分训练集和测试集，即假设初始数据集 Q 包含 M 个样本，每次使用有放回采样方式从数据集 Q 中随机选取一个样本，将其复制放入新数据集 Q'中，每个样本每次未被采样到的概率为 $1-1/M$，重复上述随机采样过程 M 次，得到一个新的数据集 Q'。显然，数据集 Q 中存在一直未被采样到的样本，本文将这些样本作为测试集，其他样本作为训练集。通过这种采样方法使得从数据集 Q 中划分的测试集约占数据集总数的 1/e，即

$$\lim_{M\to\infty}\left(1-\frac{1}{M}\right)^{M}=\frac{1}{\mathrm{e}} \tag{1}$$

通过上述方法本文将食品包装图像划分为训练集和测试集，如表 1 所示，其中训练集图像为 327 幅且包含 8 453 个文本框，测试集图像为 181 幅且包含 4 752 个文本框。

表 1　数据集数量

Tab.1　Number of datasets

参数	数据集	训练集	测试集
图像数	509	327	181
文本框数	13 205	8 453	4 752

1.2　数据集标注

为了有效表示食品包装图像中标签文本区域，本文将属于同一标签内容的空间区域视为一个单独的文本区域。鉴于食品包装可能存在褶皱和弯曲的现象，采用由若干个坐标点顺时针连接构成的多边形表示文本区域，这样可以满足任意形状的标签区域标识需要，因此未区分规则矩形和弯曲标签文本。原始图像和人工标注如图 1 所示。

2　基于语义分割的距离场模型

针对食品包装图像中的标签文本检测，提出一种基于语义分割的距离场模型（Semantic segmentation based distance field model，DFM），该模型包含像素分类任务和距离场回归任务。DFM 能

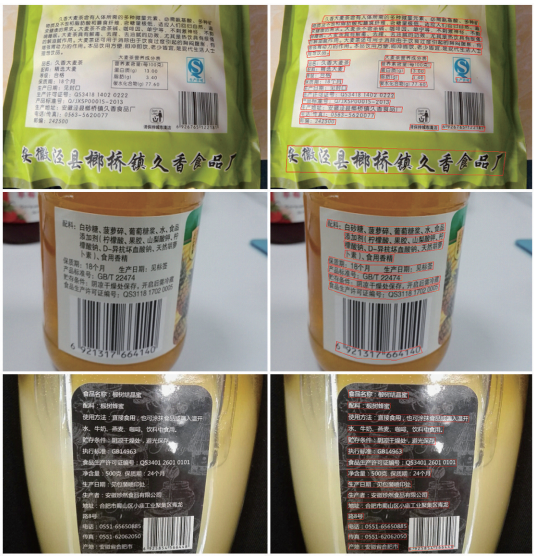


图 1 原始图像和标注图像示例

Fig. 1 Examples of original images and annotated images

够根据神经网络模型预测图像中文本区域内像素点至其边界的归一化距离以及文本与背景分割图。为了有效利用像素分类任务与距离场回归任务之间存在的联系以提升模型的检测效果,在回归预测模块中增加了注意力模块,并且针对距离场回归任务损失值过小导致模型训练优化不佳的问题,对其损失函数提出改进,以提高模型的预测准确率。最后根据 DFM 两类任务的预测结果利用扩展分组算法获得食品包装的标签文本区域。

2.1 模型框架

DFM 包含两部分:多尺度特征提取模块和回归预测模块,多尺度特征提取模块利用基础网络提取并融合不同维度的图像特征,再将融合特征输入到回归预测模块。回归预测模块包含像素分类和距离场回归任务,分别预测食品包装图像中的文本分割图以及文本区域内像素到最近边界的归一化距离(简称距离场),模型检测流程如图 2 所示。

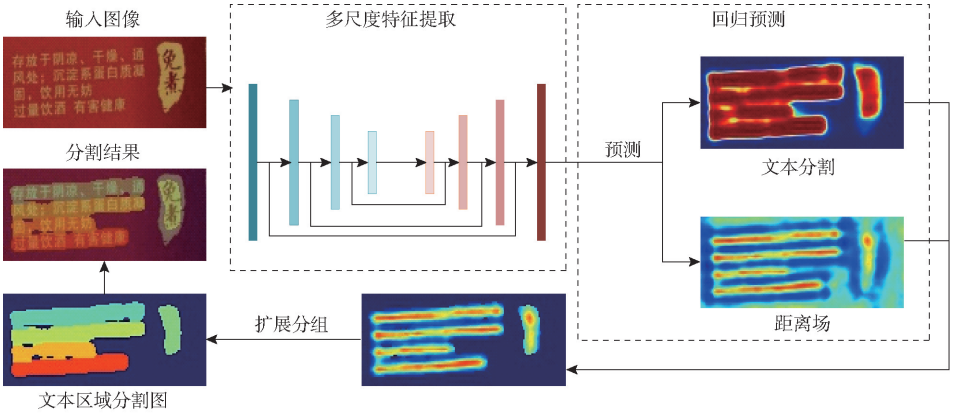


图 2 基于语义分割的距离场模型检测流程

Fig. 2 Detection process of distance field model based on semantic segmentation

2.1.1 多尺度特征提取模块

DFM 使用 ResNet[26] 网络作为基础特征提取网络,该网络是当前应用最广泛的深度卷积特征提取网络之一。ResNet[26] 主要由 stage1 ~ stage5 共 5 个网络模块组成,多尺度特征提取模块采用基于 FPN[27](特征金字塔网络)的方式,对 ResNet[25] 提取的特征进行融合作为后续回归预测模块的输入,即融合低维图像精细特征和高维图像粗略特征。

如图 3 所示,将 ResNet-50 网络中不同阶段得到语义特征 f_1、f_2、f_3 和 f_4 通过卷积和上采样运算得到不同维度融合特征 P_1、P_2、P_3 和 P_4,其计算方式为

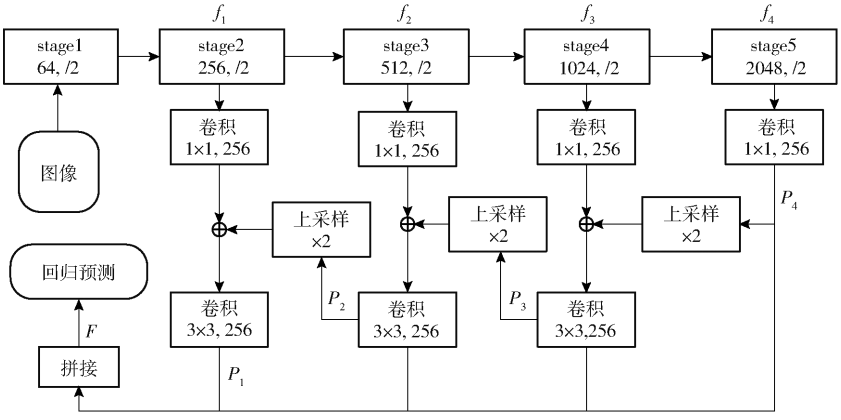


图 3 基于语义分割的距离场模型结构

Fig. 3 Structure diagram of distance field model based on semantic segmentation

$$P_4 = \mathrm{Conv}_{1\times1}(f_4) \tag{2}$$

$$P_i = \mathrm{Conv}_{3\times3}(\mathrm{Conv}_{1\times1}(f_i) + \mathrm{Up}_{\times2}(P_{i+1})) \quad (i=1,2,3) \tag{3}$$

式中　$\mathrm{Up}_{\times s}$——s 倍上采样函数

$\mathrm{Conv}_{k\times k}$——核尺寸 $k\times k$ 的卷积函数

最后通过上采样运算和拼接操作将 P_1、P_2、P_3 和 P_4 合并得到多尺度融合特征图 F，其计算方式为

$$F = [P_1; \mathrm{Up}_{\times2}[P_2; \mathrm{Up}_{\times2}[P_3; \mathrm{Up}_{\times2}(P_4)]]] \tag{4}$$

式中　$[X_1; X_2]$——X_1 与 X_2 拼接

2.1.2　回归预测模块及其改进

回归预测模块对特征 F 进行卷积运算并上采样至原图像大小，得到预测图 $\boldsymbol{P} \in \mathbf{R}^{H\times W\times3}$，其中 $\mathbf{R}^{H\times W\times2}$ 用于食品包装图像的像素分类预测任务，$\mathbf{R}^{H\times W\times1}$ 用于距离场回归任务，H 和 W 分别表示输入图像的高度和宽度。

DFM 预测任务中，像素分类任务与距离场回归任务具有很强的关联性。对于距离场真值结果 T_d^* 和像素分类真值结果 T_c^*，显然距离场真值结果值大于 0 的像素集合 P_d 应该与像素分类真值结果为文本的像素集合 P_c 相等，即

$$P_d = \{p_i \mid p_i \in T_d^* > 0\} \tag{5}$$

$$P_c = \{p_i \mid p_i \in T_c^* = 1\} \tag{6}$$

$$P_d = P_c \tag{7}$$

在计算距离场损失值时，本文仅考虑图像文本区域中像素点距离预测值与真值的差距。因此，距离场的预测结果仅关注食品图像中像素分类结果为文本区域的特征，鉴于 DFM 两类预测任务之间存在这样的关联性，本文在回归预测模块增加注意力模块(Attention module，AM)以进一步强化任务之间的关系，并利用这种关系提高模型的预测准确率。如图 4b 所示，考虑到距离场的预测任务要比像素分类任务更加困难，本文在多尺度融合特征 F 的基础上利用宽高均为 3 的卷积核提取图像复杂特征 F_1，提高距离场预测结果准确性。然后将像素分类任务文本区域预测图与特征 F_1 按元素相乘得到特征 F_d，最后再对特征 F_d 进行卷积运算并通过 Sigmoid 激活函数得到距离场的预测结果 T_d。

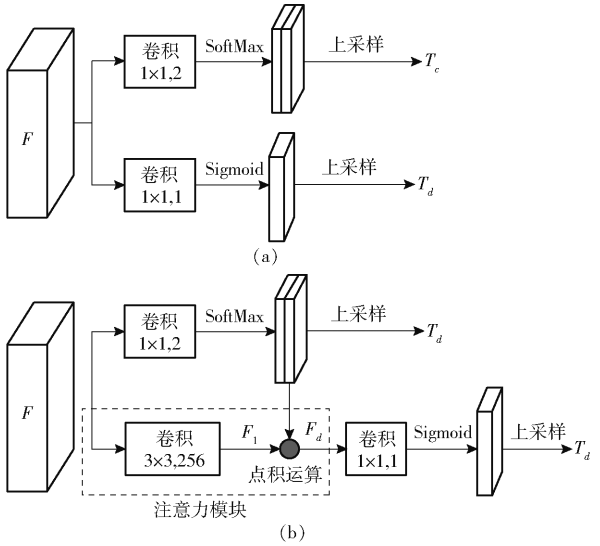


图 4　回归预测模块结构

Fig. 4　Structure diagram of regression prediction module

2.1.3　模型损失函数及其改进

DFM 模型的损失函数 L 为

$$L = L_t + \lambda L_d \tag{8}$$

式中　L_t——像素分类任务的损失函数

L_d——距离场回归任务的损失函数

λ——两类任务损失的平衡系数

(1)像素分类任务的损失函数

像素分类任务预测食品包装图像中标签文本区域，本文将文本区域内的像素点标记为正样本，背景区域的像素点标记为负样本，通常食品包装图像中背景区域占比较大，这种类别不平衡使得网络偏向学习背景特征从而影响食品标签的检测。当前有不同的解决方法处理该问题，例如困难负样本挖掘(Hard example mining，HEM)[28]、焦点损失函数(Focal loss)[29]和平衡交叉熵等，本文采用 HEM 策略筛选预测错误的负样本，使正负样本的比例为 1∶3。并使用平衡交叉熵作为像素分类的损失函数，公式为

$$L_t = \frac{1}{\Omega}\sum_{i\in\Omega}[c_i \lg c_i^* + (1-c_i)\lg(1-c_i^*)] \tag{9}$$

式中　Ω——经过困难负样本挖掘获得的所有正负样本的集合

c_i^*——第 i 个像素点位置是文本区域的真值置信度

c_i——预测置信度

(2)距离场回归任务的损失函数

仅依赖文本与背景的像素分类任务很难区分排列紧凑的文本行或者单词，本文提出距离场回归任务，即预测图像中属于文本区域内的第 i 个像素点与该文本区域边界的归一化距离 d_i(图 5a)，将 d_i 大于阈值的部分视为文本中心区域，再向边缘区域按 d_i 递减扩展像素点，直至 d_i 不再递增。d_i 的计算方式为

$$d_i = \frac{r_{i,j}}{\max(r_j)} \tag{10}$$

其中

$$r_{i,j} = \min\left(\bigcup_{b\in B_j} \| p_{i,j} - b \|_2\right) \tag{11}$$

式中　$p_{i,j}$——属于文本区域 j 的第 i 个像素点的坐标

B_j——文本区域 j 的边界像素点坐标集合

$r_{i,j}$——属于文本区域 j 的第 i 个像素点到其

边界 B_j 的欧氏距离

r_j——属于文本区域 j 的所有像素点至边界 B_j 欧氏距离的集合

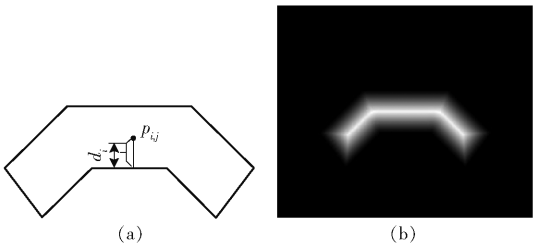


图 5 距离场标签生成

Fig. 5 Label generation of distance field

为了消除距离对文本区域和图像尺寸的敏感性,本文将欧氏距离变换至范围为[0,1]的归一化距离 d_i,将文本区域内像素点到边界的欧氏距离 $r_{i,j}$ 除以 r_j 的最大值得到归一化距离 d_i。对于背景区域的像素点,本文将其距离场的标签值设为零,最后生成的距离场标签如图 5b 所示。显然文本中心到边界区域的 d_i 逐渐减小,且相邻的不同文本区域边界区域的 d_i 呈现先递减再递增的变化趋势,这种特征有助于区分相邻的文本。最后基于分类预测和距离场回归结果使用后处理的距离场分组算法得到食品标签文本并分离相邻的文本,DFM 的初始距离场损失定义为

$$L_d = \frac{1}{2 \mid T_c^* \mid} \sum_i c_i^* (d_i - d_i^*)^2 \tag{12}$$

式中 d_i^*——距离场真值

在距离场真值标注中,由于越靠近边界区域的归一化距离值越接近于零,这种不一致性导致边界区域的距离场预测结果损失通常偏小,因而不能保证预测结果的准确性,这种损失值与期望准确率的不匹配性会影响距离场回归预测结果,进而影响后处理分组算法获取食品标签文本。

本文使用两个不同的方案解决该问题,从细节考虑,由于距离场损失和预测准确率的不匹配性源于误差参考尺度的不一致性,使用相对误差替换每个像素位置距离回归预测的绝对误差,即增加权重的方式,其计算方式为

$$L_{d_w} = \frac{1}{2 \mid T_c^* \mid} \sum_i c_i^* \left(\frac{d_i - d_i^*}{d_i^*} \right)^2 \tag{13}$$

从整体考虑,回归任务比分类任务有更高的预测精度要求,然而为了消除食品标签文本大小对距离的影响,本文使用归一化距离作为预测目标,这导致食品标签文本检测的距离场损失值很小,因此,在损失函数基础上增加开方运算自适应放大损失值,即增加开方的方式,而不仅依赖于固定的距离场损失系数,其计算方式为

$$L_{d_s} = \sqrt{\frac{1}{2 \mid T_c^* \mid} \sum_i c_i^* (d_i - d_i^*)^2} \tag{14}$$

为了对比验证不同函数对于模型训练优化的作用,本文将在距离场回归任务中分别采用损失函数 L_d、L_{d_w}、L_{d_s} 以及 L_{d_ws} 进行实验。L_{d_ws} 为增加权重与开方的组合方式,其计算方式为

$$L_{d_ws} = \sqrt{\frac{1}{2 \mid T_c^* \mid} \sum_i c_i^* \left(\frac{d_i - d_i^*}{d_i^*} \right)^2} \tag{15}$$

2.2 扩展分组算法

为了区分彼此靠近的食品标签文本,本文根据距离场任务的预测结果利用扩展分组算法来解决这个问题,其中心思想是源于广度优先搜索算法。模型预测结果中包含图像像素分类 T_c 和距离场 T_d,通过设置阈值 α 和 β,将 $T_c > \alpha$ 且 $T_d > \beta$ 的区域通过连通域计算得到食品标签文本的中心区域 T_k,如算法 1 第 3 ~ 11 行所示,其中 α 取值为 0.9,β 取值为 0.5。

算法 1 扩展分组算法

输入:像素分类预测 T_c 和距离场预测 T_d,文本区域分割阈值 α,中心区域分割阈值 β,边界阈值 γ

输出:文本区域分割图 T_s

1 $T_k \leftarrow T_c > \alpha$ and $T_d > \beta$
2 $visited \leftarrow \varnothing$, $queue \leftarrow \varnothing$, $T_s \leftarrow \varnothing$, $N \leftarrow 0$
3 for $p_i \in T_k$ do
4 if $p_i \notin visited$ then
5 $queue$. put(p_i); $visited \leftarrow visited \cup p_i$; $N \leftarrow N + 1$; $K \leftarrow \varnothing$
6 while $queue \neq \varnothing$ do
7 $p_i \leftarrow queue$. pop(); $K \leftarrow K \cup (p_i, N)$
8 if $\exists h \in$ neibor(p_i) and $h \notin visited$ and $h \in T_k$ then
9 $queue$. put(h); $visited \leftarrow visited \cup h$
10 end if
11 end while
12 $queue$. put(boundary(K))
13 while $queue \neq \varnothing$ do
14 $p_i \leftarrow queue$. pop()
15 if $\exists h \in$ neibor(p_i) and $h \notin visited$ and $T_d(p_i) \geqslant T_d(h) \geqslant \gamma$ then
16 $queue$. put(h); $K \leftarrow K \cup (h, N)$; $visited \leftarrow visited \cup h$
17 end if
18 end while
19 $T_s \leftarrow T_s \cup K$
20 end if
21 end for
22 return T_s

然后将 $T_c > \alpha$ 的非中心区域像素点进行分组将其划分入不同的标签文本，如算法 1 第 12 ~ 17 行所示。该分组过程为：将 T_k 的边缘像素点作为初始点，选择每个像素点的相邻且未分组的点作为候选扩展点，根据标签文本区域距离场从内向外的递减特征，若候选像素点位置的距离场预测值不大于当前点，将其作为扩展像素点划分入 T_k 中，循环上述步骤寻找新增点的候选扩展点并判断是否符合分组条件，直至扩展到边界时停止。由于 Sigmoid 激活函数预测值大于零，算法实现中将距离场边界的阈值 γ 设置为很小的值（本文中 γ 取值为 0.001）而不是零。

3　实验与结果对比分析

3.1　模型训练与评估

3.1.1　训练细节及参数

由于本文数据集偏小，故采用以下数据增广策略提高模型训练效果：将图像随机垂直翻转或旋转大小在（-10°，10°）区间内的角度；将图像按照比例{0.5，1.0，2.0}随机缩放；从图像中随机裁剪尺寸为 512 像素 ×512 像素子区域作为训练样本。此外，还将图像按通道内的均值和标准差进行归一化以统一训练数据的分布，增加模型的训练速度。训练阶段使用权重（decay）为 0.001、动量为 0.9 的随机梯度下降算法作为优化器，损失函数平衡系数 λ 取值为 5，每次训练图像数量为 8，数据集训练轮次（epoch）为 600，初始学习率为 10^{-3}，且每 150 轮次缩小 90%。

3.1.2　评估方法

目前文本检测包含 3 个评价指标：召回率 R_{recall}、准确率 $P_{precision}$ 和调和平均值 $F_{measure}$，调和平均值的计算方法为

$$F_{measure} = \frac{2R_{recall}P_{precision}}{R_{recall} + P_{precision}} \tag{16}$$

真值框与检测框的交并比 $I_{i,j}$ 的计算方法为

$$I_{i,j} = \frac{\text{area}(\text{intersection}(G_i, D_j))}{\text{area}(\text{union}(G_i, D_j))} \tag{17}$$

式中　intersection（·）——交集函数

union（·）——并集函数

area（·）——面积函数

G_i——真值框　　D_j——检测框

3.2　实验结果对比

为了验证注意力模块对模型的影响以及不同损失函数的预测效果，通过消融实验对比改进前后的模型检测性能。并且为了证明模型检测食品标签文本的准确性，将 DFM 与其他文本检测模型进行实验对比，并介绍分析实验结果。

3.2.1　不同回归预测模块的对比

首先分别对改进前的预测结构（图 4a）以及改进后包含注意力模块的预测结构（图 4b）进行训练，以此验证注意力模块对于模型性能的提升。为了消除损失函数优化对于实验结果的影响，本实验的损失函数采用均方差损失函数 L_d，其具体计算方法如式（12）所示。实验结果如表 2 所示，增加空间注意力模块使得模型在准确率以及调和平均均有所提升，分别提高了 4.39、2.23 个百分点。相较于改进前直接预测像素分类结果和距离场回归结果的方式，通过增加注意力模块使得距离场的回归预测与像素分类结果存在关联关系，即利用像素分类任务的预测结果弱化模型提取的图像背景区域特征，使得距离场回归任务减少对于非文本区域的错误预测，而当距离场回归任务出现假文本区域的距离场预测结果时，表明模型的像素分类任务存在错误的背景与文本预测结果，通过间接反馈使得模型在训练的过程中进一步优化像素分类任务。其意义在于使得模型的像素分类任务更加关注图像的背景与文本的二分类预测，降低模型对于食品图像中背景区域的错误预测，从而增加模型整体的预测准确率。

表 2　是否包含注意力模块的模型对比

Tab. 2　Model comparison of whether or not attention modulewas included　　%

是否包含注意力模块	召回率	准确率	调和平均值
否	86.22	68.31	76.22
是	85.19	72.70	78.45

如图 6 所示，注意力模块使得模型在训练过程中收敛更快，并且模型预测任务的损失值略小于未包含注意力模块的模型。这说明增加注意力模块有效减少了像素分类任务以及距离场任务对于图像中非文本区域的错误预测结果。鉴于注意力模块对模型预测准确率的有效提升，回归预测模块采用包含

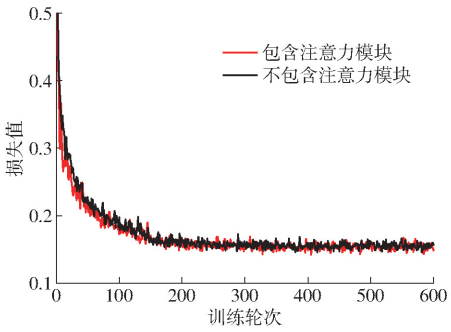


图 6　模型训练过程的损失值对比

Fig. 6　Comparison of loss value during process of model training

注意力模块的架构设计。

3.2.2 不同损失函数的对比

在增加注意力模块的基础上,为了验证距离场回归任务中损失函数修改的有效性,对 4 种距离场损失函数分别进行了实验,实验结果如表 3 所示。其中利用开方自适应增加损失值的方式将准确率提升了 3.80 个百分点。增加权重的方法使得模型在检测性能上没有明显的提升,由于增加权重的方法中权重值大于等于 1,因此一定程度上等效于增加了损失值,但增加边缘区域的损失权重意味着中心区域的损失权重相对降低,进而影响距离场中心区域距离值的预测准确性,弱化了模型预测准确率的提升效果。这也使得同时使用两种方法相较于仅使用开方自适应增加损失值的方式准确率减低了 1.85 个百分点。因此本文最终采用 L_{d_s} 作为模型的距离场回归任务损失函数,易于模型的训练优化,提高模型的预测准确率。

表 3 不同距离场损失函数训练的模型对比

Tab.3 Comparison of model trained with different distance field loss functions %

损失函数	召回率	准确率	调和平均值
L_d	85.19	72.70	78.45
L_{d_w}	85.75	72.42	78.52
L_{d_ws}	87.18	74.65	80.43
L_{d_s}	87.61	76.50	81.68

3.2.3 不同模型方法的对比

为了验证 DFM 对食品标签文本检测的性能,最后将 DFM 与基于语义分割的自然场景文本检测方法 TextSnake[22]、TextField[25]和 PSENet[23]作对比如表 4 所示,实验结果表明 DFM 对食品标签文本检测具有更高的性能,改进前的初始模型准确率较其他模型提升约 2 个百分点,增加注意力模块以及损失损失函数的改进使得 DFM 的召回率、准确率和调和平均值分别为 87.61%、76.50%和 81.68%,比其他模型的召回率提升超过 2 个百分点,调和平均值的提升超过 7 个百分点。相较于改进前,改进后的 DFM 预测结果比其他模型更加准确。

表 4 不同模型方法的对比

Tab.4 Comparison of different models %

方法	召回率	准确率	调和平均值
TextSnake	84.25	66.59	74.38
TextField	85.40	66.37	74.70
PSENet	85.42	66.23	74.61
DFM + L_d (without AM)	86.22	68.31	76.22
DFM + L_{d_s} (with AM)	87.61	76.50	81.68

对食品标签文本检测效果的影响主要在于标签文本检测不准确或者丢失。如图 7 所示,TextSnake[22]模型中利用圆盘参数表征文本区域,其圆盘连接不稳定,极大影响了检测效果。TextField[25]模型根据预测结果生成的文本区域代表像素存在丢失的情况,导致检测文本失败。PSENet[23]模型检测营养成分等较小的文本存在丢失,该方法依赖于标签中心区域的文本和背景二分类预测,对于很小的文本其中心区域极小,很容易被忽略。DFM 对标签文本的检测比其他方法更加准确,PSENet[23]检测到图像中的文本,但并没有将属于同一标签的内容定位为整体,并且丢失了强光照附近的标签内容。



(a) TextSnake (b) TextField (c) PSENet (d) DFM

图 7 不同模型的检测效果对比

Fig.7 Comparison of detection result of different models

4 结束语

针对食品包装上的食品标签检测提出一种基于语义分割的距离场模型,对模型回归预测模块结构和距离场回归任务损失函数的改进使得模型的预测准确率分别提升了 4.39、3.80 个百分点。改进后的 DFM 模型在食品包装图像数据集上的调和平均值达到 81.68%,优于近两年流行的其他自然场景文本检测模型。

参 考 文 献

[1] 张素智,陈小妮,李鹏辉,等. 食品安全大数据的融合及分类技术综述[J]. 计算机技术与发展, 2020,30(2): 159 – 165.
ZHANG Suzhi,CHEN Xiaoni,LI Penghui,et al. Review on food safety big data fusion and classification technology[J]. Computer Technology and Development,2020,30(2):159 – 165. (in Chinese)

[2] 袁云,食品安全监督存在的问题与解决措施分析[J]. 食品安全导刊, 2018(6): 15 – 16.
YUAN Yun. Analysis on the problems and solutions of food safety supervision[J]. China Food Safety Magazine, 2018(6): 15 – 16. (in Chinese)

[3] EPSHTEIN B, OFEK E, WEXLER Y. Detecting text in natural scenes with stroke width transform[C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, 2010: 2963 – 2970.

[4] MATAS J, CHUM O, URBAN M, et al. Robust wide-baseline stereo from maximally stable extremal regions[J]. Image and Vision Computing, 2004,22(10): 761 – 767.

[5] WANG K, BELONGIE S. Word spotting in the wild[C]//European Conference on Computer Vision, 2010: 591 – 604.

[6] TIAN Shangxuan, PAN Yifeng, HUANG Chang, et al. Text Flow: a unified text detection system in natural scene images[C]//IEEE International Conference on Computer Vision, Santiago, 2015: 4651 – 4659.

[7] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, 2005: 886 – 893.

[8] OJALA T, PIETIKAINEN M, MAENPAA T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2002,24(7): 971 – 987.

[9] 王建新,王子亚,田萱. 基于深度学习的自然场景文本检测与识别综述[J]. 软件学报, 2020,31(5): 1465 – 1496.
WANG Jianxin, WANG Ziya, TIAN Xuan. Review of natural scene text detection and recognition based on deep learning[J]. Journal of Software, 2020,31(5):1465 – 1496. (in Chinese)

[10] YAO Cong, BAI Xiang, LIU Wenyu, et al. Detecting texts of arbitrary orientations in natural images[C]//IEEE Conference on Computer Vision and Pattern Recognition, Providence, 2012: 1083 – 1090.

[11] LIU Chunmei, WANG Chunheng, DAI Ruwei. Text detection in images based on unsupervised classification of edge-based features[C]//International Conference on Document Analysis and Recognition, Seoul, 2005: 610 – 614.

[12] BUTA M, NEUMANN L, MATAS J. FASText: efficient unconstrained scene text detector[C]//IEEE International Conference on Computer Vision, Santiago, 2015: 1206 – 1214.

[13] SHIVAKUMARA P, PHAN T, TAN C. A laplacian approach to multi-oriented text detection in video[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011,33(2): 412 – 419.

[14] LIAO Minghui, SHI Baoguang, BAI Xiang. TextBoxes + +: asingle-shot oriented scene text detector[J]. IEEE Transactions on Image Processing, 2018,27(8):3676 – 3690.

[15] 孙红,李松,李民赞,等. 农业信息成像感知与深度学习应用研究进展[J/OL]. 农业机械学报,2020,51(5):1 – 17.
SUN Hong,LI Song,LI Minzan,et al. Research progress of image sensing and deep learning in agriculture[J/OL]. Transactions of the Chinese Society for Agricultural Machinery,2020,51(5):1 – 17. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? file_no=20200501&flag=1. DOI:10.6041/j.issn.1000-1298.2020.05.001. (in Chinese)

[16] LIAO Minghui, ZHU Zhen, SHI Baoguan, et al. Rotation-sensitive regression for oriented scene text detection[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018: 5909 – 5918.

[17] 田萱,王亮,丁琪. 基于深度学习的图像语义分割方法综述[J]. 软件学报, 2019,30(2): 440 – 468.
TIAN Xuan, WANG Liang, DING Qi. Review of image semantic segmentation based on deep learning[J]. Journal of Software, 2019,30(2):440 – 468. (in Chinese)

[18] LONG J, SHELHAMER E, DARREL T. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017,39(4): 640 – 651.

[19] BADRINARAYANAN V, KENDALL A, CIPOLLA R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017,39(12): 2481 – 2495.

[20] CHEN L, PAPANDREOU G, KOKKINOS I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834 – 848.

[21] ZHAO Hengshuang, SHI Jianping, QI Xiaojuan, et al. Pyramid scene parsing network[C]//IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017: 6230 – 6239.

[22] LONG Shangbang, RUAN Jiaqiang, ZHANG Wenjie, et al. TextSnake: aflexible representation for detecting text of arbitrary shapes[C]//European Conference on Computer Vision, 2018: 19 – 35.

[23] LI Xiang, WANG Wenhai, HOU Wenbo, et al. Shape robust text detection with progressive scale expansion network[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, 2019: 9328 – 9337.

[24] DENG Dan, LIU Haifeng, LI Xuelong, et al. PixelLink: detecting scene text via instance segmentation[C]// National Conference on Artificial Intelligence, 2018: 6773 – 6780.

[25] XU Yongchao, WANG Yukang, ZHOU Wei, et al. TextField: learning a deep direction field for irregular scene text detection [J]. IEEE Transactions on Image Processing, 2019,28(11): 5566 – 5579.

[26] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016: 770 – 778.

[27] LIN T, DOLLAR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 2017: 936 – 944.

[28] SHRIVASTAVA A, GUPTA A, GIRSHICK R. Training region-based object detectors with online hard example mining[C]// IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016: 761 – 769.

[29] LIN T, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020,42(2): 318 – 327.