绿洲农业科学与工程　2025，10（1）：160-163
OASIS AGRICULTURE SCIENCE AND ENGINEERING

doi：10.3969/j.issn.2096-2177.2025.01.024

基于双流自适应特征融合的多模态烟草文档分类

孙首名[1]，张琦[1]，王喆[1]，苏娜[2]，沈奇[2]

（1. 辽宁省烟草公司铁岭市公司，辽宁 铁岭 112000；2. 东北大学，辽宁 沈阳 110000）

摘要：针对烟草文档自动化分类的需求，提出一种基于双流自适应特征融合的多模态烟草文档分类网络，名为DSAFFNet。该网络结合烟草文档的文本模态和图像模态，采用DSAFF（Dual-Stream Adaptive Feature Fusion）模块对不同模态特征的重要性自适应调整权重，实现灵活而精确的多模态融合。试验结果表明，所提网络在烟草文档数据集上的表现优于以往分类方法。

关键词：烟草文档分类；多模态学习；双流网络

中图分类号：Q953　　**文献标识码**：A　　**文章编号**：2096-2177（2025）01-160-04

烟草是我国重要经济作物，种植面积和产量均居世界首位。随着烟草行业文档数量激增，传统人工分类方法已难以满足需求，亟需开发智能文档分类技术。一般来说，现有文档图像分类方法主要分为4类：基于结构特征、视觉特征、文本特征和多模态特征融合的方法。其中，结构特征的算法设计较复杂且实现难度大，单一特征的方法都存在明显的局限性。相比之下，多模态方法可以捕获到烟草文档中更为丰富的信息并能识别更为细粒度的特征表示，进而利用文本与图像特征的互补性提升分类效果。因此，本文提出一种基于双流自适应特征融合的多模态烟草文档分类网络。

1　材料与方法

1.1　试验数据集

试验数据集为Tobacco3482[1]，包含3 482张烟草公司商业文档的单页扫描图像，涵盖电子邮件、信函、备忘录等10个类别。随机选取1 500张作为训练集，500张作为验证集，1 482张作为测试集。为了评估多模态网络的效果，使用QS-OCR-Small文本数据集[2]。

1.2　双流自适应特征融合网络

多模态烟草文档分类模型（见图1）通过融合文本流和图像流提取的特征，经过DSAFF模块自适应融合不同模态信息，增强与分类任务相关的信息，同时消除冗余数据并抑制无关信息。

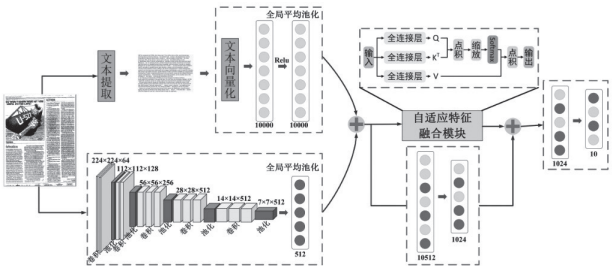

图1　DBSAFNet网络结构图

Fig.1 Network architecture of DBSAFNet

1.3　模态特定编码器

图像流负责处理输入的图像数据并提取特征。烟草图像预处理为224 × 224 × 3的大小，并在此基础上训练CNN。通过VGG16处理后，特征图大小为7 × 7 × 512。为保留全局信息并减少参数量，全局平均池化层用于将特征图缩减为大小为512的特征向量。

文本流负责处理经过文本提取后的数据，词向量大小为10 000。使用TextVectorization进行向量化处理，提取TF-IDF特征。TF-IDF是一种衡量词语在文档中的重要性相对于整个文档集合的方法。最后，进行特征标准化以获取文本特征向量。

收稿日期：2024-12-16

基金项目：国家自然科学基金面上项目—多时延全基因组调控网络的大规模动态概率图建模与分析（62072089）

作者（通信作者）简介：孙首名（1979-），男，汉族，辽宁铁岭人，助理工程师，本科，研究方向：智能烟草数据处理。

1.4　自适应特征融合模块

DSAFF（Dual-Stream Adaptive Feature Fusion）模块主要从图像流和文本流的融合特征中提取关键信息，融合后的特征维度为10 512。该模块在收到特征后，将它们分别映射为Q、K和V，通过计算QK^T来获得图像特征和文本特征之间的相关性。通过动态调整输入特征的权重，DSAFF能够突出对模型决策有显著贡献的重要特征，并过滤掉对最终结果影响较小的部分。其公式如下：

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

其中，Q是查询矩阵，K是键矩阵，V是值矩阵，d_k是键向量的维度，QK^T表示查询矩阵Q和键矩阵K之间的点积。

在实际应用中单头注意力往往扩展到多头注意力，多头注意力公式如下：

$$\text{MultiHead}(Q, K, V) = \text{Concat}(head_1, \cdots, head_h)W^O$$

其中，投影是参数矩阵$W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$，$W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$，$W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_k}$，$W_i^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$。采用$h=8$头注意力，对于其中的每一个头，使用$d_k = d_v = d_{\text{model}}/h = 64$。

1.5　损失函数优化

稀疏分类交叉熵用于测量模型预测分布与真实标签间的差异，适用于整数编码的多分类任务。与传统交叉熵相比，它无需将标签转换为独热编码，具有计算效率高、存储需求低的优势。

2　结果与分析

2.1　试验环境配置

试验在Kaggle平台上完成，所用显卡为16GB显存的NVIDIA Tesla P100，使用Python版本为3.10，网络开发框架TensorFlow 2.15.0。使用Adam算法作为优化器，设置初始学习率为0.000 1，批次大小为16，迭代次数为100个周期，并设置早停机制。

2.2　对比试验

将DSAFFNet与InceptionV3[3]、ResNet50[4]、DenseNet121[5]、Xception[6]和MobileNetV2[7]框架复现的单图像流模型进行对比试验（见表1）。DSAFFNet的分类效果稳定，误分类情况少。与单图像流模型相比，DSAFFNet在深层特征提取和泛化能力上更强。与前人提出的TobaccoV&T[8]的双流模型进行对比（见表1），DSAFFNet在处理分类任务时的表现也更为优异。

表1　对比试验

Tab.1　Comparative experiments

模型 Model	准确率 Accuracy（%）	精确率 Precision（%）	召回率 Recall（%）	F1分数 F1-score（%）
InceptionV3[3]	80.84	80.73	80.84	80.66
ResNet50[4]	83.27	83.10	83.27	83.00
DenseNet121[5]	84.68	84.49	84.68	84.45
Xception[6]	80.84	80.69	80.84	80.47
MobileNetV2[7]	79.69	80.76	79.69	78.60
TobaccoV&T[8]	79.80	-	-	86.00
DSAFFNet	87.92	88.05	87.92	87.88

2.3　混淆矩阵可视化分析

在测试集上评估的混淆矩阵（见图2），Resume类别的表现最佳，准确率达到100%，Email类别的准确率为97%。这2个类别的表现较好，主要由于样本数量较多，特征明确且易于区分。相对而言，Report和Scientific类别的表现较差。主要有2个原因：（1）类别之间存在难以区分的样本，导致模型难以做出准确的分类；（2）类别在数据集中样本量较少，模型无法充分学习到特征。

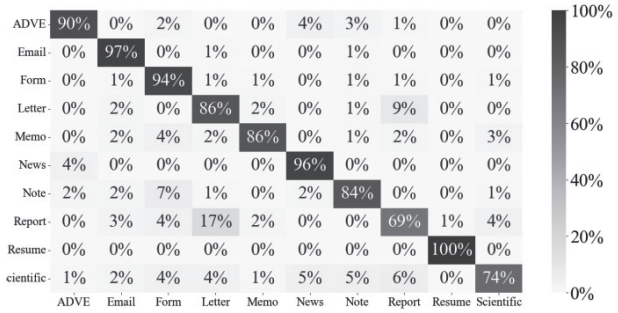

图2　在Tobacco3482测试集上评估的混淆矩阵

Fig.2 Confusion matrix evaluated on Tobacco3482 test set

2.4　消融试验

对模型进行消融试验（见表2），相比单文本流或单图像流模型，文本流和图像流融合模型准确率提升至少4.66%，精确率提升至少4.78%，召回率提升至少4.66%，F1分数提升至少4.75%。DSAFFNet相较于文本流和图像流融合模型，准确率提升1.55%，精确率提升1.65%，召回率提升1.55%，F1分数提升1.6%。以上消融试验的结果进

一步验证了所提出各模块的有效性。

表2 消融试验

Tab.2 Ablation studies

模型 Model	准确率 Accuracy（%）	精确率 Precision（%）	召回率 Recall（%）	F1分数 F1-score（%）
Text_Stream	75.57	76.56	75.57	75.22
Image_Stream	81.71	81.62	81.71	81.53
Text_Stream+Image_Stream	86.37	86.40	86.37	86.28
DSAFFNet	87.92	88.05	87.92	87.88

3 结论

本文提出了一种基于双流自适应特征融合的多模态烟草文档分类网络，其自适应地融合多模态特征，充分从多模态网络中获取重要的特征表示。与多种模型相比，所提方法性能表现得更加优异。该模型可以为烟草文档的分类提供有力技术辅助。

参考文献

[1] Kumar J，Ye P，Doermann D. Structural similarity for document image classification and retrieval[J]. Pattern Recognition Letters，2014，43：119-126.

[2] Audebert N，Herold C，Slimani K，et al. Multimodal Deep Networks for Text and Image-Based Document Classification[C]. Machine Learning and Knowledge Discovery in Databases：International Workshops of Ecml Pkdd，2019，2020：427-443.

[3] Szegedy C，Vanhoucke V，Ioffe S，et al. Rethinking the Inception Architecture for Computer Vision[C]. Proceedings of the IEEE conference on computer vision and pattern recognition（CVPR），2016：2 818-2 826.

[4] He K，Zhang X，Ren S，et al. Deep Residual Learning for Image Recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition（CVPR），2016：770-778.

[5] Huang G，Liu Z，Laurens V D M，et al. Densely Connected Convolutional Networks[C]. Proceedings of the IEEE conference on computer vision and pattern recognition（CVPR），2017：4 700-4 708.

[6] Chollet F. Xception：Deep Learning with Depthwise Separable Convolutions[C]. Proceedings of the IEEE conference on computer vision and pattern recognition（CVPR），2017：1 251-1 258.

[7] Sandler M，Howard A，Zhu M，et al. MobileNetV2：Inverted Residuals and Linear Bottlenecks[C]. Proceedings of the IEEE conference on computer vision and pattern recognition（CVPR），2018：4 510-4 520.

[8] Noce L，Gallo I，Zamberletti A，et al. Embedded Textual Content for Document Image Classification with Convolutional Neural Networks[C]. Proceedings of the 2016 ACM symposium on document engineering，2016：165-173.

Multi-Modal Tobacco Document Classification Based on Dual-Stream Adaptive Feature Fusion

SUN Shouming[1]，ZHANG Qi[1]，WANG Zhe[1]，SU Na[2]，SHEN Qi[2]

（1. Liaoning Tobacco Company Tieling Branch，Tieling Liaoning 112000，China；

2. Northeastern University，Shenyang Liaoning 110000，China）

Abstract：To address the need for automated tobacco document classification，we propose DSAFFNet，a multimodal classification network with dual-stream adaptive feature fusion. The network integrates both textual and visual modalities of tobacco documents，utilizing a DSAFF（Dual-Stream Adaptive Feature Fusion）module to adaptively adjust the importance weights of different modal features，enabling flexible and accurate multimodal fusion. Experimental results show the proposed network outperforms previous classification methods on the tobacco document dataset.

Keywords：tobacco document classification，multi-modal learning，dual-stream network

Fund project：National Natural Science Foundation of China（NSFC），Modeling and Analysis of Large-scale Dynamic Probability Maps for Multi-duration Whole-genome Regulatory Networks（62072089）

Correspondence author：SUN Shouming（1979-），male，Han nationality，from Tieling，Liaoning，assistant engineer，undergraduate，research direction：intelligent tobacco data processing.