面向中文变体字识别的共享权重多模态翻译模型

孙元康[1,2]，李冰[1,2]，李乐翔[1,2]，杨鹏[1,2]，杨冬梅[3]
[1]东南大学计算机科学与工程学院，中国南京市，210000
[2]东南大学计算机网络和信息集成教育部重点实验室，中国南京市，210000
[3]北京科技大学计算机与通信工程学院，中国北京市，100083

摘要：中文变体字识别任务旨在解决中文字符中存在的语义模糊和混淆问题，这些问题对网页内容的安全性构成潜在风险，并加剧敏感词汇管理的复杂性。大多数现有方法在预训练阶段侧重于从中文语料库和词汇中获取上下文语义，往往忽视了中文固有的音韵和形态特征。基于上述问题，本文提出一种面向中文变体字识别的共享权重多模态翻译模型。该模型将拼音的音韵特征和字体的形态特征整合到每个中文词元中，以学习变体文本的深层语义特征。具体来说，通过嵌入层对中文拼音音韵特征进行编码，并利用卷积神经网络学习中文字体形态特征。考虑到中文变体字识别任务中源句与目标句之间的多模态特征相似性，设计了共享权重嵌入机制，在训练过程中利用源句的启发式信息生成目标句。实验结果表明，本文所提出的共享权重多模态翻译模型在双语评估测试（BLEU）和 F1 值方面分别达到 89.550%和 79.480%，与当前最先进的基线模型相比有显著提升。

关键词：中文变体字；多模态模型；翻译模型；音韵和形态
https://doi.org/10.1631/FITEE.2400504