姚瑶, 张森, 杨华英. 基于 Transformer 的外语图书资源的文字识别研究[J]. 智能计算机与应用,2026,16(4):126-130.
DOI:10.20169/j.issn.2095-2163.24060601

基于 Transformer 的外语图书资源的文字识别研究

姚 瑶[1], 张 森[2], 杨华英[1]
(1 南京邮电大学 外国语学院, 南京 210023; 2 南京邮电大学 通信与信息工程学院, 南京 210023)

摘 要: 目前计算机技术以及人工智能已经广泛应用在外语学习中,计算机可以整合多种多媒体资源,并且提供了丰富的智能交互学习环境。本文研究了当前主流外语图书资源在线网站的使用情况,设计了一种基于 Transformer 的外语图书资源文字识别算法,该算法摒弃传统卷积神经网络(如 CNN),采用视觉模型作为编码器(encoder)完成图像特征抽取,利用 BERT 模型作为 decoder 进行文字转换。本文使用 COMICS 英文漫画数据集和 Manga109s 日文漫画数据集训练了英文和日文的 Transformer OCR 模型,训练结果表明,英文 Transformer OCR 模型准确率达 93.57%,日文 Transformer OCR 模型准确率达 89.53%。通过对比其他传统 OCR 模型在 COMICS 数据集上的表现,验证了该文字识别算法在单词基本的准确率上较传统 OCR 模型有明显提升。

关键词: Transformer; 人工智能; 外语学习; 文字识别
中图分类号: TP391.4 **文献标志码**: A **文章编号**: 2095-2163(2026)04-0126-05

Research and application on intelligent learning of foreign languages based on Transformer

YAO Yao[1], ZHANG Sen[2], YANG Huaying[1]
(1 School of Foreign Languages, Nanjing University of Posts and Telecommunications, Nanjing 210023, China;
2 School of Communications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

Abstract: Currently, computer technology and artificial intelligence have been widely used in foreign language learning. Computers can integrate a variety of multimedia resources and provide a rich intelligent interactive learning environment. This paper investigates the application of mainstream online foreign language book resources, and designs a text recognition algorithm for foreign language book resources based on Transformer, which uses ViT visual model as encoder for image feature extraction and BERT model as decoder for text conversion. For COMICS English comics dataset and Manga109s Japanese comics dataset, the English and Japanese Transformer OCR model is trained, auxiliary screen word recognition and query are carried out, and the training shows that the English Transformer OCR model reaches 93.57% accuracy. The Japanese Transformer OCR model achieves 89.53% accuracy. By comparing with the performance of other traditional OCR models on the COMICS dataset, it is verified that the proposed text recognition algorithm has a signficant improvement in the accuracy of word recognition compared with traditional OCR models.

Key words: Transformer; AI; foreign language learning; OCR

0 引 言

随着计算机技术的迅速发展,许多教师开始利用电子外文书籍作为外语课堂之外的补充以及学生的自主学习工具。电子书籍具有灵活性与便捷性,学习者通过电脑或者手机就可以随时随地进行外语学习,为外语学习者提供了不同以往的阅读体验[1],目前常见的免费外文书籍阅读平台包括 Audible、Gutenberg、OpenLibrary[2]等。这些平台包含了多种类型的外文阅读资源,在电子书籍阅读时嵌入文字识别功能可极大提高学习的效率以及简化单词查询的过程。然而传统的 OCR 模型在处理漫画文本时面临诸多挑战,

基金项目: 南京邮电大学 2024 年实验室工作研究课题(2024XSG10)。
作者简介: 张 森(2003—),男,本科生,主要研究方向:网站编程;杨华英(1978—),女,学士,助理研究员,主要研究方向:英语教育,实验室技术。
通信作者: 姚 瑶(1993—),女,硕士,助理实验师,主要研究方向:人工智能,网站技术。Email:yaoyao@njupt.edu.cn。
收稿日期: 2024-06-06

如图像背景复杂，字体多变等问题，导致与绘本和图画相关的文字识别准确率不高。因此，本文基于 Transformer 模型的序列到序列核心特性，提出一种高效的外语图书文本 OCR 算法。

日常生活场景中的文本识别已经成为计算机视觉和模式识别领域的活跃研究领域[3]，传统 OCR 多采用深度学习技术，Prommas 等[4]使用 CNN 网络模型实现了手写体识别，Shinde 等[5]使用卷积循环神经网络 CRNN 实现了票据手写体识别。受图像文本自身或扫描机器的噪声影响，OCR 模型可能会导致错误识别结果[6]。Transformer 是一种用于处理序列数据的模型，可用来解决传统循环神经网络（RNN）以及长短期记忆网络（LSTM）中的一些限制。Transformer 最初由 Vaswani 等[7]于 2017 年提出，并且在机器翻译任务中取得了显著成功。Transformer 在 NLP 领域的广泛应用，成功的启发了其在计算机视觉（CV）领域的应用，Google 团队 Dosovitskiy 等[8]提出基于 Transformer 的视觉模型 ViT（Vision Transformer），其不依赖 CNN 结构，并大幅度节约了算力，在 ViT 的基础上，Touvron 等[9]改进了 ViT 模型，在一个大规模的 ImageNet－1k 图像数据集上以监督学习的方式进行了预训练和微调，输入图像分辨率为 224×224 像素，输出为图像的特征。Devlin 等[10]提出了 BERT，BERT 是一个预训练的语言模型，可用于多种 NLP 任务，如机器翻译、语言理解、语言生成等[11]。为解决训练以及推理时运行速度，Turc 等[12]提出了不区分大小写的自监督预训练 BERT 模型，该模型是经知识蒸馏后（即减小模型参数量）的 BERT 模型。roBERTa[13]是 BERT 的改进版本，通过更大的数据训练量和 batch size，可更好地进行下游任务推广。对于基于 Transformer 的 OCR 算法，Li 等[14]提出了 TrOCR 算法，将 ViT 模型作为 Encoder，BERT 模型作为 Decoder 完成图像到文字的转换。

本文基于 TrOCR 架构设计并实现了一种基于 Transformer 模型的外语图书文字识别算法，实现了英语、日语双语的图书资源文字识别功能。针对不同类型，不同清晰度的外文漫画，均取得较好的识别效果，提升了学生学习外语的效率，扩展了外语阅读的相关视野。

1　文字识别模型设计与实现

如图 1 所示，本文基于端到端的 TrOCR 模型，采用 Transformer 编码器－解码器结构，编码器选择预训练的 CV 模型（ViT－style）提取图像块特征表示，解码器选择两种不同语言的预训练 NLP 模型（Bert－style），指导生成单词片段序列。

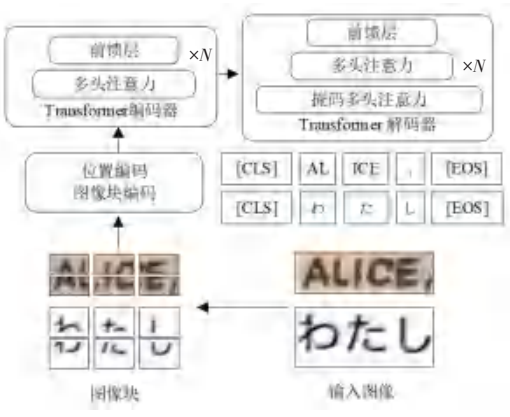


图 1　绘本 OCR 模型架构图

Fig. 1　Architecture of comics OCR model

1.1　Embedding 层设计

Transformer 模型无法直接将图像数据作为输入，因此需要对图像做预处理，并将其作为编码器输入。

首先将原始图像 $\Re^{3\times H_0\times W_0}$ 随机进行图像像素级变换，随机变换包括模糊、高斯噪声、旋转、锐化、对比度图像压缩、随机下采样等，图像随机变换流程如图 2 所示。

图 2　图像基础变换图

Fig. 2　Flow of image transforms

图像转换的目的在于增加文字识别算法的鲁棒性，对于输入的图像，设置图像变换顺序流程为旋转、下采样、模糊、对比度调整、高斯噪声、图像压缩以及灰度变换。为增加随机性，根据图像变换的程度分为不变组、轻量组和中度组，不变组不对图像做变换操作，轻量组和中度组的旋转角度、模糊系数、对比度变换值、噪声大小不同，图像变换以及分布概率设置的表达式如下：

$$\begin{aligned} \text{image}T &= \text{Transform}_i(\text{image}),\ p = [0.3, 0.6, 0.1] \\ \text{Transform}_0 &= \text{None} \\ \text{Transform}_{1,2} &= \text{compose}([\text{rotate},\ \text{downscale},\ \text{blur},\ \text{contrast},\ \text{gaussNoise},\ \text{gray}]) \end{aligned} \quad (1)$$

图像变换后,需要对图像进行张量转换,首先转换为 (H,W) 大小,转换后将图像分解为 N 个固定尺寸 (P,P) 的图像块,为保证处理速度,取 H,W 为 224,图像块尺寸 P 为 16, N 为 14,每一个图像块可作为一个词向量,共有 14×14 = 196 个词向量,每个词向量的维度为 16×16×3 = 768。转换公式如下式:

$$N = HW/P^2 \tag{2}$$

将每一个图像块信息做线性投影,转换为 D 维张量(D 为隐藏尺寸 768),同时随机生成[cls]token 在张量起始处进行拼接。图像块位置信息编码(position embedding)作为基础 Transformer 模型的序列输入,通过组合图像块信息和图像块位置信息后,完成模型的 embedding,过程可表示如下;

$$E_p = [x_{\text{cls}};x_p^1E;x_p^2E,\cdots,x_p^NE] \tag{3}$$

$$Z = E_p + E_{\text{pos}} \tag{4}$$

1.2 编码器设计

Tranformer 的编码器主要为多头注意力机制,自注意力机制可分成 3 个可学习的权重矩阵: $\boldsymbol{W}^{\boldsymbol{Q}}$、$\boldsymbol{W}^{\boldsymbol{K}}$、$\boldsymbol{W}^{\boldsymbol{V}}$, 对于 embedding 层输入,传入全连接层后,每一个 Z 与学习矩阵相乘,得到 3 个向量: $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$,表示如下:

$$\boldsymbol{Q} = \boldsymbol{Z}\boldsymbol{W}^{\boldsymbol{Q}},\boldsymbol{K} = \boldsymbol{Z}\boldsymbol{W}^{\boldsymbol{K}},\boldsymbol{V} = \boldsymbol{Z}\boldsymbol{W}^{\boldsymbol{V}} \tag{5}$$

多头注意力机制示意图如图 3 所示。

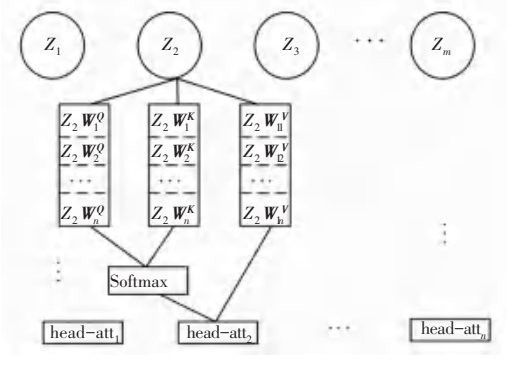


图 3 多头注意力机制图

Fig. 3 Illustration on multi-head attention

将每一个 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 均分别通过相应矩阵计算,得到多头注意力,其中 d_k 为词向量的维度,计算过程如下式:

$$\text{head}-\text{att}_i = \text{Softmax}\left(\frac{\boldsymbol{Z}\boldsymbol{W}_i^{\boldsymbol{Q}}\cdot(\boldsymbol{Z}\boldsymbol{W}_i^{\boldsymbol{K}})^{\text{T}}}{\sqrt{d_k}}\right)\times\boldsymbol{Z}\boldsymbol{W}_i^{\boldsymbol{V}} \tag{6}$$

$$\text{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Concat}(\text{head}-\text{att}_1,\cdots,\text{head}-\text{att}_h)\boldsymbol{W}^{\boldsymbol{O}} \tag{7}$$

经过多头注意力层后,还添加了一层前馈层,前馈层为两层全连接层,连接后作为编码器输出。

1.3 解码器设计

解码器结构与编码器类似,具有多头注意力层和前馈层。不同的是,解码器多了一层掩码多头注意力层。解码器自注意力机制利用带有 mask 掩码的操作防止在训练期间获取不需要的未来信息,从而确保模型仅依赖当前已知的信息。掩码矩阵为下三角矩阵,如下所示:

$$\begin{bmatrix}1&0&0&0\\1&1&0&0\\1&1&0&0\\1&1&1&1\end{bmatrix}$$

掩码矩阵作用在 Softmax 之前,在 $\boldsymbol{Q}\boldsymbol{K}^{\text{T}}$ 之后,计算公式表示如下:

$$\text{MultiHeadMasked}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{Softmax}(\boldsymbol{Q}\boldsymbol{K}^{\text{T}}\boldsymbol{M})\boldsymbol{V} \tag{8}$$

其中, $\boldsymbol{M}$ 为掩码矩阵。

因此在使用注意力矩阵来预测下一个词的概率分布时,仅使用前面词汇的信息。

由于解码器的输出需要右移一个位置,仅需要考虑当前输入向量的左侧部分, 即输入位置要小于 i,公式如下:

$$\boldsymbol{h}_i = \text{Proj}(\text{Emb}(\text{Token}_i)) \tag{9}$$

其中, $\boldsymbol{h}_i$ 为第 i 个位置 token 的嵌入向量。

通过 Softmax 函数计算输出词汇的概率, 计算维度为 V,具体计算过程如下:

$$\sigma(h_{ij}) = \frac{e^{h_{ij}}}{\sum_{k=1}^{V}e^{h_{ik}}},j = 1,2,\cdots,V \tag{10}$$

最后通过束搜索算法对词汇表(vocabulary)索引进行筛选,输出正确的文字识别结果。

2 实验及结果分析

2.1 数据集

针对外语漫画学习的文字识别 OCR 数据集包含以下两个数据源:

英文黄金时代 COMICS 数据集[15]:由 1930 年代开始直到 1950 年代中期(美国漫画的黄金时代),涵盖 4 000 本大量的超级英雄、浪漫主义题材的漫画,从中抽取 17 110 张对话框图片以及对应的文本信息。

日文 Manga-109 数据集[16]:收集由 1970 年代至 2010 年代的 109 本日本漫画,涵盖多种题材。数据集中标注了多种类型,本文仅利用标注的对话框

和文本信息，共计 113 664 张。

两个数据集的样式和对应文字见表 1。

表 1　英文 COMICS 与日文 Manga109 数据集样例

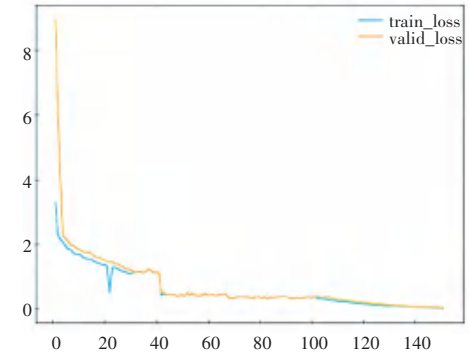
Table 1　Examples of English COMICS and Japanese Manga 109 datasets

数据集	漫画数	图片数	示例图片	对应文字
COMICS	4 000+	17 110		ALICE
Manga109	109	113 664		わたし

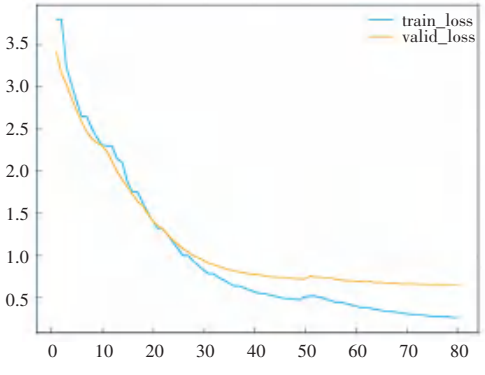
2.2　实验设置及结果

训练过程运行在 Ubuntu 系统，硬件配置内存 64 GB、GPU 为 NVIDIA L4，显存 24 GB，根据数据量不同，为防止过拟合，分别设置 COMICS 数据集和 Manga109 s 数据集的训练轮次（epoch）为 150 和 80，训练集和测试集比例为 0.8∶0.2。

定义损失函数为交叉熵函数（CrossEntropy），两种数据集训练过程中 loss 均呈现下降趋势，具体训练过程中每一轮的训练损失（validation）以及验证损失（evaluate loss）如图 4 所示。



（a）英文训练 OCR loss



（b）日文训练 OCR loss

图 4　训练过程 loss 图

Fig. 4　Loss curve of training and validation per epoch

训练完成后，在忽略标点符号的情况下，对字符串级别计算整个数据集上的准确率，最终的准确率（Accuracy）和损失值（loss）具体值见表 2。

表 2　算法的准确率与损失值

Table 2　Loss and accuracy of algorithm

数据集	Epoch	训练 loss	评估 loss	Accuracy
COMICS	150	0.020 0	0.182 3	0.935 7
Manga109	80	0.254 8	0.640 6	0.895 3

受限于巨大的训练成本以及 GPU 算力，日文漫画 OCR 可对比的深度学习算法较少，英文漫画数据集 COMICS 在其他常用 OCR 模型训练后，包括 SATRMsm[17]、ABINet[18]、CRNN－TPS[19] 以及 CRNN[20] 算法。通常单词识别难度要高于单一字符识别，本文算法和其他算法在单词层面上的准确率表现见表 3。

表 3　单词准确率对比表

Table 3　Comparisons of word accuracy

OCR 模型	Word Accuracy
本文模型	0.935 7
SATRNsm	0.930 0
ABINet	0.723 7
CRNN-TPS	0.708 3
CRNN	0.697 1

由表 3 可知，本文提出的 OCR 模型在英文 COMICS 漫画数据集上的单词识别准确率达到 93.57%，显著优于其他英文 OCR 模型。

2.3　应用实例

如图 5 所示，在数据集中随机抽取 3 个待识别的图像以及识别文字展示如下，对于横版且清晰度较差的英文单词有着较好的识别结果。而日文单词中，背景复杂，有边框干扰时也可有着较好的识别结果。

图 5　OCR 应用示例

Fig. 5　Examples of OCR application

3　结束语

本文成功实现了一种基于 Transformer 的外文图书资源文字识别算法，并探索了该算法在外语学

习场景中的实际应用价值。通过一系列的实验和评估,证明该算法能够有效提取文字信息,可辅助外语学习者提高阅读理解能力。但是本算法仍存在以下不足:

(1)面对超大规模数据集时模型的识别准确率仍有提升空间,后续可通过引入更大规模的预训练数据、优化模型轻量化结构来解决该问题;

(2)模型的计算速度尚未达到实时识别的要求,可通过模型剪枝、量化等轻量化技术进一步优化推理性能。

后续研究将持续对算法进行迭代与更新,重点围绕大数据集适配性与实时推理性能展开优化,以期为外语学习提供更全面的技术支撑。

参考文献

[1] 张浩,钱冬明,祝智庭. 电子阅读方式分类研究[J]. 中国电化教育, 2011,32(9):5. DOI:10. 3969/j. issn. 1006-9860. 2011. 09. 006.

[2] BRIGGS A, BURKE P. A social history of the media: From gutenberg to the internet[J]. European Journal of Communication, 2010, 25(4):423-424.

[3] CHEN X, JIN L, ZHU Y, et al. Text recognition in the wild: A survey[J]. ACM Computing Surveys, 2022,55(2):54.

[4] PROMMAS S, SIRIBORVORNRATANAKUL T. CNN - based Thai handwritten OCR: An application for automated mail sorting [J]. International Journal of Information Technology, 2024, 16 (2):16-28.

[5] SHINDE S, SARAIYA T, JAIN J, et al. Using CRNN to perform OCR over forms[J]. International Journal of Engineering Research and Technology, 2021,10(9):215-221.

[6] NGUYEN T, JATOWT A, COUSTATY M, et al. Survey of Post-OCR processing approaches [J]. ACM Transactions on Knowledge Discovery from Data (TKDD), 2021,15(6):1-32. DOI:10. 1145/3453476.

[7] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//Proceedings of Advances in Neural Information Processing Systems. NeurIPS, 2017:5998-6008.

[8] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv,2010. 11929, 2020.

[9] TOUVRON H, CORD M, DOUZE M, et al. Training data-efficient image transformers & distillation through attention[C]// Proceedings of the International Conference on Machine Learning. PMLR, 2021: 10347-10357.

[10] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv preprint arXiv,1810. 04805, 2018.

[11] KOROTEEV M. BERT: A review of applications in natural language processing and understanding [J]. Artificial Intelligence Review,2022,55(3):2091-2148. DOI:10. 48550/arXiv. 2103. 11943.

[12] TURC C I, CHANG M W, LEE K, et al. Well-read students learn better: On the importance of pre-training compact models [J]. arXiv preprint arXiv,1908. 08962, 2019.

[13] LIU Y, OTT M, GOYAL N, et al. Roberta: A robustly optimized bert pretraining approach [J]. arXiv preprint arXiv, 1907. 11692, 2019.

[14] LI M, LV T, CUI L, et al. TrOCR: Transformer-based optical character recognition with pre-trained models[J]. arXiv preprint arXiv,2109. 10282,2021. DOI:10. 48550/arXiv. 2109. 10282.

[15] IYYER M, MANJUNATHA V, GUHA A , et al. The amazing mysteries of the gutter: Drawing inferences between panels, in comic book narratives[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway,NJ:IEEE, 2017:1639-1648. DOI:10. 1109/CVPR. 2017. 686.

[16] AIZAWA K, FUJIMOTO A, OTSUBO A, et al. Building a manga dataset " Manga109 " with annotations for multimedia applications[J]. IEEE MultiMedia, 2020, 27(2):8-18. DOI:10. 1109/MMUL. 2020. 2987895.

[17] LEE J, PARK S, BAEK J, et al. On recognizing texts of arbitrary shapes with 2D self-attention[J]. arXiv preprint arXiv, 1910. 04396,2019. DOI:10. 48550/arXiv. 1910. 04396.

[18] FANG S, XIE H, WANG Y, et al. Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition [J]. arXiv preprint arXiv, 2103. 06495, 2021. DOI:10. 48550/arXiv. 2103. 06495.

[19] SHI B, WANG X, LYU P, et al. Robust scene text recognition with automatic rectification [C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2016: 4168-4176. DOI: 10. 1109/CVPR. 2016. 452.

[20] SHI B, BAI X, YAO C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017,30(11):2298-2304. DOI:10. 1109/TPAMI. 2016. 2646371.