基于Nougat的多行数学公式识别

基于Nougat的多行数学公式识别
1、背景
2、方法综述
3、模型描述
4、结果和讨论
5、结论
参考文献

1、背景

数学表达式是重要的信息来源，特别是对于科学研究、监管部门和教育部门。从文档图像中自动识别数学表达式具有广泛的下游应用，如自动评分、办公自动化和学术辅助。多行数学表达式在这些表达式中占很大比例，由于符号之间复杂的空间关系，使得多行数学公式识别具有一定挑战性。

2、方法综述

近年来，基于编解码器的深度学习数学公式识别领域的主流框架。根据解码器的结构，过去的深度学习方法可以分为基于RNN模型和基于Transformer模型的方法。此外，基于解码策略，这些方法可以分为基于序列解码的方法和基于树形结构解码的方法。

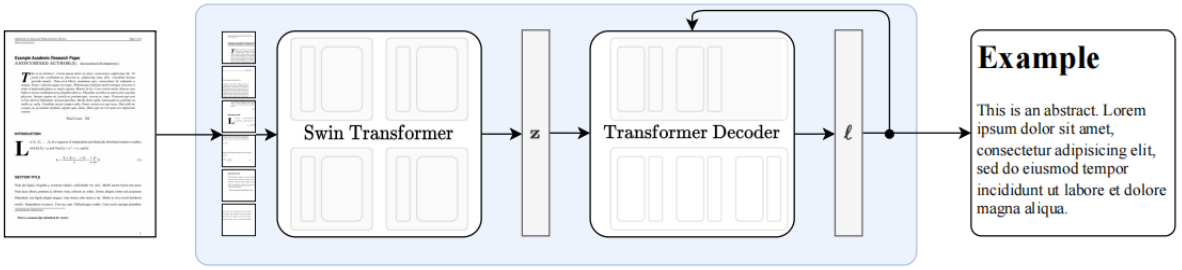
基于RNN的方法 在2017年，Zhang等人提出了一个端到端深度学习模型，WAP[1]，来解决HMER的问题。该模型的编码器部分是一个类似于VGGnet [2]的全卷积神经网络。解码器部分使用GRU [3]模型从提取的视觉特征中生成预测的LATEX序列。WAP不仅避免了由不准确的符号分割引起的问题，而且还消除了手动预定义的LATEX语法，从而成为后续深度学习方法的基准模型。在WAP之后，Zhang等人进一步提出了DenseWAP [4]，它用DenseNet[5]取代了VGGnet。后续的工作通常采用DenseNet 作为编码器的主干网络。CAN[6]模型引入了一个多尺度计数模块，利用符号计数任务作为辅助任务，与表达式识别任务进行联合优化。

基于Transformer的方法 为了缓解输出不平衡的问题，并充分利用双向语义信息，BTTR[7]在基于变压器的解码器的基础上采用了双向训练策略。随后，CoMER [8]将覆盖信息整合到Transformer解码器中，并引入了一个注意力细化模块。该模块利用来自Transformer解码器内的多头注意机制的注意权值来计算覆盖向量，同时保持并行解码的特性。基于CoMER，GCN[9]整合了额外的符号分类信息，利用通用类别识别任务作为与HMER任务联合优化的补充任务，取得了显著的性能。然而，GCN引入的类别识别任务需要手动构建符号类别，并且仅限于特定的数据集。基于CoMER，ICAL[10]引入了隐式字符构造模块（ICCM）来对隐式字符信息进行建模，有效地利用了LATEX中的全局信息，同时提出了融合模块来聚合ICCM的输出，从而校正Transformer的预测。

基于树形结构解码的方法 LATEX作为一种标记语言，由于方括号等分隔符的影响，可以很容易地解析为树状表达式。因此，通过利用数学表达式固有的二维结构，模型可以为预测过程提供一定的可解释性。Zhang等人提出了DenseWAP-TD [11]，它用一个基于二维树状结构的解码器取代了直接回归LATEX序列的GRU解码器。TDv2模型[12]在训练过程中，对同一LATEX字符串使用不同的转换方法，削弱了上下文依赖性，赋予解码器更强的泛化能力。SAN模型[13]将LATEX序列转化为解析树，并设计了一系列的语法规则，将预测LATEX序列的问题转化为树的遍历过程。此外，SAN还引入了一个新的语法感知注意模块，以更好地利用LATEX中的语法信息。

可视化文档理解（VDU）是深度学习数学公式识别领域研究的另一个相关主题，主要关注于提取多种文档类型的相关信息。以前的工作依赖于预先训练的模型，这些模型通过使用Transformer架构联合建模文本和布局信息来提取信息。LayoutLM[14]模型家族使用掩蔽布局预测任务来捕获不同文档元素之间的空间关系。在分数、指数和矩阵等数学符号中，字符的相对位置是至关重要的，为此，Nougat[15]，一个基于Transformer的模型，它可以将文档页面的图像转换为格式化的标记文本，这是一个能够将PDF转换为轻量级标记语言的预训练模型。

3、模型描述



Nougat是一个编码器-解码器Transformer架构，允许端到端的训练过程。在Donut[16]架构的基础上，不需要任何与OCR相关的输入或模块，文本会被网络隐式识别。

编码器 编码器接收一个多行公式图像$x \in R^{3\times H_0\times W_0}$，裁剪边缘，并调整图像大小到（560，560）。如果图像小于矩形，则添加额外的填充，以确保每个图像具有相同的维数。我们使用一个Swin Transformer[17]，它将图像分割成固定大小的不重叠窗口，并应用一系列自注意层来聚合这些窗口之间的信息。

解码器 编码的图像z使用交叉注意的Transformer解码器架构解码成一系列token序列。这些token以自回归的方式生成，使用自注意和交叉注意分别关注输入序列和编码器输出的不同部分，最终输出解码结果。解码器使用了mBART[18]来实现解码。

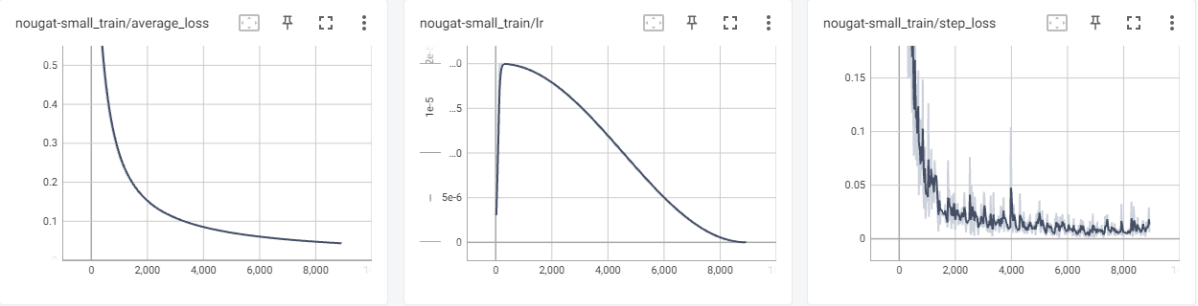
4、结果和讨论

目前，我们将15000数据集按照9：1进行切分。

设置超参如下：

参数	值
input	3x560x560
max_len	1050
dict_len	339
lr	2e-05
optimizer	adamw
weight_decay	0.05
beta	[0.9,0.98]
eps	1e-06
epoch	20

目前训练20个epoch，并达到最优，训练过程如下：



最终,我们在验证集上的测试结果如下：

Expression recall	Character recall
0.66	0.98

5、结论

实验证明，将文档理解用于多行公式识别具有良好的效果，对未来公式识别模型改进提供了新的思路。

参考文献

1. Zhang, J., Du, J., Zhang, S., Liu, D., Hu, Y., Hu, J., Wei, S., Dai, L.: Watch, attend and parse: An end-to-end neural network based approach to handwritten mathematical expression recognition. Pattern Recognition 71, 196–206 (2017)
2. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
3. Cho, K., Van Merri¨enboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
4. Zhang, J., Du, J., Dai, L.: Multi-scale attention with dense encoder for handwritten mathematical expression recognition. In: 2018 24th international conference on pattern recognition (ICPR). pp. 2245–2250 (2018)
5. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
6. Li, B., Yuan, Y., Liang, D., Liu, X., Ji, Z., Bai, J., Liu, W., Bai, X.: When counting meets hmer: counting-aware network for handwritten mathematical expression recognition. In: European Conference on Computer Vision. pp. 197–214. Springer (2022)
7. Zhao, W., Gao, L., Yan, Z., Peng, S., Du, L., Zhang, Z.: Handwritten mathematical expression recognition with bidirectionally trained transformer. In: Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16. pp. 570–584. Springer (2021)
8. Zhao, W., Gao, L.: Comer: Modeling coverage for transformer-based handwritten mathematical expression recognition. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII. pp. 392–408. Springer (2022)

9. Zhang, X., Ying, H., Tao, Y., Xing, Y., Feng, G.: General category network: Handwritten mathematical expression recognition with coarse-grained recognition task. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
10. Zhu J, Gao L, Zhao W. ICAL: Implicit Character-Aided Learning for Enhanced Handwritten Mathematical Expression Recognition[J]. arXiv preprint arXiv:2405.09032, 2024.
11. Zhang, J., Du, J., Yang, Y., Song, Y.Z., Wei, S., Dai, L.: A tree-structured decoder for image-to-markup generation. In: ICML. p. In Press (2020)
12. Wu, C., Du, J., Li, Y., Zhang, J., Yang, C., Ren, B., Hu, Y.: Tdv2: A novel tree□structured decoder for offline mathematical expression recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 2694–2702 (2022)
13. Yuan, Y., Liu, X., Dikubab, W., Liu, H., Ji, Z., Wu, Z., Bai, X.: Syntax-aware network for handwritten mathematical expression recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4553–4562 (2022)
14. Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. LayoutLMv3: Pre-training for Document AI
with Unified Text and Image Masking, July 2022. URL http://arxiv.org/abs/2204.08387. arXiv:2204.08387 [cs].
15. Blecher L, Cucurull G, Scialom T, et al. Nougat: Neural Optical Understanding for Academic Documents,(2023)[J]. arXiv preprint arXiv:2308.13418.
16. Geewook Kim, Teakgyu Hong, Moonbin Yim, Jeongyeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang,
Sangdoo Yun, Dongyoon Han, and Seunghyun Park. OCR-free Document Understanding Transformer, October
2022. URL http://arxiv.org/abs/2111.15664. arXiv:2111.15664 [cs].
17. Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin
Transformer: Hierarchical Vision Transformer using Shifted Windows, August 2021. URL http://arxiv.org/abs/
2103.14030. arXiv:2103.14030 [cs].
18. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoy□anov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Genera□tion, Translation, and Comprehension, October 2019. URL http://arxiv.org/abs/1910.13461. arXiv:1910.13461
[cs, stat].