Mathematical Formula Identification Method Based on Transformer

Zhongxi Chen,36920221153073
Zhipeng Qian 36920221153109
Xuyang Song,36920221153112
Zhiyao Wang,36920221153123
Wei Yu,36920221153142
[1]AI Class

Abstract

Mathematical formulas are crucial in academic research. For those who need to write LaTeX, it is necessary to be able to identify formulas quickly and conveniently. LaTeX formula recognition can be viewed as a subtask of image recognition. With the development of deep learning technology, the previous traditional character recognition methods can no longer meet our requirements. This paper proposes a new framework to recognize formulas in images, based on the model consisting of a ViT encoder with a ResNet backbone and a Transformer decoder. On this basis, we added the hybrids, beam search, data enhancement and preprocessing network resizer to improve the accuracy of formula recognition. Finally, experiments are conducted to verify the improvement of model performance on the two metrics of BLEU and ACC.

Introduction

Mathematical formulas are widely used in education, science and engineering. For master and doctoral students, especially engineering students, when using Latex to write course papers or project summary papers and reports, they often encounter very complex mathematical formulas, which are also very time-consuming when completing the final paper. Sometimes, we need to refer to some formulas that have been used by others, but it is also inconvenient to edit in MathType. If we can have a tool to recognize screenshots or handwritten formulas and automatically convert them into LaTeX code expressions, it will greatly facilitate our use of LaTeX mathematical formulas. Mathematical Expression Recognition (MER) technology is based on this problem. In this paper, we study the image recognition method of mathematical formula based on Vision Transformer.

With the popularization and development of computer technology, fast and accurate character recognition method has become the demand. For the recognition of ordinary texts in literature and books, there are already relatively mature Optical Character Recognition (OCR) methods, but the application of this technology to the recognition of mathematical formulas is not effective. For character recognition, handwritten characters have various styles, rich handwriting information, and need to pay attention to context information. What's more, mathematical formulas have many special symbols, and it is difficult to distinguish between symbols. At the same time, mathematical expressions have complex structures. Many mathematical formulas have two-dimensional structures, such as subscripts, summation formulas, limit formulas, open formulas, fractions, series, and other forms. These forms have multi-layer structures nested, and different spatial structures, which greatly increase the difficulty of identification. Therefore, it is necessary to find an effective method to identify the numerical formula.

The mathematical formulas in current websites and articles are mostly in LaTeX format. LaTeX is a Tex based typesetting system with powerful and beautiful formula typesetting and simple and universal rendering. It can not only apply mathematical formulas to LaTeX document typesetting, but also apply LaTeX to the display of mathematical formulas on web pages. Therefore, our research content is to convert the input mathematical formula picture into LaTeX formula. To realize such a recognition system, three functions need to be realized: formula detection, formula recognition and formula retrieval. The traditional formula recognition method is carried out step by step. The recognition of mathematical expressions is divided into symbol segmentation and structure analysis. This method has poor generalization performance and low accuracy. With the development of deep learning technology, more and more end-to-end learning methods have achieved significant improvement in traditional tasks.

We have adopted the Vision Transformer (ViT) method in this article. ViT applies the Transformer model to the vision field, completely adopts the standard structure of Transformer in structure, and has achieved the most advanced level in many benchmark tasks of image recognition. The research on the recognition of mathematical formulas is not only conducive to the reuse and editing of mathematical formulas in the form of images in the literature, but also can check academic misconduct through the retrieval and duplicate checking of mathematical formulas, which can make up for the shortcomings of the current duplicate checking system.

Related work

Handwritten Formula Recognition (HMER) or Mathematical Formula Recognition (MER) can be said to be a

special optical character recognition (OCR) task. The input formula images have superscript and subscript, special symbols, and nested structures and other intractable structures. The image is then represented as a structured language or markup that defines both the text itself and its presentation semantics.

Despite the great successes of the current OCR systems, MER(mathematical expression recognition)still remains a very challenging problem due to the complex structures of formulas.

OCR

Existing OCR approaches could be roughly divided into word-level classification based, sequence-to-label-based and sequence-to-sequence-based methods(Li, Wang, and Shen 2017). Now sequence-to-sequence-based methods are flourishing for their excellent performance in exploiting the context. With the development of RNN techniques, several approaches (Liao et al. 2017; He et al. 2016b) propose deep recurrent models to encode the output features of CNN and adopt CTC as the text decoder. In (Lee and Osindero 2016), they utilize an attention-based sequence-to-sequence framework to focus on specified CNN features when decoding individual characters. For challenging irregularly shaped text, and this work (Shi, Bai, and Yao 2016) utilizes a spatial transformer network to correct the input image. The above works focus on the recognition of single-line text.

Image Caption

MER task is suitable to be solved using Image Caption, similar to automatic caption generation tasks (Xu et al. 2015) and image-to-markup systems (Deng et al. 2017). Unlike traditional Text OCR tasks, which assume left-to-right sequential text localization, Image Caption allows the neural network to focus its attention wherever it contributes to formula recognition. There has also been work on applying Image Caption to MER, (Deng et al. 2017; Hu et al. 2020) using the sequence-to-sequence model of the encoder-decoder structure, using a scalable coarse-to-fine attention mechanism based on converting images into formula text, In (Li et al. 2022), they add a symbol counting module to constrain the pairs of symbols, and this work (Zhao and Gao 2022) proposed a way to use coverage information in the Transformer's decoder to refine the attention weights. The above work has achieved good results in the MER task.

Image Recognition

Image recognition is a technology that is capable of identifying places, people, objects and many other types of elements within an image, and drawing conclusions from them by analyzing them. Handwritten formula recognition is also a subtask of image recognition. With the development of deep learning methods, image recognition methods based on deep learning have also begun to emerge.

The CNN network was first used in the task of handwritten digit recognition, e.g, AlexNet (Krizhevsky, Sutskever, and Hinton 2012), ZFnet (Zeiler and Fergus 2014), VGG-16 (Simonyan and Zisserman 2014), ResNet (He et al. 2016a), etc. With the optimization of network depth and structure, the image recognition performance based on CNN method has gradually reached a bottleneck. People began to think about whether some methods could be borrowed from other fields and applied to the image field.

In the field of NLP, the proposal of the RNN encoder-Decoder (RNNenc)(Cho et al. 2014b) has pioneering significance for NLP tasks to solve sequence-to-sequence problems. The encoder is responsible for encoding a variable-length sequence into a fixed-length vector representation, and the decoder is responsible for converting a given fixed-length vector representation back into a variable-length sequence. The article mainly applies rnn encoder-Decoder to machine translation tasks.

Self-attention architectures, especially transformers, have become the model of choice for natural language processing, capable of pre-training on large textual corpora and then fine-tuning on smaller task-specific datasets. Transformer (Vaswani et al. 2017) follows the encoder-decoder structure, which abandons the traditional CNN and RNN neural networks. The entire network structure consists of the Attention mechanism and the feedforward neural network. Moreover, the performance of machine translation has been greatly improved. However, this method is suitable for processing sequence information, and how to apply it to image processing tasks has become a problem.

Inspired by the transformer, we want to be able to apply it to image tasks with as little modification as possible. The vision transformer (Dosovitskiy et al. 2020) is able to achieve excellent results when pre-trained at a sufficient scale and transferred to tasks with less data. When processing an image, the input image x is first converted into a two-dimensional patch sequence, that is, the image is divided into multiple fixed-size patches, and then these patches are tiled and mapped to the D dimension, and encoded by the Transformer. The overall structure of ViT can refer to the Figure1. Moreover, The hybrid structure is mentioned in the article. The author proposes the hybrids, which apply the patch embedding map to the patches extracted by the CNN feature map, such as ResNet, etc, which can further improve performance.

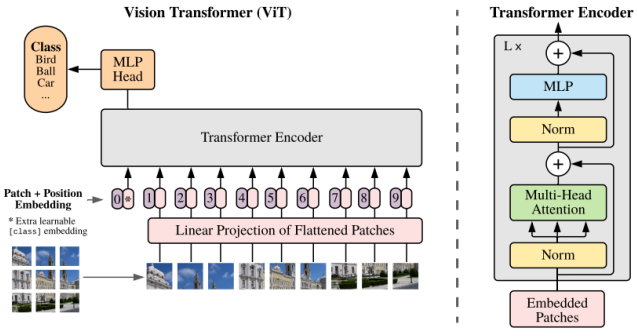

Figure 1: Vision Transformer structure

Beam-search is used in Gated recursive convolutional neural network (grConv)(Cho et al. 2014a), which is an im-

provement to the greedy strategy. At each time step, instead of only retaining the output with the highest current score, it retains the output with the highest score of the first beam numbers.

The task of this article is LaTeX formula recognition, which is equivalent to translating images containing formulas into corresponding LaTeX symbols. According to previous methods, our model consists of a ViT encoder with a ResNet backbone and a Transformer decoder to solve the LaTeX formula recognition problem.

Proposed Solution

Research Design

Compared with the traditional formula recognition, we put forward several problems: (1) First, in the inference stage, the local environment, including the resolution of the screen, the size of the formula display and other factors, will lead to a decrease in recognition performance. (2) The Latex formula we need to identify may be deformed due to printing, may be inverted due to the dark mode of the browser, and there may be a watermark in the recognition area. (3) The search algorithm can continue to be optimized.

Methodology

For the above problems, we propose corresponding solutions: First, in order to reduce the difference between the training domain and the actual inference image domain (screen resolution, formula display size), we train a preprocessing network resizer. Enter formula images with different size resolutions, and output a size that needs to be resized.

In view of the problems of Latex formula deformation, inverted color and watermark that need to be identified, we use image enhancement to expand the data set to make the trained model more suitable for various occasions. The image enhancement method chosen this time is to use Randaugment, which adds random distortion, perspective transformation, texture/watermark, inverted color, etc. to the image.

We adopted beam search in the search phase, and Beam Search improved the greedy search, expanding the search space and making it easier to get the global optimal solution. Beam Search contains a parameter, beam size k, which means that the k sequences with the highest scores are retained at each moment, and then continue to be generated with those k sequences at the next moment. At each time step, only the 1 output with the highest current score is retained, but num-beams.

Solution

Our work is based on the ViT algorithm, and the formula is identified by image caption and transformed into LaTeX code, so that the formula can be quickly transferred to LaTeX when we use latex for paper writing. This task can be regarded as an extension of the number recognition task, which can provide convenience for our scientific research writing and is of great practical significance. However, due to the limitations of data sets and practical requirements, our task is only limited to identify formulas and generate LaTeX code.

Unlike RNN, each step in Transformer is computed independently of each other. While this feature improves parallelism in Transformer, it also makes it difficult to directly use the override mechanism from previous work in the Transformer decoder. To solve the above problems, we propose a novel Attention Refinement Module (ARM), which can refine the attention weight according to the past alignment information without affecting the parallelism.

CNN encoder. In the encoder part, ResNet V2 is used as the encoder. For ResNet V2, the identity mapping branch does not have a ReLU activation function, which can be unblocked during forward propagation and backward propagation, and truly realizes identity mapping. The "final presentation result" of ResNet V2 is also relatively simple, that is, the convolution, BN and ReLU in the original residual block are reversed in order, and the larger part of the article is experimentally proving the effect of various structures, so it is not too much introduction, just know to ensure that the identity mapping branch is as "clean" as possible, it is easier to optimize. The overall structure of ResNet V2 can refer to the Figure2.

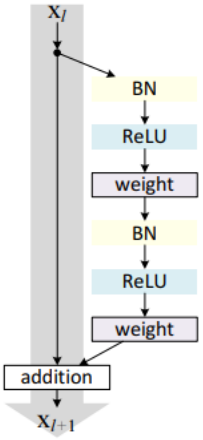

Figure 2: ResNet V2 structure

Position Encoding. Unlike RNN decoders, additional location information is necessary since Transformer decoders do not have spatial location relationships between tokens. The design is consistent with BTTR(Zhao et al. 2021), using both image position coding and character position coding. For character position encodings, 1D position encodings introduced in Transformer(Vaswani et al. 2017) are used. Given the encoding dimension d, position p, and feature dimension index i, the character position encoding vector $\mathbf{p}^{\mathbf{w}}_{p,d} \in \mathbb{R}^d$ can be expressed as

$$\mathbf{p}^{\mathbf{W}}_{p,d}[2i] = \sin\left(p/10000^{2i/d}\right) \quad (1)$$

$$\mathbf{p}^{\mathbf{W}}_{p,d}[2i+1] = \cos\left(p/10000^{2i/d}\right) \quad (2)$$

Attention Refinement Module. If the coverage attention mechanism of RNN-type is directly adopted in Transformer. Then a coverage matrix $F \in \mathbb{R}^{T*L*h*d_{attn}}$ with o(TLhd) space complexity will result, which is of unacceptable size. The bottleneck of the problem is that the covering matrix needs to be added to other eigenvectors and then multiplied by the vector $v_a \in \mathbb{R}^{d_{attn}}$. If we can first multiply the coverage matrix with v_a, together with the result of LuongAttention(Luong, Pham, and Manning 2015), the space complexity will be greatly reduced to o(TLhd). So the attention mechanism is modified to

$$\begin{aligned} \mathbf{e}'_t &= \tanh\left(\mathbf{H}_t\mathbf{W}_h + \mathbf{X}_f\mathbf{W}_x\right)\mathbf{v}_a + \mathbf{F}_t\mathbf{v}_a \\ &= \underbrace{\tanh\left(\mathbf{H}_t\mathbf{W}_h + \mathbf{X}_f\mathbf{W}_x\right)\mathbf{v}_a}_{\text{attention}} + \underbrace{\mathbf{r}_t}_{\text{refinement}} \end{aligned} \tag{3}$$

The similarity vector e'_t can be divided into the attention term and the refinement term $r_t \in \mathbb{R}^L$. Note that the refining term can be generated directly from the cumulative c_t vector via the cover function, thus avoiding the intermediate term with dimension d_{attn}. The above formula is named the attention refining framework.

Experiment

Datasets and Implementation

Datasets and Metrics. We evaluate the proposed method on the public im2latex-100k and CROHME.Here we give a brief introduction to their composition.

The purpose of a dataset is to provide data and its labels.Then the data in the im2latex100k is the picture of the mathematical formula. The label corresponding to the picture is a mathematical formula, and each mathematical formula is represented by a string, each line is a mathematical formula. There is a one-to-one correspondence between these mathematical formulas and pictures.The specific corresponding method is to rely on the formula id number and the image name to correspond.

The CROHME is the most widely used public data set in handwritten mathematical formula recognition, generated from an online handwritten mathematical formula recognition competition (CROHME). The training set in the CROHME data set consists of the following three parts: CROHME 2014 (986), CROHME 2016(1147), CROHME 2019(1199). The number of mathematical formulas is in parentheses. The number of symbol classes identified in the CROHME dataset is 111, including "eos" and "sos".

For evaluation, we adopt the BLEU and Edit Distance(ED) as the metrics, consistent with existing methods in OCR. In machine translation tasks in natural language processing, BLEU is very common, and it is an indicator for evaluating the difference between the sentence (candidate) generated by the model and the actual sentence (reference).Its value ranges from 0.0 to 1.0. If the two sentences perfectly match (perfect match), then BLEU is 1.0. Conversely, if the two sentences do not match perfectly (perfect mismatch), then BLEU is 0.0.Edit Distance (ED) is to quantify the degree of difference between two strings S1 and S2. The calculation method is to see how many times of processing is required to change S1 into S2. If the distance between them is greater, it means that they are more different.

Implementation Details. Our method is implemented with the PyTorch framework. The input images are resized to 128x128. For the data augmentation, we adopt some common strategies, including the random horizontal flip and random erasing.Following existing method, we use the Vision Transformer as the feature extractor with 158480 Latex formulas and 9822 pictures of handwritten formulas on the datasets. We use the Adam optimizer to train the model for 200 epochs with weight decay set to 0.0005, and the learning rate is initialized to 0.00035 with decay by 0.1 at epoch 70 and 110.

Evaluate our model

In experiments, we compare our proposed method with the existing OCR methods.In Table 1, we show the evaluation results on the im2latex100k and CROHME datasets. We compare the BLUE and accuracy(ACC) metrics of these methods.

We must admit that our modified model is still not as effective as the best model in the current OCR field, but our efforts are not in vain. According to the comparison experiments of the last five groups in Table 1, we can see that our self-designed resizer preprocessing network and beam search search strategy are helpful to improve the performance of the model. At the same time, based on objective facts, it is not difficult to find that the effect of using data enhancement will be worse. Finally, we concluded that the modified model achieves the best effect when using beam search and resizer without data argumentation, which is close to SOTA.

	resizer	BLEU	ACC
Ori Model on Ori Dataset*	no	0.803	0.537
Ori Model on aug Dataset	no	0.135	0.068
New Model on Ori Dataset	yes	0.788	0.483
New Model on ori Dataset + beam search	yes	0.790	0.514
New Model on aug Dataset	yes	0.727	0.388
New Model on aug Dataset + beam search	yes	0.747	0.408

Table 1: Comparison with existing method methods on the datasets.

Ablation Study

Through ablation experiments, we verify the influence of each module on the experimental results

Discussion

In our tried designs, we mainly achieved performance improvement by adding beam-search and resize networks in the inference stage,and image augment in the training stage.However,even though we spend much longer time in inference stage and more resource consumption in training stage,we can see that the improvement is very limited

	w/t resize		w/ resizer	
scale	BLEU	ACC	BLEU	ACC
1.5X	0.643	0.296	0.789	0.515
1.25X	0.761	0.493	0.785	0.510
1X	0.786	0.519	0.786	0.518
0.75X	0.711	0.394	0.755	0.463
0.5X	0.285	0.101	0.304	0.101

Table 2: Comparing the effect of using the resizer network at different image sizes

	BLEU	ACC
VIT	0.719	0.412
ResnetV2+VIT	0.803	0.537
ResnetV2+VIT+Beam Search	0.810	0.543

Table 3: Comparing the effect of models under different backbone feature extraction networks

and takes longer. Compared with this little improvement,the cost seems unacceptable,especially the increased inference time,which will greatly influence users' experience.So,in the future, we look forward to making other attempts,such as modifying the backbone of the network,to achieve more improvements.

In terms of application, we look forward to applying the model to a wider range of fields, and the datasets currently used are only LaTeX formulas and handwritten formulas, which cannot be recognized in other languages.With more datasets including different kinds of formulas,including Chinese formulas and so on,our proposed may have a wider usage in transferring different formulas into LaTeX code and obtain a better performance.

	w/t data augment	w/ data augment
time consuming	1h30min	2h15min

Table 4: Time consuming with or without data augment during training

Conclusion

Most of the mathematical formulas on current websites and articles are in Latex format, the purpose of our project is to create a learning based system that takes an image of a math formula and returns corresponding LaTeX code, which will make it easier for people to write papers. This model is mainly based on the ViT (Vision Transformer) algorithm, and the model consist of a ViT encoder with a ResNet backbone and a Transformer decoder. Based on the original model,we tried several tricks on it in order to get better performance,and the experiments shows those tricks make sense.Though the final model performs well but it still has great space to improve,just as we discussed in the above section.We hope our work will help more scientific workers in writing papers and get convenience especially in editing formulas.

	FPS
ResnetV2+VIT	10.0
ResnetV2+VIT+Beam Search	6.67

Table 5: Inference speed with or without Beam Search

References

Cho, K.; Van Merriënboer, B.; Bahdanau, D.; and Bengio, Y. 2014a. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014b. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Deng, Y.; Kanervisto, A.; Ling, J.; and Rush, A. M. 2017. Image-to-markup generation with coarse-to-fine attention. In *International Conference on Machine Learning*, 980–989. PMLR.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016a. Deep Residual Learning for Image Recognition. *IEEE*.

He, P.; Huang, W.; Qiao, Y.; Loy, C. C.; and Tang, X. 2016b. Reading scene text in deep convolutional sequences. In *Thirtieth AAAI conference on artificial intelligence*.

Hu, Y.; Zheng, Y.; Liu, H.; Jiang, D.; Liu, Y.; and Ren, B. 2020. Accurate structured-text spotting for arithmetical exercise correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 686–693.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. 2012. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in neural information processing systems*, 25(2).

Lee, C.-Y.; and Osindero, S. 2016. Recursive recurrent nets with attention modeling for ocr in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2231–2239.

Li, B.; Yuan, Y.; Liang, D.; Liu, X.; Ji, Z.; Bai, J.; Liu, W.; and Bai, X. 2022. When Counting Meets HMER: Counting-Aware Network for Handwritten Mathematical Expression Recognition. In *European Conference on Computer Vision*, 197–214. Springer.

Li, H.; Wang, P.; and Shen, C. 2017. Towards end-to-end text spotting with convolutional recurrent neural networks. In *Proceedings of the IEEE international conference on computer vision*, 5238–5246.

Liao, M.; Shi, B.; Bai, X.; Wang, X.; and Liu, W. 2017. Textboxes: A fast text detector with a single deep neural network. In *Thirty-first AAAI conference on artificial intelligence*.

Luong, M.-T.; Pham, H.; and Manning, C. D. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Shi, B.; Bai, X.; and Yao, C. 2016. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11): 2298–2304.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, 2048–2057. PMLR.

Zeiler, M. D.; and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*, 818–833. Springer.

Zhao, W.; and Gao, L. 2022. CoMER: Modeling Coverage for Transformer-Based Handwritten Mathematical Expression Recognition. In *European Conference on Computer Vision*, 392–408. Springer.

Zhao, W.; Gao, L.; Yan, Z.; Peng, S.; Du, L.; and Zhang, Z. 2021. Handwritten mathematical expression recognition with bidirectionally trained transformer. In *International Conference on Document Analysis and Recognition*, 570–584. Springer.