Global Context-Based Network with Transformer for Image2latex

Nuo Pang, Chun Yang* , Xiaobin Zhu, Jixuan Li and Xu-Cheng Yin

Introduction

Image2latex usually means converts mathematical formulas in images into latex markup. It is a very challenging job due to the complex two-dimensional structure, variant scales of input, and very long representation sequence. Many researchers use encoder-decoder based model to solve this task and achieved good results. However, these methods don't make full use of the structure and position information of the formula. To solve this problem, we propose a global context-based network with transformer.

Our Method

Our method is based on encoder decoder architecture as shown in Figure, it consists of three key modules: a global context-based feature extractor, a transformer-based encoder and a mask attention-based decoder.

The feature extractor is responsible for extracting the features of the input image and outputting a feature map. One of our contribution in this work is that we introduce the global context block in the feature extractor to enhance the ability to extract contextual features, which is beneficial for the model to understand the complex spatial relationship of mathematical formula images.

The transformer-based encoder is inserted between the feature extractor and decoder, which can make better use of the relative position relationship between formula symbols and encode different types of feature dependencies.

The mask attention-based decoder can mask the past alignment information to solve over-parsing problem, which denotes that some parts in feature map are repeatedly parsed multiple times when decoding.

Visual

$$_p = E_{p\sigma}(x)\omega_\sigma = N(n)e^{i(p_0x^0+p_2x^2+p_3x^3)}D_n(\rho)\omega_\sigma$$

baseline: p=E_{p \sigma}(x) \omega_{\sigma}=N(n) e^{i}\left(p_{0} x^{0}+p_{2} x^{2}+p_{3} x^{3}\right) D_{n}(\rho) \omega_{\sigma}

ours: p=E_{p \sigma}(x) \omega_{\sigma}=N(n) e^{i\left(p_{0} x^{0}+p_{2} x^{2}+p_{3} x^{3}\right)} D_{n}(\rho) \omega_{\sigma}

(a)

$$U, U_\Omega \sim O(\hbar^{1/2}), \qquad V_\infty, \overline{V}_\infty \sim O(\hbar^1),$$

baseline: U, U_{\Omega} \sim O\left(\hbar^{1 / 2}\right), \quad V_{\infty}, V V_{\infty} \sim O\left(\hbar^{1}\right)

ours: U, U_{\Omega} \sim O\left(\hbar^{1 / 2}\right), \quad V_{\infty}, V_{\infty} \sim O\left(\hbar^{1}\right)

(b)

Figure 3. Visualization of two cases.

Visualization of two cases is shown in Figure 3. The red tokens denote wrong prediction, and the green denote true. (a) denotes the recognition error of spatial structure relationship that the right bracket close too early; (b) denotes the over-parsing error that V is parsed twice in succession.

Model architecture

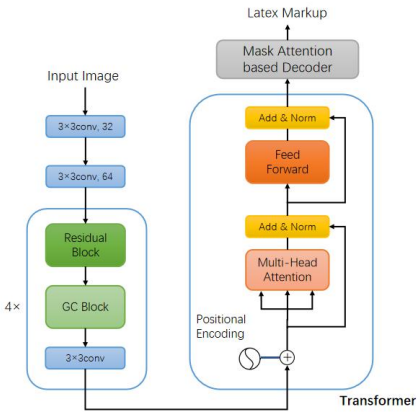

Figure 1. The whole architecture of our model.

Results

Table 1 COMPARISON WITH DIFFERENT MODELS ON THE IM2LATEX-100K

MODEL	BLEU	EDA	EMA
INFTY [1]	66.65	53.82	26.66
WYGIWYS [22]	87.73	87.60	79.88
Coarse-to-Fine Attention [23]	87.07	-	78.10
Double Attention [29]	88.42	88.57	-
our model	**89.72**	**90.07**	**82.13**

Table 2 ABLATION EXPERIMENTS FOR OUR THREE KEY MODULES

MODEL	BLEU	EDA	EMA
Baseline	86.93	86.48	78.37
Baseline + GC	87.71	87.43	79.24
Baseline + Transformer	88.37	88.24	79.91
Baseline + Mask Attention	87.85	87.61	79.46

The results of ablation experimentsablation experiments are shown in Table 2. We use ResNet without any additional modules and an attention-based decoder as our baseline model.

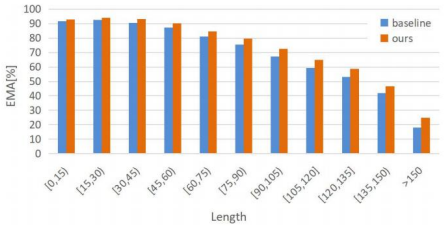

Figure 2. Exact Match Accuracy(EMA) for different formula length. While the recognition performance of long formulas has been greatly improved.

In Figure 2, we show the exact match accuracy of different formula lengths. Our model has a significant improvement (more than 4%) when the formula length is greater than 75.