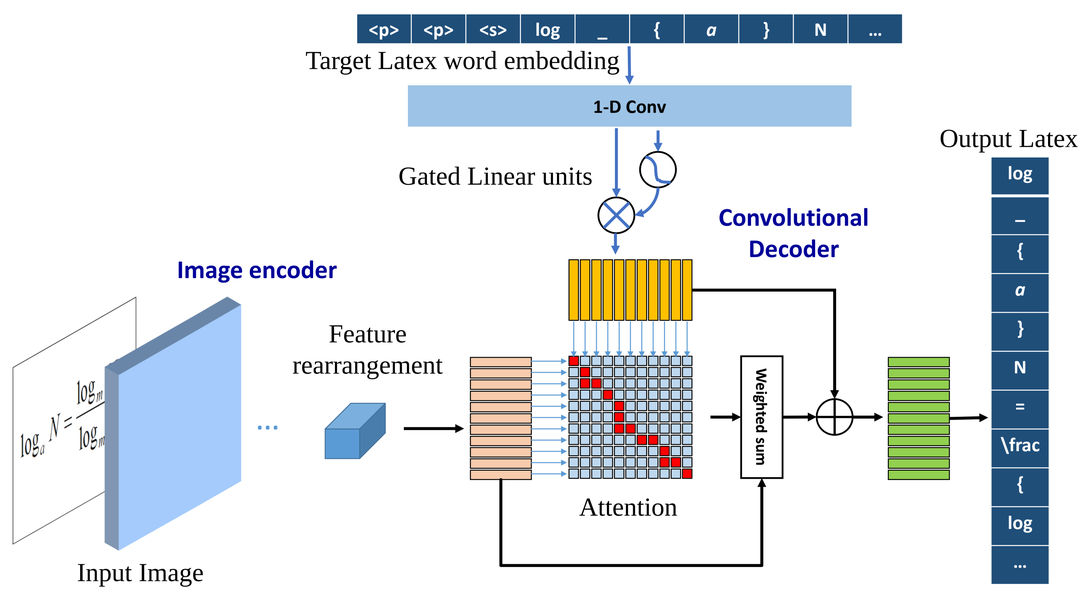
ConvMath : A Convolutional Sequence Network for Mathematical Expression Recognition

Zuoyu Yan, Xiaode Zhang, Liangcai Gao, Ke Yuan and Zhi Tang
Wangxuan Institute of Computer Technology, Peking University
{yanzuoyu3,zhangxiaode,gaoliangcai,yuanke,tangzhi}@pku.edu.cn

Introduction

- Main Task: converts the mathematical expression description in an image into a LaTeX sequence
- Existing problems: math expression exhibits complicated 2-D layout, variant scales and different symbols can be similar
- We propose an entirely convolutional encoder-decoder model with a residual encoder and a convolutional decoder with multi-layer attention
- Combine attention mechanism with the convolutional decoder to alleviate the problem of lacking coverage
- The model achieves state-of-the-art result and much faster efficiency on an open dataset named IM2LATEX-100K

Model



Residual encoder:

- Input image X, output feature vector $V = \{v_i\}$, where $v_i \in R^D$
- combine high-level and low-level features and easy to optimize

Convolutional decoder:

- input: feature vector $V = \{v_i\}$, output latex sequence $Y = \{y_i\}$
- embedding: latex embedding: $W = \{w_i\}$,(same as [1]) and position embedding $P = \{p_i\}$ (same as [2]) where $w_i, p_i \in R^D$ final embedding $G = \{g_i\} = \{w_i + p_i\}$
- stack multiple basic blocks using residual connection, each block with a output of $H = \{h_i\}$
- basic block: a 1-dim convolution and a gated linear units(GLU)
- 1-dim convolution: input: k continuous elements $\in R^{kD}$ output: $M = [A; B] \in R^{2D}$ where $A, B \in R^D$
- GLU: $GLU(M) = A \otimes \sigma(B)$ where $\otimes$ is the point-wise multiplication and σ is sigmoid function : used to select important parts
- Loss function: $L_c = -\frac{1}{|D|}\sum_D \sum_{i=1}^{N} logp(y_i|y_{<i}, X)$

Attention Mechanism:

- content vector: $c_i^l = \sum_{j=1}^{W'*H'} a_{ij}^l v_j$, here c_i^l is the content vector of the l-th decoder layer corresponding to the i-th state
- attention score: $a_{ij}^l = \frac{exp(d_i^l, v_j)}{\sum_{t=1}^{W'*H'} exp(d_i^l, v_t)}$
- decoder state summary: $d_i^l = W_d^l h_i^l + b_d^l + g_i$
- $c_i^l + h_i^l$ as the input of the next layer
- apply to each decoder layer and alleviate the problem of lacking coverage

Conclusion

- Propose a convolution based model which achieves SOTA results and much higher speed
- combine multi-layer attention mechanism with the decoder, which solves the problem of lacking coverage

Experiments

Method	BLEU	time($s/batch$)	Edit Distance	Exact Match
WYGIWYS[3]	87.73	0.129	87.60	79.88
WAP[4]	88.21	0.135	89.58	82.08
ConvMath	88.33	0.083	90.80	83.41

- Dataset: IM2LATEX-100K, contains Latex expressions from over 60000 papers from arxiv
- training/validation/test set: 65995/8181/8301 expressions
- symbol dictionary: 583, embedding size: 512
- evaluation: BLEU score, column-wise edit distance, exact match accuracy, the elapsed time to finish a forward inference for a batch(batch size 10)

Image	$M_g = M_{c_1} M_{c_2} M_{c_3} M_{c_4} M_{c_5} M_{r=\infty} = 1$
Ground truth	M _ { g } = M _ { c _ { 1 } } M _ { c _ { 2 } } M _ { c _ { 3 } } M _ { c _ { 4 } } M _ { c _ { 5 } } M _ { r = \infty } = 1
WYGIWYS	M _ { g } = M _ { c _ { 1 } } M _ { c _ { 2 } } M _ { c _ { 3 } } M _ { c _ { 2 } } M _ { c _ { 5 } } M _ { r = \infty } = 1 \qquad \qquad
ConvMath	M _ { g } = M _ { c _ { 1 } } M _ { c _ { 2 } } M _ { c _ { 3 } } M _ { c _ { 4 } } M _ { c _ { 5 } } M _ { r = \infty } = 1
Image	$V(z, \bar{z}) = e^{-q\Phi(z)} e^{i\alpha \cdot H} e^{i(P_R \cdot X_R - P_L \cdot X_L)}$,
Ground truth	V (z , \bar { z }) = e ^ { - q \Phi (z) } e ^ { i \alpha \cdot H } e ^ { i (P _ { R } \cdot X _ { R } - P _ { L } \cdot X _ { L }) } \; ,
WYGIWYS	V (z , \bar { z }) = e ^ { - q \Phi (z) } e ^ { i \alpha \cdot H } e ^ { i (P _ { R } \rightarrow X _ { R } - P _ { L } X _ { L }) } \; , \hspace { 1 c m }
ConvMath	V (z , \bar { z }) = e ^ { - q \Phi (z) } e ^ { i \alpha \cdot H } e ^ { i (P _ { R } \cdot X _ { R } - P _ { L } \cdot X _ { L }) } \;
Image	$R(e_1) = \epsilon^{-J_{67}+J_{89}}, \quad R(e_2) = \epsilon^{J_{45}-J_{89}}.$
Ground truth	R (e _ { 1 }) = \epsilon ^ { - J _ { 6 7 } + J _ { 8 9 } } , \quad R (e _ { 2 }) = \epsilon ^ { J _ { 4 5 } - J _ { 8 9 } } .
WYGIWYS	R (e _ { 1 }) = \epsilon ^ { - J _ { 0 } + J _ { 8 } } , \quad R (e _ { 2 }) = \epsilon ^ { J _ { 3 } - J _ { 8 } } .
ConvMath	R (e _ { 1 }) = \epsilon ^ { - J _ { 0 } + J _ { 8 9 } } , \quad R (e _ { 2 }) = \epsilon ^ { I _ { 4 } } - J _ { 8 9 } } .

- over parsing(feature vectors generate multiple times) rarely happens
- under parsing(feature vectors not parsed) is common(refer to the third example), can be a future direction

Reference

[1] Sennrich R, Haddow B. Linguistic input features improve neural machine translation[J]. arXiv preprint arXiv:1606.02892, 2016.
[2] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.
[3] Deng Y, Kanervisto A, Rush A M. What you get is what you see: A visual markup decompiler[J]. arXiv preprint arXiv:1609.04938, 2016, 10: 32-37.
[4] Zhang J, Du J, Zhang S, et al. Watch, attend and parse: An end-to-end neural network based approach to handwritten mathematical expression recognition[J]. Pattern Recognition, 2017, 71: 196-206.