Recognizing handwritten mathematical expressions via paired dual loss attention network and printed mathematical expressions

Anh Duc Le
The center for Open Data in Humanities
Tokyo, Japan
anh@ism.ac.jp

Abstract

Recognition of Handwritten Mathematical Expressions (HMEs) is a challenging problem because of the complicated structure and uncommon math symbols contained in HMEs. Moreover, the lack of training data is a serious issue, especially for deep learning-based systems. In this paper, we proposed a dual loss attention model that utilizes the existing latex corpus to improve accuracy. The proposed dual loss attention has two losses, including decoder loss and context matching loss to learn semantic invariant features for the encoder and latex grammar for the decoder from handwritten and printed MEs. The results of experiments on the CROHME 2014 and 2016 databases demonstrate the superiority and effectiveness of our proposed model. These results are competitive compared to others reported in recent literature.

1. Introduction

Mathematical expressions (MEs) are commonly used in scientific documents, books, and examinations. However, it is uneasy to input MEs into computers, since the complicated structure and uncommon math symbols contained in MEs. Handwritten is the natural way to input MEs for humans, but it is hard for computers to recognize them. Complicated two-dimensional structures, ambiguous spatial relations, and unstable writing style are the main challenges on recognition of HMEs. Recognition of HMEs has recently focused since it has many potential applications in computer-assisted scoring systems, self-study math applications, and handwriting input systems for scientific writing.

Traditional recognition methods can be divided into three main processes: symbol segmentation, symbol recognition, and structural analysis. Most of them rely on a human-defined grammar such as Context Grammar. Inspired by recent successes of the attention-based encoder-decoder model in neural machine translation and image caption generation, researchers focus on solving HMEs recognition as an end-to-end trainable system. The end-to-end trainable system requires only input data and their corresponding Latex for training. All the processes of symbol segmentation/recognition and structural analysis are incorporated in the attention-based encoder-decoder system. This method outperformed the traditional method.

Many approaches have been proposed for recognizing HMEs, especially during the last two decades. They are summarized in the survey papers [2, 9] and the recent competition papers [7, 6, 5]. Most of them employed the Context Free Grammar (CFG) [4, 1] and attention-based encoder-decoder [3, 10, 8]. In the following, we will review a few recent approaches evaluated on the recent Competition on Recognition of Online Handwritten Mathematical Expressions (CROHME).

For CFG based method, A 2D Stochastic CFG based method was proposed by Alvaro et al.[1]. The Cocke-Younger-Kasami parsing algorithm is modified to parse an input online HME in two dimensions. This system was the best system at CROHME 2011 and the best system with using only CROHME data at CROHME 2013 and 2014. Le et al. proposed a recognition method for HMEs based on SCFG [4]. Stroke order is employed to reduce the search space and the CYK algorithm is used to parse a sequence of input strokes. They extended the grammar rules to cope with multiple symbol order variations. The system participated in CROHME 2013, 2014, and 2016.

For attention-based encoder-decoder method, Zhang et al. proposed an end-to-end approach based on neural network to recognize HMEs, which is called WAP [10]. They employed a DenseNet encoder to extract features from an input HME and a GRU decoder with an attention-based parser to generate LaTeX sequences.

Le et al. proposed data generation strategies that incorporate with attention-based encoder-decoder system [3]. The experiment shows superior by the additional generated data.

Recently, Wu et al. proposed a Paired Adversarial Learning to learn semantic invariant features between handwritten and printed MEs in the feature space. In this paper, we employ similar paired images to train our proposed system.

Since the training dataset for HMEs are small and the collection and annotation processes are labor-intensive and time-consuming. So that improving the recognition by using other existing resources such as latex corpus is a potential research direction. For recognition of HMEs, we have a large corpus of latex which is easy to collect from the internet. If we could take advantage of printed MEs rendered from the latex corpus, it will be helpful. The core idea of the paper is to learn from the printed ME domain to improve the handwritten ME domain. The system learns not only semantic invariant features between handwritten and printed MEs, but also latex grammar from printed MEs.

In this paper, we propose a new method for recognizing HMEs, named dual loss attention. As shown in Figure 1, the proposed dual loss attention has two losses including decoder loss (the loss to map context vectors to target symbols) and context matching loss (the loss to map context vectors between handwritten and printed MEs to extract semantic invariant features). To learn semantic invariant features, we train the network with paired handwritten and printed MEs. To learn grammar-based features for the decoder, we train the network with printed MEs from the latex corpus.

We summarize our contributions as follows: We propose a dual loss attention for robust HMEs recognition, which could learn semantic invariant features and grammar-based features between handwritten and printed MEs. We propose a context matching module to map context vectors cross domains. We proposed two scenarios: paired MEs training scenario to learn semantic invariant features and printed MEs training scenario to learn grammar-based features for mathematics. We show the effectiveness of our proposed model and training scenarios though experiments. The results are competitive compared to others reported in recent literature.

2. Dual loss attention network

In this paper, we develop dual loss attention for robust recognition of HMEs. We have paired images of handwritten and printed MEs from the CROHME training set and printed MEs rendered from the Latex corpus. Formally, we assume that there are N paired annotated samples $X_{pair} = (x_i^h, x_i^p)$ with the corresponding labels $Y_{pair} = (y_i)$ and M printed MEs rendered from latex corpus $\bar{X} = (\bar{x}_i)$ with the corresponding labels $\bar{Y} = (\bar{y})$. Where x_i^h is HMEs images, x_i^p and $\bar{x}_i$ are printed MEs images. Given a training set $D = (X_{pair}, Y_{pair}), (\bar{X}, \bar{Y})$, we try to maximize the prediction probability w.r.t. parameters θ of the recognition system. As shown in Figure 1, the proposed dual loss attention encodes a pair of handwritten and printed MEs into two sequences of context vectors (attended math symbol-level features). The sequence of context vector is used for predicting the target latex sequence by a GRU decoder. The context vectors are the representation of handwritten and printed math symbols. We employ the Mean Squared Error (MSE) as Context Matching to map the corresponding representation of handwritten and printed math symbols. When the representation of handwritten and printed math symbols are mapped, we can train the network for recognizing HMEs by printed MEs. Figure 2 shows the illustration of the representation of handwritten and printed domains before and after using Context Matching. The role of Context Matching is to make the representation of handwritten and printed symbols closer. As a result, the decoder can predict target symbols from whatever handwritten or printed MEs.

2.1. Attention based ME Recognition

Our recognition system is based on the attention-based encoder-decoder. It contains three modules: a convolution neural network for feature extraction, attention module for generating context vectors, and a GRU decoder for generating the target latex symbols. In this research, we employ the open source of the related work [10] as the baseline system.

DenseNet based Feature Extraction: DenseNet outperforms the VGG and ResNet by proposing direct connections from any preceding layers to succeeding layers. The i^{th} layer receives the feature maps of all preceding layers, $x_0, \ldots, x_{i-1}$, as input:

$$x_i = H_i([x_0, x_1, ..., x_{i-1}]) \tag{1}$$

where H_i refers to the convolutional function of the i^{th} layer. The detail setting of the DenseNet is described in []. The extracted features from an input x is $F(x)$. The size of the extracted features is $H \times W \times C$, where C denotes the number of channels, H and W are the height and width of extracted features, respectively.

Attention Module: The role of the attention model is to learn the corresponding between a location in an input image and a decoding symbol. At time-step t, the context vector c_t is a result of a weighted sum of the extracted features F and the attention probability α.

$$c_t = \sum_{u,v} \alpha_{t(u,v)} * F_{u,v} \tag{2}$$

where, the attention probability $\alpha_{u,v}$ is calculated by the following softmax function:

$$\alpha_{t(u,v)} = \frac{exp(e_{t(u,v)})}{\sum_{i,j} exp(e_{t(i,j)})} \tag{3}$$

where $e_{t(u,v)}$ denotes the energy of $F_{u,v}$ at time step t. The energy is calculated from the input feature $F_{u,v}$, the previous hidden state of the decoder h_{t-1}, and coverage vector

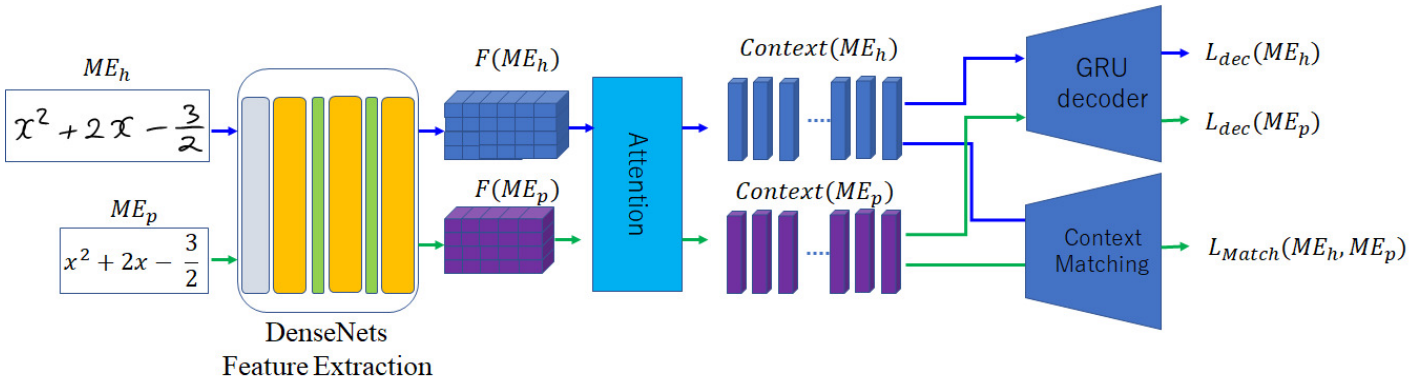

Figure 1. The structure of dual loss attention.

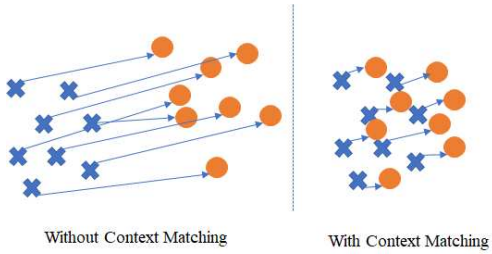

Figure 2. Illustration of the representation of handwritten and printed domains before and after using Context Matching. Circle points and X point represent the context vectors of handwritten and printed symbols, respectively.

$cov_{t(u,v)}$. The coverage vector is initialized as a zero vector and we compute it based on the summation of all past attention probabilities α. v_{att}^T, W_h, W_F, and W_{cov} are learnable parameters.

$$e_{t(u,v)} = v_{att}^T \tanh(W_h * h_{t-1} + W_F * F_{u,v} + W_{cov} * cov_{t(u,v)}) \quad (4)$$

$$cov_{t(u,v)} = \sum_{l=1}^{t-1} \alpha_{l(u,v)} \quad (5)$$

GRU Decoder: The GRU decoder predicts one symbol at a time. At each time step t, the decoder predicts symbol y_t based on the embedding vector of the previous decoded symbol $E_{y_{t-1}}$, the current hidden state of the decoder h_t, and the current context vector c_t as the following equation:

$$p(y_t) = softmax(W_E * E_{y_{t-1}} + W_h * h_t + W_c * c_t) \quad (6)$$

where W_E, W_h, W_c are learnable parameters. The hidden state is calculated by a GRU. It is based on the previous hidden state of the LSTM, the context vector, and the embedding vector of the previous decoded symbol Eyt-1 as the following equation:

$$h_t = GRU(h_{t-1}, c_t, E_{y_{t-1}}) \quad (7)$$

We employ cross-entropy as the objective function to maximize the probability of predicted symbols for a target latex sequence.

$$L_{dec}(y|x) = -\sum_{i=1}^{|y|} \log p(y_i|F(x)) = -\sum_{i=1}^{|y|} \log p(y_i|y_{i-1}, c_i) \quad (8)$$

2.2. Context Matching

Given a pair of handwritten and printed MEs (ME_h, ME_p), they are represented as two sequences of context vectors $Context(ME_h)$ and $Context(ME_p)$ after feature extraction and attention processes. Note that $Context(ME_h)$ and $Context(ME_p)$ have the same length. The i^{th} elements of the two sequences of context vectors are corresponding to each other and aligned with the target symbols y_i. If we can make the distance between each pair of the context vectors $Context(ME_p)_i$ and $Context(ME_p)_i$ become zero, the decoder can learn to recognize handwritten MEs by printed MEs and vice versa. Therefore, we can improve the handwritten ME recognition by using printed ME rendered from the latex corpus. We employ simple Mean Squared Error for Context Matching module. The context matching loss is calculated by the following equation:

$$L_{Match}(ME_h, ME_p) = \frac{1}{n} \sum_{i=1}^{n} (p_i - q_i)^2 \quad (9)$$

where p and q are $Context(ME_h)$ and $Context(ME_p)$, respectively.

2.3. Training algorithm

Given N paired annotated samples $X_{pair} = (X^h, X^p)$ and M printed MEs rendered from latex corpus $\bar{X}$, we train the dual loss attention model by minimizing the following

loss function:

$$L = L_D(X^h) + L_D(X^p) + L_D(\bar{X}) + \lambda L_{Match}(X^h, X^p) \quad (10)$$

where λ is a hyperparameter that controls the trade-off between decoder loss and context matching loss. We employ the AdaDelta algorithm with gradient clipping to learn the parameters. The batch size is set to 8. The training process is stopped when the expression rate on the validation set does not improve after 15 epochs.

3. Evaluation

3.1. Datasets

We employed the CROHME training dataset and printed MEs rendered from latex corpus for training, the CROHME 2013 test set for validation, and the CROHME 2014 as well as CROHME 2016 test sets for testing. The CROHME training dataset contains 8,835 online HMEs for training, and the CROHME 2013, 2014 and 2016 testing datasets contain 671 online HMEs, 986 online HMEs, and 1127 online HMEs, respectively. We select 40,000 latex equations in the latex corpus. Then, we render them to get printed MEs ($\bar{X}$). From the CROHME training dataset, we render ground truth latex to get printed MEs (X_{pair}). The training datasets are presented in Table 1. The number of symbol classes is 101.

		Num. of MEs
X_{pair}	Handwritten MEs	8,835
	Printed MEs from training set	8,835
$\bar{X}$	Printed MEs from latex corpus	40,000

Table 1. Training datasets.

3.2. Experimental results

In order to measure the performance of our proposed system, we employ Word Error Rate (WER) and Expression Rate (Exp. Rate) metrics which are generally employed for evaluating attention based HME recognition systems. An HME was judged to be recognized correctly in terms of Exp. Rate if all of its symbols, relations, and structures were recognized correctly. WER and Exp. Rate are calculated as the following equations:

$$WER = \frac{N_{sub} + N_{del} + N_{ins}}{N_Y} \quad (11)$$

$$WER = \frac{N_{correct}}{|Y|} \quad (12)$$

where N_{sub}, N_{del}, and N_{ins} are the number of substitutions, deletions, insertions. N_Y is the number of symbol in the target set Y. $N_{correct}$ is the number of equations recognized correctly.

The first experiment is to evaluate the attention-based ME recognition as the baseline system. We employ handwritten MEs from the CROHME training dataset for this experiment. Table 2 shows the results on the CROHME 2014 and 2016 test sets.

Testing set	Metrics	
	WER	Exp. Rate
2014	16.63	46.65
2016	15.82	44.64

Table 2. Performance of the attention based ME recognition as baseline system.

The second experiment is to evaluate the dual loss attention with different values of the hyperparameter λ. We train the recognition system on X_{pair} which contains handwritten and printed MEs. We set $\lambda = 0.0, 0.1, 0.2, 0.3, 0.4$ for the experiment. We achieved the best Exp. rates (49.85 on the CROHME 2014 test set) and (47.34 on the CROHME 2016 test set) which are better than Exp rates of baseline system.

Testing set	Metrics	$Hyperparameter \lambda$			
		0.1	0.2	0.3	0.4
2014	WER	14.52	14.70	15.48	13.9
	Exp. Rate	49.85	46.8	47.01	49.44
2016	WER	15.08	15.58	16.70	15.45
	Exp. Rate	47.34	43.94	42.46	43.59

Table 3. Performance of the dual loss attention on X_{pair} set.

The third experiment is to evaluate the dual loss attention with different values of the hyperparameter λ on X_{pair} and $\bar{X}$. We set $\lambda = 0.0, 0.1, 0.2, 0.3, 0.4$ for the experiment. We achieved the best Exp. rates (51.88 on the CROHME 2014 test set) and (51.53 on the CROHME 2016 test set). In conclusion, the context matching loss leads to a 3% accuracy improvement while the context matching loss and additional printed MEs lead to a 5% - 7% accuracy improvement over the baseline system.

Testing set	Metrics	$Hyperparameter \lambda$			
		0.1	0.2	0.3	0.4
2014	WER	12.48	12.16	12.99	12.74
	EXP Rate	50.66	51.88	50.15	49.44
2016	WER	12.34	12.28	13.62	12.77
	EXP Rate	49.78	51.09	48.3	51.53

Table 4. Performance of the dual loss attention on X_{pair} and $\bar{X}$ sets.

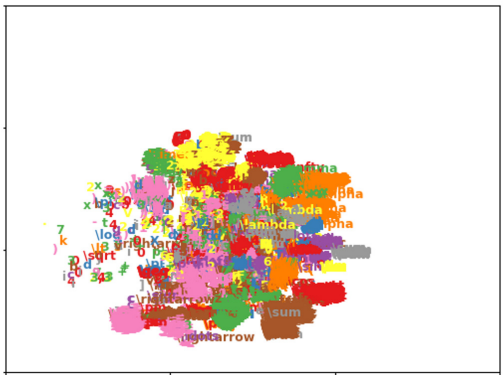

Figure 3. Visualize the context vectors extracted by the baseline system on the CROHME 2016 test set.

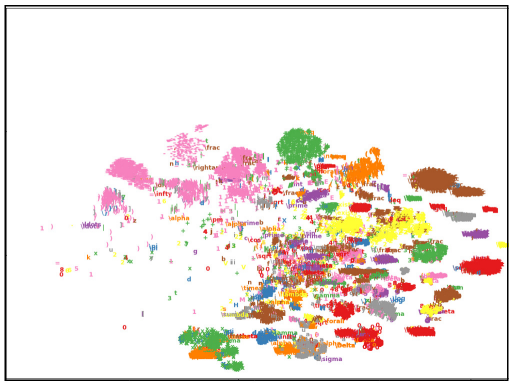

Figure 4. Visualize the context vectors extracted by the proposed system on the CROHME 2016 test set.

In the final experiment, we compare the proposed model with three best participants of CROHME 2014, 2016 and other recently attention-based models. The system with the sign * used extra handwritten MEs for training. For a fair comparison, all the systems do not utilize the ensemble of multiple models. Our proposed dual loss attention outperforms Alvaro et al., IRCCyN (2014), TUAT (2016), End-to-End, WAP, and PAL-V2. Moreover, Myscript, PAL-v2 systems employed a statistical language model as a post processing while our proposed model does not employ any post processing.

System	Exp. rate	
	2014	2016
Myscript*	62.68	67.65
Alvaro et al.	37.22	49.61
IRCCyN (2014), TUAT (2016)	26.06	43.94
End-to-End (Le and Nakagawa 2017)*	48.78	45.60
WAP (Zhang et al. 2017)	48.38	46.82
PAL-v2 (Wu et al. 2020)	48.88	49.61
Dual loss attention (our)	51.88	51.53

Table 5. Comparison of our proposed system and the state-of-the-art recognition systems on CROHME 2014 and 2016.

To visualize the learned features between baseline system and the proposed system (the best system in table 4), we show the context vectors of each symbol class in the CROHME 2016 test set. We employ t-SNE to visualize context vectors from D dimensions to 2 dimensions. We observed that the proposed system provides better representation for context vectors. The distance between classes in figure 4 is larger than that in figure 3. Therefore, the decoder is easier to recognize them.

4. Conclusion

In this paper, we have proposed the dual loss attention to recognize HMEs. The proposed model has two losses, including decoder loss and context matching loss to learn semantic invariant features and grammar-based features from handwritten and printed MEs. The efficiency of the proposed model was demonstrated through experiments. The recognition rate is improved when we employ the dual loss and printed MEs generated from the latex corpus. Our best recognition system is competitive with state-of-the-art recognition systems in recent literature.

References

[1] F. Alvaro, J. Sanchez, and J. Benedi. Recognition of on-line handwritten mathematical expressions using 2d stochastic context-free grammars and hidden markov models. *Pattern Recognition Letters*, pages 58–67, 2014.

[2] K. Chan and D. Yeung. Mathematical expression recognition: A survey. *International Journal of Document Analysis and Recognition*, pages 3–15, 2000.

[3] A. Le, I. Bipin, and M. Nakagawa. Pattern generation strategies for improving recognition of handwritten mathematical expressions. *Pattern Recognition Letters*, pages 255–262, 2019.

[4] A. Le and M. Nakagawa. A system for recognizing online handwritten mathematical expressions by using improved structural analysis. *International Journal of Document Analysis and Recognition*, page 305–319, 2016.

[5] H. Mouchere, C. Viard-Gaudin, R. Zanibbi, and U. Garain. ICFHR 2016 CROHME: Competition on recognition of online handwritten mathematical expressions. *International Conference on Frontiers in Handwriting Recognition*, page 607–612, 2016.

[6] H. Mouchere, C. Viard-Gaudin, R. Zanibbi, and U. Garainn. ICFHR 2014 competition on recognition of on-line handwritten mathematical expressions. *International Conference on Frontiers in Handwriting Recognition*, pages 791–796, 2014.

[7] H. Mouchère, C. Viard-Gaudin, R. Zanibbi, U. Garain, D. H. Kim, and J. H. Kim. ICDAR 2013 CROHME: Third internationalcompetition on recognition of online handwritten mathematical expressions. *International Conference on Document Analysis and Recognition*, 2013.

[8] J. Wu, F. Yin, Y. Zhang, X. Zhang, and C. Liu. Handwritten mathematical expression recognition via paired adversarial learning. *International Journal of Computer Vision*, 2020.

[9] R. Zanibbi and D. Blostein. Recognition and retrieval of mathematical expressions. *International Journal of Document Analysis and Recognition*, pages 331–357, 2012.

[10] J. Zhang, J. Du, and L. Dai. Multi-scale attention with dense encoder for handwritten mathematical expression recognition. *24th International Conference on Pattern Recognition*, page 2245–2250, 2018.