# Multi-modal Attention Network for Handwritten Mathematical Expression Recognition

Jiaming Wang, Jun Du, Jianshu Zhang, Zi-Rui Wang
National Engineering Laboratory for Speech and Language Information Processing
University of Science and Technology of China, Hefei, Anhui, P. R. China
Email: jmwang66@mail.ustc.edu.cn, jundu@ustc.edu.cn, xysszjs@mail.ustc.edu.cn, cs211@mail.ustc.edu.cn

***Abstract*—In this paper, we propose a novel multi-modal attention network (MAN), which is based on encoder-decoder framework, for handwritten mathematical expression recognition (HMER). Here, multi-modal means two specific modalities: online and offline, where online modality employs dynamic trajectories as input and offline modality employs static images as input. More specifically, the proposed method first feeds dynamic trajectories and static images into online and offline channels of the multi-modal encoder respectively. The output of the encoder is then transferred to the multi-modal decoder to generate a LaTeX sequence as the mathematical expression recognition result. To make full use of the complementary information that comes from the two modalities, we propose a re-attention mechanism as an enhanced version of the multi-modal attention mechanism which can further improve the recognition performance. Evaluated on a benchmark published by CROHME competition, the proposed approach achieves an expression recognition accuracy of 54.05% on CROHME 2014 and 50.56% on CROHME 2016 which substantially outperforms the state-of-the-arts using the single online or offline modality.**

***Index Terms*—Multi-modal, Handwritten Mathematical Expression Recognition, Encoder-Decoder, Attention**

## I. Introduction

The recognition of handwritten mathematical expression is an indivisible branch of optical character recognition. Different from the original character recognition, handwritten mathematical expression recognition (HMER) is confronted with more challenges due to the complicated two-dimensional structural analysis [1]–[3].

The recent studies of HMER can be roughly divided into two tasks: online HMER [4], [5] and offline HMER [6]. The difference between these two tasks is that the input of online HMER is dynamic handwriting trajectories while the input of offline HMER is static images. Therefore, the primary issue of these two tasks varies severely. In general, online HMER can achieve better performance than offline HMER as the input of online HMER has rich dynamic (spatial and temporal) information which is extremely helpful for handwriting recognition. Owing to these rich dynamic information, online HMER meets fewer difficulties coming from ambiguous handwriting which are usually hard to deal with in offline HMER. Nevertheless, the lack of global image information in online HMER may lead to incorrect recognition of delayed strokes or inserted strokes [7], [8] which can be naturally solved in offline HMER. Therefore, it is an intuitive way to utilize both dynamic trajectories and static images to build a more powerful recognition system for HMER.

Inspired by recent work in multi-modal researches [9]–[12], we propose a novel multi-modal attention network (MAN) that attempts to utilize both the advantages of dynamic handwriting trajectories and static images for HMER. To our best knowledge, this is the first multi-modal learning study for HMER. The proposed multi-modal attention network is based on the encoder-decoder framework [13]–[15]. The encoder which we call multi-modal encoder, comprises two channels: online channel and offline channel. The online channel employs dynamic handwriting trajectories as input and the offline channel employs static images as input. More specifically, the online channel is convolutional neural networks (CNN) [16] following a stack of bidirectional recurrent neural networks with gated recurrent units (GRU-RNN) while the offline channel is a deeper CNN. The output of the encoder is then transferred to the multi-modal decoder to generate a math symbol sequence including spatial structure in LaTeX format [17] for recognition. The multi-modal attention mechanism equipped in the decoder adopts the output of online and offline channels to compute a multi-modal context vector that only contains the useful parts of input to describe one math symbol at each decoding step. Therefore, the multi-modal context vector can contain both dynamic information from online modality and static global information from offline modality simultaneously. Besides, to make full use of the complementary information that comes from the two modalities, we propose a novel re-attention mechanism as an enhanced version of the multi-modal attention mechanism to help control the flow of one modality information into the other one to further improve the recognition performance. In experiments, we evaluate our method on a benchmark published by the competition for handwritten mathematical expression (CROHME), and we achieve expression recognition accuracy of 54.05% on CROHME 2014 and 50.56% on CROHME 2016 which substantially outperforms the state-of-the-arts using the single online or offline modality. Besides, the proposed re-attention mechanism can be easily adopted to other multi-modal tasks. The contributions of this paper are as follows:

- We propose a novel multi-modal attention network (MAN) which can exploit both advantages of online and offline modalities to improve the performance of handwritten mathematical expression recognition.

2379-2140/19/$31.00 ©2019 IEEE
DOI 10.1109/ICDAR.2019.00191

- We propose to utilize a multi-modal encoder to encode dynamic trajectories and static images simultaneously. Instead of employing a stack of GRU-RNNs as introduced in [7], we employ CNN following a fewer stack of GRU-RNN as the online encoder which helps reduce overfitting and achieve better recognition results.
- We show the advantages of multi-modal attention against single-modal attention through experimental analysis and attention visualization.

## II. The Proposed Approach

In this section, we elaborate the proposed MAN based encoder-decoder framework for HMER. As illustrated in Fig. 1, the raw data is a sequence of points containing xy-coordinates and stroke information. First, data processing should be applied to raw data to get trajectory sequences and greyscale images. Then the online channel and the offline channel of multi-modal encoder, extract high-level features from trajectory sequences and greyscale images respectively. The multi-modal decoder is unidirectional GRU-RNN equipped with a multi-modal attention mechanism. To acquire better alignments between input handwriting traces and output LaTeX symbols, we propose a re-attention mechanism as an enhanced version of the multi-modal attention mechanism which can make online and offline modalities complementary to each other and further improve the recognition performance.

### A. Processing

The raw data of handwriting traces are collected during the writing process, which can be represented as a variable length sequence:

$$\{[x_1, y_1, s_1], [x_2, y_2, s_2], \cdots, [x_N, y_N, s_N]\} \tag{1}$$

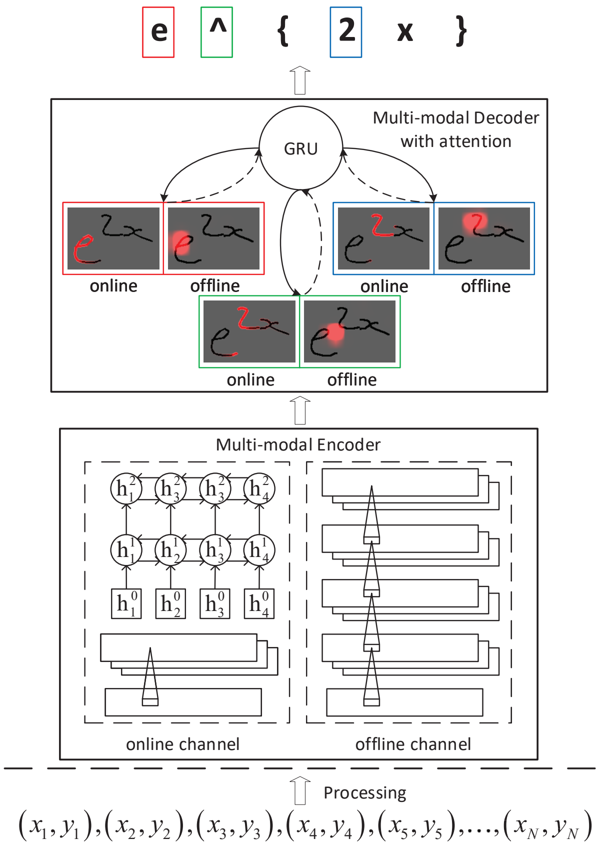
where $x_i$ and $y_i$ are the xy-coordinates of the pen movements and $s_i$ indicates which stroke the $i^{\text{th}}$ point belongs to.

As for the online input, following [18], we remove the redundant points and normalize the xy-coordinates into a standard interval because of the non-uniform sampling and the variable size of handwriting input. Then we obtain an 8-dimension feature vector for each point $i$:

$$\left[x_i, y_i, \Delta x_i, \Delta y_i, \Delta^2 x_i, \Delta^2 y_i, \delta\left(s_i = s_{i+1}\right), \delta\left(s_i \neq s_{i+1}\right)\right] \tag{2}$$

where $\Delta x_i = x_{i+1} - x_i$, $\Delta y_i = y_{i+1} - y_i$, $\Delta^2 x_i = x_{i+2} - x_i$, $\Delta^2 y_i = y_{i+2} - y_i$. $\delta(\cdot) = 1$ means that the condition is true otherwise zero. The two terms, $\delta\left(s_i = s_{i+1}\right)$ and $\delta\left(s_i \neq s_{i+1}\right)$ indicate whether the point is the final point of a stroke, which we usually call pen down, i.e. $[1, 0]$ and pen up, i.e. $[0, 1]$. To simplify the description of the following sections, we will use $\mathbf{X}_{\text{on}} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N)$ to represent the trajectory sequence, which is used as the input for online channel of multi-modal encoder. Note that $\mathbf{x}_i$ here is actually an 8-dimension vector.

As for the offline input, we first calculate the heights of all strokes. Then we count the average height of strokes with the height greater than one tenth of the highest stroke. Furthermore, we normalize xy-coordinates of all points in accordance with the average height. After that, to obtain the static image from strokes, we simply line trajectory points of each stroke to transform traces into the image-like representation. We record this static image as $\mathbf{X}_{\text{off}}$ and treat it as the input for offline channel of the multi-modal encoder.



e
^
{
2
x
}
Multi-modal Decoder with attention
GRU
online
offline
online
offline
online
offline
Multi-modal Encoder
$h_1^2$ $h_2^2$ $h_3^2$ $h_4^2$
$h_1^1$ $h_2^1$ $h_3^1$ $h_4^1$
$h_1^0$ $h_2^0$ $h_3^0$ $h_4^0$
online channel
offline channel
Processing
$(x_1,y_1),(x_2,y_2),(x_3,y_3),(x_4,y_4),(x_5,y_5),\ldots,(x_N,y_N)$

Fig. 1. Architectures of multi-modal attention network (MAN) for handwritten mathematical expression recognition. We show three pairs of images to visualize the decoding procedure. The left image denotes the attention probabilities on online input and the right image denotes the attention probabilities on offline input.

### B. Multi-modal Encoder

Since our model is designed for multi-modal HMER and to deal with both online and offline input, we equip the multi-modal encoder with two channels, namely online channel and offline channel.

Considering online channel, the input is a sequence, $\mathbf{X}_{\text{on}} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N)$. Different from [7], we do not choose to use a stack of RNNs. Instead, we employ CNN following a fewer stack of RNNs, which can acquire better local information and improve the recognition performance. In addition, simple RNN meets the problems of the exploding gradient and the vanishing gradient [19], [20]. Therefore, we actually apply GRU as an improved architecture of simple RNN. The hidden state of GRU [21] can be calculated as:

$$\mathbf{h}_t = \text{GRU}\left(\mathbf{x}_t, \mathbf{h}_{t-1}\right) \tag{3}$$

Furthermore, unidirectional GRU cannot exploit the future context information, so we adopt bidirectional GRU which can utilize both past and future context information. As for the offline channel, the input is the static image. Thus, we choose to use DenseNet [22] which has shown its superiority on image processing.

The output of the online channel is a variable-length sequence, namely $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_{L_{\text{on}}})$ and each element is a $D$-dimension vector. As for the offline channel, the output is a 3D tensor of size $D \times H \times W$. Then we transform the 3D tensor into a variable-length vector sequence of $L_{\text{off}}$ elements, $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_{L_{\text{off}}})$, where $L_{\text{off}} = H \times W$ and each element is a $D$-dimension vector as well.

### C. Multi-modal Decoder

As shown in Fig. 1, the multi-modal decoder generates a LaTeX sequence for recognition:

$$\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_C\}, \mathbf{y}_i \in \mathbb{R}^K \tag{4}$$

where $K$ is the number of total math symbols in the vocabulary and $C$ is the length of LaTeX sequence. Note that the decoder also has two channels to accept the features from online and offline channels of the encoder, which are the online trajectory sequence feature $\mathbf{A}$ and the offline static image feature $\mathbf{B}$.

Since the length of the LaTeX string is not fixed and the output of two encoder channels has variable length, we employ an intermediate fixed-size vector $\mathbf{c}_t$ namely multi-modal context vector [23] generated by a unidirectional GRU with a multi-modal attention mechanism which will be described later. Then another unidirectional GRU is adopted to produce the LaTeX sequence symbol by symbol. The decoder structure can be denoted as:

$$\hat{\mathbf{h}}_t = \text{GRU}_1(\mathbf{y}_{t-1}, \mathbf{h}_{t-1}) \tag{5}$$

$$\mathbf{c}_t = f_{\text{matt}}\left(\hat{\mathbf{h}}_t, \mathbf{A}, \mathbf{B}\right) \tag{6}$$

$$\mathbf{h}_t = \text{GRU}_2\left(\mathbf{c}_t, \hat{\mathbf{h}}_t\right) \tag{7}$$

where $\text{GRU}_1$, $\text{GRU}_2$ indicate two GRU layers, $f_{\text{matt}}$ denotes the multi-modal attention mechanism and $\hat{\mathbf{h}}_t$, $\mathbf{h}_t$ represent the hidden state of the first and the second GRU layers.

To obtain the probability of each predicted symbol, we exploit an additional two-layer perceptron using the previous target symbol $\mathbf{y}_{t-1}$, the hidden state of the second GRU layer, $\mathbf{h}_t$ and the multi-modal context vector $\mathbf{c}_t$ as input:

$$p(\mathbf{y}_t|\mathbf{X}_{\text{on}}, \mathbf{X}_{\text{off}}, \mathbf{y}_{t-1}) = g\left(\mathbf{W}_o \phi\left(\mathbf{E}\mathbf{y}_{t-1} + \mathbf{W}_h\mathbf{h}_t + \mathbf{W}_c\mathbf{c}_t\right)\right) \tag{8}$$

where $g$ represents the softmax activation function and $\phi$ represents the maxout activation function. Then $\mathbf{W}_o \in \mathbb{R}^{K \times \frac{m}{2}}$, $\mathbf{W}_h \in \mathbb{R}^{m \times n}$, $\mathbf{W}_c \in \mathbb{R}^{m \times D}$, $\mathbf{E} \in \mathbb{R}^{m \times K}$.

Attention mechanism [24]–[26] is usually adopted in sequence learning. The difference in our study is that our attention mechanism has more than one modalities. Therefore, we propose a multi-modal attention mechanism which can exploit information from both online and offline modalities. The attention mechanism can be denoted as:

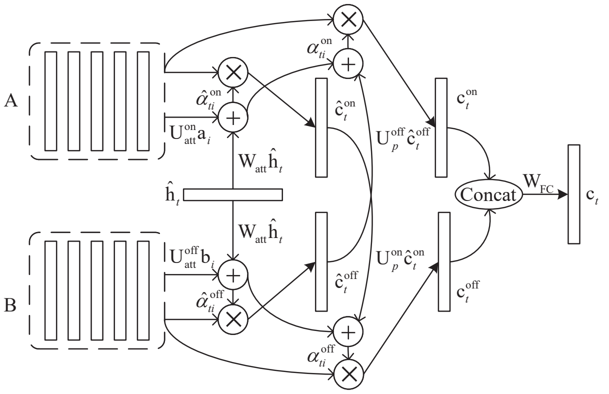

A
B
$\hat{\alpha}_{ti}^{\text{on}}$
$\text{U}_{\text{att}}^{\text{on}}\text{a}_i$
$\text{W}_{\text{att}}\hat{\text{h}}_t$
$\hat{\text{h}}_t$
$\text{W}_{\text{att}}\hat{\text{h}}_t$
$\text{U}_{\text{att}}^{\text{off}}\text{b}_i$
$\hat{\alpha}_{ti}^{\text{off}}$
$\alpha_{ti}^{\text{on}}$
$\hat{\text{c}}_t^{\text{on}}$
$\hat{\text{c}}_t^{\text{off}}$
$\text{U}_p^{\text{off}}\hat{\text{c}}_t^{\text{off}}$
$\text{U}_p^{\text{on}}\hat{\text{c}}_t^{\text{on}}$
$\alpha_{ti}^{\text{off}}$
$\text{c}_t^{\text{on}}$
$\text{c}_t^{\text{off}}$
Concat
$\text{W}_{\text{FC}}$
$\text{c}_t$

Fig. 2. Re-attention mechanism. “Concat” denotes concatenation operation.

$$\hat{\alpha}_{ti}^{\text{on}} = g\left(\mathbf{v}_{\text{att}}^T \tanh\left(\mathbf{W}_{\text{att}}\hat{\mathbf{h}}_t + \mathbf{U}_{\text{att}}^{\text{on}}\mathbf{a}_i + \mathbf{U}_f^{\text{on}}\hat{\mathbf{f}}_i^{\text{on}}\right)\right) \tag{9}$$

$$\hat{\alpha}_{ti}^{\text{off}} = g\left(\mathbf{v}_{\text{att}}^T \tanh\left(\mathbf{W}_{\text{att}}\hat{\mathbf{h}}_t + \mathbf{U}_{\text{att}}^{\text{off}}\mathbf{b}_i + \mathbf{U}_f^{\text{off}}\hat{\mathbf{f}}_i^{\text{off}}\right)\right) \tag{10}$$

where $\mathbf{v}_{\text{att}} \in \mathbb{R}^{n'}$, $\mathbf{W}_{\text{att}} \in \mathbb{R}^{n' \times n}$ are shared parameters while $\mathbf{U}_{\text{att}}^{\text{on}} \in \mathbb{R}^{n' \times D}$, $\mathbf{U}_{\text{att}}^{\text{off}} \in \mathbb{R}^{n' \times D}$, $\mathbf{U}_f^{\text{on}} \in \mathbb{R}^{n' \times k}$, $\mathbf{U}_f^{\text{off}} \in \mathbb{R}^{n' \times k}$ are unshared parameters. $\hat{\alpha}_{ti}^{\text{on}}$, $\hat{\alpha}_{ti}^{\text{off}}$ are the attention coefficients of $\mathbf{a}_i$, $\mathbf{b}_i$ at time step $t$. $\hat{\mathbf{f}}_i^{\text{on}}$ and $\hat{\mathbf{f}}_i^{\text{off}}$ denote the coverage vectors at the location $i$ of $\hat{\mathbf{F}}^{\text{on}}$ and $\hat{\mathbf{F}}^{\text{off}}$, which can be calculated as:

$$\hat{\mathbf{F}}^{\text{on}} = \mathbf{Q}^{\text{on}} * \sum_{l=1}^{t-1} \hat{\boldsymbol{\alpha}}_l^{\text{on}}, \quad \hat{\mathbf{F}}^{\text{off}} = \mathbf{Q}^{\text{off}} * \sum_{l=1}^{t-1} \hat{\boldsymbol{\alpha}}_l^{\text{off}} \tag{11}$$

where $\hat{\boldsymbol{\alpha}}_l^{\text{on}}$, $\hat{\boldsymbol{\alpha}}_l^{\text{off}}$ represent the attention probabilities of online and offline part at time step $l$. $\mathbf{Q}^{\text{on}}$ represents a $1D$ convolution filter with $k$ output channels while $\mathbf{Q}^{\text{off}}$ represents a $2D$ convolution filter with $k$ output channels as well.

Once the attention weights are calculated, the single modal context vectors of online and offline can be obtained as:

$$\hat{\mathbf{c}}_t^{\text{on}} = \sum_{i=1}^{L_{\text{on}}} \hat{\alpha}_{ti}^{\text{on}}\mathbf{a}_i, \quad \hat{\mathbf{c}}_t^{\text{off}} = \sum_{i=1}^{L_{\text{off}}} \hat{\alpha}_{ti}^{\text{off}}\mathbf{b}_i \tag{12}$$

Finally, we can obtain the multi-modal context vector by concatenating the single online and offline modal vectors as:

$$\mathbf{c}_t = \tanh\left(\mathbf{W}_{\text{FC}}\begin{bmatrix}\hat{\mathbf{c}}_t^{\text{on}} \\ \hat{\mathbf{c}}_t^{\text{off}}\end{bmatrix}\right) \tag{13}$$

where $\mathbf{W}_{\text{FC}} \in \mathbb{R}^{D \times 2D}$. In addition, to make full use of the information that comes from the online and offline modalities, we propose a re-attention mechanism as an enhanced version of the multi-modal attention mechanism which is shown in Fig. 2. Note that to simplify the illustration, we have omitted the coverage vectors. The re-attention mechanism can be divided into two parts, namely pre-attention model and fine-attention model. The pre-attention model is the same as the multi-modal attention mechanism. The fine-attention model is similar to pre-attention model and the difference is that we utilize single modal vectors of another modality which are computed by pre-attention model as an additional item to calculate the query of the fine-attention. Therefore, the computation of the fine-attention query can be represented as:

$$\mathbf{q}^{\text{on}}_{ti} = \mathbf{W}_{\text{att}}\hat{\mathbf{h}}_t + \mathbf{U}^{\text{on}}_{\text{att}}\mathbf{a}_i + \mathbf{U}^{\text{on}}_f \mathbf{f}^{\text{on}}_i + \mathbf{U}^{\text{off}}_p \hat{\mathbf{c}}^{\text{off}}_t \tag{14}$$

$$\mathbf{q}^{\text{off}}_{ti} = \mathbf{W}_{\text{att}}\hat{\mathbf{h}}_t + \mathbf{U}^{\text{off}}_{\text{att}}\mathbf{b}_i + \mathbf{U}^{\text{off}}_f \mathbf{f}^{\text{off}}_i + \mathbf{U}^{\text{on}}_p \hat{\mathbf{c}}^{\text{on}}_t \tag{15}$$

where $\mathbf{U}^{\text{on}}_p \in \mathbb{R}^{n' \times D}$, $\mathbf{U}^{\text{off}}_p \in \mathbb{R}^{n' \times D}$. Note that the same part of the pre-attention and the fine-attention models share parameters. Obviously, the fine-attention model can be executed repeatedly any times. We denote the multi-modal attention network with multi-modal attention mechanism as MAN and which with re-attention mechanism as E-MAN (Enhanced MAN).

## III. Experiments

Our experiments are conducted on CROHME competition database. We use the training set of CROHME 2014 as our training set, which consists of 8836 expressions. After training, we evaluate the performance of our models on CROHME 2014 and CROHME 2016 testing set. The CROHME 2014 testing set has 986 expressions while CROHME 2016 testing set has 1147 expressions.

Our model aims to minimize the predicted symbol probability as show in Eq. (8) and employs cross entropy (CE) as the criterion. Besides. we set weight decay as $10^{-5}$ to reduce overfitting. The multi-modal encoder has two channels, namely online channel and offline channel, respectively. The online channel is a DenseNet followed by two layers of bidirectional GRU and each GRU layer has 250 forward and 250 backward GRU units. The offline channel is a deeper DenseNet. The DenseNet in online channel has 5 dense blocks. Each block without bottleneck layer has 3 convolutional layers with kernel size of $(1 \times 3)$. The growth rate is 24 while the $\theta$ in transition layer is 1. The DenseNet in offline channel is actually DenseNet-BC [22] with $\theta = 0.5$ and has 3 dense blocks. Each block has 16 convolutional layers with kernel size of $(3 \times 3)$. The growth rate is also 24. The multi-modal decoder is two layers of unidirectional GRU layers with a multi-modal attention or re-attention mechanism. Each GRU layer in the decoder has 256 forward units. Note that there is an additional fully connected layer on top of online and offline channel of the encoder to convert the output dimension of these two channels to be the same. The embedding dimension $m$ and the number of output channels $k$ are both 256 while the attention dimension $n'$ and annotation dimension $D$ are both 500. The kernel sizes of convolution filters $\mathbf{Q}^{\text{on}}$, $\mathbf{Q}^{\text{off}}$ for computing coverage vectors are set to $(1 \times 7)$ and $(11 \times 11)$.

We train our network by AdaDelta algorithm and the corresponding hyperparameters are set as $\rho = 0.9$, $\varepsilon = 10^{-8}$. During the decoding stage, we expect to obtain the most likely LaTeX sequence. To achieve better performance, we employ beam search algorithm and maintain the 10 most likely candidate LaTeX symbols at each decoding step.

TABLE I
Performance comparisons on CROHME 2014 between original TAP, WAP and our implemented TAP, WAP.

| System | ExpRate(%) | $\leq$ 1(%) | $\leq$ 2(%) | $\leq$ 3(%) |
|---|---|---|---|---|
| TAP-Original | 46.86 | 61.87 | 65.82 | 66.63 |
| **TAP-Ours** | **48.47** | **63.28** | **67.34** | **67.95** |
| WAP-Original | 43.71 | 58.42 | 62.88 | 64.20 |
| **WAP-Ours** | **48.38** | **66.13** | **70.18** | **70.79** |

TABLE II
Comparisons of expression recognition rate (ExpRate) in % on CROHME 2014 and CROHME 2016.

| System | CROHME 2014 | | | CROHME 2016 | | |
|---|---|---|---|---|---|---|
| | ExpRate | $\leq 1$ | $\leq 2$ | ExpRate | $\leq 1$ | $\leq 2$ |
| UPV | 37.22 | 44.42 | 47.26 | - | - | - |
| Wiris | - | - | - | 49.61 | 60.42 | 64.69 |
| IM2TEX | 35.90 | - | - | - | - | - |
| TAP-Ours | 48.47 | 63.28 | 67.34 | 44.81 | 59.72 | 62.77 |
| WAP-Ours | 48.38 | 66.13 | 70.18 | 46.82 | 64.64 | 65.48 |
| **MAN** | **52.43** | **68.25** | **71.81** | **49.87** | **64.52** | **67.13** |
| **E-MAN** | **54.05** | **68.76** | **72.21** | **50.56** | **64.78** | **67.13** |

### *A. Recognition performance*

To be fairly comparable, the single-modal recognition system and multi-modal recognition system are implemented using the same configuration. In Table I, we first show our implemented TAP and WAP models for HMER are better than original TAP and WAP introduced in [7], [8]. Considering TAP, rather than only employing GRU-RNN as the encoder, we also employ a CNN layer for processing input feature to further improve the ability of feature extraction and help alleviate over-fitting. As for WAP, the encoders used in original WAP and our WAP are both DenseNet although the encoder is VGG [27] in [8]. Furthermore, our models are all implemented by Pytorch while original TAP and WAP are implemented by Theano, which may also influence the performance. The source codes of our own TAP and WAP are publicly available[1].

The TAP-Original and WAP-Original represent original TAP and WAP respectively while TAP-Ours and WAP-Ours represent our own TAP and WAP respectively. As we can see, our implemented TAP and WAP substantially outperform original TAP and WAP. We use TAP-Ours and WAP-Ours as default in the following experiments.

The overall ExpRate comparisons are demonstrated in Table II. The system UPV denotes the best system in all of submitted systems to CROHME 2014 competition, while the system Wiris denotes the best system in all of submitted systems to CROHME 2016 competition. The details can be seen in [28], [29]. IM2TEX [30] is a system which employs a coarse-to-fine attention and achieves comparable performance to UPV. We use MAN to denote our multi-modal attention network with the multi-modal attention mechanism and use E-MAN to denote our multi-modal attention network with the re-attention mechanism.

[1] https://github.com/jmwang66

The system MAN achieves an ExpRate of 52.43% on CROHME 2014 and 49.87% on CROHME 2016, which significantly outperforms all the published modal-specific models. With the re-attention mechanism, E-MAN achieves an ExpRate of 54.05% on CROHME 2014 and 50.56% on CROHME 2016 which further improves the performance over MAN. We also show the results of the expression recognition accuracies with one, two (and three) errors per expression, represented by "$\leq$ 1", "$\leq$ 2" (and "$\leq$ 3") in Table I and Table II, but they are not totally comparable with the submitted systems to CROHME competitions as we do not consider the segmentation error in our models.

### B. Attention visualization and modal complementarity

In this section, we show how the proposed attention mechanism is capable to attend the useful parts of the input at each decoding step through attention visualization. We only show the results of re-attention mechanism as it achieves better performance than multi-modal attention mechanism. Since our model has two modality inputs, the attention visualization also has two parts, namely online and offline attention. In addition, we utilize the red color to describe the attention probabilities. As shown in Fig. 3, the correctly recognized expression, $\int 2x^{-2}dx$ is used to show how the model translates this handwritten mathematical expression from a trajectory sequence and the corresponding static image into a LaTeX sequence " \int 2 x ∧ { - 2 } d x " step by step. When decoding basic math symbols, such as "$\int$", "2", "x", "d" and "-", the re-attention mechanism helps the model attend the corresponding parts well. For the spatial relationship in $x^{-2}$, the re-attention mechanism accurately distinguishes the superscript relationship to decode the symbol "∧". At the same time, the decoder generates a pair of braces "{}" right after detecting the superscript spatial relationship, which are necessary for LaTeX grammar.

Furthermore, the proposed re-attention mechanism can help the model generate more accurate attention by fully utilizing the information that comes from the online and offline modalities and making them complementary to each other, which can acquire better recognition performance than TAP or WAP. In Fig. 4, we show how the re-attention mechanism can acquire better alignments between features and symbols than using single online modality (TAP). The basic symbol "$x$" is unparsed in single online modality as the model focuses attention on both the current symbol "$x$" and the following symbols "$d$", "$x$". However, the online attention in multi-modal well distinguishes the last three symbols, "$x$", "$d$", "$x$".

In addition to better attention, the proposed model can also improve recognition performance by mitigating ambiguity problems even though the attentions of single modality and multi-modal are both accurate. As shown in Fig. 5, two representative examples are shown. For the online example, the writing trajectory of symbols, " $\in$ " and "t" is similar which can easily lead to recognition error. The similar problem exists in offline modality example. The symbol of the ground truth " $\beta$ " is similar to the symbol "p" due to the bad writing.

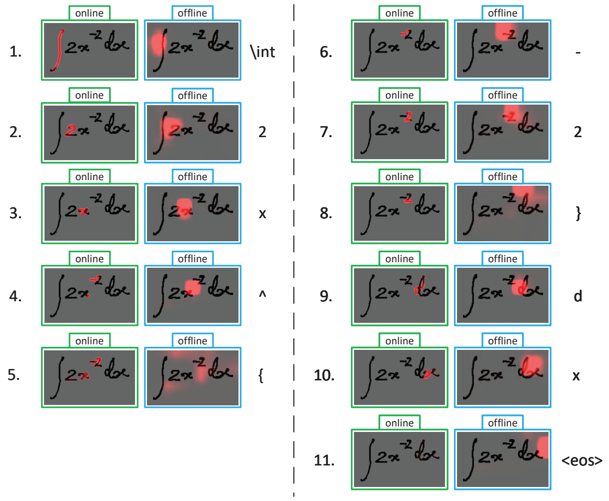


Fig. 3. Attention visualization for a correctly recognized example of a handwriting mathematical expression with the LaTeX ground truth " \int 2 x ∧ { - 2 } d x ". Numbers of 1 to 11 at the left of the images denote the orders of each decoding step.

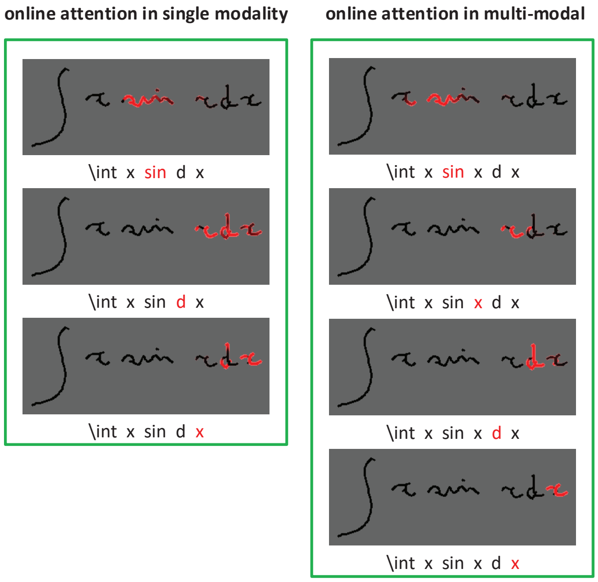


Fig. 4. Examples of online attention in single modal (TAP) and multi-modal (re-attention). Parts of the recognized LaTeX sequence are printed below each image (the red areas in the images indicate the attended regions, and the red text in the LaTeX sequence indicates the corresponding words).

As a result, it is hard to distinguish only by using the static information from offline modality. However, our multi-modal attention network can acquire information from both online and offline modalities to solve ambiguity problems above and achieve better recognition results.

## IV. Conclusions

In this study, a multi-modal attention network (MAN) based on encoder-decoder framework for HMER is introduced. To make full use of the information that comes from the online

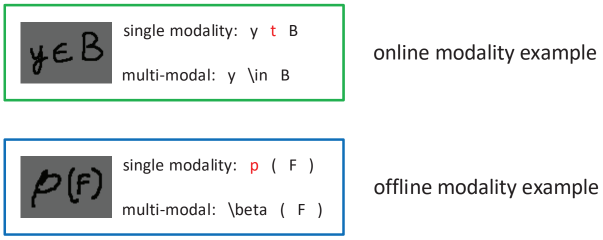


Fig. 5. Examples to show that re-attention mechanism can help mitigate ambiguity problems in single online modality (TAP) or single offline modality (WAP). The symbols in red color represent incorrectly recognized symbols.

and offline modalities, we propose a re-attention mechanism as an enhanced version of the multi-modal attention mechanism. We achieve significant improvements on both CROHME 2014 competition and CROHME 2016 competition compared with state-of-the-art single-modal systems. Through attention visualization, we show how our model improves recognition performance with better alignments. In the future, we aim to investigate a better way to combine two modalities to achieve a higher performance and efficiency.

## V. Acknowledgment

This work was supported in part by the National Key R&D Program of China under contract No. 2017YFB1002202, the National Natural Science Foundation of China under Grants No. 61671422 and U1613211, the Key Science and Technology Project of Anhui Province under Grant No. 17030901005. This work was also funded by Huawei Noah's Ark Lab.

## References

[1] E. G. Miller and P. A. Viola, "Ambiguity and constraint in mathematical expression recognition," in *AAAI/IAAI*, 1998, pp. 784–791.

[2] R. H. Anderson, "Syntax-directed recognition of hand-printed two-dimensional mathematics," in *Symposium on Interactive Systems for Experimental Applied Mathematics: Proceedings of the Association for Computing Machinery Inc. Symposium*. ACM, 1967, pp. 436–459.

[3] A. Belaid and J.-P. Haton, "A syntactic approach for handwritten mathematical formula recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 1, pp. 105–111, 1984.

[4] U. Garain and B. B. Chaudhuri, "Recognition of online handwritten mathematical expressions," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 34, no. 6, pp. 2366–2376, 2004.

[5] A.-M. Awal, H. Mouchère, and C. Viard-Gaudin, "A global learning approach for an online handwritten mathematical expression recognition system," *Pattern Recognition Letters*, vol. 35, pp. 68–77, 2014.

[6] F. Álvaro, J.-A. Sánchez, and J.-M. Benedí, "Offline features for classifying handwritten math symbols with recurrent neural networks," in *Pattern recognition (icpr), 2014 22nd international conference on*. IEEE, 2014, pp. 2944–2949.

[7] J. Zhang, J. Du, and L. Dai, "Track, attend and parse (TAP): An end-to-end framework for online handwritten mathematical expression recognition," *IEEE Transactions on Multimedia*, vol. 21, no. 1, pp. 221–233, 2019.

[8] J. Zhang, J. Du, S. Zhang, D. Liu, Y. Hu, J. Hu, S. Wei, and L. Dai, "Watch, attend and parse: an end-to-end neural network based approach to handwritten mathematical expression recognition," *Pattern Recognition*, vol. 71, pp. 196–206, 2017.

[9] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 689–696.

[10] Y. Gu, K. Yang, S. Fu, S. Chen, X. Li, and I. Marsic, "Hybrid attention based multimodal network for spoken language classification," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 2379–2390.

[11] C. Hori, T. Hori, T.-Y. Lee, Z. Zhang, B. Harsham, J. R. Hershey, T. K. Marks, and K. Sumi, "Attention-based multimodal fusion for video description," in *Computer Vision (ICCV), 2017 IEEE International Conference on*. IEEE, 2017, pp. 4203–4212.

[12] H. Nam, J.-W. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 299–307.

[13] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *arXiv preprint arXiv:1511.00561*, 2015.

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

[15] K. Cho, A. Courville, and Y. Bengio, "Describing multimedia content using attention-based encoder-decoder networks," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1875–1886, 2015.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[17] L. Lamport, *LaTeX: A document preparation system: User's guide and reference. illustrations by Duane Bibby*. Reading, Mass: Addison-Wesley Professional. ISBN 0-201-52983-1, 1994.

[18] X.-Y. Zhang, F. Yin, Y.-M. Zhang, C.-L. Liu, and Y. Bengio, "Drawing and recognizing chinese characters with recurrent neural network," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 849–862, 2018.

[19] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

[20] J. Zhang, J. Tang, and L.-R. Dai, "RNN-BLSTM based multi-pitch estimation." in *INTERSPEECH*, 2016, pp. 1785–1789.

[21] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[22] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[23] J. Zhang, J. Du, and L. Dai, "A gru-based encoder-decoder approach with attention for online handwritten mathematical expression recognition," in *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on*, vol. 1. IEEE, 2017, pp. 902–907.

[24] K. Cho, A. Courville, and Y. Bengio, "Describing multimedia content using attention-based encoder-decoder networks," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1875–1886, 2015.

[25] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 4945–4949.

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

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

[28] H. Mouchere, C. Viard-Gaudin, R. Zanibbi, and U. Garain, "Icfhr 2014 competition on recognition of on-line handwritten mathematical expressions (crohme 2014)," in *2014 14th International Conference on Frontiers in Handwriting Recognition*. IEEE, 2014, pp. 791–796.

[29] H. Mouchère, C. Viard-Gaudin, R. Zanibbi, and U. Garain, "Icfhr2016 crohme: Competition on recognition of online handwritten mathematical expressions," in *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 2016, pp. 607–612.

[30] Y. Deng, A. Kanervisto, J. Ling, and A. M. Rush, "Image-to-markup generation with coarse-to-fine attention," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 980–989.