Track, Attend and Parse (TAP): An End-to-end Framework for Online Handwritten Mathematical Expression Recognition

何福铿

中山大学数学学院

2019 年 9 月 20 日

Outline

1. 论文介绍
2. 网络结构
   - 预处理
   - Tracker
   - Parser
   - Guided hybrid attention
     - Spatial attention
     - Coverage model
     - Temporal attention
     - Attention guider
   - TAP 和 WAP
3. Training
   - Training
4. Experiment Results
   - Experiment Results

论文介绍

- 数学公式在科学文档中扮演重要的角色。
- 数学公式识别分为在线手写数学公式识别和离线手写数学公式识别。本文研究的是在线手写数学公式识别 (OHMER)。
- 目前，OHMER 的识别率不高，仅在 40% 到 60% 之间。
- OHMER 的目的是将数学公式转化为 LaTeX 表达式。

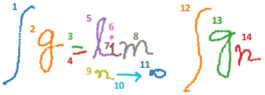


Figure: 在线手写数学公式

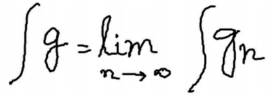


Figure: 离线手写数学公式

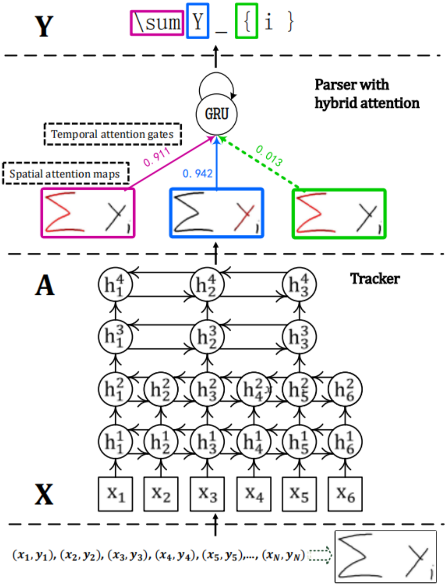
论文介绍

- OHMER 通常包括两个主要问题：符号识别和结构分析。
- OHMER 可以用分步识别和整体识别方法。分步识别方法通常有三个步骤符号识别，符号分割以及符号结构分析，但是符号识别无法利用上下文的信息。而整体识别方法则过于耗时。
- (P. A. Chou, 1989；S. Lavirotte and L. Pottier, 1998) 设计数学公式语法进行符号结构分析。

论文介绍

- 本文的方法为 Track, Attend and Parse(TAP)，包括一个 tracker，一个 parser 并且附带 guided hybrid attention(GHA) 机制。
- TAP 是数据驱动的，不需要预定一个数学语法；TAP 是一个 End-to-end 的识别方法；符号划分可以由注意力机制自动得到。
- tracker 以二维手写轨迹作为输入，使用双向 RNN（GRU 单元），把手写信息转化为 high-level representations。
- praser 是带有 GHA 机制的单向 RNN（GRU 单元），可以把 high-level representations 转化为 LaTeX 符号，每次转化一个符号。
- GHA 包含了一个 converge based spatial attention，temporal attention 以及 attention guider。

论文介绍


Y
\sum Y _ { i }
Parser with
hybrid attention
GRU
Temporal attention gates
Spatial attention maps
0. 911
0. 942
0. 013
A
Tracker
h_1^4 h_2^4 h_3^4
h_1^3 h_2^3 h_3^3
h_1^2 h_2^2 h_3^2 h_4^2 h_5^2 h_6^2
h_1^1 h_2^1 h_3^1 h_4^1 h_5^1 h_6^1
X
x_1 x_2 x_3 x_4 x_5 x_6
$(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4), (x_5, y_5), \ldots, (x_N, y_N)$

论文介绍

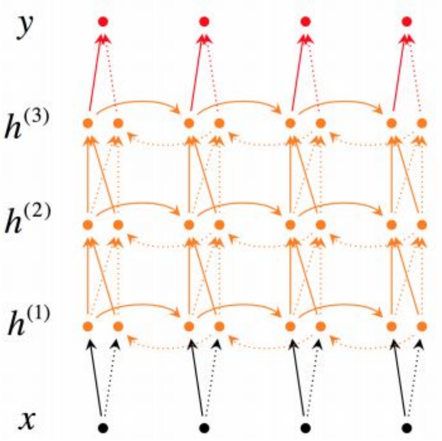


Figure: BiRNN

论文介绍

这篇文章的创新点在于：

- TAP 框架是一个 end-to-end 的方法，不设置特定的数学语法，克服了传统算法的缺点。
- TAP 框架使用 attention guider 去确定需要 attend 的地方。
- TAP 框架聚合了 FCN-based watcher 和 stronger GRU-based language model，充分利用了公式中的 offline 和 online 信息。
- 我们对 GHA 的可视化可以解释 TAP 如何实现自动符号分割以及结构分析。

Outline

1. 论文介绍
2. 网络结构
   - 预处理
   - Tracker
   - Parser
   - Guided hybrid attention
     - Spatial attention
     - Coverage model
     - Temporal attention
     - Attention guider
   - TAP 和 WAP
3. Training
   - Training
4. Experiment Results
   - Experiment Results

设输入的笔画（raw data）为

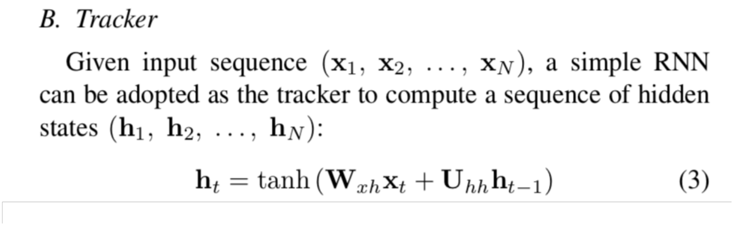
$$\{[x_1, y_1, s_1], [x_2, y_2, s_2], \cdots, [x_N, y_N, s_N]\} \quad (1)$$

长度为 N，其中 x_i 和 y_i 分别为手写输入笔画的横纵坐标，s_i 表示的是该点所在笔画的序号。
在标准化和归一化之后，我们对每个点提取了八维的特征向量：

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

Outline

1. 论文介绍
2. 网络结构
    - 预处理
    - Tracker
    - Parser
    - Guided hybrid attention
        - Spatial attention
        - Coverage model
        - Temporal attention
        - Attention guider
    - TAP 和 WAP
3. Training
    - Training
4. Experiment Results
    - Experiment Results



B. Tracker

Given input sequence $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$, a simple RNN can be adopted as the tracker to compute a sequence of hidden states $(\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_N)$:

$$\mathbf{h}_t = \tanh\left(\mathbf{W}_{xh}\mathbf{x}_t + \mathbf{U}_{hh}\mathbf{h}_{t-1}\right) \qquad (3)$$

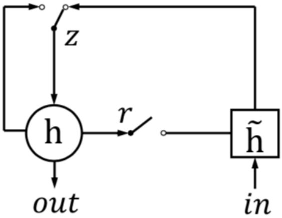
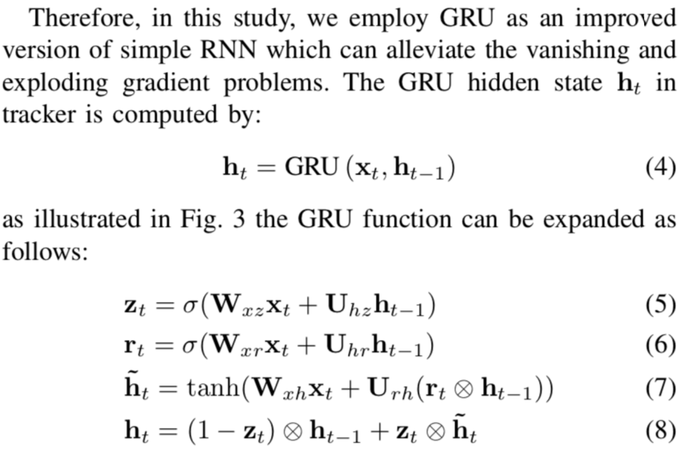


Fig. 3. Illustration of GRU function, $\mathbf{z}$ denotes update gate, $\mathbf{r}$ denotes reset gate, $\tilde{\mathbf{h}}$ denotes candidate activation and $\mathbf{h}$ denotes the output activation.



Therefore, in this study, we employ GRU as an improved version of simple RNN which can alleviate the vanishing and exploding gradient problems. The GRU hidden state $\mathbf{h}_t$ in tracker is computed by:

$$\mathbf{h}_t = \mathrm{GRU}\left(\mathbf{x}_t, \mathbf{h}_{t-1}\right) \tag{4}$$

as illustrated in Fig. 3 the GRU function can be expanded as follows:

$$\mathbf{z}_t = \sigma(\mathbf{W}_{xz}\mathbf{x}_t + \mathbf{U}_{hz}\mathbf{h}_{t-1}) \tag{5}$$

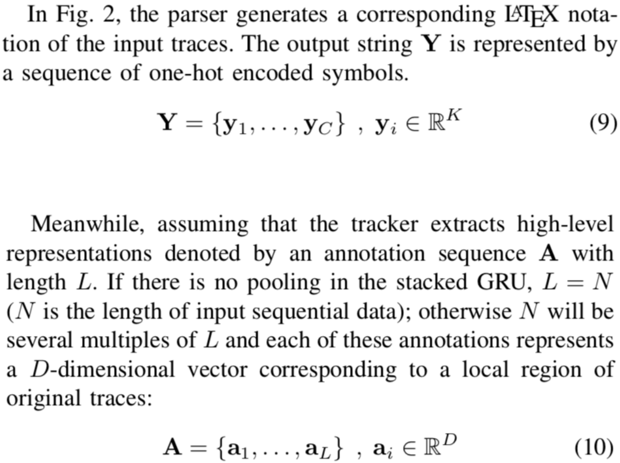
$$\mathbf{r}_t = \sigma(\mathbf{W}_{xr}\mathbf{x}_t + \mathbf{U}_{hr}\mathbf{h}_{t-1}) \tag{6}$$

$$\tilde{\mathbf{h}}_t = \tanh(\mathbf{W}_{xh}\mathbf{x}_t + \mathbf{U}_{rh}(\mathbf{r}_t \otimes \mathbf{h}_{t-1})) \tag{7}$$

$$\mathbf{h}_t = (1 - \mathbf{z}_t) \otimes \mathbf{h}_{t-1} + \mathbf{z}_t \otimes \tilde{\mathbf{h}}_t \tag{8}$$

Outline

1. 论文介绍
2. 网络结构
   - 预处理
   - Tracker
   - Parser
   - Guided hybrid attention
     - Spatial attention
     - Coverage model
     - Temporal attention
     - Attention guider
   - TAP 和 WAP
3. Training
   - Training
4. Experiment Results
   - Experiment Results

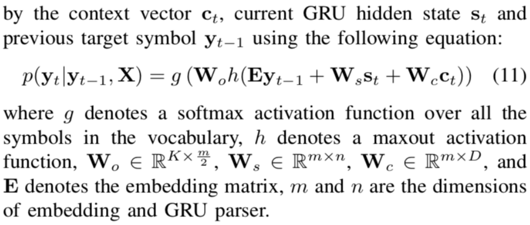


In Fig. 2, the parser generates a corresponding LaTeX notation of the input traces. The output string $\mathbf{Y}$ is represented by a sequence of one-hot encoded symbols.

$$\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_C\} \ , \ \mathbf{y}_i \in \mathbb{R}^K \tag{9}$$

Meanwhile, assuming that the tracker extracts high-level representations denoted by an annotation sequence $\mathbf{A}$ with length L. If there is no pooling in the stacked GRU, $L = N$ (N is the length of input sequential data); otherwise N will be several multiples of L and each of these annotations represents a D-dimensional vector corresponding to a local region of original traces:

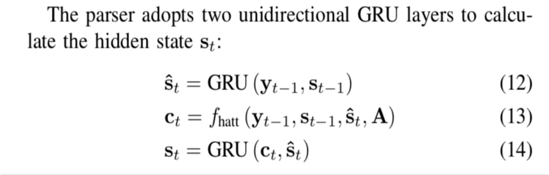
$$\mathbf{A} = \{\mathbf{a}_1, \ldots, \mathbf{a}_L\} \ , \ \mathbf{a}_i \in \mathbb{R}^D \tag{10}$$



by the context vector $\mathbf{c}_t$, current GRU hidden state $\mathbf{s}_t$ and previous target symbol $\mathbf{y}_{t-1}$ using the following equation:

$$p(\mathbf{y}_t|\mathbf{y}_{t-1},\mathbf{X}) = g\left(\mathbf{W}_o h(\mathbf{E}\mathbf{y}_{t-1} + \mathbf{W}_s\mathbf{s}_t + \mathbf{W}_c\mathbf{c}_t)\right) \quad (11)$$

where g denotes a softmax activation function over all the symbols in the vocabulary, h denotes a maxout activation function, $\mathbf{W}_o \in \mathbb{R}^{K\times\frac{m}{2}}$, $\mathbf{W}_s \in \mathbb{R}^{m\times n}$, $\mathbf{W}_c \in \mathbb{R}^{m\times D}$, and $\mathbf{E}$ denotes the embedding matrix, m and n are the dimensions of embedding and GRU parser.

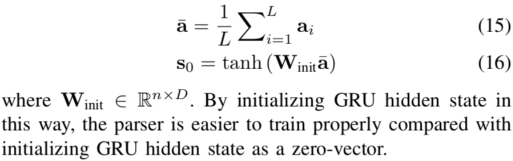


The parser adopts two unidirectional GRU layers to calculate the hidden state $\mathbf{s}_t$:

$$\hat{\mathbf{s}}_t = \mathrm{GRU}\left(\mathbf{y}_{t-1}, \mathbf{s}_{t-1}\right) \quad (12)$$

$$\mathbf{c}_t = f_{\mathrm{hatt}}\left(\mathbf{y}_{t-1}, \mathbf{s}_{t-1}, \hat{\mathbf{s}}_t, \mathbf{A}\right) \quad (13)$$

$$\mathbf{s}_t = \mathrm{GRU}\left(\mathbf{c}_t, \hat{\mathbf{s}}_t\right) \quad (14)$$



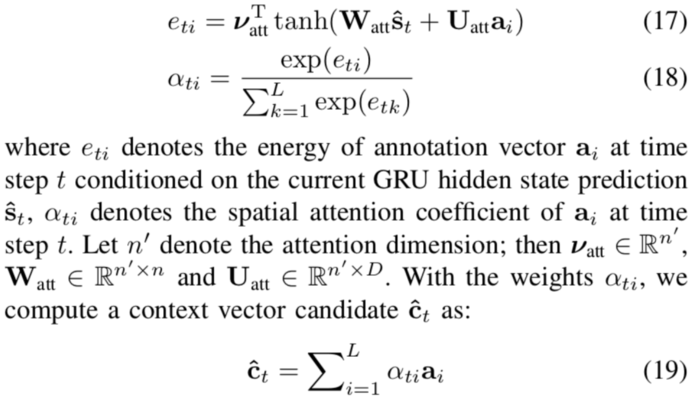
$$\bar{\mathbf{a}} = \frac{1}{L}\sum\nolimits_{i=1}^{L} \mathbf{a}_i \quad (15)$$

$$\mathbf{s}_0 = \tanh\left(\mathbf{W}_{\mathrm{init}}\bar{\mathbf{a}}\right) \quad (16)$$

where $\mathbf{W}_{\mathrm{init}} \in \mathbb{R}^{n\times D}$. By initializing GRU hidden state in this way, the parser is easier to train properly compared with initializing GRU hidden state as a zero-vector.

Outline

1. 论文介绍
2. 网络结构
   - 预处理
   - Tracker
   - Parser
   - Guided hybrid attention
     - Spatial attention
     - Coverage model
     - Temporal attention
     - Attention guider
   - TAP 和 WAP
3. Training
   - Training
4. Experiment Results
   - Experiment Results

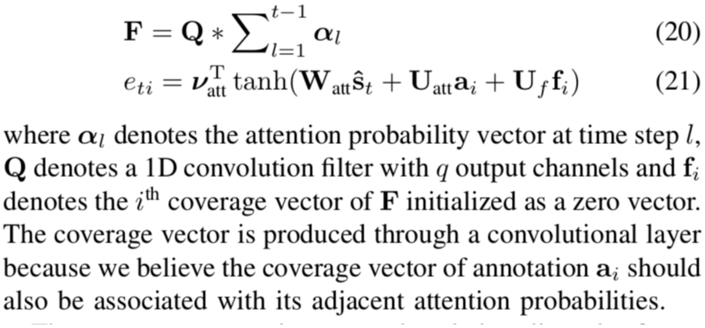


$$e_{ti} = \boldsymbol{\nu}_{\text{att}}^{\text{T}} \tanh(\mathbf{W}_{\text{att}}\hat{\mathbf{s}}_t + \mathbf{U}_{\text{att}}\mathbf{a}_i) \tag{17}$$

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{L} \exp(e_{tk})} \tag{18}$$

where e_{ti} denotes the energy of annotation vector $\mathbf{a}_i$ at time step t conditioned on the current GRU hidden state prediction $\hat{\mathbf{s}}_t$, α_{ti} denotes the spatial attention coefficient of $\mathbf{a}_i$ at time step t. Let n' denote the attention dimension; then $\boldsymbol{\nu}_{\text{att}} \in \mathbb{R}^{n'}$, $\mathbf{W}_{\text{att}} \in \mathbb{R}^{n' \times n}$ and $\mathbf{U}_{\text{att}} \in \mathbb{R}^{n' \times D}$. With the weights α_{ti}, we compute a context vector candidate $\hat{\mathbf{c}}_t$ as:

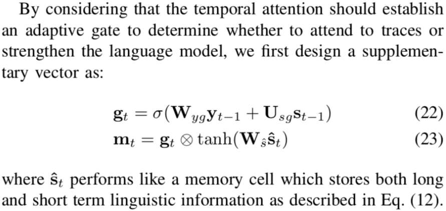
$$\hat{\mathbf{c}}_t = \sum_{i=1}^{L} \alpha_{ti}\mathbf{a}_i \tag{19}$$



$$\mathbf{F} = \mathbf{Q} * \sum_{l=1}^{t-1} \boldsymbol{\alpha}_l \tag{20}$$

$$e_{ti} = \boldsymbol{\nu}_{\text{att}}^{\text{T}} \tanh(\mathbf{W}_{\text{att}} \hat{\mathbf{s}}_t + \mathbf{U}_{\text{att}} \mathbf{a}_i + \mathbf{U}_f \mathbf{f}_i) \tag{21}$$

where $\boldsymbol{\alpha}_l$ denotes the attention probability vector at time step l, $\mathbf{Q}$ denotes a 1D convolution filter with q output channels and $\mathbf{f}_i$ denotes the i^{th} coverage vector of $\mathbf{F}$ initialized as a zero vector. The coverage vector is produced through a convolutional layer because we believe the coverage vector of annotation $\mathbf{a}_i$ should also be associated with its adjacent attention probabilities.

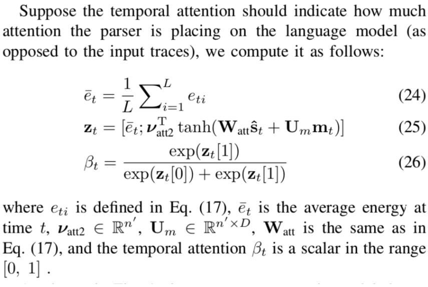


By considering that the temporal attention should establish an adaptive gate to determine whether to attend to traces or strengthen the language model, we first design a supplementary vector as:

$$\mathbf{g}_t = \sigma(\mathbf{W}_{yg}\mathbf{y}_{t-1} + \mathbf{U}_{sg}\mathbf{s}_{t-1}) \tag{22}$$

$$\mathbf{m}_t = \mathbf{g}_t \otimes \tanh(\mathbf{W}_{\hat{s}}\hat{\mathbf{s}}_t) \tag{23}$$

where $\hat{\mathbf{s}}_t$ performs like a memory cell which stores both long and short term linguistic information as described in Eq. (12).



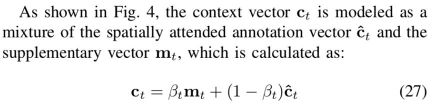
Suppose the temporal attention should indicate how much attention the parser is placing on the language model (as opposed to the input traces), we compute it as follows:

$$\bar{e}_t = \frac{1}{L}\sum\nolimits_{i=1}^{L} e_{ti} \tag{24}$$

$$\mathbf{z}_t = [\bar{e}_t; \boldsymbol{\nu}_{\text{att2}}^{\mathrm{T}} \tanh(\mathbf{W}_{\text{att}}\hat{\mathbf{s}}_t + \mathbf{U}_m\mathbf{m}_t)] \tag{25}$$

$$\beta_t = \frac{\exp(\mathbf{z}_t[1])}{\exp(\mathbf{z}_t[0]) + \exp(\mathbf{z}_t[1])} \tag{26}$$

where e_{ti} is defined in Eq. (17), $\bar{e}_t$ is the average energy at time t, $\boldsymbol{\nu}_{\text{att2}} \in \mathbb{R}^{n'}$, $\mathbf{U}_m \in \mathbb{R}^{n' \times D}$, $\mathbf{W}_{\text{att}}$ is the same as in Eq. (17), and the temporal attention β_t is a scalar in the range $[0,\ 1]$.



As shown in Fig. 4, the context vector $\mathbf{c}_t$ is modeled as a mixture of the spatially attended annotation vector $\hat{\mathbf{c}}_t$ and the supplementary vector $\mathbf{m}_t$, which is calculated as:

$$\mathbf{c}_t = \beta_t\mathbf{m}_t + (1 - \beta_t)\hat{\mathbf{c}}_t \tag{27}$$

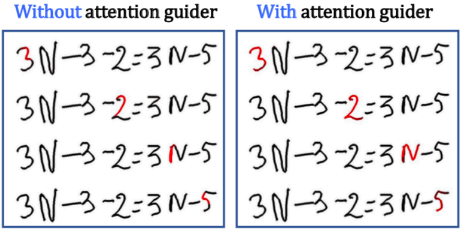
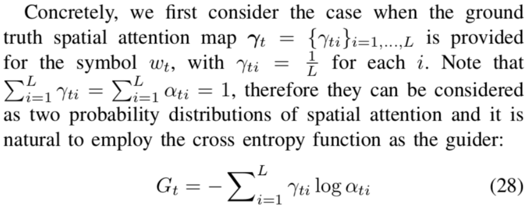


Fig. 5. Examples of attention maps with and without attention guider.

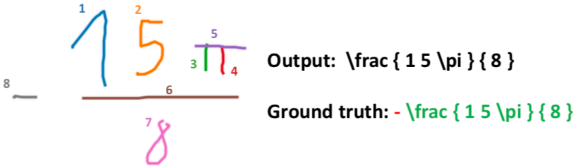


Concretely, we first consider the case when the ground truth spatial attention map $\gamma_t = \{\gamma_{ti}\}_{i=1,\ldots,L}$ is provided for the symbol w_t, with $\gamma_{ti} = \frac{1}{L}$ for each i. Note that $\sum_{i=1}^{L} \gamma_{ti} = \sum_{i=1}^{L} \alpha_{ti} = 1$, therefore they can be considered as two probability distributions of spatial attention and it is natural to employ the cross entropy function as the guider:

$$G_t = -\sum_{i=1}^{L} \gamma_{ti} \log \alpha_{ti} \tag{28}$$

Outline

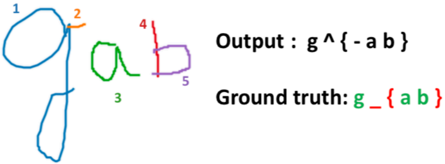
1. 论文介绍
2. 网络结构
   - 预处理
   - Tracker
   - Parser
   - Guided hybrid attention
     - Spatial attention
     - Coverage model
     - Temporal attention
     - Attention guider
   - TAP 和 WAP
3. Training
   - Training
4. Experiment Results
   - Experiment Results



Output: \frac { 1 5 \pi } { 8 }

Ground truth: - \frac { 1 5 \pi } { 8 }

Fig. 6. An incorrectly recognized example due to the delayed stroke.



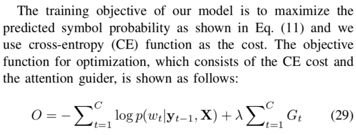
Output : g ^ { - a b }

Ground truth: g _ { a b }

Fig. 7. An incorrectly recognized example due to the inserted stroke.

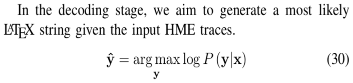
Outline

1. 论文介绍
2. 网络结构
   - 预处理
   - Tracker
   - Parser
   - Guided hybrid attention
     - Spatial attention
     - Coverage model
     - Temporal attention
     - Attention guider
   - TAP 和 WAP
3. Training
   - Training
4. Experiment Results
   - Experiment Results



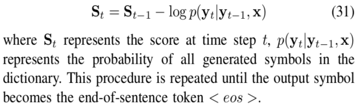
The training objective of our model is to maximize the predicted symbol probability as shown in Eq. (11) and we use cross-entropy (CE) function as the cost. The objective function for optimization, which consists of the CE cost and the attention guider, is shown as follows:

$$O = -\sum_{t=1}^{C} \log p(w_t|\mathbf{y}_{t-1}, \mathbf{X}) + \lambda \sum_{t=1}^{C} G_t \tag{29}$$



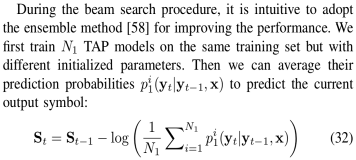
In the decoding stage, we aim to generate a most likely LaTeX string given the input HME traces.

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} \log P\left(\mathbf{y}|\mathbf{x}\right) \tag{30}$$



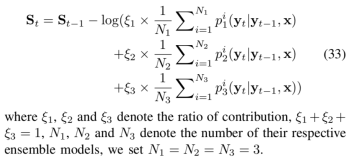
$$\mathbf{S}_t = \mathbf{S}_{t-1} - \log p(\mathbf{y}_t|\mathbf{y}_{t-1}, \mathbf{x}) \tag{31}$$

where $\mathbf{S}_t$ represents the score at time step t, $p(\mathbf{y}_t|\mathbf{y}_{t-1}, \mathbf{x})$ represents the probability of all generated symbols in the dictionary. This procedure is repeated until the output symbol becomes the end-of-sentence token $< eos >$.



During the beam search procedure, it is intuitive to adopt the ensemble method [58] for improving the performance. We first train N_1 TAP models on the same training set but with different initialized parameters. Then we can average their prediction probabilities $p_1^i(\mathbf{y}_t|\mathbf{y}_{t-1}, \mathbf{x})$ to predict the current output symbol:

$$\mathbf{S}_t = \mathbf{S}_{t-1} - \log\left(\frac{1}{N_1}\sum_{i=1}^{N_1} p_1^i(\mathbf{y}_t|\mathbf{y}_{t-1}, \mathbf{x})\right) \tag{32}$$

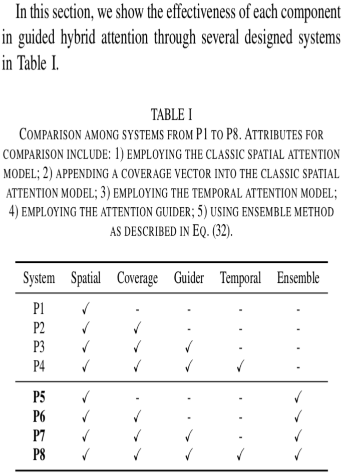


$$\begin{aligned} \mathbf{S}_t = \mathbf{S}_{t-1} - \log(\xi_1 \times \frac{1}{N_1}\sum_{i=1}^{N_1} p_1^i(\mathbf{y}_t|\mathbf{y}_{t-1}, \mathbf{x}) \\ + \xi_2 \times \frac{1}{N_2}\sum_{i=1}^{N_2} p_2^i(\mathbf{y}_t|\mathbf{y}_{t-1}, \mathbf{x}) \\ + \xi_3 \times \frac{1}{N_3}\sum_{i=1}^{N_3} p_3^i(\mathbf{y}_t|\mathbf{y}_{t-1}, \mathbf{x})) \end{aligned} \tag{33}$$

where ξ_1, ξ_2 and ξ_3 denote the ratio of contribution, $\xi_1 + \xi_2 + \xi_3 = 1$, N_1, N_2 and N_3 denote the number of their respective ensemble models, we set $N_1 = N_2 = N_3 = 3$.

Outline

1. 论文介绍
2. 网络结构
   - 预处理
   - Tracker
   - Parser
   - Guided hybrid attention
     - Spatial attention
     - Coverage model
     - Temporal attention
     - Attention guider
   - TAP 和 WAP
3. Training
   - Training
4. Experiment Results
   - Experiment Results



In this section, we show the effectiveness of each component in guided hybrid attention through several designed systems in Table I.

TABLE I
COMPARISON AMONG SYSTEMS FROM P1 TO P8. ATTRIBUTES FOR COMPARISON INCLUDE: 1) EMPLOYING THE CLASSIC SPATIAL ATTENTION MODEL; 2) APPENDING A COVERAGE VECTOR INTO THE CLASSIC SPATIAL ATTENTION MODEL; 3) EMPLOYING THE TEMPORAL ATTENTION MODEL; 4) EMPLOYING THE ATTENTION GUIDER; 5) USING ENSEMBLE METHOD AS DESCRIBED IN EQ. (32).

System	Spatial	Coverage	Guider	Temporal	Ensemble
P1	√	-	-	-	-
P2	√	√	-	-	-
P3	√	√	√	-	-
P4	√	√	√	√	-
P5	√	-	-	-	√
P6	√	√	-	-	√
P7	√	√	√	-	√
P8	√	√	√	√	√

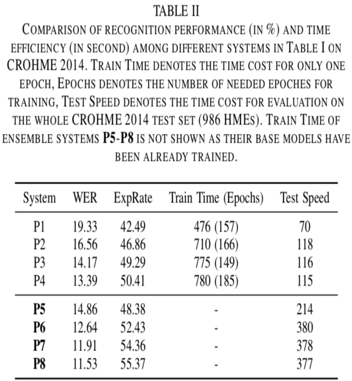


TABLE II
COMPARISON OF RECOGNITION PERFORMANCE (IN %) AND TIME EFFICIENCY (IN SECOND) AMONG DIFFERENT SYSTEMS IN TABLE I ON CROHME 2014. TRAIN TIME DENOTES THE TIME COST FOR ONLY ONE EPOCH, EPOCHS DENOTES THE NUMBER OF NEEDED EPOCHES FOR TRAINING, TEST SPEED DENOTES THE TIME COST FOR EVALUATION ON THE WHOLE CROHME 2014 TEST SET (986 HMES). TRAIN TIME OF ENSEMBLE SYSTEMS **P5**-**P8** IS NOT SHOWN AS THEIR BASE MODELS HAVE BEEN ALREADY TRAINED.

System	WER	ExpRate	Train Time (Epochs)	Test Speed
P1	19.33	42.49	476 (157)	70
P2	16.56	46.86	710 (166)	118
P3	14.17	49.29	775 (149)	116
P4	13.39	50.41	780 (185)	115
P5	14.86	48.38	-	214
P6	12.64	52.43	-	380
P7	11.91	54.36	-	378
P8	11.53	55.37	-	377

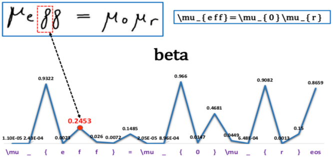


Fig. 8. An illustration of using temporal attention, symbol in the red rectangle is incorrectly recognized as “8” without temporal attention and correctly recognized as “*f*” with temporal attention.

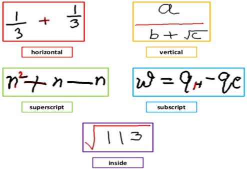


Fig. 9. Learning of five spatial relationships (horizontal, vertical, subscript, superscript and inside) through attention visualization.

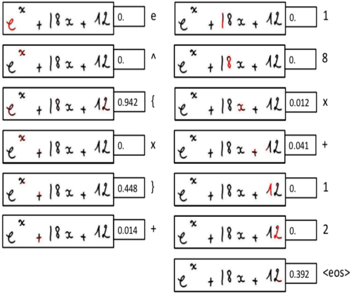


Fig. 10. Hybrid attention visualization for an example of an HME with the LaTeX ground truth “e ∧ { x } + 1 8 x + 1 2”, spatial attention is shown through red color in images and temporal attention is shown in the attached boxes, on the right of boxes are predicted symbols.

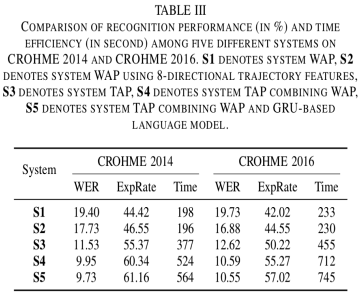


TABLE III
COMPARISON OF RECOGNITION PERFORMANCE (IN %) AND TIME EFFICIENCY (IN SECOND) AMONG FIVE DIFFERENT SYSTEMS ON CROHME 2014 AND CROHME 2016. **S1** DENOTES SYSTEM WAP, **S2** DENOTES SYSTEM WAP USING 8-DIRECTIONAL TRAJECTORY FEATURES, **S3** DENOTES SYSTEM TAP, **S4** DENOTES SYSTEM TAP COMBINING WAP, **S5** DENOTES SYSTEM TAP COMBINING WAP AND GRU-BASED LANGUAGE MODEL.

System	CROHME 2014			CROHME 2016		
	WER	ExpRate	Time	WER	ExpRate	Time
S1	19.40	44.42	198	19.73	42.02	233
S2	17.73	46.55	196	16.88	44.55	230
S3	11.53	55.37	377	12.62	50.22	455
S4	9.95	60.34	524	10.59	55.27	712
S5	9.73	61.16	564	10.55	57.02	745

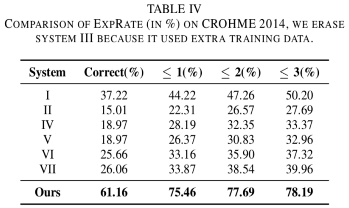


TABLE IV
COMPARISON OF EXPRATE (IN %) ON CROHME 2014, WE ERASE SYSTEM III BECAUSE IT USED EXTRA TRAINING DATA.

System	Correct(%)	$\leq$ 1(%)	$\leq$ 2(%)	$\leq$ 3(%)
I	37.22	44.22	47.26	50.20
II	15.01	22.31	26.57	27.69
IV	18.97	28.19	32.35	33.37
V	18.97	26.37	30.83	32.96
VI	25.66	33.16	35.90	37.32
VII	26.06	33.87	38.54	39.96
Ours	**61.16**	**75.46**	**77.69**	**78.19**

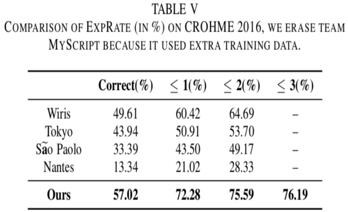


TABLE V
COMPARISON OF EXPRATE (IN %) ON CROHME 2016, WE ERASE TEAM MYSCRIPT BECAUSE IT USED EXTRA TRAINING DATA.

	Correct(%)	$\leq$ 1(%)	$\leq$ 2(%)	$\leq$ 3(%)
Wiris	49.61	60.42	64.69	–
Tokyo	43.94	50.91	53.70	–
São Paolo	33.39	43.50	49.17	–
Nantes	13.34	21.02	28.33	–
Ours	**57.02**	**72.28**	**75.59**	**76.19**

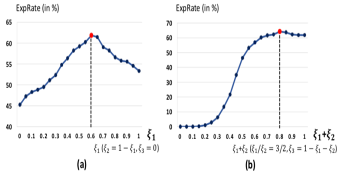
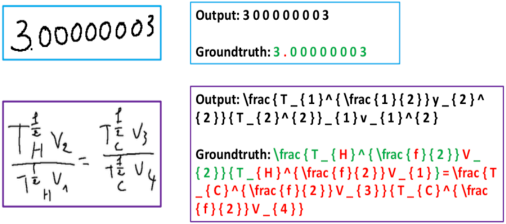


Fig. 11. (a) The curve of ExpRate with respect to ξ_1 (the contribution of TAP) in the ensemble of TAP and WAP (with $\xi_2 = 1 - \xi_1, \xi_3 = 0$); (b) The curve of ExpRate with respect to $(\xi_1 + \xi_2)$ in the ensemble of (TAP+WAP) and GRU-LM (with $\xi_1/\xi_2 = 3/2, \xi_3 = 1 - \xi_1 - \xi2$). We draw the two curves on the validation set.



Output: 3 0 0 0 0 0 0 0 3

Groundtruth: 3 . 0 0 0 0 0 0 0 3

Output: \frac { T _ { 1 } ^ { \frac { 1 } { 2 } } y _ { 2 } ^ { 2 } } { T _ { 2 } ^ { 2 } } _ { 1 } v _ { 1 } ^ { 2 }

Groundtruth: \frac { T _ { H } ^ { \frac { f } { 2 } } V _ { 2 } } { T _ { H } ^ { \frac { f } { 2 } } V _ { 1 } } = \frac { T _ { C } ^ { \frac { f } { 2 } } V _ { 3 } } { T _ { C } ^ { \frac { f } { 2 } } V _ { 4 } }

Fig. 12. Two examples of HME which are incorrectly recognized, besides the HME images are their predicted output and ground truth, in the ground truth green texts are predicted correctly while red texts are predicted incorrectly.

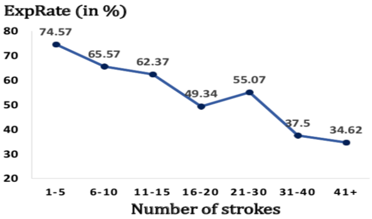


Fig. 13. Number of strokes vs. ExpRate (in %) on CROHME 2014.

Thank you!