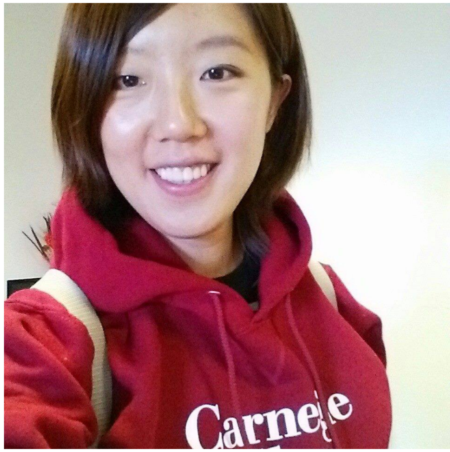
Paper Reading

Xinwei Geng

2018-07-02

Focused Hierarchical RNNs for Conditional Sequence Processing



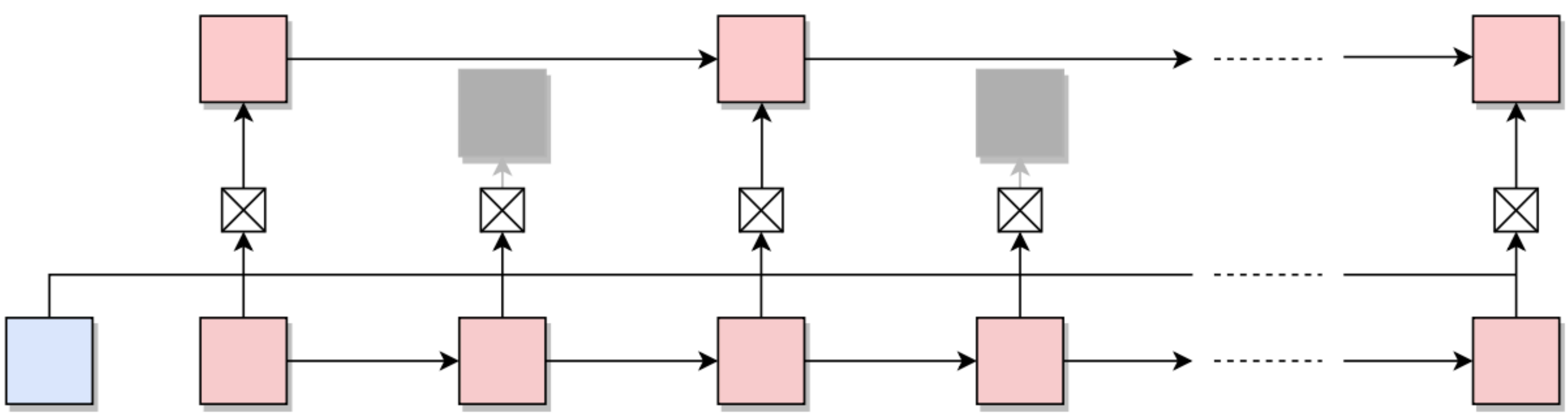
- Nan Rosemary Ke
- A second year PhD student at the Montreal Institute for Learning Algorithms MILA
- Received a Bachelors in Computer Science at the University of Auckland
- At Carnegie Mellon University working on speech recognition and deep learning
- Now at Microsoft Research, Montreal, where I work on improved RNN training, generative models and language related research.
- Interested in new ways of training Recurrent Neural Networks, generative models and causal inference learning.
- ICML 2018
- Homepage: https://nke001.github.io/

Issues

- Recurrent neural network with attention
  - the attention looks over the entire sequence and assign a soft weight to each token
  - in some case such as complex tasks, selectively process relevant information within input is crucial
- Real-word human cognitive processes
  - Reading a Wikipedia article and trying to identify information that is relevant to answering a question
  - Given a context or question, it’s much easier to read over the article, identify relevant information, group items and selectively process relevant information to answer the question
- It’s beneficial to selectively process the input sequence conditioned on a question or input context

Focused Hierarchical RNN

- Focused hierarchical encoder is modeled by a two-layer LSTM
  - The lower layer operates at the input token level
  - The upper layer focuses on tokens relevant to the context
  - A conditional boundary gate to whether it is useful to update the upper-level LSTM with lower layer hidden state



$$b_t = \sigma(\mathbf{w}_b^\top \, \mathrm{LReLU}(\mathbf{W}_b \mathbf{z}_t + \mathbf{b}_b)), \qquad \mathbf{z}_t = [\mathbf{q} \odot \mathbf{h}_t^l, \mathbf{h}_t^l, \mathbf{q}],$$

Training

- Maximize the log-likelihood of the answer (A) given the context (Q) and the passage (P)

$$\mathcal{R} = \log p(A \mid Q, P).$$

- Policy gradient
  - the discrete decisions involved in sampling the boundary variable make it impossible to use standard gradient back-propagation to learn the parameters of the boundary gate

$$\sum_{\mathbf{b}} \nabla \pi_b(\mathbf{b}) \mathcal{R}_{\mathbf{b}} = E_{\mathbf{b} \sim \pi_b} [\nabla \log \pi_b(\mathbf{b}) \mathcal{R}_{\mathbf{b}}],$$

  - Negative entropy regularization
  - Sparsity Constraints

$$\beta G(\mathbf{b}) = \beta ReLU \left((\sum_{t=1}^{T} b_t) - \gamma T \right)$$

Synthetic Experiment

- Picking task
  - A sequence of randomly generated digits of length n
  - The goal of the picking task is to determine the most frequent digit within the first k digits

Input Sequence	k	Target Mode
random examples		
80560201708283837170131630447З	10	**0**
638733290890396690255937986485	23	**3**
164551937579373896813981125982	26	**1**
malicious examples		
666333666288882888819999999990	6	**6**
666333666288882888819999999990	10	**6**
666333666288882888819999999990	20	**8**
666333666288882888819999999990	30	**9**

- Pixel-by-Pixel MNIST QA task
  - the passage encoder reads in MNIST digits one pixel at a time
  - the question asked is whether the image is a specific digit and the answer is either True or False

Experimental Results of Picking Task

Table 2. Accuracy (%) for *picking task* for LSTM1, LSTM2 and FHE-fixed. Our model and LSTM2 are on par with performing while LSTM1 is behind for longer input sequences.

LENGTH	LSTM1	LSTM2	FHE-FIXED
100	99.4	**99.7**	99.5
200	97.0	99.2	**99.4**
400	92.9	**97.5**	96.9

Table 4. Test accuracy (%) for longer sequence length for *picking task* on model trained on sequence length $n = 200$.

LENGTH	LSTM1	LSTM2	FHE-FIXED
200	97.1	99.2	**99.4**
400	55.9	61.4	**97.6**
800	39.6	39.7	**95.6**
1600	29.5	28.6	**93.3**
10000	18.5	14.8	**66.8**

Table 3. Accuracy (%) for *picking task* for the models providing a level of control over the accuracy-sparsity trade-off at a cost of slightly lower performance.

LENGTH	FHE80	FHE90	FHE95	FHE98
100	93.4	94.2	96.6	**98.7**
200	92.3	92.4	**93.6**	**93.6**
400	87.2	90.5	90.0	**91.0**

Visualization of Picking Task

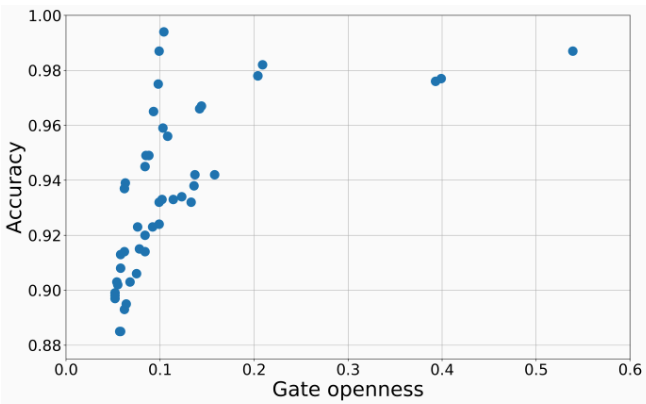


Figure 2. A relationship between accuracy and gate openness for *picking task* and sequence length $n = 100$. The best performance is achieved for gate openness around 10%.

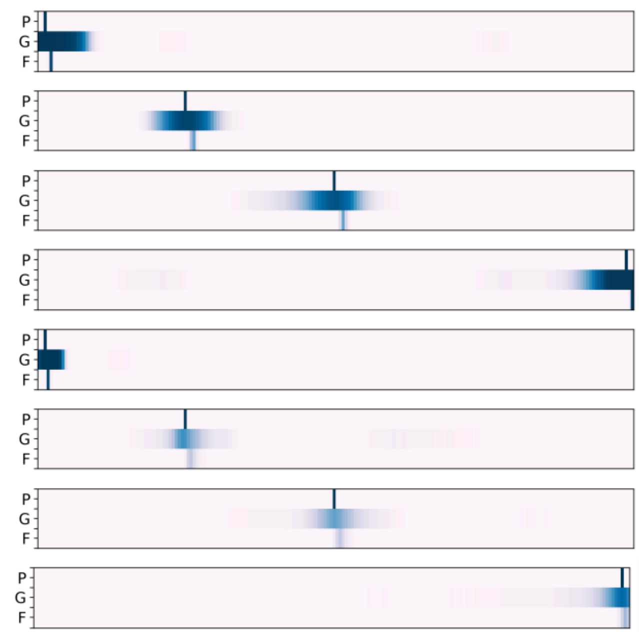


Figure 3. Gate openness (G) conditioned on the position asked (P). Focus (F) is the average of final attention weight set for a given step. Hence, focus sums to one and it is always lower than gate openness (because our model attends only over unique states). Result showed for sequence length $n = 200$. The first four plots illustrate FHE model having 99.4% accuracy and 10% gate openness, while the last four are for FHE model having 97% accuracy but 5% gate openness.

Experimental Results of Pixel-by-Pixel MNIST QA Task

LSTM1	LSTM2	FHE-FIXED
97.3	98.4	**99.1**

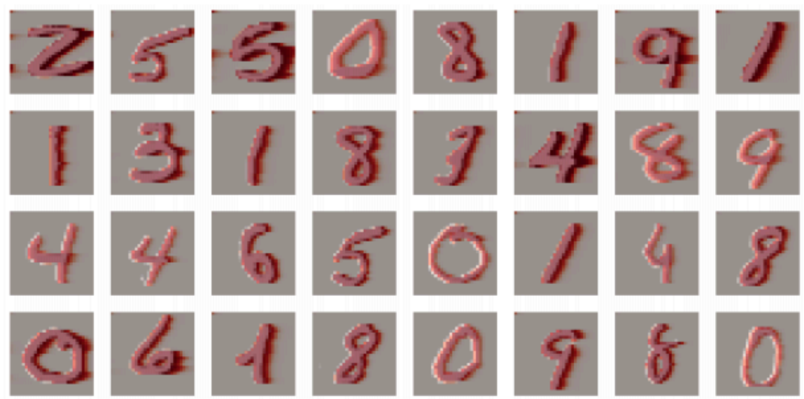


Figure 4. A visualization of the gating mechanism learned using the Pixel-by-Pixel MNIST dataset. Red pixels indicate a gate opening and are overlayed on top of the digit which is white on a gray background. The digits are vectorized row-wise which explains why white pixels appear left of the red pixels.

SearchQA Question and Answering Task

Models	Validation F1	Validation EM	Test F1	Test EM
TF-IDF Max (Dunn et al., 2017)	-	13.0	-	12.7
ASR (Dunn et al., 2017)	24.1	43.9	22.8	41.3
AQA (Buck et al., 2018)	47.7	40.5	45.6	38.7
Human (Dunn et al., 2017)	-	-	43.9	-
LSTM1 + pointer softmax	52.8	41.9	48.7	39.7
LSTM2 + pointer softmax	55.3	44.7	51.9	41.7
Our Model	56.7	**49.6**	53.4	**46.8**
Concurrent work				
AMANDA (Kundu & Ng, 2018)	**57.7**	48.6	**56.6**	**46.8**

MS MARCO Question and Answering Task

Generative Models	Validation BLEU-1	Validation ROUGE-L	Test BLEU-1	Test ROUGE-L
Seq-to-Seq (Nguyen et al., 2016)	-	8.9	-	-
Memory Network (Nguyen et al., 2016)	-	11.9	-	-
Attention Model (Higgins & Nho, 2017)	9.3	12.8	-	-
LSTM1 + pointer softmax	24.8	26.5	28	28
LSTM2 + pointer softmax	24.3	23.3	27	28
Our Model	**27.3**	**26.7**	**30**	**30**
Ablation study				
Our Model – dot-product between question and context	18.5	19.3	-	-
Our Model – pointer softmax	20.5	18.7	-	-
Our Model – learned boundaries	23.5	24	-	-

Thanks & QA