Attention-over-Attention Neural Networks for Reading Comprehension

Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu and Guoping Hu

Joint Laboratory of HIT and iFLYTEK (HFL), Beijing

2017-04-22
OUTLINE

• Introduction: Cloze-style Reading Comprehension
• Related Works
• Attention-over-Attention Reader (AoA Reader)
• Experiments & Analysis
• Conclusions & Future Works
Preface

• This work was first made publicly available on June, 2016 @arXiv

• Citations: 18 (based on Google Scholar)

• Propose a novel mechanism called “Attention-over-Attention”
Introduction

• Recently, *Reading Comprehension* has become enormously popular in the community

• A main obstacle in the RC research

• NO MUCH DATA!

• The related works are often started from providing the relevant corpus, and then proposing some technical insights in solving them
INTRODUCTION

• Key points in RC

• Document

• Query

• Candidates

• Answer

*Example is chosen from the MCTest dataset (Richardson et al., 2013)
INTRODUCTION

• Key points in RC
• Document
• → Query
• Candidates
• Answer

James the Turtle was always getting in trouble. Sometimes he'd reach into the freezer and empty out all the food. Other times he'd sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back.

One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn't pay, and instead headed home.

His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.

After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

1) What is the name of the trouble making turtle?
   A) Fries
   B) Pudding
   C) James
   D) Jane

*Example is chosen from the MCTest dataset (Richardson et al., 2013)
INTRODUCTION

• Key points in RC

• Document

• Query

• → Candidates

• Answer

*Example is chosen from the MCTest dataset (Richardson et al., 2013)
INTRODUCTION

• Key points in RC
• Document
• Query
• Candidates

→ Answer

James the Turtle was always getting in trouble. Sometimes he’d reach into the freezer and empty out all the food. Other times he’d sled on the deck and get a splinter. His aunt Jane tried as hard as she could to keep him out of trouble, but he was sneaky and got into lots of trouble behind her back.

One day, James thought he would go into town and see what kind of trouble he could get into. He went to the grocery store and pulled all the pudding off the shelves and ate two jars. Then he walked to the fast food restaurant and ordered 15 bags of fries. He didn’t pay, and instead headed home.

His aunt was waiting for him in his room. She told James that she loved him, but he would have to start acting like a well-behaved turtle.

After about a month, and after getting into lots of trouble, James finally made up his mind to be a better turtle.

1) What is the name of the trouble making turtle?
A) Fries
B) Pudding
C) James
D) Jane

*Example is chosen from the MCTest dataset (Richardson et al., 2013)
INTRODUCTION

- Specifically, in cloze-style RC
  - Document: the same as the general RC
  - Query: a sentence with a blank
  - Candidate (optional): several candidates to fill in
  - Answer: a single word that exactly match the query (the answer word should appear in the document)

*Example is chosen from the CNN dataset (Hermann et al., 2015)
INTRODUCTION

• CBT dataset (Hill et al., 2015)

Step 1: Choose 21 sentences

Step 2: Choose first 20 sentences as Context

Step 3: Choose 21st sentence as Query

Step 4: Choose other 9 similar words from Context as Candidate

Step 3: With a BLANK

Step 3: The word removed from Query
RELATED WORKS

- Attentive Reader (Hermann et al., NIPS2015)
- Attention Sum Reader (Kadlec et al., ACL2016)
- Consensus Attention Reader (Cui et al., COLING2016)
- Gated-attention Reader (Dhingra et al., ICLR2017)
- ...
ATTENTIVE READER

- Teaching Machines to Read and Comprehend (Hermann et al., 2015)

\[
m(t) = \tanh(W_{ym}y_d(t) + W_{um}u),
\]
\[
s(t) \propto \exp(w_{ms}^T m(t)),
\]
\[
r = y_ds,
\]
\[
g^{AR}(d, q) = \tanh(W_{rg}r + W_{ug}u).
\]
**ATTENTION SUM READER**

- Text Understanding with the Attention Sum Reader Network (Kadlec et al., 2016)

<table>
<thead>
<tr>
<th>Document</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input text</td>
<td>XXXXX visited Prague</td>
</tr>
<tr>
<td>Embeddings</td>
<td>e(Obama) e(and) e(Putin) e(said) e(Obama) e(in) e(Prague)</td>
</tr>
<tr>
<td>Recurrent neural networks</td>
<td>f, g</td>
</tr>
<tr>
<td>Dot products</td>
<td></td>
</tr>
<tr>
<td>Softmax $s_i$ over words in the sentence</td>
<td></td>
</tr>
<tr>
<td>Probability of the answer</td>
<td>$P(Obama</td>
</tr>
</tbody>
</table>
Consensus Attention Reader

- Consensus Attention-based Neural Networks for Chinese Reading Comprehension (Cui et al., 2016)
GATED-ATTENTION READER

- Gated-Attention Reader for Text Comprehension (Dhingra et al., 2017)
AoA Reader

• Motivated by
  
  • AS Reader (Kadlec et al., ACL2016)
  
  • CAS Reader (Cui et al., COLING2016)
  
• Some of the components in AoA Reader has been widely adopted in the follow-up works (see the papers cited)


**AoA Reader**

- **Contextual Embedding**

- Transform document and query into contextual representations using GRU

\[
e(x) = W_e \cdot x, \text{ where } x \in \mathcal{D}, \mathcal{Q} \quad (1)
\]

\[
\hat{h}_s(x) = GRU(e(x)) \quad (2)
\]

\[
\hat{h}_s(x) = GRU(e(x)) \quad (3)
\]

\[
h_s(x) = [\hat{h}_s(x); \hat{h}_s(x)] \quad (4)
\]
AoA Reader

• Pair-wise Matching Score

• Calculate ‘similarity’ between each document word and query word

\[ M(i, j) = h_{doc}(i)^T \cdot h_{query}(j) \]  

\( P("Mary"|D,Q) = \sum_{i \in \text{Word("Mary",D)}} s_i = s_f + s_k \)
AoA Reader

• Individual Attentions

• Calculate attention with respect to each query word

\[
\alpha(t) = \text{softmax}(M(1, t), \ldots, M(|D|, t)) \quad (6)
\]

\[
\alpha = [\alpha(1), \alpha(2), \ldots, \alpha(|Q|)] \quad (7)
\]
**AoA Reader**

- **Attention-over-Attention**
  - Dynamically assign weights to individual attentions

\[
\beta(t) = \text{softmax}(M(t, 1), \ldots, M(t, |Q|)) \tag{8}
\]
\[
\beta = \frac{1}{n} \sum_{t=1}^{\left|\mathcal{D}\right|} \beta(t) \tag{9}
\]
\[
s = \alpha^T \beta \tag{10}
\]
AoA Reader

- Final Predictions

- Apply sum-attention mechanism (Kadlec et al., 2016) to get the final probability of the answer

\[
P(w|D, Q) = \sum_{i \in I(w, D)} s_i, \ w \in V 
\] (11)

\[
\mathcal{L} = \sum_{i} \log(p(x)) , x \in A 
\] (12)
**Experiments**

- **Dataset**
  - CNN (Hermann et al., 2015) and CBT-NE/CN (Hill et al., 2015)

- **Parameters**
  - Embedding: uniform distribution \([-0.05, 0.05]\) with l2-regularization, dropout 0.1
  - Hidden Layer: bi-GRU
  - Optimization: Adam(l\(r\)=0.001), gradient clipping 5, batch 32

- **Framework**: Theano + Keras
**Experimental Results**

- Single model performance

<table>
<thead>
<tr>
<th>Model</th>
<th>CNN Valid</th>
<th>CNN Test</th>
<th>CBTest NE Valid</th>
<th>CBTest NE Test</th>
<th>CBTest CN Valid</th>
<th>CBTest CN Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep LSTM Reader (Hermann et al., 2015)</td>
<td>55.0</td>
<td>57.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Attentive Reader (Hermann et al., 2015)</td>
<td>61.6</td>
<td>63.0</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Human (context+query) (Hill et al., 2015)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>81.6</td>
<td>81.6</td>
</tr>
<tr>
<td>MemNN (window + self-sup.) (Hill et al., 2015)</td>
<td>63.4</td>
<td>66.8</td>
<td>70.4</td>
<td>66.6</td>
<td>64.2</td>
<td>63.0</td>
</tr>
<tr>
<td>AS Reader (Kadlec et al., 2016)</td>
<td>68.6</td>
<td>69.5</td>
<td>73.8</td>
<td>68.6</td>
<td>68.8</td>
<td>63.4</td>
</tr>
<tr>
<td>CAS Reader (Cui et al., 2016)</td>
<td>68.2</td>
<td>70.0</td>
<td>74.2</td>
<td>69.2</td>
<td>68.2</td>
<td>65.7</td>
</tr>
<tr>
<td>Stanford AR (Chen et al., 2016)</td>
<td>72.4</td>
<td>72.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GA Reader (Dhingra et al., 2016)</td>
<td>73.0</td>
<td>73.8</td>
<td>74.9</td>
<td>69.0</td>
<td>69.0</td>
<td>63.9</td>
</tr>
<tr>
<td>Iterative Attention (Sordoni et al., 2016)</td>
<td>72.6</td>
<td>73.3</td>
<td>75.2</td>
<td>68.6</td>
<td>72.1</td>
<td>69.2</td>
</tr>
<tr>
<td>EpiReader (Trischler et al., 2016)</td>
<td>73.4</td>
<td>74.0</td>
<td>75.3</td>
<td>69.7</td>
<td>71.5</td>
<td>67.4</td>
</tr>
<tr>
<td>AoA Reader</td>
<td>73.1</td>
<td>74.4</td>
<td>77.8</td>
<td>72.0</td>
<td>72.2</td>
<td>69.4</td>
</tr>
</tbody>
</table>
## Experimental Results

- Ensemble performance

- We use 4-model greedy ensemble approach

<table>
<thead>
<tr>
<th>Model</th>
<th>CNN Valid</th>
<th>CNN Test</th>
<th>CBT Test NE Valid</th>
<th>CBT Test NE Test</th>
<th>CBT Test CN Valid</th>
<th>CBT Test CN Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>MemNN (Ensemble)</td>
<td>66.2</td>
<td>69.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>AS Reader (Ensemble)</td>
<td>73.9</td>
<td>75.4</td>
<td>74.5</td>
<td>70.6</td>
<td>71.1</td>
<td>68.9</td>
</tr>
<tr>
<td>GA Reader (Ensemble)</td>
<td>76.4</td>
<td>77.4</td>
<td>75.5</td>
<td>71.9</td>
<td>72.1</td>
<td>69.4</td>
</tr>
<tr>
<td>EpiReader (Ensemble)</td>
<td>-</td>
<td>-</td>
<td>76.6</td>
<td>71.8</td>
<td>73.6</td>
<td>70.6</td>
</tr>
<tr>
<td>Iterative Attention (Ensemble)</td>
<td>74.5</td>
<td>75.7</td>
<td>76.9</td>
<td>72.0</td>
<td>74.1</td>
<td>71.0</td>
</tr>
<tr>
<td>AoA Reader (Ensemble)</td>
<td>-</td>
<td>-</td>
<td>78.9</td>
<td>74.5</td>
<td>74.7</td>
<td>70.8</td>
</tr>
</tbody>
</table>
Analysis

- Accuracy vs. Length of Document
  - AoA Reader shows consistent improvements over AS Reader on different length of document
  - The improvements become larger when the length of document increases
**Analysis**

- Accuracy v.s. Frequency of answer
  - Most of the answers are the top frequent word among candidates
  - Tend to choose either high or low frequency word
CONCLUSIONS

• Propose a novel mechanism called “Attention-over-Attention” to dynamically calculate weights between individual attentions

• Adopting both doc-to-query and query-to-doc attentions for final predictions

• Experimental results show significant improvements over various systems

• Future Works

  • Investigate more complex attention mechanism

  • Look into the problems that need comprehensive reasoning over several sentences
INTERACTIVE AoA Reader

- As a step further of our work, we upgrade our model as ‘interactive’
- Shows good performance on Stanford SQuAD RC Task
- No.2 in single model ranking
CCL-CMRC2017

- The 1st Evaluation Workshop on Chinese Machine Reading Comprehension (CMRC2017)

- Hosted by CIPS, organized by Joint Laboratory of HIT and iFLYTEK (HFL)

- Co-located with CCL2017(2017.10.13 ~ 2017.10.15, Nanjing)

- Welcome to join us!

<table>
<thead>
<tr>
<th>事件</th>
<th>时间</th>
</tr>
</thead>
<tbody>
<tr>
<td>预报名</td>
<td>2017年4月5日 ~ 2017年4月17日</td>
</tr>
<tr>
<td>正式报名</td>
<td>2017年4月18日 ~ 2017年4月25日</td>
</tr>
<tr>
<td>发布训练集和开发集</td>
<td>2017年5月3日</td>
</tr>
<tr>
<td>系统搭建及调整</td>
<td>2017年5月3日 ~ 2017年7月31日</td>
</tr>
<tr>
<td>提交系统验证开发集</td>
<td>2017年7月1日 ~ 2017年7月31日</td>
</tr>
<tr>
<td>提交系统验证测试集</td>
<td>2017年8月1日 ~ 2017年8月3日</td>
</tr>
<tr>
<td>撰写系统描述</td>
<td>2017年8月18日</td>
</tr>
<tr>
<td>召开CMRC2017大会</td>
<td>2017年10月13日</td>
</tr>
</tbody>
</table>
MORE INFORMATION

• Paper download (through arXiv)
  • https://arxiv.org/abs/1607.04423

• General training tips & Leaderboard of Cloze-style RC (updates irregularly)
  • https://github.com/ymcui/Eval-on-NN-of-RC

• Personal Website (slides and new paper will be uploaded soon)
  • http://ymcui.github.io
REFERENCES


REFERENCES


REFERENCES


REFERENCES


THANK YOU!

AND TIME TO REVISE CAMERA-READY PAPER…

CONTACT: admin [at] ymcui [dot] com