FusionNet: Fusing Complete Information via Fully-aware Attention in Machine Reading Comprehension

Feb. 6, 2018
Chenguang Zhu
Microsoft AI+Research
Turing Test, 1950
Answer:
Madame Curie was born on November 7, 1867.

NLP platform people want to build

Question:
When was Madame Curie born?
Major challenge

- Accurately answering *tricky* questions
  - Not of the factoid type (e.g., Marie Curie birthday)
  - Rather questions requiring comprehension and reasoning
    - Go beyond Alexa, Google Home or Cortana type of platforms
Private school
The Stanford Question Answering Dataset

Private schools, also known as independent schools, non-governmental, or nonstate schools, are not administered by local, state or national governments; thus, they retain the right to select their students and are funded in whole or in part by charging their students tuition, rather than relying on mandatory taxation through public (government) funding; at some private schools students may be able to get a scholarship, which makes the cost cheaper, depending on a talent the student may have (e.g. sport scholarship, art scholarship, academic scholarship), financial need, or tax credit scholarships that might be available.

Along with non-governmental and nonstate schools, what is another name for private schools?
Ground Truth Answers: independent, independent schools, independent schools

Along with sport and art, what is a type of talent scholarship?
Ground Truth Answers: academic, academic, academic

Rather than taxation, what are private schools largely funded by?
Ground Truth Answers: tuition, charging their students tuition, tuition
Machine Reading Comprehension

- Context Integration
  - Embed words into high-dimension vectors
  - Get high-level understanding of a word using its context
    “The meaning of a word is obtained from the company it keeps.”
    --- Robert Firth, 1957

- Attention
  - Discover relationship between context and question words
Embed words

- refrigerator
- oven
- microwave
- kitchen
- sink
- bathroom
- toilet
- bathtub
- faucet
- table
- fan
- light
- led
- bulb
- charger
- battery
- saw
- tool
- drill
- garden
- hose
- sprinkler
- color
- paint
Context Integration
Attention

- Obtain an understanding of word from one text (context) based on the information from the other text (question)

Forming attention weight over a set of objects

Global Fusion
Attention in Machine Translation

I am Chenguang Zhu

我是朱晨光
Attention in MRC

Reweighted context: \( \{v_1, \ldots, v_i, \ldots, v_m\} \)

Typical Usage:

Context: \( \{c_1, \ldots, c_i, \ldots, c_m\} \)  
Question: \( \{q_1, \ldots, q_j, \ldots q_n\} \)

\[ S_{ij} = f(c_i, q_j), \quad 1 \leq i \leq m, \quad 1 \leq j \leq n \quad : \text{Attention Score} \]

\[ \alpha_{ij} = \frac{e^{S_{ij}}}{\sum_j e^{S_{ij}}} \quad : \text{Attention Weight} \]

\[ v_i = \sum_j \alpha_{ij} q_j \quad : \text{Reweighted context vector from question} \]
## Existing Architectures for Reweighting

<table>
<thead>
<tr>
<th>Architectures</th>
<th>(1)</th>
<th>(2)</th>
<th>(2’)</th>
<th>(3)</th>
<th>(3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Match-LSTM (Wang and Jiang, 2016)</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>DCN (Xiong et al., 2017)</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>FastQA (Weissenborn et al., 2017)</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>FastQAExt (Weissenborn et al., 2017)</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>BiDAF (Seo et al., 2017)</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>RaSoR (Lee et al., 2016)</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>DrQA (Chen et al., 2017)</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>MPCM (Wang et al., 2016)</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Mnemonic Reader (Hu et al., 2017)</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>R-net (Wang et al., 2017b)</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Problem: either ignore low-level or high-level or middle-level representations

Reason: not enough space or time
Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.

What was the name of a quarterback whose number was 38 in Super Bowl?
Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.

When performing attention in previous architectures:

- (High-level) Name of an athlete with some number
- (High-level) What name of athlete with number
- (High-level) Some number in a competition
- (High-level) Number on a cloth in a competition

Simplified Attention Score Matrix

What was the name of a quarterback whose number was 38 in Super Bowl?
Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.

When performing attention in previous architectures:

- (High-level) Name of an athlete with some number
- (High-level) What name of athlete with number
- (High-level) Some number in a competition
- (High-level) Number on a cloth in a competition

What was the name of a quarterback whose number was 38 in Super Bowl?
Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.

When performing attention in previous architectures

(High-level) Name of an athlete with some number

(High-level) What name of athlete with number

(High-level) Some number in a competition

(High-level) Number on a cloth in a competition

What was the name of a quarterback whose number was 38 in Super Bowl?
History of Word

Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.

When performing attention in previous architectures

(High-level) Name of an athlete with some number
(High-level) What name of athlete with number
(High-level) Some number in a competition

(High-level) Number on a cloth in a competition

Perfect Match!

What was the name of a quarterback whose number was 38 in Super Bowl?
Attention Scoring Function

Additive form:
\[ S_{ij} = s^T \tanh(W_1 c_i + W_2 q_j) \]

Space: \( O(mnk) \), \( W \) is \( kxd \)

Product form:
\[ S_{ij} = c_i^T W q_j \]
\[ S_{ij} = c_i^T U^T V q_j \]
\[ S_{ij} = c_i^T W^T D W q_j \]

Space: \( O((m+n)k) \)

1. Smaller space
2. Non-linearity

\[ S_{ij} = \text{Relu}(c_i^T W^T) D \text{Relu}(W q_j) \]
“Long” Scores, Short Combination

7-Level History-of-Word Attention
Multi-level Attention

When Context is long, Self-boosted Fusion can be used

Multi-level Fusion

- Understanding
- High-level Concept
- Low-level Concept
- Input Word

Word-level Fusion

Context

Question

BiLSTM

3-Level History-of-Word Attention
FusionNet

Fully-aware Fusion Network

Fully-aware Self-boosted Fusion

Context Understanding

Question Understanding

When Context is long, Self-boosted Fusion can be used

Multi-level Fusion

Understanding

High-level Concept

Low-level Concept

Input Word

Context

Word-level Fusion

Question

BiLSTM

7-Level History-of-Word Attention

BiLSTM

3-Level History-of-Word Attention
Answer Generation

\[ v = \sum_i \alpha_i v_i, \alpha_i \propto \exp(q^T v_i) \]

Context Understanding

\[ P_S^i \propto \exp(v^T W_S u_i) \]
\[ P_E^i \propto \exp(v^T W_T u_i) \]

Question Understanding
SQuAD: Stanford Question Answering Dataset
Evaluation Criteria

Evaluation Criteria: **Exact Match (EM)** & **F1 score (F1)**

What impact does workers working harder have on productivity of a business?

*Ground Truth Answers:* less workers are required raises the productivity of each worker,

In Marxist analysis, capitalist firms increasingly substitute capital equipment for labor inputs (workers) under competitive pressure to reduce costs and maximize profits. Over the long-term, this trend increases the organic composition of capital, meaning that less workers are required in proportion to capital inputs, increasing unemployment (the "reserve army of labour"). This process exerts a downward pressure on wages. The substitution of capital equipment for labor (mechanization and automation) raises the productivity of each worker, resulting in a situation of relatively stagnant wages for the working class amidst rising levels of property income for the capitalist class.
Our Result on SQuAD

Leaderboard

Since the release of our dataset, the community has made rapid progress! Here are the ExactMatch (EM) and F1 scores of the best models evaluated on the test and development sets of v1.1. Will your model outperform humans on the QA task?

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AIR-FusionNet (ensemble)</td>
<td>78.842</td>
<td>85.936</td>
</tr>
<tr>
<td></td>
<td>Microsoft Business AI Solutions Team</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>DCN+ (ensemble)</td>
<td>78.706</td>
<td>85.619</td>
</tr>
<tr>
<td></td>
<td>Salesforce Research</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Interactive AoA Reader (ensemble)</td>
<td>77.845</td>
<td>85.297</td>
</tr>
<tr>
<td></td>
<td>Joint Laboratory of HIT and iFLYTEK Research</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>r-net (ensemble)</td>
<td>78.244</td>
<td>85.206</td>
</tr>
<tr>
<td></td>
<td>Microsoft Research Asia</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><a href="http://aka.ms/rnet">http://aka.ms/rnet</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Reinforced Mnemonic Reader (ensemble)</td>
<td>77.678</td>
<td>84.888</td>
</tr>
<tr>
<td></td>
<td>NUDT and Fudan University</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>AIR-FusionNet (single model)</td>
<td>75.968</td>
<td>83.900</td>
</tr>
<tr>
<td></td>
<td>Microsoft Business AI Solutions team</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>r-net (single model)</td>
<td>75.705</td>
<td>83.496</td>
</tr>
<tr>
<td></td>
<td>Microsoft Research Asia</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><a href="http://aka.ms/rnet">http://aka.ms/rnet</a></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Experiments

<table>
<thead>
<tr>
<th>Single Model</th>
<th>Test Set</th>
<th>AddSent</th>
<th>EM / F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR Baseline (Rajpurkar et al., 2016)</td>
<td>40.4 / 51.0</td>
<td>LR Baseline</td>
<td>17.0 / 23.2</td>
</tr>
<tr>
<td>Match-LSTM (Wang and Jiang, 2016)</td>
<td>64.7 / 73.7</td>
<td>Match-LSTM (E)</td>
<td>24.3 / 34.2</td>
</tr>
<tr>
<td>BiDAF (Seo et al., 2017)</td>
<td>68.0 / 77.3</td>
<td>BiDAF (E)</td>
<td>29.6 / 34.2</td>
</tr>
<tr>
<td>SEDT (Liu et al., 2017)</td>
<td>68.2 / 77.5</td>
<td>SEDT (E)</td>
<td>30.0 / 35.0</td>
</tr>
<tr>
<td>RaSoR (Lee et al., 2016)</td>
<td>70.8 / 78.7</td>
<td>Mnemonic Reader (E)</td>
<td>40.7 / 46.2</td>
</tr>
<tr>
<td>DrQA (Chen et al., 2017)</td>
<td>70.7 / 79.4</td>
<td>ReasoNet (E)</td>
<td>34.6 / 39.4</td>
</tr>
<tr>
<td>ReasoNet (Shen et al., 2017)</td>
<td>70.6 / 79.4</td>
<td>DCN+ (S)</td>
<td>39.4 / 44.5</td>
</tr>
<tr>
<td>R. Mnemonic Reader (Hu et al., 2017)</td>
<td>73.2 / 81.8</td>
<td>FusionNet (E)</td>
<td>46.2 / 51.4</td>
</tr>
<tr>
<td>DCN+</td>
<td>74.9 / 82.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R-net (Wang et al., 2017b)</td>
<td>75.7 / 83.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>FusionNet</strong></td>
<td><strong>76.0 / 83.9</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ensemble Model</th>
<th>Test Set</th>
<th>AddOneSent</th>
<th>EM / F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ReasoNet (Shen et al., 2017)</td>
<td>75.0 / 82.3</td>
<td>LR Baseline</td>
<td>22.3 / 30.4</td>
</tr>
<tr>
<td>MEMEN (Pan et al., 2017)</td>
<td>75.4 / 82.7</td>
<td>Match-LSTM (E)</td>
<td>34.8 / 41.8</td>
</tr>
<tr>
<td>R. Mnemonic Reader (Hu et al., 2017)</td>
<td>77.7 / 84.9</td>
<td>BiDAF (E)</td>
<td>40.7 / 46.9</td>
</tr>
<tr>
<td>R-net (Wang et al., 2017b)</td>
<td>78.2 / 85.2</td>
<td>SEDT (E)</td>
<td>40.0 / 46.5</td>
</tr>
<tr>
<td>DCN+</td>
<td>78.7 / 85.6</td>
<td>Mnemonic Reader (E)</td>
<td>48.7 / 55.3</td>
</tr>
<tr>
<td><strong>FusionNet</strong></td>
<td><strong>78.8 / 85.9</strong></td>
<td>ReasoNet (E)</td>
<td>43.6 / 49.8</td>
</tr>
<tr>
<td>Human (Rajpurkar et al., 2016)</td>
<td>82.3 / 91.2</td>
<td>DCN+ (S)</td>
<td>47.4 / 54.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>FusionNet (E)</strong></td>
<td><strong>54.7 / 60.7</strong></td>
</tr>
</tbody>
</table>

Table 2: The performance of FusionNet and competing approaches on SQuAD hidden test set.

Table 3: Comparison on AddSent. (S: Single model, E: Ensemble)

Table 4: Comparison on AddOneSent. (S: Single model, E: Ensemble)
Conclusions

- Machine reading comprehension
- FusionNet
  - Introduce the “History-of-Word” concept
  - Create a new form of attention scoring function
  - Multi-level attention
Paper

- FusionNet: Fusing via Fully-Aware Attention with Application to Machine Comprehension


- To appear in ICLR 2018

- Joint work With Hsin-Yuan Huang, Yelong Shen and Weizhu Chen
Thank you!

- Any questions?