Dynamic Coattention Networks for Question Answering

Caiming Xiong    Victor Zhong    Richard Socher

Salesforce Research

ICLR, 2017
Presenter: Tianlu Wang
Outline

1 Introduction
   - Question Answering
   - Related Work

2 Model
   - Overview
   - Document and Question Encoder
   - Coattention Encoder
   - Dynamic Pointing Decoder

3 Experiments
   - Qualitative Examples
   - Results

4 Summary
Outline

1 Introduction
   • Question Answering
   • Related Work

2 Model
   • Overview
   • Document and Question Encoder
   • Coattention Encoder
   • Dynamic Pointing Decoder

3 Experiments
   • Qualitative Examples
   • Results

4 Summary
Human annotated high quality but small dataset

Large scale dataset through semi-annotated techniques but far from natural language

Stanford Question Answering dataset (SQuAD)
  - Larger than all previous hand-annotated datasets
  - Various qualities
  - Answers are spans in a reference document
Outline

1 Introduction
   - Question Answering
   - Related Work

2 Model
   - Overview
   - Document and Question Encoder
   - Coattention Encoder
   - Dynamic Pointing Decoder

3 Experiments
   - Qualitative Examples
   - Results

4 Summary
Related Work

- **Statistical QA**
  - Rule-based algorithms
  - Linear classifiers over feature sets: lexical features (bag of words), word distance, word order, pos_tag, dependency parse

- **Neural QA**
  - NLI (natural language inference): match LSTM encoder + pointer network decoder,
  - dynamic chunk reader: extract answer candidates and rank
  - hierarchical co-attention model
Q: what is the man holding a snowboard on top of a snow covered? A: **mountain**

Q: what is the color of the bird? A: **white**

Q: how many snowboarders in formation in the snow. four is sitting? A: **5**
Outline

1. Introduction
   - Question Answering
   - Related Work

2. Model
   - Overview
   - Document and Question Encoder
   - Coattention Encoder
   - Dynamic Pointing Decoder

3. Experiments
   - Qualitative Examples
   - Results

4. Summary
Overview

End-to-end neural network for question answering:

- A coattention encoder captures the interaction between the question and the document
- A dynamic pointing decoder alternates between estimating the start and end of the answer span
Overview

End-to-end neural network for question answering:

- A coattention encoder captures the interaction between the question and the document
- A dynamic pointing decoder alternates between estimating the start and end of the answer span

![Diagram of the Dynamic Coattention Network]

Figure 1: Overview of the Dynamic Coattention Network.
Outline

1 Introduction
   - Question Answering
   - Related Work

2 Model
   - Overview
   - Document and Question Encoder
     - Coattention Encoder
     - Dynamic Pointing Decoder

3 Experiments
   - Qualitative Examples
   - Results

4 Summary
Document and Question Encoder

- Sequence of word vectors in document:
  \((x_1^D, x_2^D, \ldots, x_n^D)\)
  \(\Rightarrow d_t = LSTM_{enc}(d_{t-1}, x_t^D)\)
  \(\Rightarrow D = [d_1 \ldots d_m d_\phi] \in \mathbb{R}^{l \times (m+1)}\)

- Sequence of word vectors in document:
  \((x_1^Q, x_2^Q, \ldots, x_m^Q)\)
  \(\Rightarrow q_t = LSTM_{enc}(q_{t-1}, x_t^Q)\)
  \(\Rightarrow Q' = [q_1 \ldots q_n q_\phi] \in \mathbb{R}^{l \times (n+1)}\)
  \(\Rightarrow Q = tanh(W^{(Q)}Q' + b^{(Q)}) \in \mathbb{R}^{l \times (n+1)}\)
  (allow for variation between question encoding space and document encoding space)

- \(d_\phi\) and \(q_\phi\): sentinel vector
Outline

1 Introduction
   • Question Answering
   • Related Work

2 Model
   • Overview
   • Document and Question Encoder
   • Coattention Encoder
   • Dynamic Pointing Decoder

3 Experiments
   • Qualitative Examples
   • Results

4 Summary
Coattention Encoder

Figure 2: Coattention encoder. The affinity matrix $L$ is not shown here. We instead directly show the normalized attention weights $A^D$ and $A^Q$. 

---

**Coattention Encoder**

- **D:**
  - $\ell$
  - $m+1$ documents

- **Q:**
  - $\ell$
  - $n+1$

- **U:**
  - $u_t = \text{Bi-LSTM}(u_{t-1}, u_{t+1}, [d_t; c_t^D], \text{size}=2I)$

- **$A^D$**
  - $\text{softmax}(L, (m+1)\times(n+1))$
  - Affinity matrix $L = D^\top Q$, $(m+1)\times(n+1)$

- **$A^Q$**
  - $\text{softmax}(L^\top, (n+1)\times(m+1))$

- **$C^Q$**
  - $[Q; C^Q]A^D$, 2lx(m+1), ??

- **$C^D$**
  - $\text{concat}$

- **Attention context of the document in light of each word, $DA^Q$, lx(n+1)**
Outline

1. Introduction
   - Question Answering
   - Related Work

2. Model
   - Overview
   - Document and Question Encoder
   - Coattention Encoder
   - Dynamic Pointing Decoder

3. Experiments
   - Qualitative Examples
   - Results

4. Summary
Dynamic Pointing Decoder

\[ h_i = \text{LSTM}_{dec}(h_{i-1}, [u_{s_{i-1}}; u_{e_{i-1}}]), \quad U = [u_1, \ldots, u_m] \in \mathbb{R}^{2l \times m} \text{ from encoder} \]

Figure 3: Dynamic Decoder. Blue denotes the variables and functions related to estimating the start position whereas red denotes the variables and functions related to estimating the end position.
\[ h_i = LSTM_{dec}(h_{i-1}, [u_{s_{i-1}}, u_{e_{i-1}}]) \]

Given current hidden state \( h_i \), previous start position \( u_{s_{i-1}} \) and previous end position \( u_{e_{i-1}} \), how to estimate the current start position \( s_i \) and current end position \( e_i \)?

\[ s_i = \arg\max(\alpha_1, \ldots, \alpha_m) \]

\[ \alpha_t = HMN_{start}(u_t, h_i, u_{s_{i-1}}, u_{e_{i-1}}) \]
Highway Maxout Network (HMN)

\[
HMN(u_t, h_i, u_{s_{i-1}}, u_{e_{i-1}}) = \max(W^{(3)}[m_t^{(1)}; m_t^{(2)}] + b^{(3)})
\] (1)

\[
r = \tanh(W^{(D)}[h_i; u_{s_{i-1}}; u_{e_{i-1}}])
\] (2)

\[
m_t^{(1)} = \max(W^{(1)}[u_t; r] + b^{(1)})
\] (3)

\[
m_t^{(2)} = \max(W^{(2)}m_t^{(1)} + b^{(2)})
\] (4)

Figure 4: Highway Maxout Network. Dotted lines denote highway connections.
Outline

1 Introduction
   - Question Answering
   - Related Work

2 Model
   - Overview
   - Document and Question Encoder
   - Coattention Encoder
   - Dynamic Pointing Decoder

3 Experiments
   - Qualitative Examples
   - Results

4 Summary
Qualitative Examples

**Question 1:** Who recovered Tolbert's fumble?

*Answer:* Danny Trevathan  
*Groundtruth:* Danny Trevathan

**Question 2:** What did the Kenyan business people hope for when meeting with the Chinese?

*Answer:* gain support from China for a planned $2.5 billion railway  
*Groundtruth:* support from China for a planned $2.5 billion railway
### Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Dev EM</th>
<th>Dev F1</th>
<th>Test EM</th>
<th>Test F1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ensemble</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DCN (Ours)</td>
<td>70.3</td>
<td>79.4</td>
<td>71.2</td>
<td>80.4</td>
</tr>
<tr>
<td>Microsoft Research Asia *</td>
<td>—</td>
<td>—</td>
<td>69.4</td>
<td>78.3</td>
</tr>
<tr>
<td>Allen Institute *</td>
<td>69.2</td>
<td>77.8</td>
<td>69.9</td>
<td>78.1</td>
</tr>
<tr>
<td>Singapore Management University *</td>
<td>67.6</td>
<td>76.8</td>
<td>67.9</td>
<td>77.0</td>
</tr>
<tr>
<td>Google NYC *</td>
<td>68.2</td>
<td>76.7</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td><strong>Single model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DCN (Ours)</td>
<td>65.4</td>
<td>75.6</td>
<td>66.2</td>
<td>75.9</td>
</tr>
<tr>
<td>Microsoft Research Asia *</td>
<td>65.9</td>
<td>75.2</td>
<td>65.5</td>
<td>75.0</td>
</tr>
<tr>
<td>Google NYC *</td>
<td>66.4</td>
<td>74.9</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Singapore Management University *</td>
<td>—</td>
<td>—</td>
<td>64.7</td>
<td>73.7</td>
</tr>
<tr>
<td>Carnegie Mellon University *</td>
<td>—</td>
<td>—</td>
<td>62.5</td>
<td>73.3</td>
</tr>
<tr>
<td>Dynamic Chunk Reader (Yu et al., 2016)</td>
<td>62.5</td>
<td>71.2</td>
<td>62.5</td>
<td>71.0</td>
</tr>
<tr>
<td>Match-LSTM (Wang &amp; Jiang, 2016b)</td>
<td>59.1</td>
<td>70.0</td>
<td>59.5</td>
<td>70.3</td>
</tr>
<tr>
<td>Baseline (Rajpurkar et al., 2016)</td>
<td>40.0</td>
<td>51.0</td>
<td>40.4</td>
<td>51.0</td>
</tr>
<tr>
<td>Human (Rajpurkar et al., 2016)</td>
<td>81.4</td>
<td>91.0</td>
<td>82.3</td>
<td>91.2</td>
</tr>
</tbody>
</table>

Table 1: Leaderboard performance at the time of writing (Nov 4 2016). * indicates that the model used for submission is unpublished. — indicates that the development scores were not publicly available at the time of writing.
## Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Dev EM</th>
<th>Dev F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dynamic Coattention Network (DCN)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pool size 16 HMN</td>
<td>65.4</td>
<td>75.6</td>
</tr>
<tr>
<td>pool size 8 HMN</td>
<td>64.4</td>
<td>74.9</td>
</tr>
<tr>
<td>pool size 4 HMN</td>
<td>65.2</td>
<td>75.2</td>
</tr>
<tr>
<td>DCN with 2-layer MLP instead of HMN</td>
<td>63.8</td>
<td>74.4</td>
</tr>
<tr>
<td>DCN with single iteration decoder</td>
<td>63.7</td>
<td>74.0</td>
</tr>
<tr>
<td>DCN with Wang &amp; Jiang (2016b) attention</td>
<td>63.7</td>
<td>73.7</td>
</tr>
</tbody>
</table>

Table 2: Single model ablations on the development set.
An end-to-end neural network architecture for question answering
On the SQuAD dataset achieves the state of the art results at 75.9% F1 with a single model and 80.4% F1 with an ensemble.