SDNet: Contextualized Attention-based Deep Network for Conversational Question Answering

Jan. 31, 2019
Chenguang Zhu
Microsoft Speech and Dialogue Research Group
Motivation

• Traditional MRC models target for single-turn QA scenario

• No correlation between different questions/answers

• Humans interact by conversation, with rich context

• Models with conversational AI better match realistic settings, like customer support chatbot, social chatbot
CoQA Dataset

• From Stanford NLP Lab

• First conversational question and answering datasets

• Context can be carried over, requiring coreference resolution

• Model should comprehend both the passage and questions/answers from previous rounds
Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80. Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie’s husband Josh were coming as well. Jessica had . . .

Q₁: Who had a birthday?
A₁: Jessica
R₁: Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80.

Q₂: How old would she be?
A₂: 80
R₂: she was turning 80

Q₃: Did she plan to have any visitors?
A₃: Yes
R₃: Her granddaughter Annie was coming over

Q₄: How many?
A₄: Three
R₄: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie’s husband Josh were coming as well.

Q₅: Who?
A₅: Annie, Melanie and Josh
R₅: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie’s husband Josh were coming as well.
Our formulation

• Passage (context): $C$
• Previous rounds of questions and answers: $Q_1, A_1, ..., Q_{k-1}, A_{k-1}$
• Current question: $Q_k$

• We prepend the latest $N$ rounds of QAs to the current question:
  • $Q_k = (Q_{k-N}, A_{k-N}, ..., Q_{k-1}, A_{k-1}, Q_k)$
Encoding Layer

- For each word in passage and question
  - 300-dim GloVe word embedding
  - 1024-dim BERT contextual embedding
Using BERT

• We utilize BERT as a contextual embedder, with some adaptation

• BERT tokenizes text into **byte pair encoding (BPE)**

• It may not align with canonical tokenizers like spaCy
Using BERT--Alignment

• Tokenize text into word sequence \((w_1, w_2, \ldots, w_n)\)

• Then use BERT to tokenize each word into BPEs:
  • \(w_i \rightarrow (b_1, b_2, \ldots, b_{t_i})\)

• After obtaining contextual embedding for each BPE, we average the BERT embeddings for all the BPEs corresponding to a word
Using BERT—Linear combination

• BERT has many layers and the paper recommends using only the last layer

• We linearly combine all layers of output

• This can greatly improve the results
Using BERT—Linear combination

\[ \text{BERT}_w = \sum_{l=1}^{L} \alpha_l \frac{\sum_{t=1}^{s} h_t^l}{s} \]

Linear combination  Averaged BPE embedding for a word
Using BERT—Lock internal weights

• Locking weights of BERT (no backpropagation)

• In our case, finetuning does not work well
Representation

• Each word in passage is represented by $h_i^C$

• Each word in question is represented by $h_i^Q$
Integration Layer

• Word-level Inter-Attention
  • Attend from question to passage, i.e. $Attn(\{h^C_i\}, \{h^Q_i\}, \{h^Q_i\})$:

$$
S_{ij} = \text{ReLU}(U h^C_i) D \text{ReLU}(U h^Q_j),
$$

$$
\alpha_{ij} \propto \exp(S_{ij}),
$$

$$
\hat{h}^C_i = \sum_j \alpha_{ij} h^Q_j,
$$

• Passage words have several more features
  • POS, NER embedding
  • 3-dim exact matching vector
  • A normalized term frequency
RNN

- Passage (context) and question words both go through K layers of Bi-LSTMs

\[
\begin{align*}
    h_{1}^{C,k}, ..., h_{m}^{C,k} &= \text{BiLSTM} (h_{1}^{C,k-1}, ..., h_{m}^{C,k-1}), \\
    h_{1}^{Q,k}, ..., h_{n}^{Q,k} &= \text{BiLSTM} (h_{1}^{Q,k-1}, ..., h_{n}^{Q,k-1}), \\
    h_{i}^{C,0} &= \tilde{w}_{i}^{C}, \quad h_{i}^{Q,0} = \tilde{w}_{i}^{Q},
\end{align*}
\]

- Question words go through one more RNN layer

\[
\begin{align*}
    h_{1}^{Q,K+1}, ..., h_{n}^{Q,K+1} &= \text{BiLSTM} (h_{1}^{Q}, ..., h_{n}^{Q}), \\
    h_{i}^{Q} &= [h_{i}^{Q,1}; ..., h_{i}^{Q,K}]
\end{align*}
\]
Question self-attention

• Question contains conversation history
• We need to establish relationship between history and current question
• Self-attention:

$$\{u_i^Q\} = \text{Attn}(\{h_{i,K+1}^Q\}, \{h_{i,K+1}^Q\}, \{h_{i,K+1}^Q\})$$

• $$\{u_i^Q\}$$ is the final representation of question words
Multi-level inter-attention

- Adapted from FusionNet (Huang et al. 2018)

\[
\{m_i^{(k),C}\}_{i=1}^m = \text{Attn}(\{\text{HoW}_i^C\}_{i=1}^m, \{\text{HoW}_i^Q\}_{i=1}^n, \{h_i^{Q,k}\}_{i=1}^n), 1 \leq k \leq K + 1
\]

\[
\text{HoW}_i^C = [\text{GloVe}(w_i^C); \text{BERT}_{w_i^C}; h_i^{C,1}; ...; h_i^{C,k}],
\]

\[
\text{HoW}_i^Q = [\text{GloVe}(w_i^Q); \text{BERT}_{w_i^Q}; h_i^{Q,1}; ...; h_i^{Q,k}].
\]

- After multi-level inter-attention, we conduct RNN, self-attention and another RNN to obtain the final representation of context: \(\{u_i^C\}\)
Generate answer – positional probability

• Firstly, we condense the question into one vector

\[ u^Q = \sum_i \beta_i u^Q_i, \beta_i \propto \exp(w^T u^Q_i) \]

• Then we generate the probability that the answer starts and ends at each position:

Start probability: \( P^S_i \propto \exp((u^Q)^T W_S u^C_i) \)

GRU: \( t^Q = GRU(u^Q, \sum_i P^S_i u^C_i) \)

End probability: \( P^E_i \propto \exp((t^Q)^T W_E u^C_i) \)

• We generate the probability of “yes”, “no”, “no answer” similarly
Overall structure -- SDNet
Training and prediction

• Loss is cross-entropy
• Batch size: 1 dialogue
• 30 epochs
• ~2 hours per epoch on one Volta-100 GPU

• Prediction of span is done by choosing maximum (start x end) probability, with length <= 15

• Ensemble model consists of 12 best single models with different random seeds
Experiments

• CoQA dataset
• Dev set: 100 passages per in-domain topics
• Test set: 100 passages per in-domain and out-of-domain topics

<table>
<thead>
<tr>
<th>Domain</th>
<th>#Passage</th>
<th>#QA turn</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child Story</td>
<td>750</td>
<td>14.0</td>
</tr>
<tr>
<td>Literature</td>
<td>1,815</td>
<td>15.6</td>
</tr>
<tr>
<td>Mid/High Sc.</td>
<td>1,911</td>
<td>15.0</td>
</tr>
<tr>
<td>News</td>
<td>1,902</td>
<td>15.1</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>1,821</td>
<td>15.4</td>
</tr>
<tr>
<td>Out of domain</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Science</td>
<td>100</td>
<td>15.3</td>
</tr>
<tr>
<td>Reddit</td>
<td>100</td>
<td>16.6</td>
</tr>
<tr>
<td>Total</td>
<td>8,399</td>
<td>15.2</td>
</tr>
</tbody>
</table>
### Leaderboard (As of 12/11/2018)

Table 2: Model and human performance (% in F1 score) on the CoQA test set.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>PGNet</td>
<td>49.0</td>
<td>43.3</td>
<td>47.5</td>
<td>47.5</td>
<td>45.1</td>
<td>38.6</td>
<td>38.1</td>
<td>44.1</td>
</tr>
<tr>
<td>DrQA</td>
<td>46.7</td>
<td>53.9</td>
<td>54.1</td>
<td>57.8</td>
<td>59.4</td>
<td>45.0</td>
<td>51.0</td>
<td>52.6</td>
</tr>
<tr>
<td>DrQA+PGNet</td>
<td>64.2</td>
<td>63.7</td>
<td>67.1</td>
<td>68.3</td>
<td>71.4</td>
<td>57.8</td>
<td>63.1</td>
<td>65.1</td>
</tr>
<tr>
<td>BiDAF++</td>
<td>66.5</td>
<td>65.7</td>
<td>70.2</td>
<td>71.6</td>
<td>72.6</td>
<td>60.8</td>
<td>67.1</td>
<td>67.8</td>
</tr>
<tr>
<td>FlowQA</td>
<td>73.7</td>
<td>71.6</td>
<td>76.8</td>
<td>79.0</td>
<td>80.2</td>
<td>67.8</td>
<td>76.1</td>
<td>75.0</td>
</tr>
<tr>
<td>SDNet (single)</td>
<td>75.4</td>
<td>73.9</td>
<td>77.1</td>
<td>80.3</td>
<td>83.1</td>
<td>69.8</td>
<td>76.8</td>
<td>76.6</td>
</tr>
<tr>
<td>SDNet (ensemble)</td>
<td>78.7</td>
<td>77.1</td>
<td>80.2</td>
<td>81.9</td>
<td>85.2</td>
<td>72.3</td>
<td>79.7</td>
<td>79.3</td>
</tr>
<tr>
<td>Human</td>
<td>90.2</td>
<td>88.4</td>
<td>89.8</td>
<td>88.6</td>
<td>89.9</td>
<td>86.7</td>
<td>88.1</td>
<td>88.8</td>
</tr>
</tbody>
</table>
Ablation Studies

Table 3: Ablation study of SDNet on CoQA development dataset.

<table>
<thead>
<tr>
<th>Model</th>
<th>$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDNet</td>
<td>77.99</td>
</tr>
<tr>
<td>− Variational dropout</td>
<td>77.75</td>
</tr>
<tr>
<td>− Question self attention</td>
<td>77.24</td>
</tr>
<tr>
<td>Using last layer of BERT output (no weighted sum)</td>
<td>76.24</td>
</tr>
<tr>
<td>BERT-base</td>
<td>75.38</td>
</tr>
<tr>
<td>ELMo</td>
<td>73.37</td>
</tr>
<tr>
<td>− BERT</td>
<td>70.84</td>
</tr>
</tbody>
</table>
Effect of history

<table>
<thead>
<tr>
<th>#previous QA rounds</th>
<th>$N$</th>
<th>$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td>69.43</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>76.70</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>77.99</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>77.39</td>
</tr>
</tbody>
</table>
Training
Summary

• SDNet leverages MRC, BERT and conversational history

• BERT is important, but adaptation is required to use it as a module

• Next step is open-domain conversational QA, with no passage given, and model may also ask questions
Introduction to SDRG

• Microsoft Speech and Dialogue Research Group

• Conduct research on speech recognition and conversational AI

• We’re from Stanford, Princeton, Cambridge, IBM, ...

• We’re now hiring researchers and machine learning engineers

• We have world-class research environment
  • DGX-1 machine and large clusters with tons of GPUs

• Welcome to learn more!
Q&A

• Thank you!

• My Email: chezhu@microsoft.com

• More details in our paper:
  • SDNet: Contextualized Attention-based Deep Network for Conversational Question Answering – Chenguang Zhu, Michael Zeng, Xuedong Huang
  • https://arxiv.org/abs/1812.03593