
Martin Raison\textsuperscript{1}, Pierre-Emmanuel Mazaré\textsuperscript{1}, Rajarshi Das\textsuperscript{2}, Antoine Bordes\textsuperscript{1}

\textsuperscript{1}FAIR, \textsuperscript{2}UMass Amherst

Presenter: Derrick Blakely

Department of Computer Science, University of Virginia

https://qdata.github.io/deep2Read/
Roadmap

1. Background
2. Motivation
3. About Weaver
4. Results
5. Conclusions/Takeaways
Roadmap

1. Background
2. Motivation
3. About Weaver
4. Results
5. Conclusions/Takeaways
High Level Background

- Jason Weston: “far off goal” is creating intelligent dialog agents
High Level Background

- Jason Weston: “far off goal” is creating intelligent dialog agents
- Requirements: long and short-term knowledge, reasoning ability, not too much supervision, transfer, efficiency
High Level Background

- Jason Weston: “far off goal” is creating intelligent dialog agents
- Requirements: long and short-term knowledge, reasoning ability, not too much supervision, transfer, efficiency
- Richard Socher: “Can we frame all of NLP as QA?”
High Level Background

- Jason Weston: “far off goal” is creating intelligent dialog agents
- Requirements: long and short-term knowledge, reasoning ability, not too much supervision, transfer, efficiency
- Richard Socher: “Can we frame all of NLP as QA?”
- Can we avoid imposing too much structure?
<table>
<thead>
<tr>
<th>Task 1: Single Supporting Fact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary went to the bathroom.</td>
</tr>
<tr>
<td>John moved to the hallway.</td>
</tr>
<tr>
<td>Mary travelled to the office.</td>
</tr>
<tr>
<td>Where is Mary? A: office</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Task 2: Two Supporting Facts</th>
</tr>
</thead>
<tbody>
<tr>
<td>John is in the playground.</td>
</tr>
<tr>
<td>John picked up the football.</td>
</tr>
<tr>
<td>Bob went to the kitchen.</td>
</tr>
<tr>
<td>Where is the football? A: playground</td>
</tr>
</tbody>
</table>
bAbI (Weston et al, 2015)

**Task 1: Single Supporting Fact**
Mary went to the bathroom.
John moved to the hallway.
Mary travelled to the office.
Where is Mary? A: office

**Task 2: Two Supporting Facts**
John is in the playground.
John picked up the football.
Bob went to the kitchen.
Where is the football? A: playground

- 6 dialog tasks, 20 QA tasks
bAbI (Weston et al, 2015)

- 6 dialog tasks, 20 QA tasks
- Good collection--necessary (but not sufficient) for dialog

**Task 1: Single Supporting Fact**
- Mary went to the bathroom.
- John moved to the hallway.
- Mary travelled to the office.
- Where is Mary? A: office

**Task 2: Two Supporting Facts**
- John is in the playground.
- John picked up the football.
- Bob went to the kitchen.
- Where is the football? A: playground
Article: Endangered Species Act
Paragraph: “...Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition was raised.”

Question 1: “Which laws faced significant opposition?”
Plausible Answer: later laws

Question 2: “What was the name of the 1937 treaty?”
Plausible Answer: Bald Eagle Protection Act
SQuAD (Rajpurkar et al, 2016)

- 87K questions

**Article:** Endangered Species Act

**Paragraph:** “… Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition was raised.”

**Question 1:** “Which laws faced significant opposition?”

Plausible Answer: *later laws*

**Question 2:** “What was the name of the 1937 treaty?”

Plausible Answer: *Bald Eagle Protection Act*
SQuAD (Rajpurkar et al, 2016)

- 87K questions
- More organic than bAbI

**Article:** Endangered Species Act

**Paragraph:** “... Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition was raised.”

**Question 1:** “Which laws faced significant opposition?”
**Plausible Answer:** later laws

**Question 2:** “What was the name of the 1937 treaty?”
**Plausible Answer:** Bald Eagle Protection Act
SQuAD (Rajpurkar et al, 2016)

- 87K questions
- More organic than bAbI
- Focus of intensive effort

**Article:** Endangered Species Act

**Paragraph:** “... Other legislation followed, including the Migratory Bird Conservation Act of 1929, a *1937 treaty* prohibiting the hunting of right and gray whales, and the *Bald Eagle Protection Act of 1940*. These later *laws* had a low cost to society—the species were relatively rare—and little *opposition* was raised.”

**Question 1:** “Which laws faced significant *opposition*?”

**Plausible Answer:** *later laws*

**Question 2:** “What was the name of the *1937 treaty*?”

**Plausible Answer:** *Bald Eagle Protection Act*
SQuAD (Rajpurkar et al, 2016)

- 87K questions
- More organic than bAbI
- Focus of intensive effort
- Some very accurate (and complex) models have beat human performance

Article: Endangered Species Act
Paragraph: “...Other legislation followed, including the Migratory Bird Conservation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940. These later laws had a low cost to society—the species were relatively rare—and little opposition was raised.”

Question 1: “Which laws faced significant opposition?”
Plausible Answer: later laws

Question 2: “What was the name of the 1937 treaty?”
Plausible Answer: Bald Eagle Protection Act
Classical QA

\[ P(w_1, \ldots, w_m) = \prod_{i=1}^{m} P(w_i \mid w_1, \ldots, w_{i-1}) \approx \prod_{i=1}^{m} P(w_i \mid w_{i-(n-1)}, \ldots, w_{i-1}) \]
Classical QA

\[ P(w_1, \ldots, w_m) = \prod_{i=1}^{m} P(w_i \mid w_1, \ldots, w_{i-1}) \approx \prod_{i=1}^{m} P(w_i \mid w_{i-(n-1)}, \ldots, w_{i-1}) \]

- Language modeling: requires n-gram counts
Classical QA

\[ P(w_1, \ldots, w_m) = \prod_{i=1}^{m} P(w_i \mid w_1, \ldots, w_{i-1}) \approx \prod_{i=1}^{m} P(w_i \mid w_{i-(n-1)}, \ldots, w_{i-1}) \]

- Language modeling: requires n-gram counts
- Hard to handle long-range dependencies
Classical QA

\[ P(w_1, \ldots, w_m) = \prod_{i=1}^{m} P(w_i \mid w_1, \ldots, w_{i-1}) \approx \prod_{i=1}^{m} P(w_i \mid w_{i-(n-1)}, \ldots, w_{i-1}) \]

- Language modeling: requires n-gram counts
- Hard to handle long-range dependencies
- Requires explicitly structuring text data via knowledge bases (e.g., WikiData or DBpedia)
Classical QA
Can be used to create a language model
RNN’s

- Can be used to create a language model
- Can be used to encode questions and contexts
RNN’s

- Can be used to create a language model
- Can be used to encode questions and contexts
- Gradient problem and better dependency modeling ➔ GRU’s and LSTM’s
RNN’s

- Can be used to create a language model
- Can be used to encode questions and contexts
- Gradient problem and better dependency modeling → GRU’s and LSTM’s
- LSTM’s alone still inadequate for long-range encoding and reasoning
Memory Networks (Weston et al, 2014/2015)

- “IGOR” model
Memory Networks (Weston et al, 2014/2015)

- “IGOR” model
- I: convert data to a feature representation
Memory Networks (Weston et al, 2014/2015)

- “IGOR” model
- I: convert data to a feature representation
- G (generalization): update memory given new input
Memory Networks (Weston et al, 2014/2015)

- “IGOR” model
- I: convert data to a feature representation
- G (generalization): update memory given new input
- O (output): use existing memories to produce new output
Memory Networks (Weston et al, 2014/2015)

- “IGOR” model
- I: convert data to a feature representation
- G (generalization): update memory given new input
- O (output): use existing memories to produce new output
  - Find the relevant memory cells using some matching function (they use $q^T U^T U_d$)
Memory Networks (Weston et al, 2014/2015)

● “IGOR” model
● I: convert data to a feature representation
● G (generalization): update memory given new input
● O (output): use existing memories to produce new output
  ○ Find the relevant memory cells using some matching function (they use $q^T U^T U d$)
  ○ Typically involves a 2 hops
Memory Networks (Weston et al, 2014/2015)

- “IGOR” model
- I: convert data to a feature representation
- G (generalization): update memory given new input
- O (output): use existing memories to produce new output
  - Find the relevant memory cells using some matching function (they use $q^T U^T Ud$)
  - Typically involves a 2 hops
- R (response): get the actual text answer
Memory Networks (Weston et al, 2014/2015)
Memory Networks (Weston et al, 2014/2015)
Dynamic Memory Networks (Socher et al, 2015)
Post-2015 Architectures

- Ideas from MemoryNets and DMN’s always used
Post-2015 Architectures

- Ideas from MemoryNets and DMN’s always used
- GRU’s and LSTM’s (typically bidirectional) always used
Post-2015 Architectures

- Ideas from MemoryNets and DMN’s always used
- GRU’s and LSTM’s (typically bidirectional) always used
- Attention mechanism sprinkled liberally
Post-2015 Architectures

- Ideas from MemoryNets and DMN’s always used
- GRU’s and LSTM’s (typically bidirectional) always used
- Attention mechanism sprinkled liberally
- Performance on bAbI and SQuAD have been great
Post-2015 Architectures

- Ideas from MemoryNets and DMN’s always used
- GRU’s and LSTM’s (typically bidirectional) always used
- Attention mechanism sprinkled liberally
- Performance on bAbI and SQuAD have been great
- Models super specialized for these select tasks
Post-2015 Architectures

- Ideas from MemoryNets and DMN’s always used
- GRU’s and LSTM’s (typically bidirectional) always used
- Attention mechanism sprinkled liberally
- Performance on bAbI and SQuAD have been great
- Models super specialized for these select tasks
- Performance degrades as the context grows
Roadmap

1. Background
2. Motivation
3. About Weaver
4. Results
5. Conclusions/Takeaways
Roadmap

1. Background
2. Motivation
3. About Weaver
4. Results
5. Conclusions/Takeaways
Motivation

- Far-off goal: intelligent dialog agents
Motivation

- Far-off goal: intelligent dialog agents
- Generalizable models (especially if most of NLP can be cast into a QA problem)
Motivation

- Far-off goal: intelligent dialog agents
- Generalizable models (especially if most of NLP can be cast into a QA problem)
- Avoid using too much attention
Motivation

- Far-off goal: intelligent dialog agents
- Generalizable models (especially if most of NLP can be cast into a QA problem)
- Avoid using too much attention
- Models aren’t working well with longer contexts
Roadmap

1. Background
2. Motivation
3. About Weaver
4. Results
5. Conclusions/Takeaways
Roadmap

1. Background
2. Motivation
3. About Weaver
4. Results
5. Conclusions/Takeaways
Weaver

1. Input word embedding with fastText trained on a large corpus
Weaver

1. Input word embedding with fastText trained on a large corpus

2. Context and question co-encoding
1. Input word embedding with fastText trained on a large corpus
2. Context and question co-encoding
3. Memory network step
Weaver

1. Input word embedding with fastText trained on a large corpus
2. Context and question co-encoding
3. Memory network step
4. Final answer prediction
Embedding

- Question:
  \[ [q_1, q_2, \ldots, q_m] \]

- Content:
  \[ [c_1, c_2, \ldots, c_n] \]
Question and Context Co-Encoding

- Coordinate map:

\[ f : (q_i, c_j) \rightarrow [q_i \| c_j] \]
Question and Context Co-Encoding

- Coordinate map:
  \[ f : (q_i, c_j) \rightarrow [q_i \| c_j] \]

- What they actually do:
  \[ f : (q_i, c_j, c_j^{extra}) \rightarrow [q_i \| q_i - c_j \| q_i^T c_j \| c_j^{extra}] \]
Question and Context Co-Encoding

3d tensor ∈ m x n x d  M₀
1. Slice in the “context direction” $\rightarrow$ n slices of size $m \times d$
Question and Context Co-Encoding

1. Slice in the “context direction” ➔ n slices of size m x d

2. Feed each slice into BiLSTM ➔ obtain $M_1$ (n slices of size m x 2h)
Question and Context Co-Encoding

1. Slice in the “context direction” → $n$ slices of size $m \times d$

2. Feed each slice into BiLSTM → obtain $M_1$ ($n$ slices of size $m \times 2h$)

3. Slice $M_1$ in the “question direction”
Question and Context Co-Encoding

1. Slice in the “context direction” $\rightarrow$ n slices of size $m \times d$

2. Feed each slice into BiLSTM $\rightarrow$ obtain $M_1$ (n slices of size $m \times 2h$)

3. Slice $M_1$ in the “question direction”

4. Feed each slice into (new) BiLSTM $\rightarrow$ obtain $M_2$
Question and Context Co-Encoding

1. Slice in the “context direction” → n slices of size m x d

2. Feed each slice into BiLSTM → obtain $M_1$ (n slices of size m x 2h)

3. Slice $M_1$ in the “question direction”

4. Feed each slice into (new) BiLSTM → obtain $M_2$

5. Repeat
Question and Context Co-Encoding
Memory Network

- Co-encoding outputs can be used directly, but using a memory network was better
Memory Network

- Co-encoding outputs can be used directly, but using a memory network was better
- Similar to end-to-end MemNets (Sukhbaatar et al, 2015) and DMN’s
Memory Network

- Co-encoding outputs can be used directly, but using a memory network was better
- Similar to end-to-end MemNets (Sukhbaatar et al, 2015) and DMN’s
- Uses T hops and attention
Memory Network

- Co-encoding outputs can be used directly, but using a memory network was better
- Similar to end-to-end MemNets (Sukhbaatar et al, 2015) and DMN’s
- Uses $T$ hops and attention

\[
x_t = C^h W^c \text{softmax}(C^h W^h s_t)
\]

\[
s_{t+1} = \text{GRU}(x_t, s_t)
\]
Answer Prediction

- Softmax to predict indices for start and end of the answer

\[
\begin{align*}
p^s &= \text{softmax}(C^h W^s s_T) \\
p^e &= \text{softmax}(C^h W^e s_T)
\end{align*}
\]
Answer Prediction

- Softmax to predict indices for start and end of the answer

\[ p^s = \text{softmax}(C^h W^s s_T) \]
\[ p^e = \text{softmax}(C^h W^e s_T) \]

- Max:

\[ p^s_i p^e_j \text{ for } i \leq j \leq i + 15 \]
Roadmap

1. Background
2. Motivation
3. About Weaver
4. Results
5. Conclusions/Takeaways
Roadmap

1. Background
2. Motivation
3. About Weaver
4. Results
5. Conclusions/Takeaways
Results

- BAbI - solves 17 out of 20 tasks (though they don’t count two of the ones Weaver couldn’t do)
- SQuAD (normal):

<table>
<thead>
<tr>
<th></th>
<th>Dev set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EM</td>
<td>F1</td>
</tr>
<tr>
<td>DrQA</td>
<td>69.5</td>
<td>78.8</td>
</tr>
<tr>
<td>Conductor-net</td>
<td>72.1</td>
<td>81.4</td>
</tr>
<tr>
<td>M-Reader+RL</td>
<td>72.1</td>
<td>81.6</td>
</tr>
<tr>
<td>DCN+</td>
<td>74.5</td>
<td>83.1</td>
</tr>
<tr>
<td>FusionNet</td>
<td>75.3</td>
<td>83.6</td>
</tr>
<tr>
<td>SAN</td>
<td>76.2</td>
<td>84.1</td>
</tr>
<tr>
<td>Weaver</td>
<td>74.1</td>
<td>82.4</td>
</tr>
</tbody>
</table>
Results

- BAbI - solves 17 out of 20 tasks (though they don’t count two of the ones Weaver couldn’t do)
- SQuAD (document-level):

<table>
<thead>
<tr>
<th></th>
<th>Train</th>
<th>Test</th>
<th>EM</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>DrQA</td>
<td>paragraph</td>
<td>full doc.</td>
<td>49.4</td>
<td>58.0</td>
</tr>
<tr>
<td>DrQA*</td>
<td>paragraph</td>
<td>full doc.</td>
<td>59.1</td>
<td>67.0</td>
</tr>
<tr>
<td>DrQA*</td>
<td>full doc.</td>
<td>full doc.</td>
<td>64.7</td>
<td>73.2</td>
</tr>
<tr>
<td>Weaver</td>
<td>paragraph</td>
<td>full doc.</td>
<td>60.6</td>
<td>69.7</td>
</tr>
<tr>
<td>Weaver</td>
<td>full doc.</td>
<td>full doc.</td>
<td><strong>67.0</strong></td>
<td><strong>75.9</strong></td>
</tr>
</tbody>
</table>
# Results - All of English Wikipedia

<table>
<thead>
<tr>
<th></th>
<th>SQuAD</th>
<th>CuratedTREC</th>
<th>WebQuestions</th>
<th>WikiMovies</th>
</tr>
</thead>
<tbody>
<tr>
<td>YodaQA</td>
<td>- addtl sources</td>
<td>-</td>
<td>31.3</td>
<td><strong>39.8</strong></td>
</tr>
<tr>
<td>DrQA</td>
<td>- SQuAD train</td>
<td>27.1</td>
<td>19.7</td>
<td>11.8</td>
</tr>
<tr>
<td></td>
<td>- fine-tuning</td>
<td>28.4</td>
<td>25.7</td>
<td>19.5</td>
</tr>
<tr>
<td>DrQA*</td>
<td>- SQuAD train</td>
<td>39.5</td>
<td>21.3</td>
<td>14.2</td>
</tr>
<tr>
<td></td>
<td>- fine-tuning</td>
<td>40.4</td>
<td>28.8</td>
<td>24.3</td>
</tr>
<tr>
<td>Reinf. reader-ranker</td>
<td>- fine-tuning</td>
<td>29.1</td>
<td>28.4</td>
<td>17.1</td>
</tr>
<tr>
<td>Weaver</td>
<td>- SQuAD train</td>
<td><strong>42.3</strong></td>
<td>21.3</td>
<td>13.0</td>
</tr>
<tr>
<td></td>
<td>- fine-tuning</td>
<td>-</td>
<td><strong>37.9</strong></td>
<td>23.7</td>
</tr>
</tbody>
</table>
Roadmap

1. Background
2. Motivation
3. About Weaver
4. Results
5. Conclusions/Takeaways
Roadmap

1. Background
2. Motivation
3. About Weaver
4. Results
5. Conclusions/Takeaways
Conclusions and Takeaways

- First step for architectures needs to be a traditional IR module
Conclusions and Takeaways

- First step for architectures needs to be a traditional IR module
- Clever use of LSTM’s reduces the need for attention
Conclusions and Takeaways

● First step for architectures needs to be a traditional IR module
● Clever use of LSTM’s reduces the need for attention
● Learning good representations for questions and contexts is where a lot of effort is going
Conclusions and Takeaways

- First step for architectures needs to be a traditional IR module
- Clever use of LSTM’s reduces the need for attention
- Learning good representations for questions and contexts is where a lot of effort is going
- Iterative attention mechanisms still important for QA tasks
Conclusions and Takeaways

● First step for architectures needs to be a traditional IR module
● Clever use of LSTM’s reduces the need for attention
● Learning good representations for questions and contexts is where a lot of effort is going
● Iterative attention mechanisms still important for QA tasks
● Still helpful to manually add in NLP features like NER and POS taggings
Questions?
Dynamic Memory Networks (Socher et al, 2015)

- Multiple passes used in the “Episodic memory module” to agglomerate the m-vectors
  - Reminiscent of bootstrapping--after a pass, it’s more confident about which parts of the input sequence matter
  - After multiple passes, model can get a more “global perspective”
- GRU’s often used instead of LSTM’s--same performance for encoding tasks but GRU’s have fewer parameters, so they’re often used instead of LSTMs
- Also interesting: Socher et al obtained good results by piping in image encodings instead of word vectors
- Dynamic Co-attention networks developed soon afterwards