Weaver: Deep Co-Encoding of Questions and Documents for Machine Reading (2018)

Martin Raison[1], Pierre-Emmanuel Mazaré[1], Rajarshi Das[2], Antoine Bordes[1]
[1] FAIR, [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?

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

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)

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
- Good collection--necessary (but not sufficient) for dialog

SQuAD (Rajpurkar et al, 2016)

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

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

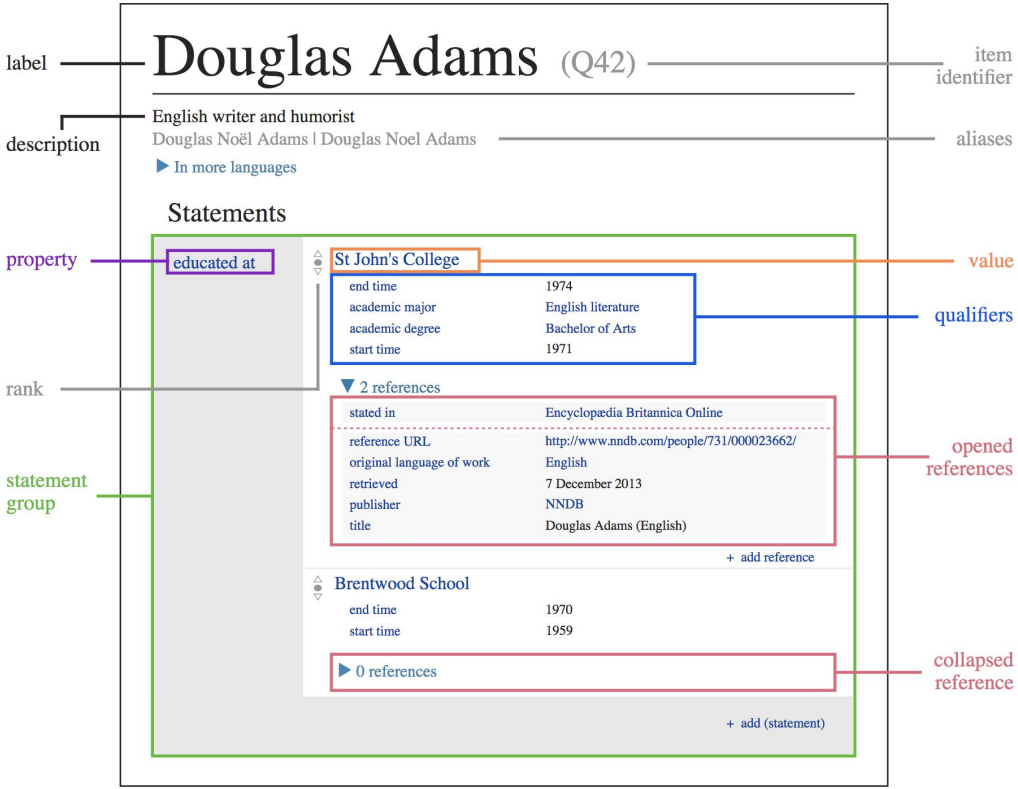
- Language modeling: requires n-gram counts
- Hard to handle long-range dependencies

Classical QA

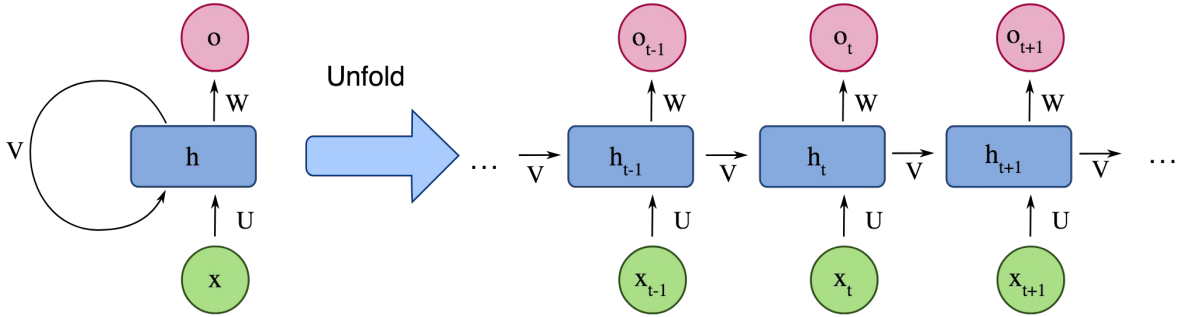
$$P(w_1, \ldots, w_m) = \prod_{i=1}^{m} P(w_i \mid w_1, \ldots, w_{i-1}) \approx \prod_{i=1}^{m} P(w_i \mid w_{i-(n-1)}, \ldots, w_{i-1})$$

- 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

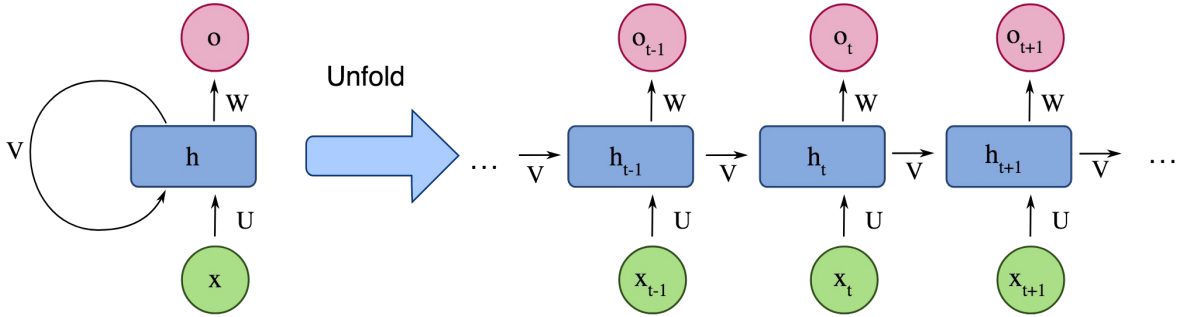


RNN's



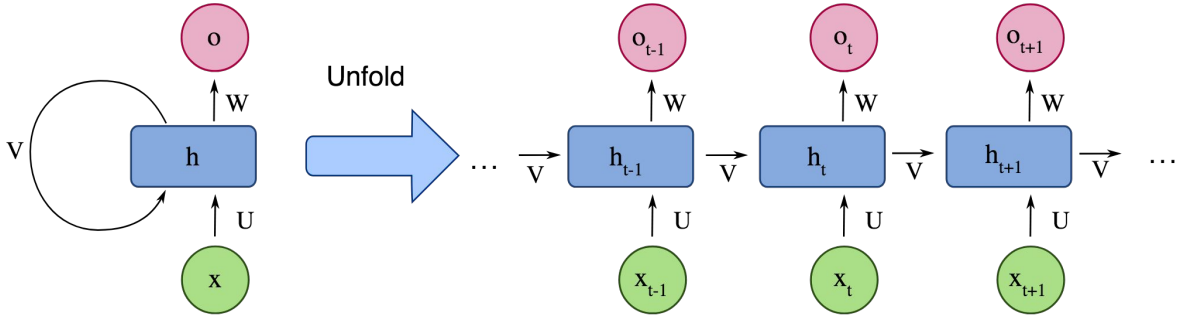
- Can be used to create a language model

RNN's



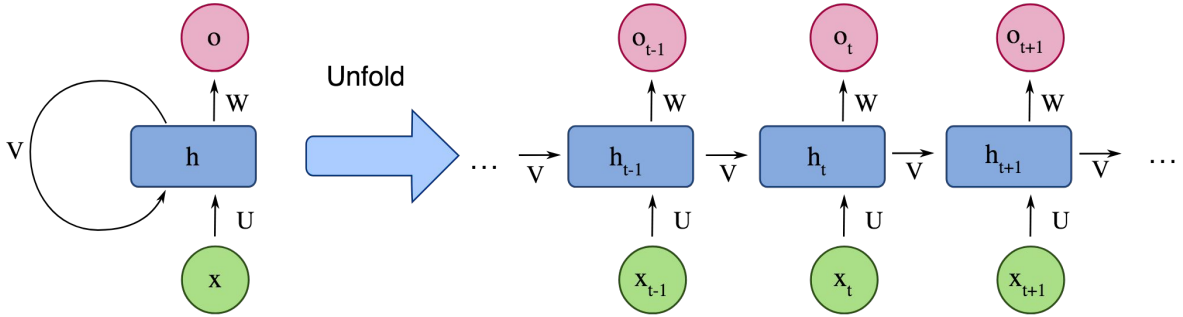
- Can be used to create a language model
- Can be used to encode questions and contexts

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 $q^T U^T U d$)

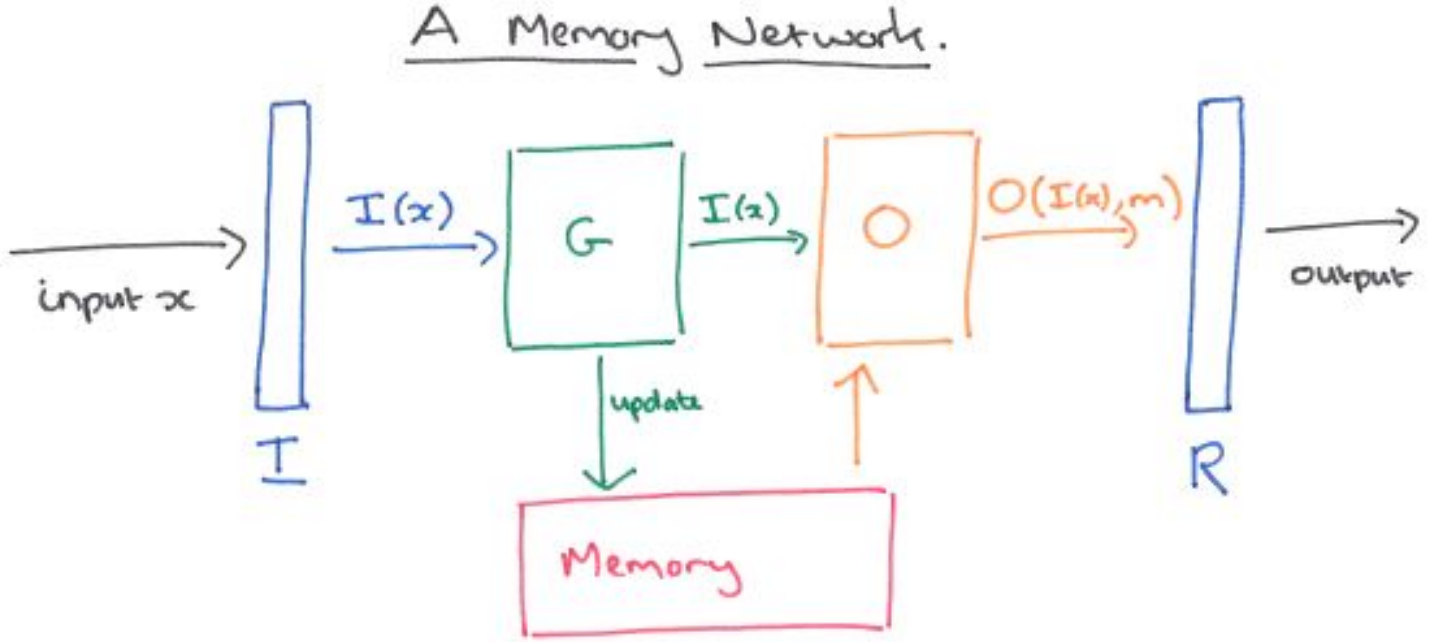
Memory Networks (Weston et al, 2014/2015)

- “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

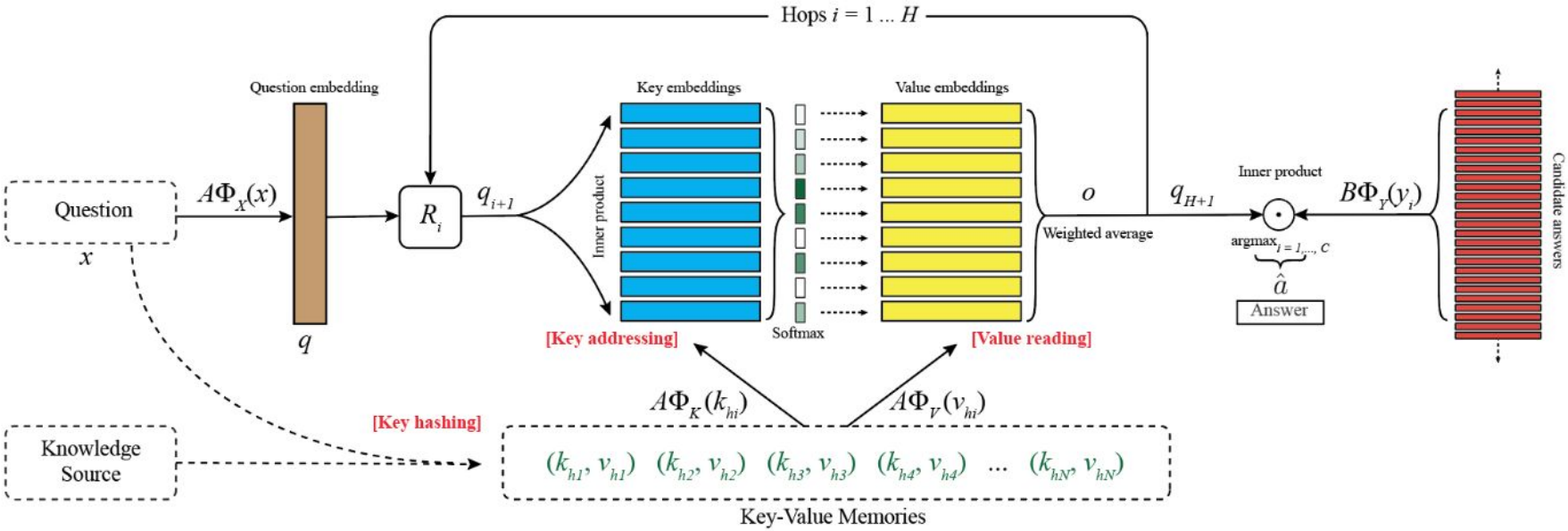
Memory Networks (Weston et al, 2014/2015)

- “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 $q^T U^T U d$)
  - Typically involves a 2 hops
- R (response): get the actual text answer

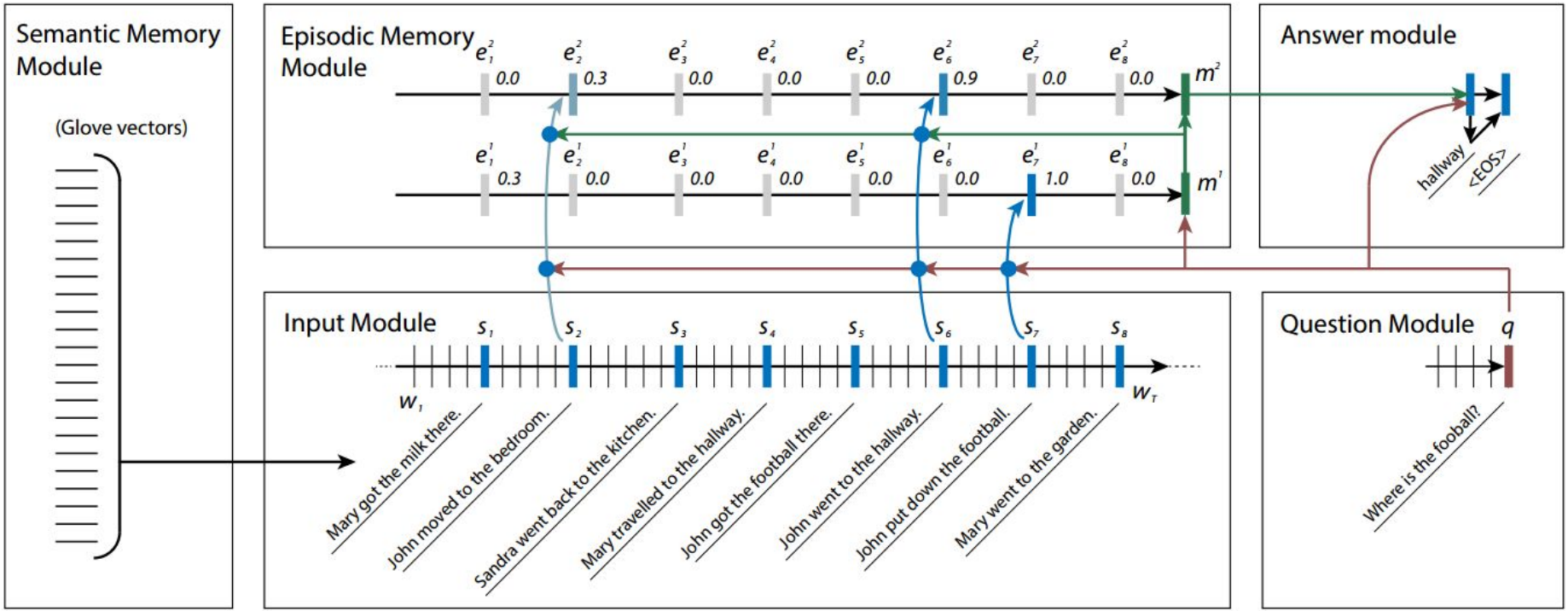
Memory Networks (Weston et al, 2014/2015)



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

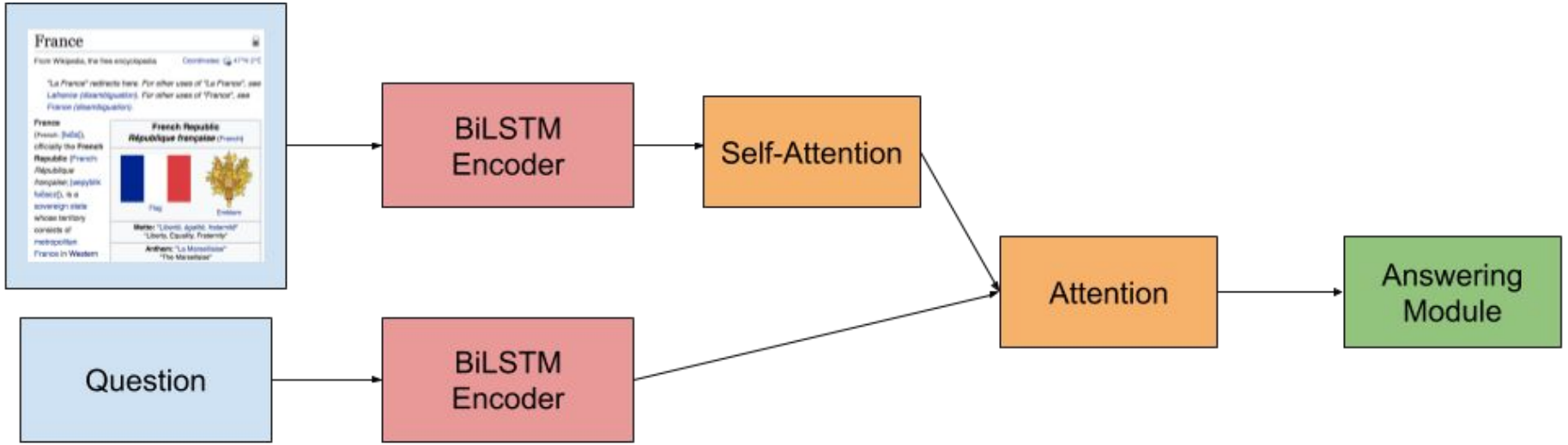
Post-2015 Architectures

- 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

Post-2015 Architectures



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

Weaver

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

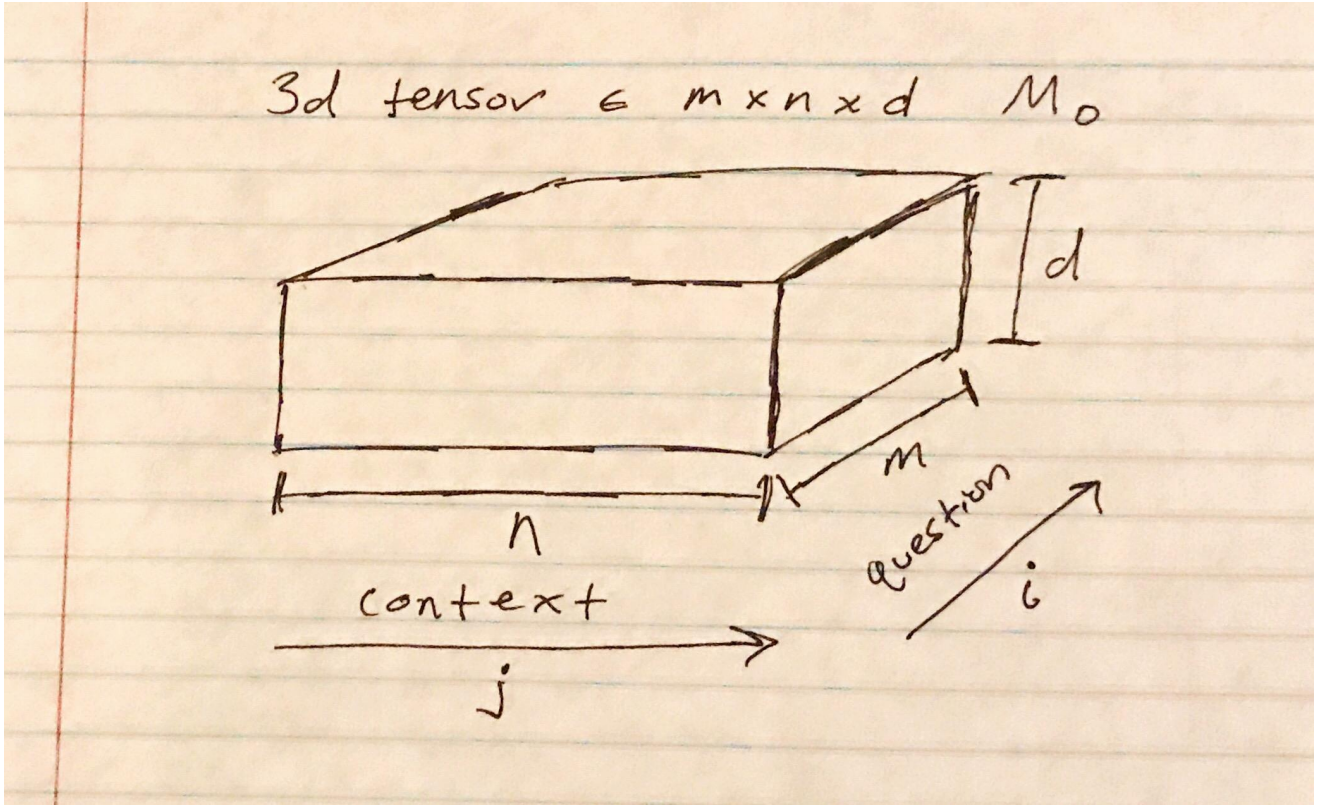
- Coordinate map:

$$f : (q_i, c_j) \rightarrow [q_i \| c_j]$$

- 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



Question and Context Co-Encoding

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

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

Question and Context Co-Encoding

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

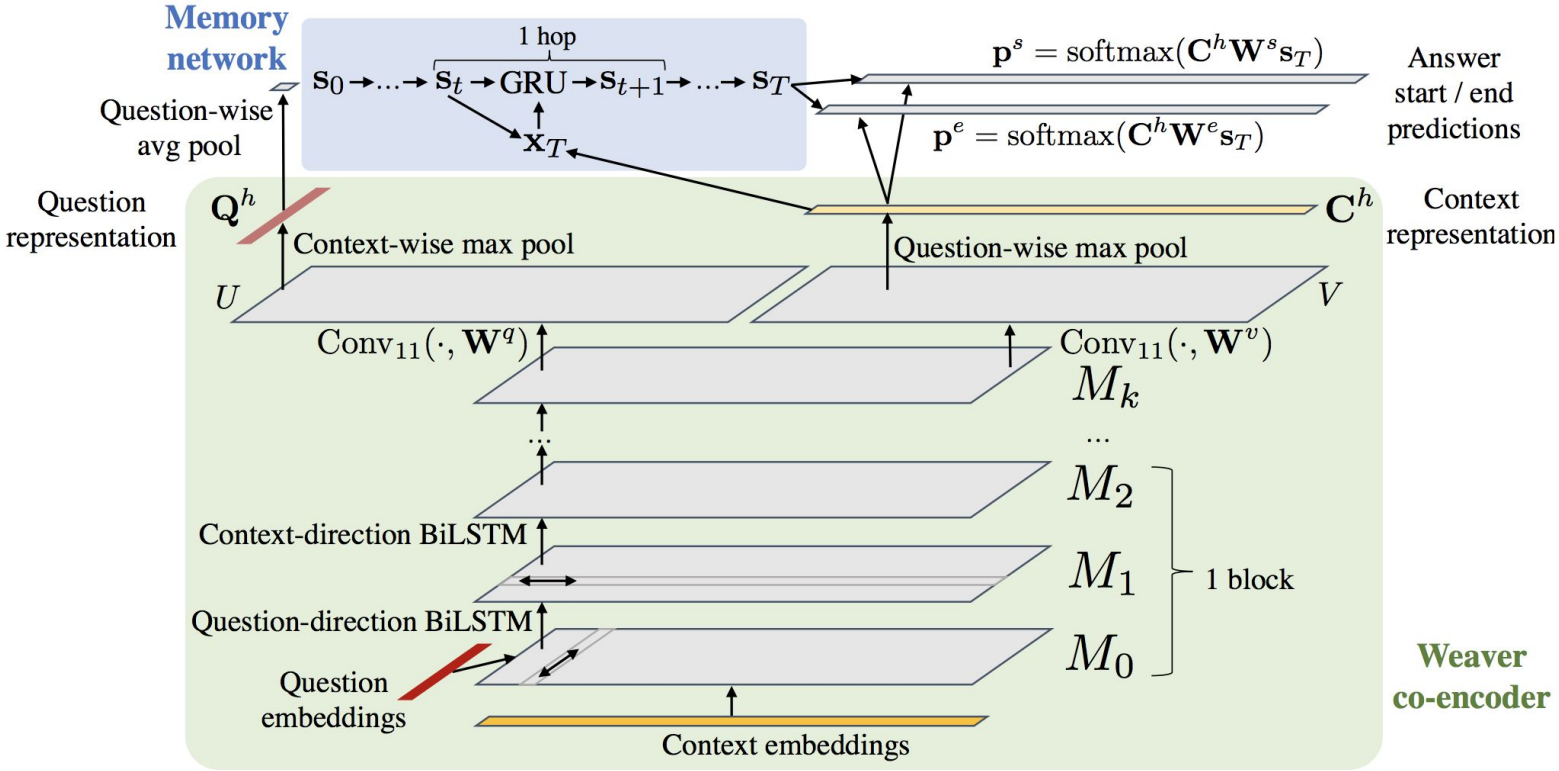
2. Feed each slice into BiLSTM ➜ obtain M_1 (n slices of size m x 2h)

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

$$\mathbf{x}_t = \mathbf{C}^h \mathbf{W}^c \text{softmax}(\mathbf{C}^h \mathbf{W}^h \mathbf{s}_t)$$

$$\mathbf{s}_{t+1} = \text{GRU}(\mathbf{x}_t, \mathbf{s}_t)$$

Answer Prediction

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

$$\mathbf{p}^s = \text{softmax}(\mathbf{C}^h \mathbf{W}^s \mathbf{s}_T)$$

$$\mathbf{p}^e = \text{softmax}(\mathbf{C}^h \mathbf{W}^e \mathbf{s}_T)$$

Answer Prediction

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

$$\mathbf{p}^s = \text{softmax}(\mathbf{C}^h \mathbf{W}^s \mathbf{s}_T)$$

$$\mathbf{p}^e = \text{softmax}(\mathbf{C}^h \mathbf{W}^e \mathbf{s}_T)$$

- Max:

$$\mathbf{p}_i^s \mathbf{p}_j^e \text{ for } i \leq j \leq i + 15$$

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):

	Dev set		Test set	
	EM	F1	EM	F1
DrQA	69.5	78.8	70.7	79.3
Conductor-net	72.1	81.4	72.6	81.4
M-Reader+RL	72.1	81.6	73.2	81.8
DCN+	74.5	83.1	75.1	83.1
FusionNet	75.3	83.6	76.0	83.9
SAN	**76.2**	**84.1**	**76.8**	**84.4**
Weaver	74.1	82.4	74.4	82.8

Results

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

	Train	Test	EM	F1
DrQA	paragraph	full doc.	49.4	58.0
DrQA*	paragraph	full doc.	59.1	67.0
DrQA*	full doc.	full doc.	64.7	73.2
Weaver	paragraph	full doc.	60.6	69.7
Weaver	full doc.	full doc.	**67.0**	**75.9**

Results - All of English Wikipedia

		SQuAD	CuratedTREC	WebQuestions	WikiMovies
YodaQA	- *addtl sources*	-	31.3	**39.8**	-
DrQA	- *SQuAD train*	27.1	19.7	11.8	24.5
	- *fine-tuning*	28.4	25.7	19.5	34.3
DrQA*	- *SQuAD train*	39.5	21.3	14.2	31.9
	- *fine-tuning*	40.4	28.8	24.3	**46.0**
Reinf. reader-ranker	- *fine-tuning*	29.1	28.4	17.1	38.8
Weaver	- *SQuAD train*	**42.3**	21.3	13.0	33.6
	- *fine-tuning*	-	**37.9**	23.7	43.9

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