End-To-End Memory Networks

Sainbayar Sukhbaatar, Arthur Szlam, Jason Wetson, Rob Fergus
Dept. Of Computer Science
Courant Institute, NYU
&
Facebook AI Research
New York

Outline

- Motivation
- Model
- Experiments
- Results
- Conclusion

Motivation

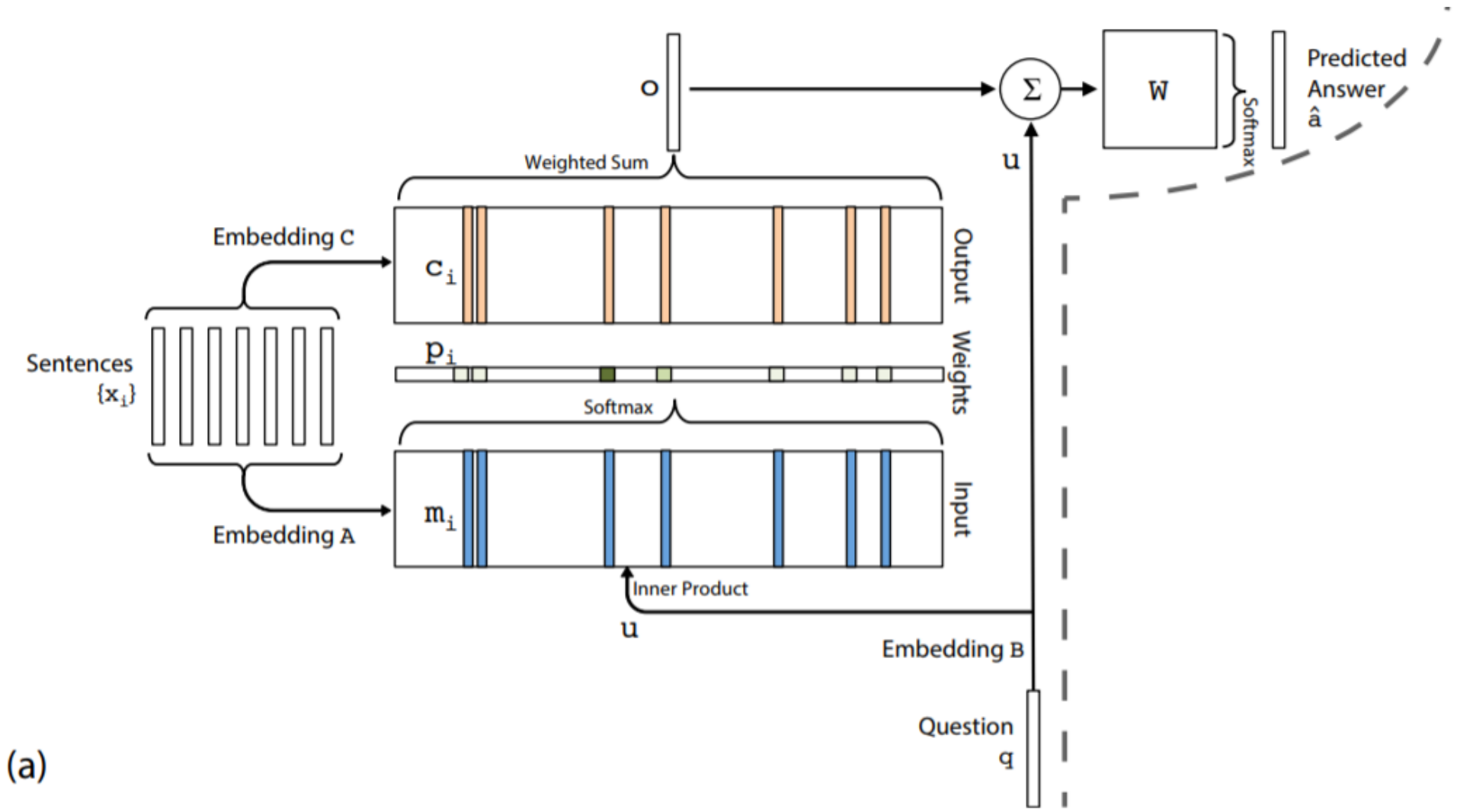
- Make a model that can perform many computational steps to answer a question.
- Make a model that describes dependencies in sequential data.
- I.E. sequential reasoning
- Lightweight & easily trainable

Motivation over MemNN

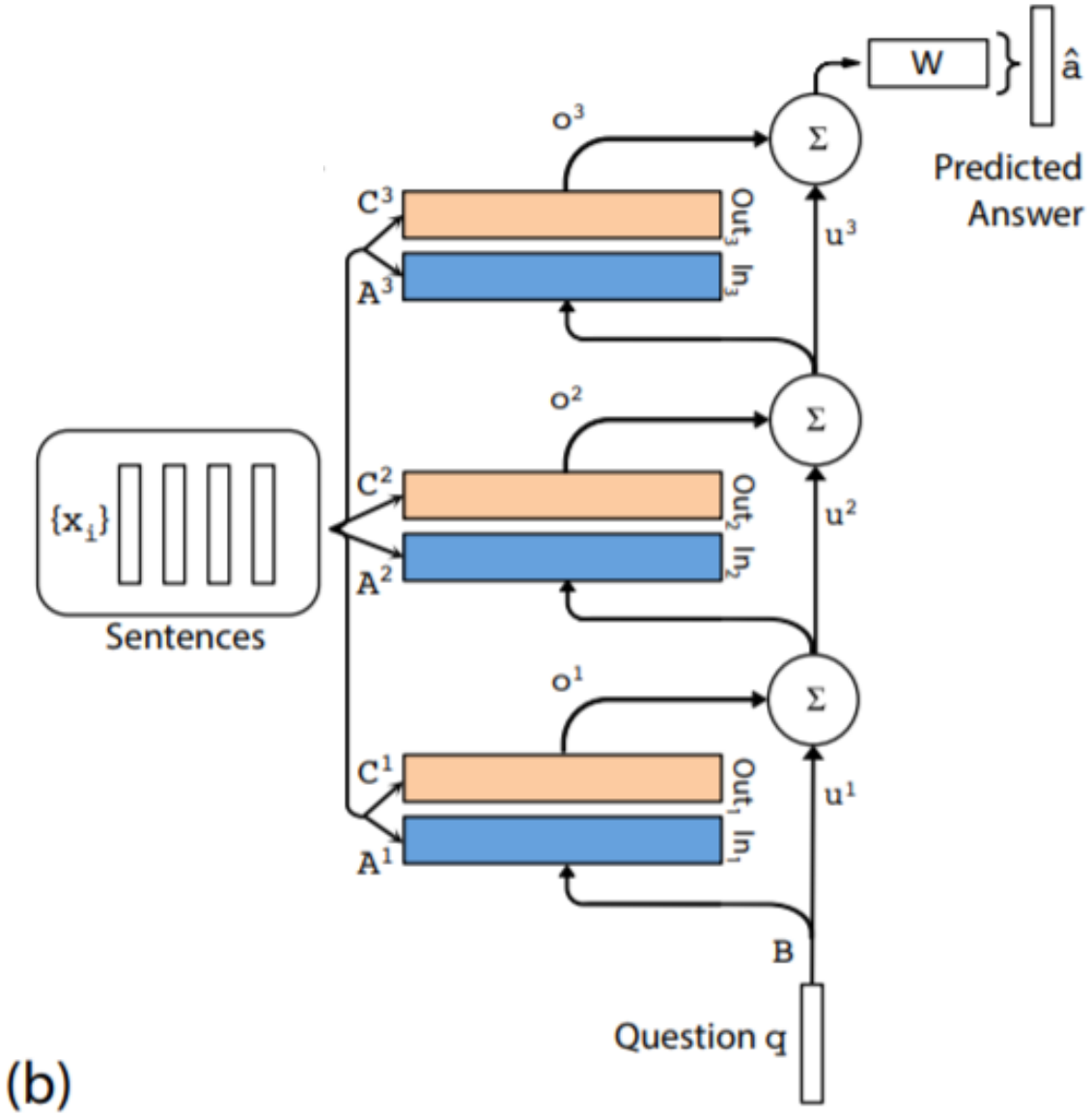
- End-To-End Trainable
- Far less supervision
- More generalizable

Overview of Model

- Variables:
  - Discrete set of inputs (x_i)
  - A query (q)
  - Produce an answer (a)
- Static Memory Bank
- Multiple Hops

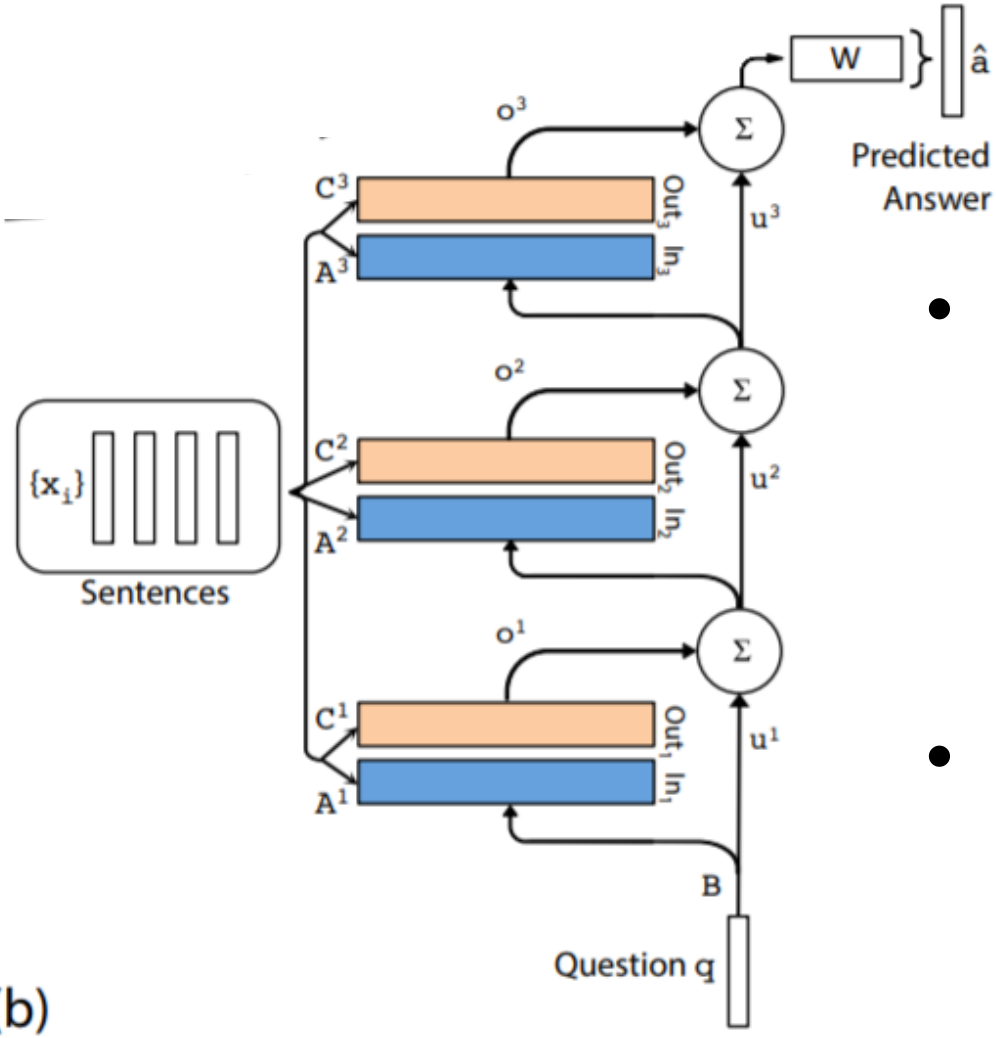


$$p_i = \text{Softmax}(u^T m_i) \qquad o = \sum_i p_i c_i \qquad \hat{a} = \text{Softmax}(W(o+u))$$


W
â
Predicted
Answer
Σ
Σ
Σ
o3
o2
o1
C3
C2
C1
A3
A2
A1
Out3
In3
Out2
In2
Out1
In1
u3
u2
u1
{xi}
Sentences
B
Question q

(b)

Weight Tying



- **Adjacent:**
  - $A^{k+1} = C^k$
  - $W^T = C^K$
  - B = A^1
- **Layer-wise (RNN-like):**
  - $A^1 = \ldots = A^k$, $C^1 = \ldots = C^k$
  - $u^{k+1} = Hu^k + o^k$

Sentence Representation

- Bag-of-words
  - $m_i = \sum_j Ax_{ij}$
- Position Encoding (PE)
  - $m_i = \sum_j l_j \cdot Ax_{ij}$
- Temporal Encoding (TE)
  - $m_i = \sum_j Ax_{ij} + T_A(i)$

Synthetic QA Experiments

Sam walks into the kitchen.
Sam picks up an apple.
Sam walks into the bedroom.
Sam drops the apple.
Q: Where is the apple?
A. Bedroom

Brian is a lion.
Julius is a lion.
Julius is white.
Bernhard is green.
Q: What color is Brian?
A. White

Similarity to Attention

Story (16: basic induction)	Support	Hop 1	Hop 2	Hop 3
Brian is a frog.	yes	0.00	0.98	0.00
Lily is gray.		0.07	0.00	0.00
Brian is yellow.	yes	0.07	0.00	1.00
Julius is green.		0.06	0.00	0.00
Greg is a frog.	yes	0.76	0.02	0.00
What color is Greg? Answer: yellow		**Prediction: yellow**		

Story (18: size reasoning)	Support	Hop 1	Hop 2	Hop 3
The suitcase is bigger than the chest.	yes	0.00	0.88	0.00
The box is bigger than the chocolate.		0.04	0.05	0.10
The chest is bigger than the chocolate.	yes	0.17	0.07	0.90
The chest fits inside the container.		0.00	0.00	0.00
The chest fits inside the box.		0.00	0.00	0.00
Does the suitcase fit in the chocolate? Answer: no		**Prediction: no**		

NOTE: This model does not use the “support” label during training

Results

	Baseline			MemN2N				
Task	Strongly Supervised MemNN [22]	LSTM [22]	MemNN WSH	1 hop PE LS joint	2 hops PE LS joint	3 hops PE LS joint	PE LS RN joint	PE LS LW joint
1: 1 supporting fact	0.0	50.0	0.1	0.8	0.0	0.1	0.0	0.1
2: 2 supporting facts	0.0	80.0	42.8	62.0	15.6	14.0	11.4	18.8
3: 3 supporting facts	0.0	80.0	76.4	76.9	31.6	33.1	21.9	31.7
4: 2 argument relations	0.0	39.0	40.3	22.8	2.2	5.7	13.4	17.5
5: 3 argument relations	2.0	30.0	16.3	11.0	13.4	14.8	14.4	12.9
6: yes/no questions	0.0	52.0	51.0	7.2	2.3	3.3	2.8	2.0
7: counting	15.0	51.0	36.1	15.9	25.4	17.9	18.3	10.1
8: lists/sets	9.0	55.0	37.8	13.2	11.7	10.1	9.3	6.1
9: simple negation	0.0	36.0	35.9	5.1	2.0	3.1	1.9	1.5
10: indefinite knowledge	2.0	56.0	68.7	10.6	5.0	6.6	6.5	2.6
11: basic coreference	0.0	38.0	30.0	8.4	1.2	0.9	0.3	3.3
12: conjunction	0.0	26.0	10.1	0.4	0.0	0.3	0.1	0.0
13: compound coreference	0.0	6.0	19.7	6.3	0.2	1.4	0.2	0.5
14: time reasoning	1.0	73.0	18.3	36.9	8.1	8.2	6.9	2.0
15: basic deduction	0.0	79.0	64.8	46.4	0.5	0.0	0.0	1.8
16: basic induction	0.0	77.0	50.5	47.4	51.3	3.5	2.7	51.0
17: positional reasoning	35.0	49.0	50.9	44.4	41.2	44.5	40.4	42.6
18: size reasoning	5.0	48.0	51.3	9.6	10.3	9.2	9.4	9.2
19: path finding	64.0	92.0	100.0	90.7	89.9	90.2	88.0	90.6
20: agent's motivation	0.0	9.0	3.6	0.0	0.1	0.0	0.0	0.2
Mean error (%)	6.7	51.3	40.2	25.8	15.6	13.3	12.4	15.2
Failed tasks (err. $>$ 5%)	4	20	18	17	11	11	11	10
On 10k training data								
Mean error (%)	3.2	36.4	39.2	24.5	10.9	7.9	7.5	11.0
Failed tasks (err. $>$ 5%)	2	16	17	16	7	6	6	6

Language Modeling

Model	Penn Treebank # of hidden	# of hops	memory size	Valid. perp.	Test perp.	Text8 # of hidden	# of hops	memory size	Valid. perp.	Test perp.
RNN [15]	300	-	-	133	129	500	-	-	-	184
LSTM [15]	100	-	-	120	115	500	-	-	122	154
SCRN [15]	100	-	-	120	115	500	-	-	-	161
MemN2N	150	2	100	128	121	500	2	100	152	187
	150	3	100	129	122	500	3	100	142	178
	150	4	100	127	120	500	4	100	129	162
	150	5	100	127	118	500	5	100	123	154
	150	6	100	122	115	500	6	100	124	155
	150	7	100	120	114	500	7	100	118	**147**
	150	6	25	125	118	500	6	25	131	163
	150	6	50	121	114	500	6	50	132	166
	150	6	75	122	114	500	6	75	126	158
	150	6	100	122	115	500	6	100	124	155
	150	6	125	120	112	500	6	125	125	157
	150	6	150	121	114	500	6	150	123	154
	150	7	200	118	**111**	-	-	-	-	-

Conclusion

- Outperforms all baselines with the same level of supervision (LSTMs etc.)
- Slightly worse than a strongly supervised Memory Network, but it was trained without supporting facts, so it can be easily trained in more general settings.
- On language modeling, outperforms RNNs and LSTMs