End to End Memory Networks a.k.a MemN2N[1]

Varun Gangal, CMU

Based on the work of

1. Sukhbaatar et al

Core Idea

Learn an **end-to-end backproppable** arch. which can **memorize** a **long input** and can then **access and aggregate** memories **multiple times** to generate output/answer.

Reading Comprehension QA setup

- Read sequence of sentences, answer single sentence Q
- Input Passage X = $x_1, x_2, x_3 \ldots x_N$
- Input units x_i - Here sentences - No segmenter needed! Memories= $m_1, m_2 \ldots m_N$
- Single sentence question Q - initialize controller
- Basic pipeline: Init u_1 to q, weigh memories w.r.t u_1, take weighted aggregate o_1, update u_1 to u_2, continue process. Finally, use 1-step softmax or decoder

Mem keys

- How to represent each sentence as mem key vector?
- Various choices, arch. independent of this
- CBOW - v(Sam)+v(ate)+v(food)
- CBOW+Temporal - v(Sam)+v(ate)+v(food)+**v(3)**
- BiLSTM - BiLSTM*(Sam ate food)+v(3)
- Etc, etc
- Here CBOW variants used in experiments
- Similarly for q - CBOW/BiLSTM etc

Mem.Values

- Simplest case: Same as key $c_i = m_i$; $C = A$
- Alternative: Use CBOW here, also but use separate embedding matrix
- Can have shared or different key-embeddings and value-embeddings for different hops. More on this later

Address-Read

Address

$$p_i = \text{Softmax}(u^T m_i)$$

Read

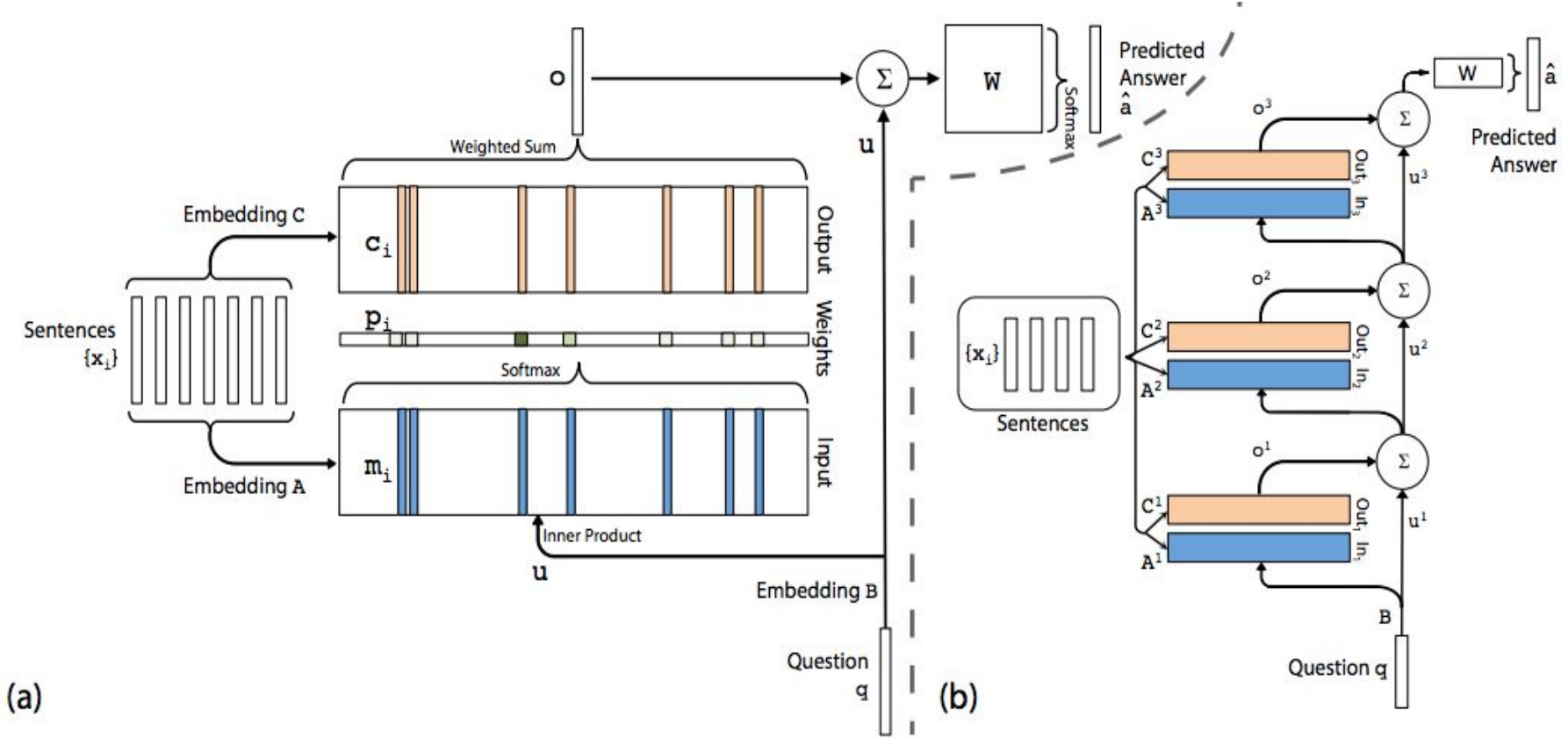
$$o = \sum_i p_i c_i.$$

Update-Controller

$$u^{k+1} = u^k + o^k.$$

Predict

$$\hat{a} = \text{Softmax}(W(o + u))$$


o
Σ
W
Softmax
Predicted Answer
â
Weighted Sum
Embedding C
Sentences
{x_i}
c_i
Output
p_i
Weights
Softmax
m_i
Input
Embedding A
Inner Product
u
Embedding B
Question q
(a)
{x_i}
Sentences
C^3
A^3
Out_3
In_3
C^2
A^2
Out_2
In_2
C^1
A^1
Out_1
In_1
o^3
o^2
o^1
u^3
u^2
u^1
B
Question q
(b)

a) 1-hop b) 3-hop

So What's New?

- ***Sequence to Sequence Attn (Bahdanau et al, 15)***
  - Very similar, authors note too, just not sold as mem
  - That was more MT-specific, this is more QA-specific.
  - But truly new: - **Multiple hops**
  - Mem can be shared across examples, attention can't. Distinction unexplored here.

So What's New?

- ***Memory Networks (Weston et al, 15)***
  - Use **explicit supervision** for memory access
  - Take **argmaxes (hard)** instead of **softmax (soft)**
  - **Single** hop or **double** hop - **Multiple hops** here
  - **Point 2** seems like **not such a big deal** now that we have **Gumbel-softmax-reparam** etc, but both works **precede** that.

Param.Sharing across hops

- Most general case: Separate A_k, C_k for each hop
- But: Ease of training, inductive bias
- Layer-Wise: $A_{k+1}=A_k$, $C_{k+1}=C_k$
- Adjacent: $A_{k+1}=C_k$

RC Dataset: 20 bAbi tasks

- Each test one type of reasoning.
- Simple language, small vocab.
- No coreference, hierarchical clauses etc. **Relevant subset** also provided, **this model doesn't use** it. Original memnets did (they needed too, else you can't take argmaxes in between and backprop, not without Gumbel reparam atleast)

Examples

Sam walks into the kitchen.	Brian is a lion.	Mary journeyed to the den.
Sam picks up an apple.	Julius is a lion.	Mary went back to the kitchen.
Sam walks into the bedroom.	Julius is white.	John journeyed to the bedroom.
Sam drops the apple.	Bernhard is green.	Mary discarded the milk.
Q: Where is the apple?	Q: What color is Brian?	Q: Where was the milk before the den?
A. Bedroom	A. White	A. Hallway

Some Tricks

- Inject random “noise” - empty memory vectors - 10% - found to help
- CBOW+Position (Within-Sent)+Temporal (Sent-Index)

Baselines

- Strong-supervised MemNet: Not really fair to compare
- Question-Answer LSTM: Of course, but
  - Didn't get why not Question LSTM+Passage LSTM etc. Would have been fairer
- Heuristic MemNets - Bias that first argmax sentence should word-overlap with question, second with answer.

Results

	Baseline			MemN2N								
Task	Strongly Supervised MemNN [22]	LSTM [22]	MemNN WSH	BoW	PE	PE LS	PE LS RN	1 hop PE LS joint	2 hops PE LS joint	3 hops PE LS joint	PE LS RN joint	PE LS LW joint
Mean error (%)	6.7	51.3	40.2	25.1	20.3	16.3	13.9	25.8	15.6	13.3	12.4	15.2
Failed tasks (err. > 5%)	4	20	18	15	13	12	11	17	11	11	11	10
On 10k training data												
Mean error (%)	3.2	36.4	39.2	15.4	9.4	7.2	6.6	24.5	10.9	7.9	7.5	11.0
Failed tasks (err. > 5%)	2	16	17	9	6	4	4	16	7	6	6	6

Story (1: 1 supporting fact)	Support	Hop 1	Hop 2	Hop 3
Daniel went to the bathroom.		0.00	0.00	0.03
Mary travelled to the hallway.		0.00	0.00	0.00
John went to the bedroom.		0.37	0.02	0.00
John travelled to the bathroom.	yes	0.60	0.98	0.96
Mary went to the office.		0.01	0.00	0.00
Where is John? Answer: bathroom Prediction: bathroom				

Story (2: 2 supporting facts)	Support	Hop 1	Hop 2	Hop 3
John dropped the milk.		0.06	0.00	0.00
John took the milk there.	yes	0.88	1.00	0.00
Sandra went back to the bathroom.		0.00	0.00	0.00
John moved to the hallway.	yes	0.00	0.00	1.00
Mary went back to the bedroom.		0.00	0.00	0.00
Where is the milk? Answer: hallway Prediction: hallway				

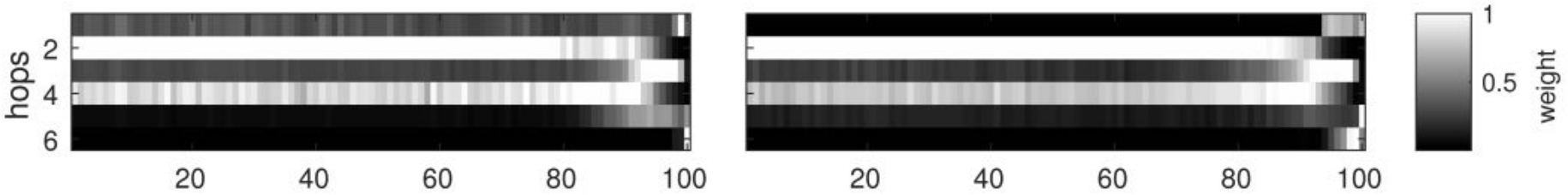
Story (16: basic induction)	Support	Hop 1	Hop 2	Hop 3
Brian is a frog.	yes	0.00	0.98	0.00
Lily is gray.		0.07	0.00	0.00
Brian is yellow.	yes	0.07	0.00	1.00
Julius is green.		0.06	0.00	0.00
Greg is a frog.	yes	0.76	0.02	0.00
What color is Greg? Answer: yellow Prediction: yellow				

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				

Language Modelling

- $P(w_i \mid w1, .. w_{(i-1)})$
- Memories = Words seen so far - Again no segmenter
- Query - Fixed question vector
  - Maybe better: Fixed+RNN composition of words
- PTB (small), Text8 (Large)
- Comparable results, not impressive
  - Only 7-hop MemN2N-LM beats LSTM-LM

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**	-	-	-	-	-


hops
2
4
6
20
40
60
80
100
20
40
60
80
100
1
0.5
weight

Conclusion, Concerns, Future Work

- Good on QA, OKish on LM
- Weak (rather, non-overstrong) supervision, multi-hop
- Concern 1: Softmax and aggregation over all memories prohibitive when mem.space is large. Hashing
- Concern 2: Hops hyperparam. Not dynamic
- Concern 3: Memories not written/updated.
- 20 min MSR Talk by Sainbayar Sukhbaatar: https://www.youtube.com/watch?v=ZwvWY9Yy76Q