MEMORY NETWORKS

Presentation by @Y. Zhou

Facebook AI research

- Jason Weston et al. Memory Networks. In ICLR, 2015.
- S. Sukhbaatar, J. Weston et al. End-To-End Memory Networks. arXiv: 1503.08895, 2015. In NIPS, 2015.
- Large-scale Simple Question Answering with Memory Networks. arXiv: 1506.02075.
- …

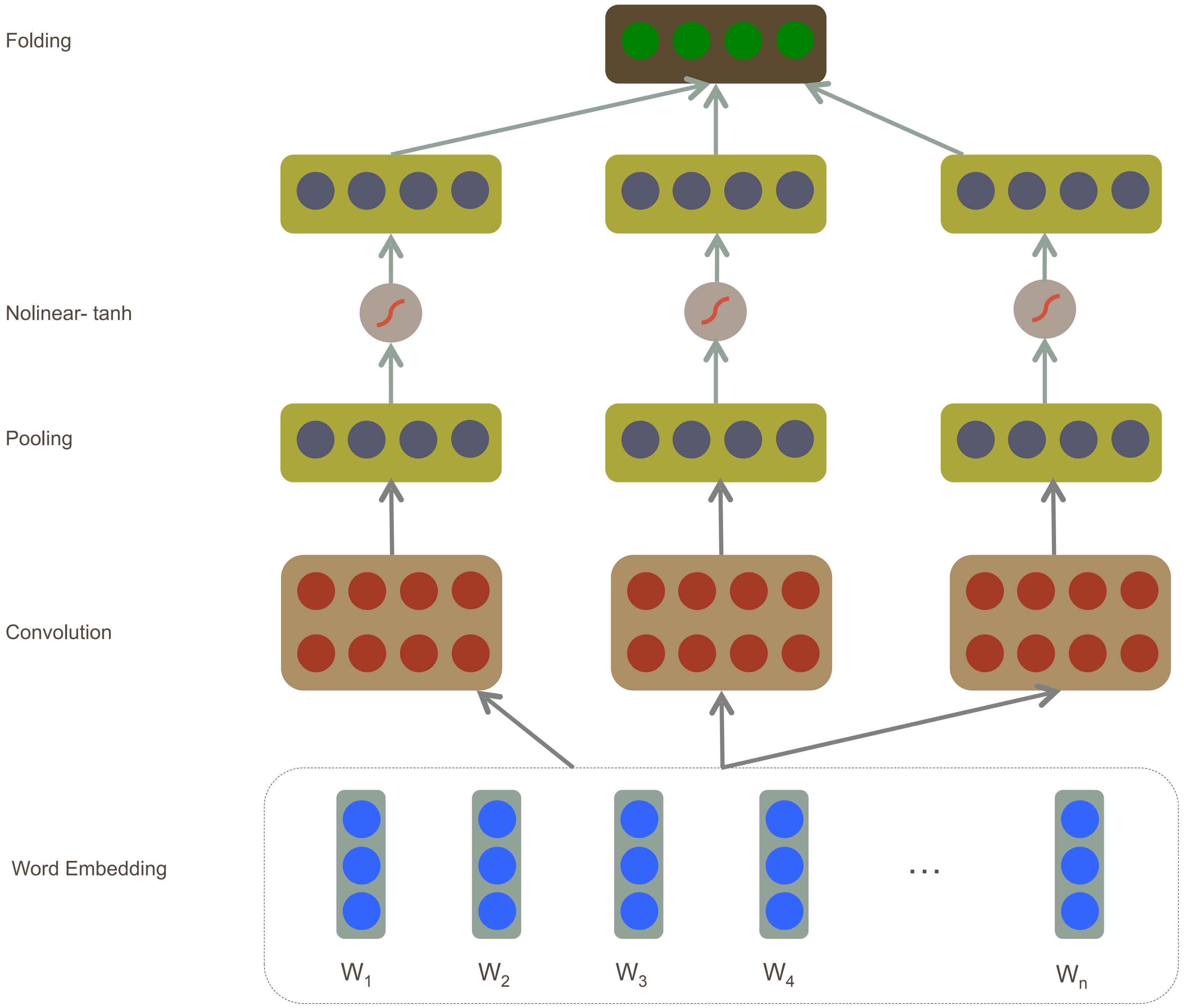
Overview

- Introduction to Neural Networks
- Single Layer of MemNN
- Multiple Layers
- Experiments
- Conclusion

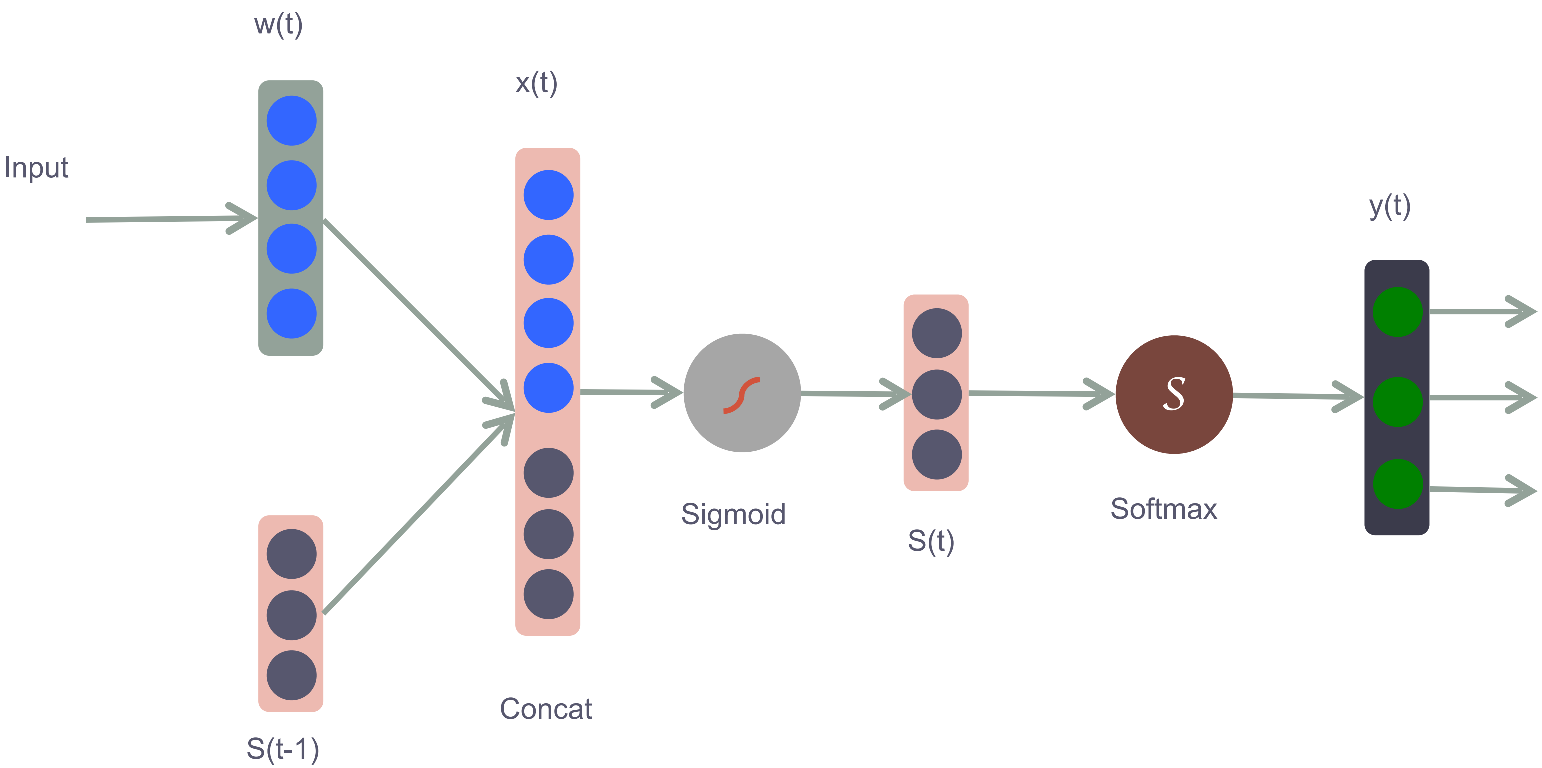
Introduction to Neural Networks

- Convolution NN for Sentence Representation
- Recurrent NN for Language Model or Translation
- LSTM for Language Model or Translation
- Recursive NN for Sentiment classification
- Attention-Based improve Recurrent NN
- Memory Networks for LM or Question Answering

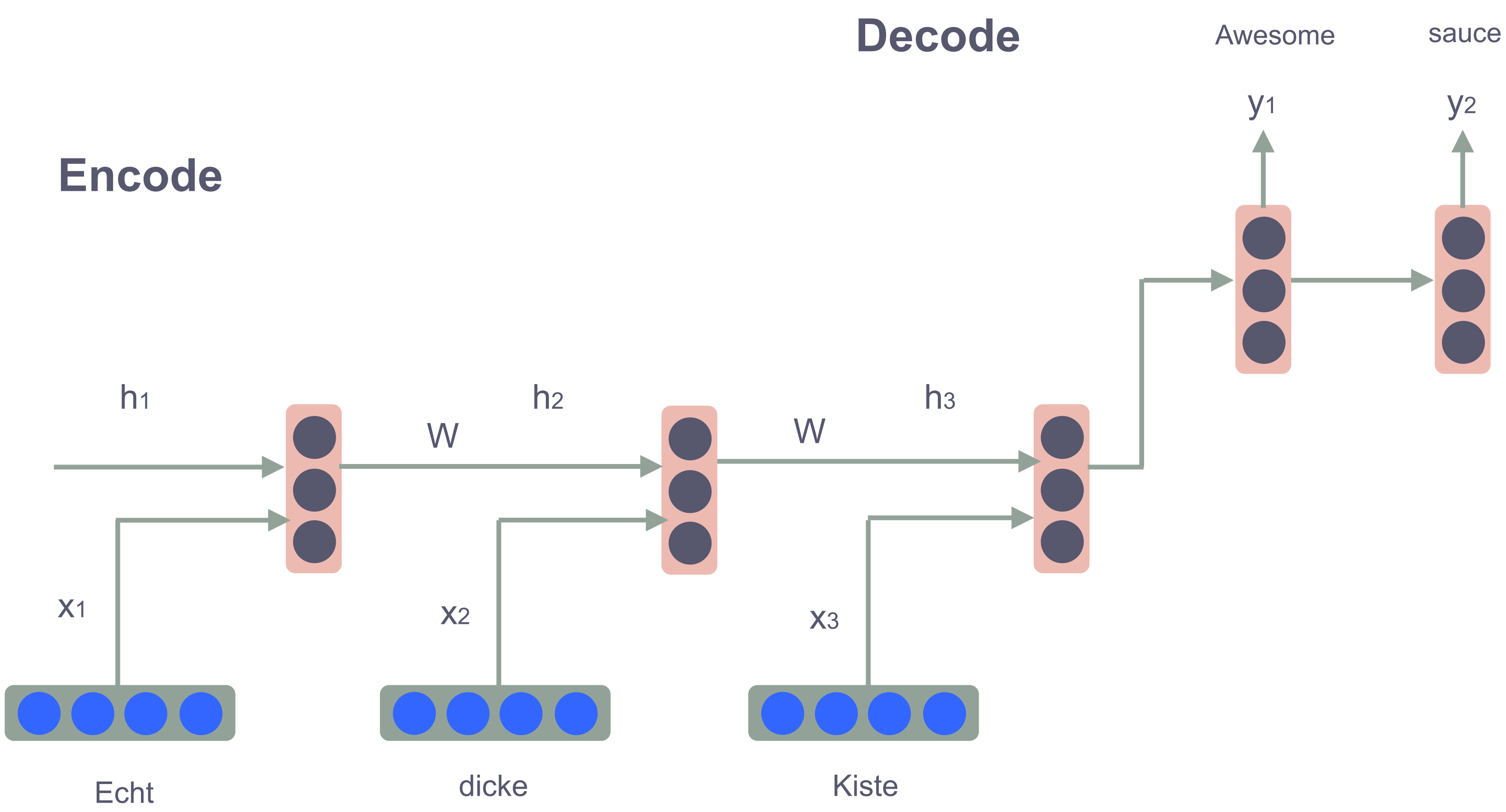
CNN


Folding
Nolinear- tanh
Pooling
Convolution
Word Embedding
W1
W2
W3
W4
...
Wn

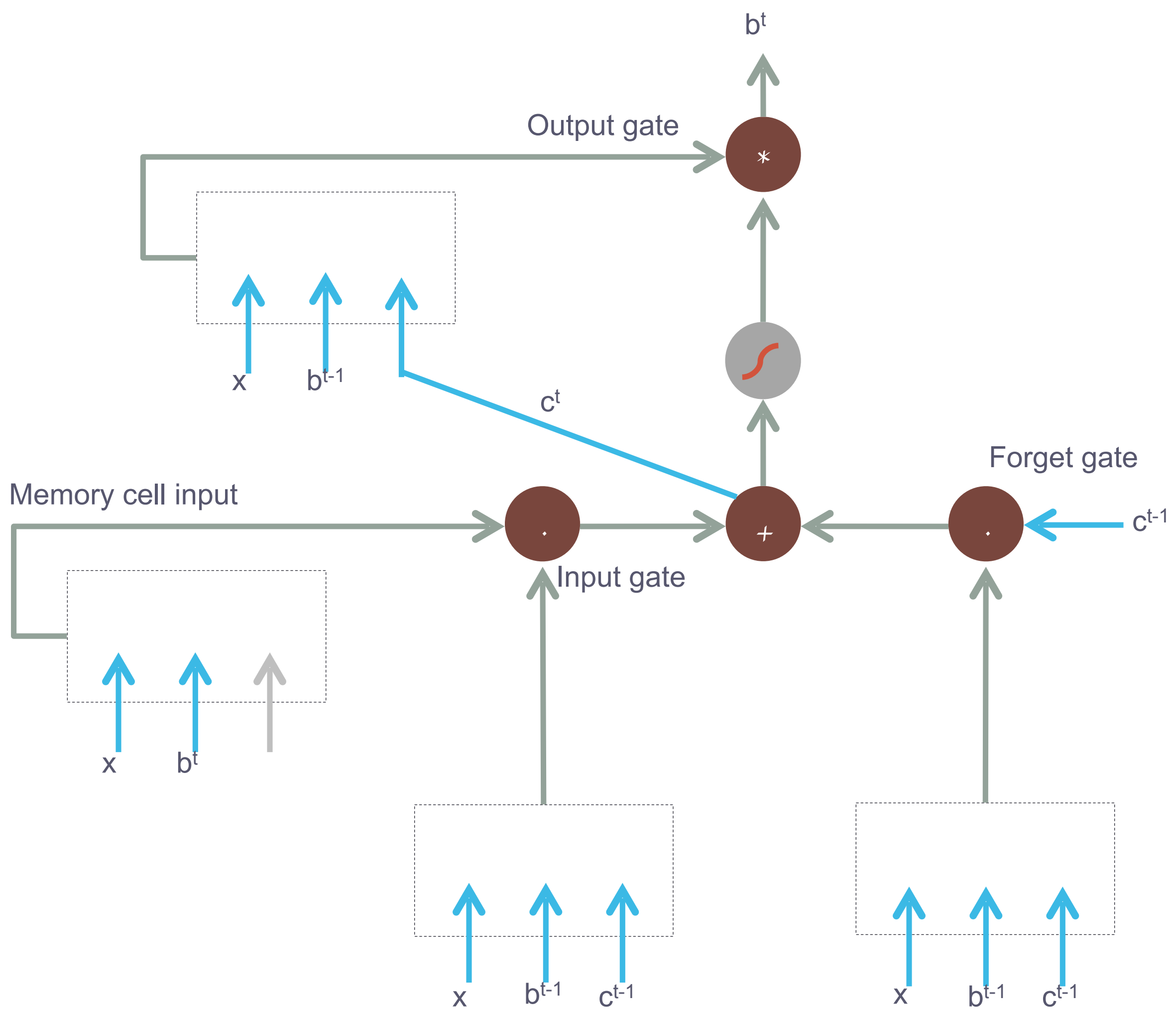
Recurrent Neural Network

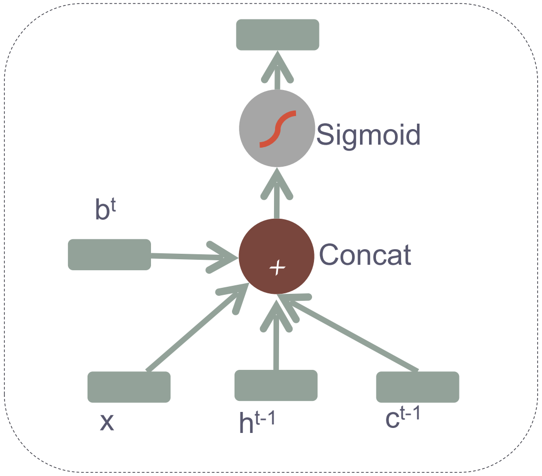


Recurrent Neural Network for MT



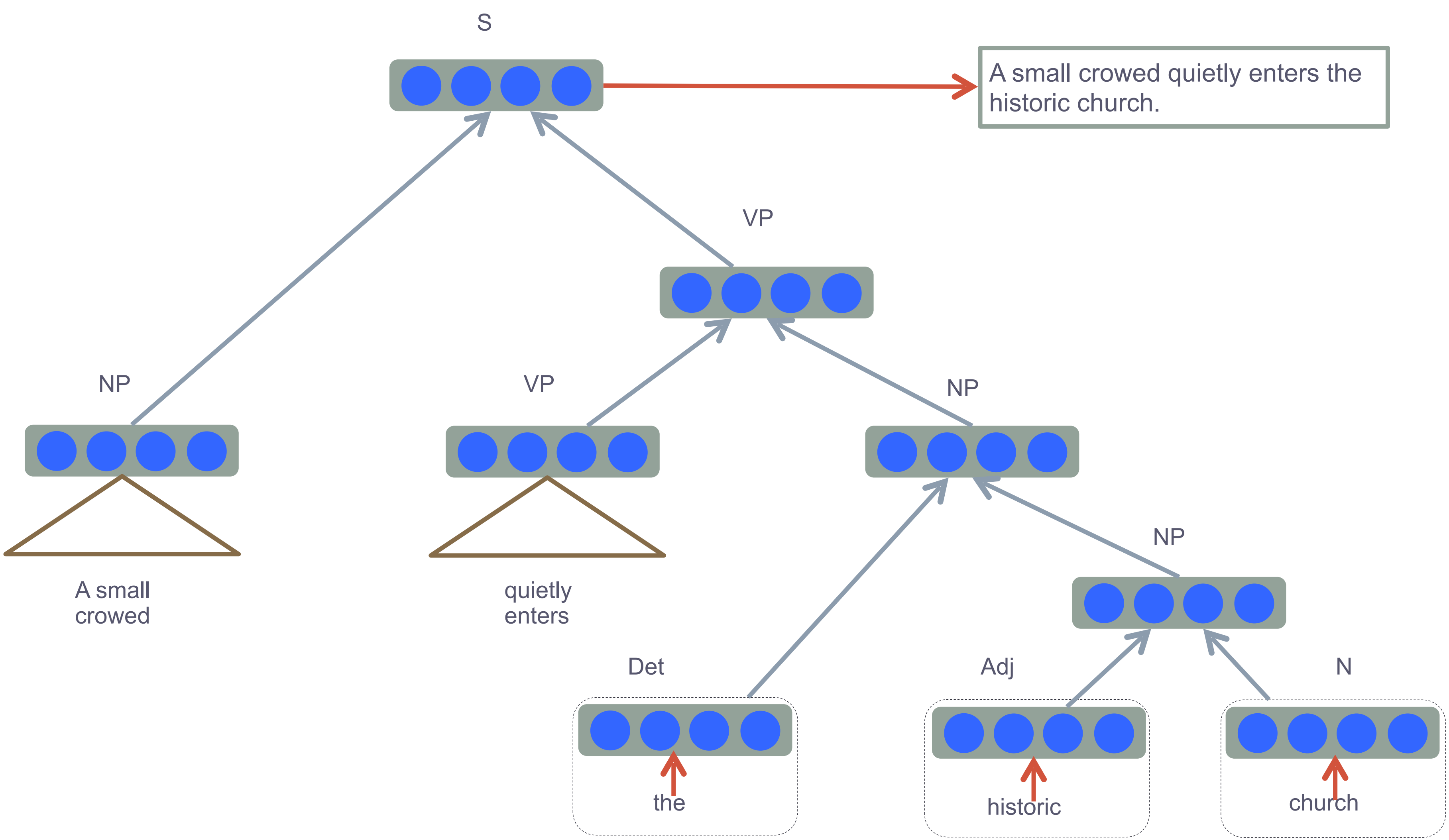
Long Short-Term Memory



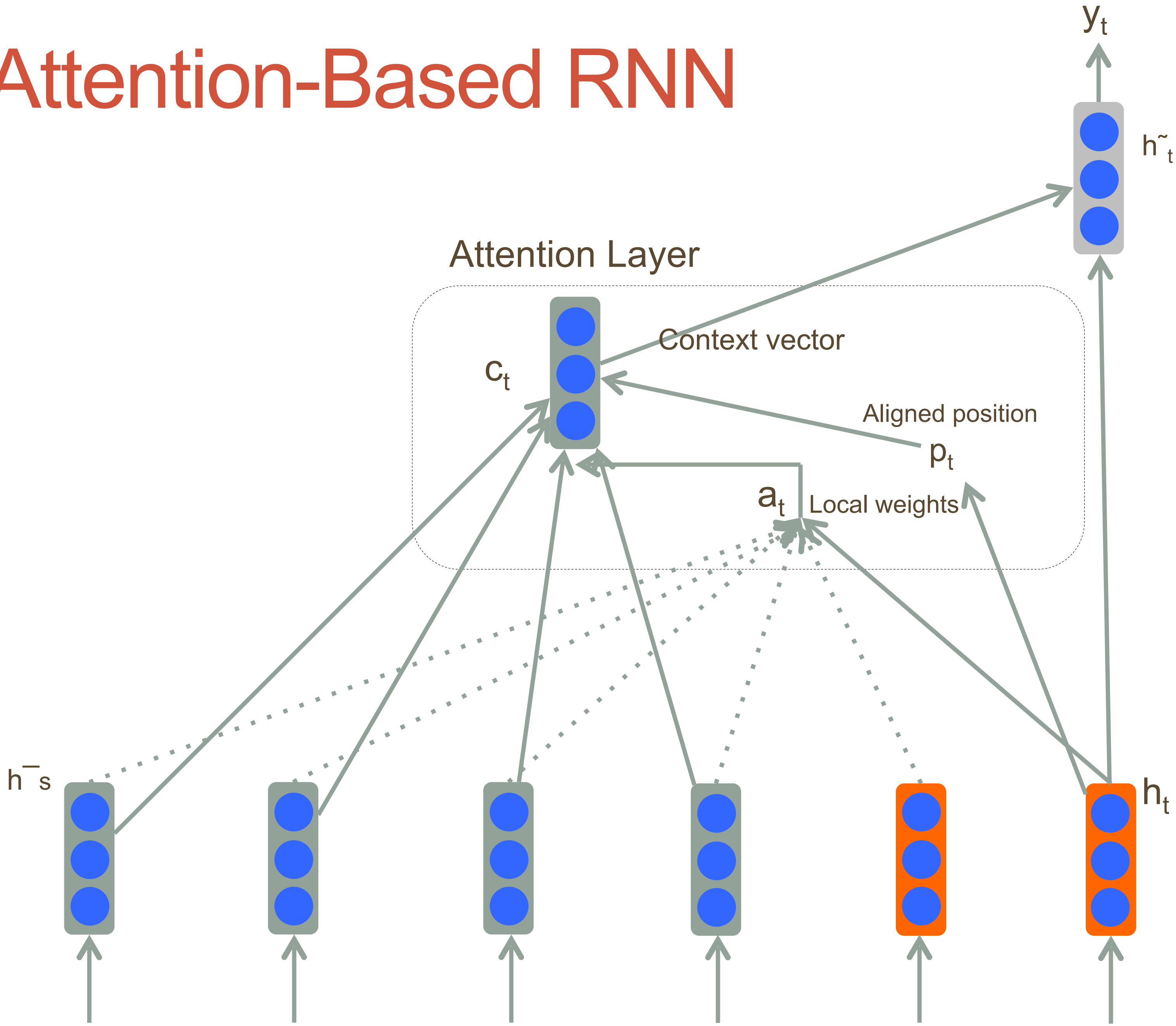


LSTM Sub-unit

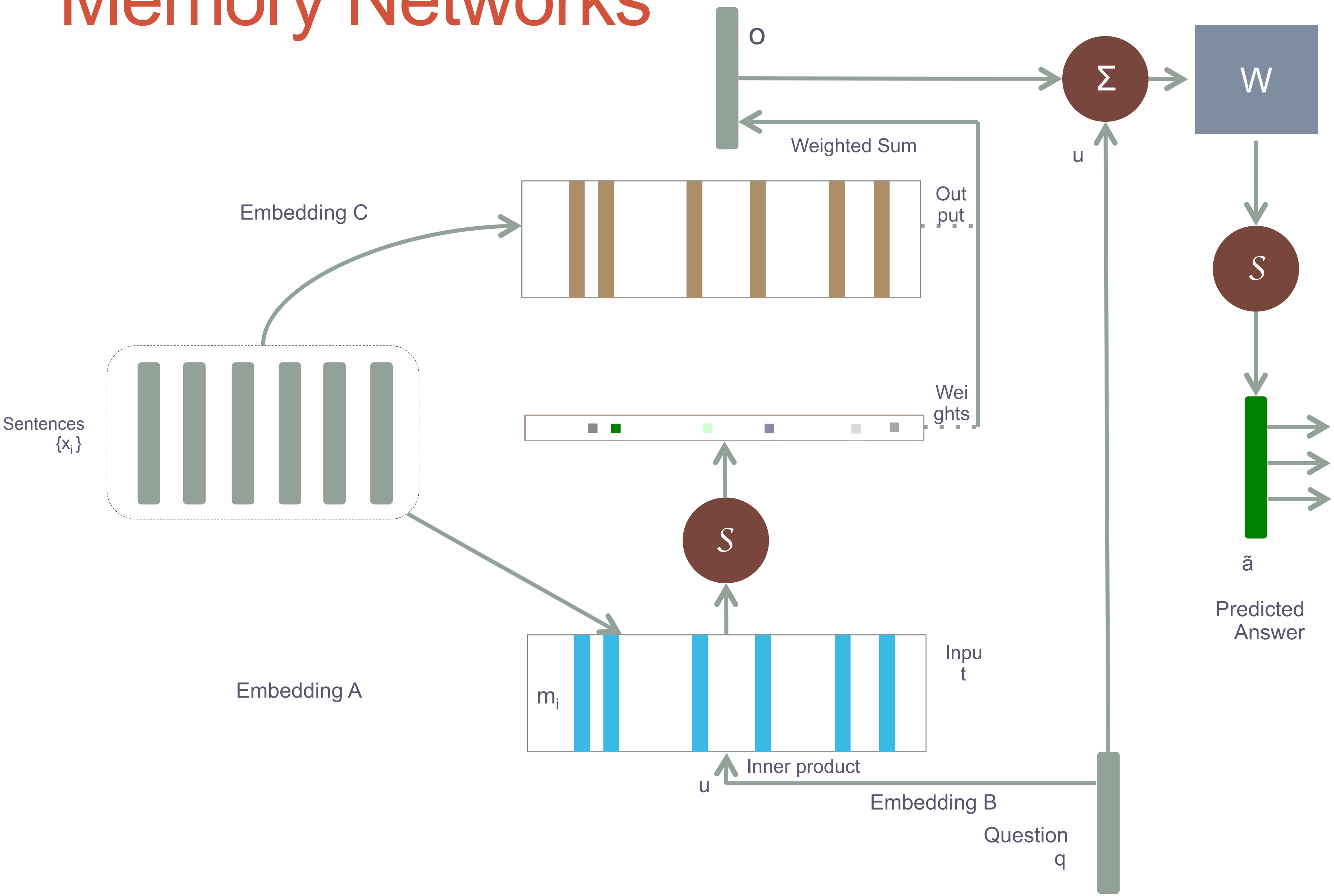
Recursive Neural Network



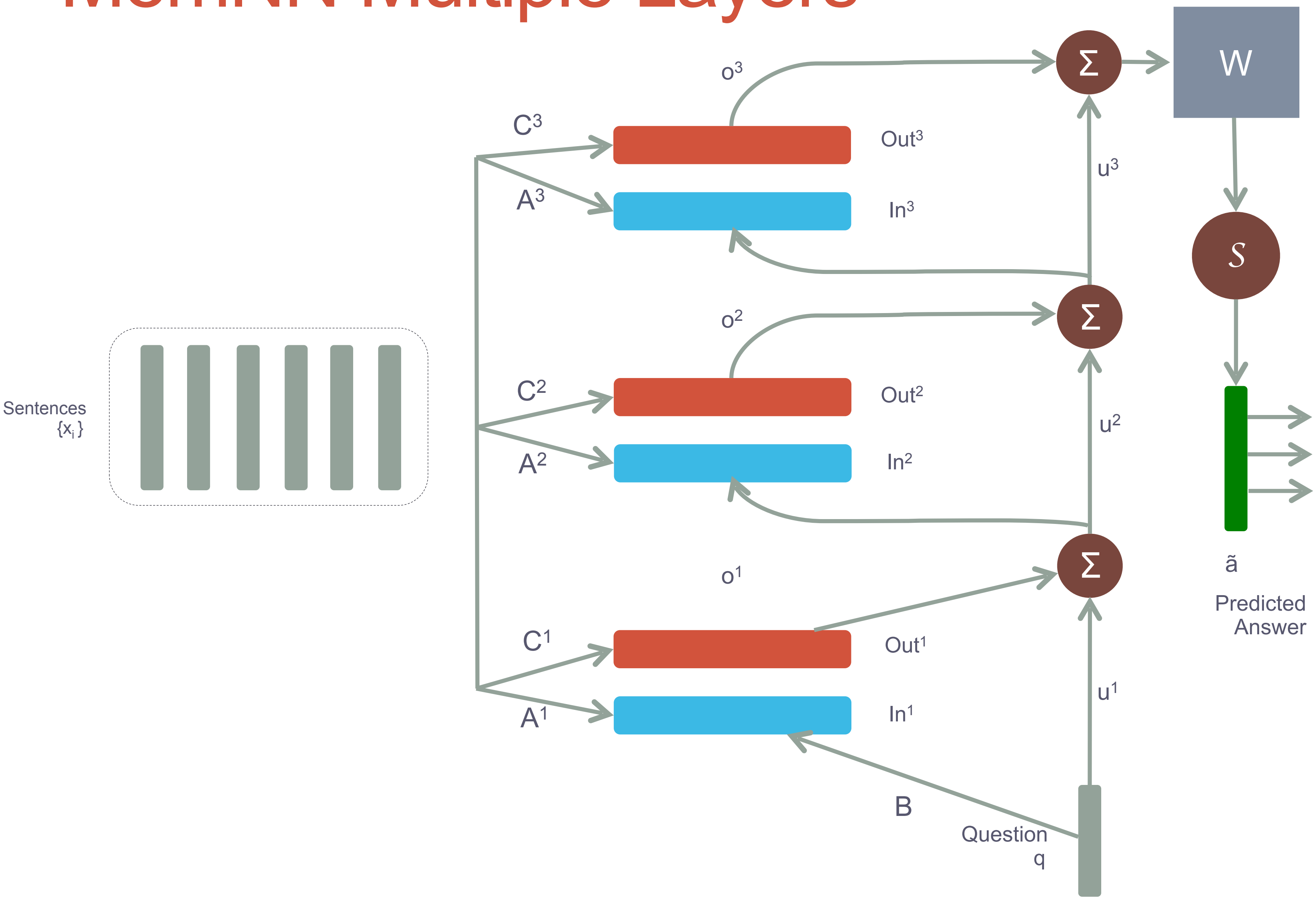
Attention-Based RNN



Memory Networks



MemNN Multiple Layers



Memory Networks

- Motivation: Long term dependencies in sequential data.
- Strategy: Store inputs to memory, read and write.
- Tasks:
  - Question Answering (Syntax) – Given documents and related questions, answering the questions
  - Language Model – Basic issue of Natural Language Processing

Details of MemNN

Temporal Encoding

- BoW (Bag of words)

$c_i = \Sigma_j Cx_{ij} + T_C(i)$

$c_i = \Sigma_j Cx_{ij}$

$o = \Sigma_j p_i c_i$

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

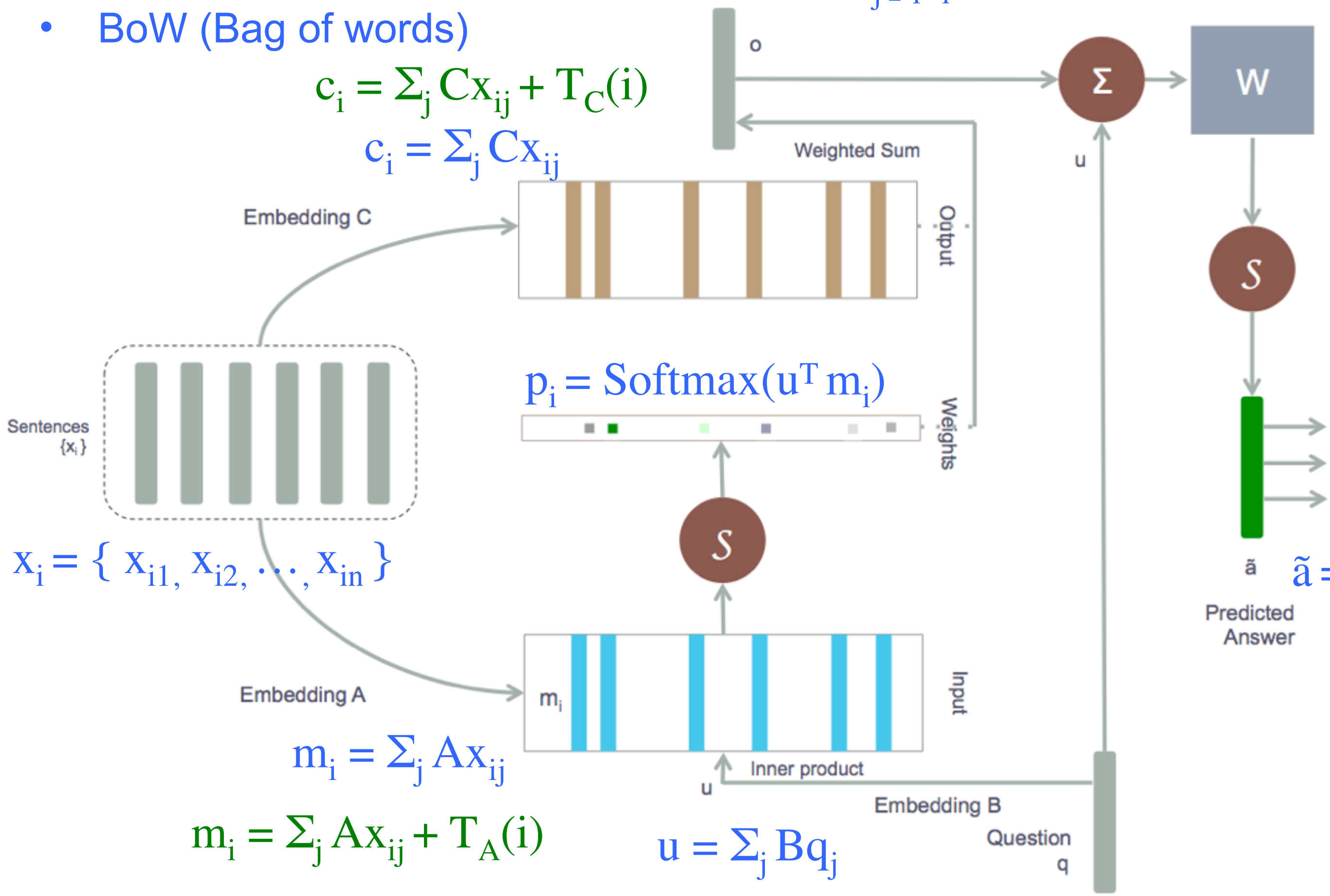
$x_i = \{ x_{i1}, x_{i2}, \ldots, x_{in} \}$

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

$m_i = \Sigma_j Ax_{ij}$

$m_i = \Sigma_j Ax_{ij} + T_A(i)$

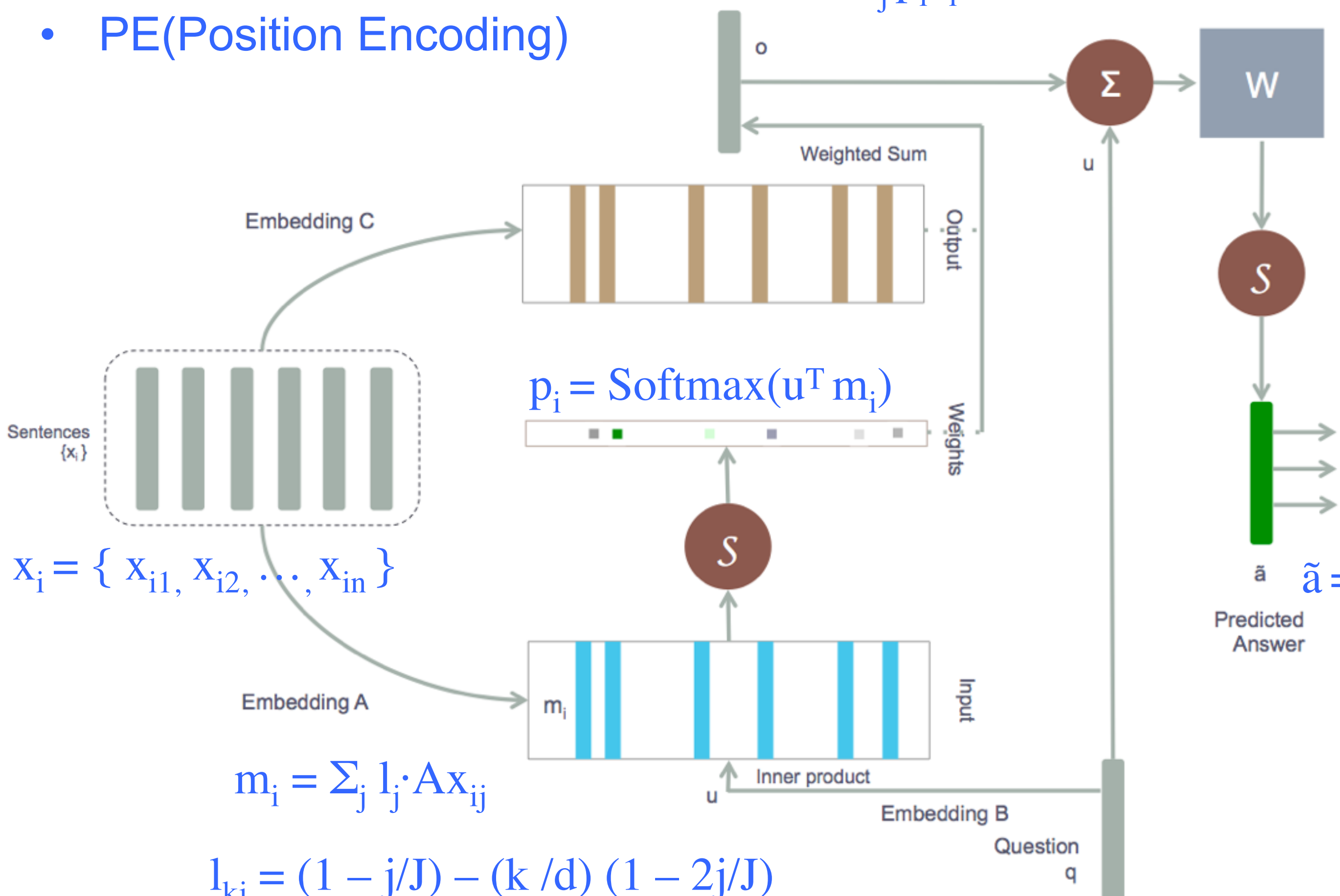
$u = \Sigma_j Bq_j$



Details of MemNN

- PE(Position Encoding)

$$o = \Sigma_j p_i c_i$$



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

$$x_i = \{ x_{i1}, x_{i2}, \ldots, x_{in} \}$$

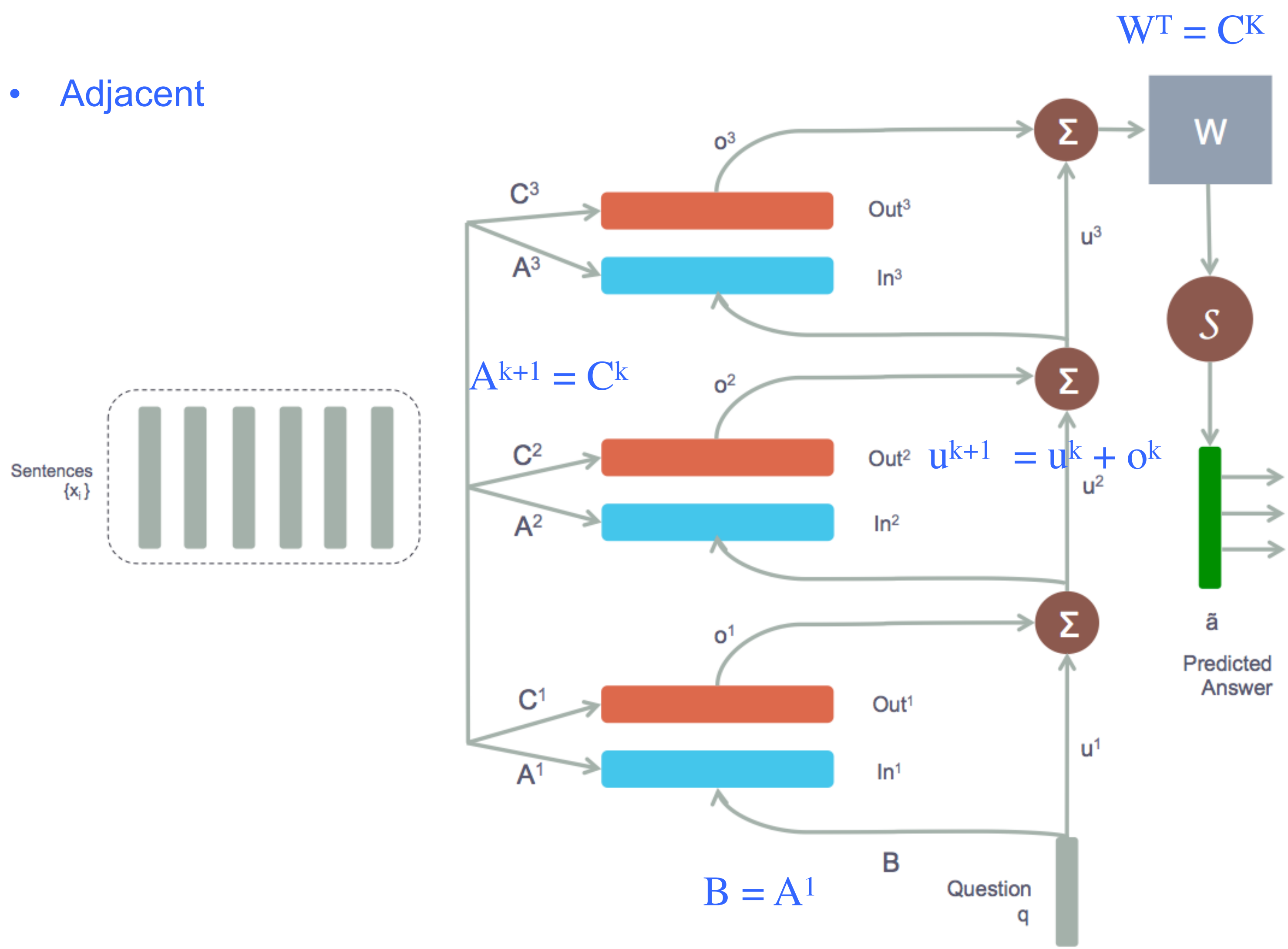
$$\tilde{a} = \mathrm{Softmax}(\, W (o+u) \,)$$

$$m_i = \Sigma_j l_j \cdot Ax_{ij}$$

$$l_{kj} = (1 - j/J) - (k/d)(1 - 2j/J)$$

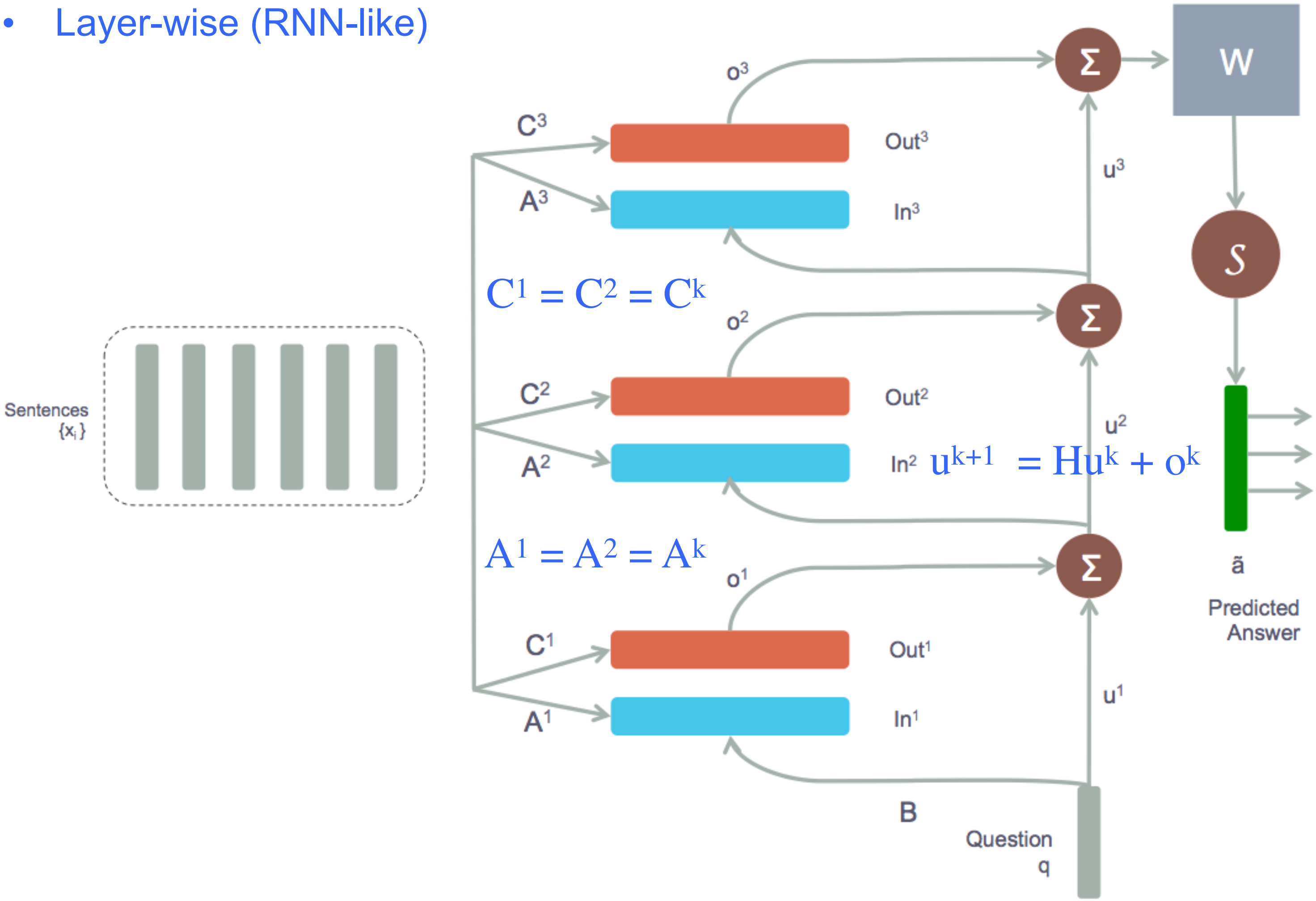
Details of Multiple MemNN

- Adjacent



Details of Multiple MemNN

- Layer-wise (RNN-like)



$C^1 = C^2 = C^k$

$A^1 = A^2 = A^k$

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

Experiments – QA tasks

	Baseline			MemN2N									
Task	Strongly Supervised MemNN [21]	LSTM [21]	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
1: 1 supporting fact	0.0	50.0	0.1	0.6	0.1	0.2	0.0	0.8	0.0	0.1	0.0	0.1
2: 2 supporting facts	0.0	80.0	42.8	17.6	21.6	12.8	8.3	62.0	15.6	14.0	11.4	18.8
3: 3 supporting facts	0.0	80.0	76.4	71.0	64.2	58.8	40.3	76.9	31.6	33.1	21.9	31.7
4: 2 argument relations	0.0	39.0	40.3	32.0	3.8	11.6	2.8	22.8	2.2	5.7	13.4	17.5
5: 3 argument relations	2.0	30.0	16.3	18.3	14.1	15.7	13.1	11.0	13.4	14.8	14.4	12.9
6: yes/no questions	0.0	52.0	51.0	8.7	7.9	8.7	7.6	7.2	2.3	3.3	2.8	2.0
7: counting	15.0	51.0	36.1	23.5	21.6	20.3	17.3	15.9	25.4	17.9	18.3	10.1
8: lists/sets	9.0	55.0	37.8	11.4	12.6	12.7	10.0	13.2	11.7	10.1	9.3	6.1
9: simple negation	0.0	36.0	35.9	21.1	23.3	17.0	13.2	5.1	2.0	3.1	1.9	1.5
10: indefinite knowledge	2.0	56.0	68.7	22.8	17.4	18.6	15.1	10.6	5.0	6.6	6.5	2.6
11: basic coreference	0.0	38.0	30.0	4.1	4.3	0.0	0.9	8.4	1.2	0.9	0.3	3.3
12: conjunction	0.0	26.0	10.1	0.3	0.3	0.1	0.2	0.4	0.0	0.3	0.1	0.0
13: compound coreference	0.0	6.0	19.7	10.5	9.9	0.3	0.4	6.3	0.2	1.4	0.2	0.5
14: time reasoning	1.0	73.0	18.3	1.3	1.8	2.0	1.7	36.9	8.1	8.2	6.9	2.0
15: basic deduction	0.0	79.0	64.8	24.3	0.0	0.0	0.0	46.4	0.5	0.0	0.0	1.8
16: basic induction	0.0	77.0	50.5	52.0	52.1	1.6	1.3	47.4	51.3	3.5	2.7	51.0
17: positional reasoning	35.0	49.0	50.9	45.4	50.1	49.0	51.0	44.4	41.2	44.5	40.4	42.6
18: size reasoning	5.0	48.0	51.3	48.1	13.6	10.1	11.1	9.6	10.3	9.2	9.4	9.2
19: path finding	64.0	92.0	100.0	89.7	87.4	85.6	82.8	90.7	89.9	90.2	88.0	90.6
20: agent's motivation	0.0	9.0	3.6	0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.2
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

Table 1: Test error rates (%) on the 20 QA tasks for models using 1k training examples (mean test errors for 10k training examples are shown at the bottom). Key: BoW = bag-of-words representation; PE = position encoding representation; LS = linear start training; RN = random injection of time index noise; LW = RNN-style layer-wise weight tying (if not stated, adjacent weight tying is used); joint = joint training on all tasks (as opposed to per-task training).

Experiments – QA tasks

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				

Figure 2: Example predictions on the QA tasks of [21]. We show the labeled supporting facts (support) from the dataset which MemN2N does not use during training, and the probabilities p of each hop used by the model during inference. MemN2N successfully learns to focus on the correct supporting sentences.

Experiments – 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**	-	-	-	-	-

Table 2: The perplexity on the test sets of Penn Treebank and Text8 corpora. Note that increasing the number of memory hops improves performance.

Experiments – LM

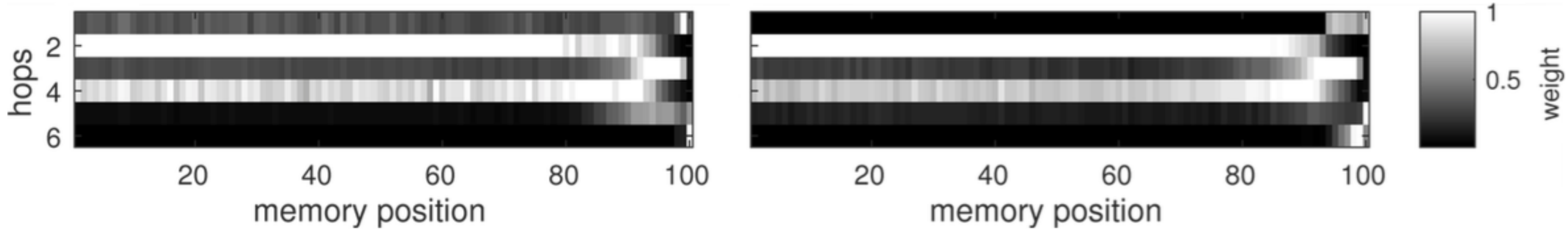


Figure 3: Average activation weight of memory positions during 6 memory hops. White color indicates where the model is attending during the k^{th} hop. For clarity, each row is normalized to have maximum value of 1. A model is trained on (left) Penn Treebank and (right) Text8 dataset.

Conclusion

- Deep learning for Natural Language Processing is booming
- End-To-End or Traditional features Composition, How to represent ?
- Embed structure information to neural networks
- More complicated network structure, learn to design
- …

Thanks!

Any question?