End-To-End Memory Networks

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, Rob Fergus (2015)

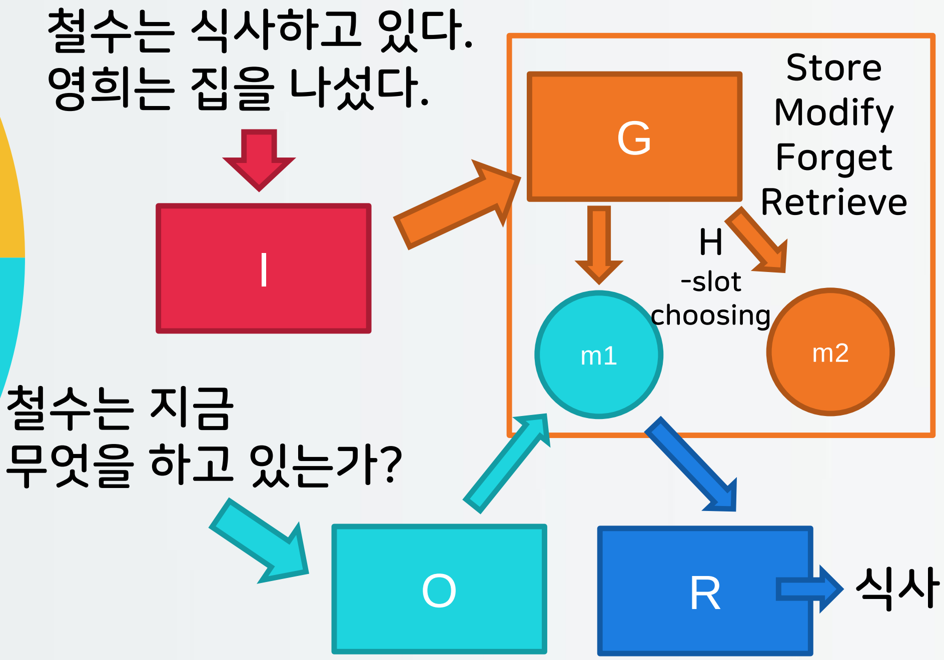
발표: 조 주 현

Contents

Motivation

Motivation

Memory Network



Motivation

RNNSearch & Attention Mechanism

Long Term Dependency Problem 해결 하는 아이디어

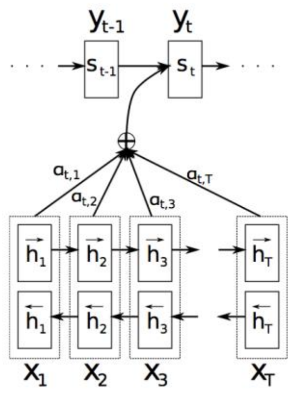


Figure 1: The graphical illustration of the proposed model trying to generate the t-th target word y_t given a source sentence $(x_1, x_2, \ldots, x_T)$.

RNN
- Multiple Hops 구조 통한 반복

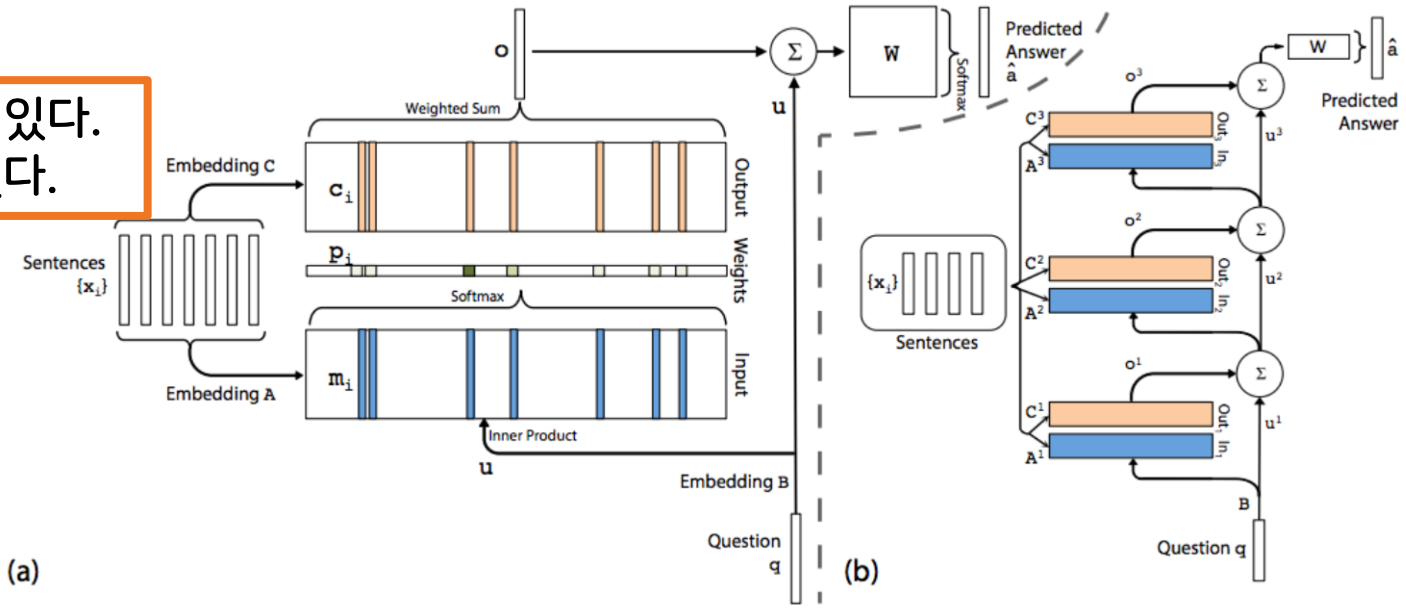
Attention
- MemoryNet에서
O function 의 역할을 대체

Base Model

Input & Output

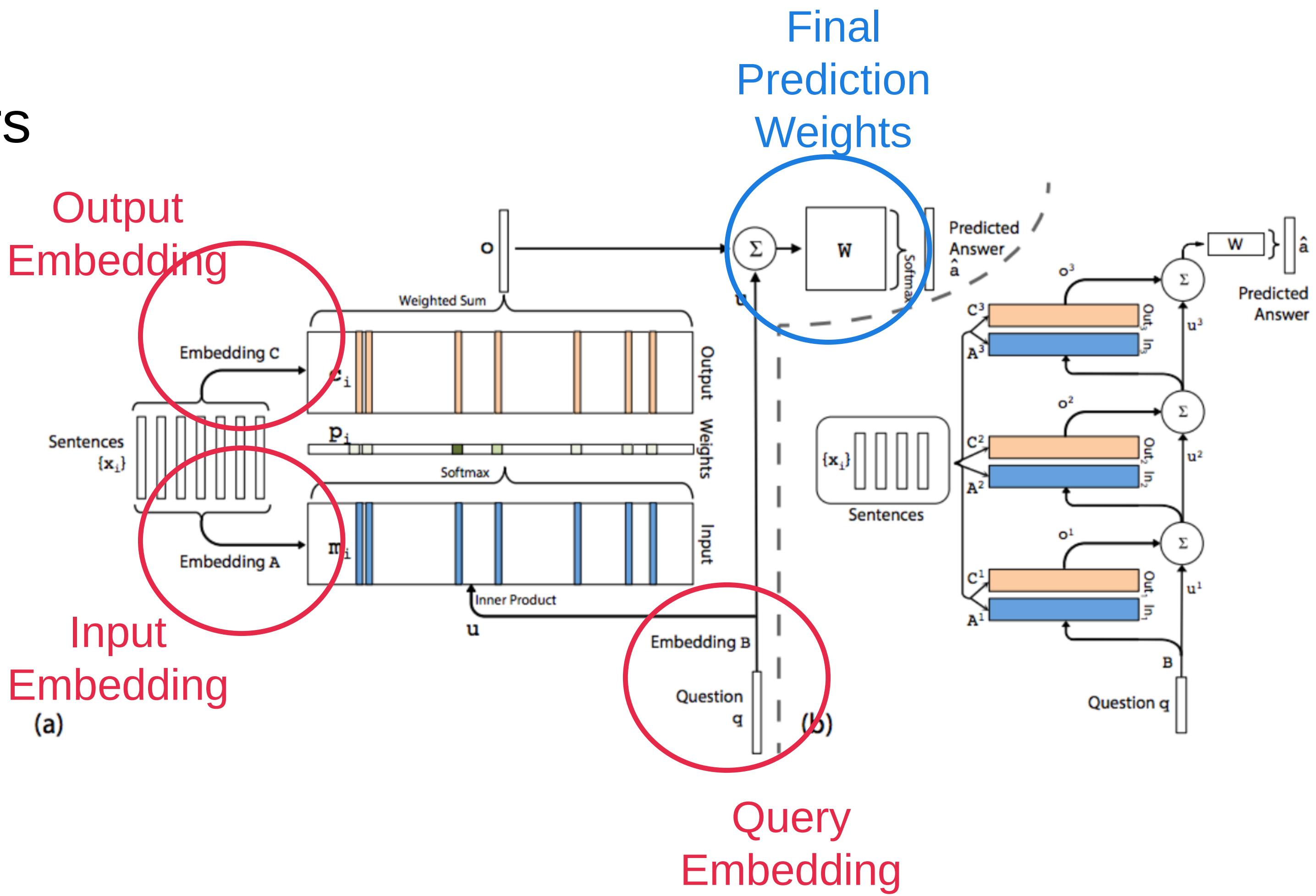
철수는 식사하고 있다.
영희는 집을 나섰다.

식사

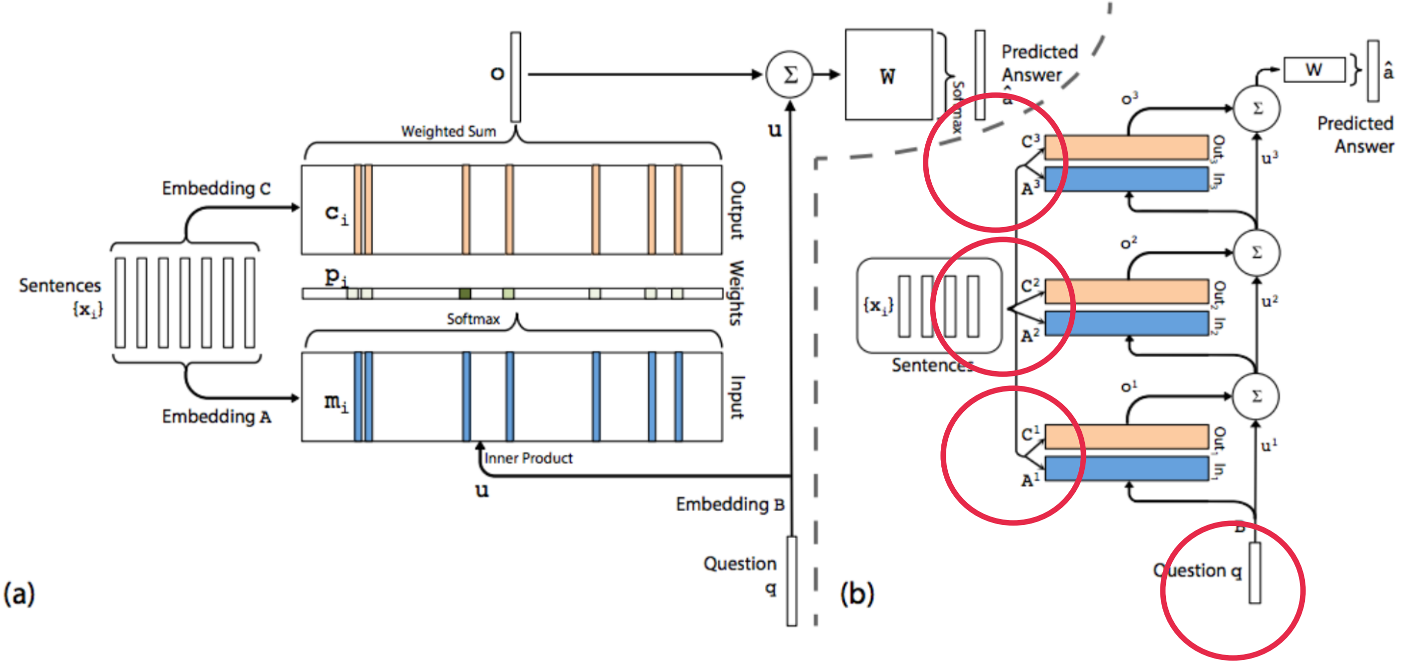


철수는 지금
무엇을 하고 있는가?

Trainable Parameters

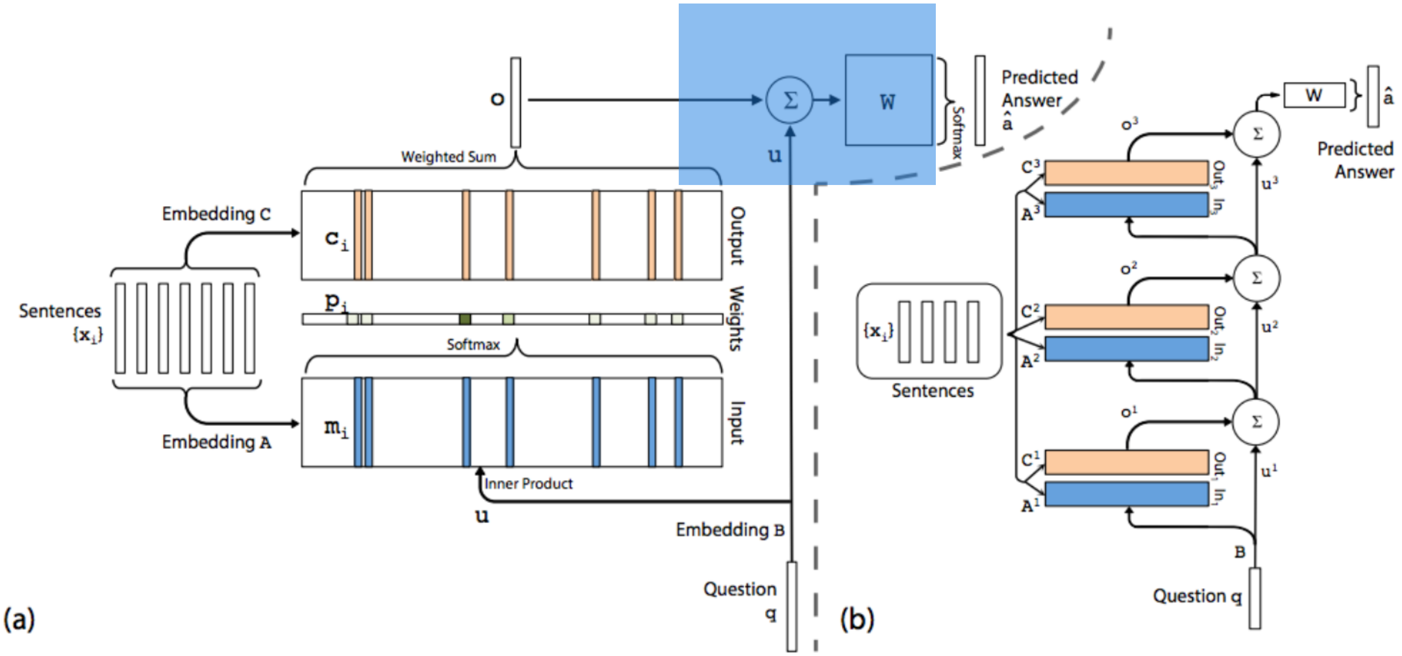


Multiple Hops



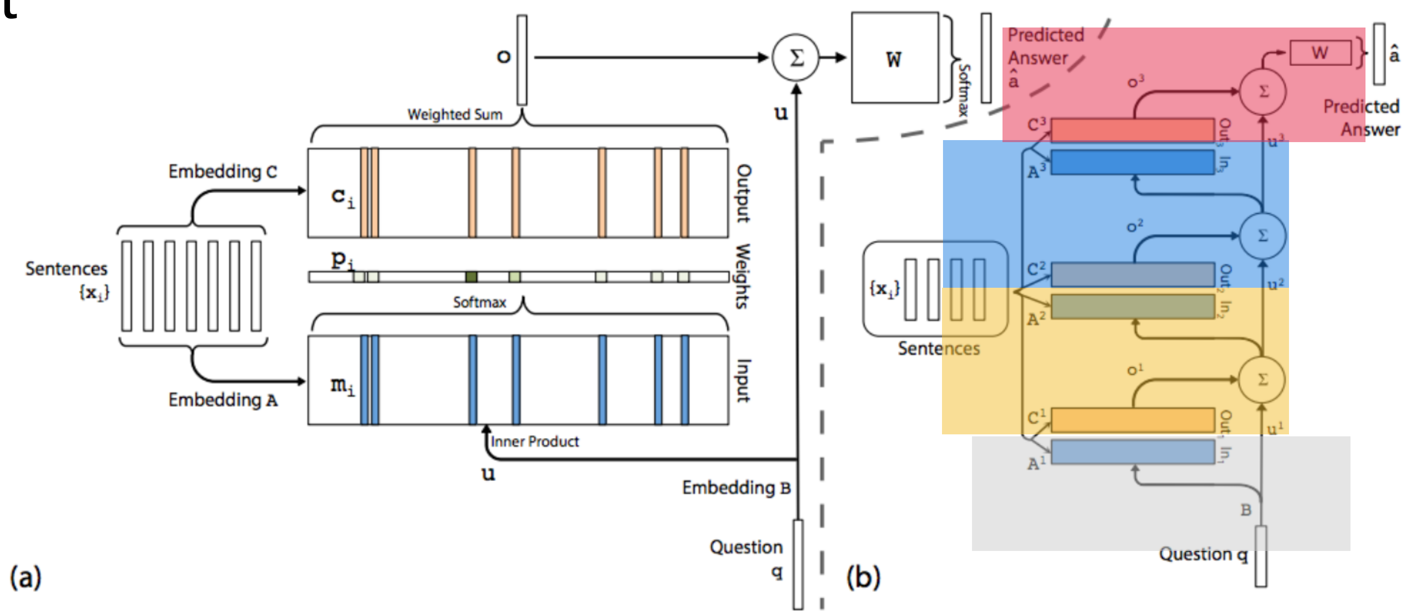
Other Tricks
1. Residual Connection

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



Other Tricks
2. Embedding Weight Tying
(1) Adjacent

$$B = A^1$$
$$A^{k+1} = C^k$$
$$C^k = W^T$$



유인: 학습 파라메터 수의 감소

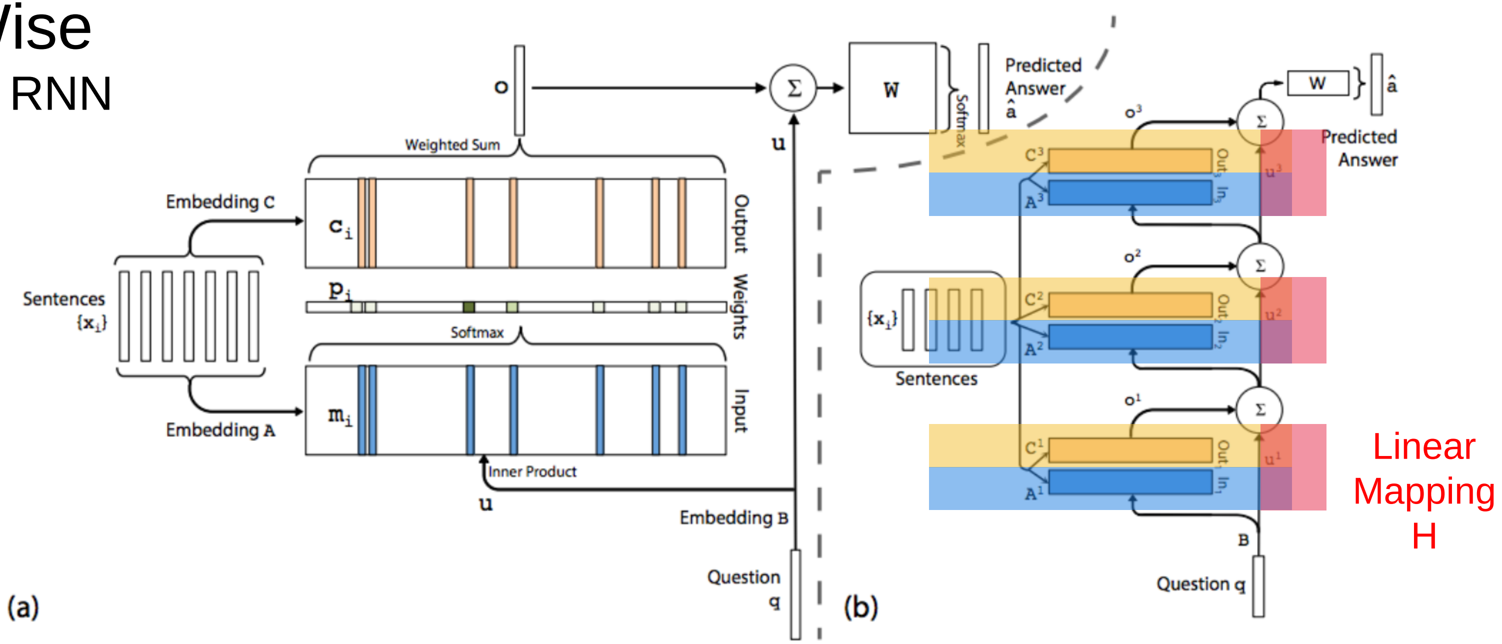
Other Tricks

2. Embedding Weight Tying

(2) Layer-Wise

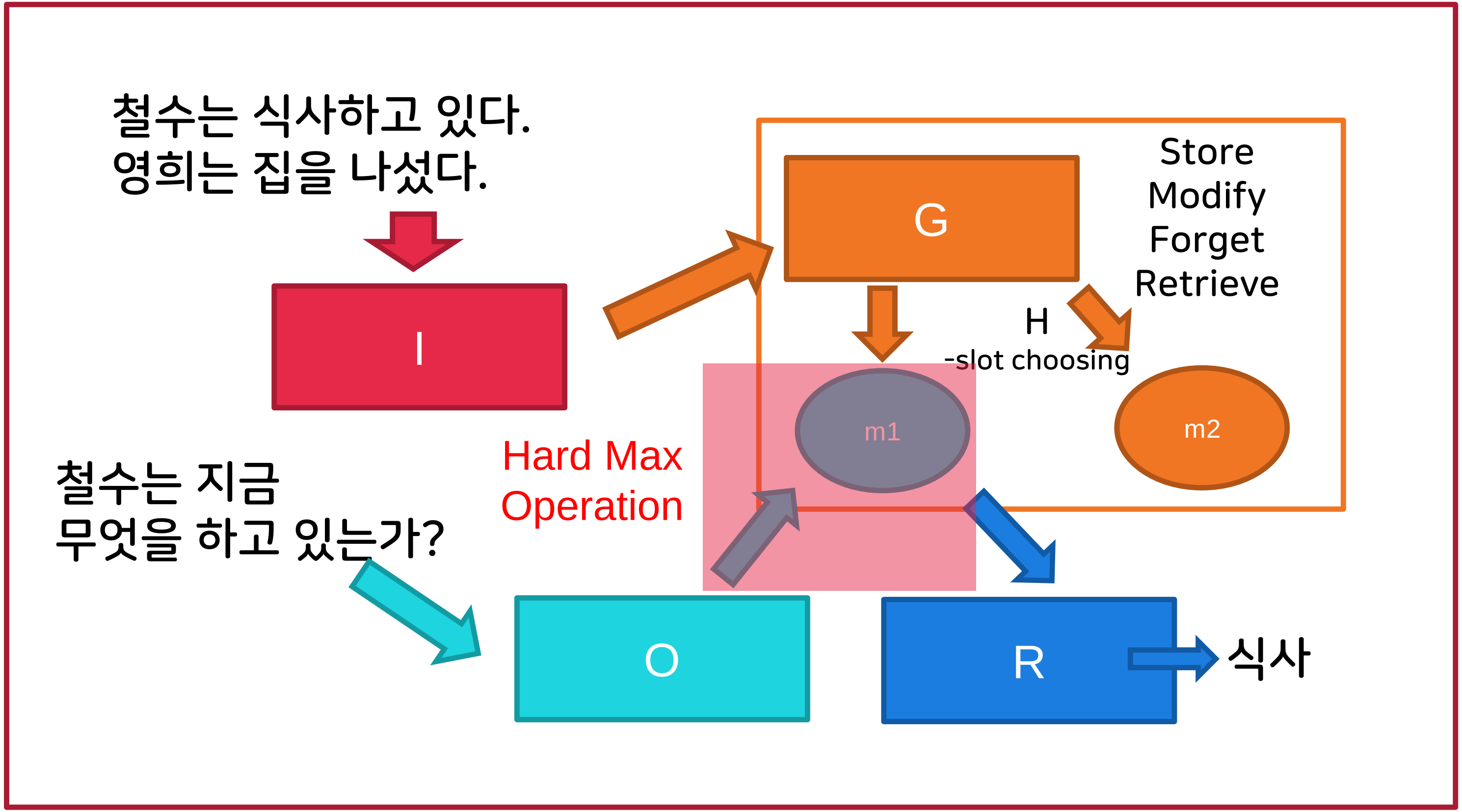
- Similar to RNN

$$A^1 = A^2 = \cdots = A^k$$
$$C^1 = C^2 = \cdots = C^k$$



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

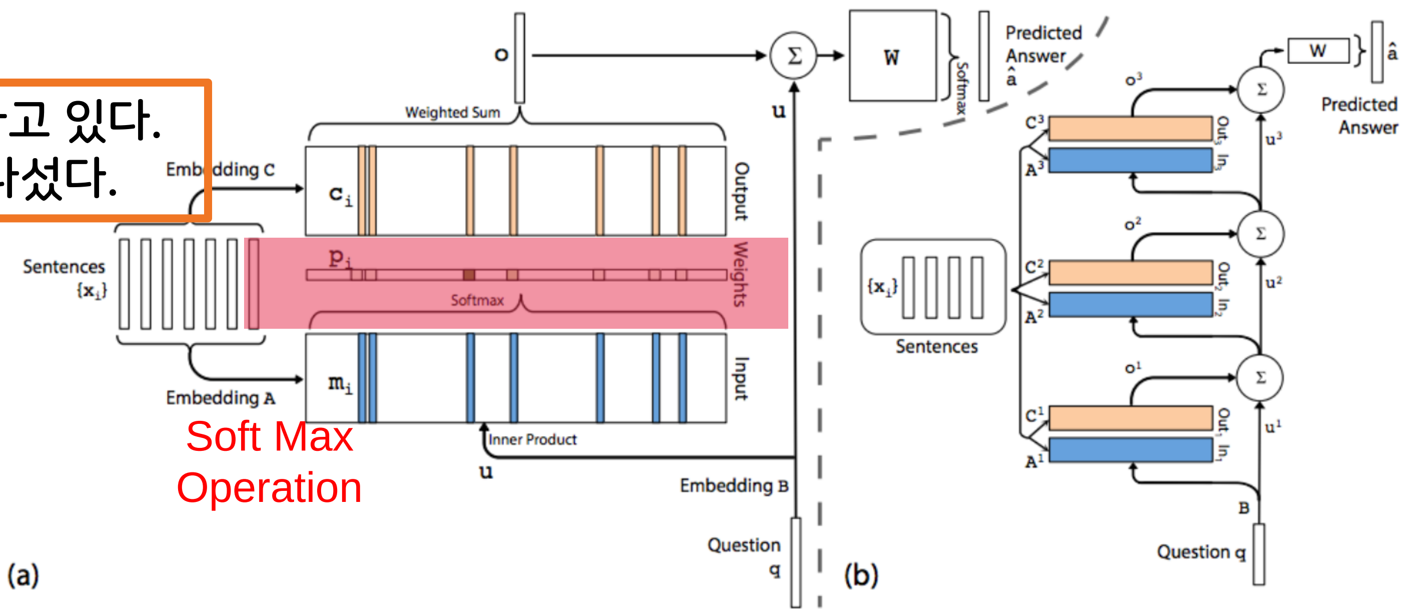
Comparison with Memory Network



Comparison with Memory Network

식사

철수는 식사하고 있다.
영희는 집을 나섰다.



Soft Max Operation

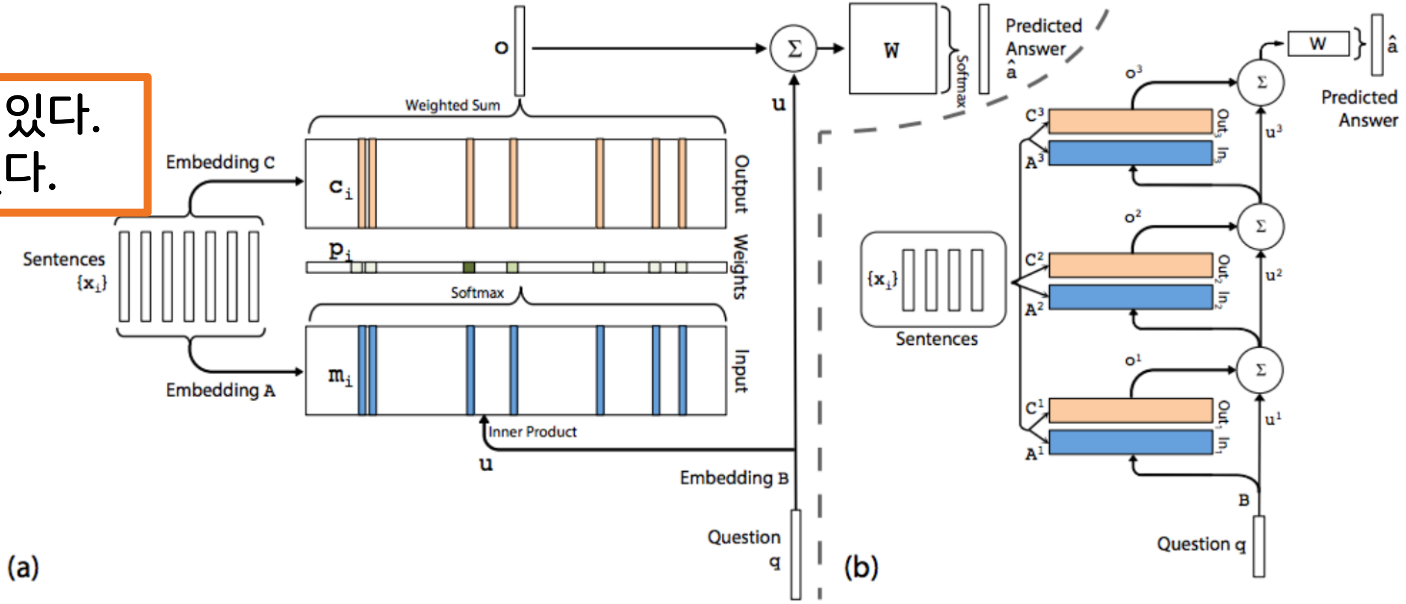
철수는 지금
무엇을 하고 있는가?

Task-Specific Model 1
Synthetic Q&A Model

Input & Output

철수는 식사하고 있다.
영희는 집을 나섰다.

식사



철수는 지금
무엇을 하고 있는가?

Dataset

Distractor

```
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
```

```
Mary journeyed to the den.
Mary went back to the kitchen.
John journeyed to the bedroom.
Mary discarded the milk.
Q: Where was the milk before the den?
A. Hallway
```

1) No Supporting Subset

2)
20 QA Tasks, I sentences(I < =320)
Question q, Answer a

Model Details

Sentence Representation

Bag-of-Words(BoW)

$m_i = \sum_j Ax_{ij}$
Sum of word embeddings
단순히 단어마다 임베딩을 적용하여 더함.

Positional Encoding

$m_i = \sum_j l_j \cdot Ax_{ij}$
$l_{kj} = (1 - j/J) - (k/d)(1 - 2j/J)$
order of words affects m_i
단어 임베딩을 하되, 각 위치마다 다른 가중치를 곱해 줌.

Temporal Encoding

Temporal information

Ex)
보섭은 공부를 하고 있다.
(30분 후) 보섭은 공부를 접었다.
$m_i = \sum_j Ax_{ij} + T_A(i)$
Sentence 간의 positional encoding

Reverse ordering

Question과 가까운 문장이 앞에 오도록.

Random Noise

Learning time invariance

"dummy" memories to regularize T_A
Add 10% of empty memories to the stories

Training Details – 특별한 것만

Selective L2 Norm

Gradient 의 L2 Norm이 특정 수치(ex/ 40)를 넘으면 ,

그 수치가 되도록 스칼라로 나누어 줌

Linear Start training

마지막 final prediction layer을 제외한 모든 부분의 softmax를 비활성화 한 채로 학습을 시작.

이후 validation loss가 줄어들지 않는 순간부터 softmax 활성화.

Jointly Training

각 Task마다 다른 모델을 사용하지 않고, 동일한 모델을 학습시킴.

Results

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				

Results

1. Position Encoding 해라
2. Linear Start 해라 – local minima avoiding effect
3. Random Noise 넣어라 –small but consistent boost in performance
4. Joint Train 해라
5. Multiple hop 써라
0. 기존 모델보다 좋다

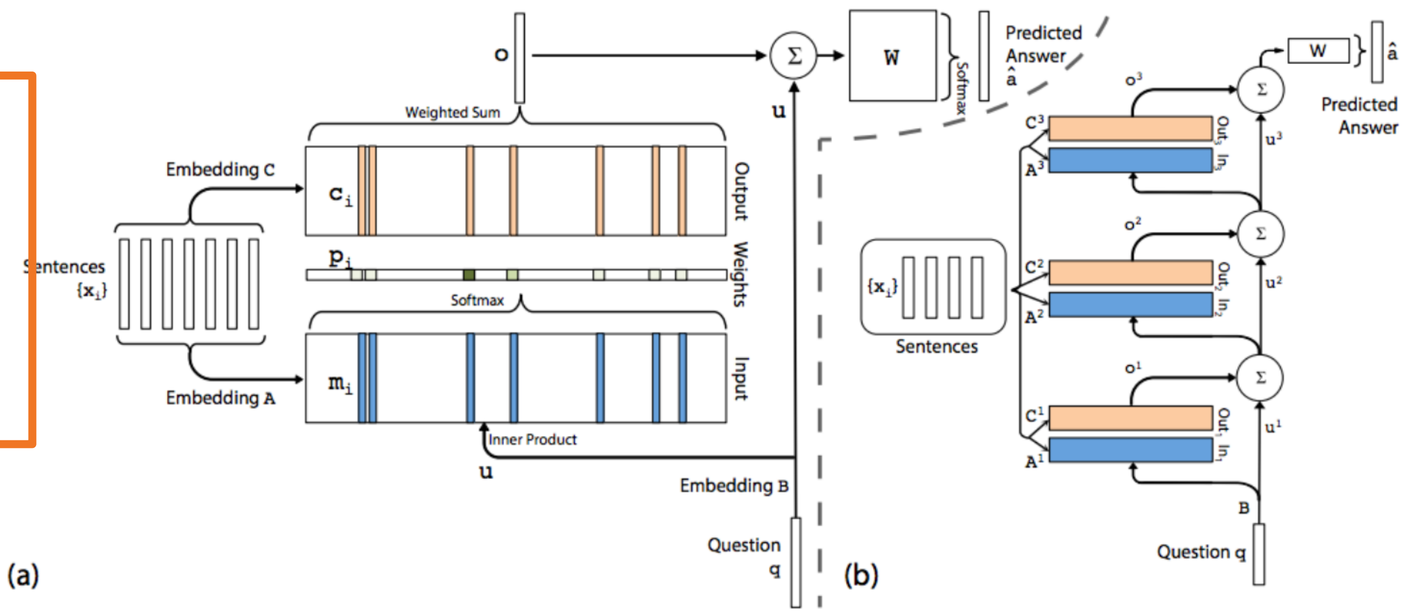
	Baseline			MemN2N								
Task	Strongly Supervised MemNN [22]	LSTM [22]	MemNN WSH	BoW	PE	PE LS	PE LS RN	1 hop PE LS joint	2 hops PE LS joint	3 hops PE LS joint	PE LS RN joint	PE LS LW joint
1: 1 supporting fact	0.0	50.0	0.1	0.6	0.1	0.2	0.0	0.8	0.0	0.1	0.0	0.1
2: 2 supporting facts	0.0	80.0	42.8	17.6	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

Task-Specific Model 2
Language Model

Input & Output

동해물
-과
백두산
-이
마르
-고

"닭"



Constant vector 0.1

Training Details – 특별한 것만

ReLU

Apply ReLU operations to half of the units in each layer

Embedding weight tying

RNN과 유사한 구조가 됨
(layer-wise weight sharing)

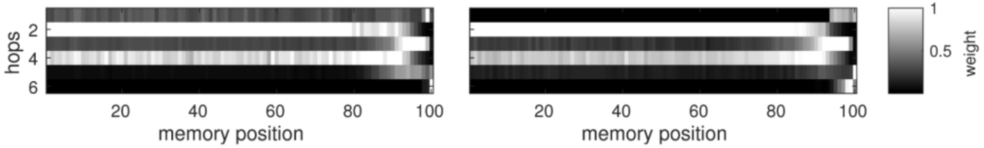
Temporal embedding approach

Word 단위로 input이 들어가므로, BoW, Linear mapping representation이 사용되지 않음.

대신, Temporal embedding을 도입하여 위치정보 사용.(개인 의견)

Results

1. Lower Perplexity를 얻었다.
2. 심지어 parameter 수도 훨씬 적다(RNN, LSTM등에 비해)
3. Hops마다 보는 곳이 다르다.(Similar to N-Gram effect)



	Penn Treebank					Text8				
Model	# of hidden	# of hops	memory size	Valid. perp.	Test perp.	# 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

Conclusion & future work

기여

1. Supporting fact에 대한 정보가 제공되어야 하는, 기존 memory network보다 더 범용적이다
2. 비슷한 Supervision 수준의 다른 모델에 비해 경쟁력이 있다.(Language model 등에서)(2015년 당시에..)

한계 및 과제

1. Strong Supervision이 제공된 기존 memory network 보다 Outperform 하지는 않음.
2. 몇몇 1k QA task(small trainset case)에 실패함
3. Larger memory가 필요한 태스크를 위해 scale up 하기에는 난항.(multiscale notion of attention, Hashing 필요)

Thank you