Tasks and Architectures for Language Understanding and Dialogue with Memory

Jason Weston

Facebook AI Research

Collaborators: A. Bordes, Y. Boureau, S. Chopra, J. Dodge, R. Fergus, A. Fisch, A. Gane, M. Henaff, F. Hill, A. Joulin, Y. LeCun, B. van Merriënboer, J. Li, T. Mikolov, A. Miller, A.M. Rush, S. Sukhbaatar, A. Szlam, X. Zhang

Intelligent Learners: Requirements

Some things we require:

- **Memory**: *long and short-term, knowledge*
- **Reasoning**: *both logical & not + commonsense*
- **Indirect Supervision**: *no labels, almost no rewards*
- **Composition**: *of learnt functions*
- **Transfer**: *between tasks*
- **Data Efficiency**: *most NNs are data hungry*
- **Planning**: *long term goals*

Intelligent Learners: Requirements

Some

- **Mer**
- **Rea**
- **Indi**
- **Con**
- **Tran**
- **Dat**
- **Plan**

Our Strategy:

1. Pick a far off goal: ***intelligent dialogue agent***
2. Identify subgoals (requirements).
3. Choose tasks to evaluate requirements; or design them if they don't exist.
4. Find/innovate models to solve them.
5. Iterate 2, 3 & 4 until we reach our goal?

Intelligent Learners: Requirements

Some things we require:

- **Memory**: *long and short-term, knowledge*
- **Reasoning**: *both logical and not*
- **Indirect Supervision**: *no labels, almost no rewards*
- **Composition**
- **Transfer**
- **Data Efficiency**
- **Planning**

Related Problems?

Intelligent Learners: Requirements

Som

- Me
- Re
- In
- Co
- Tr
- Da
- Pl

<u>Why Dialogue?:</u>

- Compared to vision "perception" (e.g. words) is simpler.
- But structure (e.g. sentences) is still complex.
- Accesses the **reasoning** problem more directly.
- Clear uses of **memory,** e.g. past knowledge /conv. history.
- **Indirect supervision** from utterance+response is powerful?
- Many different tasks can be posed in one format: dialogue!
  Good for measuring **transfer**, **composition**, etc.

Memory-Augmented Networks

(that I've worked on)

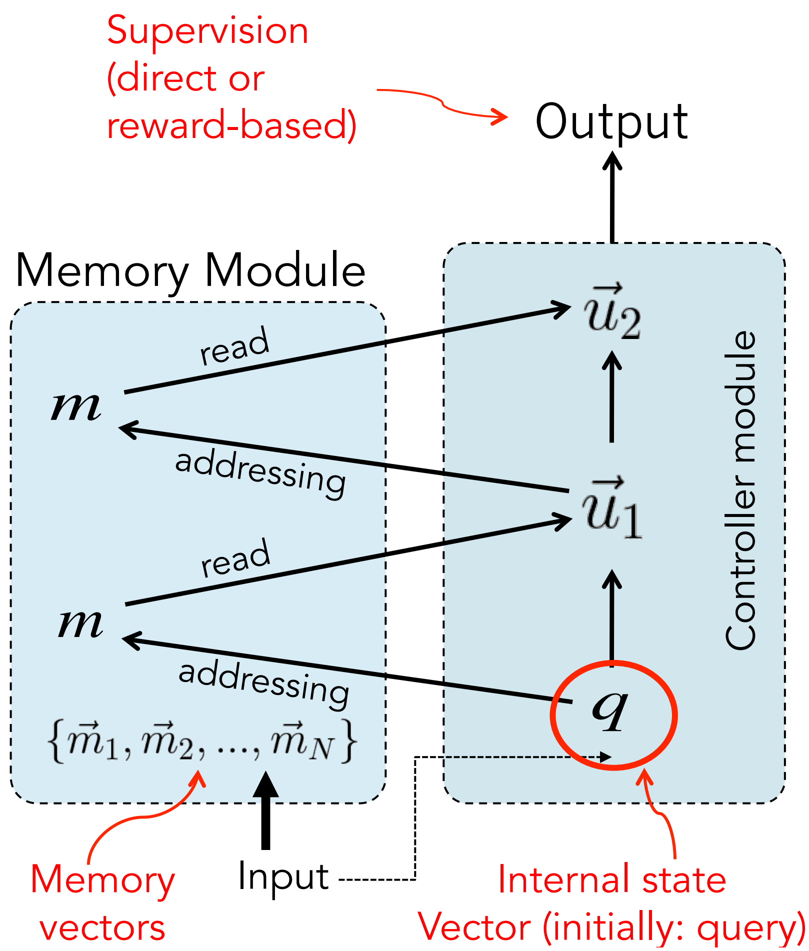
- Memory Network (Weston et al,'14)
- End-To-End Memory Network (Sukhbaatar et al., '15)
- Key-Value Memory Network (Miller et al., '16)
- Forward Prediction Memory Network (Weston, '16)
- Recurrent Entity Network (Henaff et al., '16)

Some related models: NTM (Graves et al, '14), Stack RNNs (Joulin & Mikolov, '15, Grefenstette et al, '15), Dynamic Mem. Nets (Kumar et al., '15), DNC (Graves et al., '16), …

Memory Network Models

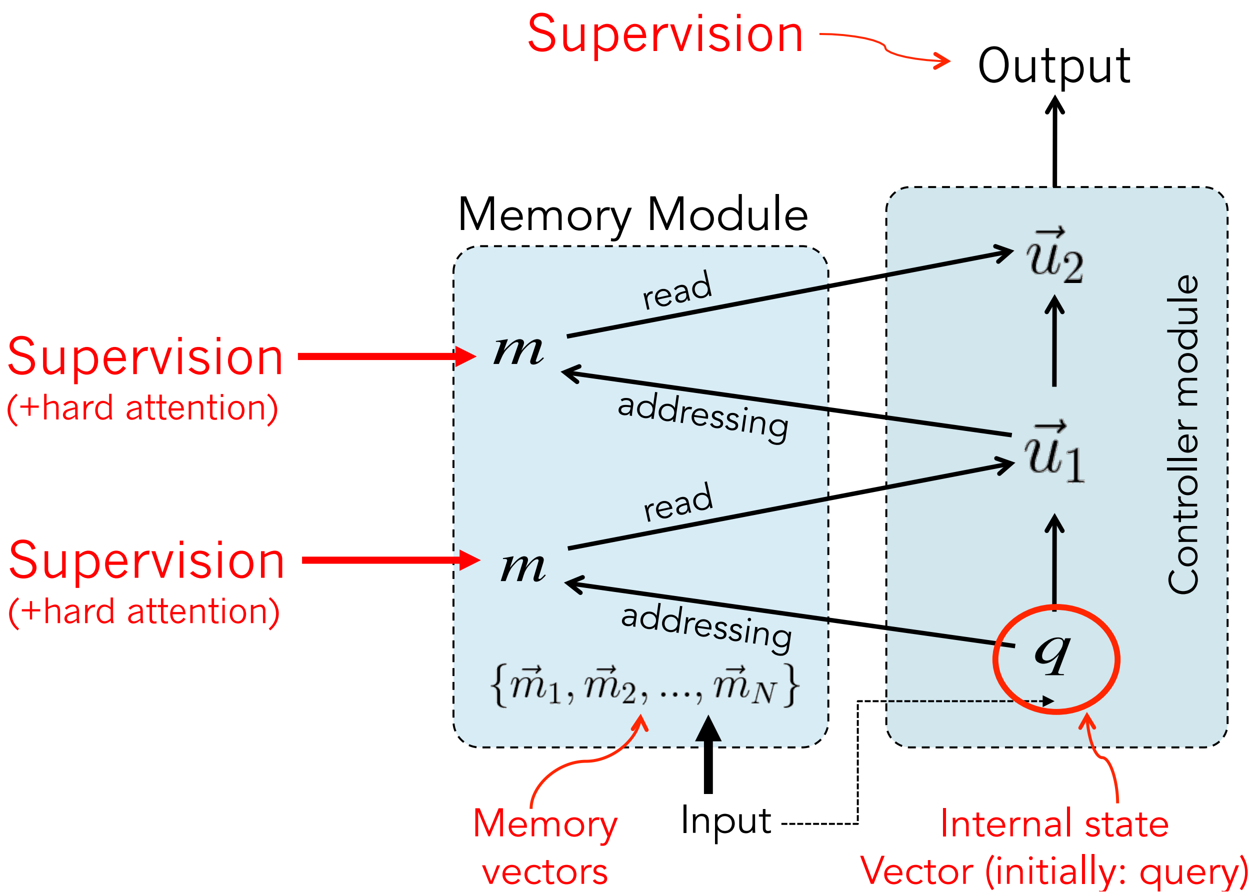
Addressing:
score m_i w.r.t. q

Read:
return best m_i



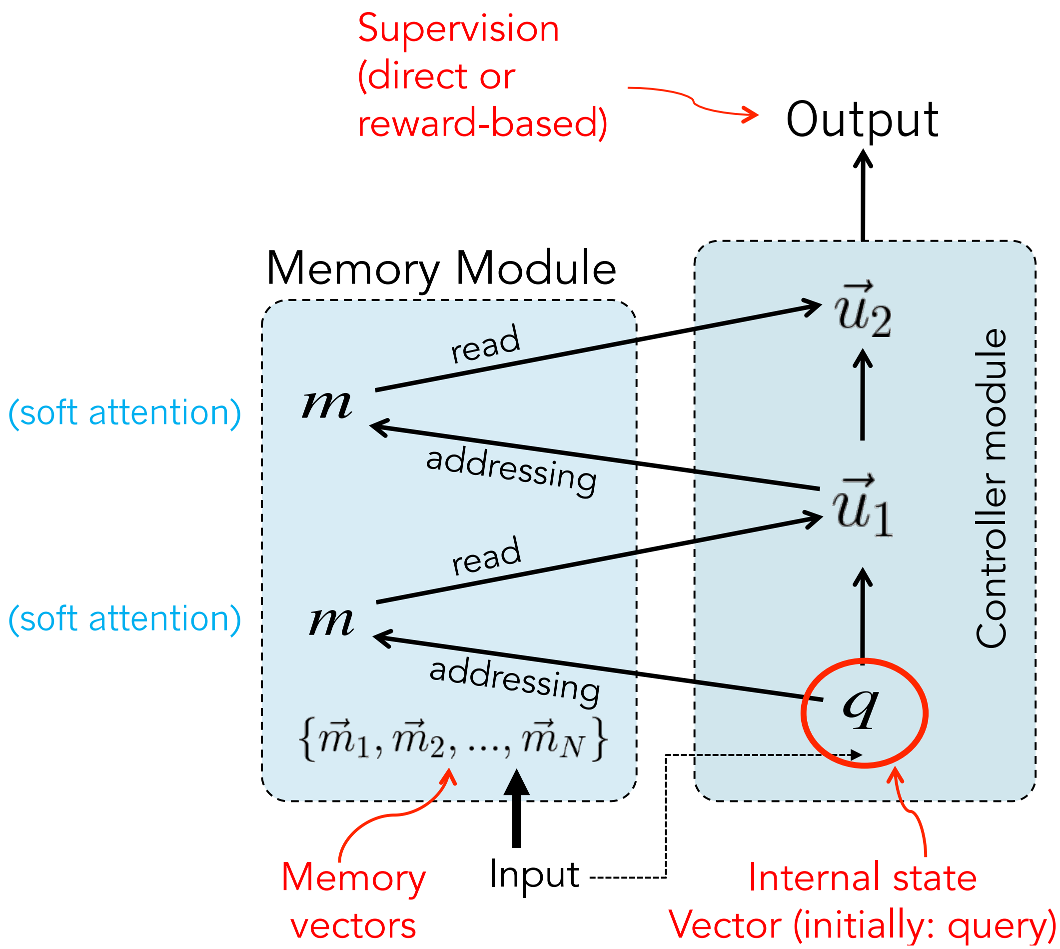
[Figure by Saina Sukhbaatar]

Memory Network (Weston et al,'14)



End-To-End Memory Network

(Sukhbaatar et al., '15)



[Figure by Saina Sukhbaatar]

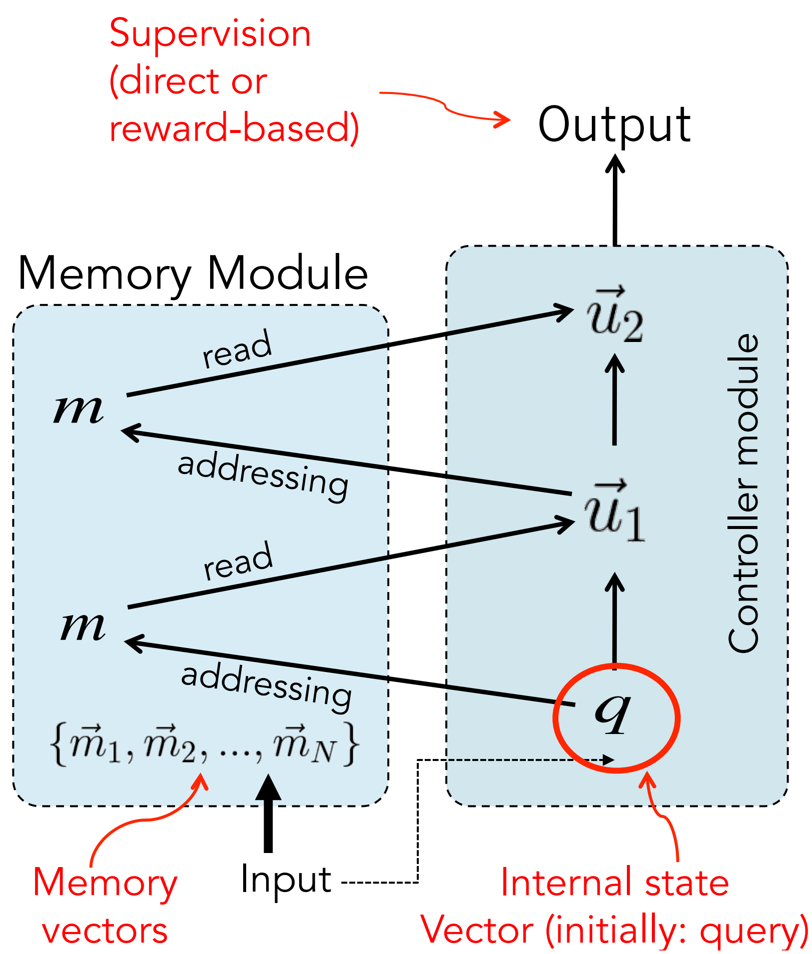
Less supervision: applicable to more problems

Key-Value Memory Network (Miller et al., '16)

$m_i = (k_i, v_i)$

Address: score k_i vs. q

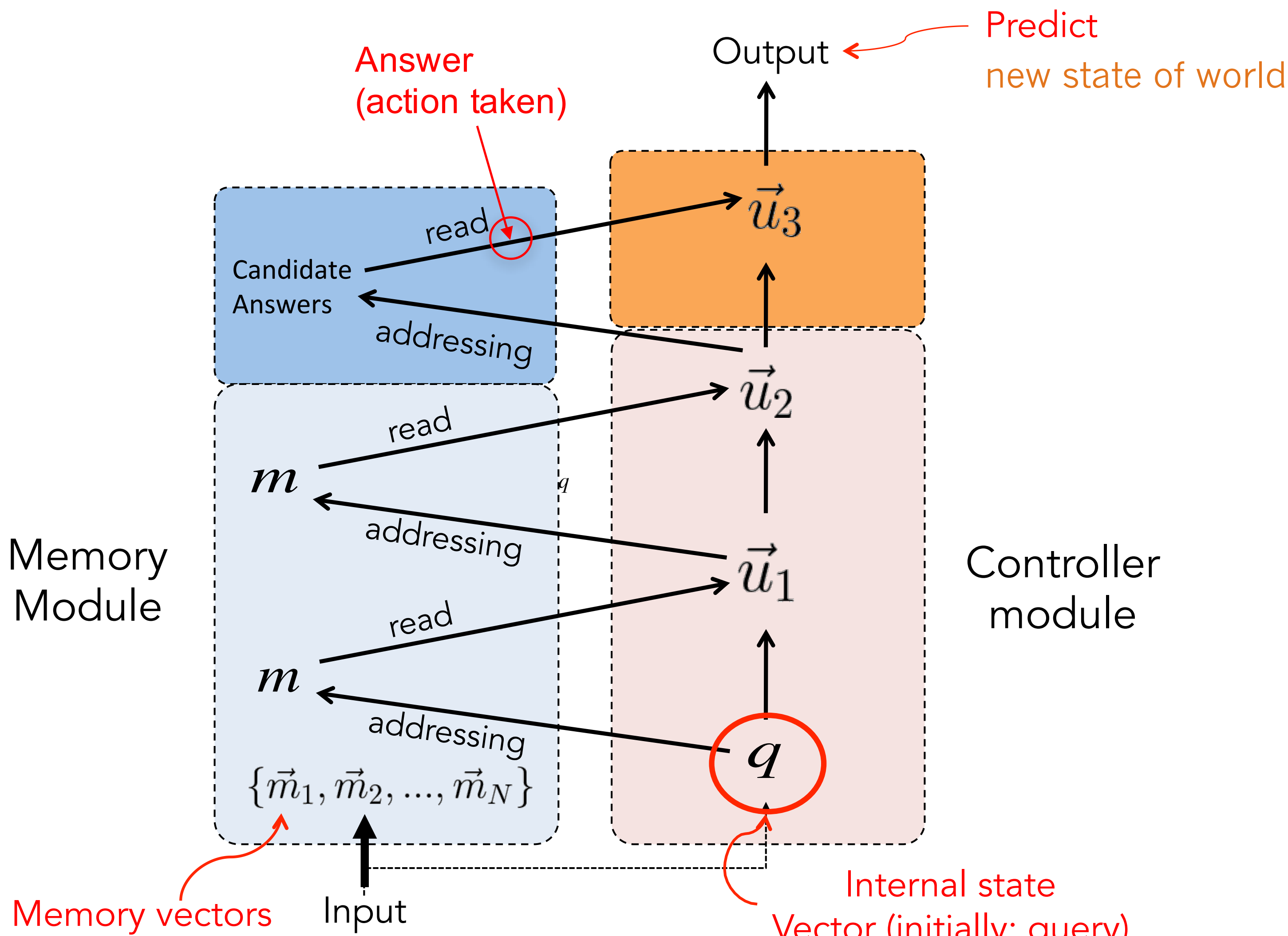
Read: return v_i



Encode Prior Knowledge: if k_i matches q, v_i matches output ➡ better results

Forward Prediction Memory Network

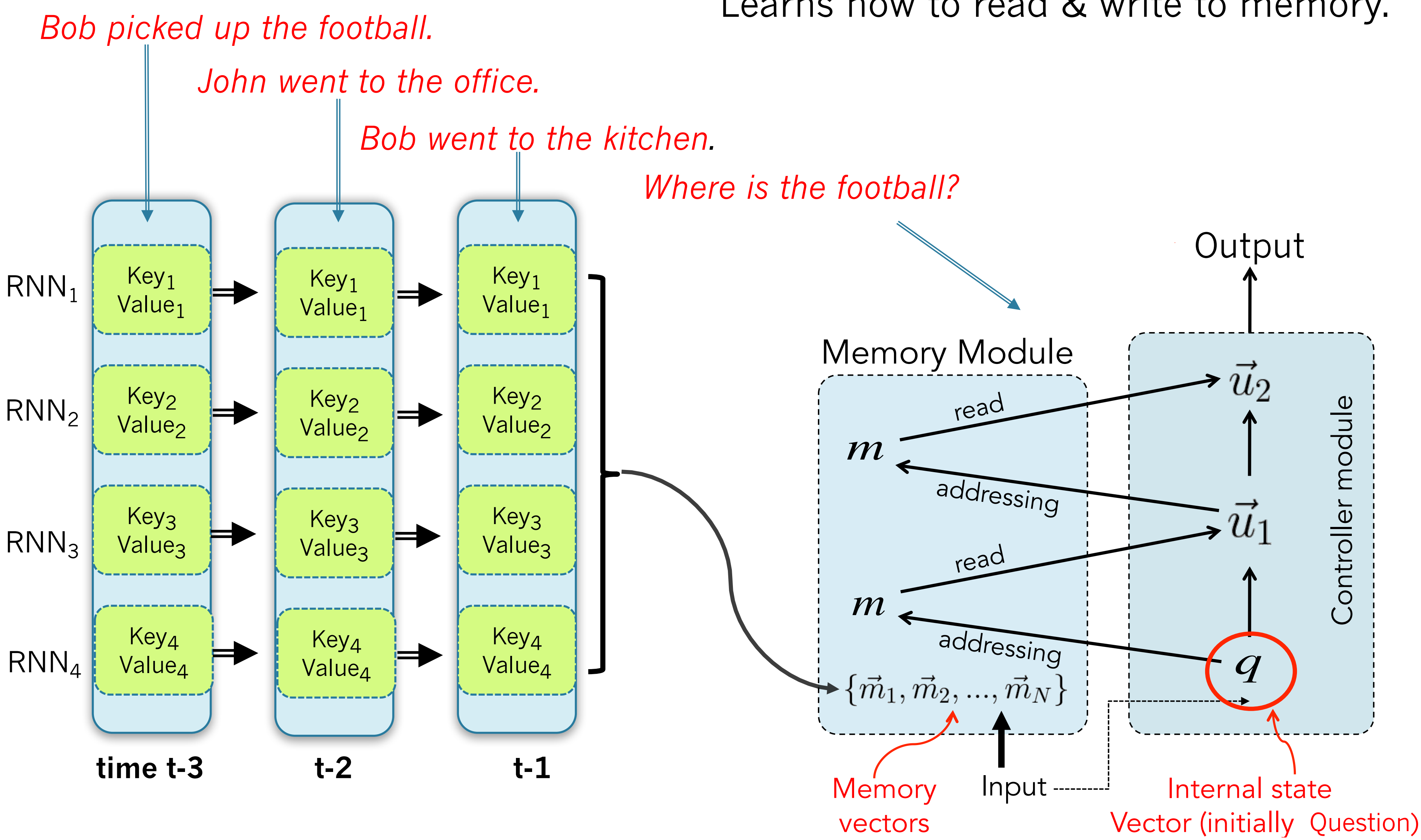
(Weston, '16)



"Unsupervised" Forward Model: does not require labeled supervision

Recurrent Entity Network (Henaff et al., '16)

Doesn't assume memories m_i are given.
Learns how to read & write to memory.



Recurrent Entity Network (Henaff et al., '16)

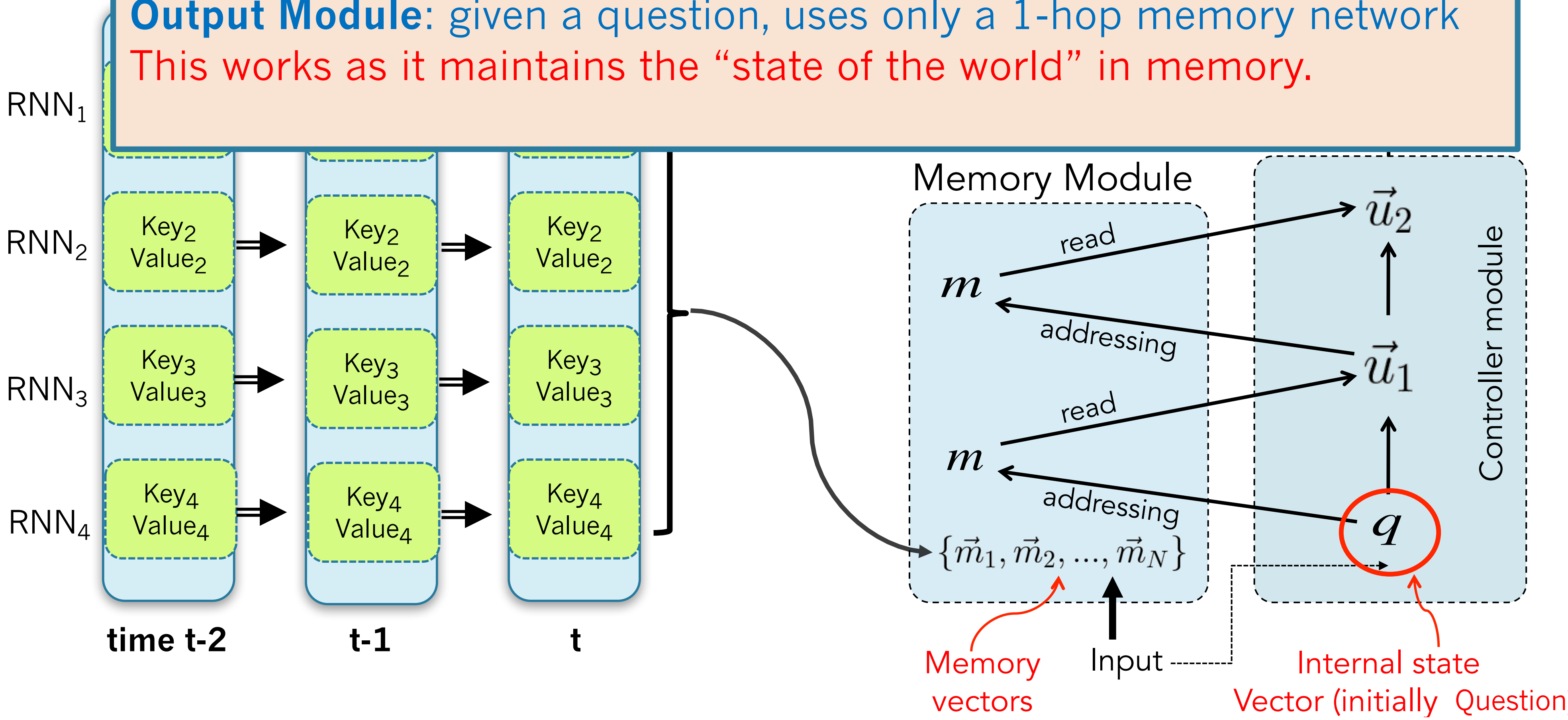
Encoder Module: encodes each input to a fixed size vector.

Memory Module: consists of key, value pairs that are learnt.

- Each memory slot is an RNN (with shared weights).
- Every RNN is fed the next input.
- Updated via gating function dependent on key and value.

Output Module: given a question, uses only a 1-hop memory network

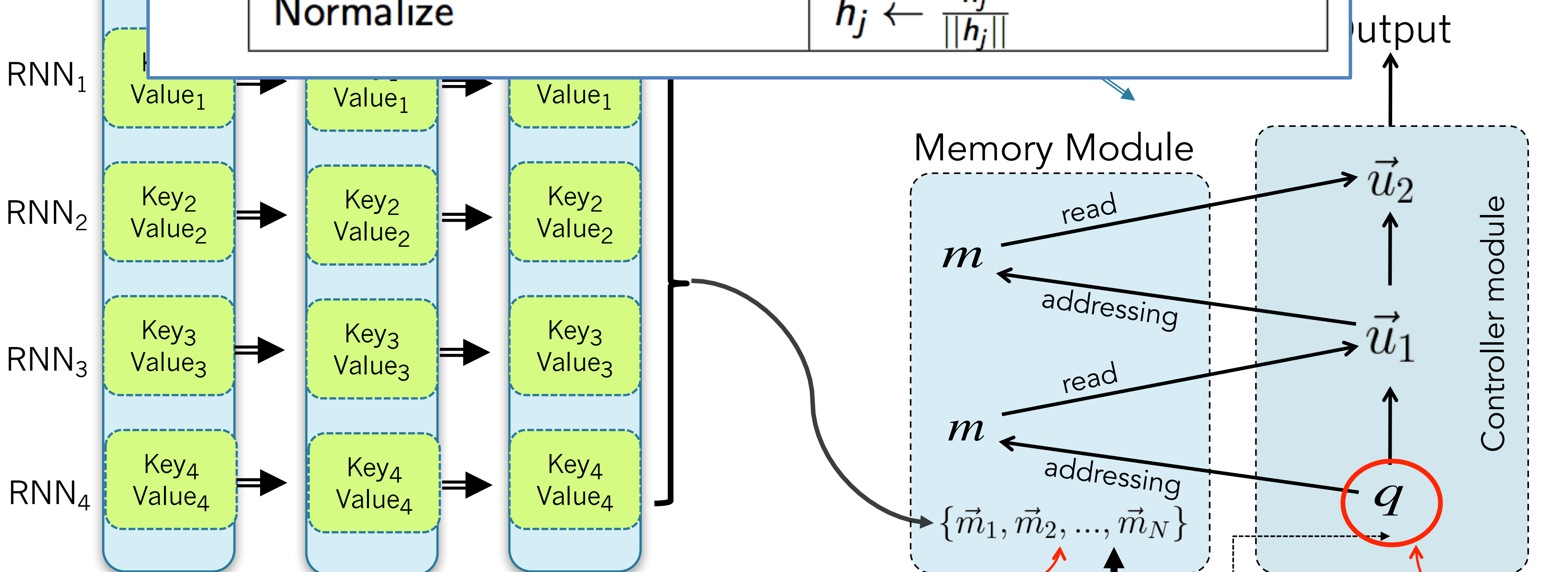
This works as it maintains the "state of the world" in memory.



Recurrent Entity Network (Henaff et al., '16)

- Input is s_t
- w_j, h_j are key/memory vectors
- Memory Update for j^{th} memory slot at timestep t:

Compute gate	$g_j \leftarrow \sigma(s_t^T w_j + s_t^T h_j)$
Compute candidate memory	$\tilde{h}_j \leftarrow \phi(Ah_j + Bs_t + Cw_j)$
Update memory	$h_j \leftarrow h_j + g_j \cdot \tilde{h}_j$
Normalize	$h_j \leftarrow \frac{h_j}{\lvert\lvert h_j \rvert\rvert}$



This model can reason while it is storing memories, MemNNs cannot.
Goal: Maintain an approximate "state of the world" in memory

Tasks (at http://fb.ai/babi)

- 20 bAbI Tasks (Weston et al., '15)
- Children's Book Test (Hill et al., '15)
- Movie Dialogue Dataset (Dodge et al., '15)
- 6 bAbI Dialogue Tasks (Bordes et al., '16)
- Dialogue-based Lang. Learning (Li et al., '16)

Related datasets: QACNN (Hermann et al., '15), Ubuntu (Lowe et al., '15), SQuAD (Rajpurkar et al., '16), BookTest (Bajgar et al., '16) MS MARCO (Nguyen et al., '16), Frames & NewsQA (Maluba)...

bAbI Tasks (Weston et al., '15)

Set of 20 tasks testing basic reasoning capabilities from simulated stories
Useful to foster innovation: cited 138+ times, used to evaluate new methods

Attention during mem lookup

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 (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				

20 bAbI Tasks
1k training set

	Test Acc	Failed tasks
LSTM	49%	20
MemN2N 1 hop	74.8%	17
2 hops	84.4%	11
3 hops	87.6.%	11

bAbI 1k and 10k comparisons

- 1k training set

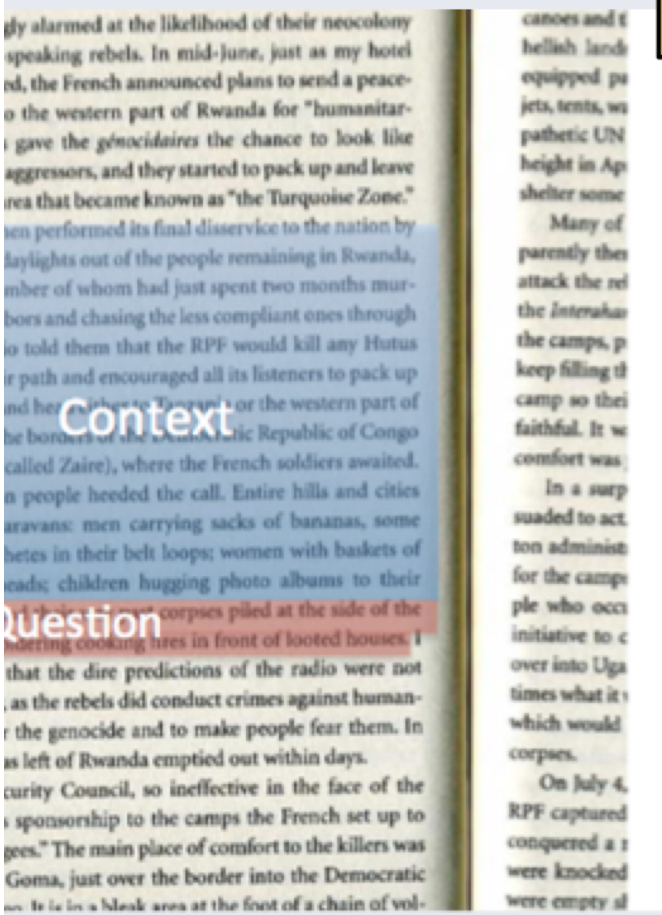
	Test Acc	Failed tasks
LSTM	51%	20
NTM	**???**	**???**
MemN2N (3 hops)	12.4.%	11
DNC	**???**	**???**
Dynamic MemNet	24.9%	12
EntNet (1 hop)	29.6%	15

10k training set

	Test Acc	Failed tasks
LSTM	36.4%	16
D-NTM	12.8%	9
MemN2N (3 hops)	4.2%	3
DNC	3.8%	2
Dynamic MemNet	2.8%	1
EntNet (1 hop)	0.5%	0

QA on REAL children's stories

missing word in a sentence given 20 previous sentences **as multiple choice task.**



New dataset based on 118 children books from project Gutenberg

1 `` phebe beckoned to him ; i saw her , '' cried rose , staring hard at the door .
2 `` is it more presents coming ? ''
3 asked jamie , just as his brother re-appeared , looking more excited than ever .
4 `` yes ; a present for mother , and here it is ! ''
5 roared archie , flinging wide the door to let in a tall man , who cried out , `` where 's my little woman ?
6 the first kiss for her , then the rest may come on as fast as they like . ''
7 before the words were out of his mouth , mrs. jessie was half-hidden under his rough great-coat , and four boys were prancing about him clamouring for their turn .
8 of course , there was a joyful tumult for a time , during which rose slipped into the window recess and watched what went on , as if it were a chapter in a christmas story .
9 it was good to see bluff uncle jem look proudly at his tall son , and fondly hug the little ones .
10 it was better still to see him shake his brothers ' hands as if he would never leave off , and kiss all the sisters in a way that made even solemn aunt myra brighten up for a minute .

11 but it was best of all to see him finally established in grandfather 's chair , with his `` little woman '' beside him , his three youngest boys in his lap , and _____ hovering over him like a large-sized cherub .

faith | brothers | rose | <u>archie</u> | rest | mouth | way | mother | sisters | george

Results on Children's Book Test

Methods	Named Entities	Common Nouns	Verbs	Prepositions
Humans (query)(*)	0.520	0.644	0.716	0.676
Humans (context+query)(*)	***0.816***	***0.816***	***0.828***	0.708
Maximum frequency (corpus)	0.120	0.158	0.373	0.315
Maximum frequency (context)	0.335	0.281	0.285	0.275
Sliding window	0.168	0.196	0.182	0.101
Word distance model	0.398	0.364	0.380	0.237
Kneser-Ney language model	0.390	0.544	0.778	0.768
Kneser-Ney language model + cache	0.439	0.577	0.772	0.679
Embedding Model (context+query)	0.253	0.259	0.421	0.315
Embedding Model (query)	0.351	0.400	0.614	0.535
Embedding Model (window)	0.362	0.415	0.637	0.589
Embedding Model (window+position)	0.402	0.506	0.736	0.670
LSTMs (query)	0.408	0.541	0.813	0.802
LSTMs (context+query)	0.418	0.560	**0.818**	0.791
Contextual LSTMs (window context)	0.436	0.582	0.805	**0.806**
MemNNs (lexical memory)	0.431	0.562	0.798	0.764
MemNNs (window memory)	0.493	0.554	0.692	0.674
MemNNs (sentential memory + PE)	0.318	0.305	0.502	0.326
MemNNs (window memory + self-sup.)	**0.666**	**0.630**	0.690	0.703

Showed that language modeling should focus on named entities / nouns, as that's the hard problem compared to human performance. Requires memory + reasoning.

Results on Children's Book Test

Many New Models And Results This Year

Text Understanding with the Attention Sum Reader Network. Kadlec et al. '16 CBT-NE: 71.0 CBT-CN: 68.9

Uses RNN style encoding of words + bypass module + 1 hop

Iterative Alternating Neural Attention for Machine Reading. Sordoni et al. '16 CBT-NE: 72.0 CBT-CN: 71.0

Natural Language Comprehension with the EpiReader. Trischler et al. '16 CBT-NE: 71.8 CBT-CN: 70.6

Gated-Attention Readers for Text Comprehension. Dhingra et al. '16 CBT-NE: 71.9 CBT-CN: 69.0

Uses RNN style encoding of words + bypass module + multiplicative combination of query + multiple hops

Metho				epositions
Human				0.676
Human				0.708
Maximu				0.315
Maximu				0.275
Sliding				0.101
Word d				0.237
Kneser				0.768
Kneser				0.679
Embedd				0.315
Embedd				0.535
Embedd				0.589
Embedd				0.670
LSTMs				0.802
LSTMs				0.791
Contex				**0.806**
MemNNs (lexical memory)	0.431	0.562	0.798	0.764
MemNNs (window memory)	0.493	0.554	0.692	0.674
MemNNs (sentential memory + PE)	0.318	0.305	0.502	0.326
MemNNs (window memory + self-sup.)	**0.666**	**0.630**	0.690	0.703

Showed that language modeling should focus on named entities / nouns, as that's the hard problem compared to human performance. Requires memory + reasoning.

QA is only a tiny part of dialogue.
What about dialogue (with knowledge)?
Reddit + OMDB

Memories	h_i	Shaolin Soccer written_by Stephen Chow Shaolin Soccer starred_actors Stephen Chow Shaolin Soccer release_year 2001 Shaolin Soccer has_genre comedy Shaolin Soccer has_tags martial arts, kung fu soccer, stephen chow Kung Fu Hustle directed_by Stephen Chow Kung Fu Hustle written_by Stephen Chow Kung Fu Hustle starred_actors Stephen Chow Kung Fu Hustle has_genre comedy action Kung Fu Hustle has_imdb_votes famous Kung Fu Hustle has_tags comedy, action, martial arts, kung fu, china, soccer, hong kong, stephen chow The God of Cookery directed_by Stephen Chow The God of Cookery written_by Stephen Chow The God of Cookery starred_actors Stephen Chow The God of Cookery has_tags hong kong Stephen Chow From Beijing with Love directed_by Stephen Chow From Beijing with Love written_by Stephen Chow From Beijing with Love starred_actors Stephen Chow, Anita Yuen . . . <and more> . . .
Short-Term Memories	c_1^u c_1^r	1) I'm looking a fun comedy to watch tonight, any ideas? 2) Have you seen Shaolin Soccer? That was zany and great.. really funny but in a whacky way.
Input	c_2^u	3) Yes! Shaolin Soccer and Kung Fu Hustle are so good I really need to find some more Stephen Chow films I feel like there is more awesomeness out there that I haven't discovered yet ...

Dodge et al. "Evaluating Prerequisite Qualities for Learning End-to-End Dialog Systems." ICLR '16

(Dialog 1) QA: *facts about movies*

Sample input contexts and target replies (in red) from Dialog Task 1:

What movies are about open source? Revolution OS
Ruggero Raimondi appears in which movies? Carmen
What movies did Darren McGavin star in? Billy Madison, The Night Stalker, Mrs. Pollifax-Spy, The Challenge
Can you name a film directed by Stuart Ortiz? Grave Encounters
Who directed the film White Elephant? Pablo Trapero
What is the genre of the film Dial M for Murder? Thriller, Crime
What language is Whity in? German

(Dialog 2) Recs: *movie recommendations*

Sample input contexts and target replies (in red) from Dialog Task 2:

Schindler's List, The Fugitive, Apocalypse Now, Pulp Fiction, and The Godfather are films I really liked. Can you suggest a film?
The Hunt for Red October

Some movies I like are Heat, Kids, Fight Club, Shaun of the Dead, The Avengers, Skyfall, and Jurassic Park. Can you suggest something else I might like? Ocean's Eleven

(Dialog 3) QA+Recs: *combination dialog*

Sample input contexts and target replies (in red) from Dialog Task 3:

I loved Billy Madison, Blades of Glory, Bio-Dome, Clue, and Happy Gilmore. I'm looking for a Music movie. School of Rock
What else is that about? Music, Musical, Jack Black, school, teacher, Richard Linklater, rock, guitar
I like rock and roll movies more. Do you know anything else?
Little Richard

(Dialog 4) Reddit: *real dialog*

Sample input contexts and target replies (in red) from Dialog Task 4:

I think the Terminator movies really suck, I mean the first one was kinda ok, but after that they got really cheesy. Even the second one which people somehow think is great. And after that... forgeddabotit.
C'mon the second one was still pretty cool.. Arny was still so badass, as was Sararah Connor's character.. and the way they blended real action and effects was perhaps the last of its kind...

Dodge et al. "Evaluating Prerequisite Qualities for Learning End-to-End Dialog Systems." ICLR '16

METHODS	QA TASK (HITS@1)	RECS TASK (HITS@100)	QA+RECS TASK (HITS@10)	REDDIT TASK (HITS@10)
QA SYSTEM (BORDES ET AL., 2014)	90.7	N/A	N/A	N/A
SVD	N/A	19.2	N/A	N/A
IR	N/A	N/A	N/A	23.7
LSTM	6.5	27.1	19.9	11.8
SUPERVISED EMBEDDINGS	50.9	29.2	65.9	27.6
MEMN2N	79.3	28.6	81.7	29.2
JOINT SUPERVISED EMBEDDINGS	43.6	28.1	58.9	14.5
JOINT MEMN2N	83.5	26.5	78.9	26.6

(Dialog ... *dation*

Sample input contex ... g Task 2:

What movies are al
Ruggero Raimondi
What movies did D
Stalker, Mrs. Pollifa
Can you name a film directed by Stuart Ortiz? Grave Encounters
Who directed the film White Elephant? Pablo Trapero
What is the genre of the film Dial M for Murder? Thriller, Crime
What language is Whity in?

n, and
lm?
Some movies I like are Heat, Kids, Fight Club, Shaun of the Dead, The Avengers, Skyfall, and Jurassic Park. Can you suggest something else I might like? Ocean's Eleven

(Dialog 3) QA+R ... *l dialog*

Sample input contexts and ta ... rom Dialog Task 4:

I loved Billy Madison, Blade
Gilmore. I'm looking for a M
What else is that about? Music, Musical, Jack Black, school, teacher, Richard Linklater, rock, guitar
I like rock and roll movies more. Do you know anything else?
Little Richard

n the first one was kinda ok, but after that they got really cheesy. Even the second one which people somehow think is great. And after that... forgeddabotit.
C'mon the second one was still pretty cool.. Arny was still so badass, as was Sararah Connor's character.. and the way they blended real action and effects was perhaps the last of its kind...

Somewhat ignored by the community so far, but I think this is closer to something real and useful compared to the previous two datasets.

Goal: an ever-expanding set of real dialogue tasks to learn from and evaluate on. This is a first step.

Learning From Human Responses

Mary went to the hallway.

John moved to the bathroom.

Mary travelled to the kitchen.

Where is Mary? A:playground

No, that's incorrect.

Where is John? A:bathroom

Yes, that's right!

If you can predict this, you are most of the way to knowing how to answer correctly.

Human Responses Give Lots of Info

Mary went to the hallway.

John moved to the bathroom.

Mary travelled to the kitchen.

Where is Mary? A:playground

No, the answer is kitchen.

Where is John? A:bathroom

Yes, that's right!

Much more signal than just “No” or zero reward.

Real Human Questions+Feedback

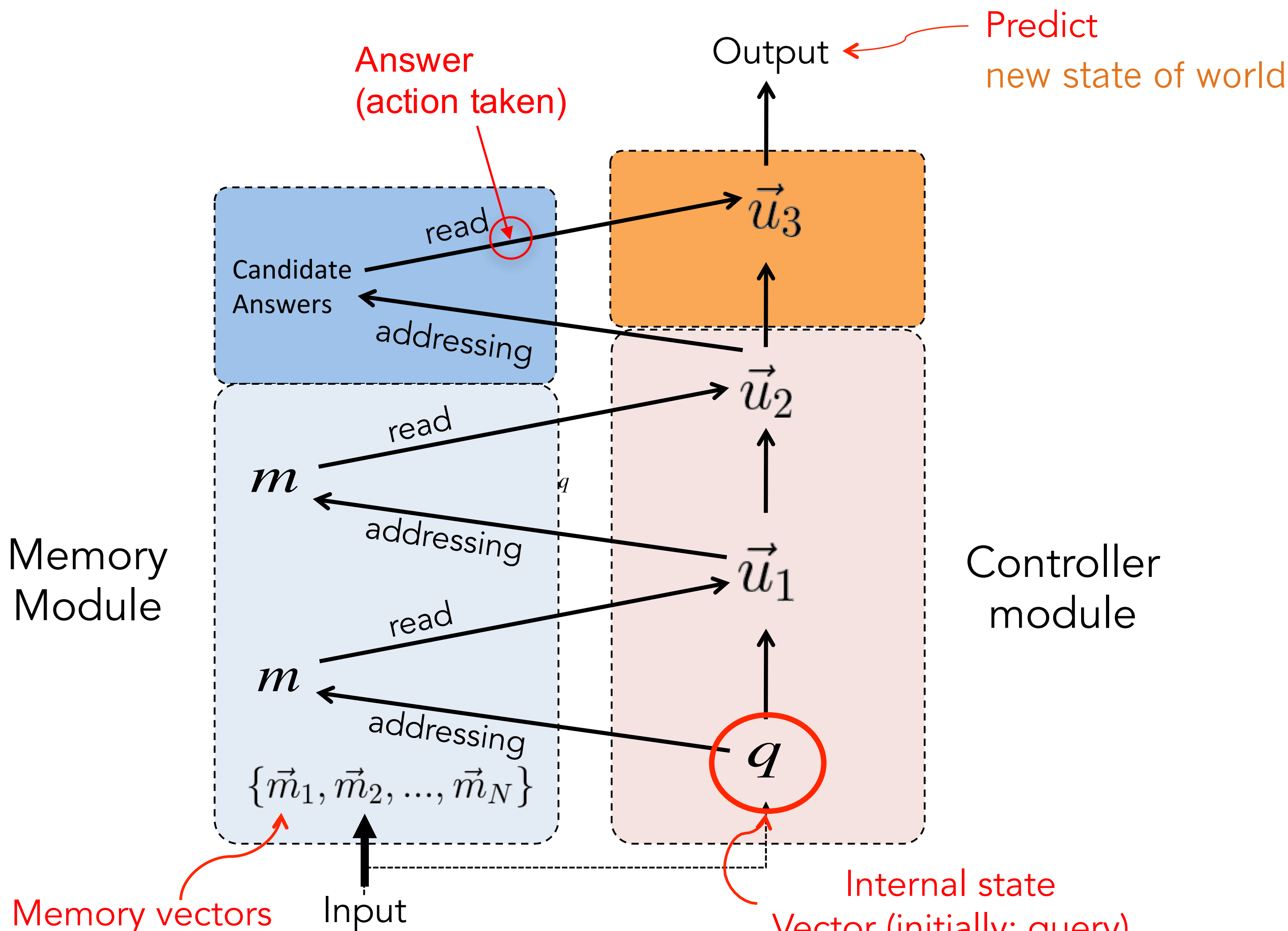
Figure 2: **Human Dialogue from Mechanical Turk (based on WikiMovies)** The human teacher's dialogue is in black and the bot is in red. We show examples where the bot answers correctly (left) and incorrectly (right). Real humans provide more variability of language in both questions and textual feedback than in the simulator setup (cf. Figure 1).

Sample dialogues with correct answers from the bot:	
Who wrote the Linguini Incident ?	richard shepard
Richard Shepard is one of the right answers here.	
What year did The World Before Her premiere?	2012
Yep! That's when it came out.	
Which are the movie genres of Mystery of the 13th Guest?	crime
Right, it can also be categorized as a mystery.	
Sample dialogues with incorrect answers from the bot:	
What are some movies about a supermarket ?	supermarket
There were many options and this one was not among them.	
Which are the genres of the film Juwanna Mann ?	kevin pollak
That is incorrect. Remember the question asked for a genre not name.	
Who wrote the story of movie Coraline ?	fantasy
That's a movie genre and not the name of the writer. A better answer would of been Henry Selick or Neil Gaiman.	

Much more diversity!!

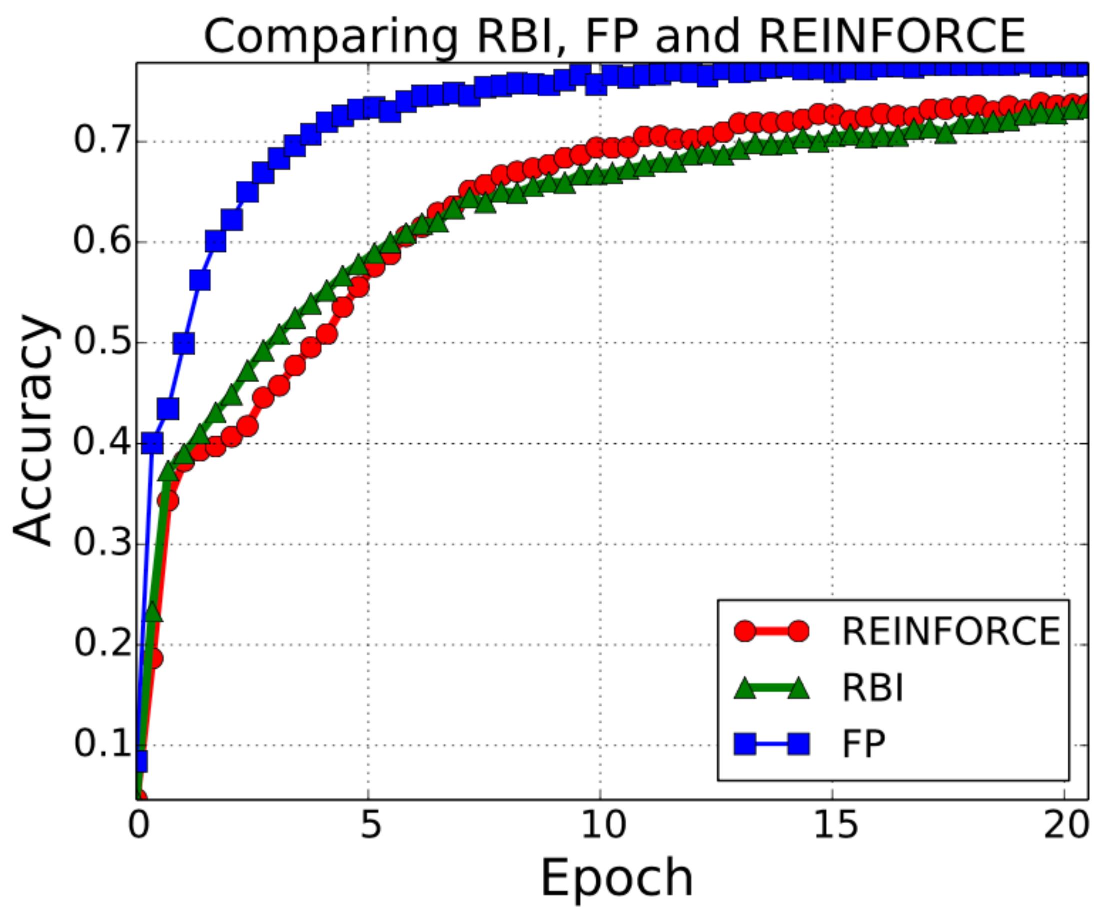
Forward Prediction Memory Network

(Weston, '16)



"Unsupervised" Forward Model: does not require labeled supervision

Dialog Feedback: Results



WikiMovies
(templated task)

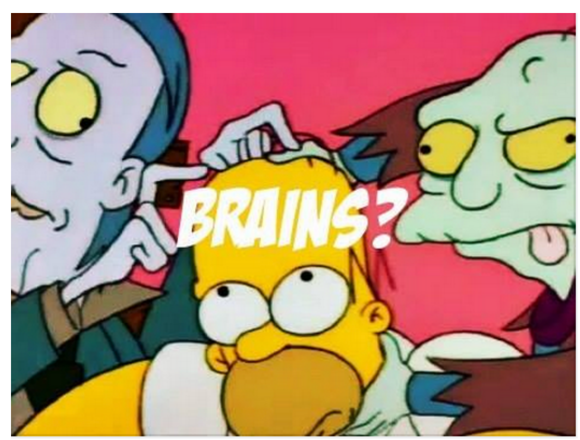
Forward Prediction MemNN (FP) which uses textual rewards
can perform better than using numerical rewards (RBI or REINFORCE)!!

Conclusion

Tasks and models for:
Memory: *long and short-term, knowledge*
Reasoning: *both logical & not + commonsense*
Indirect Supervision: *no labels, almost no rewards*

.. Many more issues still to explore...

Thanks!



FAIR: paper / data / code

- Papers:
  - bAbI tasks: arxiv.org/abs/1502.05698
  - Memory Networks: http://arxiv.org/abs/1410.3916
  - End-to-end Memory Networks: http://arxiv.org/abs/1503.08895
  - Large-scale QA with MemNNs: http://arxiv.org/abs/1506.02075
  - Reading Children's Books: http://arxiv.org/abs/1511.02301
  - Evaluating End-To-End Dialog: http://arxiv.org/abs/1511.06931
  - Dialog-based Language Learning: http://arxiv.org/abs/1604.06045
- Data:
  - bAbI tasks: fb.ai/babi
  - SimpleQuestions dataset (100k questions): fb.ai/babi
  - Children's Book Test dataset: fb.ai/babi
  - Movie Dialog Dataest: fb.ai/babi
- Code:
  - Memory Networks variants: https://github.com/facebook/MemNN
  - Simulation tasks generator: https://github.com/facebook/bAbI-tasks

Some Memory Network-related Publications

- J. Weston, S. Chopra, A. Bordes. Memory Networks. ICLR 2015 (and arXiv:1410.3916).
- S. Sukhbaatar, A. Szlam, J. Weston, R. Fergus. End-To-End Memory Networks. NIPS 2015 (and arXiv:1503.08895).
- J. Weston, A. Bordes, S. Chopra, A. M. Rush, B. van Merriënboer, A. Joulin, T. Mikolov. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. arXiv:1502.05698.
- A. Bordes, N. Usunier, S. Chopra, J. Weston. Large-scale Simple Question Answering with Memory Networks. arXiv:1506.02075.
- J. Dodge, A. Gane, X. Zhang, A. Bordes, S. Chopra, A. Miller, A. Szlam, J. Weston. Evaluating Prerequisite Qualities for Learning End-to-End Dialog Systems. arXiv:1511.06931.
- F. Hill, A. Bordes, S. Chopra, J. Weston. The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. arXiv:1511.02301.
- J. Weston. Dialog-based Language Learning. arXiv:1604.06045.
- A. Bordes, Jason Weston. Learning End-to-End Goal-Oriented Dialog. arXiv:1605.07683.

RAM Issues

- How to decide what to write and what not to write in the memory?
- How to represent knowledge to be stored in memories?
- Types of memory (arrays, stacks, or stored within weights of model), when they should be used, and how can they be learnt?
- How to do fast retrieval of relevant knowledge from memories when the scale is huge?
- How to build hierarchical memories, e.g. multiscale attention?
- How to build hierarchical reasoning, e.g. composition of functions?
- How to incorporate forgetting/compression of information?
- How to evaluate reasoning models? Are artificial tasks a good way? Where do they break down and real tasks are needed?
- Can we draw inspiration from how animal or human memories work?