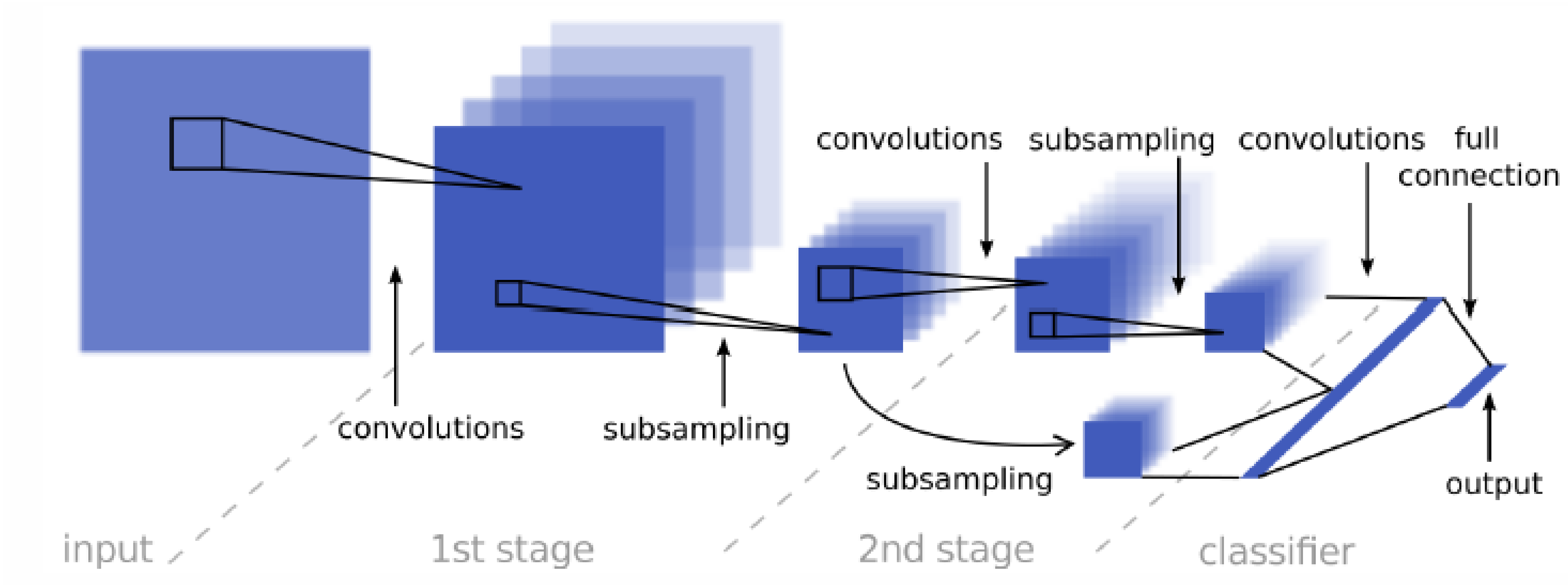
Reasoning, Attention and Memory

Sumit Chopra
Facebook AI Research

Deep Learning for Vision



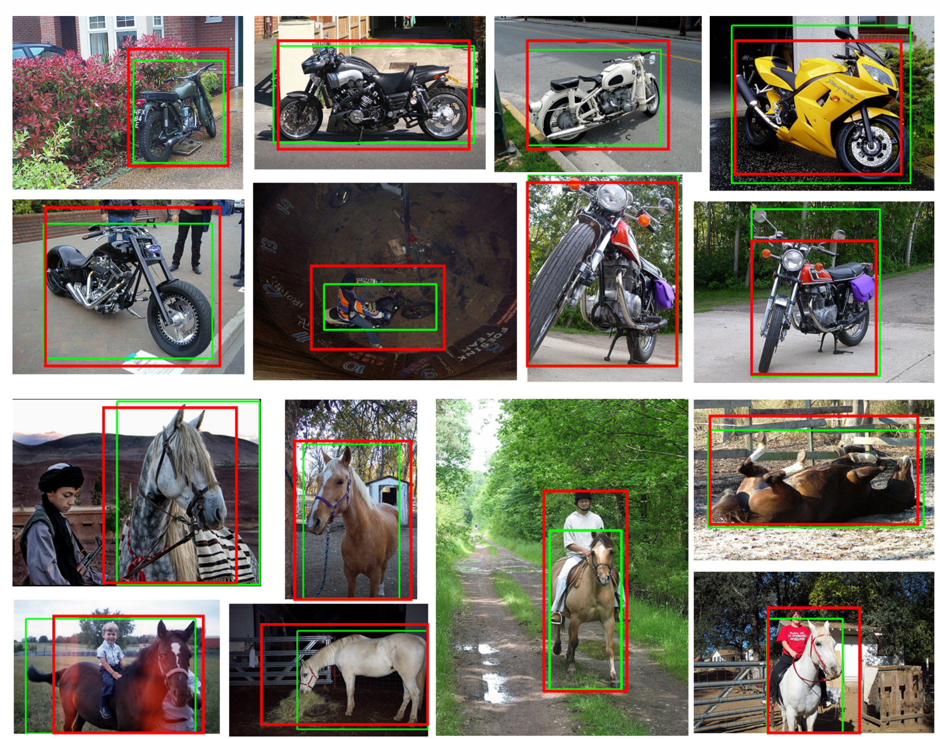
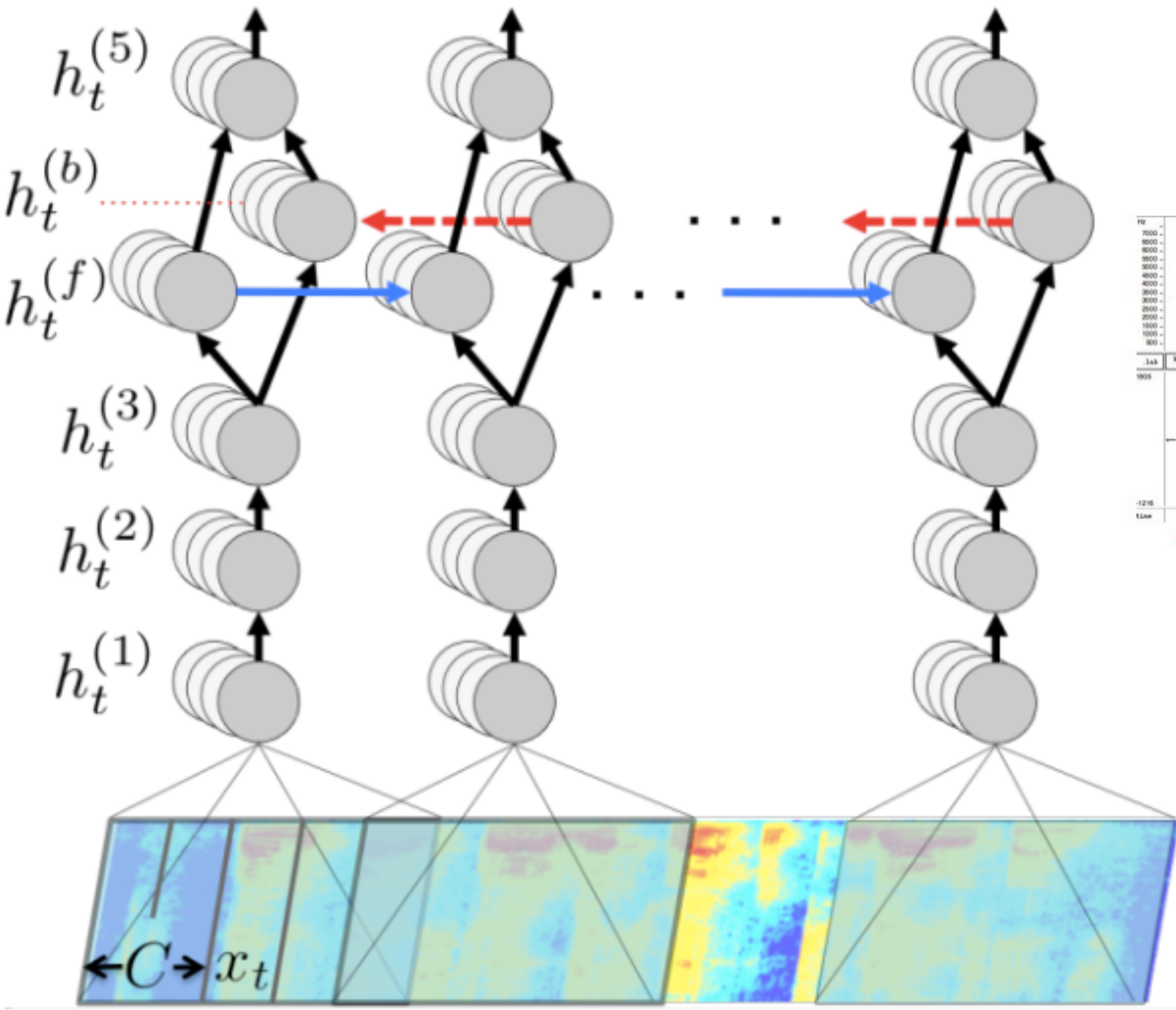


Figure Credit: Xiaogang Wang

Deep Learning for Speech



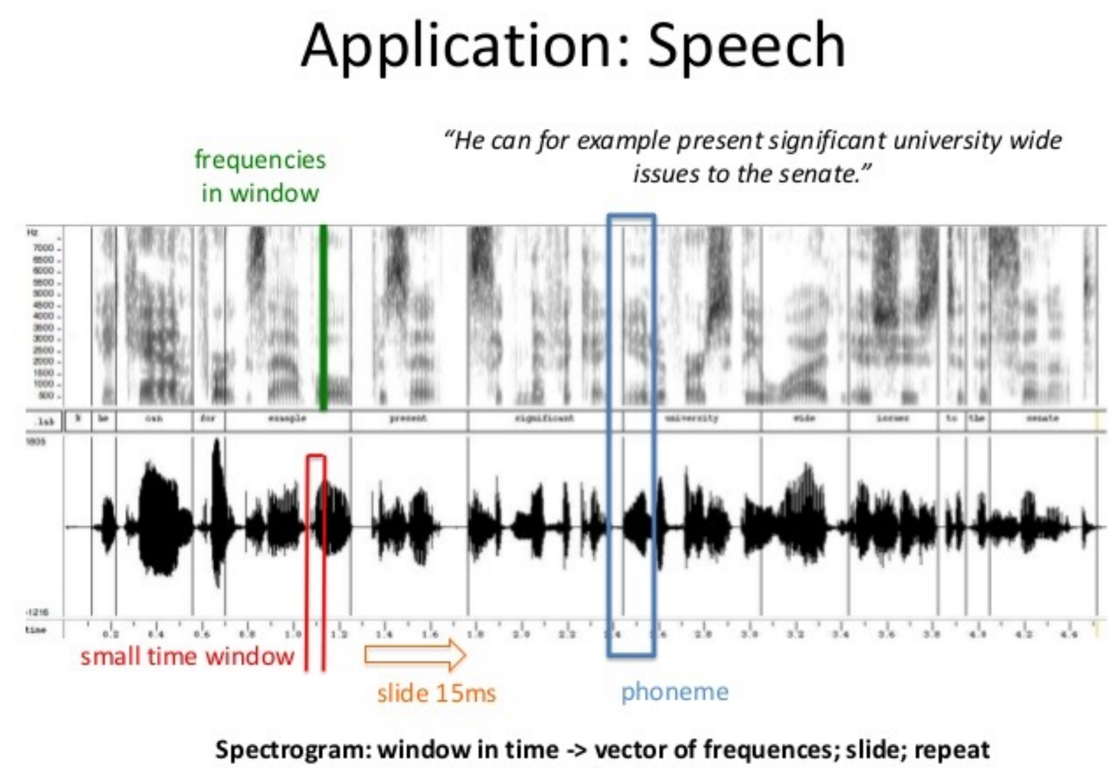
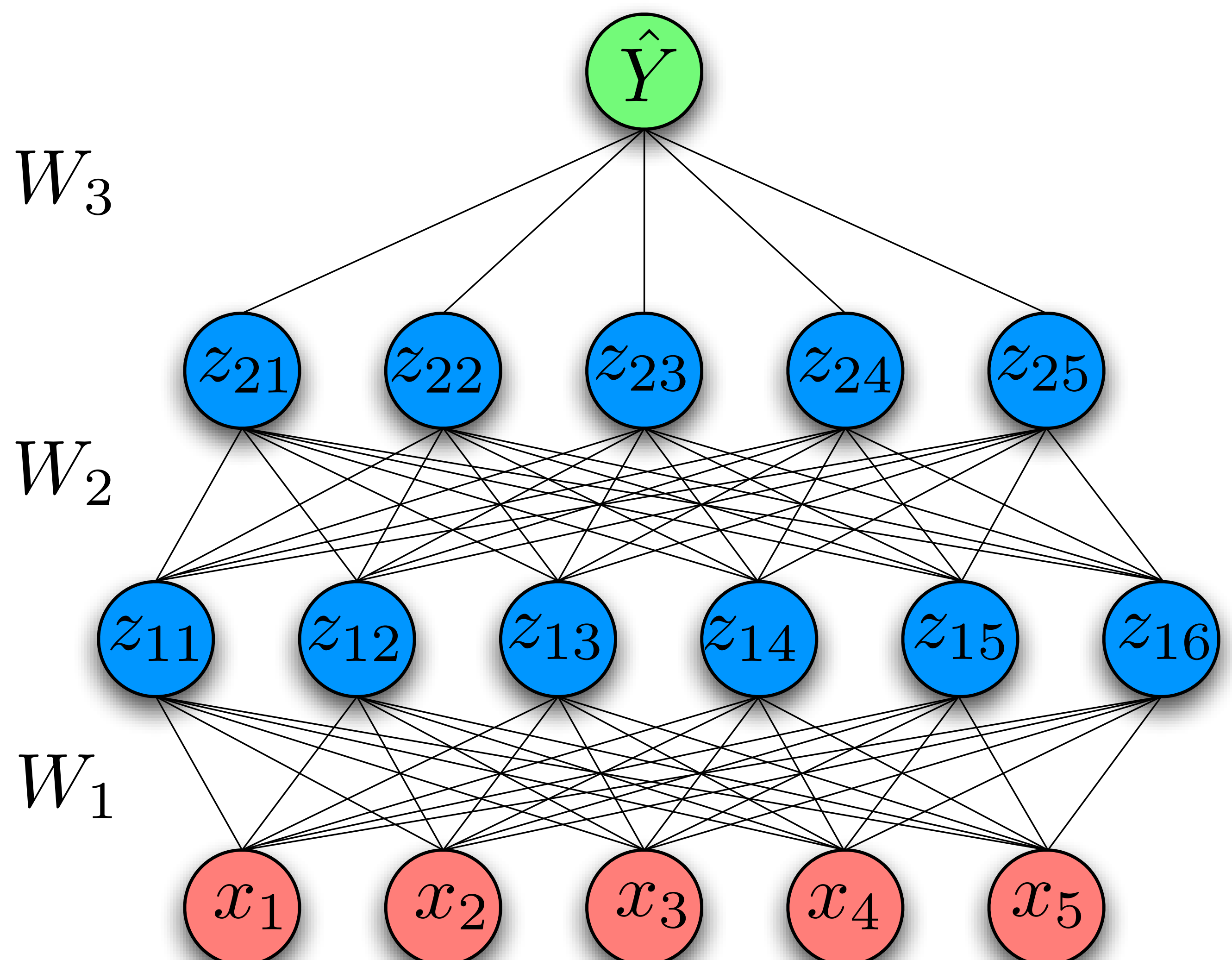


Figure Credit: Nvidia

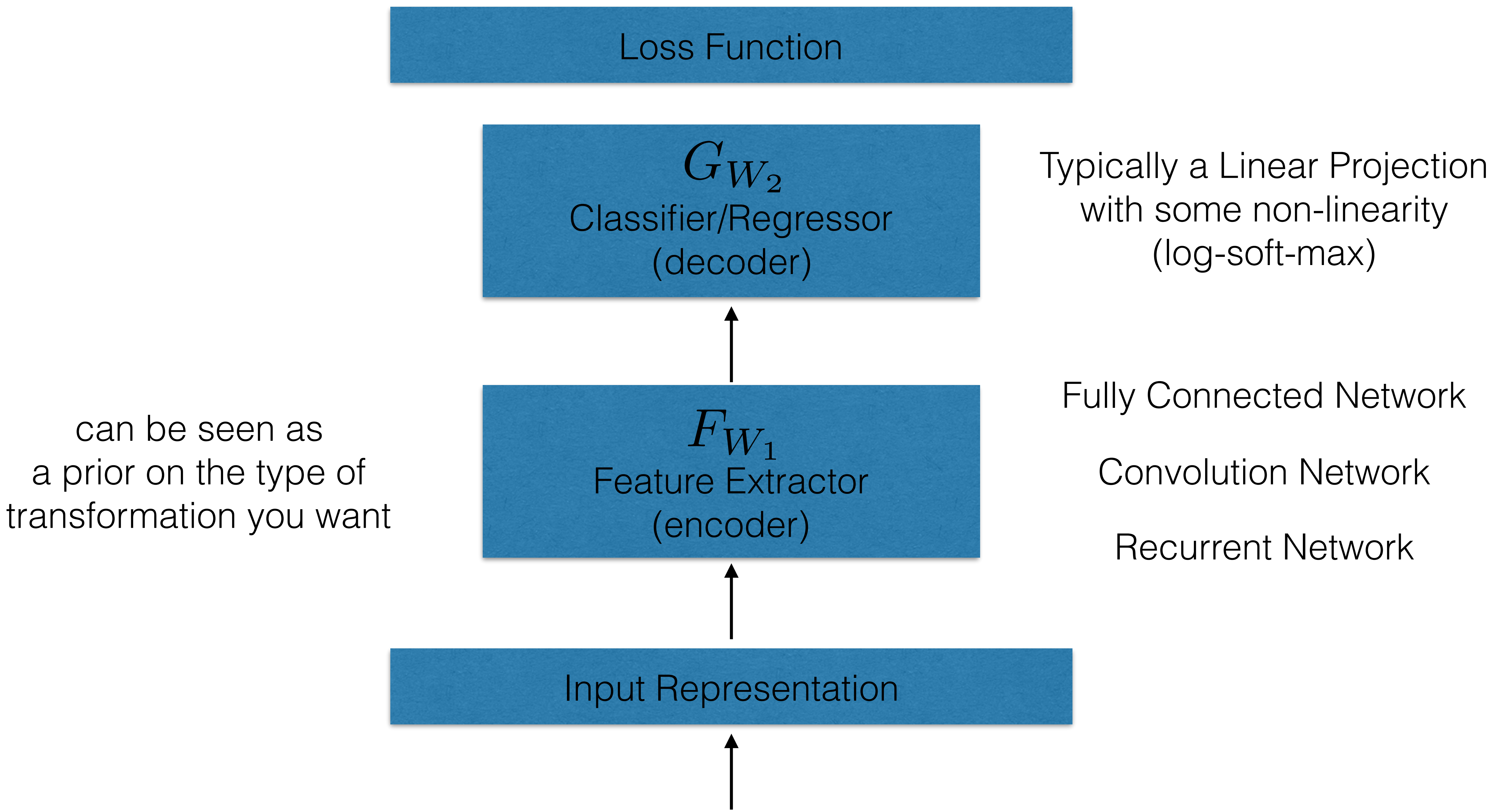
Deep Learning for Text

positive



“The movie was not bad at all. I had fun.”

Deep Models



“The movie was not bad at all. I had fun.”

Deep Models

Loss Function

G_{W_2}

Classifier/Regressor (decoder)

Typically a Linear Projection with some non-linearity (log-soft-max)

Learnable parametric function

Inputs: generally considered I.I.D.

Outputs: classification or regression

F_{W_1}

Feature Extractor (encoder)

can be seen as a prior on the type of transformation you want

Fully Connected Network

Convolution Network

Recurrent Network

Input Representation

Embedding Matrix

"The movie was not bad at all. I had fun."

Scenario 1

Joe went to the kitchen. Fred went to the kitchen. Joe picked up the milk. Joe travelled to the office. Joe left the milk. Joe went to the bathroom.

Scenario 1

Joe went to the kitchen. Fred went to the kitchen. Joe picked up the milk.
Joe travelled to the office. Joe left the milk. Joe went to the bathroom.
Where is the milk now?
Where is Joe?
Where was Joe before the office?

Scenario 1

Joe went to the kitchen. Fred went to the kitchen. Joe picked up the milk.
Joe travelled to the office. Joe left the milk. Joe went to the bathroom.
Where is the milk now? A: office
Where is Joe?
Where was Joe before the office?

Scenario 1

Joe went to the kitchen. Fred went to the kitchen. Joe picked up the milk.
Joe travelled to the office. Joe left the milk. Joe went to the bathroom.
Where is the milk now? A: office
Where is Joe? A: bathroom
Where was Joe before the office?

Scenario 1

Joe went to the kitchen. Fred went to the kitchen. Joe picked up the milk.
Joe travelled to the office. Joe left the milk. Joe went to the bathroom.
Where is the milk now? A: office
Where is Joe? A: bathroom
Where was Joe before the office? A: kitchen

Scenario 2

S: 1 Mr. Cropper was opposed to our hiring you .
2 Not , of course , that he had any personal objection to you , but he is set against female teachers , and when a Cropper is set there is nothing on earth can change him .
3 He says female teachers ca n't keep order .
4 He 's started in with a spite at you on general principles , and the boys know it .
5 They know he 'll back them up in secret , no matter what they do , just to prove his opinions .
6 Cropper is sly and slippery , and it is hard to corner him . ''
7 `` Are the boys big ? ''
8 queried Esther anxiously .
9 `` Yes .
10 Thirteen and fourteen and big for their age .
11 You ca n't whip 'em -- that is the trouble .
12 A man might , but they 'd twist you around their fingers .
13 You 'll have your hands full , I 'm afraid .
14 But maybe they 'll behave all right after all . ''
15 Mr. Baxter privately had no hope that they would , but Esther hoped for the best.
16 She could not believe that Mr. Cropper would carry his prejudices into a personal application .
17 This conviction was strengthened when he overtook her walking from school the next day and drove her home .
18 He was a big , handsome man with a very suave , polite manner .
19 He asked interestedly about her school and her work , hoped she was getting on well , and said he had two young rascals of his own to send soon .
20 Esther felt relieved .

Scenario 2

S: 1 Mr. Cropper was opposed to our hiring you .
2 Not , of course , that he had any personal objection to you , but he is set against female teachers , and when a Cropper is set there is nothing on earth can change him .
3 He says female teachers ca n't keep order .
4 He 's started in with a spite at you on general principles , and the boys know it .
5 They know he 'll back them up in secret , no matter what they do , just to prove his opinions .
6 Cropper is sly and slippery , and it is hard to corner him . ''
7 `` Are the boys big ? ''
8 queried Esther anxiously .
9 `` Yes .
10 Thirteen and fourteen and big for their age .
11 You ca n't whip 'em -- that is the trouble .
12 A man might , but they 'd twist you around their fingers .
13 You 'll have your hands full , I 'm afraid .
14 But maybe they 'll behave all right after all . ''
15 Mr. Baxter privately had no hope that they would , but Esther hoped for the best.
16 She could not believe that Mr. Cropper would carry his prejudices into a personal application .
17 This conviction was strengthened when he overtook her walking from school the next day and drove her home .
18 He was a big , handsome man with a very suave , polite manner .
19 He asked interestedly about her school and her work , hoped she was getting on well , and said he had two young rascals of his own to send soon .
20 Esther felt relieved .

q: She thought that Mr. ________ had exaggerated matters a little .

Scenario 2

S: 1 Mr. Cropper was opposed to our hiring you .
2 Not , of course , that he had any personal objection to you , but he is set against female teachers , and when a Cropper is set there is nothing on earth can change him .
3 He says female teachers ca n't keep order .
4 He 's started in with a spite at you on general principles , and the boys know it .
5 They know he 'll back them up in secret , no matter what they do , just to prove his opinions .
6 Cropper is sly and slippery , and it is hard to corner him . ''
7 `` Are the boys big ? ''
8 queried Esther anxiously .
9 `` Yes .
10 Thirteen and fourteen and big for their age .
11 You ca n't whip 'em -- that is the trouble .
12 A man might , but they 'd twist you around their fingers .
13 You 'll have your hands full , I 'm afraid .
14 But maybe they 'll behave all right after all . ''
15 Mr. Baxter privately had no hope that they would , but Esther hoped for the best.
16 She could not believe that Mr. Cropper would carry his prejudices into a personal application .
17 This conviction was strengthened when he overtook her walking from school the next day and drove her home .
18 He was a big , handsome man with a very suave , polite manner .
19 He asked interestedly about her school and her work , hoped she was getting on well , and said he had two young rascals of his own to send soon .
20 Esther felt relieved .

q: She thought that Mr. Baxter had exaggerated matters a little .

Scenario 3

Shaolin Soccer directed_by Stephen Chow
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> . . .

Scenario 3

Shaolin Soccer directed_by Stephen Chow
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> ...

Who wrote Kung Fu Hustle?

Scenario 3

Shaolin Soccer directed_by Stephen Chow
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

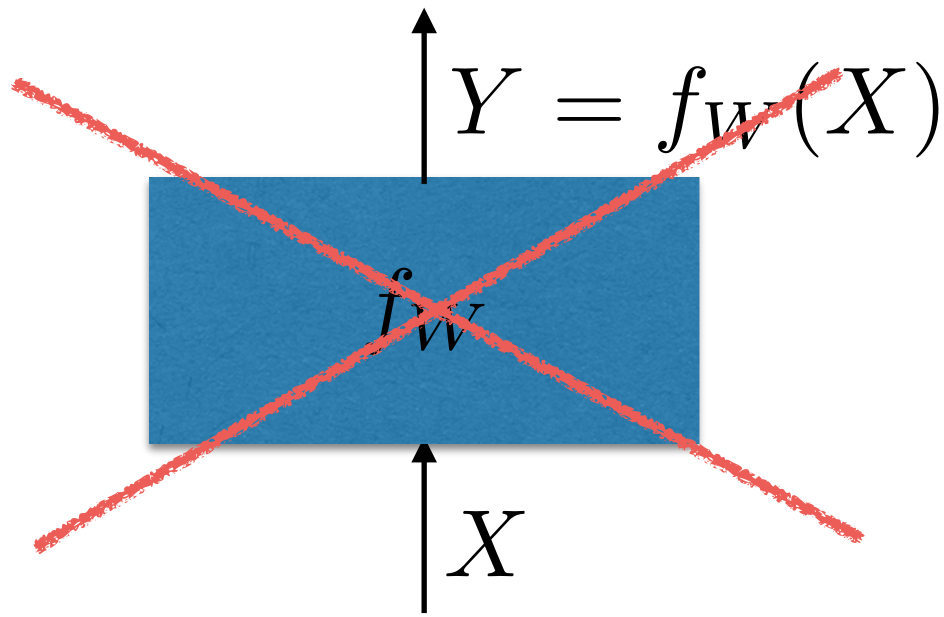
. . . <and more> . . .

I'm interested in watching a Stephen Chow movie other than Kung Fu Hustle. Can you suggest something?

Scenario 4

User	Bot
Hi!	
	Hello. What can I help you with today?
I’d like to book a table for 6 in an expensive restaurant which serves French food.	
	I’m on it. Where should it be?
In London	
	Ok. Let me look into some options for you.
Actually I would prefer for four.	
	Sure. Is there anything else to update?
No.	
	Ok. Let me look into some options for you.
	What do you think of The_Fancy_Pub
No. I don’t like that.	
	Sure. Let me find another option.
	How about The_palace
Lets do it.	
	Great. Let me make the reservation.
Can you provide me with the address?	
	Here it is: The_Palace_Address You roci
You rock!	
	Is there anything else i can help you with
No. Thanks.	
	You are welcome.

What is Required?



Not all problems can be mapped to y = f(x)

The model needs to remember external context

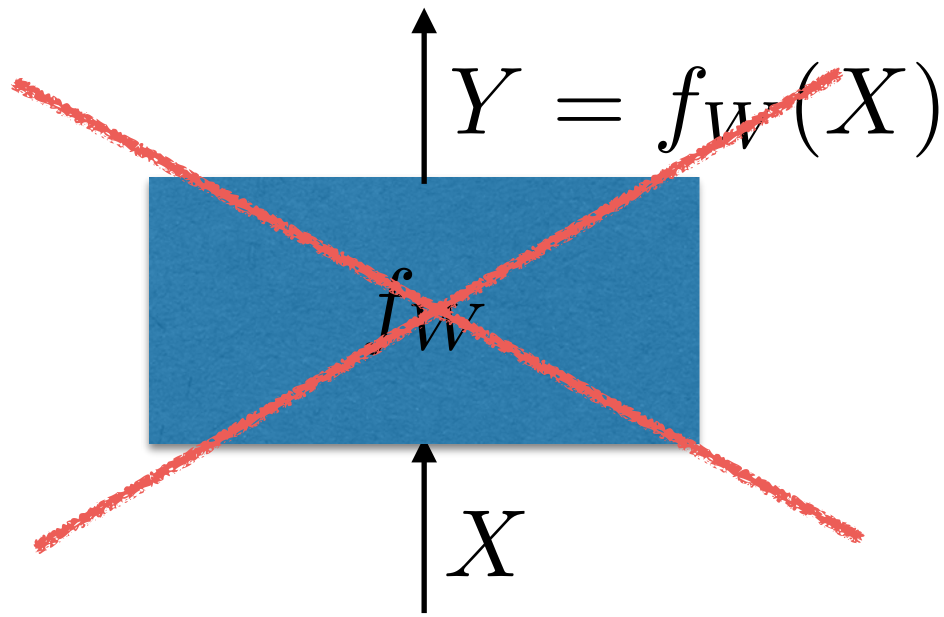
Given an input, the model needs to know where to look for in the context

It needs to know what to look for in the context

It needs to know how to reason using this external context

It needs to handle the potentially changing external context

What is Required?



Not all problems can be mapped to y = f(x)

The model needs to remember external context

Given an input, th … for in the context

It ne … ntext

Needs to have a notion of Memory

It needs t … al context

It needs to handle the potentially changing external context

Possible Solution

Hidden states of RNNs have memory

Run an RNN on the context/story/KB and get its representation

Use the representation to map question to answers/response

Possible Solution

Hidden states of RNNs have memory

Run an RNN on the context/story/KB and get its representation

Use the representation to map question to answers/response

We know this will not scale!

Outline

Memory Networks

Fully Supervised MemNNs

End2End MemNNs

Key-Value MemNNs

Architecture - How to reason - Advantages/Disadvantages

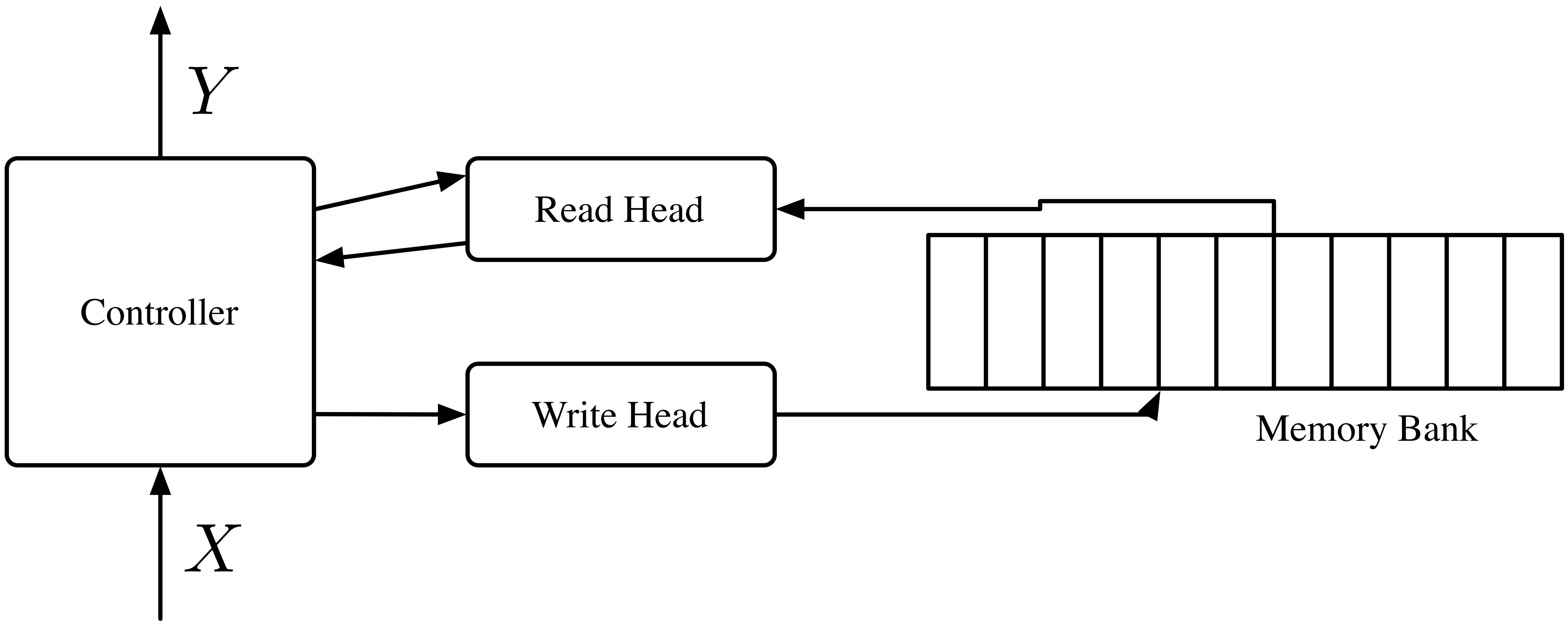
Neural Turing Machines

Architecture - How to reason - Advantages/Disadvantages

Stack/List/Queue Augmented RNNs

If time permits - otherwise you'll hear about this in lot more detail tomorrow

General Architecture



Controller takes external inputs and controls the heads

Heads read from and write to the memory

Controller combines memory reads with external input to produce an external output

What goes inside each of these components defines the model

Memory Networks

Class of models which combine large memory with learning component which can read and write to it

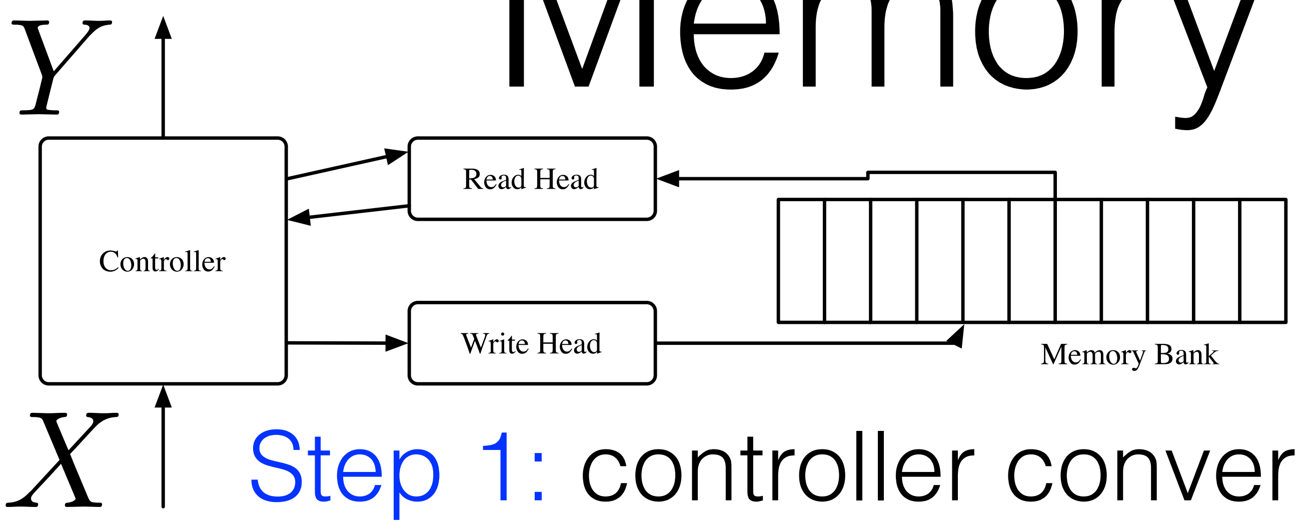
Incorporates reasoning via attention over memory

The model framework is flexible enough to store rich representations of input in memory

Models are scalable - can store and read large amount of data in memory - entire KB

Memory specification is flexible - can have both long-term memory and short-term memory - consider dialog modeling

Memory Networks



Step 1: controller converts incoming data to internal feature representation (I)

Step 2: write head updates the memories and writes the data into memory (G)

Step 3: given the external input, the read head reads the memory and fetches relevant data (O)

Step 4: controller combines the external data with memory contents returned by read head to generate output (O, R)

Memory Networks (Fully Supervised)

John was in the bathroom.
Bob was in the office.
John went to kitchen.
Bob travelled back home.

Context

Memory Networks (Fully Supervised)

John was in the bathroom.
Bob was in the office.
John went to kitchen.
Bob travelled back home.
Where is John? A: kitchen

Context

Question, Answer Pair

Memory Networks (Fully Supervised)

John was in the bathroom.
Bob was in the office.
John went to kitchen.
Bob travelled back home.
Where is John? A: kitchen

Context
Supporting Fact
Question, Answer Pair

Memory Networks (Fully Supervised)

John was in the bathroom.
Bob was in the office.
John went to kitchen.
Bob travelled back home.
Where is John? A: kitchen

Memories

$$m_i = f(John\ was\ in\ the\ bathroom.)$$
$$m_{i+1} = f(Bob\ was\ in\ the\ office.)$$
$$m_{i+2} = f(John\ went\ to\ the\ kitchen.)$$
$$m_{i+3} = f(Bob\ travelled\ back\ home.)$$

Step 1

Store the representations of facts in the memory
Free to choose what representations you store
Individual words - window of words - full sentences
Bag-of-words - CNN - RNN - LSTM

Memory Networks (Fully Supervised)

John was in the bathroom.
Bob was in the office.
John went to kitchen.
Bob travelled back home.
Where is John? A: kitchen

Memories

$$m_i = f(John\ was\ in\ the\ bathroom.)$$
$$m_{i+1} = f(Bob\ was\ in\ the\ office.)$$
$$m_{i+2} = f(John\ went\ to\ the\ kitchen.)$$
$$m_{i+3} = f(Bob\ travelled\ back\ home.)$$

$$x = f(Where\ is\ John?)$$

Step 2

Represent the question using similar function.

Memory Networks (Fully Supervised)

John was in the bathroom.
Bob was in the office.
John went to kitchen.
Bob travelled back home.
Where is John? A: kitchen

Memories

$$m_i = f(John\ was\ in\ the\ bathroom.)$$
$$m_{i+1} = f(Bob\ was\ in\ the\ office.)$$
$$m_{i+2} = f(John\ went\ to\ the\ kitchen.)$$
$$m_{i+3} = f(Bob\ travelled\ back\ home.)$$

$$x = f(Where\ is\ John?)$$

Step 3

Define a scoring function ***S*** and score the memories with the question

Scoring function should be such that it gives a high score to the relevant memories:

S(Where is John?, John went to the kitchen.) > S(Where is John?, Bob travelled back home.)

Memory Networks (Fully Supervised)

John was in the bathroom.
Bob was in the office.
John went to kitchen
Bob travelled b
Where is John?

Memories

$m_i = f(John\ was\ in\ the\ bathroom.)$

$s\ in\ the\ office.)$

$ent\ to\ the\ kitchen.)$

$velled\ back\ home.)$

$ere\ is\ John?)$

Example Choices

$$qU^tUd$$

$$G_w(q,d)$$

Define a scor   the question
Scoring functio   to the relevant
memories:

S(Where is John?, John went to the kitchen.) > S(Where is John?, Bob travelled back home.)

Memory Networks (Fully Supervised)

John was in the bathroom.
Bob was in the office.
John went to kitchen.
Bob travelled back home.
Where is John? A: kitchen

Memories

$$m_i = f(John\ was\ in\ the\ bathroom.)$$
$$m_{i+1} = f(Bob\ was\ in\ the\ office.)$$
$$m_{i+2} = f(John\ went\ to\ the\ kitchen.)$$
$$m_{i+3} = f(Bob\ travelled\ back\ home.)$$

$$x = f(Where\ is\ John?)$$

Step 4

Define another parametric function which maps the current question and relevant memories to the final response

In the first experiments, this was another scoring function which scored all possible responses against the given input and memories

Memory Networks (Fully Supervised)

John was in the bathroom.
Bob was in the office.
John went to kitchen.
Bob travelled back home.
Where is John? A: kitchen

Memories

$$m_i = f(John\ was\ in\ the\ bathroom.)$$
$$m_{i+1} = f(Bob\ was\ in\ the\ office.)$$
$$m_{i+2} = f(John\ went\ to\ the\ kitchen.)$$
$$m_{i+3} = f(Bob\ travelled\ back\ home.)$$

$$x = f(Where\ is\ John?)$$

Inference

Given the question, pick the memory which scores the highest

Use the selected memory and the question to generate the answer

Memory Networks (Fully Supervised)

Training

It involves training the memory representations and the scoring functions to generate answer

We do so my minimizing the following loss

$$L = \sum_{\bar{f} \neq m_{o1}} max(0, \gamma - S_o(x, m_{o1}) + S_o(x, \bar{f})) +$$

$$\sum_{\bar{r} \neq r} max(0, \gamma - S_r([x, m_{o1}], r) + S_r([x, m_{o1}], \bar{r}))$$

Memories

$$m_i = f(John\ was\ in\ the\ bathroom.)$$

$$m_{i+1} = f(Bob\ was\ in\ the\ office.)$$

$$m_{i+2} = f(John\ went\ to\ the\ kitchen.)$$

$$m_{i+3} = f(Bob\ travelled\ back\ home.)$$

$$x = f(Where\ is\ John?)$$

Memory Networks (Fully Supervised)

Training

It involves training the memory representations and the scoring functions to generate answer

We do so my minimizing the following loss

We had access to true supporting fact during training that's what we mean by "Fully Supervised"

$$L = \sum_{\bar{f} \neq m_{o1}} max(0, \gamma - S_o(x, m_{o1}) + S_o(x, \bar{f})) +$$

$$\sum_{\bar{r} \neq r} max(0, \gamma - S_r([x, m_{o1}], r) + S_r([x, m_{o1}], \bar{r}))$$

Memories

$$m_i = f(John\ was\ in\ the\ bathroom.)$$

$$m_{i+1} = f(Bob\ was\ in\ the\ office.)$$

$$m_{i+2} = f(John\ went\ to\ the\ kitchen.)$$

$$m_{i+3} = f(Bob\ travelled\ back\ home.)$$

$$x = f(Where\ is\ John?)$$

$S_o :$ $scoring\ function\ for\ memories$

$S_r :$ $scoring\ function\ for\ responses$

Memory Networks (Fully Supervised)

Training

It involves training the memory representations and the scoring functions to generate answer

We do so my minimizing the following loss

We had access to true supporting fact during training that's what we mean by "Fully Supervised"

$$L = \sum_{\bar{f} \neq m_{o1}} max(0, \gamma - S_o(x, m_{o1}) + S_o(x, \bar{f})) +$$

$$\sum_{\bar{r} \neq r} max(0, \gamma - S_r([x, m_{o1}], r) + S_r([x, m_{o1}], \bar{r}))$$

Memories

$$m_i = f(John\ was\ in\ the\ bathroom.)$$
$$m_{i+1} = f(Bob\ was\ in\ the\ office.)$$
$$m_{i+2} = f(John\ went\ to\ the\ kitchen.)$$
$$m_{i+3} = f(Bob\ travelled\ back\ home.)$$

$$x = f(Where\ is\ John?)$$

$S_o :$ *scoring function for memories*

$S_r :$ *scoring function for responses*

This was the case when we have a single supporting fact!

Memory Networks (Fully Supervised)

John is in the playground. ← Supporting Fact 2

Bob is in the office.

John picked up the football. ← Supporting Fact 1

Bob went to the kitchen.

Where is the football? A: playground.

Memory Networks (Fully Supervised)

John is in the playground. ← Supporting Fact 2
Bob is in the office.
John picked up the football. ← Supporting Fact 1
Bob went to the kitchen.
Where is the football? A: playground.

The current loss function will not work

$$L = \sum_{\bar{f} \neq m_{o1}} max(0, \gamma - S_o(x, m_{o1}) + S_o(x, \bar{f}))+$$
$$\sum_{\bar{r} \neq r} max(0, \gamma - S_r([x, m_{o1}], r) + S_r([x, m_{o1}], \bar{r}))$$

Memory Networks (Fully Supervised)

John is in the playground. ← Supporting Fact 2
Bob is in the office.
John picked up the football. ← Supporting Fact 1
Bob went to the kitchen.
Where is the football? A: playground.

The current loss function will not work

$$L = \sum_{\bar{f} \neq m_{o1}} max(0, \gamma - S_o(x, m_{o1}) + S_o(x, \bar{f})) +$$
$$\sum_{\bar{r} \neq r} max(0, \gamma - S_r([x, m_{o1}], r) + S_r([x, m_{o1}], \bar{r}))$$

But the cool thing is that we can iterate!

Memory Networks (Fully Supervised)

John is in the playground. ← Supporting Fact 2
Bob is in the office.
John picked up the football. ← Supporting Fact 1
Bob went to the kitchen.
Where is the football? A: playground.

$$
\begin{aligned}
Loss = &\sum_{\bar{f} \neq m_{o1}} max(0, \gamma - S_o(x, m_{o1}) + S_o(x, \bar{f})) \\
&+ \sum_{\bar{f}' \neq m_{o2}} max(0, \gamma - S_o([x, m_{o1}], m_{o2}) + S_o([x, m_{o1}], \bar{f}')) \\
&+ \sum_{\bar{r} \neq r} max(0, \gamma - S_r([x, m_{o1}, m_{o2}], r) + S_r([x, m_{o1}, m_{o2}], \bar{r}))
\end{aligned}
$$

Memory Networks (Fully Supervised)

John is in the playground. ← Supporting Fact 2
Bob is in the office.
John picked up the football. ← Supporting Fact 1
Bob went to the kitchen.
Where is the football? A: playground.

$$Loss = \sum_{\bar{f} \neq m_{o1}} max(0, \gamma - S_o(x, m_{o1}) + S_o(x, \bar{f}))$$
$$+ \sum_{\bar{f}' \neq m_{o2}} max(0, \gamma - S_o([x, m_{o1}], m_{o2}) + S_o([x, m_{o1}], \bar{f}'))$$
$$+ \sum_{\bar{r} \neq r} max(0, \gamma - S_r([x, m_{o1}, m_{o2}], r) + S_r([x, m_{o1}, m_{o2}], \bar{r}))$$

Memory Networks (Fully Supervised)

John is in the playground. ← Supporting Fact 2

Bob is in the office.

John picked up the football. ← Supporting Fact 1

Bob went to the kitchen.

Where is the football? A: playground.

$$Loss = \sum_{\bar{f} \neq m_{o1}} max(0, \gamma - S_o(x, m_{o1}) + S_o(x, \bar{f}))$$
$$+ \sum_{\bar{f}' \neq m_{o2}} max(0, \gamma - S_o([x, m_{o1}], m_{o2}) + S_o([x, m_{o1}], \bar{f}'))$$
$$+ \sum_{\bar{r} \neq r} max(0, \gamma - S_r([x, m_{o1}, m_{o2}], r) + S_r([x, m_{o1}, m_{o2}], \bar{r}))$$

Memory Networks (Fully Supervised)

John is in the playground. ← Supporting Fact 2

Bob is in the office.

John picked up the football. ← Supporting Fact 1

Bob went to the kitchen.

Where is the f

We call these "Hops"

And they are not limited to two

$$S_o(x, \bar{f}))$$

$$+ \sum_{\bar{f}' \neq m_{o2}} max(0, \gamma - S_o([x, m_{o1}], m_{o2}) + S_o([x, m_{o1}], \bar{f}'))$$

$$+ \sum_{\bar{r} \neq r} max(0, \gamma - S_r([x, m_{o1}, m_{o2}], r) + S_r([x, m_{o1}, m_{o2}], \bar{r}))$$

bAbI Dataset: Slight Digression

While working on MemNNs we also defined 20 simulated tasks to test models which have long-term memory — can do complex reasoning using those memories

The objective was to generate a set of tasks which can act "unit tests" in software engineering

Each task would test a single (or may be a couple of) "skills" which we think are natural to humans w.r.t. text understanding and reasoning

Language skills - conjunction, coreference, negation etc

Reasoning skills - counting, path finding etc

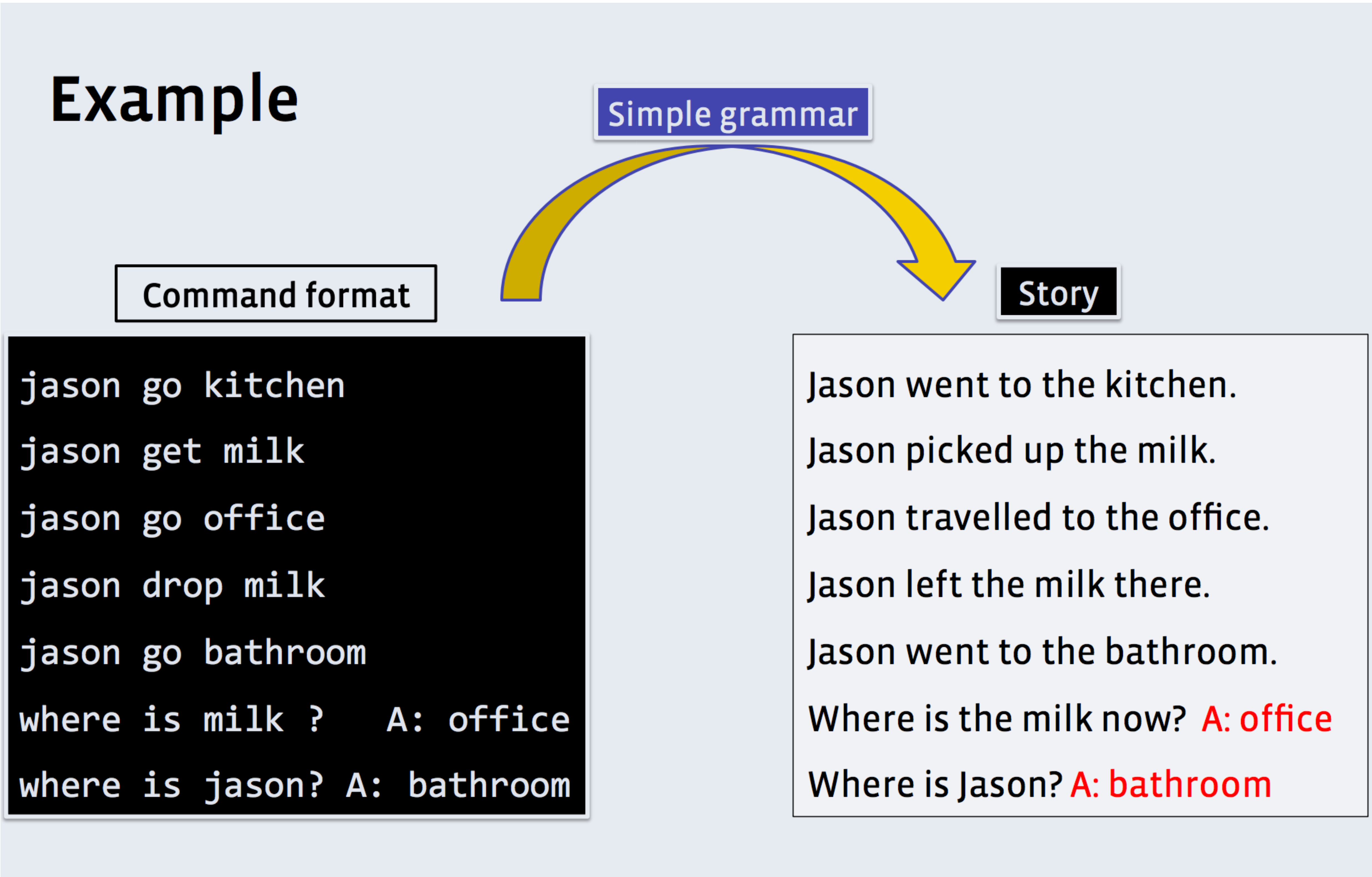
bAbI Dataset: Simulator

- go <place>
- get <object>
- get <object1> from <object2>
- put <object1> in/on <object2>
- give <object> to <person>
- drop <object>
- look
- inventory
- examine <object>

+ 2 commands for "gods" (superusers):

- create <object>
- set <obj1> <relation> <obj2>

bAbI Dataset: Simulator



bAbI Dataset

Factoid QA with Single Supporting Fact

Questions where a single supporting fact is used and it is given in the context

We test this by asking for location of a person

John is in the playground.
Bob is in the office.
Where is John? A:playground

SUPPORTING FACT

bAbI Dataset

Factoid QA with Two Supporting Facts

Questions where two supporting facts have to be chained together in order to find the answer

John is in the playground. ← SUPPORTING FACT
Bob is in the office.
John picked up the football. ← SUPPORTING FACT
Bob went to the kitchen.
Where is the football? A:playground

Factoid QA with Three Supporting Facts

Questions where Three supporting facts have to be chained together in order to find the answer

John picked up the apple.
John went to the office.
John went to the kitchen.
John dropped the apple.
Where was the apple before the kitchen? A:office

bAbI Dataset

Two Argument Relations: Subject vs. Object

Questions where the model learns the ability to differentiate and recognize subjects and objects

We make the problem harder by having sentences which have re-ordered words

For example the two questions below have same words but different meaning

The office is north of the bedroom.
The bedroom is north of the bathroom.
What is north of the bedroom? A:office
What is the bedroom north of? A:bathroom

bAbI Dataset

Three Argument Relations

Questions where the model learns the ability to differentiate and recognize two subjects and an object

Mary gave the cake to Fred.
Fred gave the cake to Bill.
Jeff was given the milk by Bill.
Who gave the cake to Fred? A: Mary
Who did Fred give the cake to? A: Bill

bAbI Dataset

Yes/No Questions

Questions where the model learns answer true/false type questions

Start with the simple case of a single supporting fact

John is in the playground.
Daniel picks up the milk.
Is John in the classroom? A:no
Does Daniel have the milk? A:yes

bAbI Dataset

Counting

Questions where the model learns to count

Daniel picked up the football.
Daniel dropped the football.
Daniel got the milk.
Daniel took the apple.
How many objects is Daniel holding? A:two

Lists/Sets

Questions where the model learns to generate a set or list of answers

Daniel picks up the football.
Daniel drops the newspaper.
Daniel picks up the milk.
What is Daniel holding? A:milk,football

bAbI Dataset

Indefinite Knowledge

Questions where the model learns to answer under uncertainty

John is either in the classroom or the playground.
Sandra is in the garden.
Is John in the classroom? A:maybe
Is John in the office? A:no

bAbI Dataset

Basic Coreference

Questions where the model learns to recognize coreferences of a single entity

Daniel was in the kitchen.
Then he went to the studio.
Sandra was in the office.
Where is Daniel? A:studio

Compound Coreferences

Questions where the model learns to recognize coreferences of multiple entities

Daniel and Sandra journeyed to the office.
Then they went to the garden.
Sandra and John travelled to the kitchen.
After that they moved to the hallway.
Where is Daniel? A:garden

bAbI Dataset

Time Manipulation

While we have an implicit notion of time already in our tasks, this particular one tests understanding the use of explicit time expressions

In the afternoon Julie went to the park.
Yesterday Julie was at school.
Julie went to the cinema this evening.
Where did Julie go after the park? A:cinema

Basic Deduction

Questions where the model learns basic deduction via inheritance of properties

Sheep are afraid of wolves.
Cats are afraid of dogs.
Mice are afraid of cats.
Gertrude is a sheep.
What is Gertrude afraid of? A:wolves

Deduction for MemNNs should be hard because it effectively involves search.

bAbI Dataset

Positional Reasoning

Questions where the model learns to do spatial reasoning

The triangle is to the right of the blue square.
The red square is on top of the blue square.
The red sphere is to the right of the blue square.
Is the red sphere to the right of the blue square? A:yes
Is the red square to the left of the triangle? A:yes

Reasoning About Size

Questions where the model learns to reason about relative sizes of objects.

Inspired by the commonsense reasoning examples in the Winograd Schema Challenge

The football fits in the suitcase.
The suitcase fits in the cupboard.
The box of chocolates is smaller than the football.
Will the box of chocolates fit in the suitcase? A:yes

Task of three supporting facts and Yes/No questions are prerequisites.

bAbI Dataset

Path Finding

Questions in which the model learns to find a path between two locations.

The kitchen is north of the hallway.
The den is east of the hallway.
How do you go from den to kitchen? A:west,north

Path Finding for MemNNs should be hard because it effectively involves search.

bAbI Dataset

Agent's Motivation

Questions in which the model learns to find the reason behind an agent's action

John is hungry.
John goes to the kitchen.
John grabbed the apple there.
Daniel is hungry.
Where does Daniel go? A:kitchen
Why did John go to the kitchen? A:hungry

MemNNs on bAbI

Baselines

Structured SVM with a collection of hand coded features - classic NLP stack

LSTM

ngram classifiers

MemNNs on bAbI

Baselines

Structured SVM with a collection of hand coded features - classic NLP stack

LSTM

ngram classifiers

	Weakly Supervised		Uses External Resources	Strong Supervision (using supporting facts)						
TASK	N-gram Classifier	LSTM	Structured SVM COREF+SRL features	MemNN Weston et al. (2014)	MemNN ADAPTIVE MEMORY	MemNN AM + N-GRAMS	MemNN AM + NONLINEAR	MemNN AM + NG + NL	No. of ex. req. ≥ 95	MultiTask Training
1 - Single Supporting Fact	36	50	99	100	100	100	100	100	250 ex.	100
2 - Two Supporting Facts	2	20	74	100	100	100	100	100	500 ex.	100
3 - Three Supporting Facts	7	20	17	20	**100**	**99**	**100**	**100**	500 ex.	**98**
4 - Two Arg. Relations	50	61	98	71	69	**100**	73	**100**	500 ex.	80
5 - Three Arg. Relations	20	70	83	83	83	86	86	**98**	1000 ex.	**99**
6 - Yes/No Questions	49	48	99	47	52	53	**100**	**100**	500 ex.	**100**
7 - Counting	52	49	69	68	78	86	83	85	FAIL	86
8 - Lists/Sets	40	45	70	77	90	88	94	91	FAIL	93
9 - Simple Negation	62	64	100	65	71	63	**100**	**100**	500 ex.	**100**
10 - Indefinite Knowledge	45	44	99	59	57	54	**97**	**98**	1000 ex.	**98**
11 - Basic Coreference	29	72	100	100	100	100	100	100	250 ex.	100
12 - Conjunction	9	74	96	100	100	100	100	100	250 ex.	100
13 - Compound Coref.	26	94	99	100	100	100	100	100	250 ex.	100
14 - Time Reasoning	19	27	99	99	100	99	100	99	500 ex.	99
15 - Basic Deduction	20	21	96	74	73	**100**	77	**100**	100 ex.	**100**
16 - Basic Induction	43	23	24	27	**100**	**100**	**100**	**100**	100 ex.	94
17 - Positional Reasoning	46	51	61	54	46	49	57	65	FAIL	72
18 - Size Reasoning	52	52	62	57	50	74	54	**95**	1000 ex.	93
19 - Path Finding	0	8	49	0	9	3	15	36	FAIL	19
20 - Agent's Motivations	76	91	95	100	100	100	100	100	250 ex.	100
Mean Performance	34	49	79	75	79	83	87	93		92

MemNNs on bAbI

Baselines

Structured SVM with a collection of hand coded features - classic NLP stack

LSTM

ngram classifiers

	Weakly Supervised		Uses External Resources	Strong Supervision (using supporting facts)						
TASK	N-gram Classifier	LSTM	Structured SVM COREF+SRL features	MemNN Weston et al. (2014)	MemNN ADAPTIVE MEMORY	MemNN AM + N-GRAMS	MemNN AM + NONLINEAR	MemNN AM + NG + NL	No. of ex. req. ≥ 95	MultiTask Training
1 - Single Supporting Fact	36	50	99	100	100	100	100	100	250 ex.	100
2 - Two Supporting Facts	2	20	74	100	100	100	100	100	500 ex.	100
3 - Three Supporting Facts	7	20	17	20	**100**	**99**	**100**	**100**	500 ex.	**98**
4 - Two Arg. Relations	50	61	98	71	69	**100**	73	**100**	500 ex.	80
5 - Three Arg. Relations	20	70	83	83	83	86	86	**98**	1000 ex.	**99**
6 - Yes/No Questions	49	48	99	47	52	53	**100**	**100**	500 ex.	**100**
7 - Counting	52	49	69	68	78	86	83	85	FAIL	86
8 - Lists/Sets	40	45	70	77	90	88	94	91	FAIL	93
9 - Simple Negation	62	64	100	65	71	63	**100**	**100**	500 ex.	**100**
10 - Indefinite Knowledge	45	44	99	59	57	54	**97**	**98**	1000 ex.	**98**
11 - Basic Coreference	29	72	100	100	100	100	100	100	250 ex.	100
12 - Conjunction	9	74	96	100	100	100	100	100	250 ex.	100
13 - Compound Coref.	26	94	99	100	100	100	100	100	250 ex.	100
14 - Time Reasoning	19	27	99	99	100	99	100	99	500 ex.	99
15 - Basic Deduction	20	21	96	74	73	**100**	77	**100**	100 ex.	**100**
16 - Basic Induction	43	23	24	27	**100**	**100**	**100**	**100**	100 ex.	94
17 - Positional Reasoning	46	51	61	54	46	49	57	65	FAIL	72
18 - Size Reasoning	52	52	62	57	50	74	54	**95**	1000 ex.	93
19 - Path Finding	0	8	49	0	9	3	15	36	FAIL	19
20 - Agent's Motivations	76	91	95	100	100	100	100	100	250 ex.	100
Mean Performance	34	49	79	75	79	83	87	93		92

MemNNs on bAbI

Baselines

Structured SVM with a collection of hand coded features - classic NLP stack

LSTM

ngram classifiers

	Weakly Supervised		Uses External Resources	Strong Supervision (using supporting facts)						
TASK	N-gram Classifier	LSTM	Structured SVM COREF+SRL features	MemNN Weston et al. (2014)	MemNN ADAPTIVE MEMORY	MemNN AM + N-GRAMS	MemNN AM + NONLINEAR	MemNN AM + NG + NL	No. of ex. req. ≥ 95	MultiTask Training
1 - Single Supporting Fact	36	50	99	100	100	100	100	100	250 ex.	100
2 - Two Supporting Facts	2	20	74	100	100	100	100	100	500 ex.	100
3 - Three Supporting Facts	7	20	17	20	**100**	**99**	**100**	**100**	500 ex.	**98**
4 - Two Arg. Relations	50	61	98	71	69	**100**	73	**100**	500 ex.	80
5 - Three Arg. Relations	20	70	83	83	83	86	86	**98**	1000 ex.	**99**
6 - Yes/No Questions	49	48	99	47	52	53	**100**	**100**	500 ex.	**100**
7 - Counting	52	49	69	68	78	86	83	85	FAIL	86
8 - Lists/Sets	40	45	70	77	90	88	94	91	FAIL	93
9 - Simple Negation	62	64	100	65	71	63	**100**	**100**	500 ex.	**100**
10 - Indefinite Knowledge	45	44	99	59	57	54	**97**	**98**	1000 ex.	**98**
11 - Basic Coreference	29	72	100	100	100	100	100	100	250 ex.	100
12 - Conjunction	9	74	96	100	100	100	100	100	250 ex.	100
13 - Compound Coref.	26	94	99	100	100	100	100	100	250 ex.	100
14 - Time Reasoning	19	27	99	99	100	99	100	99	500 ex.	99
15 - Basic Deduction	20	21	96	74	73	**100**	77	**100**	100 ex.	**100**
16 - Basic Induction	43	23	24	27	**100**	**100**	**100**	**100**	100 ex.	94
17 - Positional Reasoning	46	51	61	54	46	49	57	65	FAIL	72
18 - Size Reasoning	52	52	62	57	50	74	54	**95**	1000 ex.	93
19 - Path Finding	0	8	49	0	9	3	15	36	FAIL	19
20 - Agent's Motivations	76	91	95	100	100	100	100	100	250 ex.	100
Mean Performance	34	49	79	75	79	83	87	93		92

MemNNs on bAbI

Baselines

Structured SVM with a collection of hand coded features - classic NLP stack

LSTM

ngram classifiers

TASK	Weakly Supervised		Uses External Resources	Strong Supervision (using supporting facts)						
	N-gram Classifier	LSTM	Structured SVM COREF+SRL features	MemNN Weston et al. (2014)	MemNN ADAPTIVE MEMORY	MemNN AM + N-GRAMS	MemNN AM + NONLINEAR	MemNN AM + NG + NL	No. of ex. req. ≥ 95	MultiTask Training
1 - Single Supporting Fact	36	50	99	100	100	100	100	100	250 ex.	100
2 - Two Supporting Facts	2	20	74	100	100	100	100	100	500 ex.	100
3 - Three Supporting Facts	7	20	17	20	**100**	**99**	**100**	**100**	500 ex.	**98**
4 - Two Arg. Relations	50	61	98	71	69	**100**	73	**100**	500 ex.	80
5 - Three Arg. Relations	20	70	83	83	83	86	86	**98**	1000 ex.	**99**
6 - Yes/No Questions	49	48	99	47	52	53	**100**	**100**	500 ex.	**100**
7 - Counting	52	49	69	68	78	86	83	85	FAIL	86
8 - Lists/Sets	40	45	70	77	90	88	94	91	FAIL	93
9 - Simple Negation	62	64	100	65	71	63	**100**	**100**	500 ex.	**100**
10 - Indefinite Knowledge	45	44	99	59	57	54	**97**	**98**	1000 ex.	**98**
11 - Basic Coreference	29	72	100	100	100	100	100	100	250 ex.	100
12 - Conjunction	9	74	96	100	100	100	100	100	250 ex.	100
13 - Compound Coref.	26	94	99	100	100	100	100	100	250 ex.	100
14 - Time Reasoning	19	27	99	99	100	99	100	99	500 ex.	99
15 - Basic Deduction	20	21	96	74	73	**100**	77	**100**	100 ex.	**100**
16 - Basic Induction	43	23	24	27	**100**	**100**	**100**	**100**	100 ex.	94
17 - Positional Reasoning	46	51	61	54	46	49	57	65	FAIL	72
18 - Size Reasoning	52	52	62	57	50	74	54	**95**	1000 ex.	93
19 - Path Finding	0	8	49	0	9	3	15	36	FAIL	19
20 - Agent's Motivations	76	91	95	100	100	100	100	100	250 ex.	100
Mean Performance	34	49	79	75	79	83	87	93		92

MemNNs on bAbI

Baselines

Structured SVM with a collection of hand coded features - classic NLP stack

LSTM

ngram classifiers

	Weakly Supervised		Uses External Resources	Strong Supervision (using supporting facts)						
TASK	N-gram Classifier	LSTM	Structured SVM COREF+SRL features	MemNN Weston et al. (2014)	MemNN ADAPTIVE MEMORY	MemNN AM + N-GRAMS	MemNN AM + NONLINEAR	MemNN AM + NG + NL	No. of ex. req. ≥ 95	MultiTask Training
1 - Single Supporting Fact	36	50	99	100	100	100	100	100	250 ex.	100
2 - Two Supporting Facts	2	20	74	100	100	100	100	100	500 ex.	100
3 - Three Supporting Facts	7	20	17	20	**100**	**99**	**100**	**100**	500 ex.	**98**
4 - Two Arg. Relations	50	61	98	71	69	**100**	73	**100**	500 ex.	80
5 - Three Arg. Relations	20	70	83	83	83	86	86	**98**	1000 ex.	**99**
6 - Yes/No Questions	49	48	99	47	52	53	**100**	**100**	500 ex.	**100**
7 - Counting	52	49	69	68	78	86	83	85	FAIL	86
8 - Lists/Sets	40	45	70	77	90	88	94	91	FAIL	93
9 - Simple Negation	62	64	100	65	71	63	**100**	**100**	500 ex.	**100**
10 - Indefinite Knowledge	45	44	99	59	57	54	**97**	**98**	1000 ex.	**98**
11 - Basic Coreference	29	72	100	100	100	100	100	100	250 ex.	100
12 - Conjunction	9	74	96	100	100	100	100	100	250 ex.	100
13 - Compound Coref.	26	94	99	100	100	100	100	100	250 ex.	100
14 - Time Reasoning	19	27	99	99	100	99	100	99	500 ex.	99
15 - Basic Deduction	20	21	96	74	73	**100**	77	**100**	100 ex.	**100**
16 - Basic Induction	43	23	24	27	**100**	**100**	**100**	**100**	100 ex.	94
17 - Positional Reasoning	46	51	61	54	46	49	57	65	FAIL	72
18 - Size Reasoning	52	52	62	57	50	74	54	**95**	1000 ex.	93
19 - Path Finding	0	8	49	0	9	3	15	36	FAIL	19
20 - Agent's Motivations	76	91	95	100	100	100	100	100	250 ex.	100
Mean Performance	34	49	79	75	79	83	87	93		92

Full Supervision in MemNNs

John was in the bathroom.
Bob was in the office.
John went to kitchen.
Bob travelled back home.
Where is John? A: kitchen

Context
Supporting Fact
Question, Answer Pair

Full Supervision in MemNNs

John was in the bathroom.
Bob was in the office.
John went to kitchen.
Bob travelled back home.
Where is John? A: kitchen

Context
Supporting Fact
Question, Answer Pair

$$x = f(Where\ is\ John?)$$

$$m_i = f(John\ was\ in\ the\ bathroom.)$$
$$m_{i+1} = f(Bob\ was\ in\ the\ office.)$$
$$m_{i+2} = f(John\ went\ to\ the\ kitchen.)$$
$$m_{i+3} = f(Bob\ travelled\ back\ home.)$$

Full Supervision in MemNNs

John was in the bathroom.

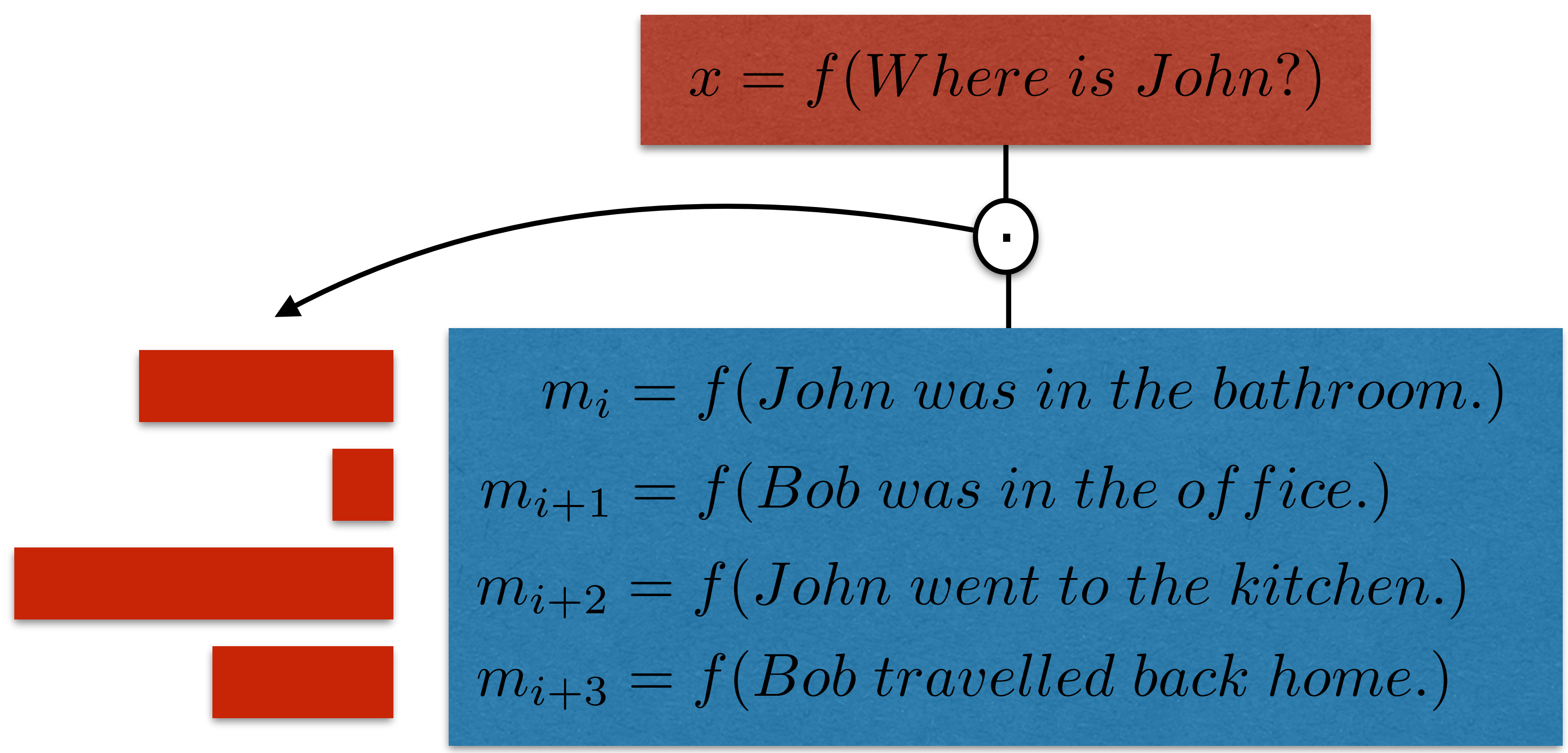
Bob was in the office.

John went to kitchen. ← Supporting Fact

Bob travelled back home.

Where is John? A: kitchen ← Question, Answer Pair

Context



Full Supervision in MemNNs

John was in the bathroom.

Bob was in the office.

John went to kitchen.

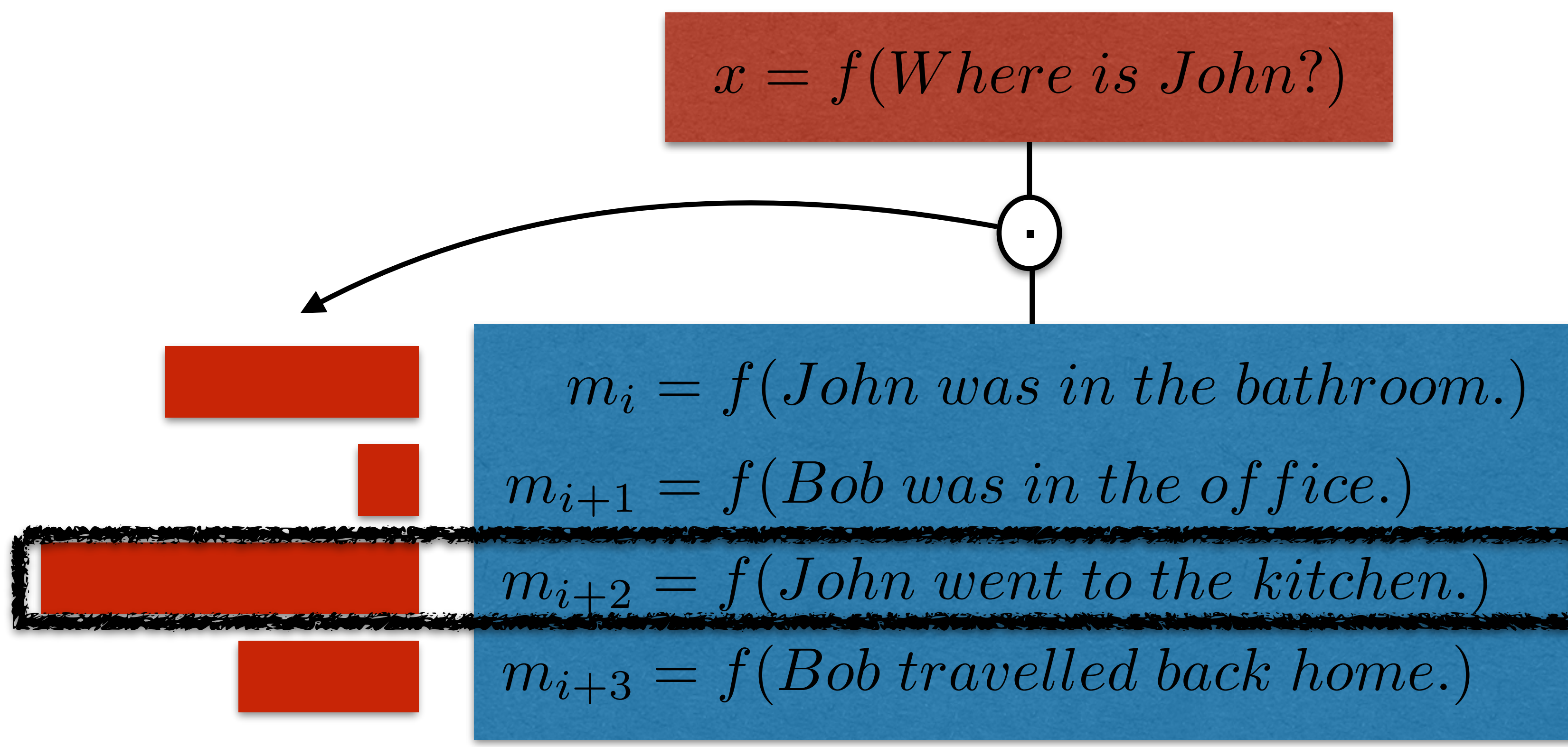
Bob travelled back home.

Where is John? A: kitchen

Context

Supporting Fact

Question, Answer Pair

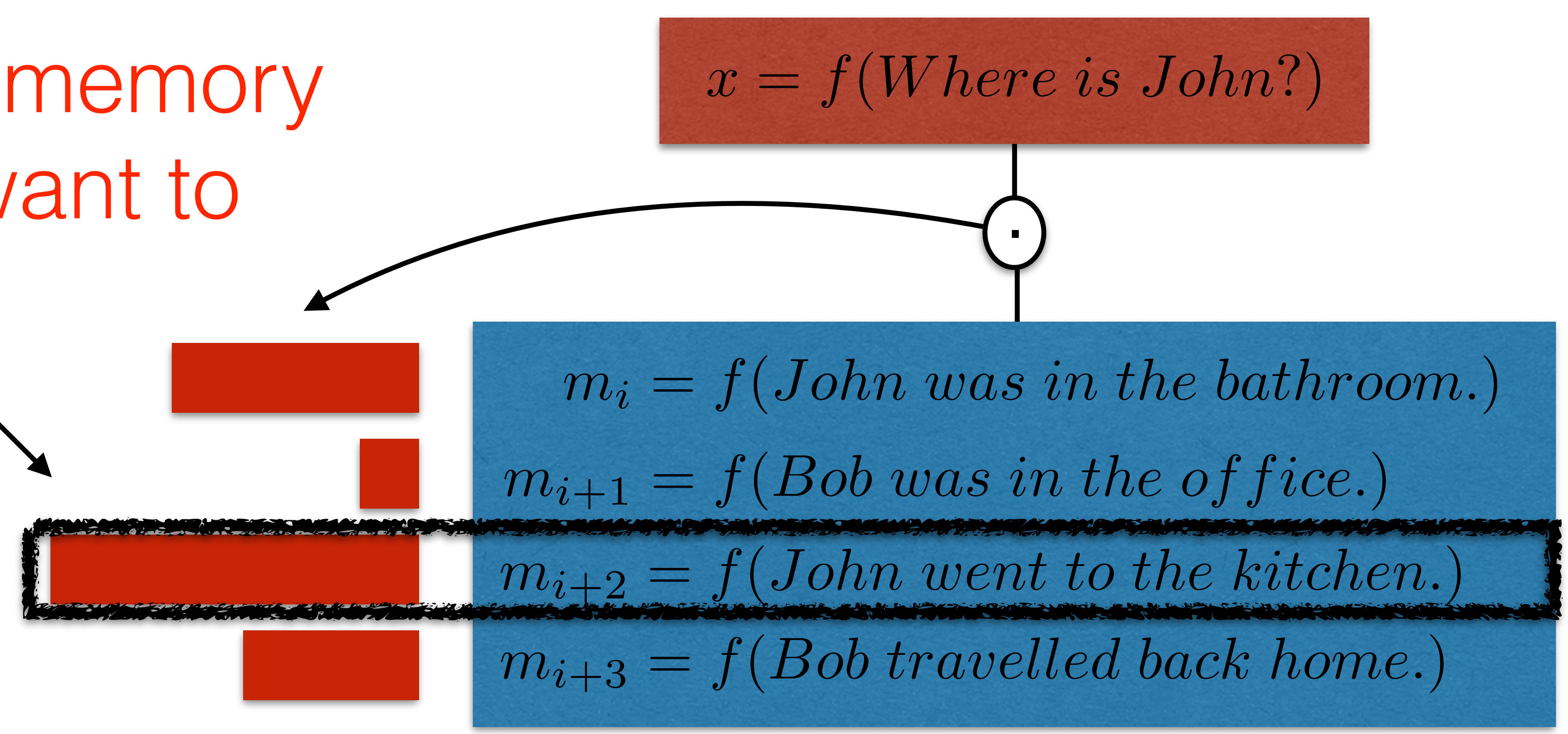


Full Supervision in MemNNs

John was in the bathroom.
Bob was in the office.
John went to kitchen.
Bob travelled back home.
Where is John? A: kitchen

Context
Supporting Fact
Question, Answer Pair

That's your retrieved memory whose score you want to push higher



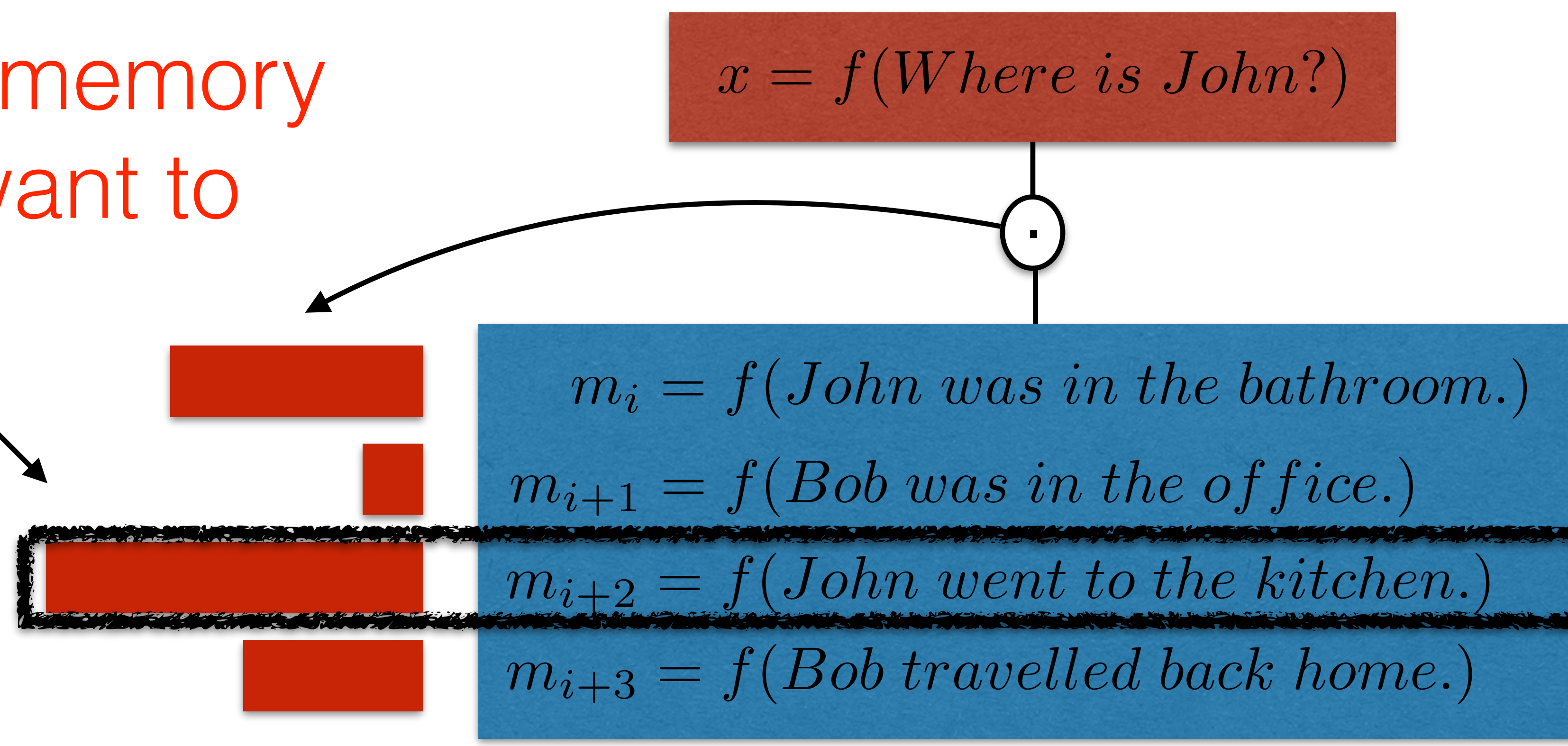
Full Supervision in MemNNs

John was in the bathroom.
Bob was in the office.
John went to kitchen.
Bob travelled back home.
Where is John? A: kitchen

Context
Supporting Fact
Question, Answer Pair

That's your retrieved memory whose score you want to push higher

This is like hard attention except that you already know where to attend!



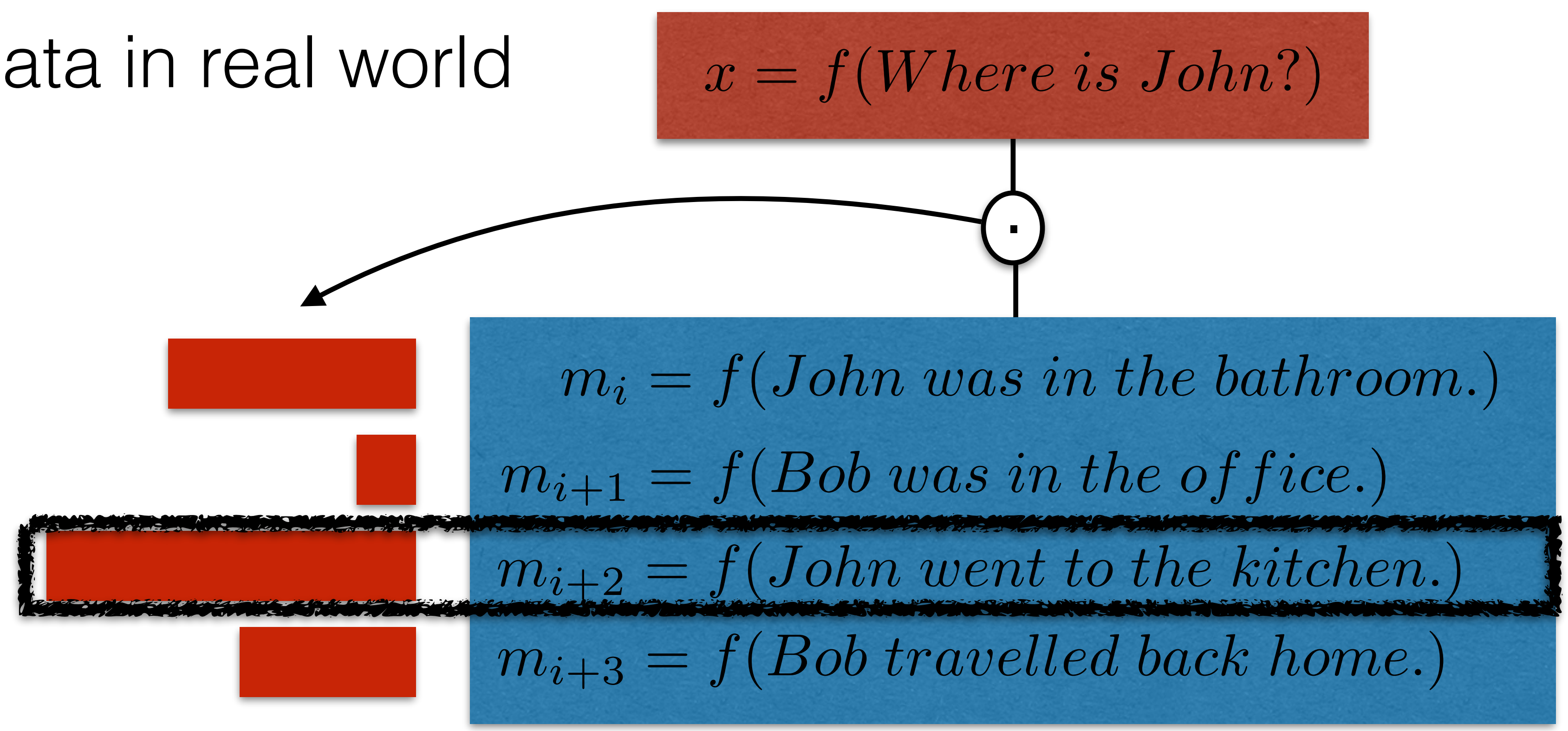
Full Supervision in MemNNs

Drawbacks

Fairly hard assumption to make

Not the most natural scenario

Expensive to get such data in real world



This is like hard attention except that you already know where to attend!

End2End MemNNs

No current supporting fact supplied

Learns which parts of the memory are relevant

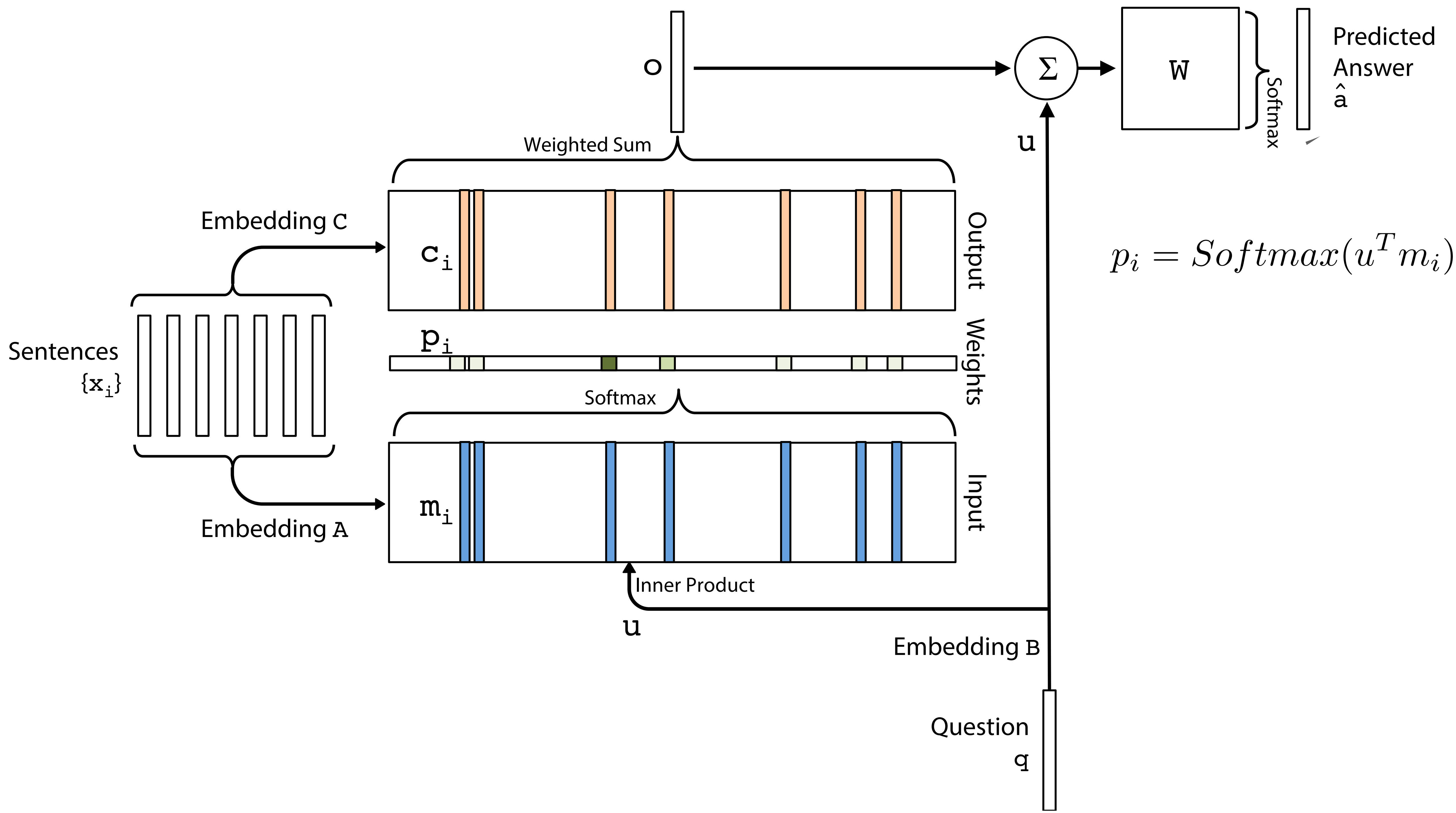
This is achieved by reading using soft attention as opposed to hard

Performs multiple lookups to refine its guess about memory relevance

The whole architecture is end-to-end differentiable

Only needs supervision at the final output

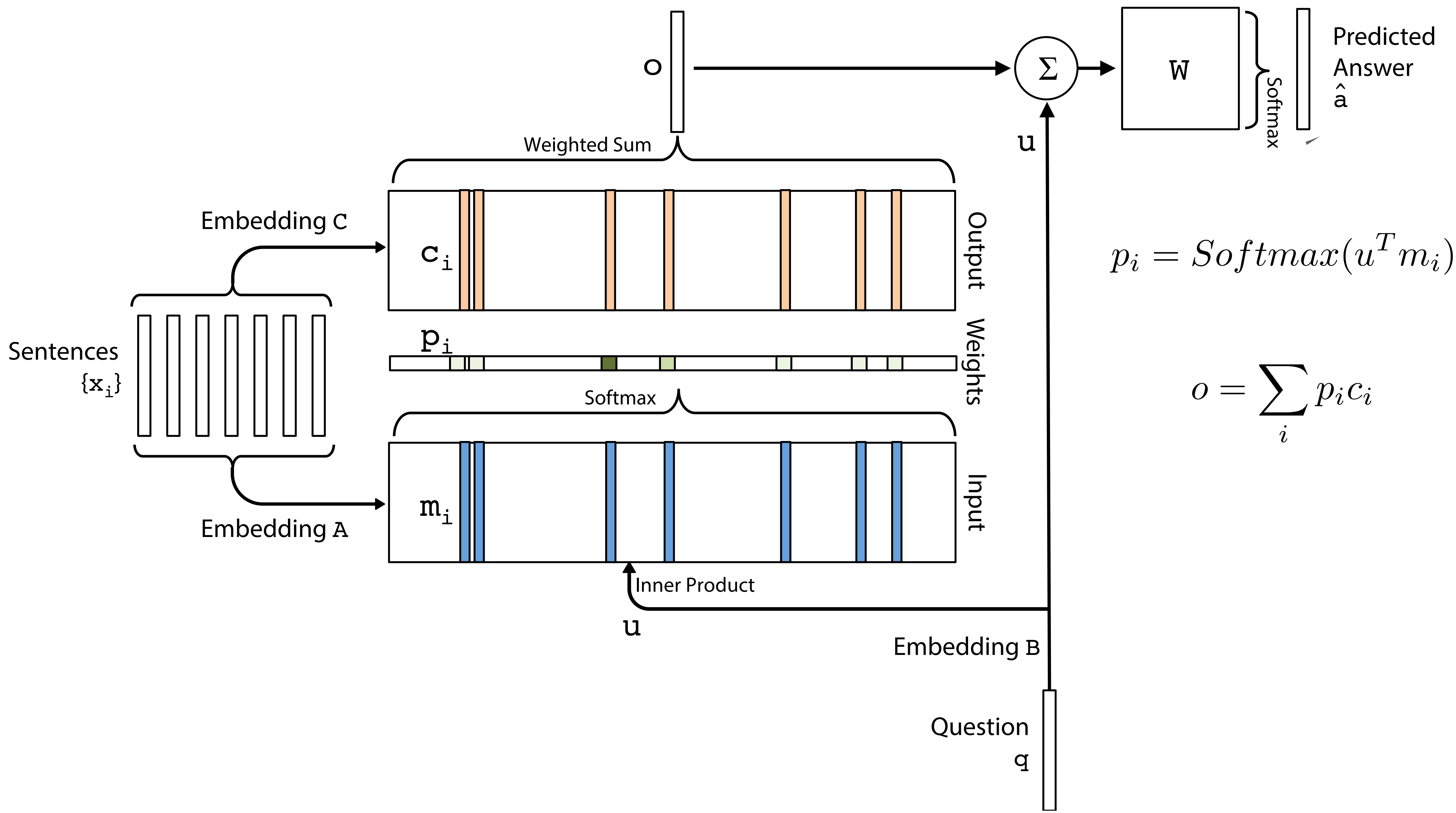
End2End MemNNs



$$p_i = Softmax(u^T m_i)$$

Single Layer

End2End MemNNs

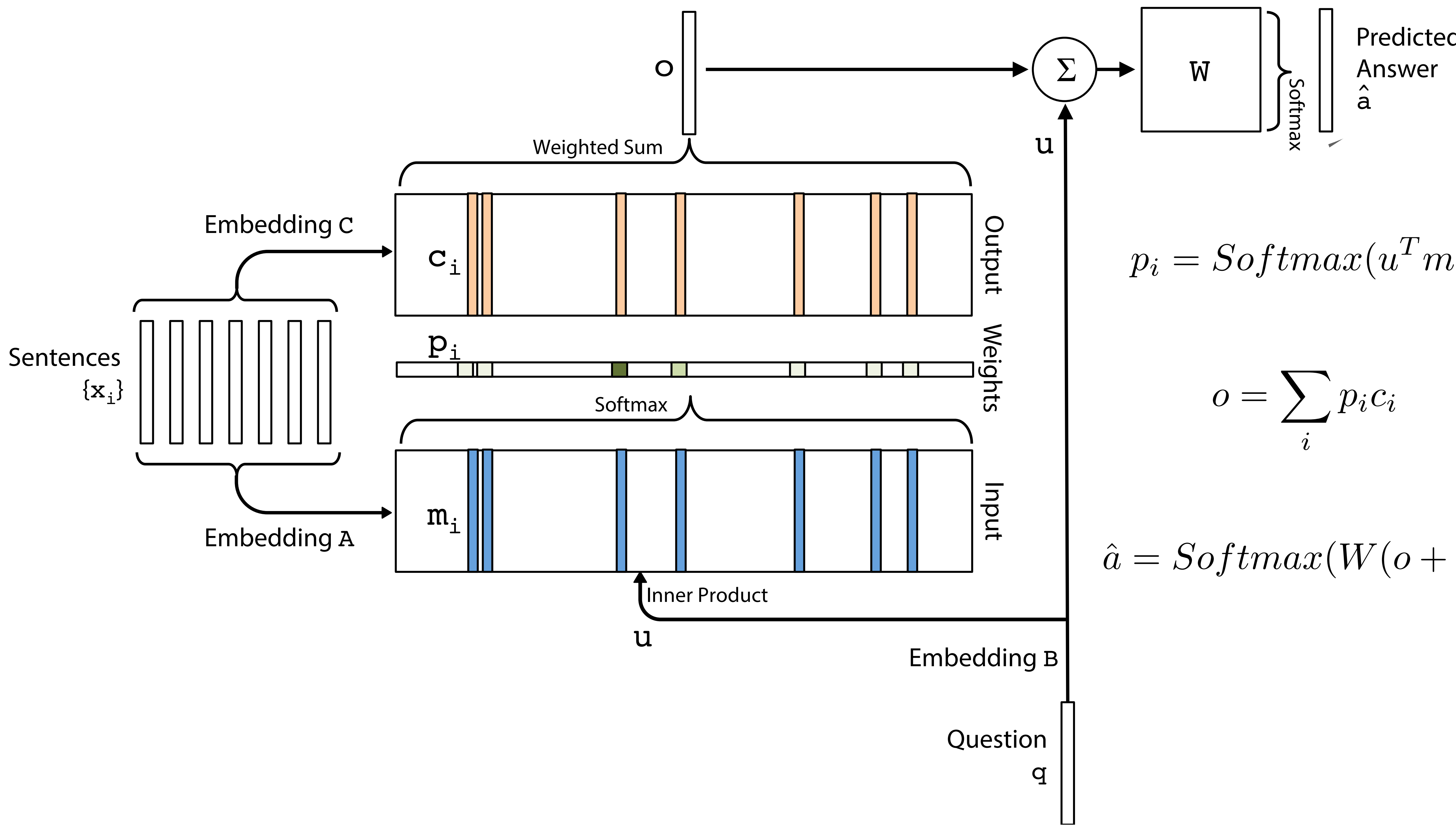


$$p_i = Softmax(u^T m_i)$$

$$o = \sum_i p_i c_i$$

Single Layer

End2End MemNNs



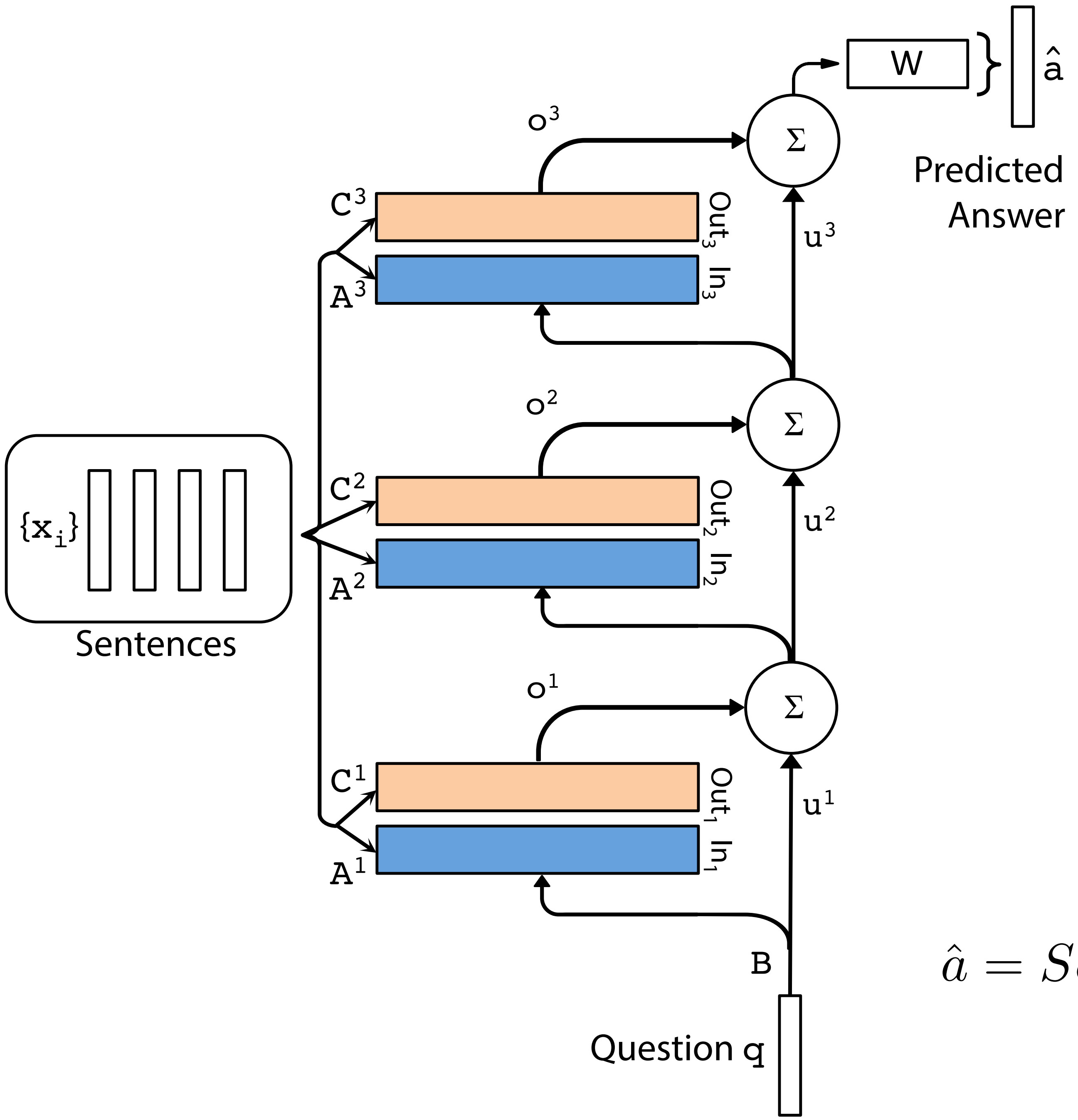
$$p_i = Softmax(u^T m_i)$$

$$o = \sum_i p_i c_i$$

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

Single Layer

End2End MemNNs



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

$$^{k}o = \sum_i {}^{k}p_i {}^{k}c_i$$

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

$$\hat{a} = Softmax(W \cdot {}^{k+1}u) = Softmax(W({}^{k}o + {}^{k}u))$$

Multiple Layer (Hops)

E2EMemNNs: Other Details

Share the input and output embeddings or not

What to store in memories — individual words, word windows, full sentences

How to represent the memories — bag-or-words, RNN style reading at words or characters

Positional Encodings - instead of modeling the sentence as a bag, the word position was modeled by a multiplicative weights on each word vector with the value of the weight being depended on the position.

E2EMemNNs: bAbI

	Weakly supervised			Supervised Supp. Facts	
TASK	**N-grams**	**LSTMs**	**MemN2N**	**Memory Networks**	**StructSVM +coref+srl**
T1. Single supporting fact	36	50	PASS	PASS	PASS
T2. Two supporting facts	2	20	87	PASS	74
T3. Three supporting facts	7	20	60	PASS	17
T4. Two arguments relations	50	61	PASS	PASS	PASS
T5. Three arguments relations	20	70	87	PASS	83
T6. Yes/no questions	49	48	92	PASS	PASS
T7. Counting	52	49	83	85	69
T8. Sets	40	45	90	91	70
T9. Simple negation	62	64	87	PASS	PASS
T10. Indefinite knowledge	45	44	85	PASS	PASS
T11. Basic coreference	29	72	PASS	PASS	PASS
T12. Conjunction	9	74	PASS	PASS	PASS
T13. Compound coreference	26	PASS	PASS	PASS	PASS
T14. Time reasoning	19	27	PASS	PASS	PASS
T15. Basic deduction	20	21	PASS	PASS	PASS
T16. Basic induction	43	23	PASS	PASS	24
T17. Positional reasoning	46	51	49	65	61
T18. Size reasoning	52	52	89	PASS	62
T19. Path finding	0	8	7	36	49
T20. Agent's motivation	76	91	PASS	PASS	PASS

E2EMemNNs: bAbI

Samples from toy 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				

20 bAbI Tasks

	Test Acc	Failed tasks
MemNN	93.3%	4
LSTM	49%	20
MemN2N 1 hop	74.82%	17
2 hops	84.4%	11
3 hops	87.6.%	11

E2EMemNNs: Language Modeling

Predict the next work given previous words in a word sequence.

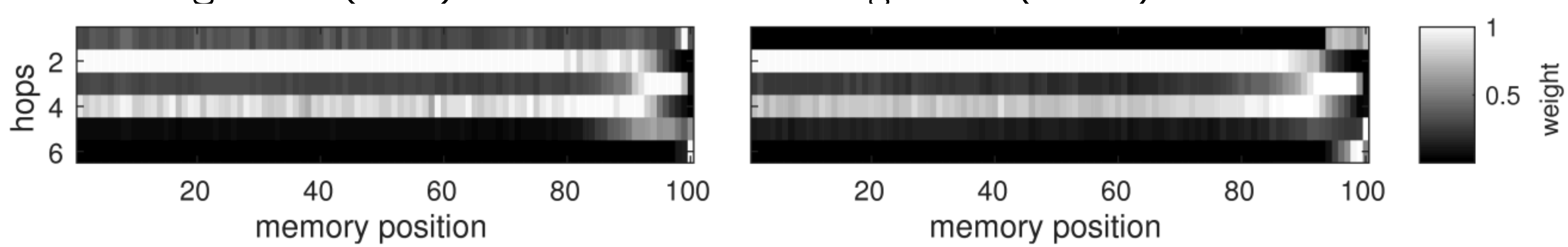
Results on PennTree Bank and Text8 data (a subset of wikipedia)

	Penn Tree	Text8
RNN	129	184
LSTM	115	154
MemN2N 2 hops	121	187
5 hops	118	154
7 hops	111	147

Test perplexity

Hops vs. Attention:
Average over (PTB)

Average over (Text8)



E2EMemNNs: Language Modeling

Same ballpark as LSTMs

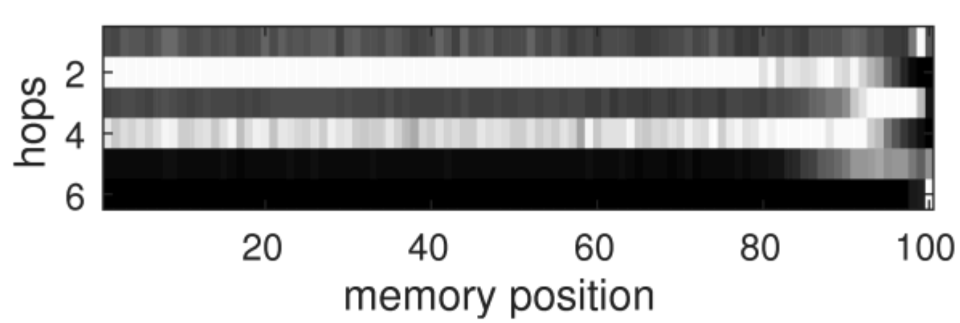
For many words we don't really need long term sequence

Might help for nouns or entities?

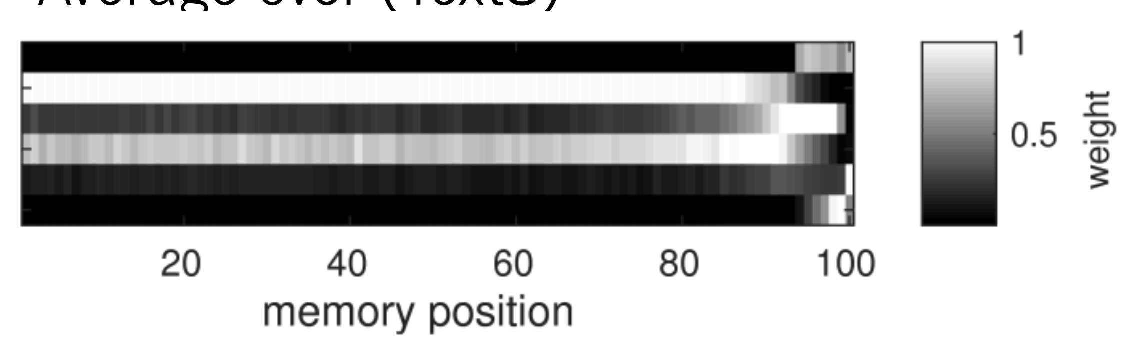
	Penn Tree	Text8
RNN	129	184
LSTM	115	154
MemN2N 2 hops	121	187
5 hops	118	154
7 hops	111	147

Test perplexity

Hops vs. Attention:
Average over (PTB)



Average over (Text8)



Relevant Literature

RNNSearch (Bahdanau et. al.) for Machine Translation

Can be seen as a Memory Network with memory storing individual words and is only a single sentence long.

At inference it reads all the memories and performs Softmax to find best alignment. It is only 1 hop though.

Generating Sequences With RNNs (Graves., 13)

Also does alignment with previous sentence to generate handwriting

Neural Turing Machines (Graves at. al., 14)

Has read/write operations over fixed small sized memory.

Until recently has only been used for toy tasks - copy, sorting etc

Earlier works by Das et. al., 92, Schmidhuber et. al., 93, DISCERN by Miikkulainen, 90) and others fall into this category

Large Scale Memories

So far we’ve only dealt with limited sized memory module

Large Scale Memories

So far we've only dealt with limited sized memory module

Shaolin Soccer directed_by Stephen Chow
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> . . .

Large Scale Memories

Write into the memories more intelligently

During the write operation, hash the memories to store in buckets

The hash functions could be a function of words in the statement: buckets would correspond to topics

Or it could be a function of the embeddings of words

The result is you avoid reading from all the memories - not only it is inefficient, it is also hard to train

Reverb Dataset

Paraphrase Driven Learning for Open Question Answering: Fader et. al., 2013

14 million facts stored as triples [subject, relation, object]

Triples are REVERB extractions mined from ClueWeb09

Statements cover diverse topics:

[milne, authored, winnie-the-pooh]

[sheep, be-afraid-of, wolf]

Training set: weakly labeled QA pairs and 35M paraphrased questions from WikiAnsweres

Who wrote the Winnie the Pooh books?

Who is Pooh's creator?

MemNNs on Reverb Dataset

Paraphrase Driven Learning for Open Question Answering: Fader et. al., 2013

14 million facts stored in memory

Single hop processing. Embedding dimension = 128

Outputs top scoring statement

Also tried adding BoW features

Method	F1
(Fader et al., 2013)	0.54
(Bordes et al., 2014)	0.73
MemNN	0.72
MemNN (with BoW features)	0.82

MemNNs on Reverb Dataset

QA reference - complete the reference

Scoring all 14 million facts in memory hard and slow

So we hash based on:

Words in the statement: inverted index

K-means in embedding space

Method	Embedding	Embed+BoW	candidates
MemNN (no hashing)	0.72	0.82	14M
MemNN (word hash)	0.63	0.68	13k (1000x)
MemNN (clust hash)	0.71	0.80	177k (80x)

Multitasked MemNNs:bAbI + Reverb

Story told to the model after training

Antoine went to the kitchen.

Antoine picked up the milk.

Antoine travelled to the office

Where is the milk? : office

Where was Antoine before the office?: kitchen

Where does milk come from?: milk come from cow

What is cow a type of?: cow be female of cattle

Where are cattle found?: cattle farm become widespread in Brazil

What does milk taste like?: milk taste like milk

What does milk go well with?: milk go with coffee

Cloze Style QA

Teaching a machine to understand language is hard

One way is to read a comprehension and answer questions pertaining to it

However the questions should be such that they cannot be answered using external world knowledge - Cloze Style QA

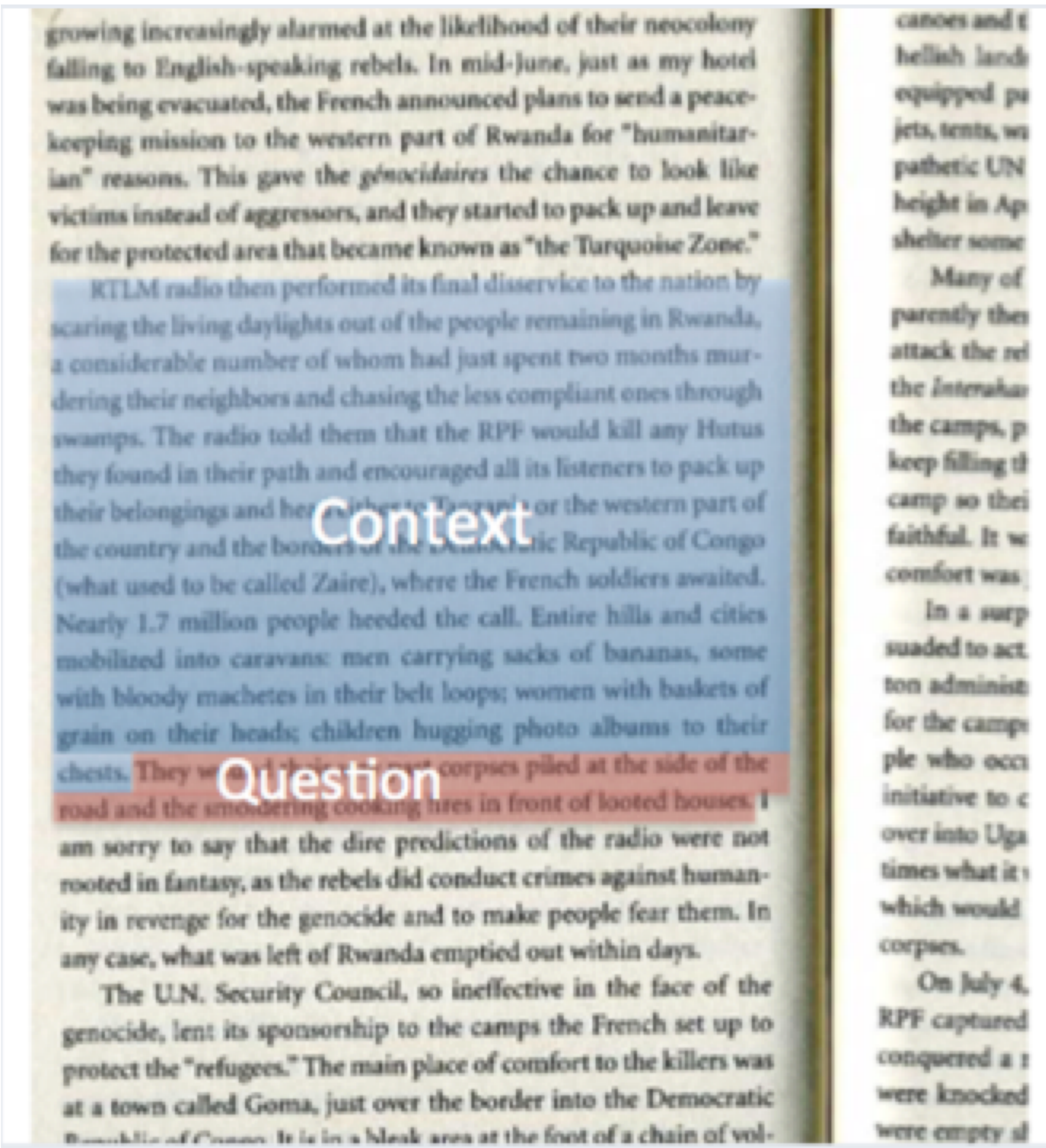
Until recently only small sized dataset existed - which were primarily used for testing - nothing to train on

Two primary efforts in this direction

Teaching Machines to Read and Comprehend: Hermann et. al.2015

The Goldilocks Principle: Reading Children's Books with Explicit Memory Representation: Hill et. al., 2015

CBT: Children's Book Dataset



growing increasingly alarmed at the likelihood of their neocolony falling to English-speaking rebels. In mid-June, just as my hotel was being evacuated, the French announced plans to send a peace-keeping mission to the western part of Rwanda for "humanitarian" reasons. This gave the *génocidaires* the chance to look like victims instead of aggressors, and they started to pack up and leave for the protected area that became known as "the Turquoise Zone."

RTLM radio then performed its final disservice to the nation by scaring the living daylights out of the people remaining in Rwanda, a considerable number of whom had just spent two months murdering their neighbors and chasing the less compliant ones through swamps. The radio told them that the RPF would kill any Hutus they found in their path and encouraged all its listeners to pack up their belongings and he[illegible] to Tanzania or the western part of the country and the bord[illegible] Republic of Congo (what used to be called Zaire), where the French soldiers awaited. Nearly 1.7 million people heeded the call. Entire hills and cities mobilized into caravans: men carrying sacks of bananas, some with bloody machetes in their belt loops; women with baskets of grain on their heads; children hugging photo albums to their chests. They w[illegible] corpses piled at the side of the road and the smoldering cooking fires in front of looted houses. I am sorry to say that the dire predictions of the radio were not rooted in fantasy, as the rebels did conduct crimes against humanity in revenge for the genocide and to make people fear them. In any case, what was left of Rwanda emptied out within days.

The U.N. Security Council, so ineffective in the face of the genocide, lent its sponsorship to the camps the French set up to protect the "refugees." The main place of comfort to the killers was at a town called Goma, just over the border into the Democratic Republic of Congo. It is in a bleak area at the foot of a chain of vol-

Dataset built from 118 freely available books from project Gutenberg

Children stories provide a clear narrative structure

Can make the role of context more salient

The Goldilocks Principle: Reading Children's Books with Explicit Memory Representation: Hill et. al., 2015

S: 1 Mr. Cropper was opposed to our hiring you .
2 Not , of course , that he had any personal objection to you , but he is set against female teachers , and when a Cropper is set there is nothing on earth can change him .
3 He says female teachers ca n't keep order .
4 He 's started in with a spite at you on general principles , and the boys know it .
5 They know he 'll back them up in secret , no matter what they do , just to prove his opinions .
6 Cropper is sly and slippery , and it is hard to corner him . ''
7 `` Are the boys big ? ''
8 queried Esther anxiously .
9 `` Yes .
10 Thirteen and fourteen and big for their age .
11 You ca n't whip 'em -- that is the trouble .
12 A man might , but they 'd twist you around their fingers .
13 You 'll have your hands full , I 'm afraid .
14 But maybe they 'll behave all right after all . ''
15 Mr. Baxter privately had no hope that they would , but Esther hoped for the best.
16 She could not believe that Mr. Cropper would carry his prejudices into a personal application .
17 This conviction was strengthened when he overtook her walking from school the next day and drove her home .
18 He was a big , handsome man with a very suave , polite manner .
19 He asked interestedly about her school and her work , hoped she was getting on well , and said he had two young rascals of his own to send soon .
20 Esther felt relieved .

q: She thought that Mr. ________ had exaggerated matters a little .

C: Baxter, Cropper, Esther, course, fingers, manner, objection, opinion, right, spite.

a: Baxter

First 20 sentences form a context - 21st sentence becomes the query.

A single word from the 21st sentence is removed, which becomes the answer.

The model must identify the answer word from a selection of 10 provided candidates

MemNNs for Story Understanding

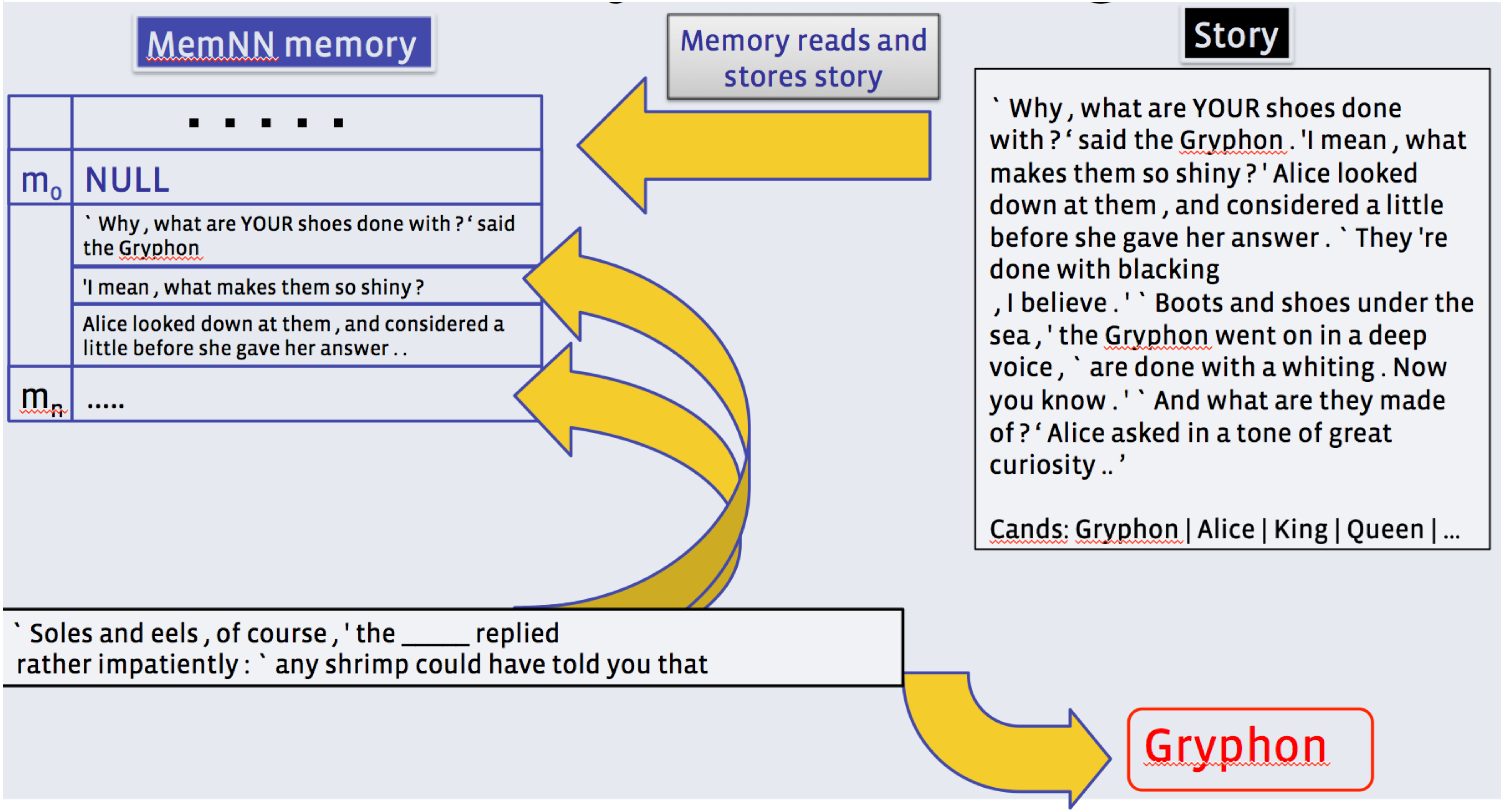


Figure: Jason Weston

MemNNs for Story Understanding

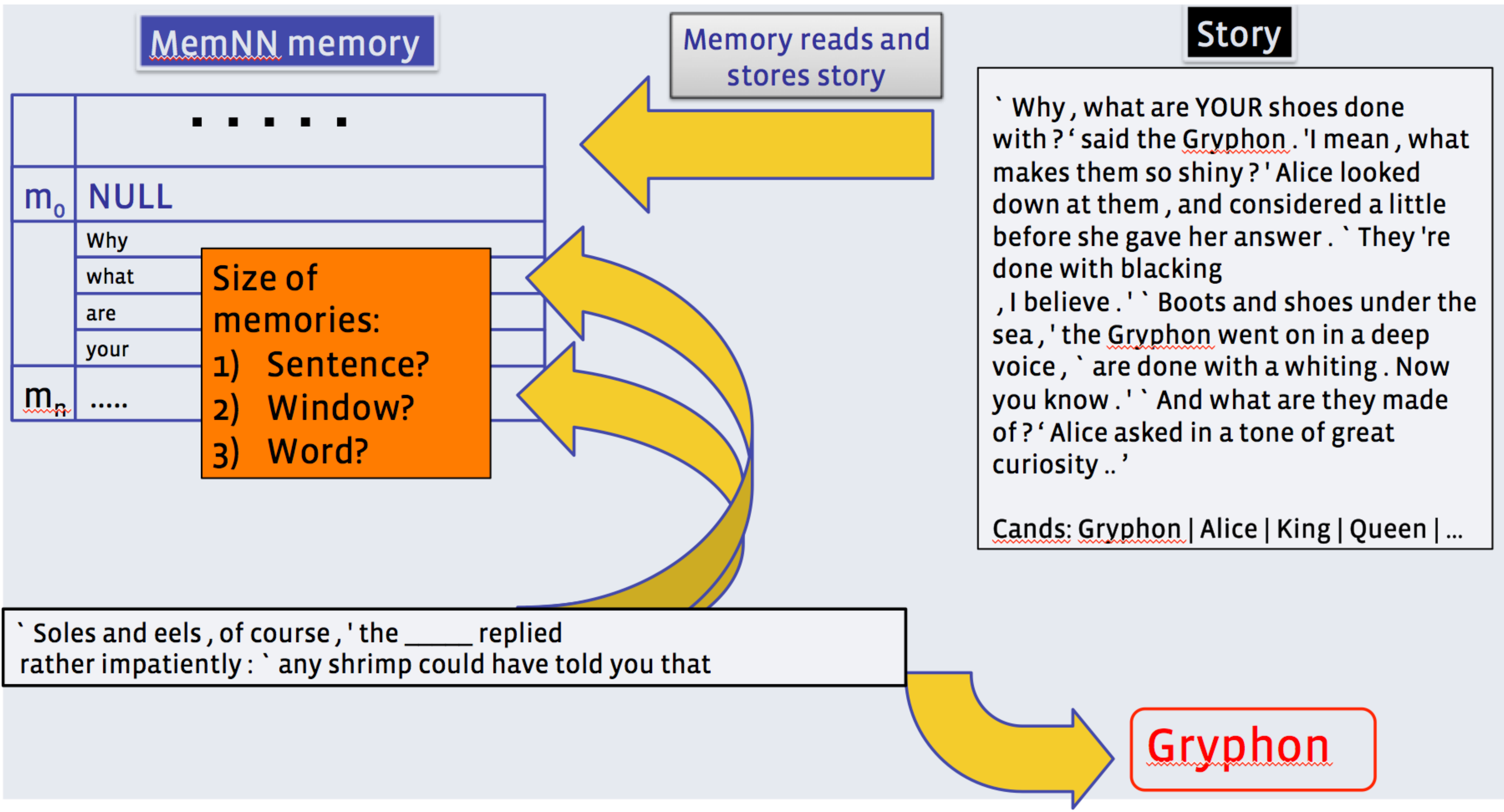


Figure: Jason Weston

MemNNs for Story Understanding

The Goldilocks Principle: Reading Children's Books with Explicit Memory Representation: Hill et. al., 2015

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

MemNNs for Story Understanding

The Goldilocks Principle: Reading Children's Books with Explicit Memory Representation: Hill et. al., 2015

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

Self Supervision in MemNNs

During training we have knowledge about the correct answer word

We can treat all the memories in which the answer word appears as the relevant supporting fact

Bump up the scores of these memories

This speeds up training

Of course this knowledge is not available at test time - so you simply pick the most relevant memory to generate your answer

QA on News Articles

Teaching Machines to Read and Comprehend: Hermann et. al.2015

	CNN			Daily Mail		
	train	valid	test	train	valid	test
# months	95	1	1	56	1	1
# documents	90,266	1,220	1,093	196,961	12,148	10,397
# queries	380,298	3,924	3,198	879,450	64,835	53,182
Max # entities	527	187	396	371	232	245
Avg # entities	26.4	26.5	24.5	26.5	25.5	26.0
Avg # tokens	762	763	716	813	774	780
Vocab size	118,497			208,045		

Table 1: Corpus statistics. Articles were collected starting in April 2007 for CNN and June 2010 for the Daily Mail, both until the end of April 2015. Validation data is from March, test data from April 2015. Articles of over 2000 tokens and queries whose answer entity did not appear in the context were filtered out.

We evaluate our models on this dataset as well

QA on News Articles

Teaching Machines to Read and Comprehend: Hermann et. al.2015

	Original Version	Anonymised Version
Context		
	The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the "Top Gear" host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broad-caster found he had subjected producer Oisin Tymon "to an unprovoked physical and verbal attack." …	the *ent381* producer allegedly struck by *ent212* will not press charges against the " *ent153* " host , his lawyer said friday . *ent212* , who hosted one of the most - watched television shows in the world , was dropped by the *ent381* wednesday after an internal investigation by the *ent180* broadcaster found he had subjected producer *ent193* " to an unprovoked physical and verbal attack . " …
Query		
	Producer **X** will not press charges against Jeremy Clarkson, his lawyer says.	producer **X** will not press charges against *ent212* , his lawyer says .
Answer		
	Oisin Tymon	*ent193*

QA on News Articles

METHODS	VALIDATION	TEST
MAXIMUM FREQUENCY (ARTICLE)(*)	0.305	0.332
SLIDING WINDOW	0.005	0.006
WORD DISTANCE MODEL(*)	0.505	0.509
DEEP LSTMS (ARTICLE+QUERY)(*)	0.550	0.570
CONTEXTUAL LSTMS ("ATTENTIVE READER")(*)	0.616	0.630
CONTEXTUAL LSTMS ("IMPATIENT READER")(*)	0.618	0.638
MEMNNS (WINDOW MEMORY)	0.580	0.606
MEMNNS (WINDOW MEMORY + SELF-SUP.)	0.634	0.668
MEMNNS (WINDOW MEMORY + ENSEMBLE)	0.612	0.638
MEMNNS (WINDOW MEMORY + SELF-SUP. + ENSEMBLE)	0.649	0.684
MEMNNS (WINDOW + SELF-SUP. + ENSEMBLE + EXCLUD. COOCURRENCES)	**0.662**	**0.694**

Dialog Modeling

So far we have focused on a single step QA potentially with long term context

How about Dialog Modeling?

We have built another large scale dataset focussed towards movie domain

Ask about movies — Ask about movie recommendation — Have dialog which combines facts and opinions — General chit-chat about movies

75k entities, and 3.5M exchanges

Evaluating Prerequisite Qualities for Learning End-toEnd Dialog Systems: Dodge et. al., 2016

Dialog Modeling

Task 1: QA on Movies

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

Evaluating Prerequisite Qualities for Learning End-toEnd Dialog Systems: Dodge et. al., 2016

Dialog Modeling

Task 2: Movie Recommendation

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

Evaluating Prerequisite Qualities for Learning End-toEnd Dialog Systems: Dodge et. al., 2016

Dialog Modeling

Task 3: Combining QA and Movie Recommendation

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

Evaluating Prerequisite Qualities for Learning End-toEnd Dialog Systems: Dodge et. al., 2016

Dialog Modeling

Task 4: Dialog from Reddit Dataset (Real Dialog)

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...

Evaluating Prerequisite Qualities for Learning End-toEnd Dialog Systems: Dodge et. al., 2016

Memory Networks for Dialog

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	c_1^u	1) I'm looking a fun comedy to watch tonight, any ideas?
Memories	c_1^r	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 ...

Results

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

Key-Value MemNNs

Key Value Memory Networks for Directly Reading Documents: Miller et. al., 2016

Facts are stored in a key value structured memory

Memory is designed so that the model learns to use keys to address relevant memories with respect to the question

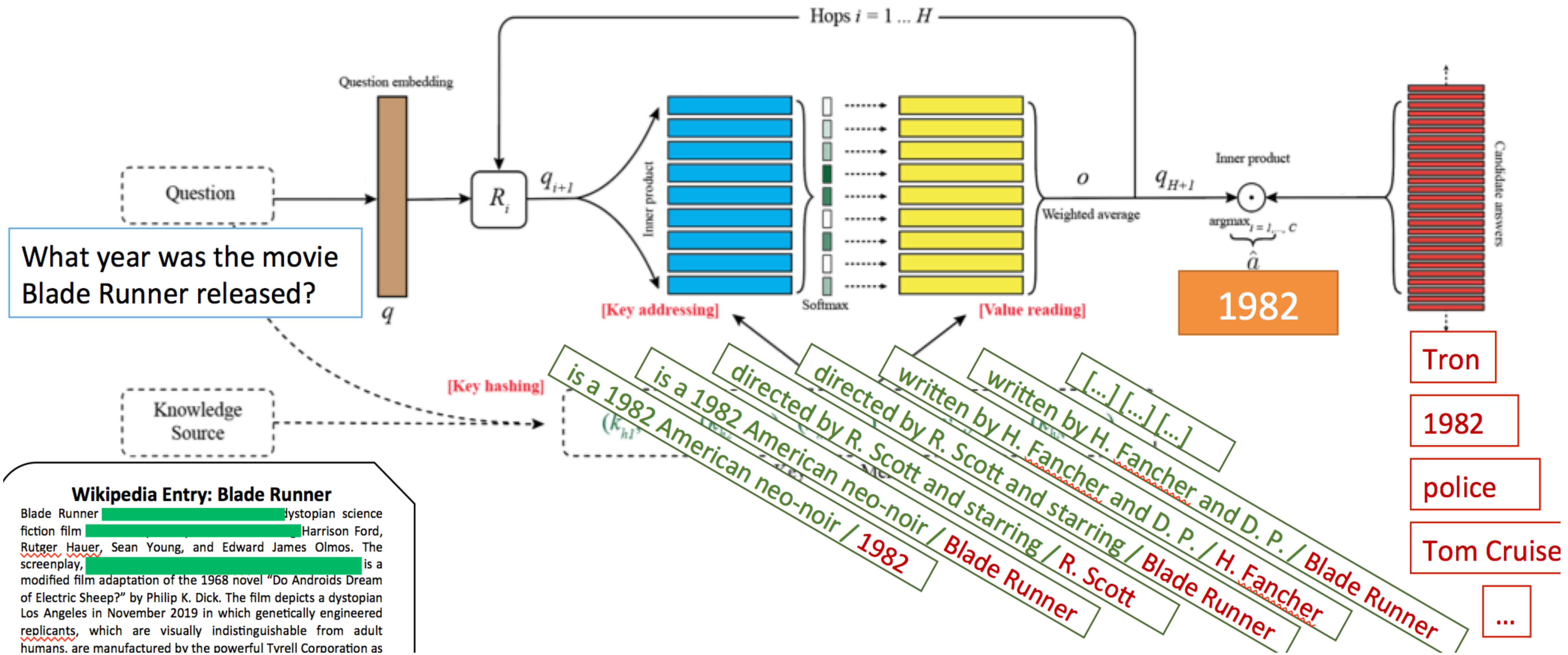
Structure allows the model to encode prior knowledge for the considered task

Structure also allows to leverage possibly complex transforms between key and value

Example: for a KB triple [subject, relation, object], Key could be [subject,relation] and value could be [object] or vice versa

Key-Value MemNNs

Key Value Memory Networks for Directly Reading Documents: Miller et. al., 2016



Key-Value MemNNs

Test results on WikiQA

Method	MAP	MRR
Word Cnt	0.4891	0.4924
Wgt Word Cnt	0.5099	0.5132
2-gram CNN (Yang *et al.*, 2015)	0.6520	0.6652
AP-CNN (Santos *et al.*, 2016)	0.6886	0.6957
Attentive LSTM (Miao *et al.*, 2015)	0.6886	0.7069
Attentive CNN (Yin and Schütze, 2015)	0.6921	0.7108
L.D.C. (Wang *et al.*, 2016)	0.7058	0.7226
Memory Network	0.5170	0.5236
Key-Value Memory Network	**0.7069**	**0.7265**

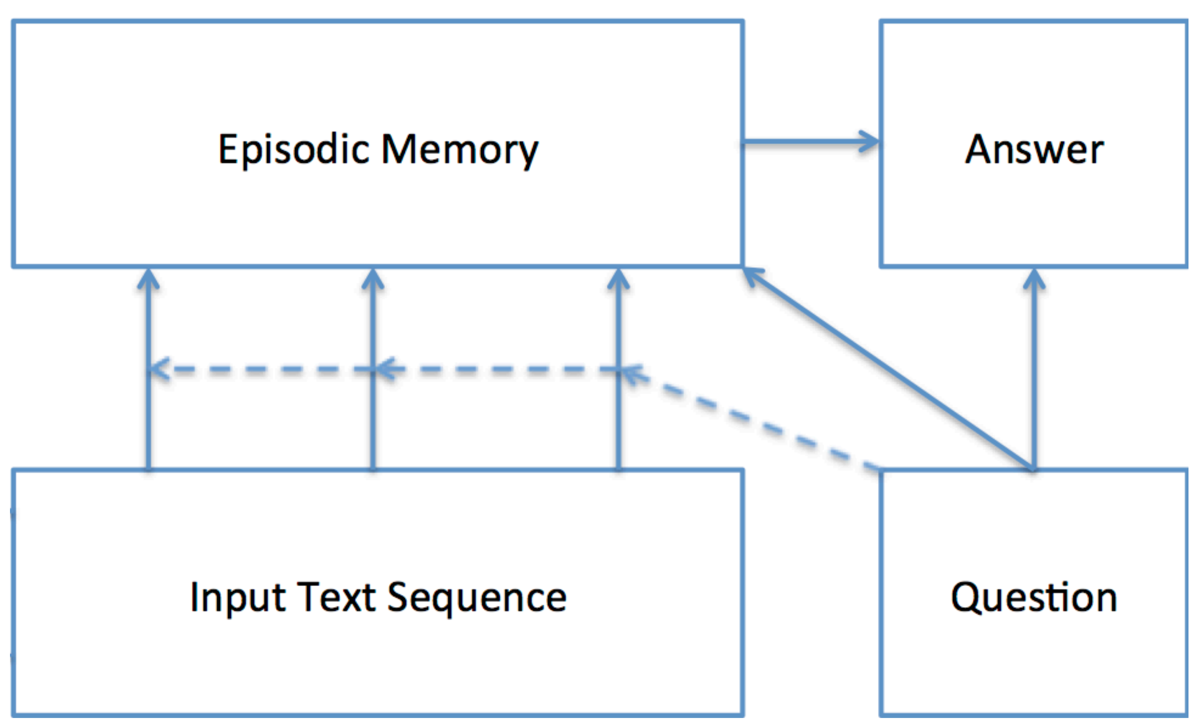
Dynamic MemNNs

Ask Me Anything: Dynamic Memory Networks for Natural Language Processing: Kumar et. al., 2016

MemNN framework allows freedom of how to represent memories, how to represent questions, and how to get the answers given the question and the input

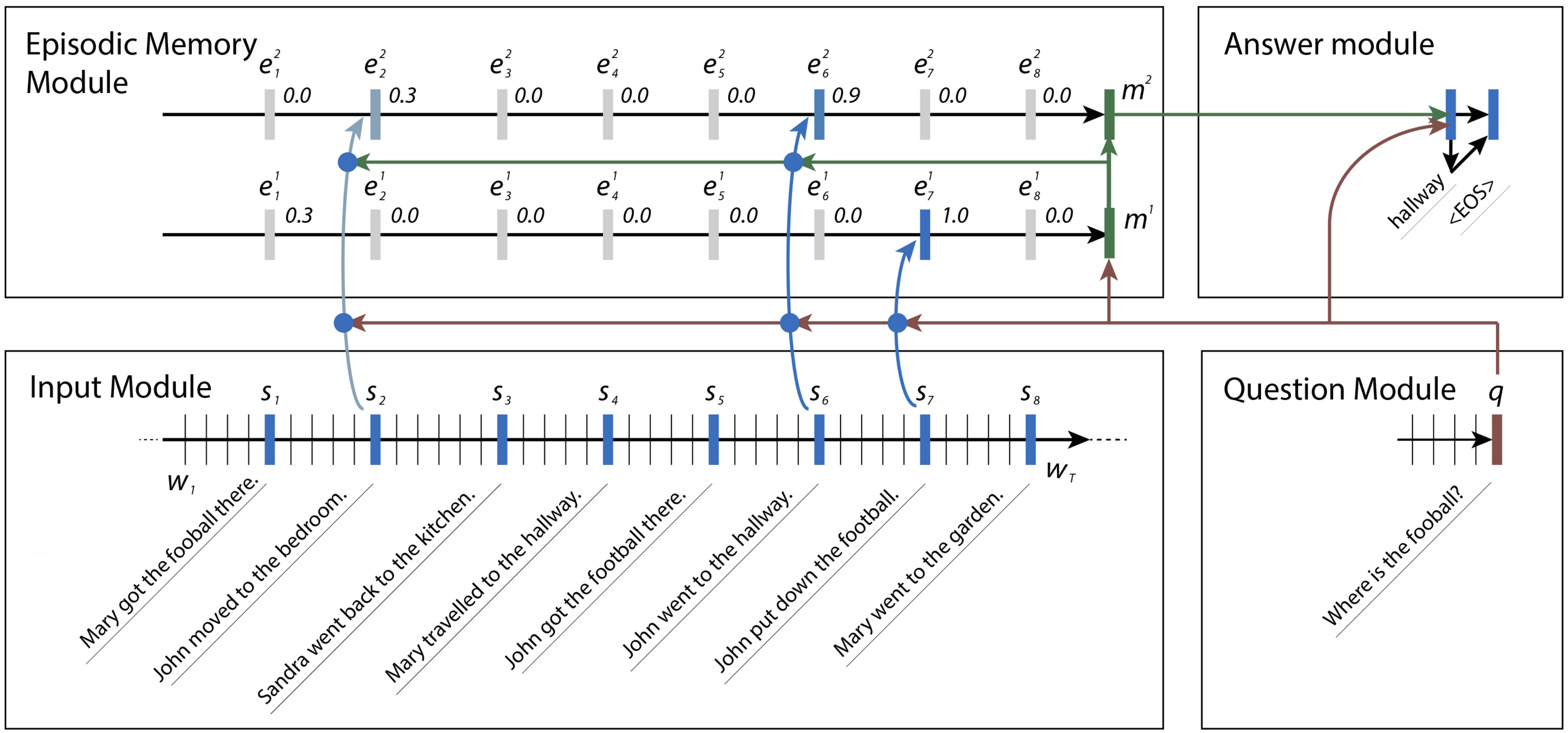
Dynamic MemNNs is a recently proposed extension along these lines

Has four modules — Input Module — Question Module — Episodic Memory Module — Answer Module



Dynamic MemNNs

Ask Me Anything: Dynamic Memory Networks for Natural Language Processing: Kumar et. al., 2016



Input Module

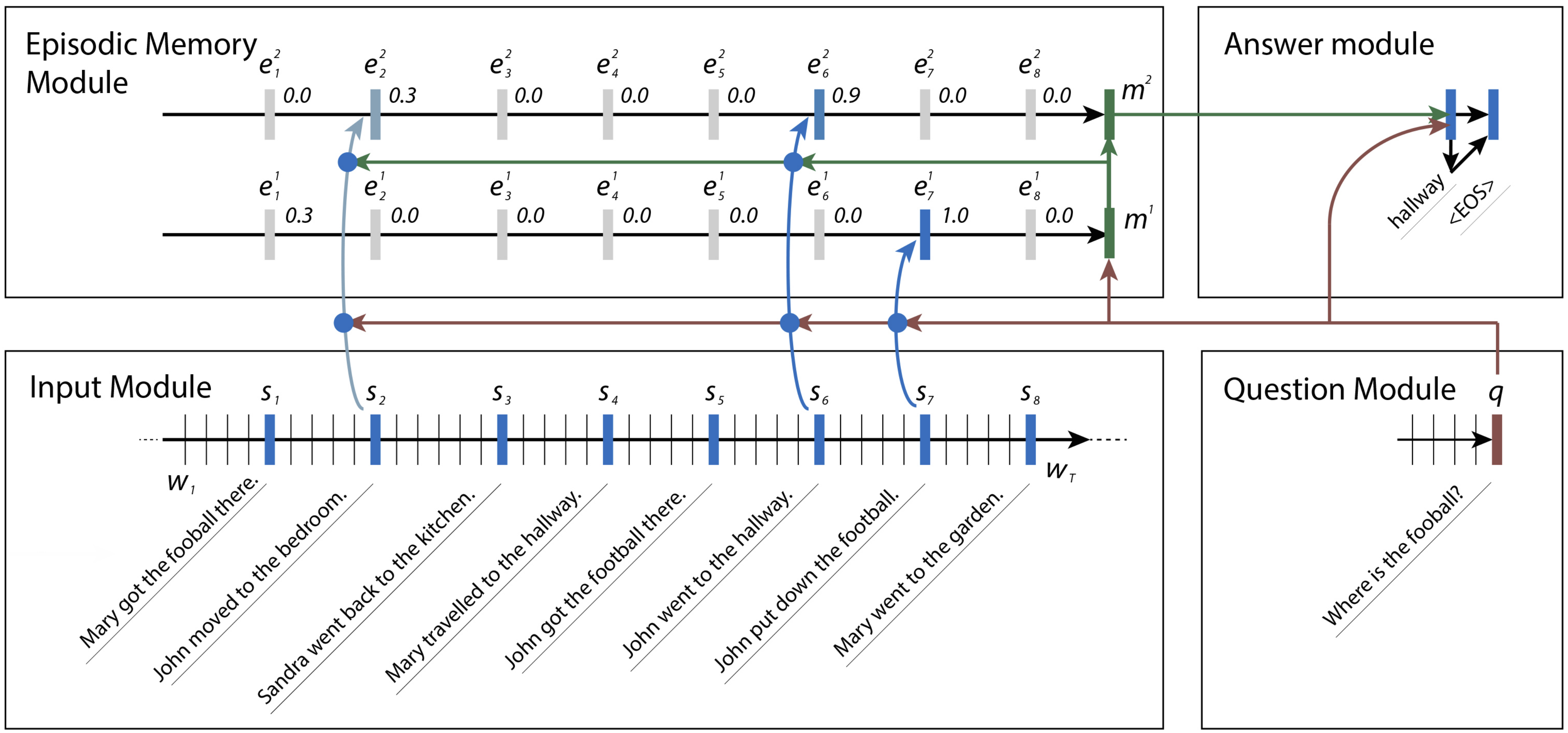
Generates and stores the representations of input statements (stories) — output of an RNN as the input representation — GRU

Question Module

Similar to the Input Module — output of an RNN as the question representation — GRU

Dynamic MemNNs

Ask Me Anything: Dynamic Memory Networks for Natural Language Processing: Kumar et. al., 2016



Episodic Memory Module

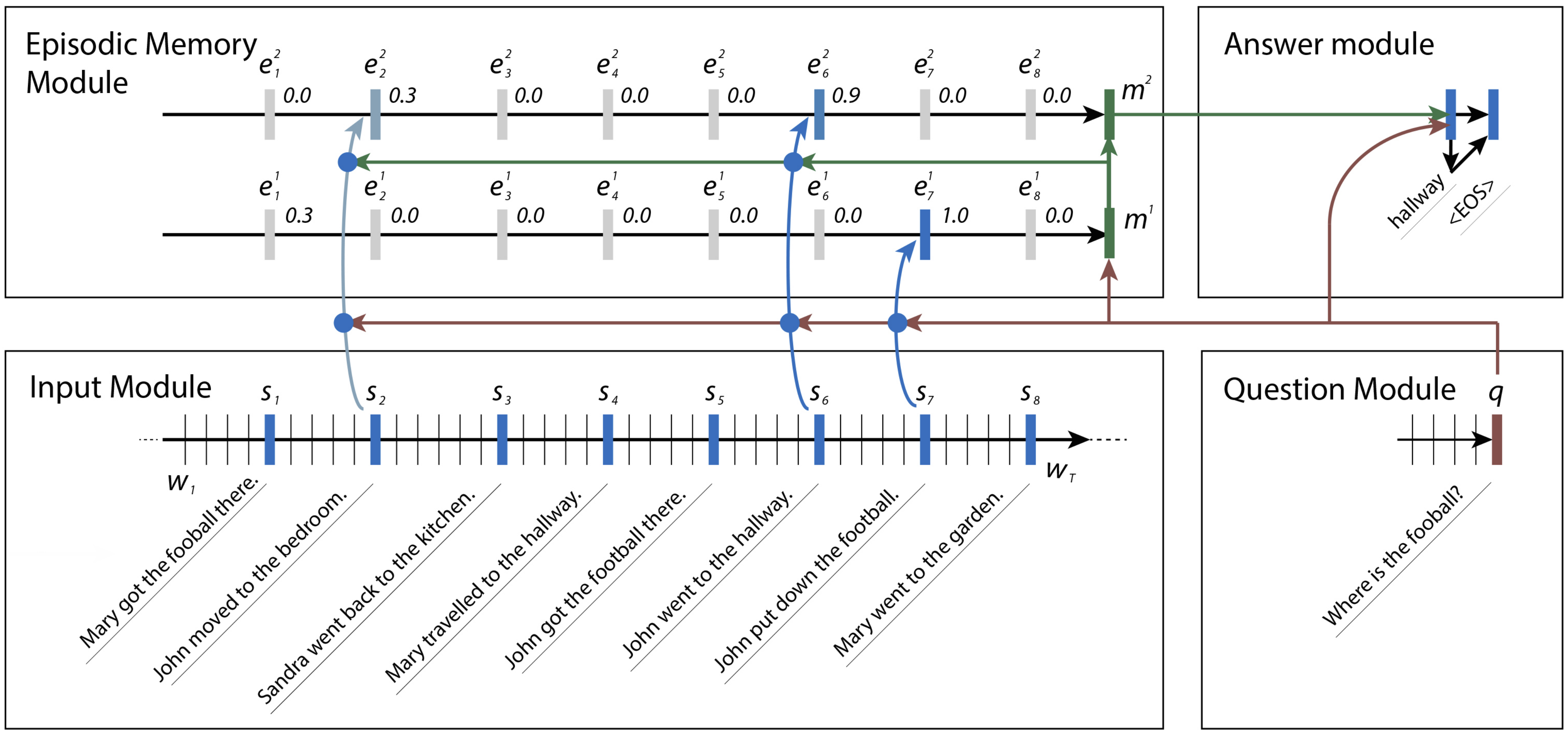
Comprises of an attention mechanism and a GRU which updates its internal memory state

given the question rep. and previous memory, this module attends over inputs to produce an episode

using new episode and previous memory the GRU generates a new memory — iterate!

Dynamic MemNNs

Ask Me Anything: Dynamic Memory Networks for Natural Language Processing: Kumar et. al., 2016



Answer Module

Given a vector the answer modules maps it to the final answer

Depending on the task the answer module is either triggered once at the end of the episode or at every time step

A typical module would have an RNN whose initial hidden state is the final memory, the inputs are the question word sequence and outputs are the answer words

Dynamic MemNNs Experiments

Ask Me Anything: Dynamic Memory Networks for Natural Language Processing: Kumar et. al., 2016

Task	MemNN	DMN
1: Single Supporting Fact	100	100
2: Two Supporting Facts	100	98.2
3: Three Supporting Facts	100	95.2
4: Two Argument Relations	100	100
5: Three Argument Relations	98	99.3
6: Yes/No Questions	100	100
7: Counting	85	96.9
8: Lists/Sets	91	96.5
9: Simple Negation	100	100
10: Indefinite Knowledge	98	97.5
11: Basic Coreference	100	99.9
12: Conjunction	100	100
13: Compound Coreference	100	99.8
14: Time Reasoning	99	100
15: Basic Deduction	100	100
16: Basic Induction	100	99.4
17: Positional Reasoning	65	59.6
18: Size Reasoning	95	95.3
19: Path Finding	36	34.5
20: Agent's Motivations	100	100
Mean Accuracy (%)	93.3	**93.6**

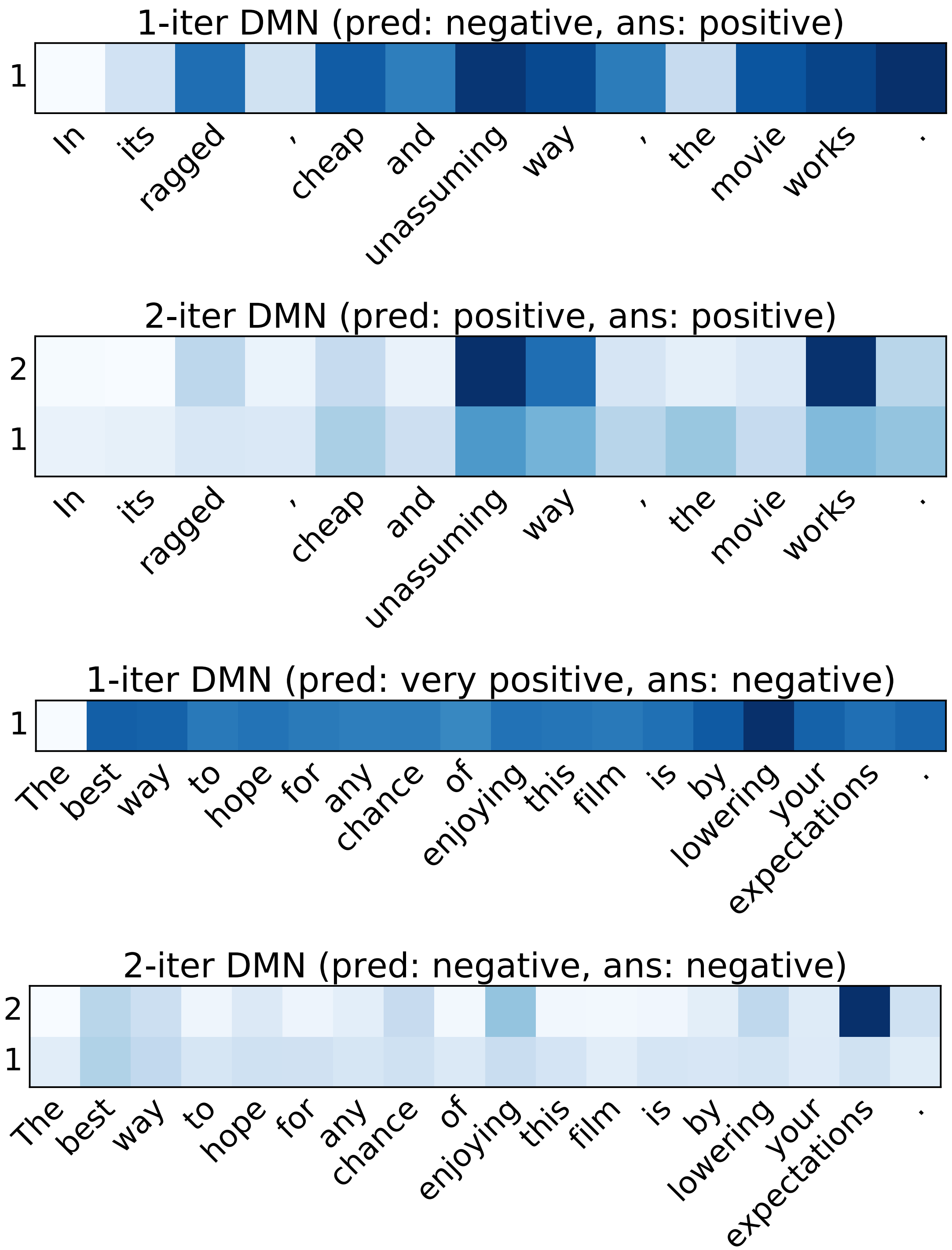
bAbI Dataset

Question: Where was Mary before the Bedroom?
Answer: Cinema.

Facts	Episode 1	Episode 2	Episode 3
Yesterday Julie traveled to the school.			
Yesterday Marie went to the cinema.			
This morning Julie traveled to the kitchen.			
Bill went back to the cinema yesterday.			
Mary went to the bedroom this morning.			
Julie went back to the bedroom this afternoon.			
[done reading]			

Dynamic MemNNs Experiments

Ask Me Anything: Dynamic Memory Networks for Natural Language Processing: Kumar et. al., 2016



Stanford Sentiment Treebank

Task	Binary	Fine-grained
MV-RNN	82.9	44.4
RNTN	85.4	45.7
DCNN	86.8	48.5
PVec	87.8	48.7
CNN-MC	88.1	47.4
DRNN	86.6	49.8
CT-LSTM	88.0	51.0
DMN	**88.6**	**52.1**

Table 2. Test accuracies for sentiment analysis on the Stanford Sentiment Treebank. MV-RNN and RNTN: Socher et al. (2013). DCNN: Kalchbrenner et al. (2014). PVec: Le & Mikolov. (2014). CNN-MC: Kim (2014). DRNN: Irsoy & Cardie (2015), 2014. CT-LSTM: Tai et al. (2015)

Dynamic MemNNs Experiments

Ask Me Anything: Dynamic Memory Networks for Natural Language Processing: Kumar et. al., 2016

WSJ-PTB Part of Speech Tagging Task

Model	Acc (%)
SVMTool	97.15
Sogaard	97.27
Suzuki et al.	97.40
Spoustova et al.	97.44
SCNN	97.50
DMN	**97.56**

Table 3. Test accuracies on WSJ-PTB

MemNNs Summary

Models which augments a standard deep network with an external readable and writable memory

These memories are learnt and used effectively in solving reasoning tasks which require long term knowledge

The architecture is quite flexible in how one represents the memories and how they are used to solve the final task

MemNNs Shortcomings

While the model is quite rich one significant drawback is that it cannot write to memory intelligently.

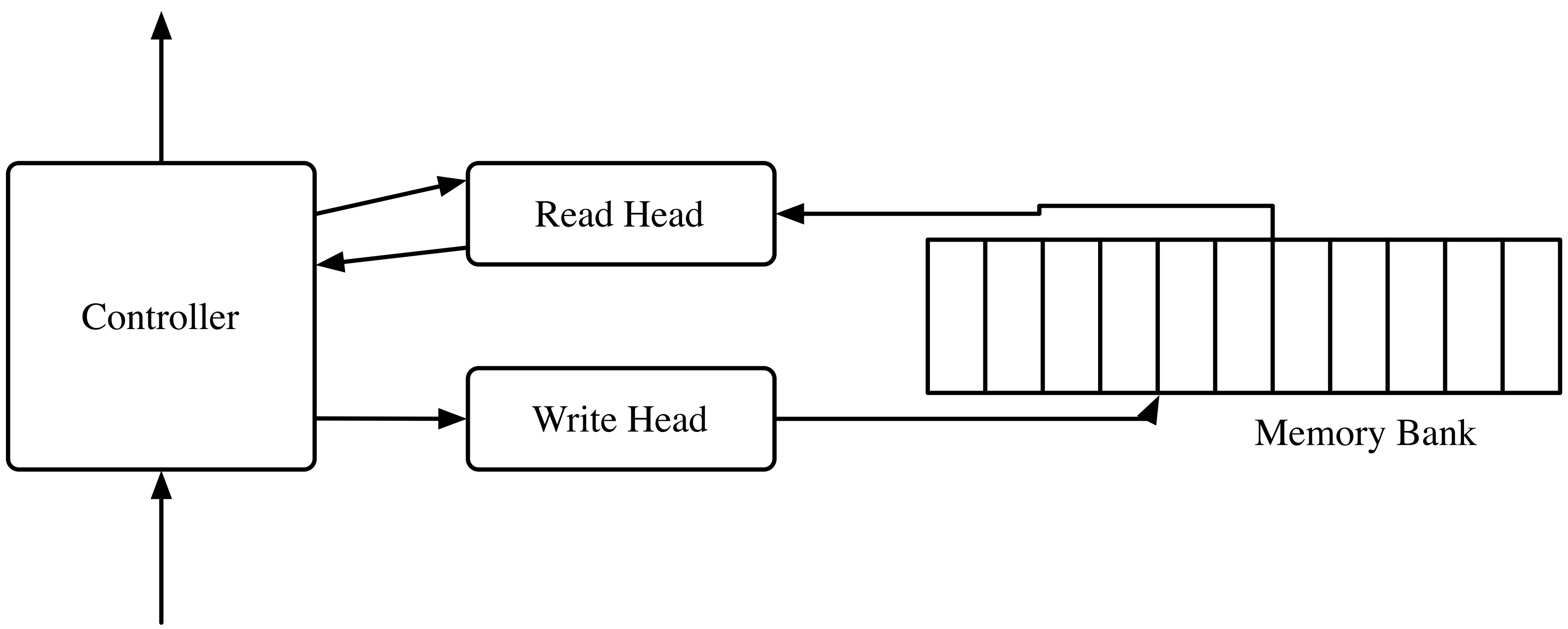
Given a new statement it simply writes it at the next available slot. If the memory is full it will cycle.

One cannot erase memories

One cannot compress memories

Neural Turing Machines

Neural Turing Machines: Graves, Wayne, Danihelka 2015



Follows the standard architecture of MemNNs

The primary difference is in the way it writes to the memory

NTM: Read Mechanism

Neural Turing Machines: Graves, Wayne, Danihelka 2015

w_t: weight vector over N memory locations emitted by the read head at time t

$$\sum_{i=1}^{N} w_t(i) = 1, \quad 0 \leq w_t(i) \leq 1, \; \forall i$$

$$r_t \leftarrow \sum_{i=1}^{N} w_t(i) M_t(i), \quad r_t \in \mathcal{R}^d$$

the read vector

contents of the i-th slot of memory at time t

NTM: Write Mechanism

Neural Turing Machines: Graves, Wayne, Danihelka 2015

w_t: weight vector over N memory locations emitted by the write head at time t

e_t : erase vector

a_t : add vector

$$\tilde{M}_t(i) \leftarrow M_{t-1}(i)[1 - w_t(i)e_t]$$

$$M_t(i) \leftarrow \tilde{M}_t(i) + w_t(i)a_t$$

NTM: Addressing Mechanism

Neural Turing Machines: Graves, Wayne, Danihelka 2015

How are the weight vectors computed?

A combination of content based addressing and location based addressing

Content based is the usual stuff: attention based on content

Location based is different. Allows for single step jumps or random location jumps

NTM: Addressing Mechanism

Neural Turing Machines: Graves, Wayne, Danihelka 2015

Content Based

$$w_t^c(i) \longleftarrow \frac{\exp\Big(\beta_t K\big[\mathbf{k}_t, \mathbf{M}_t(i)\big]\Big)}{\sum_j \exp\Big(\beta_t K\big[\mathbf{k}_t, \mathbf{M}_t(j)\big]\Big)}.$$

Scoring function $$K\big[\mathbf{u}, \mathbf{v}\big] = \frac{\mathbf{u} \cdot \mathbf{v}}{||\mathbf{u}|| \cdot ||\mathbf{v}||}.$$

NTM: Addressing Mechanism

Neural Turing Machines: Graves, Wayne, Danihelka 2015

Location Based

Step 1: compute an interpolation vector

$$\mathbf{w}_t^g \longleftarrow g_t \mathbf{w}_t^c + (1 - g_t)\mathbf{w}_{t-1}.$$

NTM: Addressing Mechanism

Neural Turing Machines: Graves, Wayne, Danihelka 2015

Location Based

Step 1: compute an interpolation vector

$$\mathbf{w}_t^g \longleftarrow g_t \mathbf{w}_t^c + (1 - g_t)\mathbf{w}_{t-1}.$$

Step 2: convolve using the shift vector

$$\tilde{w}_t(i) \longleftarrow \sum_{j=0}^{N-1} w_t^g(j)\, s_t(i-j)$$

NTM: Addressing Mechanism

Neural Turing Machines: Graves, Wayne, Danihelka 2015

Location Based

Step 1: compute an interpolation vector

$$\mathbf{w}_t^g \longleftarrow g_t \mathbf{w}_t^c + (1 - g_t)\mathbf{w}_{t-1}.$$

Step 2: convolve using the shift vector

$$\tilde{w}_t(i) \longleftarrow \sum_{j=0}^{N-1} w_t^g(j)\, s_t(i-j)$$

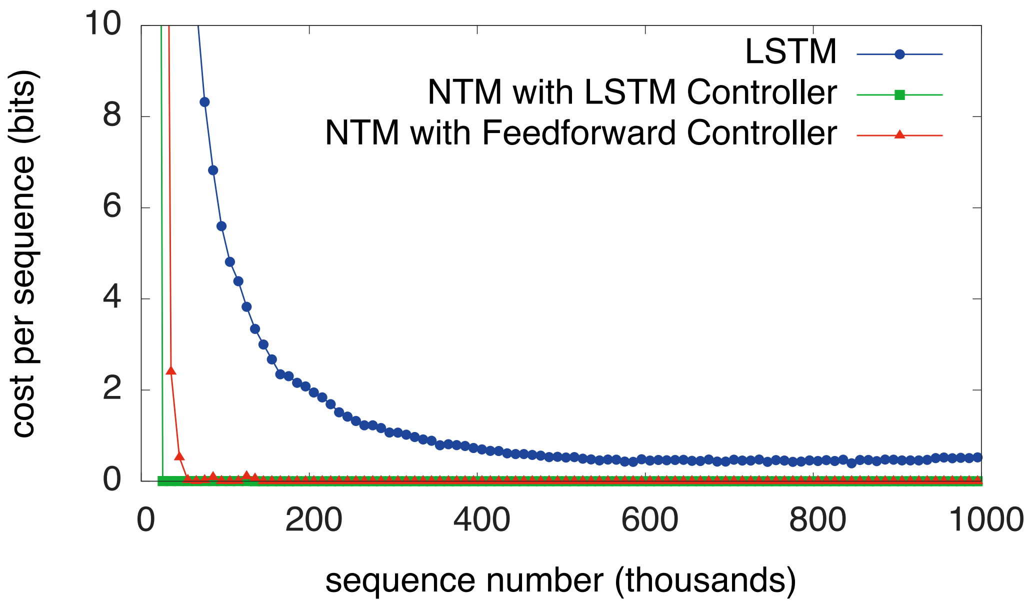
Step 3: sharpen the weight vector

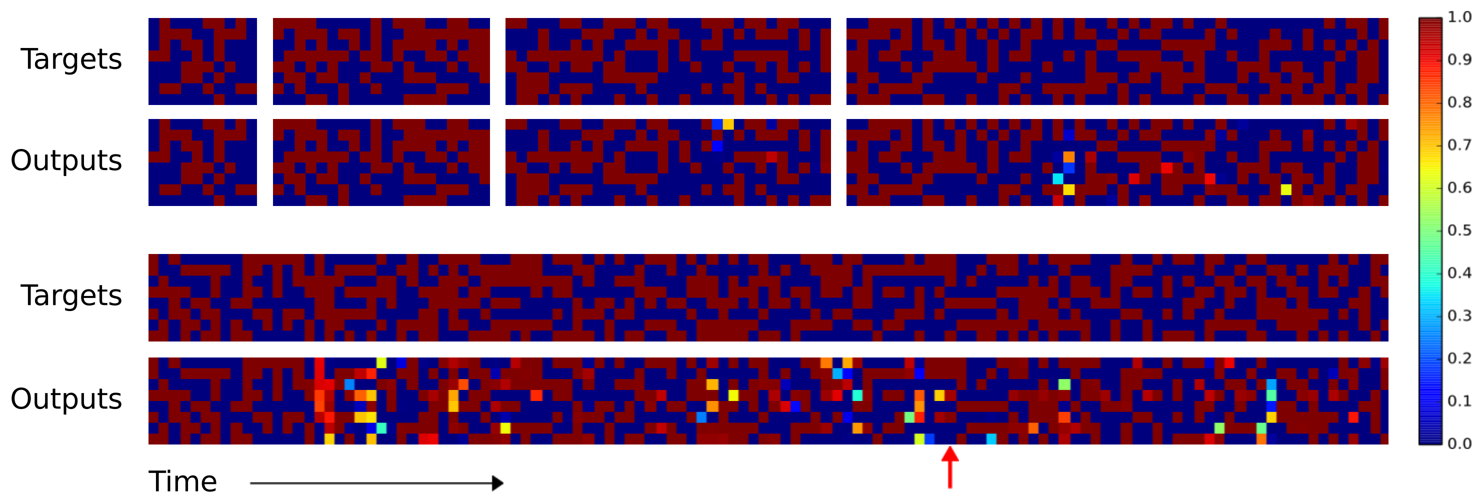
$$w_t(i) \longleftarrow \frac{\tilde{w}_t(i)^{\gamma_t}}{\sum_j \tilde{w}_t(j)^{\gamma_t}}$$

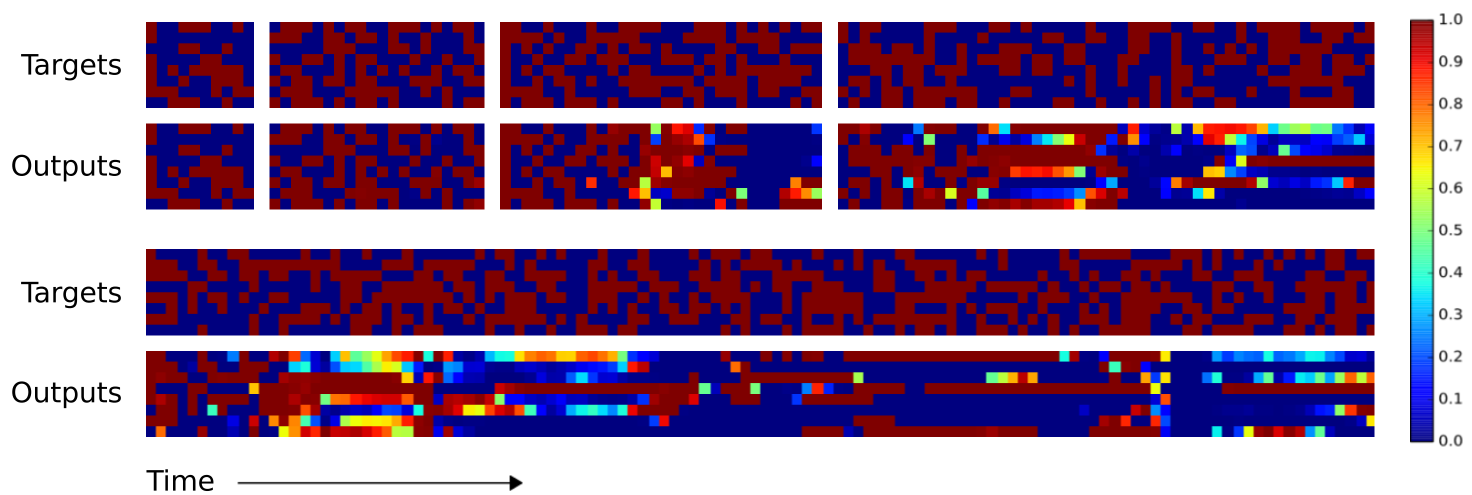
NTM: Experiments

Copy Experiment

Read the input sequence and re-generate it after finished reading it



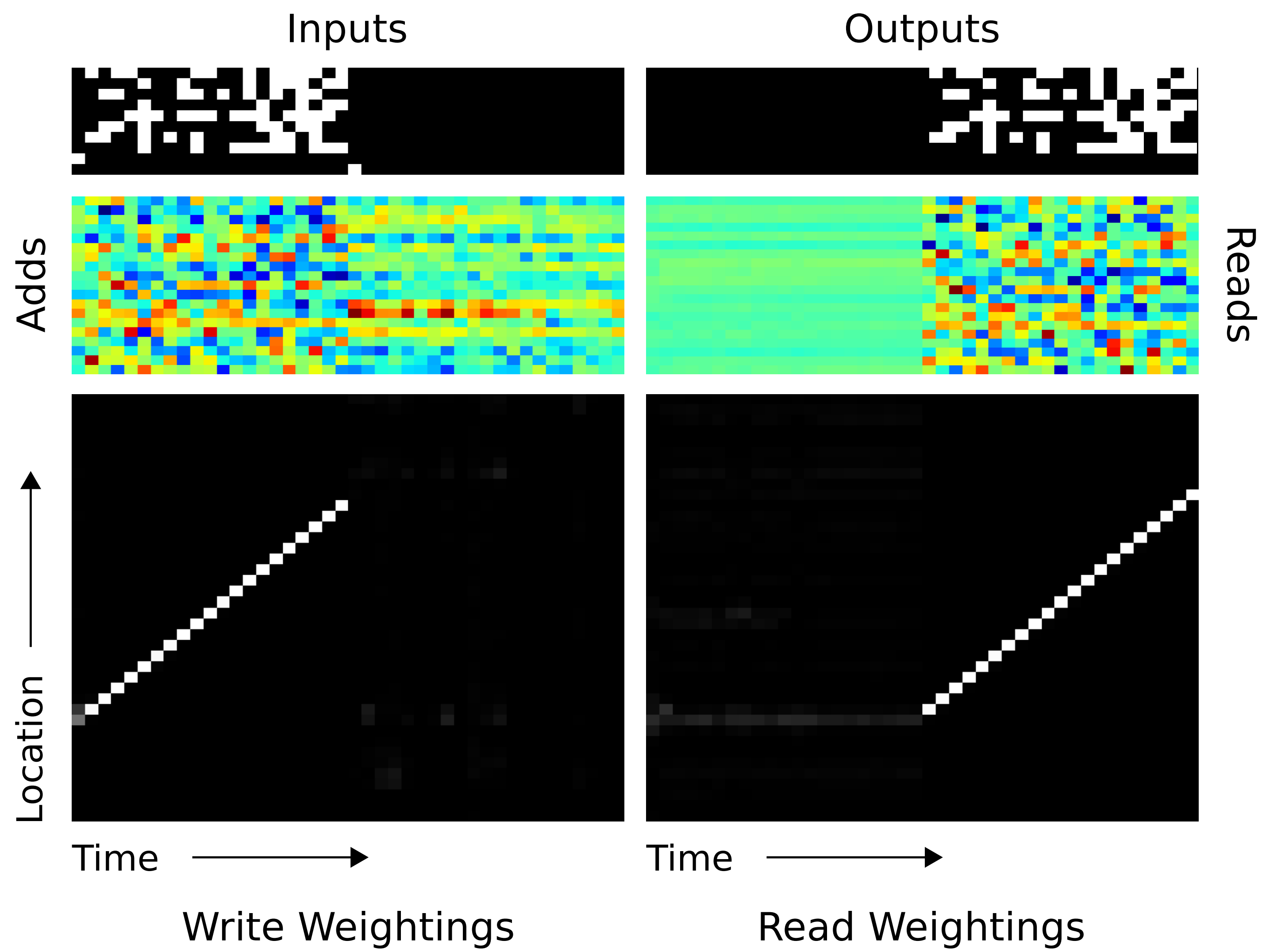




NTM: Experiments

Copy Experiment

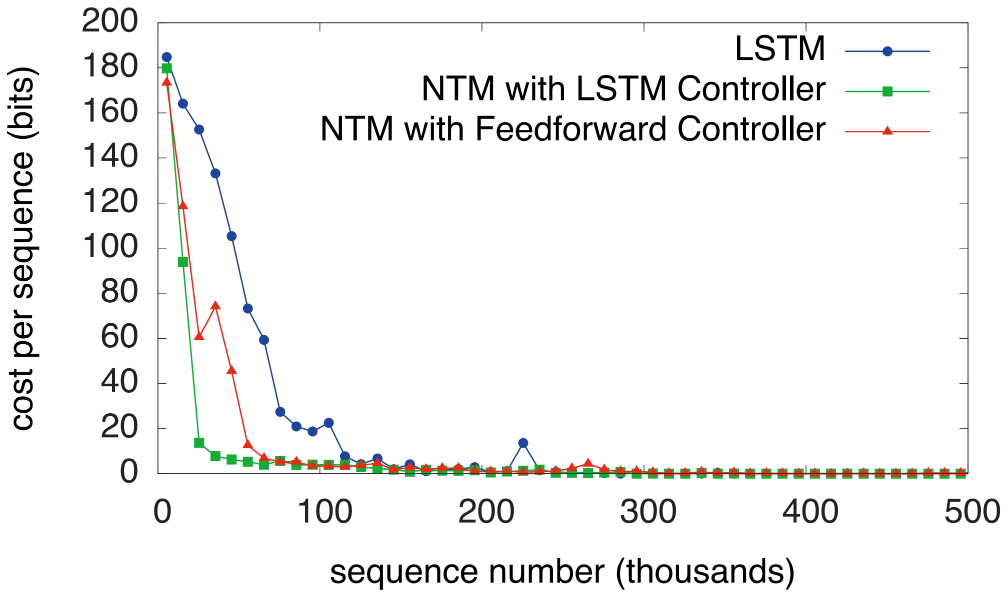
Read the input sequence and re-generate it after finishing reading it

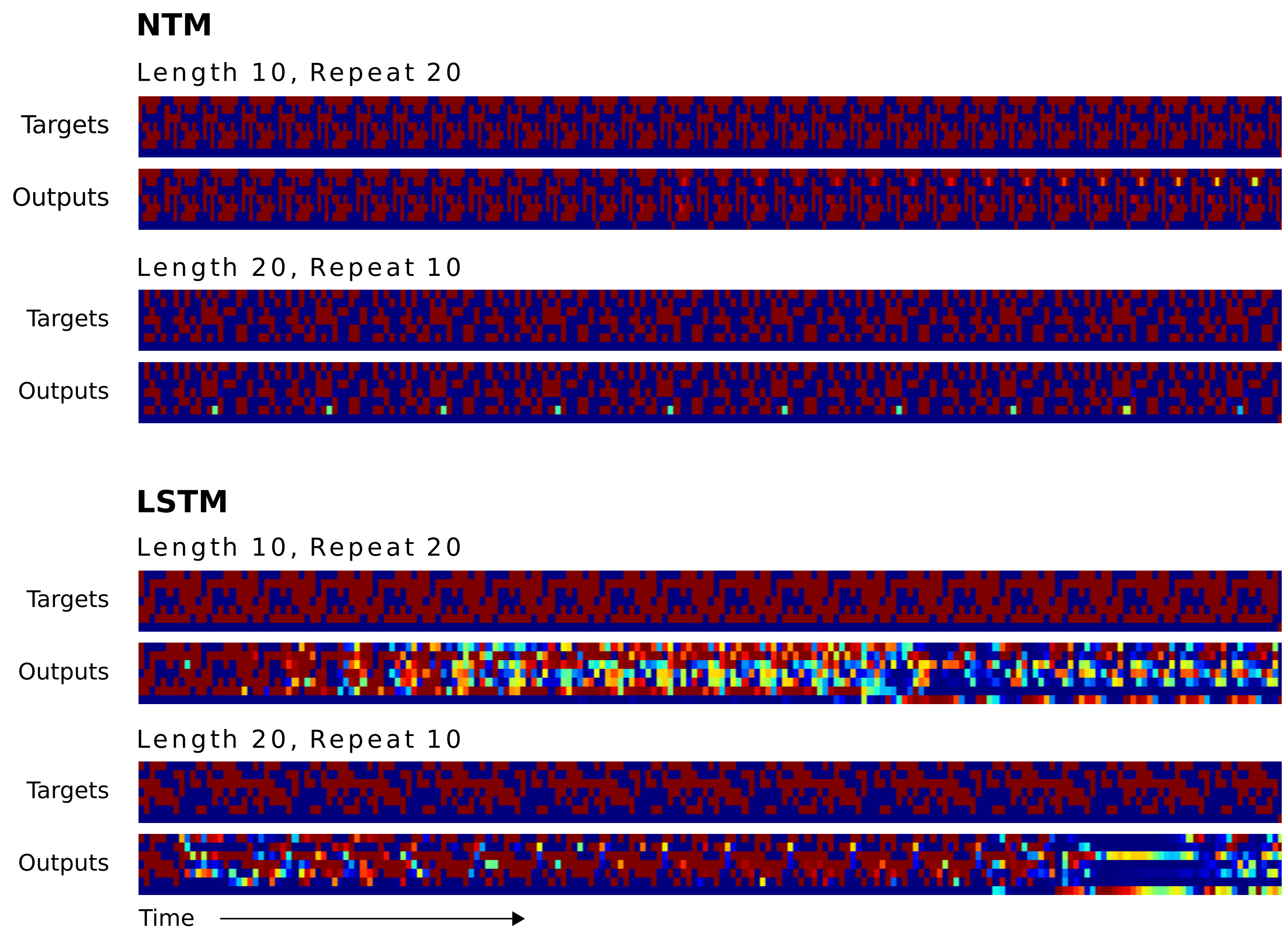


NTM: Experiments

Repeat Copy Experiment

Read the input sequence and re-generate it after finishing reading it N number of times

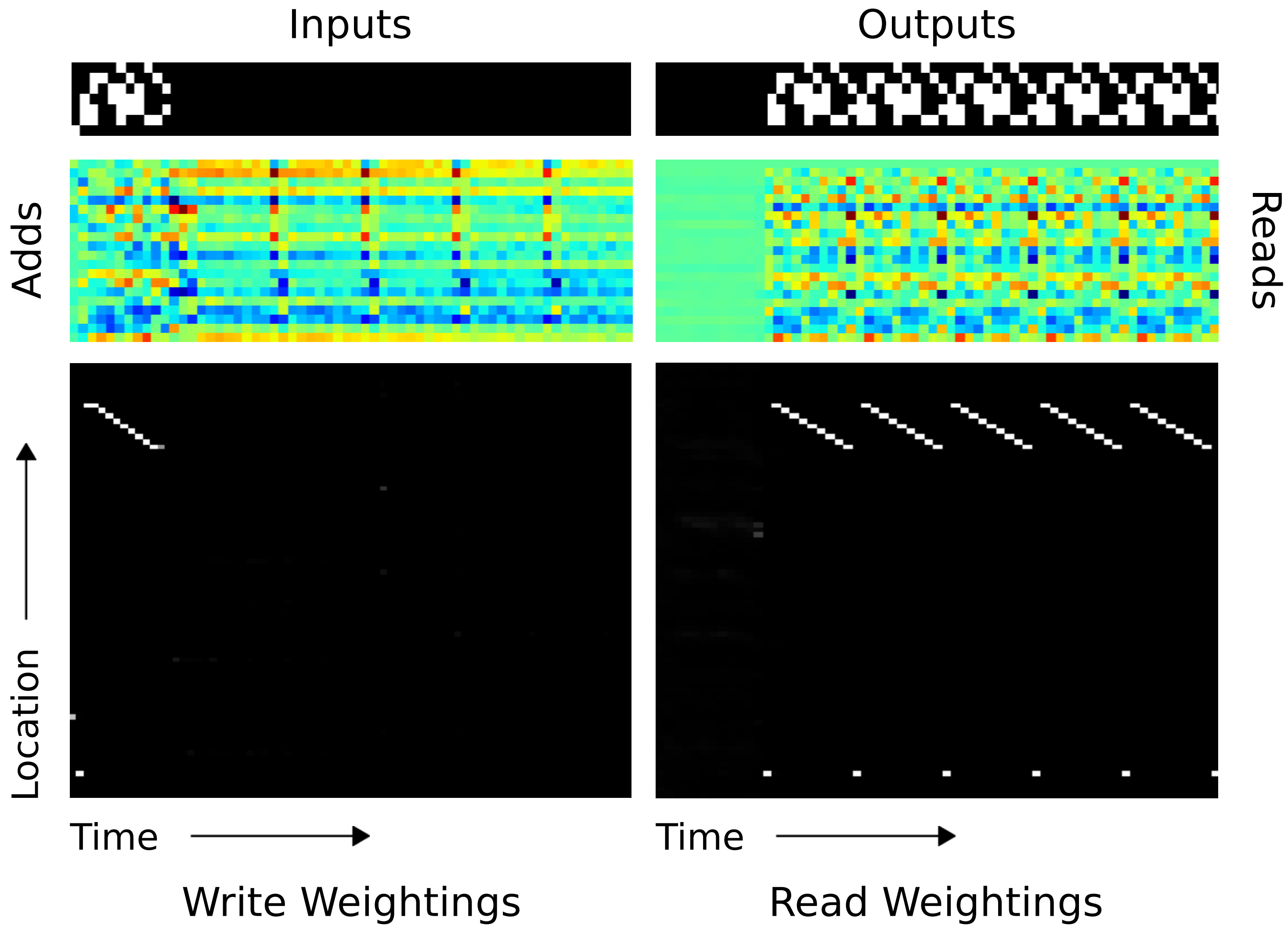




NTM: Experiments

Repeat Copy Experiment

Read the input sequence and re-generate it after finishing reading it N number of times



NTM: Experiments

Sorting Experiment

Sort a collection of vectors according to their given priority

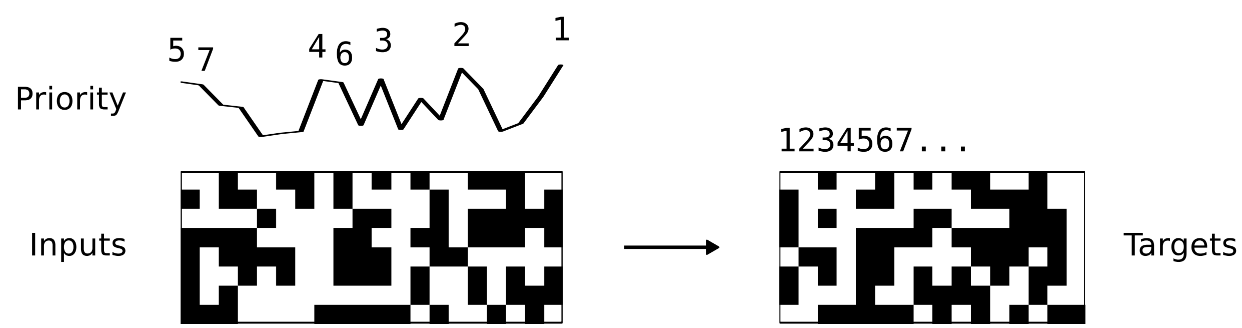


Figure 16: Example Input and Target Sequence for the Priority Sort Task. The input sequence contains random binary vectors and random scalar priorities. The target sequence is a subset of the input vectors sorted by the priorities.

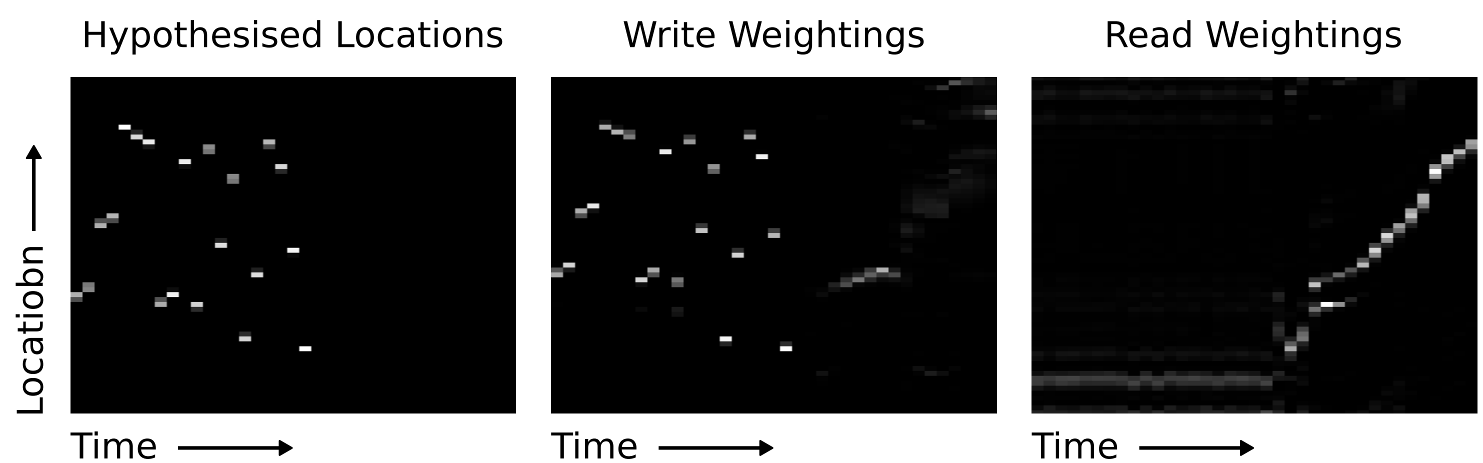
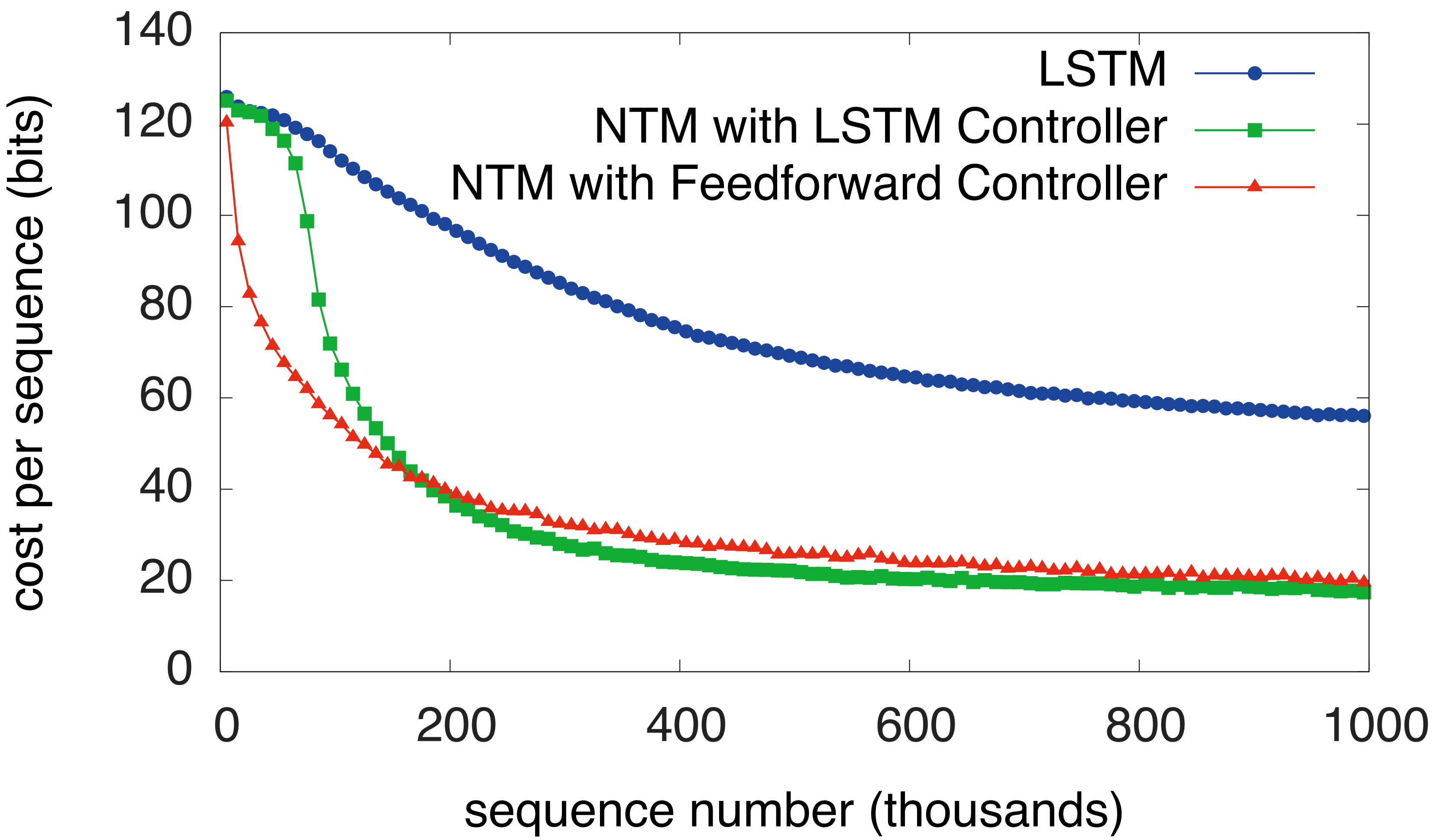


Figure 17: NTM Memory Use During the Priority Sort Task. Left: Write locations returned by fitting a linear function of the priorities to the observed write locations. Middle: Observed write locations. Right: Read locations.

NTM: Experiments

Sorting Experiment

Sort a collection of vectors according to their given priority



NTM: Summary

Another way to augment external memory with a standard deep network

The writer is general enough that it can erase the previous contents of the memory and write new content

Addressing mechanism is more sophisticated than MemNNs

As yet, shown only to work on toy problems which require only small amounts of memory.*

NTM: Summary

Another way to augment external memory with a standard deep network

The writer is general enough that it can erase the previous contents of the memory and write new content

Addressing mechanism is more sophisticated than MemNNs

As yet, shown only to work on toy problems which require only small amounts of memory.*

Very recently there has been some new developments in this area

Dynamic Neural Turing Machine with Soft and Hard Addressing Schemes: Gulcehre et. al., 2016

One-Shot Learning with Memory Augmented Neural Networks: Santoro et. al., 2016

Stack Augmented RNNs

So far we've dealt with memories which are like tapes

For MemNNs the tapes are write-once read-multiple

For NTM tapes are write-multiple read multiple

Natural to think of other forms of memory data structures - stacks, lists, queues, de-queues and more

Stack Augmented RNNs

A number of people have worked on such architectures

Learning Context-Free Grammars: Capabilities and Limitations of a Recurrent Neural Network with External Stack Memory: Das et. al., 1992

A Connectionist Symbol Manipulator that Discovers the Structure of Context Free Languages: Mozer and Das, 1993

The Induction of Dynamical Recognizers: Pollack, 1991

Discrete Recurrent Neural Networks for Grammatical Inference: Zeng et. al., 1994

Inferring Algorithmic Patterns with Stack-Augmented Recurrent Nets: Joulin and Mikolov, 2015

Learning to Transduce with Unbounded Memory: Grefenstette et. al., 2015

Stack Augmented RNNs

Inferring Algorithmic Patterns with Stack-Augmented Recurrent Nets, Joulin and Mikolov, 2015

Standard Recurrent Net

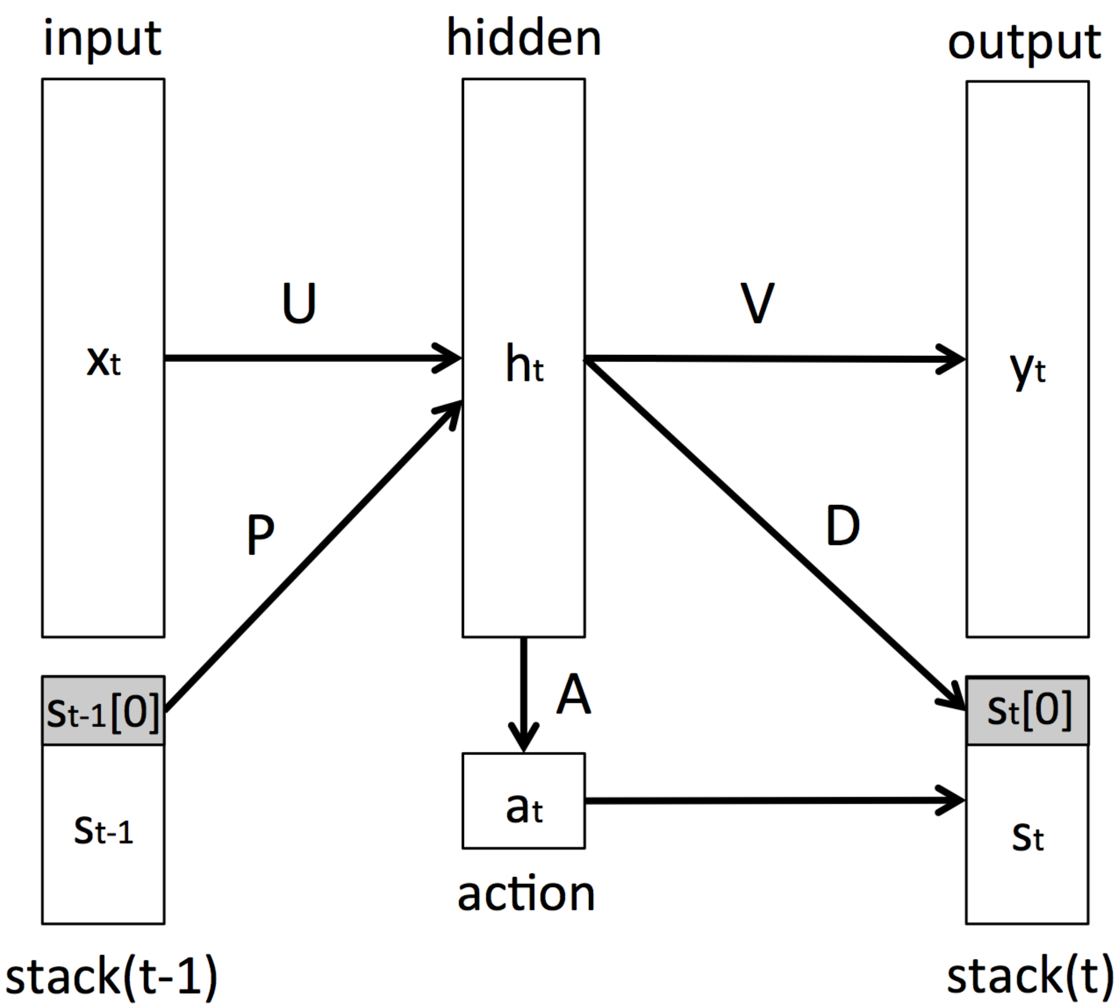
$$h_t = \sigma(Ux_t + Rh_{t-1})$$

$$y_t = g(Vh_t)$$

Stack Augmented RNNs

Inferring Algorithmic Patterns with Stack-Augmented Recurrent Nets, Joulin and Mikolov, 2015

Stack Augmented Recurrent Net



$$a_t = f\left(Ah_t\right)$$

$$s_t[0] = a_t[\mathrm{PUSH}]\sigma(Dh_t) + a_t[\mathrm{POP}]s_{t-1}[1],$$

$$s_t[i] = a_t[\mathrm{PUSH}]s_{t-1}[i-1] + a_t[\mathrm{POP}]s_{t-1}[i+1].$$

$$h_t = \sigma\left(Ux_t + Rh_{t-1} + Ps_{t-1}^k\right),$$

Stack Augmented RNNs

Inferring Algorithmic Patterns with Stack-Augmented Recurrent Nets, Joulin and Mikolov, 2015

method	$a^n b^n$	$a^n b^n c^n$	$a^n b^n c^n d^n$	$a^n b^{2n}$	$a^n b^m c^{n+m}$
RNN	25%	23.3%	13.3%	23.3%	33.3%
LSTM	100%	100%	68.3%	75%	100%
List RNN 40+5	100%	33.3%	100%	100%	100%
Stack RNN 40+10	100%	100%	100%	100%	43.3%
Stack RNN 40+10 + rounding	100%	100%	100%	100%	100%

Table 2: Comparison with RNN and LSTM on sequences generated by counting algorithms. The sequences seen during training are such that $n < 20$ (and $n + m < 20$), and we test on sequences up to $n = 60$. We report the percent of n for which the model was able to correctly predict the sequences. Performance above 33.3% means it is able to generalize to never seen sequence lengths.

Wrapping Up

We discussed the importance of having a persistent memory in models for a number of problems

Memory Networks — Neural Turing Machines — Stack Augmenting RNNs

Attention Mechanism (soft/hard) seems to be one fundamental way of implementing things

Quite a bit lacking still

Wrapping Up

How to decide what to write and what not to write

How to decide which type of memory to use and when?

How to represent knowledge stored in memory

How to incorporate forgetting/compression of information

How to build hierarchical memories: multi scale attention?

How to build hierarchical reasoning: composition of functions?

Thank You!