Dynamic Memory Networks for Question Answering over Text and Images

Richard Socher

Joint work with the MetaMind team
Caiming Xiong, Stephen Merity, James Bradbury, Ankit Kumar, Ozan Irsoy and others

MetaMind
A Salesforce Company

Current Research

All NLP/AI tasks can be reduced to question answering

QA Examples

I: Mary walked to the bathroom.
I: Sandra went to the garden.
I: Daniel went back to the garden.
I: Sandra took the milk there.
Q: Where is the milk?
A: garden
I: Everybody is happy.
Q: What's the sentiment?
A: positive

I: Jane has a baby in Dresden.
Q: What are the named entities?
A: Jane - person, Dresden - location
I: Jane has a baby in Dresden.
Q: What are the POS tags?
A: NNP VBZ DT NN IN NNP .
I: I think this model is incredible
Q: In French?
A: Je pense que ce modèle est incroyable.

Goal

A joint model for general QA

First Major Obstacle

- For NLP no single model **architecture** with consistent state of the art results across tasks

Task	State of the art model
Question answering (babI)	Strongly Supervised MemNN (Weston et al 2015)
Sentiment Analysis (SST)	Tree-LSTMs (Tai et al. 2015)
Part of speech tagging (PTB-WSJ)	Bi-directional LSTM-CRF (Huang et al. 2015)

Second Major Obstacle

- Fully joint multitask learning* is hard:
  - Usually restricted to lower layers
  - Usually helps only if tasks are related
  - Often hurts performance if tasks are not related

* meaning: same decoder/classifier
  and not only transfer learning

Tackling First Obstacle

Dynamic Memory Networks

An architecture for any QA task

High level idea for harder questions

- Imagine having to read an article, memorize it, then get asked various questions → Hard!
- You can't store everything in working memory
- **Optimal:** give you the input data, give you the question, allow as many glances as possible

```
1 Mary moved to the bathroom.
2 John went to the hallway.
3 Where is Mary?        bathroom        1
4 Daniel went back to the hallway.
5 Sandra moved to the garden.
6 Where is Daniel?      hallway         4
7 John moved to the office.
8 Sandra journeyed to the bathroom.
9 Where is Daniel?      hallway         4
10 Mary moved to the hallway.
11 Daniel travelled to the office.
12 Where is Daniel?     office          11
13 John went back to the garden.
14 John moved to the bedroom.
15 Where is Sandra?     bathroom        8
1 Sandra travelled to the office.
2 Sandra went to the bathroom.
3 Where is Sandra?      bathroom        2
```

Basic Lego Block: RNNs

- Gated Recurrent Unit (GRU), Cho et al. 2014
- A type of recurrent neural network (RNN), similar to the LSTM
- Consumes and/or generates sequences (chars, words, ...)
- The GRU updates an internal state *h* according to the existing state *h* and the current input *x*: $h_t = GRU(x_t, h_{t-1})$

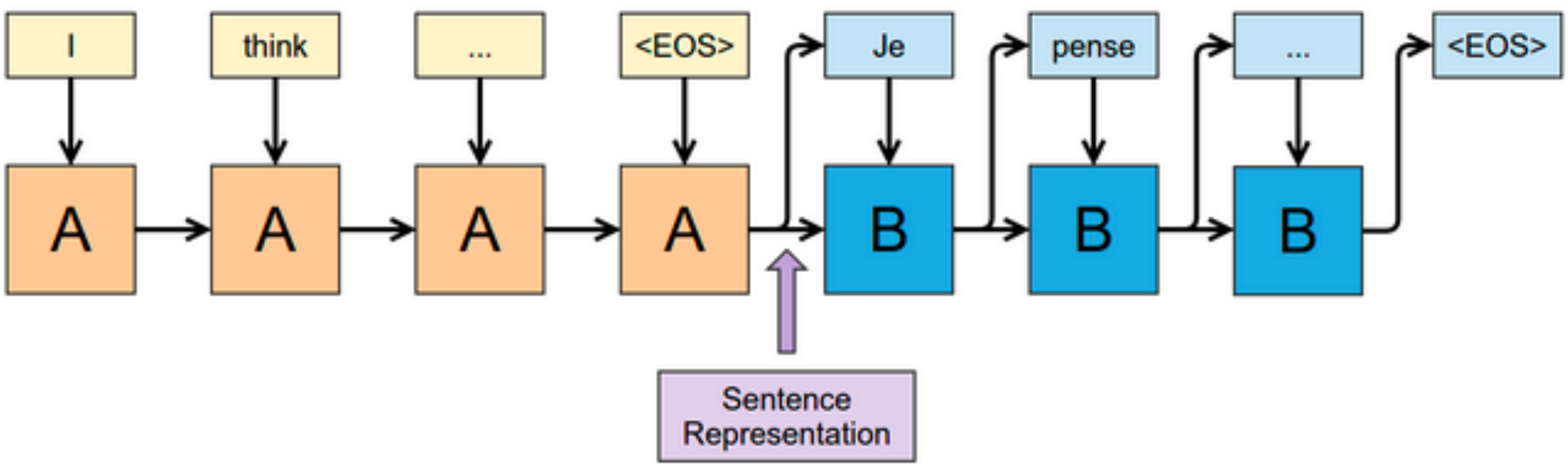
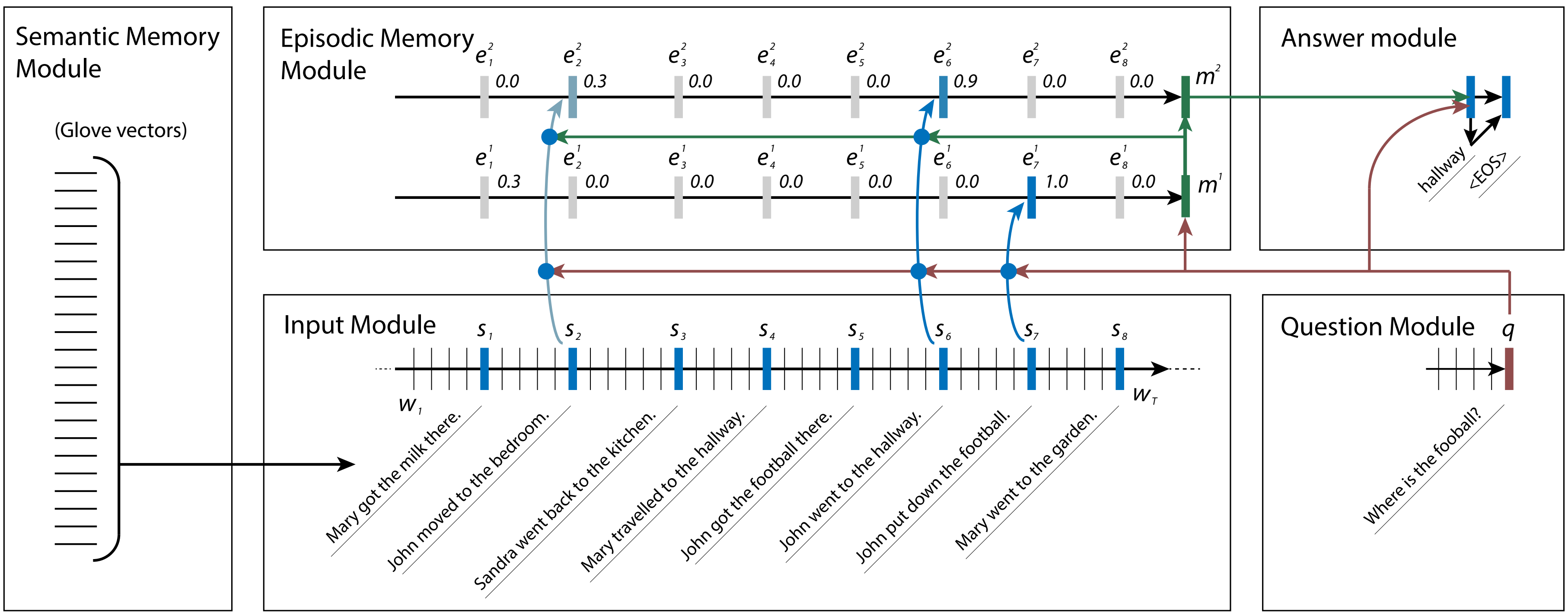
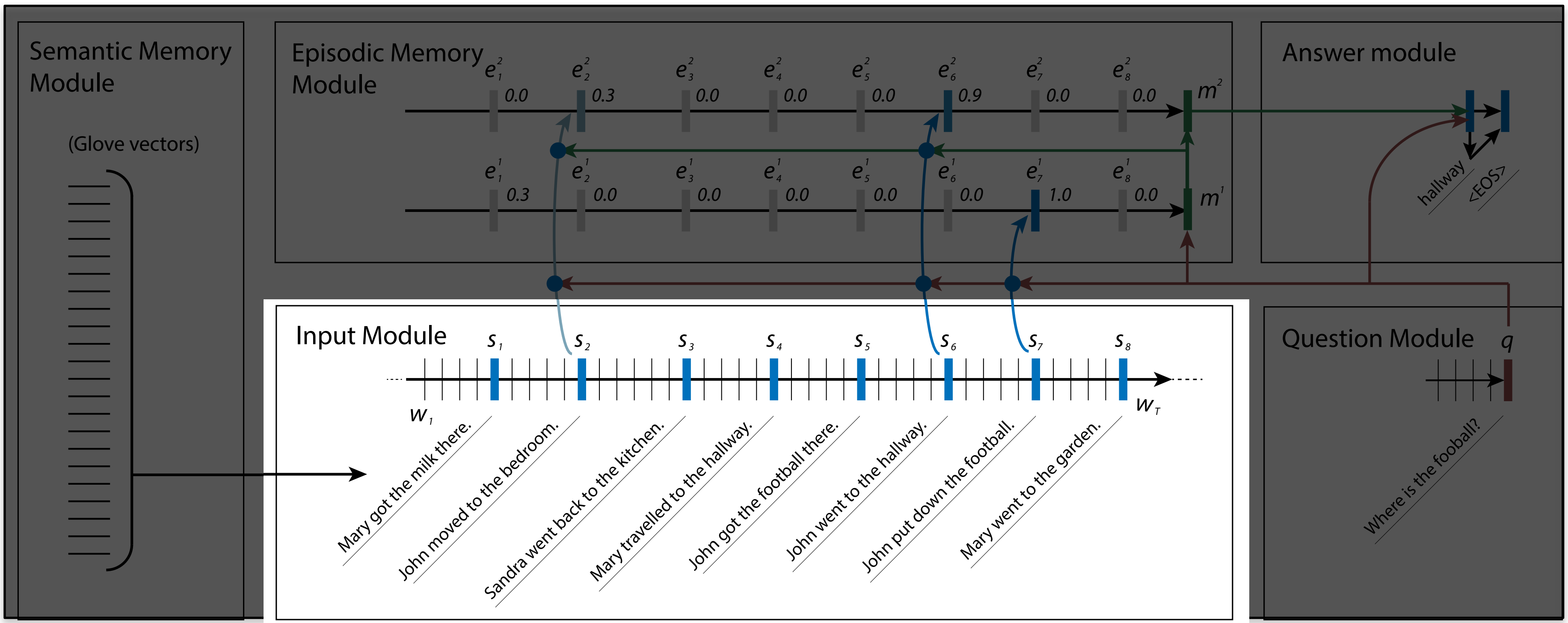


Figure from Chris Olah's Visualizing Representations

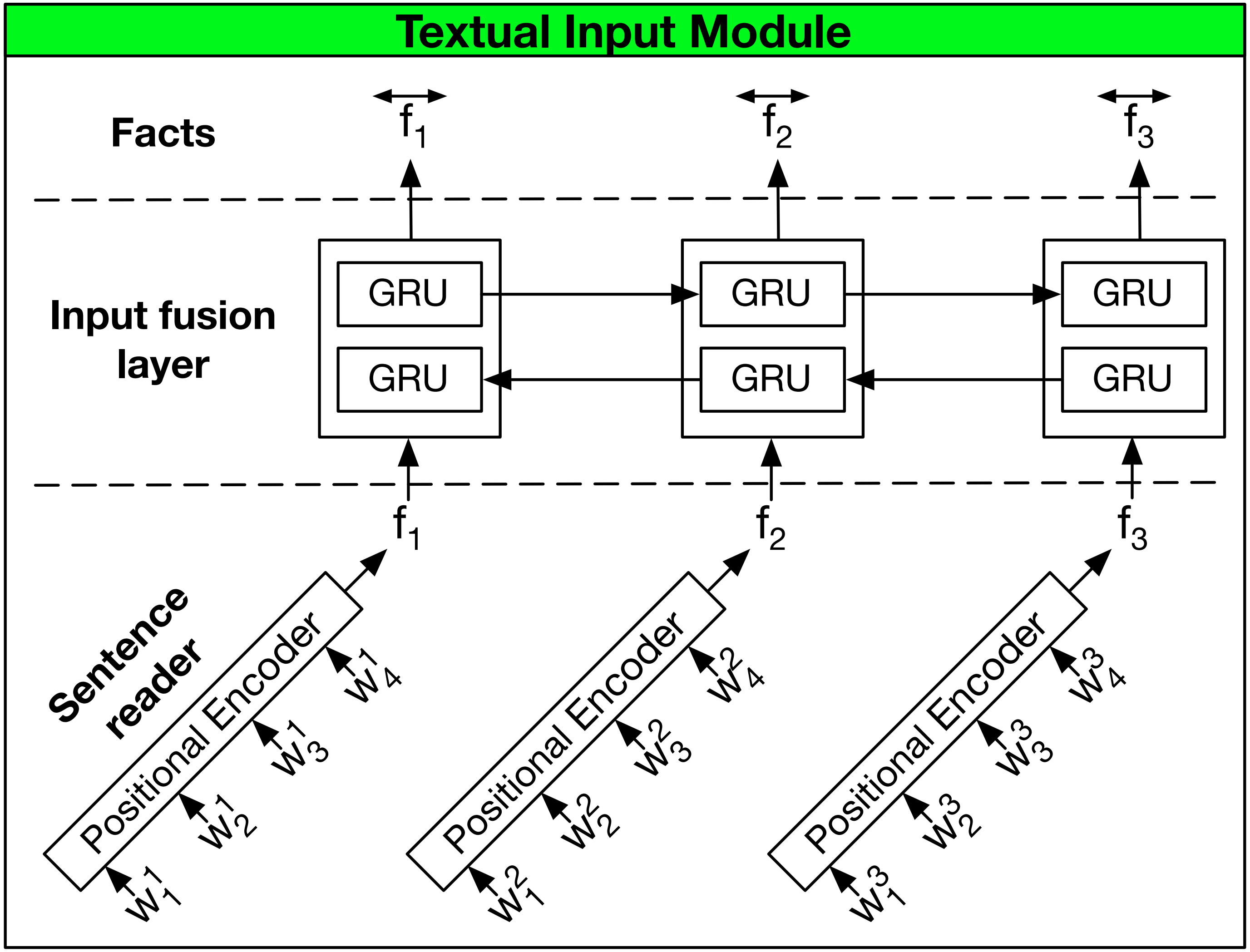
DMN Overview



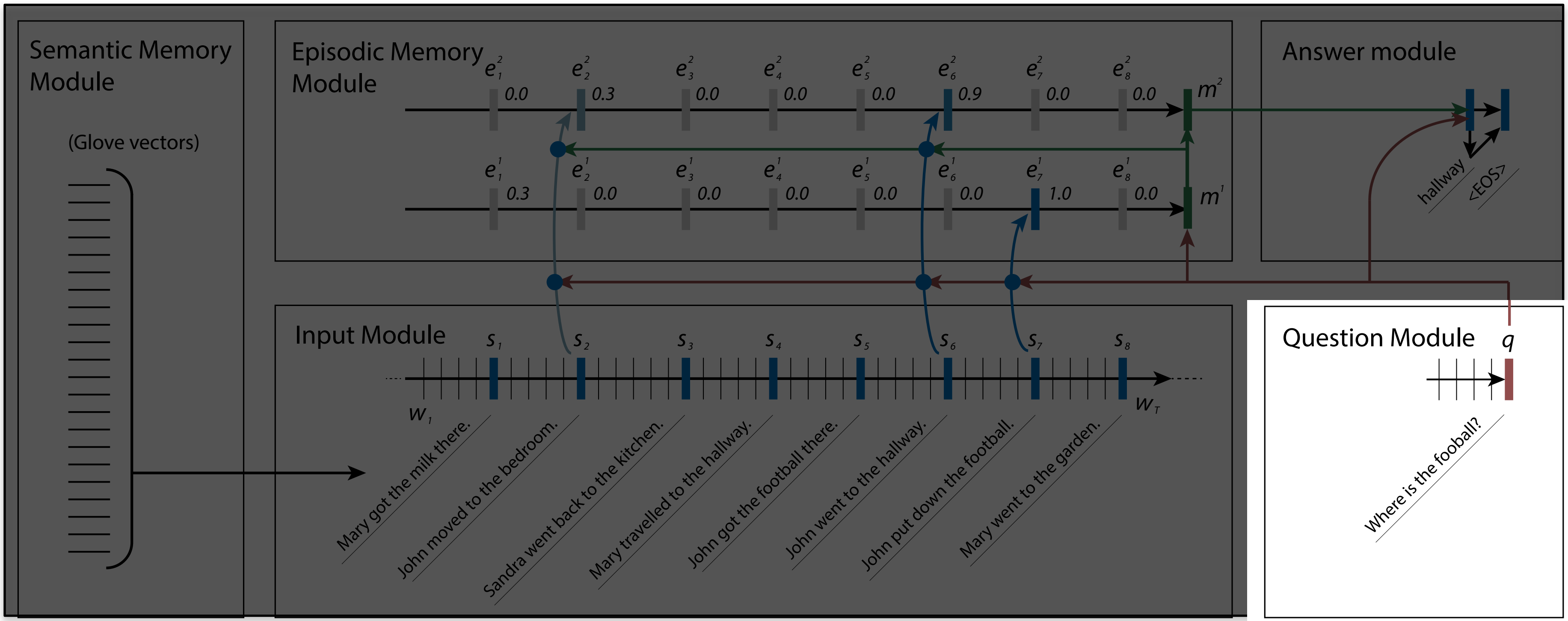
The Modules: Input



Further Improvement: BiGRU

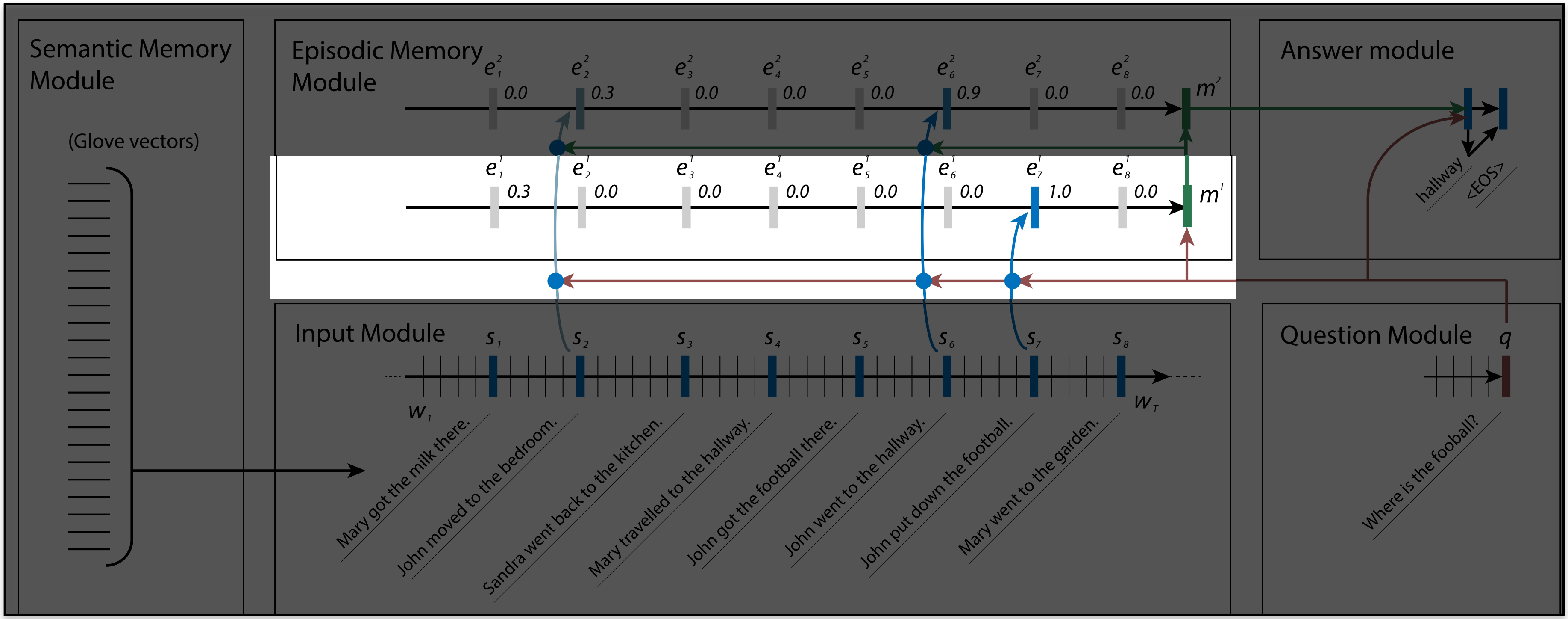


The Modules: Question

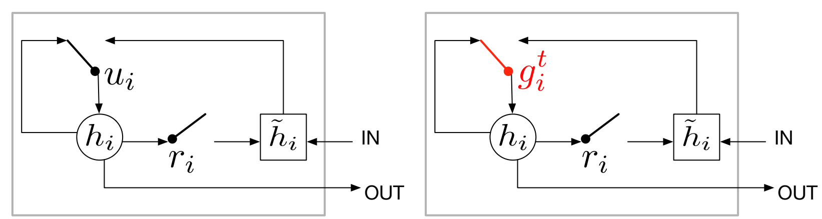


Standard GRU. Output: last hidden state → q

The Modules: Episodic Memory



$$h_i = g_i^t \circ \tilde{h}_i + (1 - g_i^t) \circ h_{i-1}$$



(a) (b)

The Modules: Episodic Memory

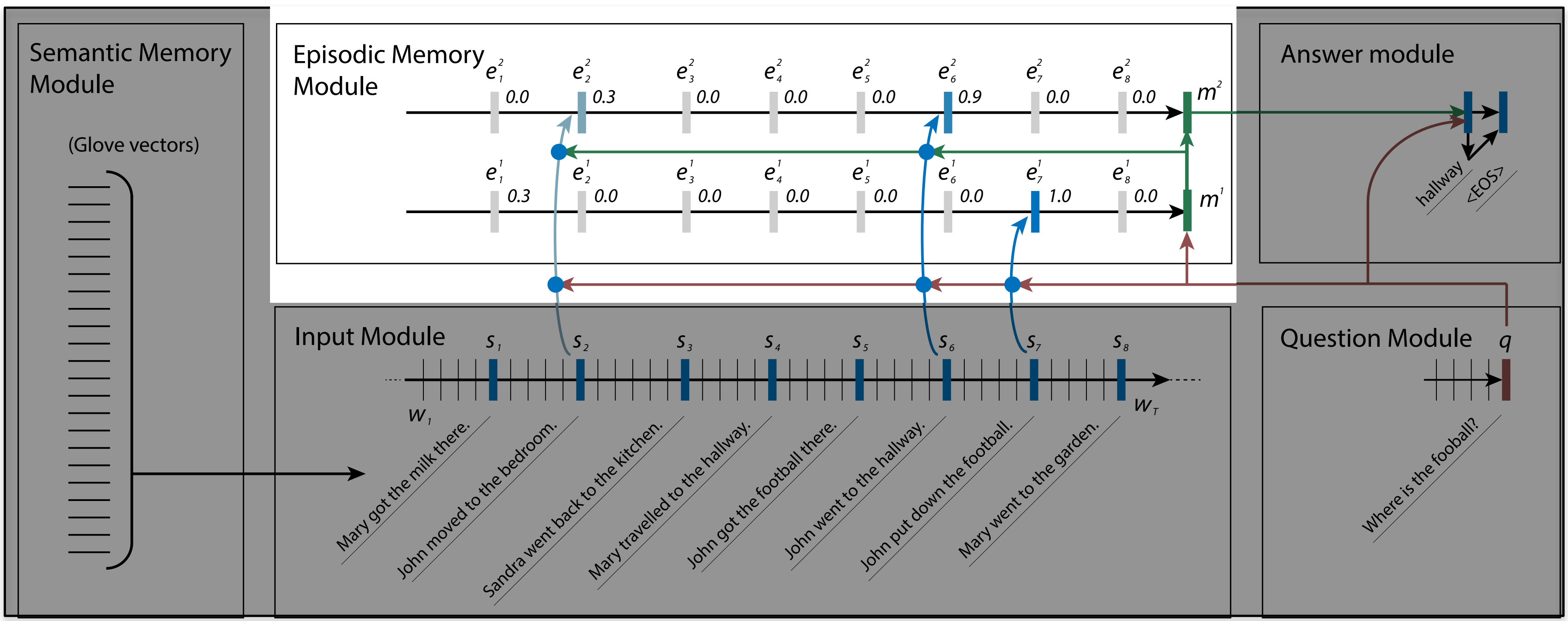
- Gates are activated if relevant to the question

$$
\begin{aligned}
z_i^t &= [\overleftrightarrow{f_i} \circ q; \overleftrightarrow{f_i} \circ m^{t-1}; |\overleftrightarrow{f_i} - q|; |\overleftrightarrow{f_i} - m^{t-1}|] \\
Z_i^t &= W^{(2)} \tanh\left(W^{(1)} z_i^t + b^{(1)}\right) + b^{(2)} \\
g_i^t &= \frac{\exp(Z_i^t)}{\sum_{k=1}^{M_i} \exp(Z_k^t)}
\end{aligned}
$$

- When the end of the input is reached, the relevant facts are summarized in another GRU or simple NNet $m^t = ReLU\left(W^t[m^{t-1}; c^t; q] + b\right)$

The Modules: Episodic Memory

- If summary is insufficient to answer the question, repeat sequence over input



Inspiration from Neuroscience

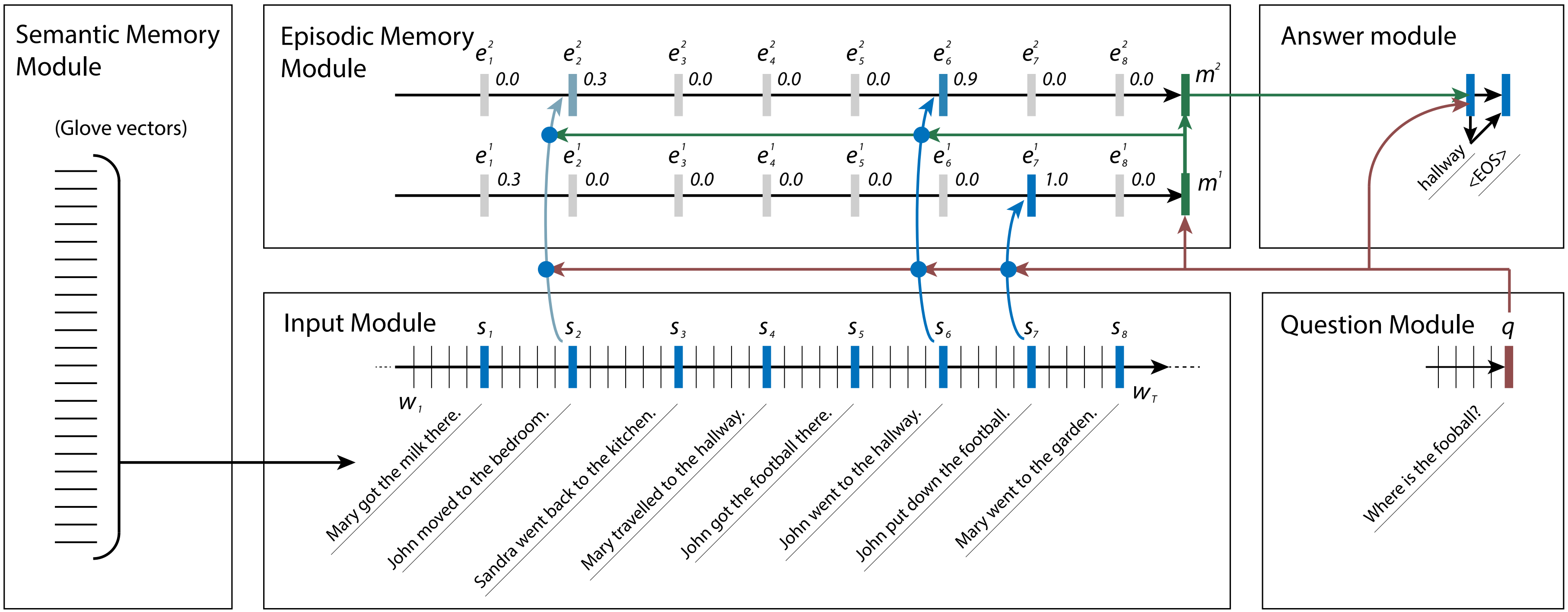
- **Episodic memory** is the **memory** of autobiographical events (times, places, etc). A collection of past personal experiences that occurred at a particular time and place.

- The hippocampus, the seat of episodic memory in humans, is active during transitive inference
- In the DMN repeated passes over the input are needed for transitive inference

The Modules: Answer

Input to Answer GRU: a = [m q]

GRU for sequences, standard softmax for single class



Academic papers and related work

- For full details:
- **Ask Me Anything: Dynamic Memory Networks for Natural Language Processing (Kumar et al., 2015)**
- **Dynamic Memory Networks for Visual and Textual Question Answering (Xiong et al., 2016)**

- Sequence to Sequence (Sutskever et al. 2014)
- Neural Turing Machines (Graves et al. 2014)
- Teaching Machines to Read and Comprehend (Hermann et al. 2015)
- Learning to Transduce with Unbounded Memory (Grefenstette 2015)
- Structured Memory for Neural Turing Machines (Wei Zhang 2015)

- Memory Networks (Weston et al. 2015)
- End to end memory networks (Sukhbaatar et al. 2015) →

Comparison to MemNets

Similarities:

- MemNets and DMNs have input, scoring, attention and response mechanisms

Differences:

- For input representations MemNets use bag of word, nonlinear or linear embeddings that explicitly encode position
- MemNets iteratively run functions for attention and response

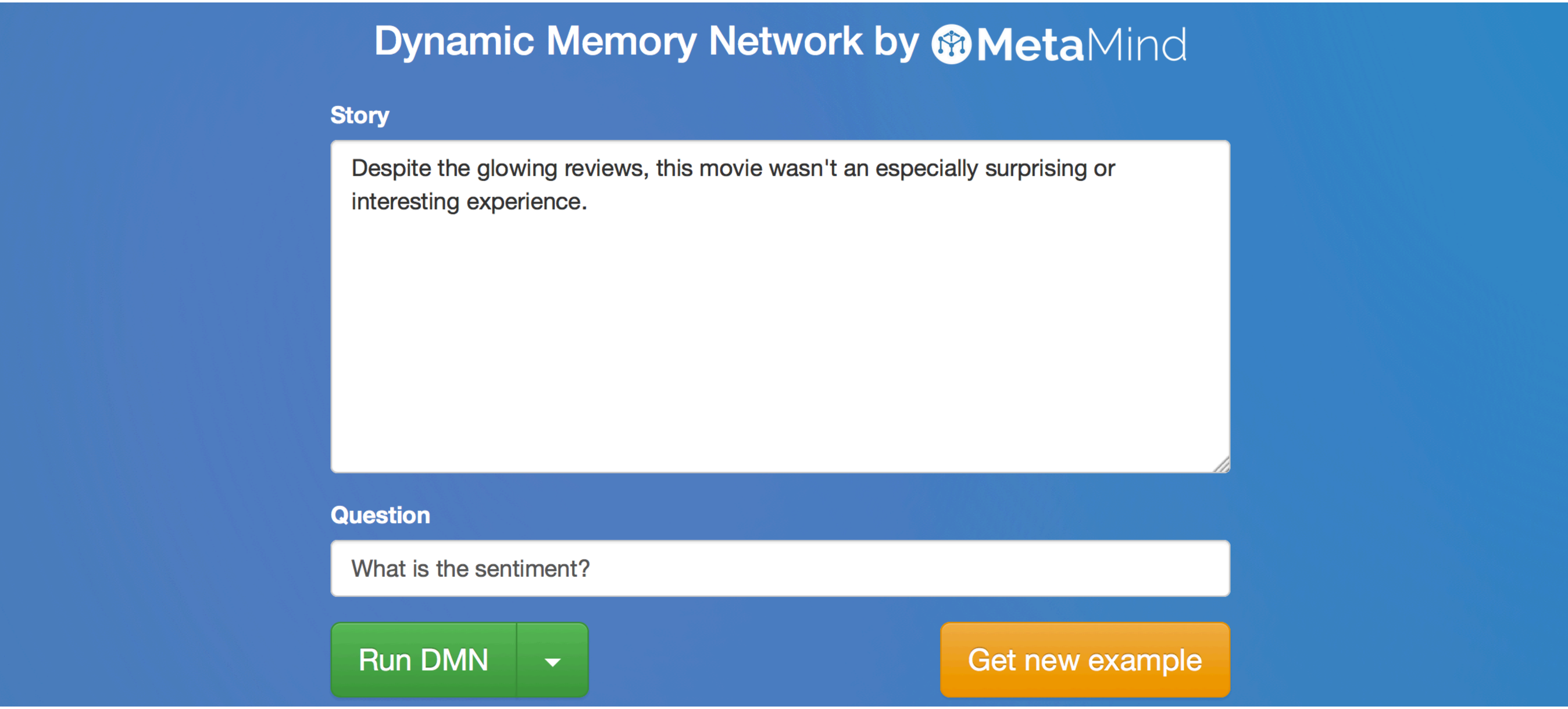
- **DMNs shows that neural sequence models can be used for input representation, attention and response mechanisms**
  → naturally captures position and temporality
- Enables broader range of applications

Experiments: QA on babI (1k)

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

This still requires that relevant facts are marked during training to train the gates.

Live Demo



Dynamic Memory Network by MetaMind
Story
Despite the glowing reviews, this movie wasn't an especially surprising or interesting experience.
Question
What is the sentiment?
Run DMN
Get new example

Experiments: Sentiment Analysis

- Stanford Sentiment Treebank

- Test accuracies:
- 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)
- CT-LSTM: Tai et al. (2015)

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**

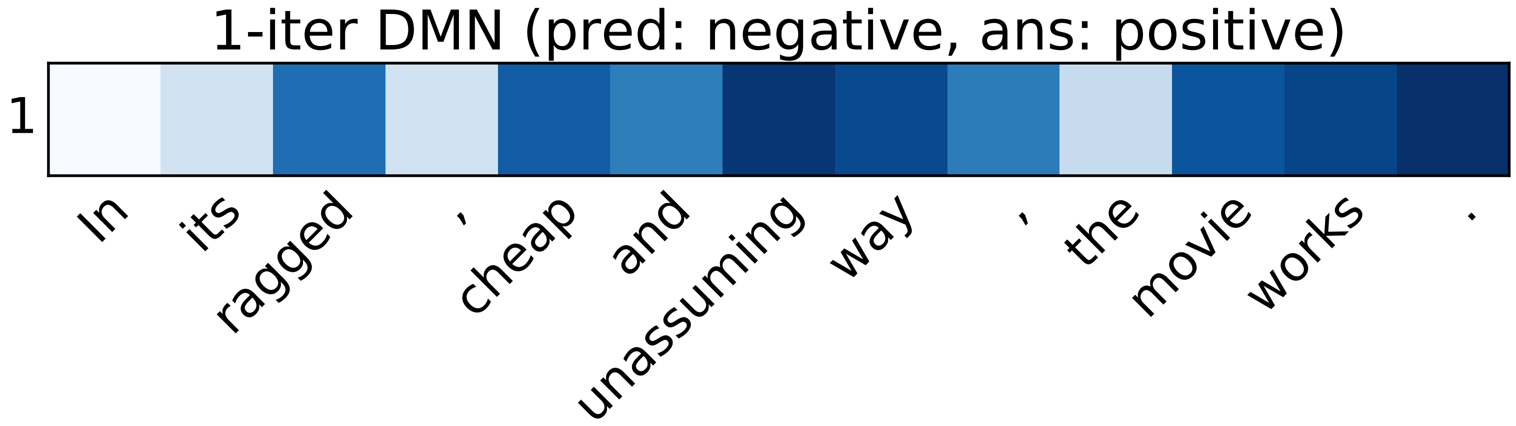
Analysis of Number of Episodes

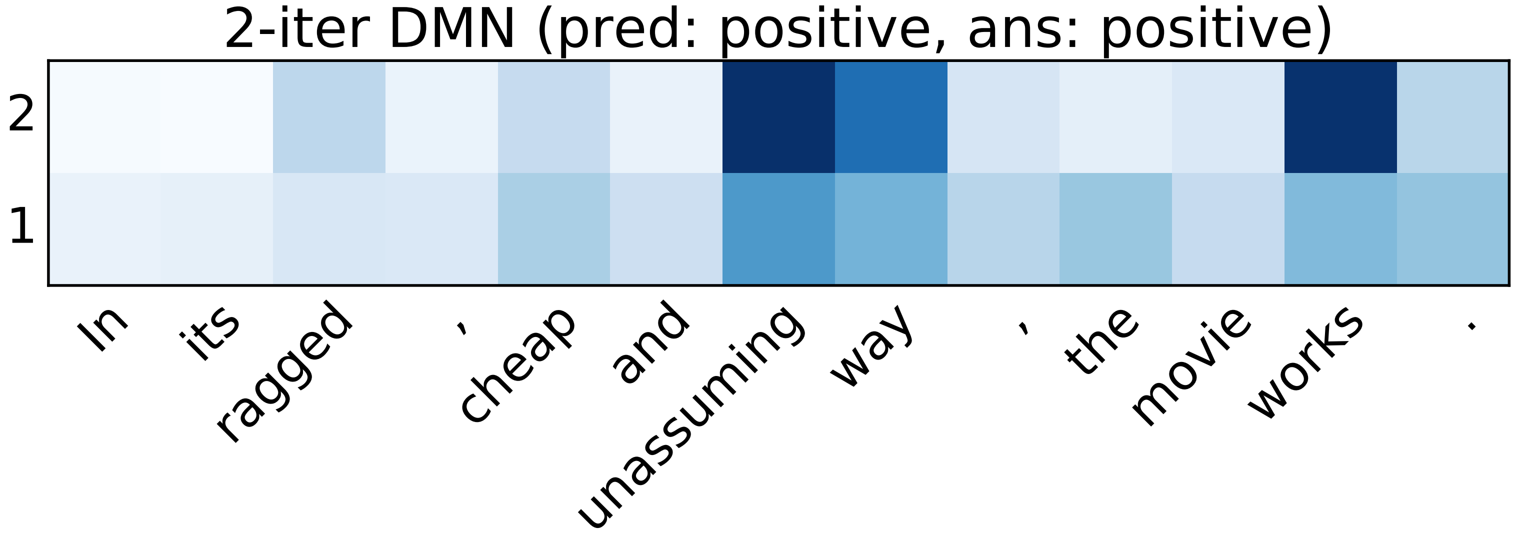
- How many attention + memory passes are needed in the episodic memory?

Max passes	task 3 three-facts	task 7 count	task 8 lists/sets	sentiment (fine grain)
0 pass	0	48.8	33.6	50.0
1 pass	0	48.8	54.0	51.5
2 pass	16.7	49.1	55.6	**52.1**
3 pass	64.7	83.4	83.4	50.1
5 pass	**95.2**	**96.9**	**96.5**	N/A

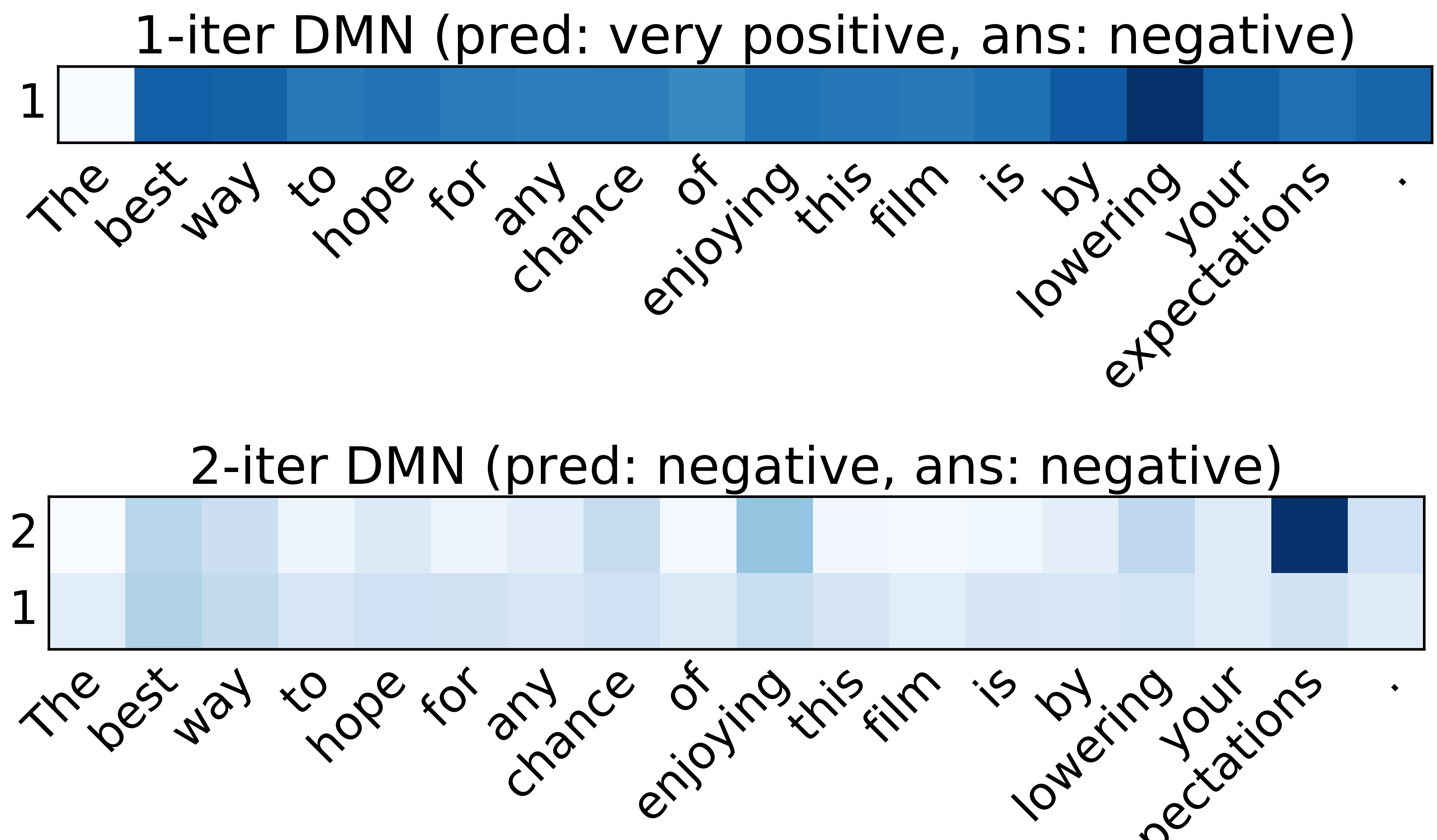
Analysis of Attention for Sentiment

- Sharper attention when 2 passes are allowed.
- Examples that are wrong with just one pass



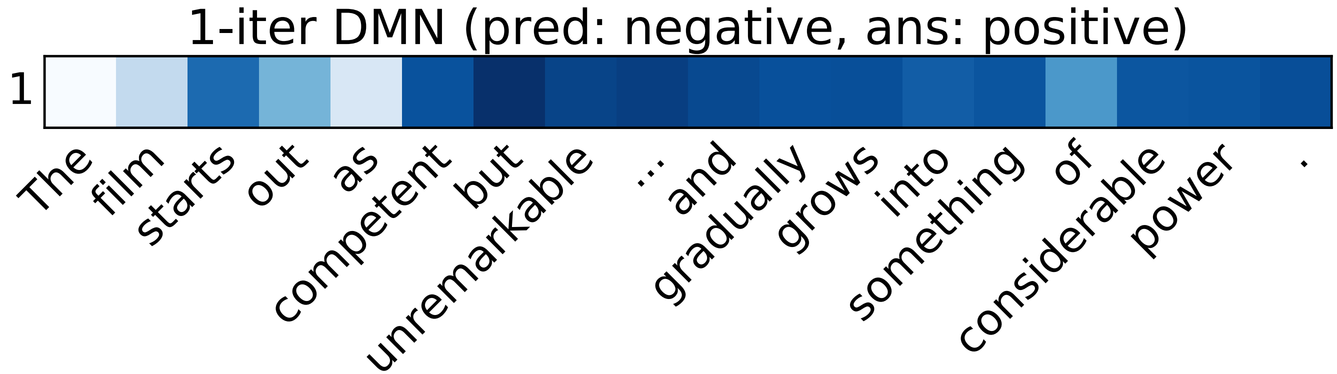


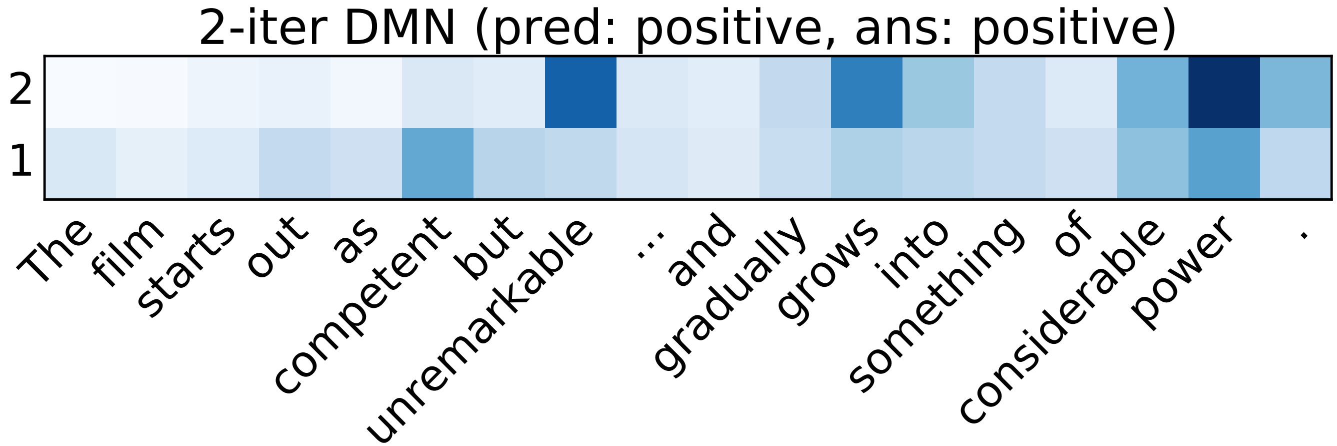
Analysis of Attention for Sentiment



Analysis of Attention for Sentiment

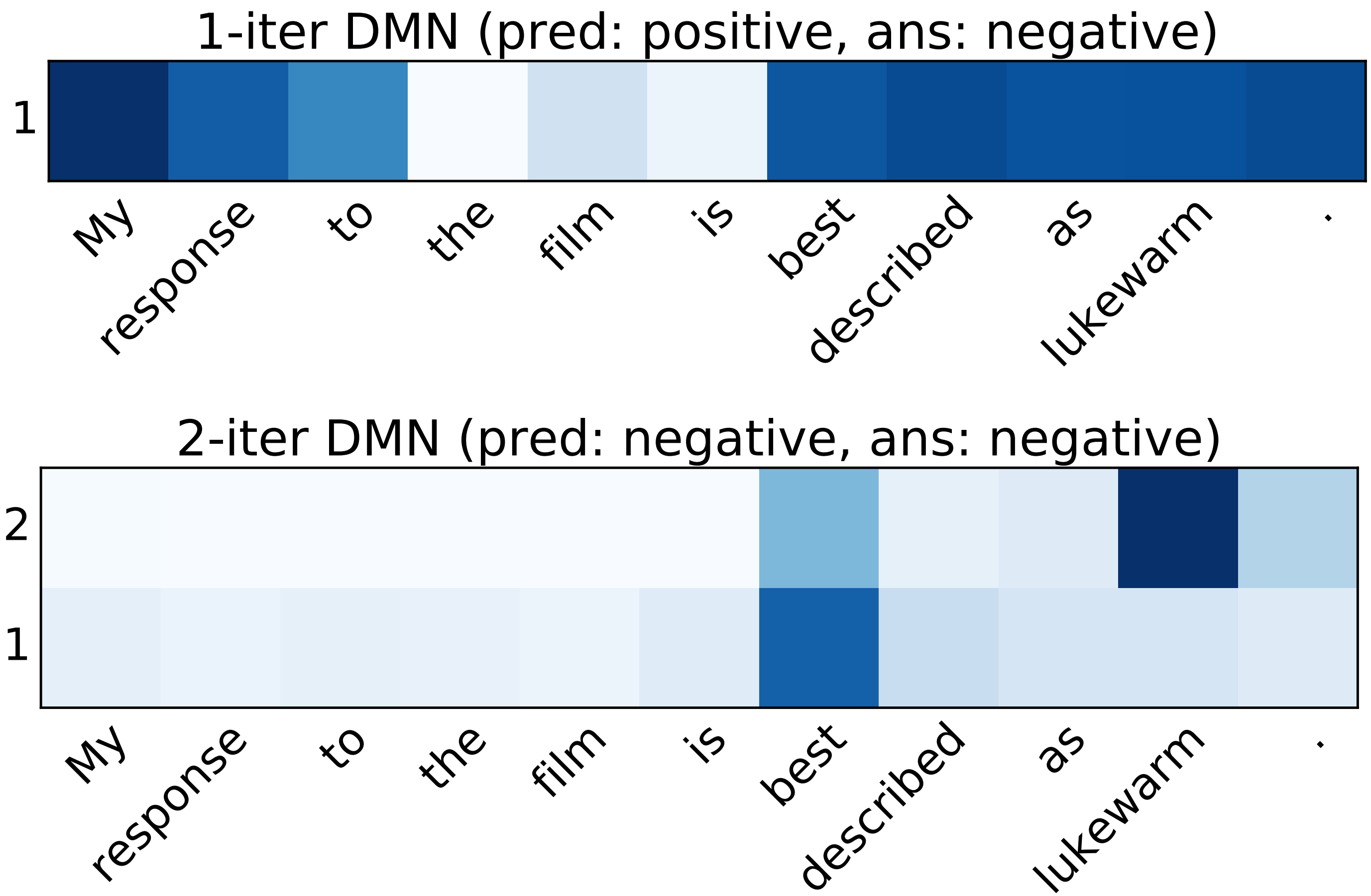
- Examples where full sentence context from first pass changes attention to words more relevant for final prediction



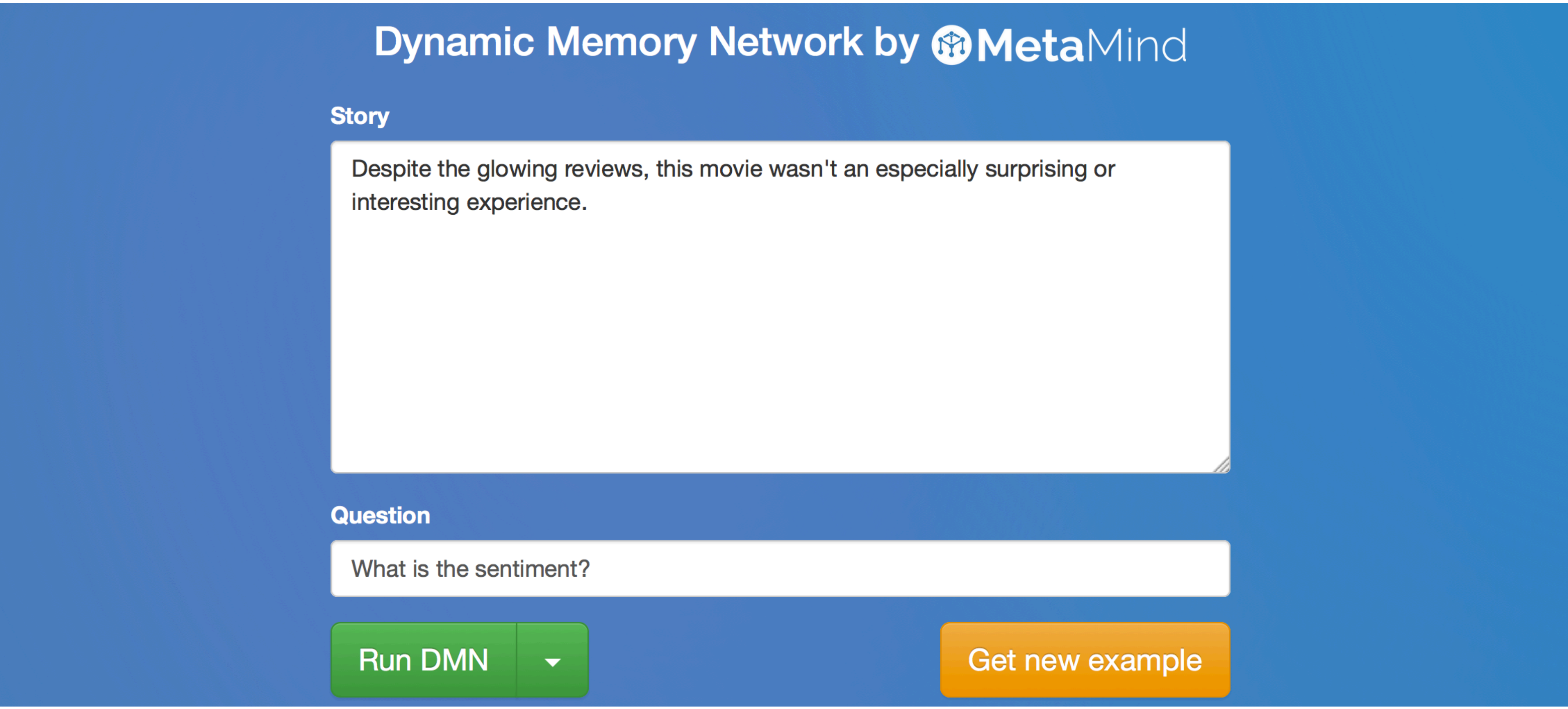


Analysis of Attention for Sentiment

- Examples where full sentence context from first pass changes attention to words more relevant for final prediction



Live Demo

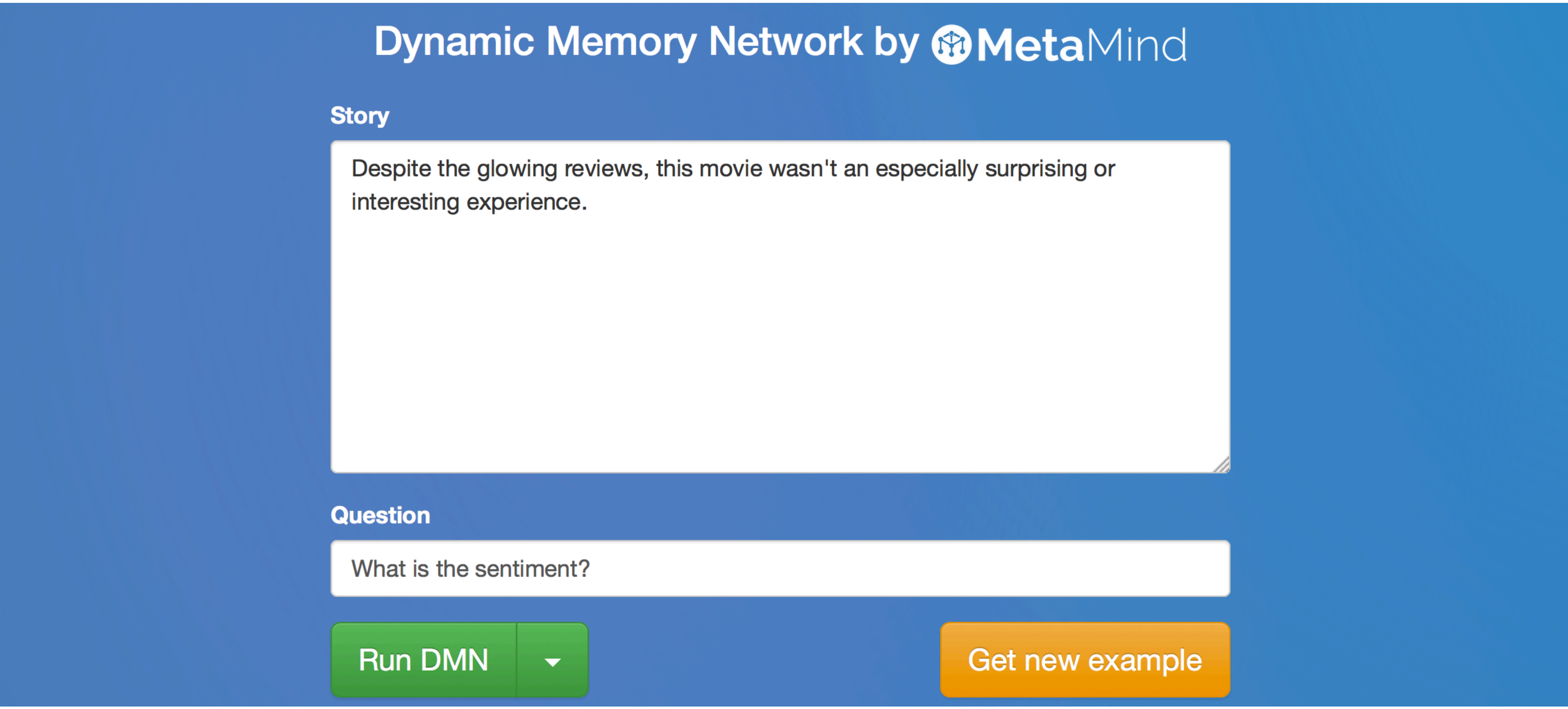

Dynamic Memory Network by MetaMind
Story
Despite the glowing reviews, this movie wasn't an especially surprising or interesting experience.
Question
What is the sentiment?
Run DMN
Get new example

Experiments: POS Tagging

- PTB WSJ, standard splits
- Episodic memory does not require multiple passes, single pass enough

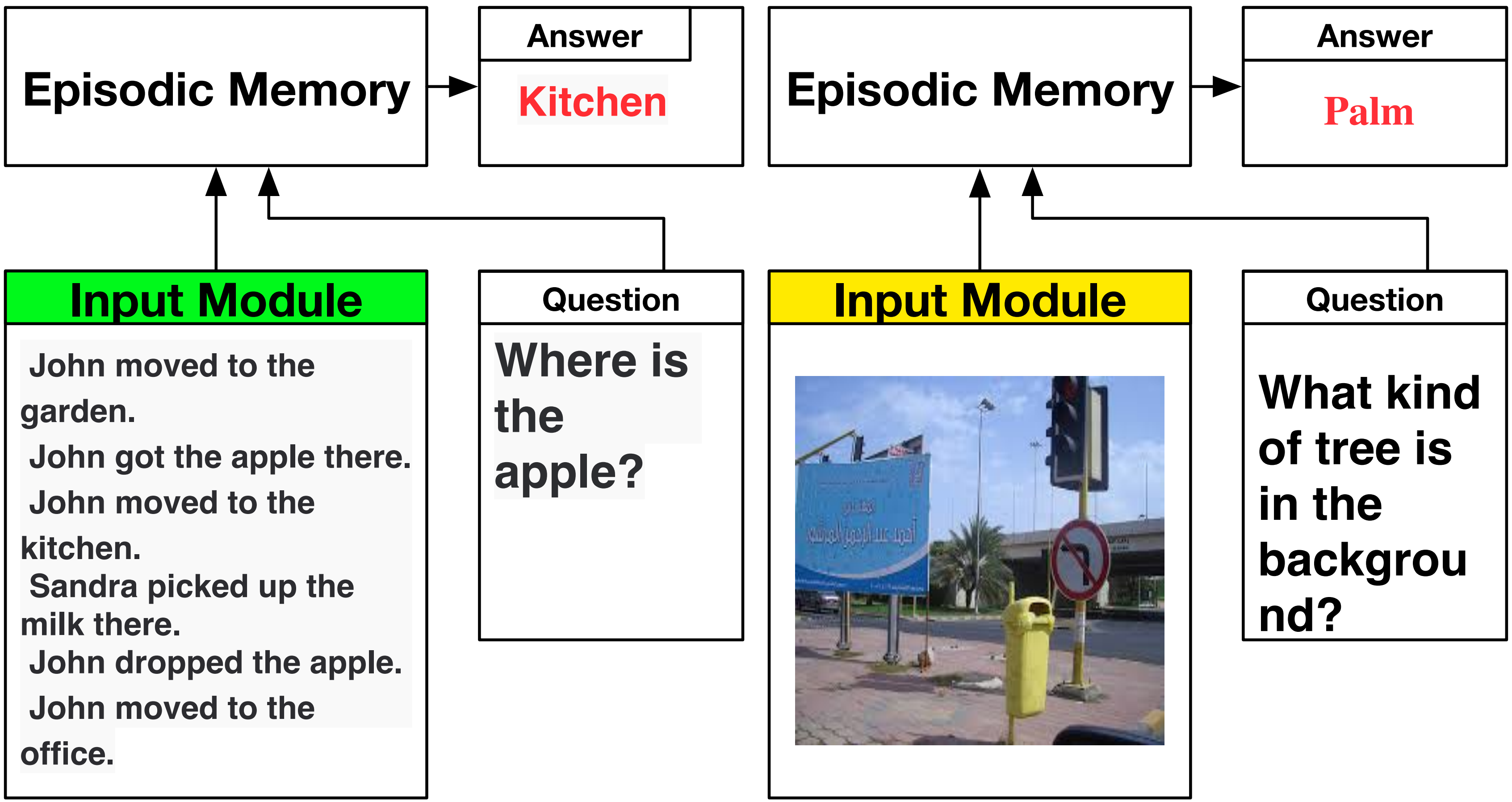
Model	SVMTool	Sogaard	Suzuki et al.	Spoustova et al.	SCNN	DMN
Acc (%)	97.15	97.27	97.40	97.44	97.50	**97.56**

Live Demo

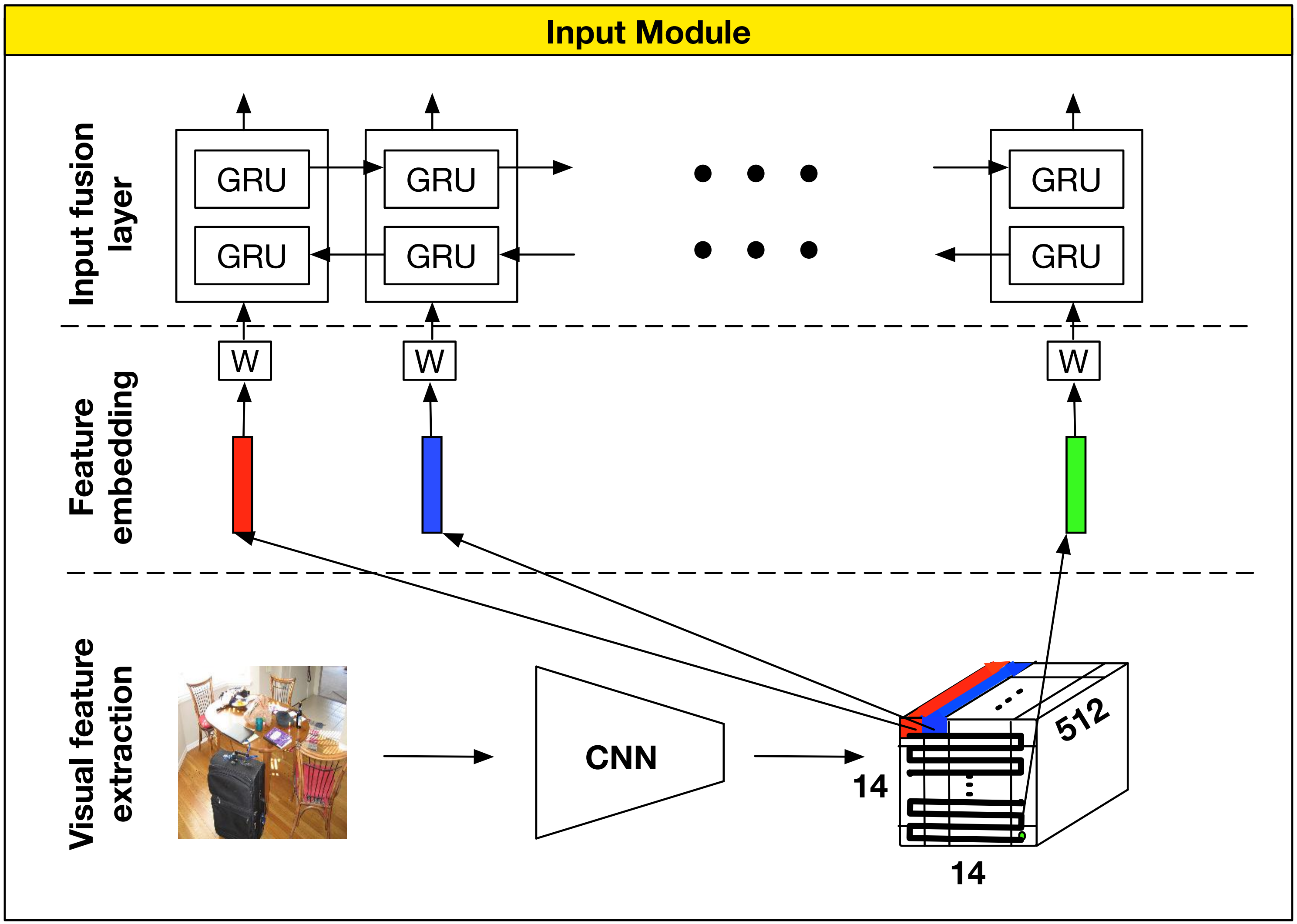


Dynamic Memory Network by MetaMind
Story
Despite the glowing reviews, this movie wasn't an especially surprising or interesting experience.
Question
What is the sentiment?
Run DMN
Get new example

Modularization Allows for Different Inputs



Input Module for Images



Accuracy: Visual Question Answering

VQA test-dev and test-standard:

- Antol et al. (2015)
- ACK Wu et al. (2015);
- iBOWIMG - Zhou et al. (2015);
- DPPnet - Noh et al. (2015); D-NMN - Andreas et al. (2016);
- SAN - Yang et al. (2015)

	test-dev				test-std
Method	All	Y/N	Other	Num	All
VQA					
Image	28.1	64.0	3.8	0.4	-
Question	48.1	75.7	27.1	36.7	-
Q+I	52.6	75.6	37.4	33.7	-
LSTM Q+I	53.7	78.9	36.4	35.2	54.1
ACK	55.7	79.2	40.1	36.1	56.0
iBOWIMG	55.7	76.5	42.6	35.0	55.9
DPPnet	57.2	80.7	41.7	37.2	57.4
D-NMN	57.9	80.5	43.1	37.4	58.0
SAN	58.7	79.3	46.1	36.6	58.9
DMN+	**60.3**	80.5	48.3	36.8	**60.4**

Attention Visualization

What is the main color on the bus ?

Answer: blue

What type of trees are in the background ?

Answer: pine

How many pink flags are there ?

Answer: 2

Is this in the wild ?

Answer: no

Attention Visualization

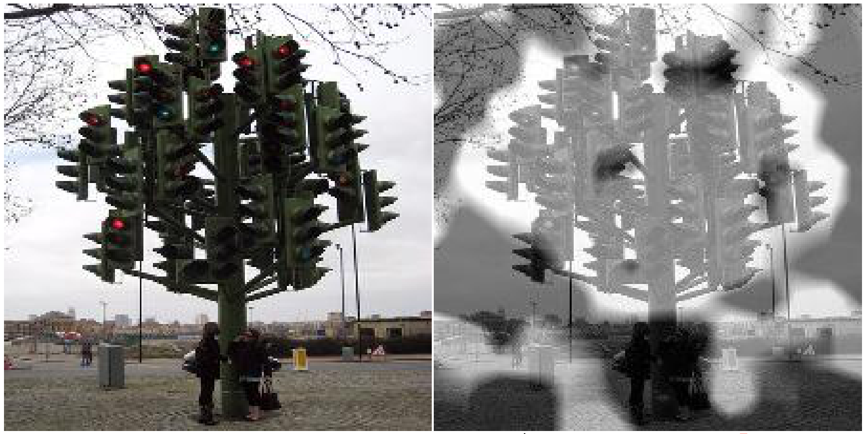
Which man is dressed more flamboyantly ? **Answer: right**

Who is on both photos ? **Answer: girl**

What time of day was this picture taken ? **Answer: night**

What is the boy holding ? **Answer: surfboard**

Attention Visualization



What is this sculpture made out of ?

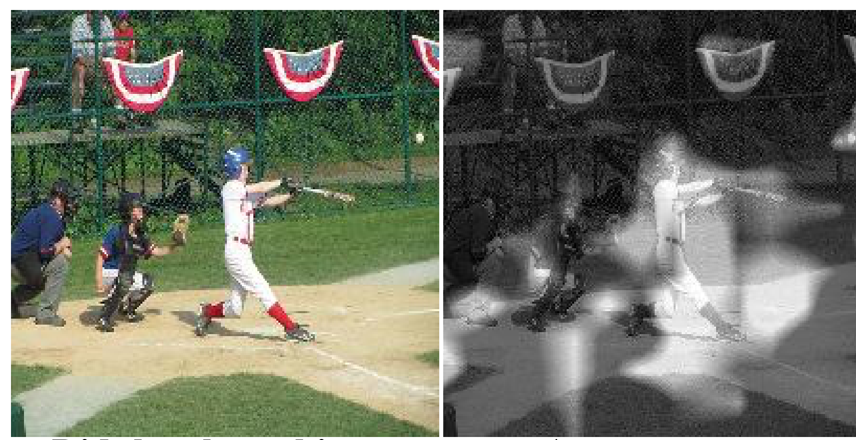
Answer: metal

What color are the bananas ?

Answer: green

What is the pattern on the cat ' s fur on its tail ?

Answer: stripes



Did the player hit the ball ?

Answer: yes

Live Demo


Dynamic Memory Network by MetaMind
Screen Shot 2015-12-12 at 4.10.04 AM.png
3.9 MB
Done
Undo Image Upload
Question
What color is the building that has a clock ?
Run DMN
VQA sample

What is the girl holding ?

tennis racket

What is the girl doing ?

playing tennis

Is the girl wearing a hat ?

yes

What is the girl wearing ?

shorts

What is the color of the ground ?

brown

What color is the ball ?

yellow

What color is her skirt ?

white

What did the girl just hit ?

tennis ball

Summary

- Most NLP tasks can be reduced to QA
- DMN accurately solves variety of QA tasks
- Next goals: One joint multitask DMN