Tackling the Limits of Deep Learning for NLP

Richard Socher

Salesforce Research

Caiming Xiong, Stephen Merity,
James Bradbury, Victor Zhong, Kazuma Hashimoto

and

Stanford: Hakan Inan, Khashayar Khosravi

The Limits of Single Task Learning

- Great performance improvements
- Projects start from random
- Single unsupervised task can't fix it
- How to express different tasks in the same framework, e.g.
  - sequence tagging
  - sentence-level classification
  - seq2seq?

Framework for Tackling NLP

A joint model for comprehensive QA

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: I think this model is incredible
Q: In French?
A: Je pense que ce modèle est incroyable.

I: 

Q: What color are the bananas?
A: Green.

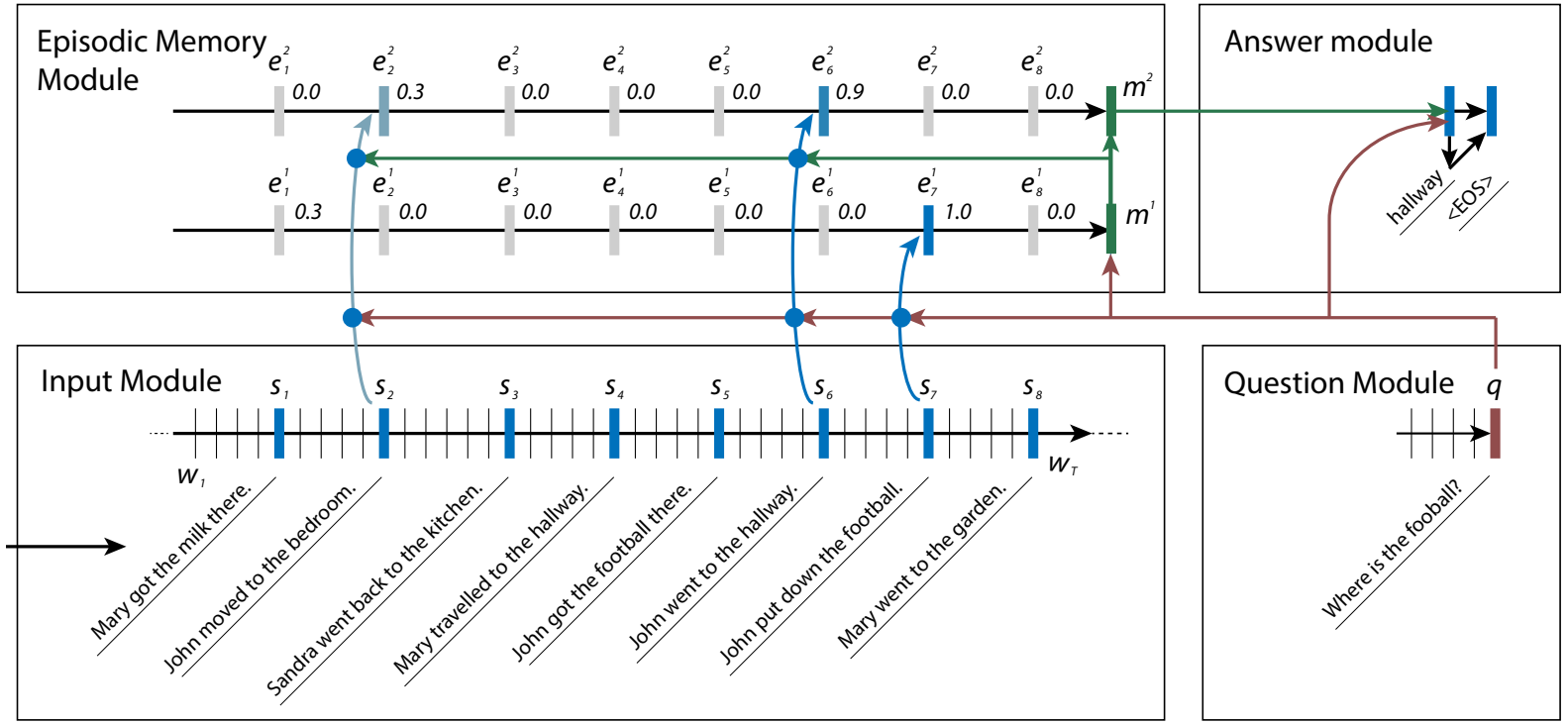
Move from $\{x_i, y_i\}$ to $\{x_i, q_i, y_i\}$

First of Six Major Obstacles

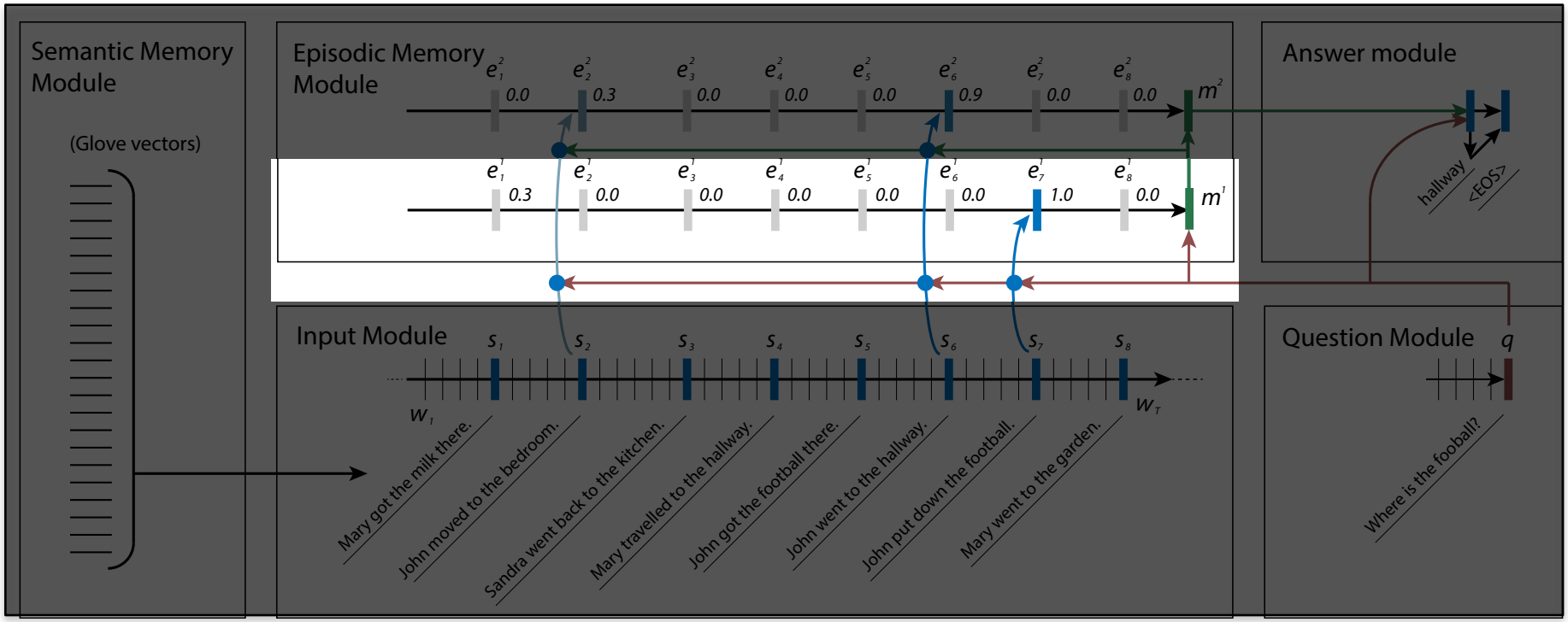
- For NLP no single model **architecture** with consistent state of the art results across tasks

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)

Tackling Obstacle 1: Dynamic Memory Network



The Modules: Episodic Memory



$$h_i^t = g_i^t GRU\left(s_i, h_{i-1}^t\right) + \left(1 - g_i^t\right) h_{i-1}^t$$

Last hidden state: m^t

The Modules: Episodic Memory

- Gates are activated if sentence relevant to the question or memory

$$z_i^t = [s_i \circ q \; ; s_i \circ m^{t-1}; |s_i - q| \; ; |s_i - m^{t-1}|]$$

$$Z_i^t = W^{(2)} \tanh\left(W^{(1)} z_i^t + b^{(1)}\right) + b^{(2)}$$

$$g_i^t = \frac{\exp(Z_i^t)}{\sum_{k=1}^{M_i} \exp(Z_k^t)}$$

- When the end of the input is reached, the relevant facts are summarized in another GRU

Related work

- 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)
  → Main difference: Sequence models for all functions in DMN, allowing for greater generality of tasks that be ”answered”

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

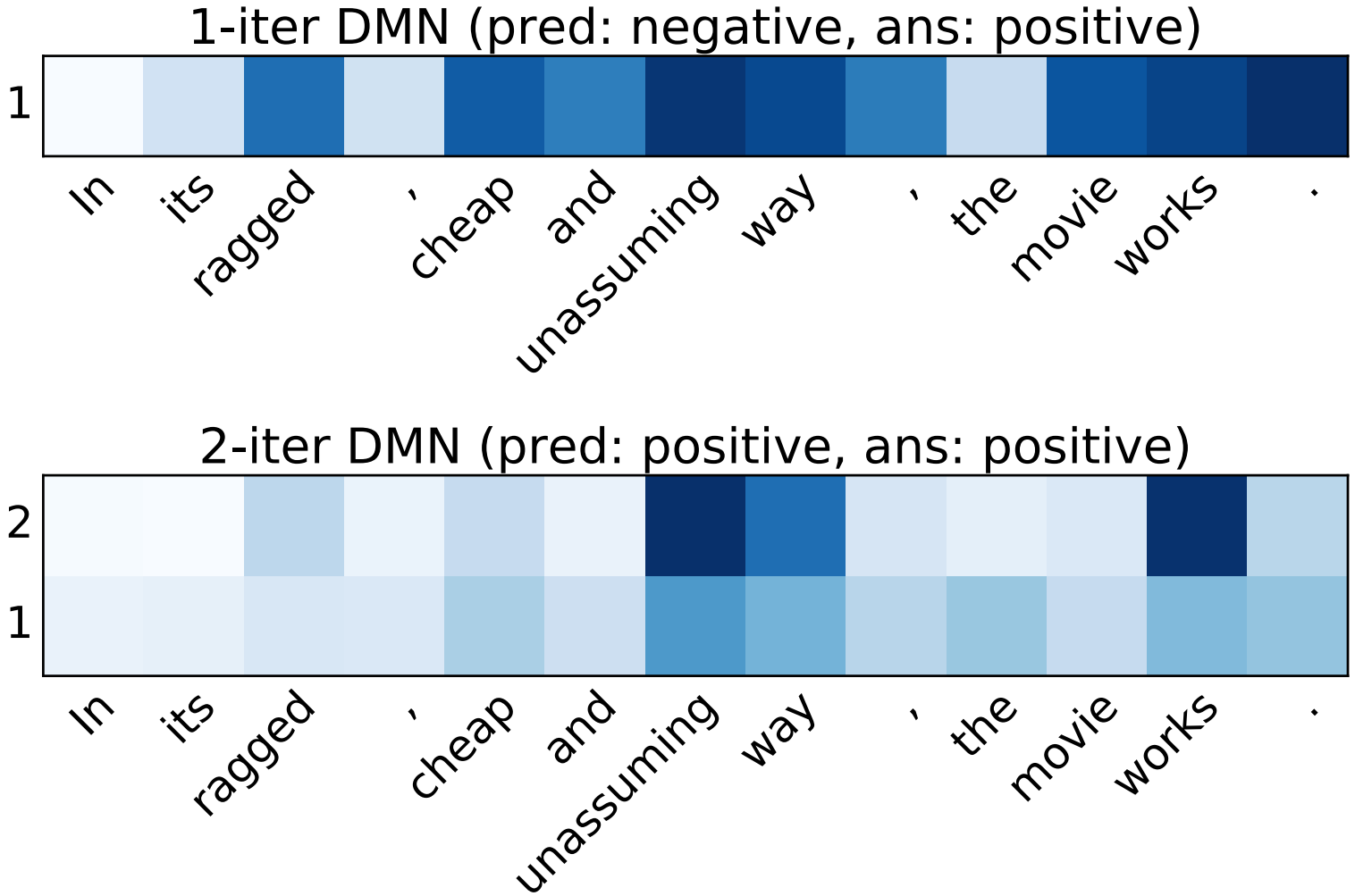
Analysis of Number of Episodes

- How many attention + memory passes are needed in the episodic memory?
- Results on Babi dataset and Stanford Sentiment

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

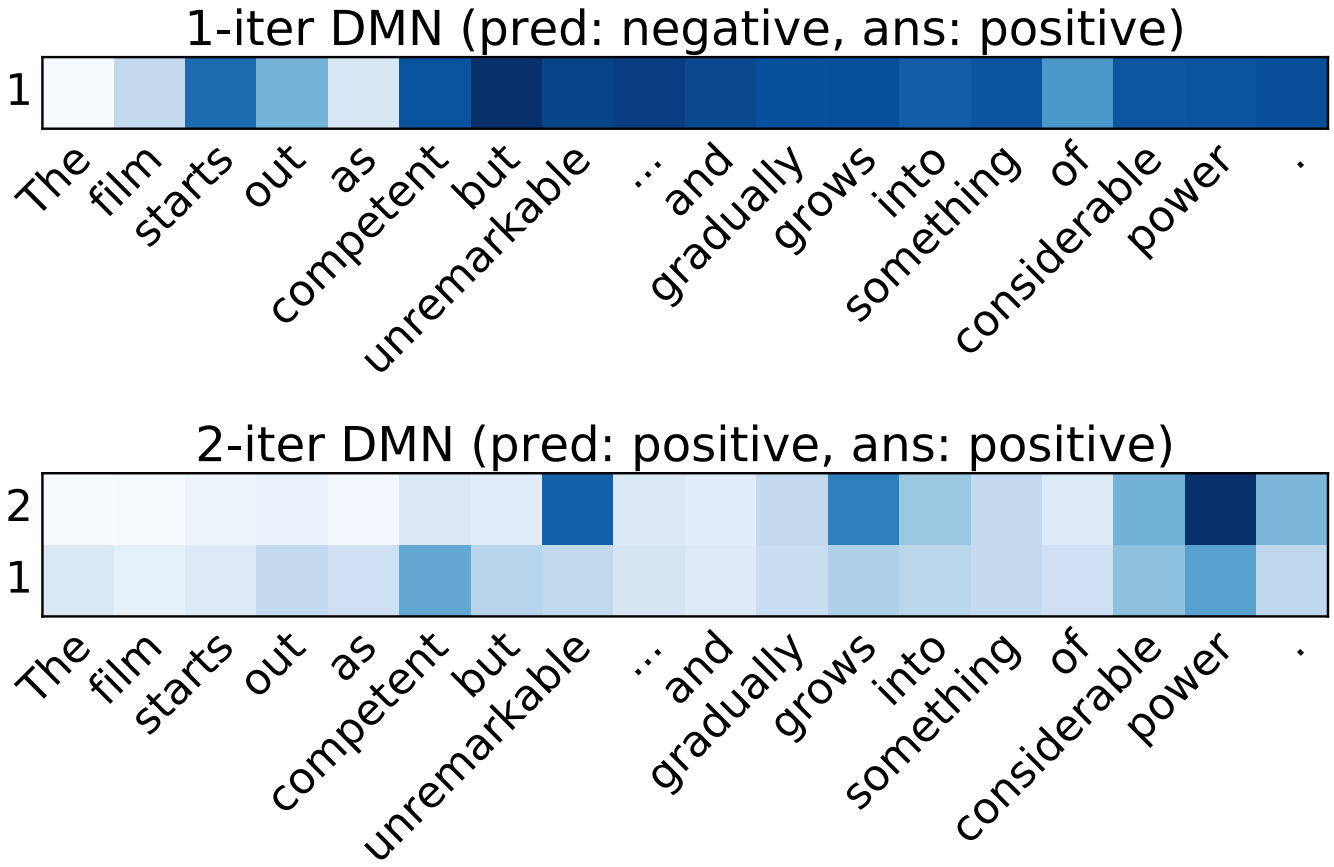
Analysis of Attention for Sentiment

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



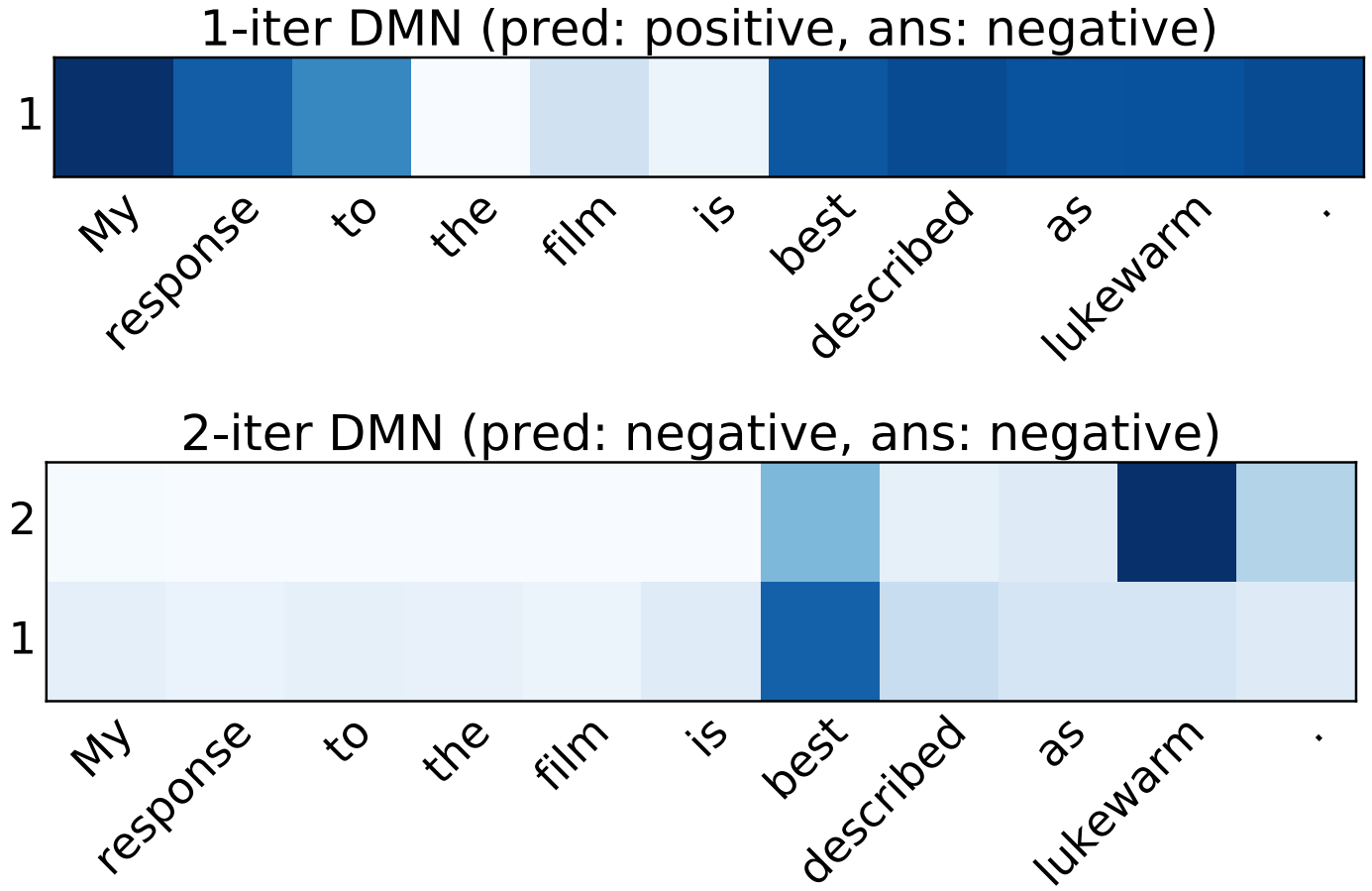
Analysis of Attention for Sentiment

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

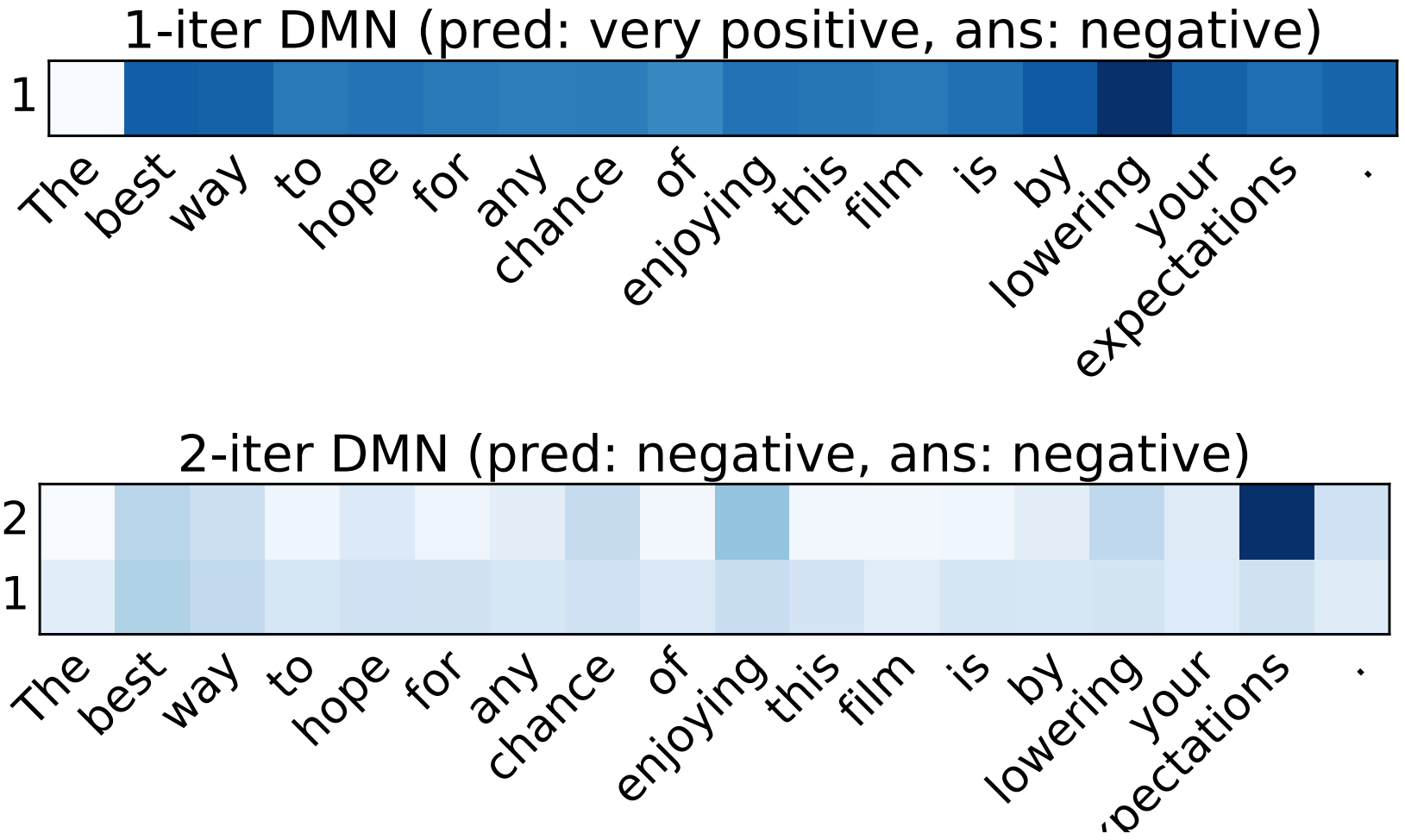


Analysis of Attention for Sentiment

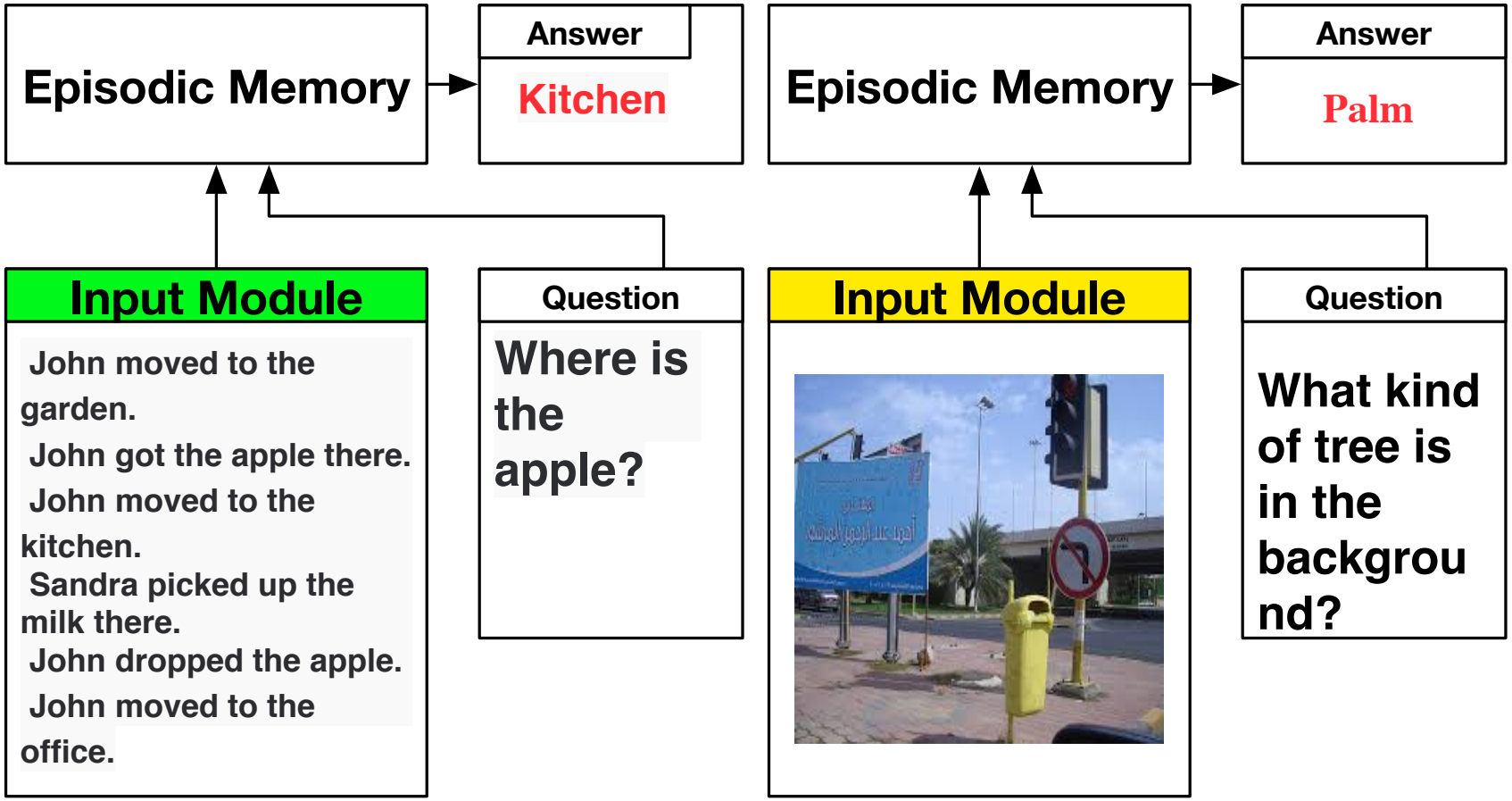
- Examples where full sentence context from first pass changes attention to words more relevant for final prediction



Analysis of Attention for Sentiment

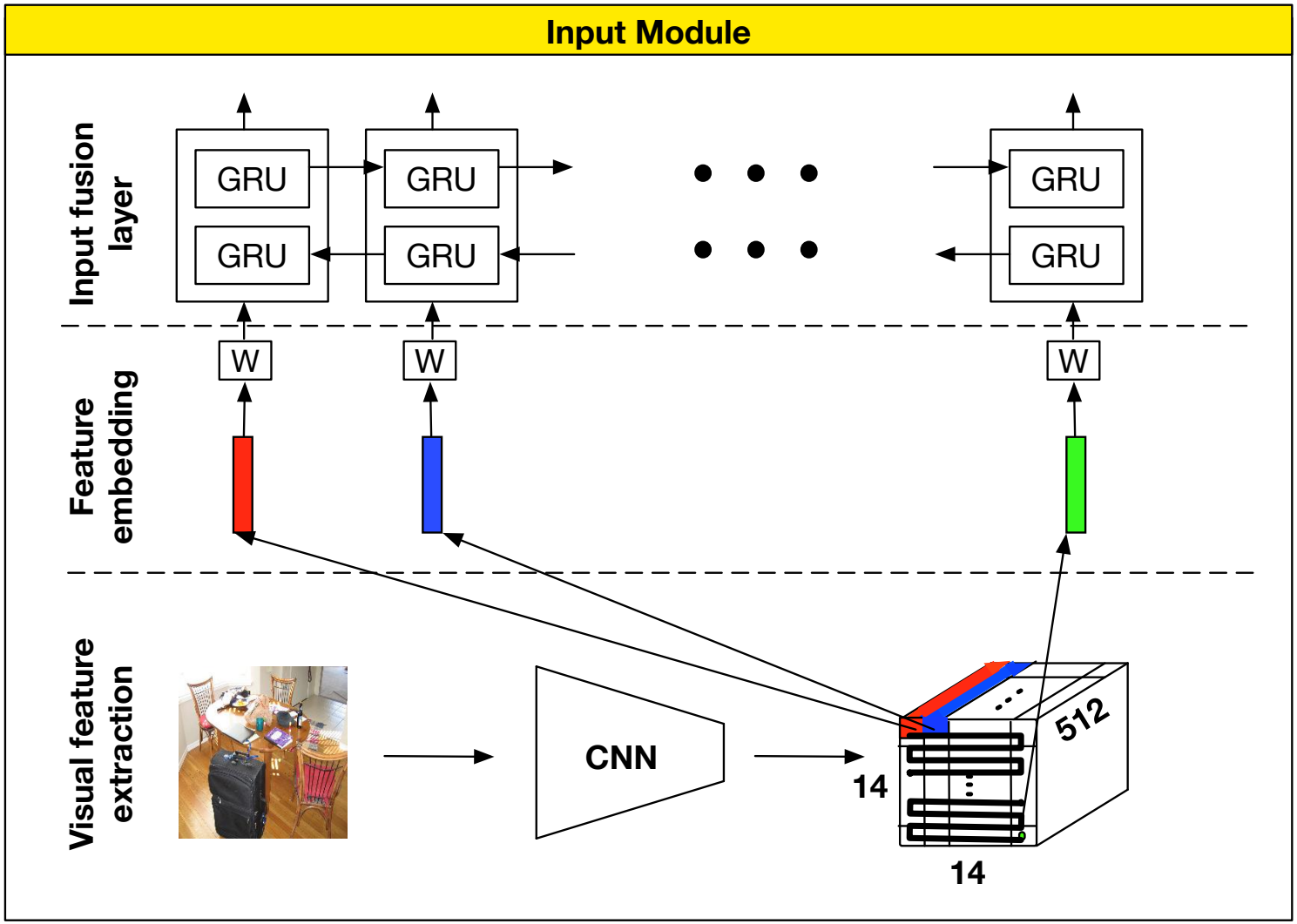


Modularization Allows for Different Inputs



Dynamic Memory Networks for Visual and Textual Question Answering, Caiming Xiong, Stephen Merity, Richard Socher

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 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

Attention Visualization


What is the main color on the bus ?

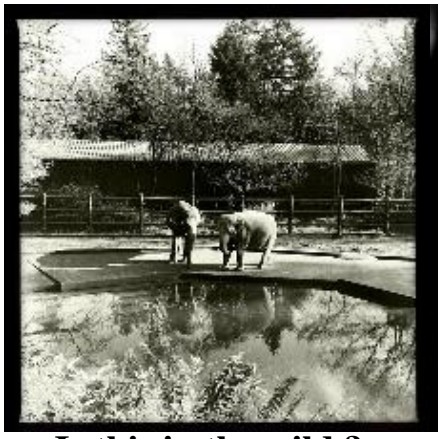
Answer: blue


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**

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

- DEMO

Obstacle 2: Joint Many-task Learning

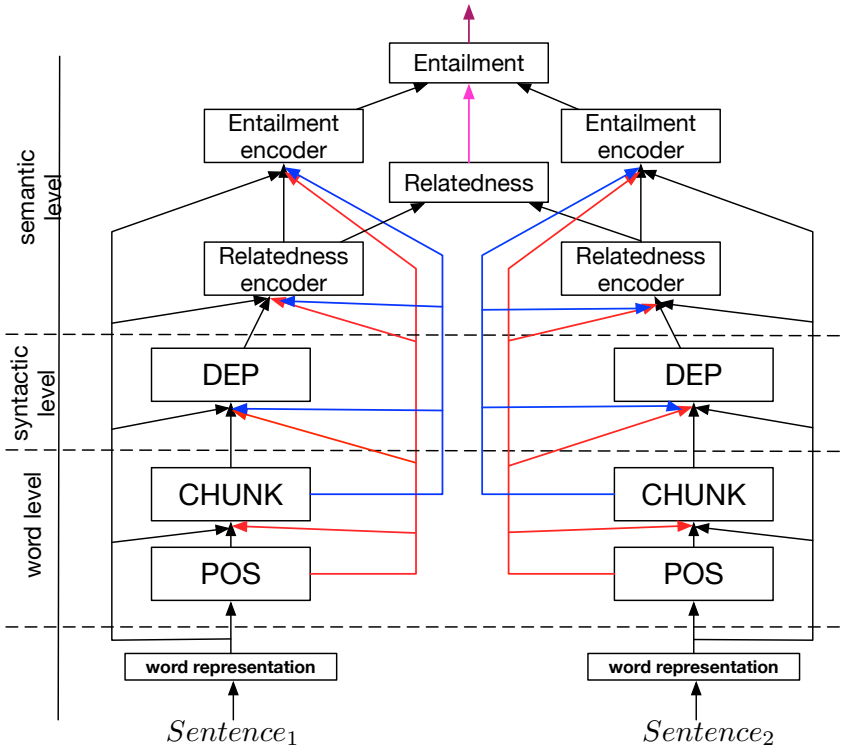
- Fully joint multitask learning* is hard:
  - Usually restricted to lower layers
  - Usually helps only if tasks are related
  - Often hurts performance if tasks are not related

* meaning: same decoder/classifier
and not only transfer learning with source target task pairs

Tackling Joint Training

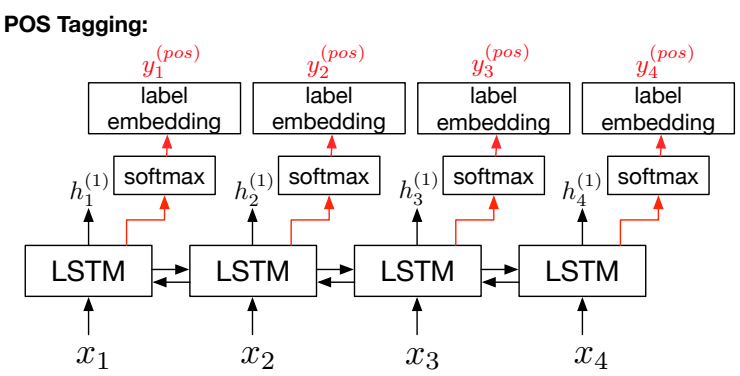
- A Joint Many-Task Model:
  Growing a Neural Network for Multiple NLP Tasks
  Kazuma Hashimoto,
  Caiming Xiong,
  Yoshimasa Tsuruoka &
  Richard Socher

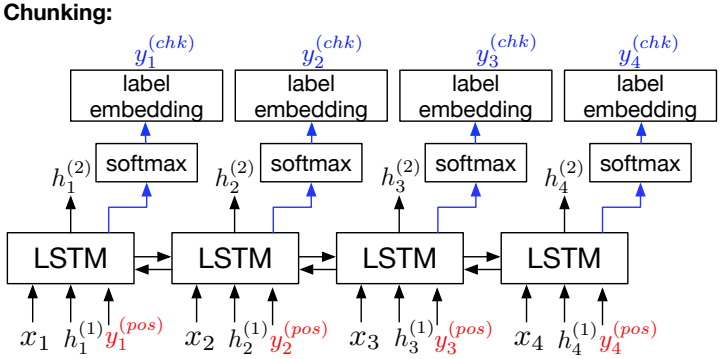
- Final Model →

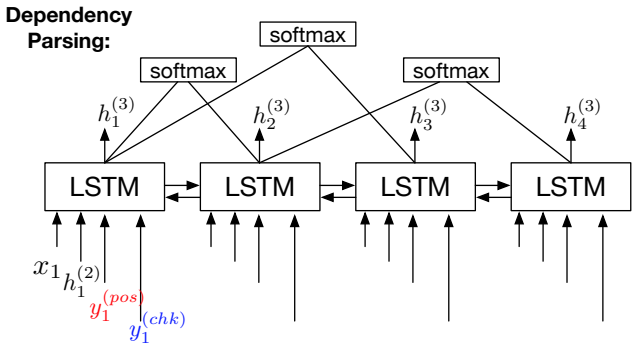


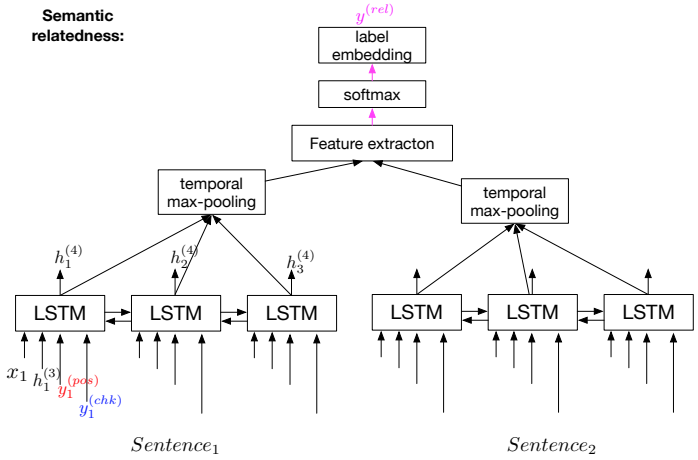
Model Details

- Include character n-grams and short-circuits
- State of the art purely feedforward parser









Training Details: Regularized Idea

Chunking training

$$-\sum_{s}\sum_{t}\log p(y_t^{(2)}=\alpha|h_t^{(2)})+\lambda\|W_{\text{chunk}}\|^2+\delta\|\theta_{\text{POS}}-\theta'_{\text{POS}}\|^2,$$

Entailment training

$$-\sum_{(s,s')}\log p(y_{(s,s')}^{(5)}=\alpha|h_s^{(5)},h_{s'}^{(5)})+\lambda\|W_{\text{ent}}\|^2+\delta\|\theta_{\text{rel}}-\theta'_{\text{rel}}\|^2,$$

New State of the Art on 4 of 5 Tasks

Method	Acc.
JMT_{all}	97.55
Ling et al. (2015)	**97.78**
Kumar et al. (2016)	97.56
Ma & Hovy (2016)	97.55
Søgaard (2011)	97.50
Collobert et al. (2011)	97.29
Tsuruoka et al. (2011)	97.28
Toutanova et al. (2003)	97.27

Table 2: POS tagging results.

Method	F1
JMT_{AB}	**95.77**
Søgaard & Goldberg (2016)	95.56
Suzuki & Isozaki (2008)	95.15
Collobert et al. (2011)	94.32
Kudo & Matsumoto (2001)	93.91
Tsuruoka et al. (2011)	93.81

Table 3: Chunking results.

Method	UAS	LAS
JMT_{all}	**94.67**	**92.90**
Single	93.35	91.42
Andor et al. (2016)	94.61	92.79
Alberti et al. (2015)	94.23	92.36
Weiss et al. (2015)	93.99	92.05
Dyer et al. (2015)	93.10	90.90
Bohnet (2010)	92.88	90.71

Table 4: Dependency results.

Method	MSE
JMT_{all}	**0.233**
JMT_{DE}	0.238
Zhou et al. (2016)	0.243
Tai et al. (2015)	0.253

Table 5: Semantic relatedness results.

Method	Acc.
JMT_{all}	86.2
JMT_{DE}	**86.8**
Yin et al. (2016)	86.2
Lai & Hockenmaier (2014)	84.6

Table 6: Textual entailment results.

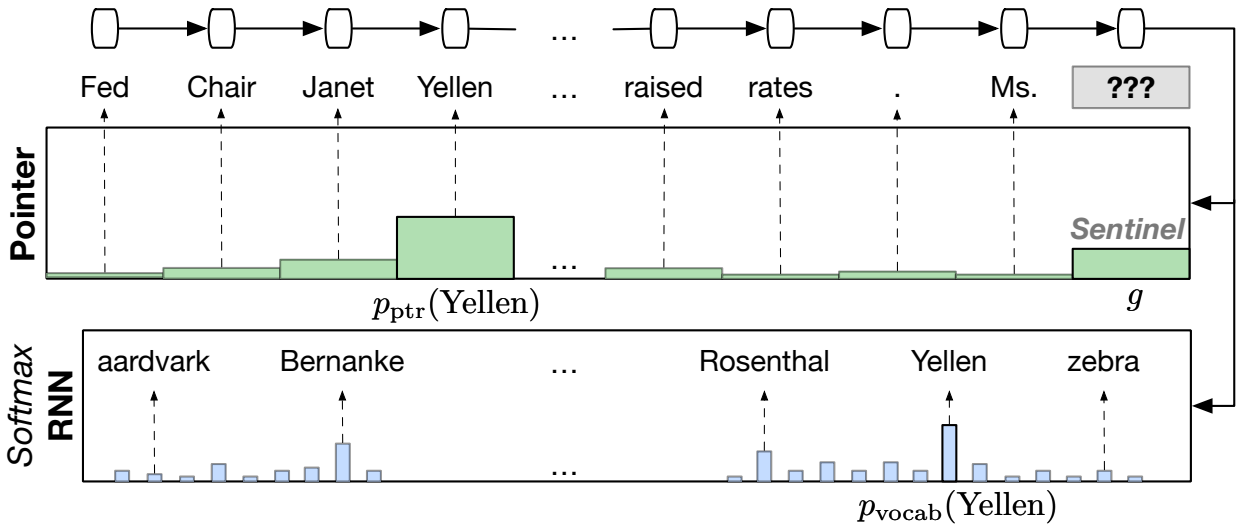
Obstacle 3: No Zero Shot Word Predictions

- Answers can only be predicted if they were seen during training and part of the softmax

- But it's natural to learn new words in an active conversation and systems should be able to pick them up

Tackling Obstacle by Predicting Unseen Words

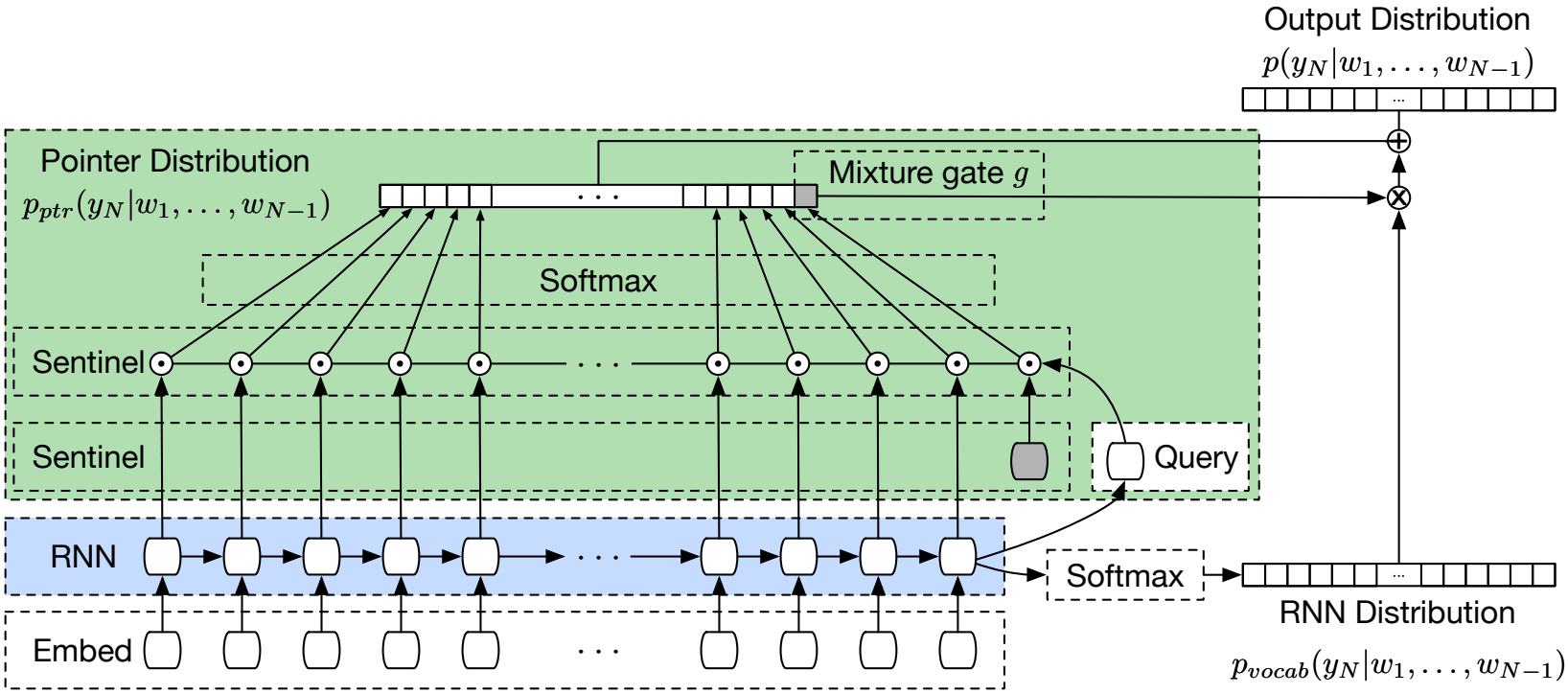
- Idea: Mixture Model of softmax and pointers:



$$p(\text{Yellen}) = g\, p_{\text{vocab}}(\text{Yellen}) + (1-g)\, p_{\text{ptr}}(\text{Yellen})$$

- Pointer Sentinel Mixture Models by Stephen Merity, Caiming Xiong, James Bradbury, Richard Socher

Pointer-Sentinel Model

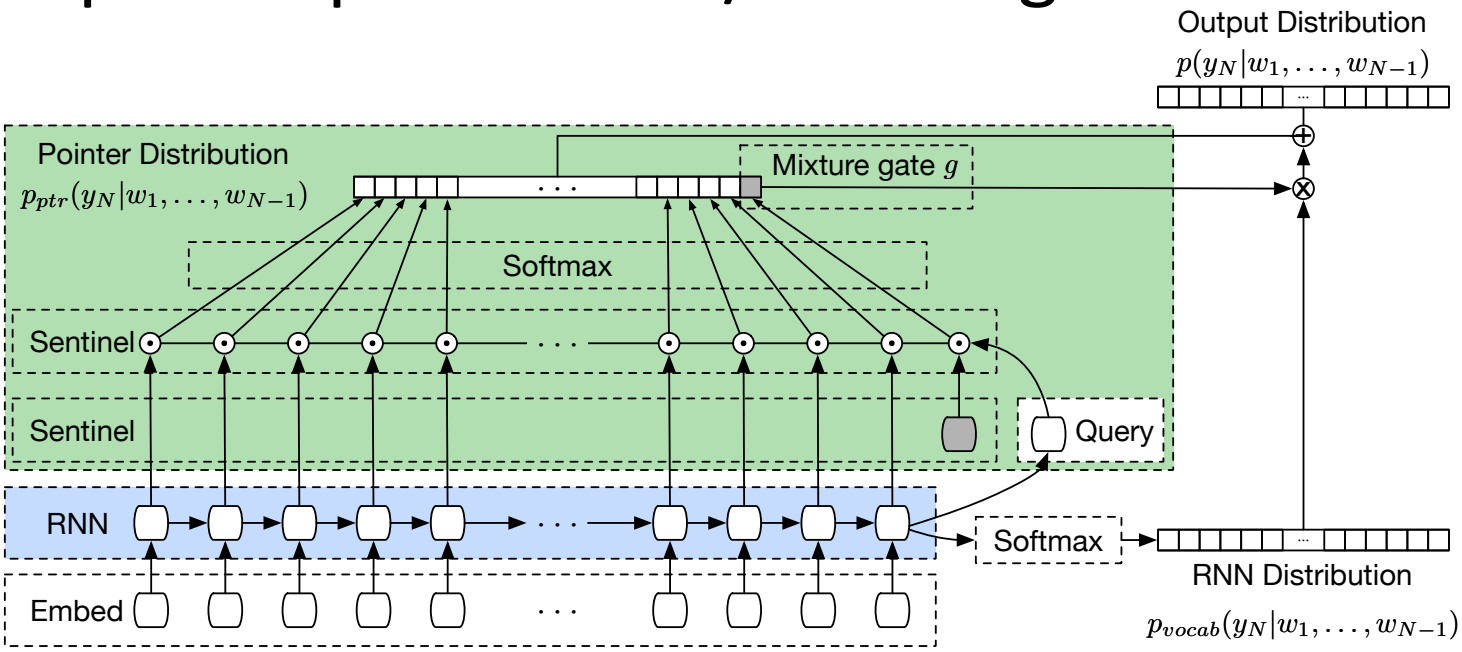


Pointer Sentinel for Language Modeling

Model	Parameters	Validation	Test
Mikolov & Zweig (2012) - KN-5	2M‡	–	141.2
Mikolov & Zweig (2012) - KN5 + cache	2M‡	–	125.7
Mikolov & Zweig (2012) - RNN	6M‡	–	124.7
Mikolov & Zweig (2012) - RNN-LDA	7M‡	–	113.7
Mikolov & Zweig (2012) - RNN-LDA + KN-5 + cache	9M‡	–	92.0
Pascanu et al. (2013a) - Deep RNN	6M	–	107.5
Cheng et al. (2014) - Sum-Prod Net	5M‡	–	100.0
Zaremba et al. (2014) - LSTM (medium)	20M	86.2	82.7
Zaremba et al. (2014) - LSTM (large)	66M	82.2	78.4
Gal (2015) - Variational LSTM (medium, untied)	20M	81.9 ± 0.2	79.7 ± 0.1
Gal (2015) - Variational LSTM (medium, untied, MC)	20M	–	78.6 ± 0.1
Gal (2015) - Variational LSTM (large, untied)	66M	77.9 ± 0.3	75.2 ± 0.2
Gal (2015) - Variational LSTM (large, untied, MC)	66M	–	73.4 ± 0.0
Kim et al. (2016) - CharCNN	19M	–	78.9
Zilly et al. (2016) - Variational RHN	32M	72.8	71.3
Zoneout + Variational LSTM (medium)	20M	84.4	80.6
Pointer Sentinel-LSTM (medium)	21M	72.4	**70.9**

Obstacle 4: Duplicate Word Representations

- Different encodings for encoder (Word2Vec and GloVe word vectors) and decoder (softmax classification weights for words)
- Duplicate parameters/meaning



Tackling Obstacle by Tying Word Vectors

- Simple but theoretically motivated idea: tie word vectors and train single weights jointly

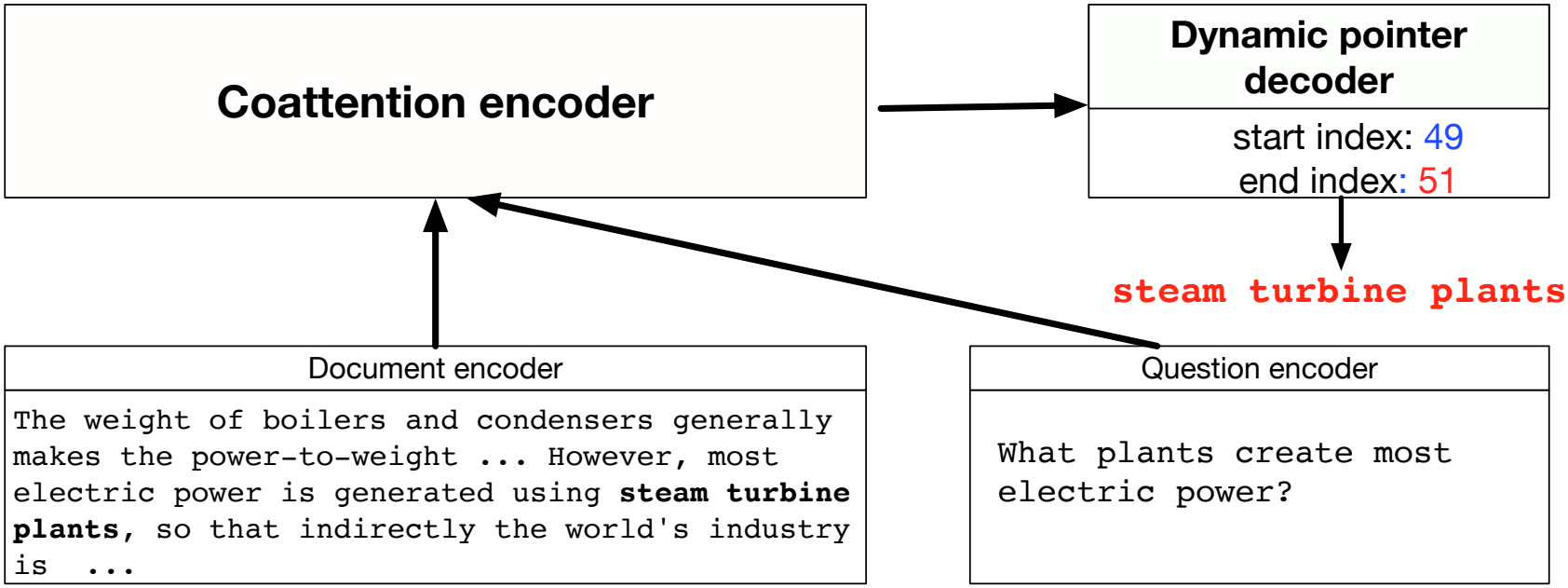
- Tying Word Vectors and Word Classifiers: A Loss Framework for Language Modeling, Hakan Inan, Khashayar Khosravi, Richard Socher

Language Modeling With Tying Word Vectors

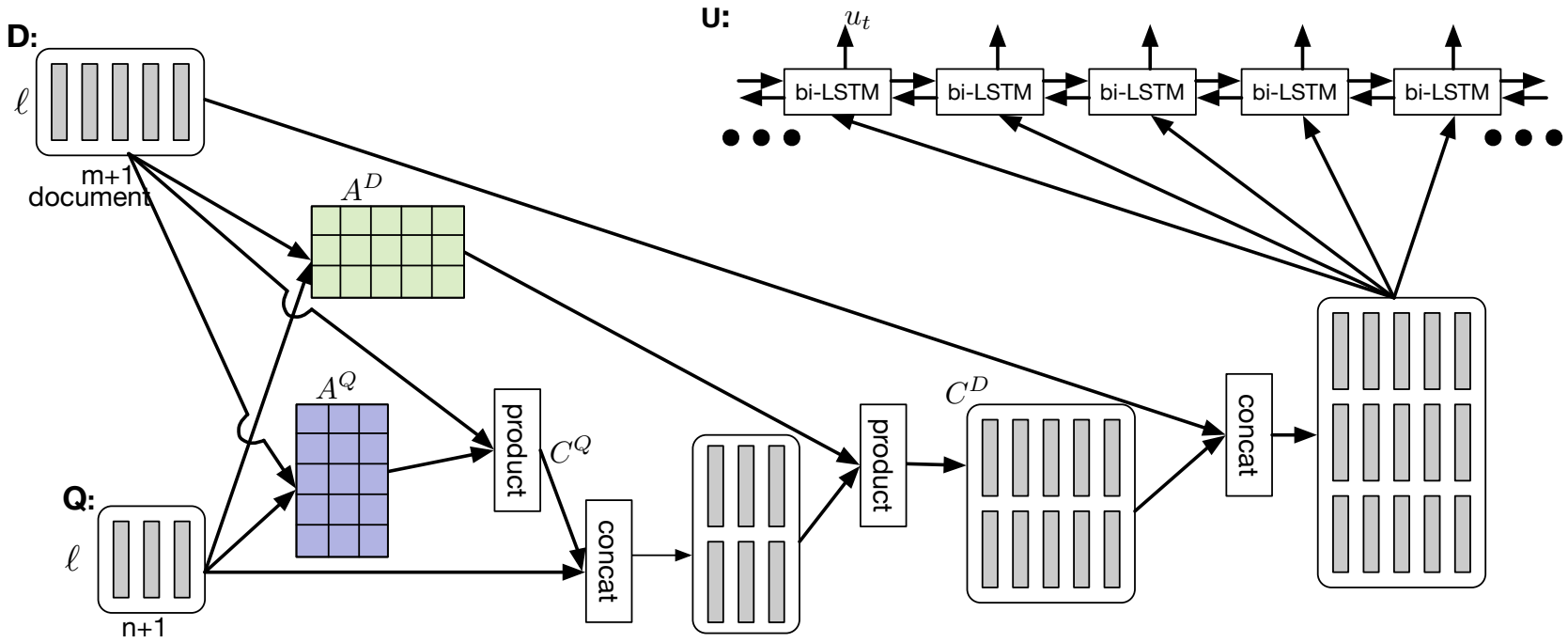
MODEL	PARAMETERS	VALIDATION	TEST
KN-5 (Mikolov & Zweig)	2M	-	141.2
KN-5 + Cache (Mikolov & Zweig)	2M	-	125.7
RNN (Mikolov & Zweig)	6M	-	124.7
RNN+LDA (Mikolov & Zweig)	7M	-	113.7
RNN+LDA+KN-5+Cache (Mikolov & Zweig)	9M	-	92.0
Deep RNN (Pascanu et al., 2013a)	6M	-	107.5
Sum-Prod Net (Cheng et al., 2014)	5M	-	100.0
LSTM (medium) (Zaremba et al., 2014)	20M	86.2	82.7
LSTM (large) (Zaremba et al., 2014)	66M	82.2	78.4
VD-LSTM (medium, untied) (Gal, 2015)	20M	81.9 ± 0.2	79.7 ± 0.1
VD-LSTM (medium, untied, MC) (Gal, 2015)	20M	-	78.6 ± 0.1
VD-LSTM (large, untied) (Gal, 2015)	66M	77.9 ± 0.3	75.2 ± 0.2
VD-LSTM (large, untied, MC) (Gal, 2015)	66M	-	73.4 ± 0.0
CharCNN (Kim et al., 2015)	19M	-	78.9
VD-RHN (Zilly et al., 2016)	32M	72.8	71.3
Pointer Sentinel-LSTM(medium) (Merity et al., 2016)	21M	72.4	70.9
38 Large LSTMs (Zaremba et al., 2014)	2.51B	71.9	68.7
10 Large VD-LSTMs (Gal, 2015)	660M	-	68.7
VD-LSTM +REAL (medium)	14M	75.7	73.2
VD-LSTM +REAL (large)	51M	**71.1**	**68.5**

Obstacle 5: Questions have input independent representations

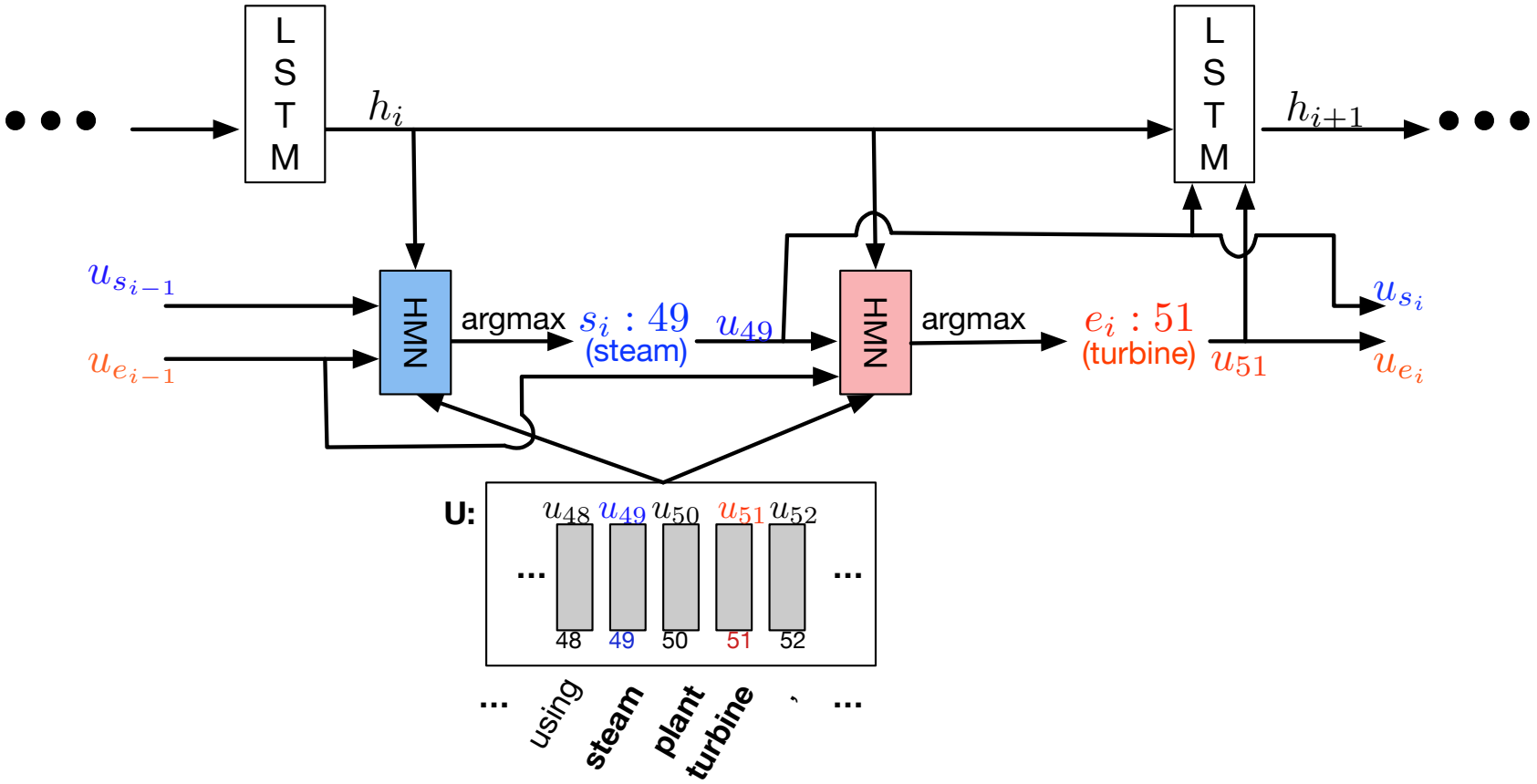
- Interdependence needed for a comprehensive QA model
- Dynamic Coattention Networks for Question Answering by Caiming Xiong, Victor Zhong, Richard Socher



Coattention Encoder



Dynamic Decoder



Stanford Question Answering Dataset

Packet_switching

The Stanford Question Answering Dataset

Starting in the late 1950s, American computer scientist Paul Baran developed the concept Distributed Adaptive Message Block Switching with the goal to provide a fault-tolerant, efficient routing method for telecommunication messages as part of a research program at the RAND Corporation, funded by the US Department of Defense. This concept contrasted and contradicted the theretofore established principles of pre-allocation of network bandwidth, largely fortified by the development of telecommunications in the Bell System. The new concept found little resonance among network implementers until the independent work of Donald Davies at the National Physical Laboratory (United Kingdom) (NPL) in the late 1960s. Davies is credited with coining the modern name packet switching and inspiring numerous packet switching networks in Europe in the decade following, including the incorporation of the concept in the early ARPANET in the United States.

What did this concept contradict

Ground Truth Answers: This concept contrasted and contradicted the theretofore established principles of pre-allocation of network bandwidth | theretofore established principles of pre-allocation of network bandwidth | principles of pre-allocation of network bandwidth

What is Donald Davies credited with

Ground Truth Answers: Davies is credited with coining the modern name packet switching and inspiring numerous packet switching networks in Europe | coining the modern name packet switching and inspiring numerous packet switching networks | coining the modern name packet switching

What did Paul Baran develop in the late 1950's

Ground Truth Answers: the concept Distributed Adaptive Message Block Switching | the concept Distributed Adaptive Message Block

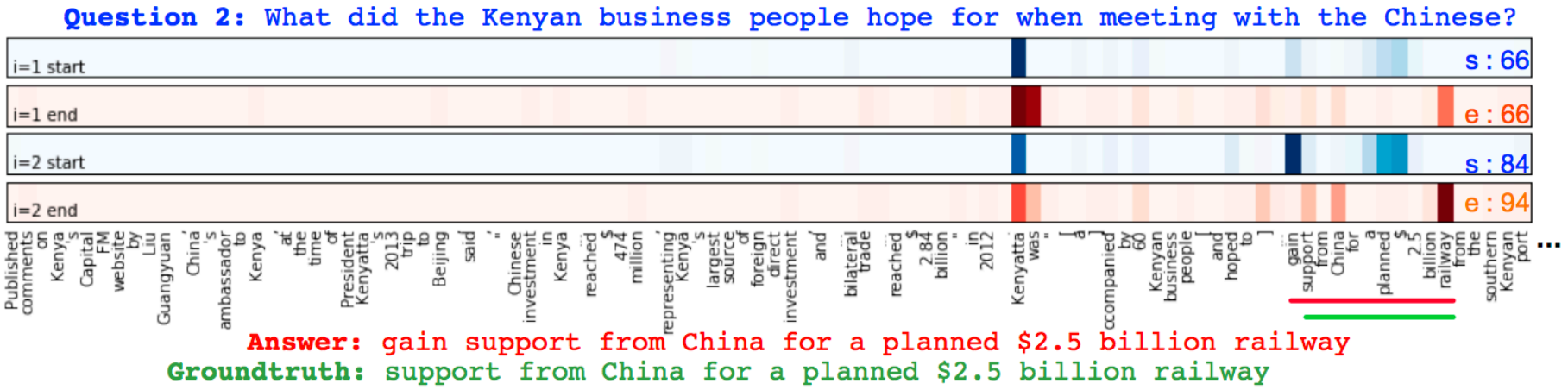
Results on SQUAD Competition

Model	Dev EM	Dev F1	Test EM	Test F1
Ensemble				
DCN (Ours)	**70.3**	**79.4**	**71.2**	**80.4**
Microsoft Research Asia *	–	–	69.4	78.3
Allen Institute *	69.2	77.8	69.9	78.1
Singapore Management University *	67.6	76.8	67.9	77.0
Google NYC *	68.2	76.7	–	–
Single model				
DCN (Ours)	65.4	**75.6**	**66.2**	**75.9**
Microsoft Research Asia *	65.9	75.2	65.5	75.0
Google NYC *	**66.4**	74.9	–	–
Singapore Management University *	–	–	64.7	73.7
Carnegie Mellon University *	–	–	62.5	73.3
Dynamic Chunk Reader (Yu et al., 2016)	62.5	71.2	62.5	71.0
Match-LSTM (Wang & Jiang, 2016)	59.1	70.0	59.5	70.3
Baseline (Rajpurkar et al., 2016)	40.0	51.0	40.4	51.0
Human (Rajpurkar et al., 2016)	81.4	91.0	82.3	91.2

Results are at time of ICLR submission

See https://rajpurkar.github.io/SQuAD-explorer/ for latest results

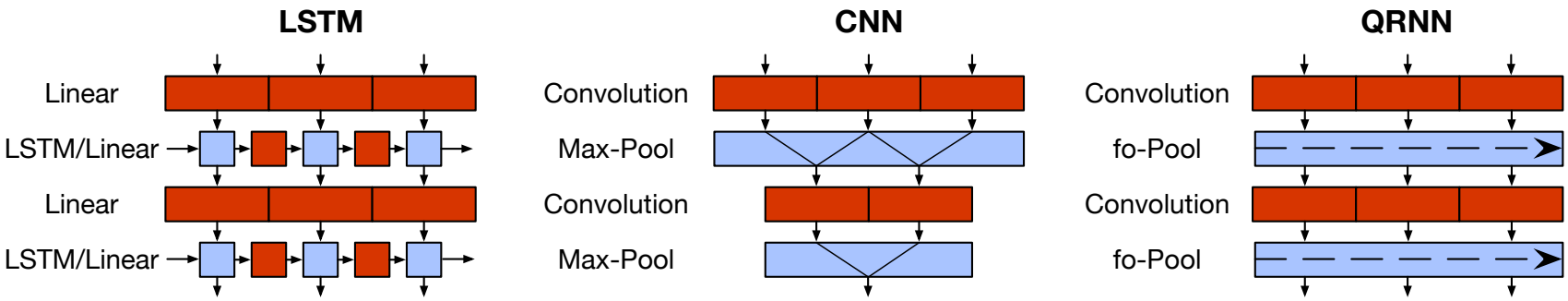
Dynamic Decoder Visualization



Obstacle 6: RNNs are Slow

- RNNs are the basic building block for deepNLP
- Idea: Take the best and parallelizable parts of RNNs and CNNs
- Quasi-Recurrent Neural Networks by James Bradbury, Stephen Merity, Caiming Xiong & Richard Socher

Quasi-Recurrent Neural Network



- Convolutions for parallelism across time:

$$\mathbf{z}_t = \tanh(\mathbf{W}_z^1 \mathbf{x}_{t-1} + \mathbf{W}_z^2 \mathbf{x}_t)$$
$$\mathbf{f}_t = \sigma(\mathbf{W}_f^1 \mathbf{x}_{t-1} + \mathbf{W}_f^2 \mathbf{x}_t)$$
$$\mathbf{o}_t = \sigma(\mathbf{W}_o^1 \mathbf{x}_{t-1} + \mathbf{W}_o^2 \mathbf{x}_t).$$

→

$$\mathbf{Z} = \tanh(\mathbf{W}_z * \mathbf{X})$$
$$\mathbf{F} = \sigma(\mathbf{W}_f * \mathbf{X})$$
$$\mathbf{O} = \sigma(\mathbf{W}_o * \mathbf{X}),$$

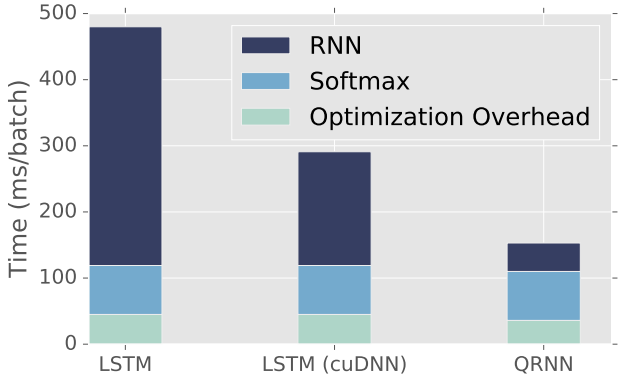
- Element-wise gated recurrence for parallelism across channels: $\mathbf{h}_t = \mathbf{f}_t \odot \mathbf{h}_{t-1} + (1 - \mathbf{f}_t) \odot \mathbf{z}_t,$

Q-RNNs for Language Modeling

- Better

Model	Parameters	Validation	Test
LSTM (medium) (Zaremba et al., 2014)	20M	86.2	82.7
Variational LSTM (medium) (Gal & Ghahramani, 2016)	20M	81.9	79.7
LSTM with CharCNN embeddings (Kim et al., 2016)	19M	–	78.9
Zoneout + Variational LSTM (medium) (Merity et al., 2016)	20M	84.4	80.6
Our models			
LSTM (medium)	20M	85.7	82.0
QRNN (medium)	18M	82.9	79.9
QRNN + zoneout ($p = 0.1$) (medium)	18M	82.1	78.3

- Faster



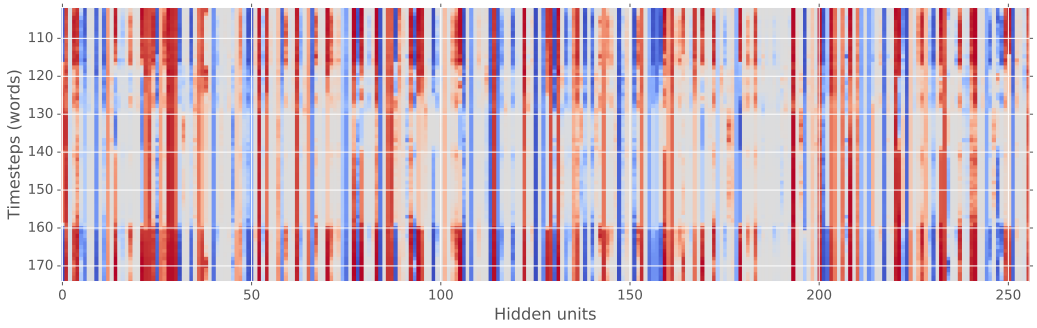
		Sequence length				
		32	64	128	256	512
Batch size	8	**5.5x**	**8.8x**	**11.0x**	**12.4x**	**16.9x**
	16	**5.5x**	**6.7x**	**7.8x**	**8.3x**	**10.8x**
	32	**4.2x**	**4.5x**	**4.9x**	**4.9x**	**6.4x**
	64	**3.0x**	**3.0x**	**3.0x**	**3.0x**	**3.7x**
	128	**2.1x**	**1.9x**	**2.0x**	**2.0x**	**2.4x**
	256	**1.4x**	**1.4x**	**1.3x**	**1.3x**	**1.3x**

Q-RNNs for Sentiment Analysis

- Better and faster than LSTMs

Model	Time / Epoch (s)	Test Acc (%)
BSVM-bi (Wang & Manning, 2012)	–	91.2
2 layer sequential BoW CNN (Johnson & Zhang, 2014)	–	92.3
Ensemble of RNNs and NB-SVM (Mesnil et al., 2014)	–	92.6
2-layer LSTM (Longpre et al., 2016)	–	87.6
Residual 2-layer bi-LSTM (Longpre et al., 2016)	–	90.1
Our models		
Deeply connected 4-layer LSTM (cuDNN optimized)	480	90.9
Deeply connected 4-layer QRNN	150	91.4
D.C. 4-layer QRNN with $k = 4$	160	91.1

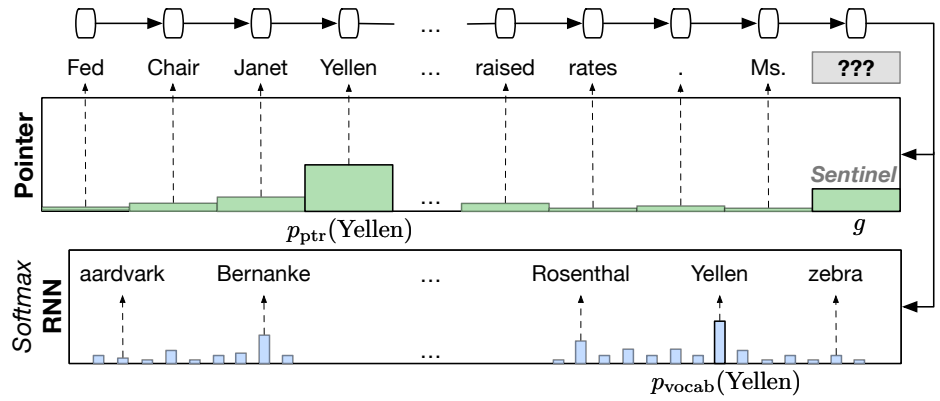
- More interpretable



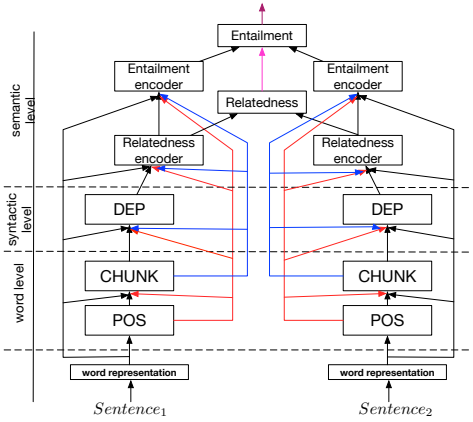
- Example:
- Initial positive review
- *Review starts out positive*
  At 117: *"not exactly a bad story"*
  At 158: *"I recommend this movie to everyone, even if you've never played the game"*

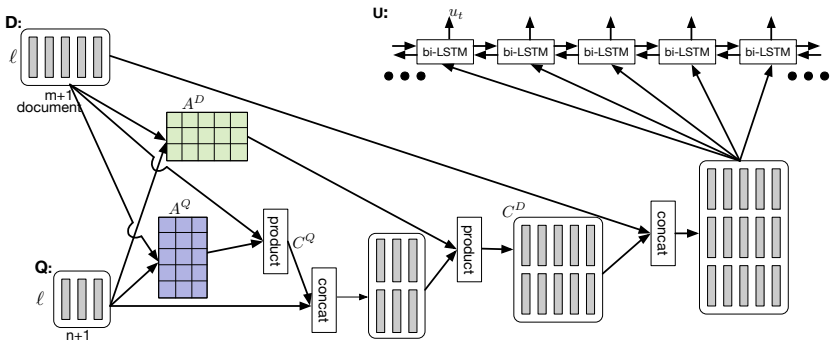
Comprehensive Question Answering

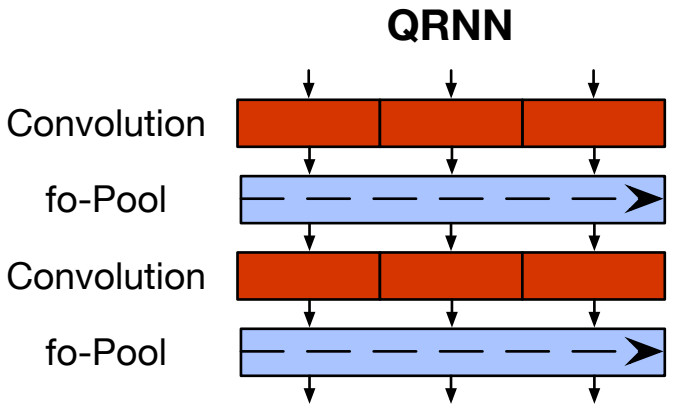
- Framework for tackling the limits of deepNLP



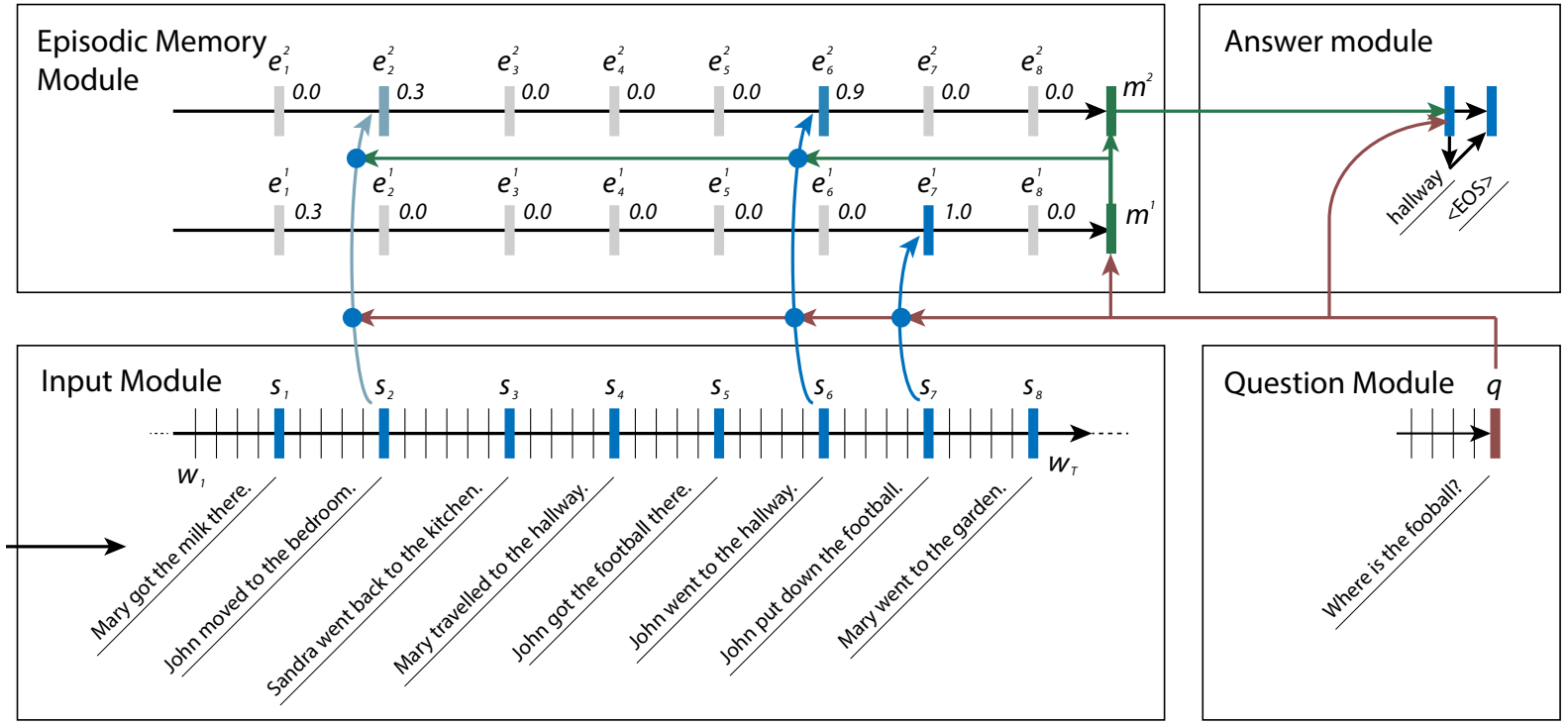
$$p(\text{Yellen}) = g\, p_{\text{vocab}}(\text{Yellen}) + (1-g)\, p_{\text{ptr}}(\text{Yellen})$$



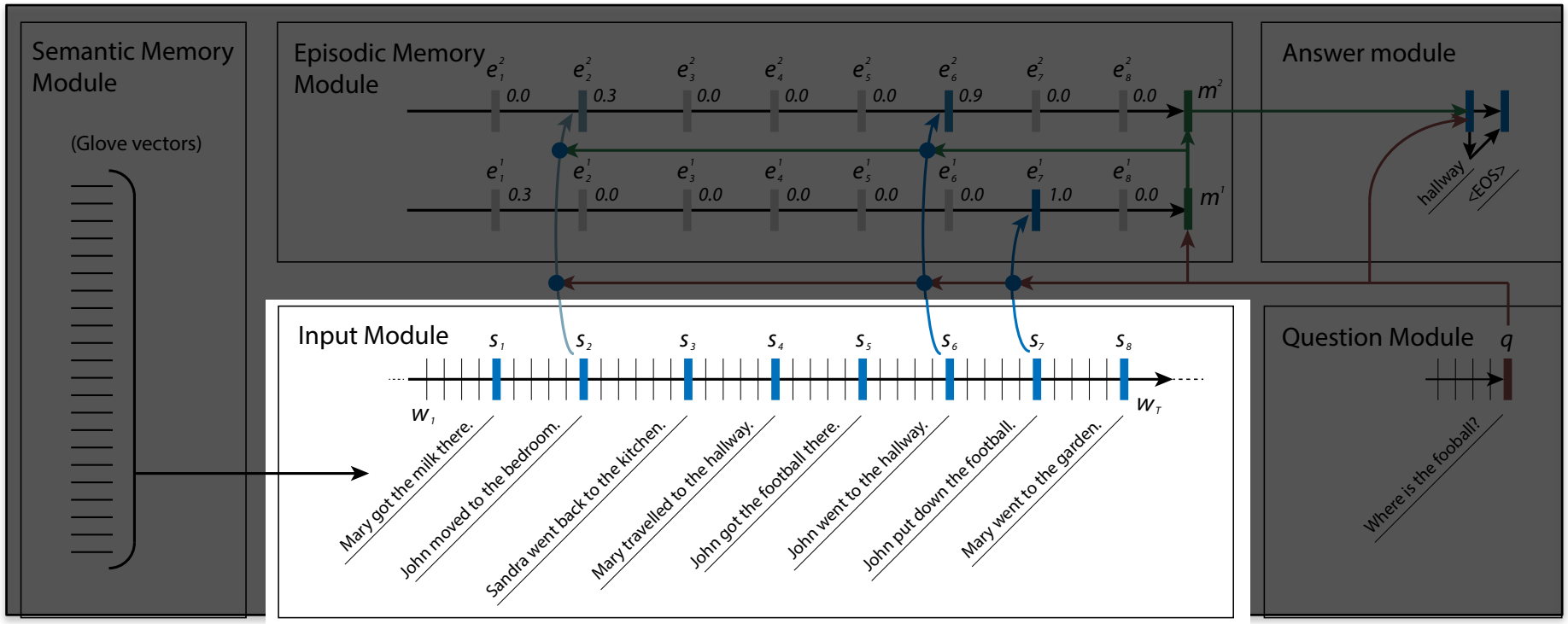




Tackling Obstacle 1: Dynamic Memory Network

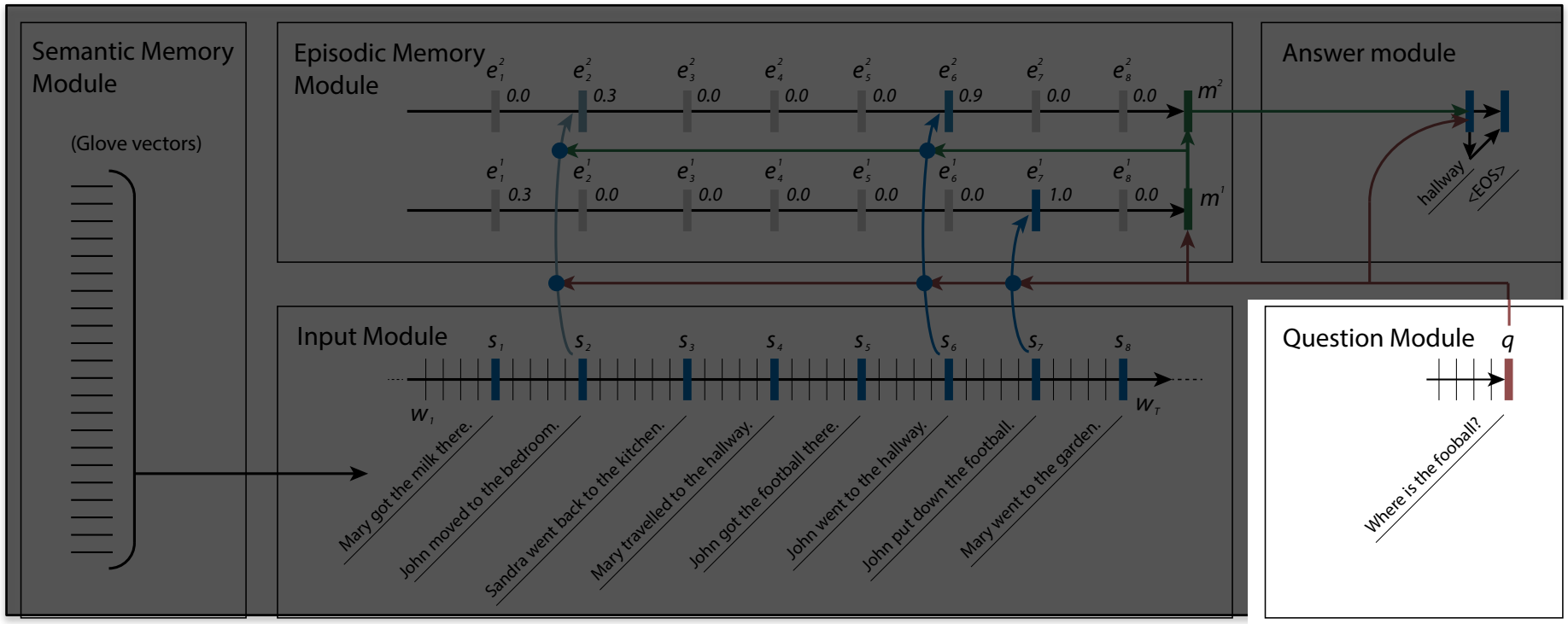


The Modules: Input



Standard GRU. The last hidden state of each sentence is accessible.

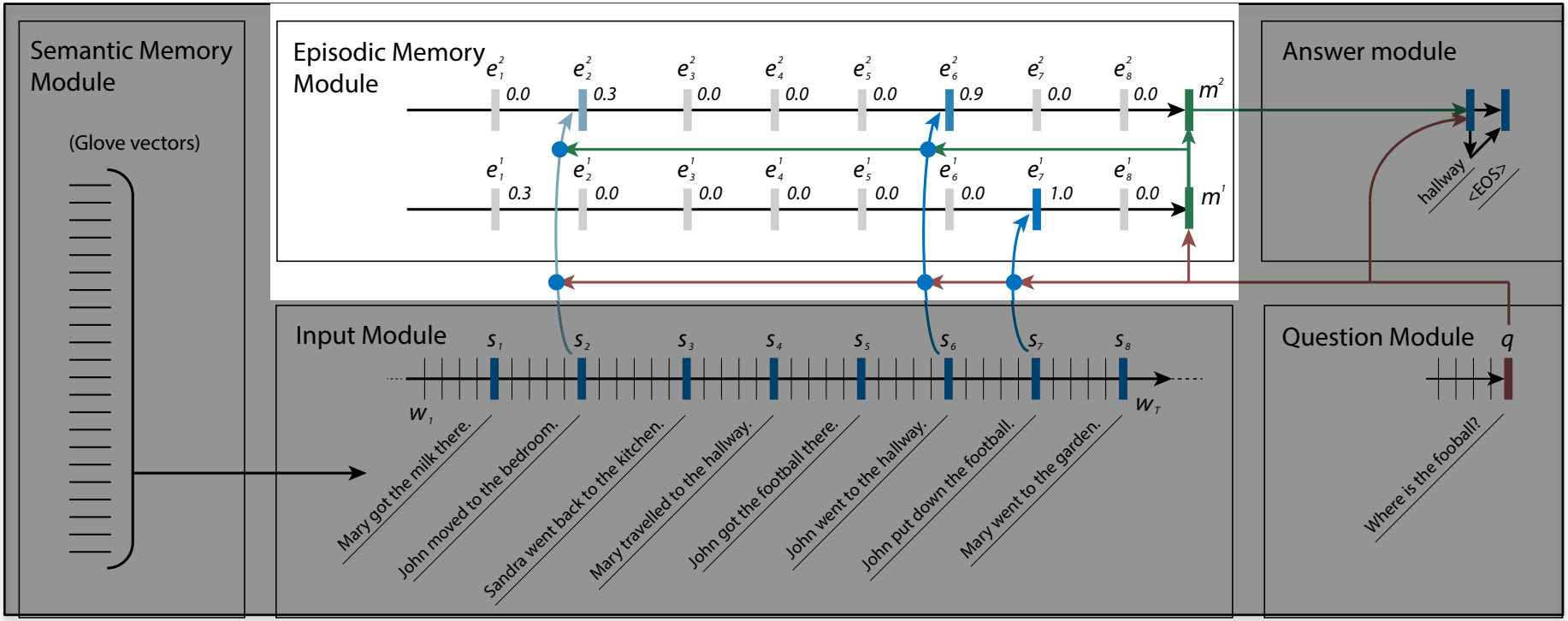
The Modules: Question



$$q_t = GRU(v_t, q_{t-1})$$

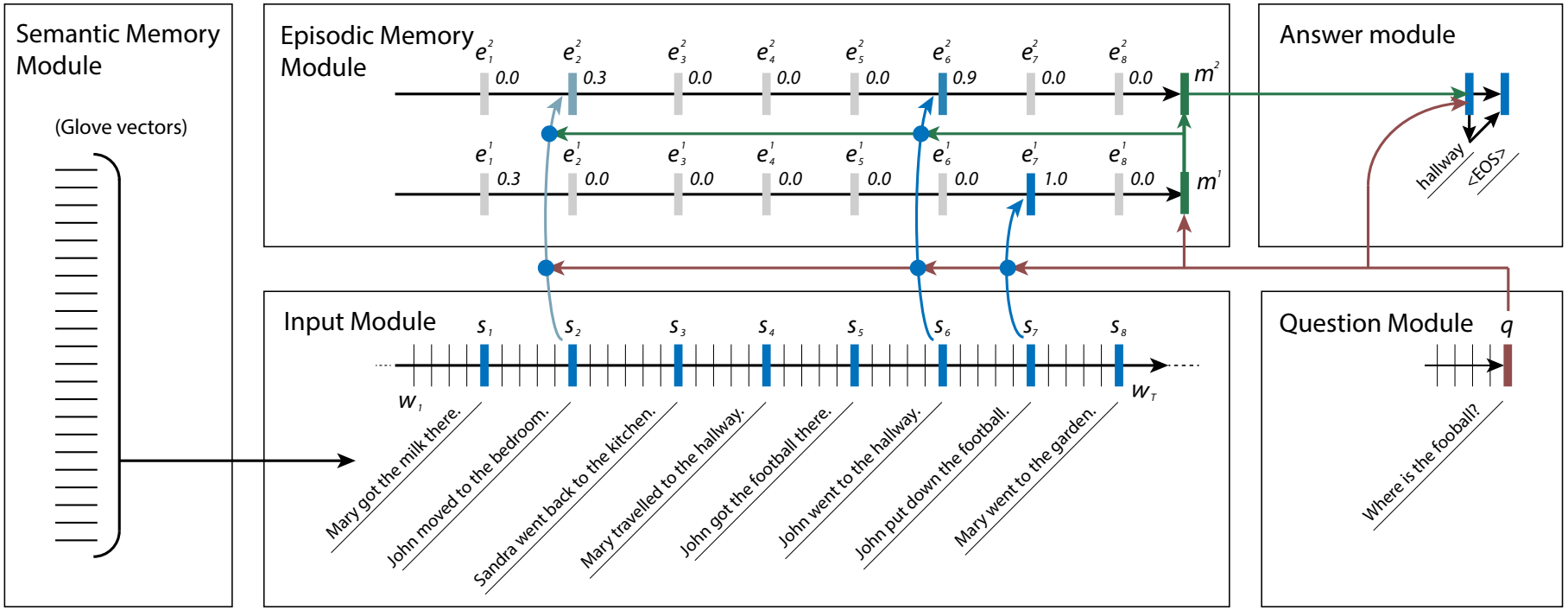
The Modules: Episodic Memory

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



The Modules: Answer

$$a_t = \texttt{GRU}([y_{t-1}, q], a_{t-1}), \qquad y_t = softmax(W^{(a)} a_t)$$



babI 1k, with gate supervision

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**

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**

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**