

Deep NLP Applications and Dynamic Memory Networks



Richard Socher
richard@metamind.io

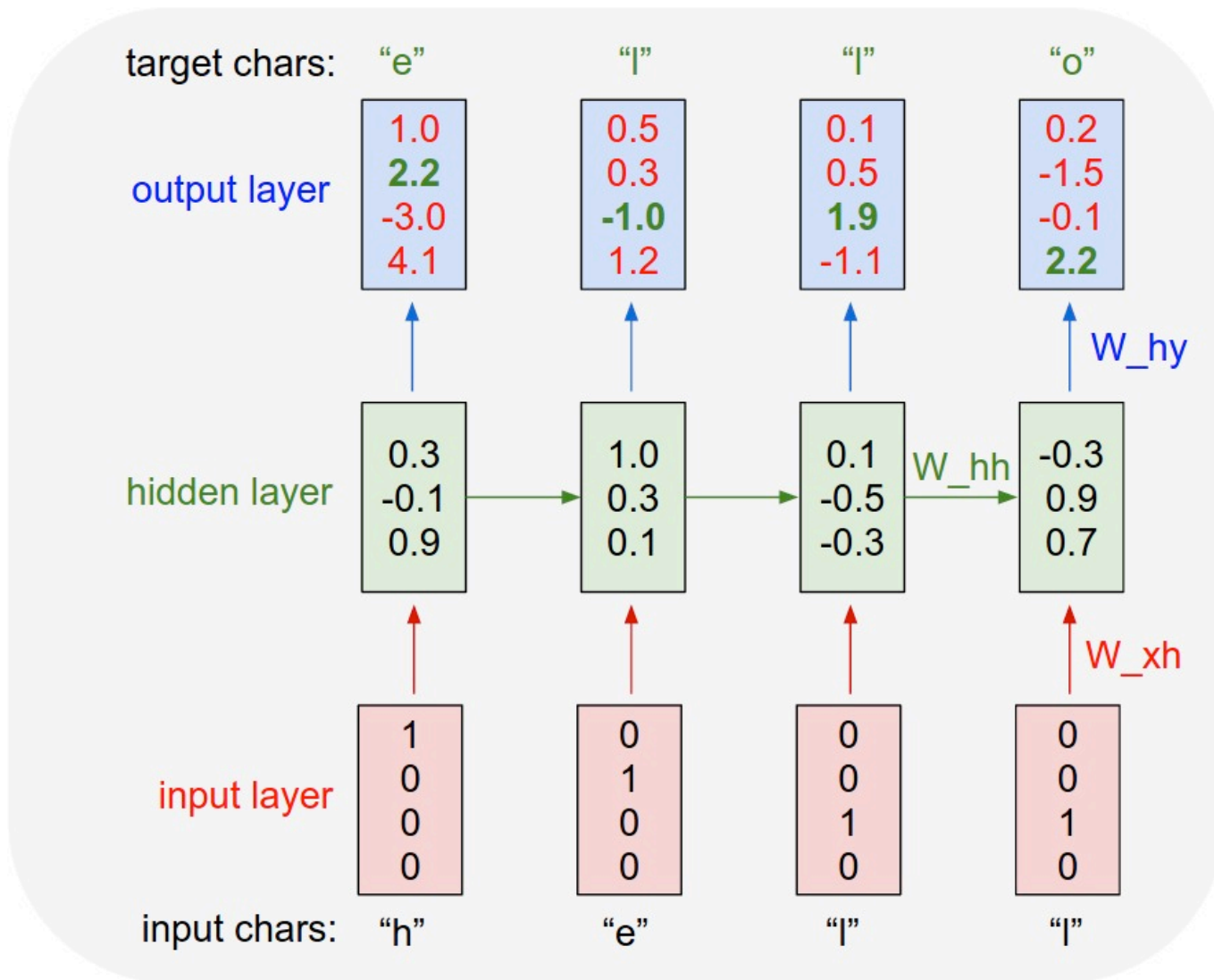
Why focus deep research on NLP?

- Image classification increasingly commoditized
 - Vision is more than classification but it's central
- Demo: <https://www.metamind.io/vision/train>

Overview

- Fun deep NLP applications:
 - Character RNNs on text and code
 - Image – Sentence mapping
 - Engagement
 - Question Answering
- Ask me Anything: **Dynamic Memory Networks** for NLP

Character RNNs on text and code



<http://karpathy.github.io/2015/05/21/rnn-effectiveness/>

Character RNNs on text and code

- Haven't yet produced useful results on real datasets
- Shows that RNNs can memorize sequences and keep memory (mostly LSTMs)
- Most interesting results simply train on dataset and sample from it afterwards (first shown by Sutskever et al. 2011: Generating Text with Recurrent Neural Networks)
- Results from an LSTM (karpathy.github.io) →

Shakespeare

PANDARUS:

Alas, I think he shall be come approached and the day
When little strain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:

They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENTIO:

Well, your wit is in the care of side and that.

Wikipedia

Naturalism and decision for the majority of Arab countries' capitalide was grounded by the Irish language by [[John Clair]], [[An Imperial Japanese Revolt]], associated with Guangzham's sovereignty. His generals were the powerful ruler of the Portugal in the [[Protestant Immineners]], which could be said to be directly in Cantonese Communication, which followed a ceremony and set inspired prison, training. The emperor travelled back to [[Antioch, Perth, October 25 | 21]] to note, the Kingdom of Costa Rica, unsuccessful fashioned the [[Thrales]], [[Cynth's Dajoard]], known in western [[Scotland]], near Italy to the conquest of India with the conflict.

Latex (had to be fixed manually)

Proof. Omitted. □

Lemma 0.1. *Let \mathcal{C} be a set of the construction.*

Let \mathcal{C} be a gerber covering. Let \mathcal{F} be a quasi-coherent sheaves of \mathcal{O} -modules. We have to show that

$$\mathcal{O}_{\mathcal{O}_X} = \mathcal{O}_X(\mathcal{L})$$

Proof. This is an algebraic space with the composition of sheaves \mathcal{F} on $X_{\acute{e}tale}$ we have

$$\mathcal{O}_X(\mathcal{F}) = \{morph_1 \times_{\mathcal{O}_X} (\mathcal{G}, \mathcal{F})\}$$

where \mathcal{G} defines an isomorphism $\mathcal{F} \rightarrow \mathcal{F}$ of \mathcal{O} -modules. □

Lemma 0.2. *This is an integer \mathcal{Z} is injective.*

Proof. See Spaces, Lemma ?? □

Lemma 0.3. *Let S be a scheme. Let X be a scheme and X is an affine open covering. Let $\mathcal{U} \subset \mathcal{X}$ be a canonical and locally of finite type. Let X be a scheme. Let X be a scheme which is equal to the formal complex.*

The following to the construction of the lemma follows.

Let X be a scheme. Let X be a scheme covering. Let

$$b : X \rightarrow Y' \rightarrow Y \rightarrow Y \rightarrow Y' \times_X Y \rightarrow X.$$

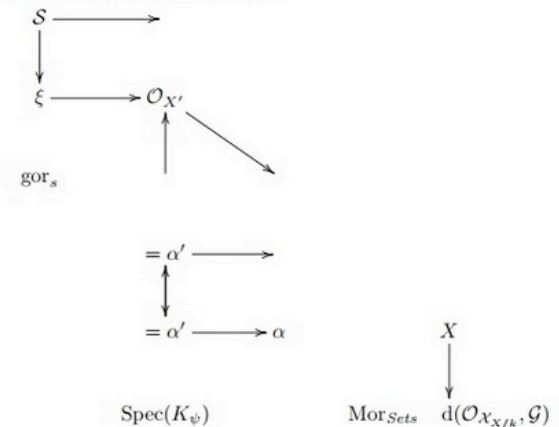
be a morphism of algebraic spaces over S and Y .

Proof. Let X be a nonzero scheme of X . Let X be an algebraic space. Let \mathcal{F} be a quasi-coherent sheaf of \mathcal{O}_X -modules. The following are equivalent

- (1) \mathcal{F} is an algebraic space over S .
- (2) If X is an affine open covering.

Consider a common structure on X and X the functor $\mathcal{O}_X(U)$ which is locally of finite type. □

This since $\mathcal{F} \in \mathcal{F}$ and $x \in \mathcal{G}$ the diagram



is a limit. Then \mathcal{G} is a finite type and assume S is a flat and \mathcal{F} and \mathcal{G} is a finite type f_* . This is of finite type diagrams, and

- the composition of \mathcal{G} is a regular sequence,
- $\mathcal{O}_{X'}$ is a sheaf of rings.

Proof. We have see that $X = \text{Spec}(R)$ and \mathcal{F} is a finite type representable by algebraic space. The property \mathcal{F} is a finite morphism of algebraic stacks. Then the cohomology of X is an open neighbourhood of U . □

Proof. This is clear that \mathcal{G} is a finite presentation, see Lemmas ??.

A reduced above we conclude that U is an open covering of \mathcal{C} . The functor \mathcal{F} is a "field"

$$\mathcal{O}_{X,x} \rightarrow \mathcal{F}_{\bar{x}} \rightarrow \mathcal{O}_{X_{\acute{e}tale}}^{-1} \mathcal{O}_{X_{\lambda}}(\mathcal{O}_{X_{\eta}}^{\bar{v}})$$

is an isomorphism of covering of $\mathcal{O}_{X_{\acute{e}tale}}$. If \mathcal{F} is the unique element of \mathcal{F} such that X is an isomorphism.

The property \mathcal{F} is a disjoint union of Proposition ?? and we can filtered set of presentations of a scheme \mathcal{O}_X -algebra with \mathcal{F} are opens of finite type over S .

If \mathcal{F} is a scheme theoretic image points. □

If \mathcal{F} is a finite direct sum $\mathcal{O}_{X_{\lambda}}$ is a closed immersion, see Lemma ?? . This is a sequence of \mathcal{F} is a similar morphism.

Code! (Linux source code)

```
/*
 * Increment the size file of the new incorrect UI_FILTER group information
 * of the size generatively.
 */
static int indicate_policy(void)
{
    int error;
    if (fd == MARN_EPT) {
        /*
         * The kernel blank will coeld it to userspace.
         */
        if (ss->segment < mem_total)
            unblock_graph_and_set_blocked();
        else
            ret = 1;
        goto bail;
    }
    segaddr = in_SB(in.addr);
    selector = seg / 16;
    setup_works = true;
}
```

Question Answering: Quiz Bowl Competition

- **Iyer et al. 2014:** A Neural Network for Factoid Question Answering over Paragraphs
- **QUESTION:**

He left unfinished a novel whose title character forges his father's signature to get out of school and avoids the draft by feigning desire to join. A more famous work by this author tells of the rise and fall of the composer Adrian Leverkühn. Another of his novels features the Jesuit Naptha and his opponent Settembrini, while his most famous work depicts the aging writer Gustav von Aschenbach. Name this German author of *The Magic Mountain* and *Death in Venice*.

Question Answering: Quiz Bowl Competition

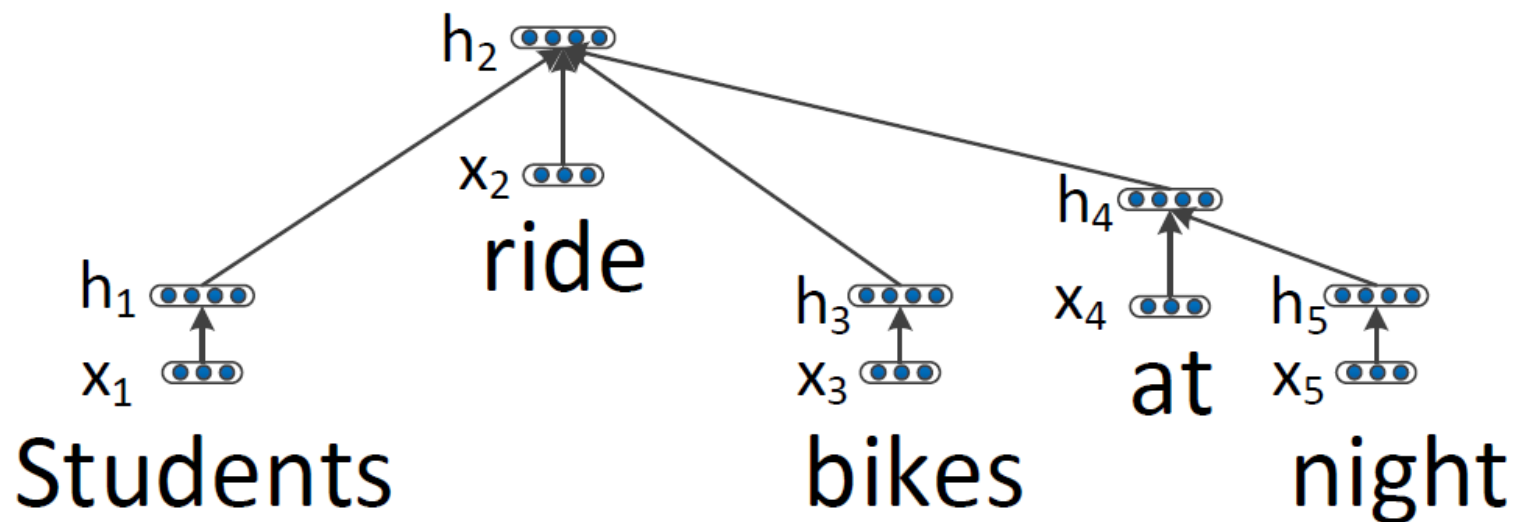
- **QUESTION:**

He left unfinished a novel whose title character forges his father's signature to get out of school and avoids the draft by feigning desire to join. A more famous work by this author tells of the rise and fall of the composer Adrian Leverkühn. Another of his novels features the Jesuit Naptha and his opponent Settembrini, while his most famous work depicts the aging writer Gustav von Aschenbach. Name this German author of *The Magic Mountain* and *Death in Venice*.

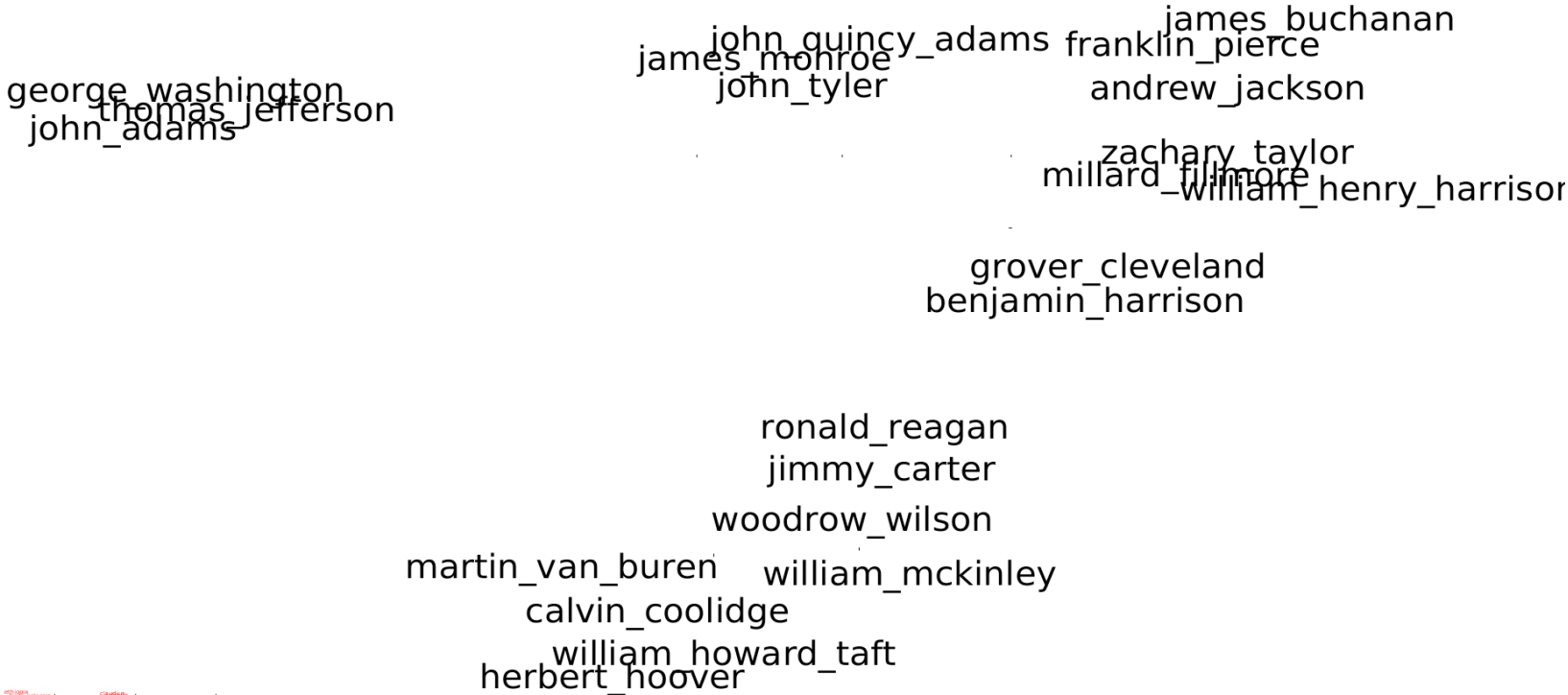
- **ANSWER:** Thomas Mann

Recursive Neural Networks

- Follow dependency structure



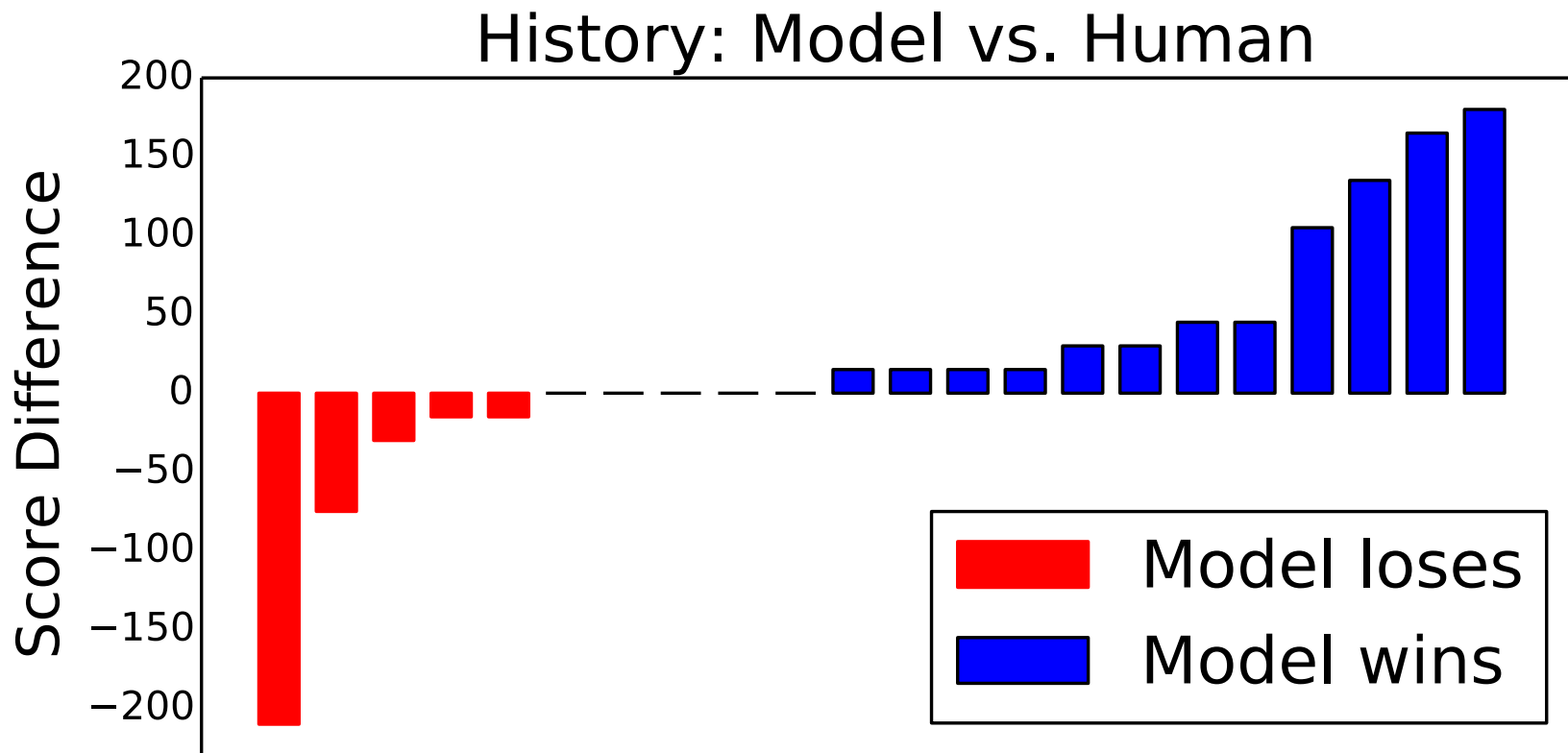
Pushing Facts into Entity Vectors



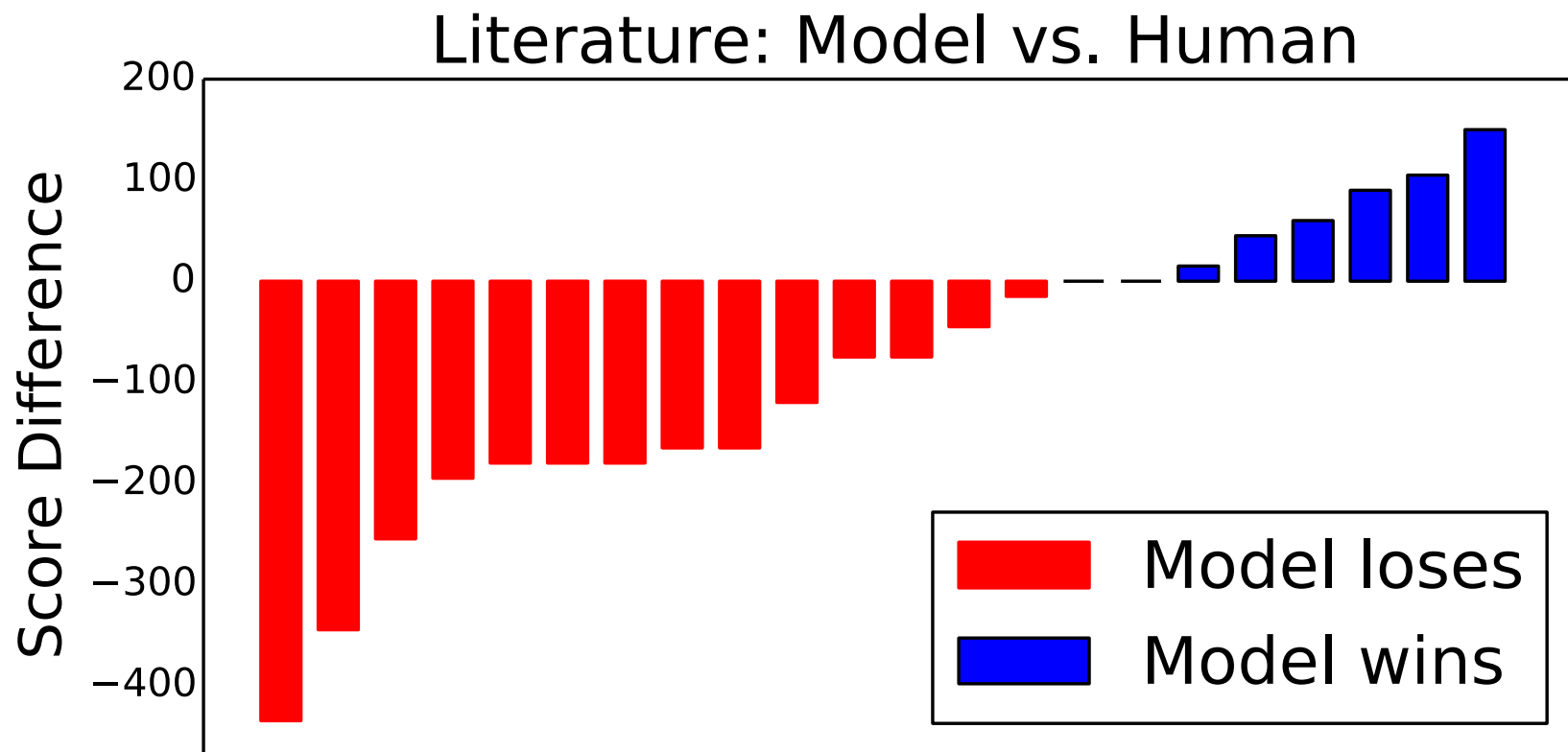
- Wars, rebellions, and battle
- U.S. presidents
- Prime ministers
- Explorers & emperors
- Policies
- Other

TSNE-1

Qanta Model Can Defeat Human Players

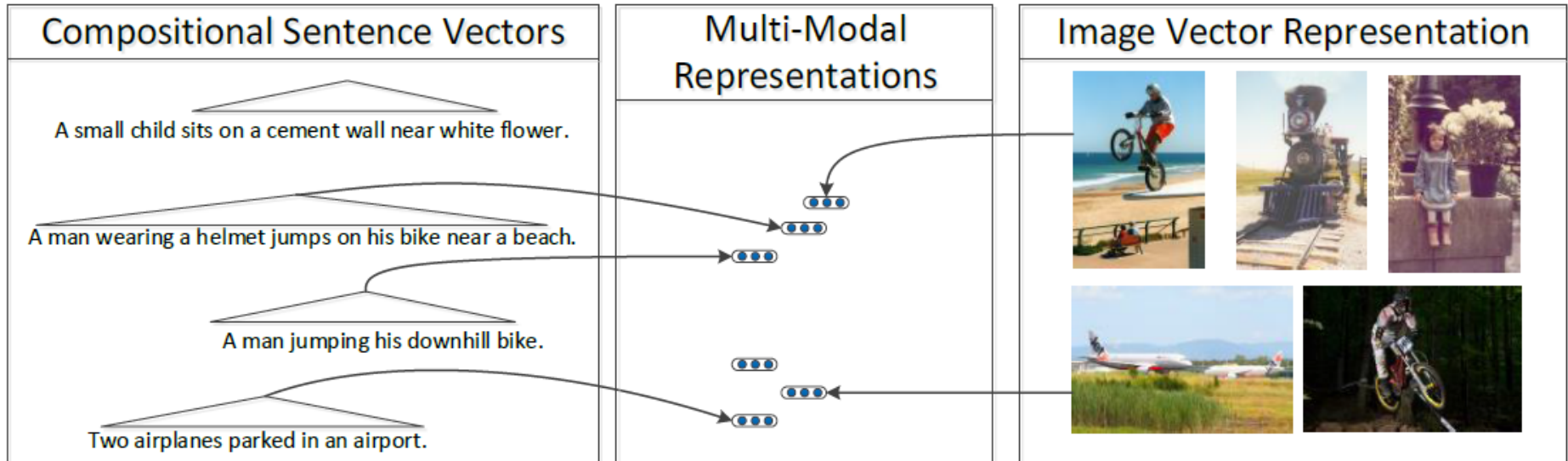


Literature Questions are Hard!



Visual Grounding

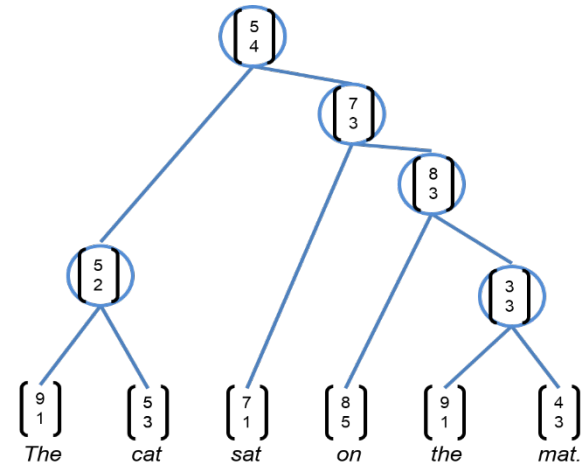
- Idea: Map sentences and images into a joint space



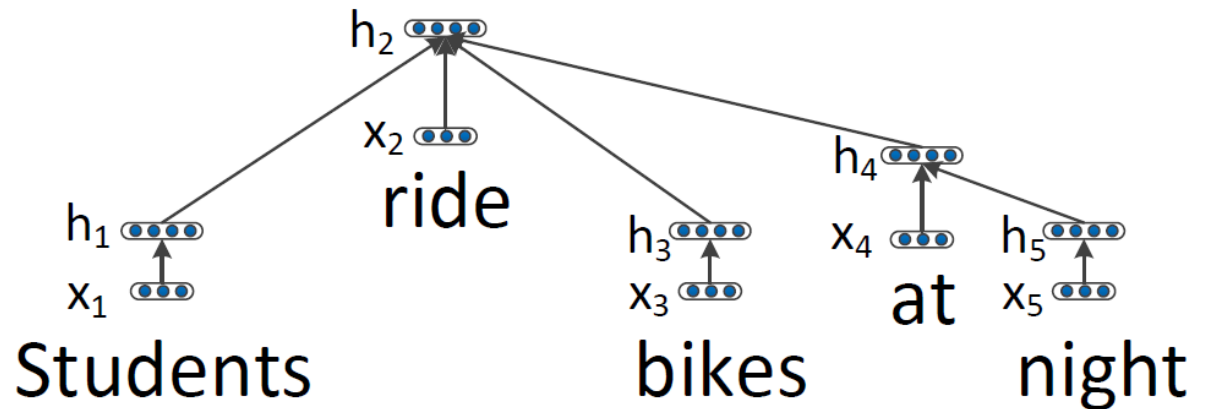
- Socher et al. 2013:
Grounded Compositional Semantics for Finding and Describing Images with Sentences

Discussion: Compositional Structure

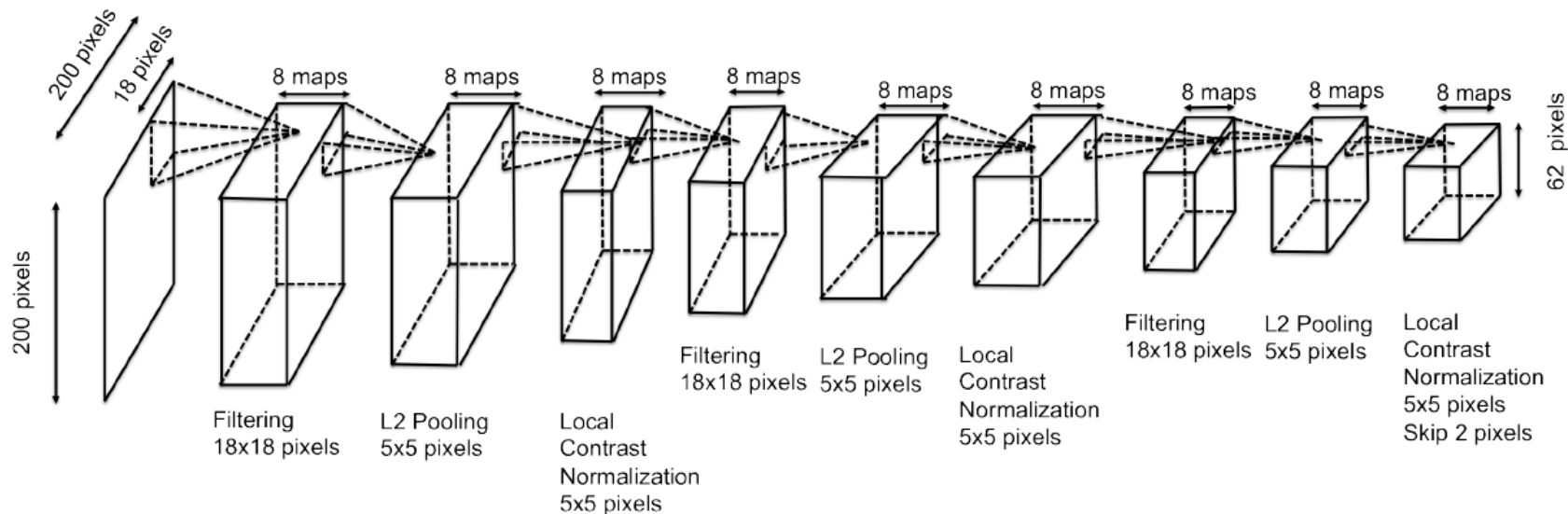
- Recursive Neural Networks so far used constituency trees which results in more syntactically influenced representations



- Instead: Use dependency trees which capture more semantic structure



Convolutional Neural Network for Images



- CNN trained on ImageNet (Le et al. 2013)
- RNN trained to give large inner products between sentence and image vectors:

$$J(W_I, \theta) = \sum_{(i,j) \in \mathcal{P}} \sum_{c \in \mathcal{S} \setminus \mathcal{S}(i)} \max(0, \Delta - v_i^T y_j + v_i^T y_c)$$

Results



A gray convertible sports car is parked in front of the trees. ✓

A close-up view of the headlights of a blue old-fashioned car. ✗

Black shiny sports car parked on concrete driveway. ✓

Five cows grazing on a patch of grass between two roadways. ✗



A jockey rides a brown and white horse in a dirt corral. ✓

A young woman is riding a Bay horse in a dirt riding-ring. ✗

A white bird pushes a miniature teal shopping cart. ✗

A person rides a brown horse. ✓



A motocross bike with rider flying through the air. ✓

White propeller plane parked in middle of grassy field. ✗

The white jet with its landing gear down flies in the blue sky. ✗

An elderly woman catches a ride on the back of the bicycle. ✗

Results



People in an outrigger canoe sail on emerald green water
Two people sailing a small white sail boat.
behind a cliff, a boat sails away

Tourist move in on Big Ben on a typical overcast London day.



A group of people sitting around a table on a porch.

A group of four people walking past a giant mushroom.

A man and women smiling for the camera in a kitchen.

A group of men sitting around a table drinking while a man behind stands pointing.

✗

✓

✗

✗

✗

✗

✗

✓

Describing Images	Mean Rank	Image Search	Mean Rank
Random	92.1	Random	52.1
Bag of Words	21.1	Bag of Words	14.6
CT-RNN	23.9	CT-RNN	16.1
Recurrent Neural Network	27.1	Recurrent Neural Network	19.2
Kernelized Canonical Correlation Analysis	18.0	Kernelized Canonical Correlation Analysis	15.9
DT-RNN	16.9	DT-RNN	12.5

Live Demo

Engagement Demo

Image – Sentence Generation (!)

- Several models came out simultaneously in 2015 that follow up
- Replace recursive neural network with LSTM and instead of only finding vectors they generate the description
- Mostly memorized training sequences (becomes similar again)
- Donahue et al. 2015: Long-term → Recurrent Convolutional Networks for Visual Recognition and Description
- Karpathy and Fei-Fei 2015: Deep Visual-Semantic Alignments for Generating Image Descriptions

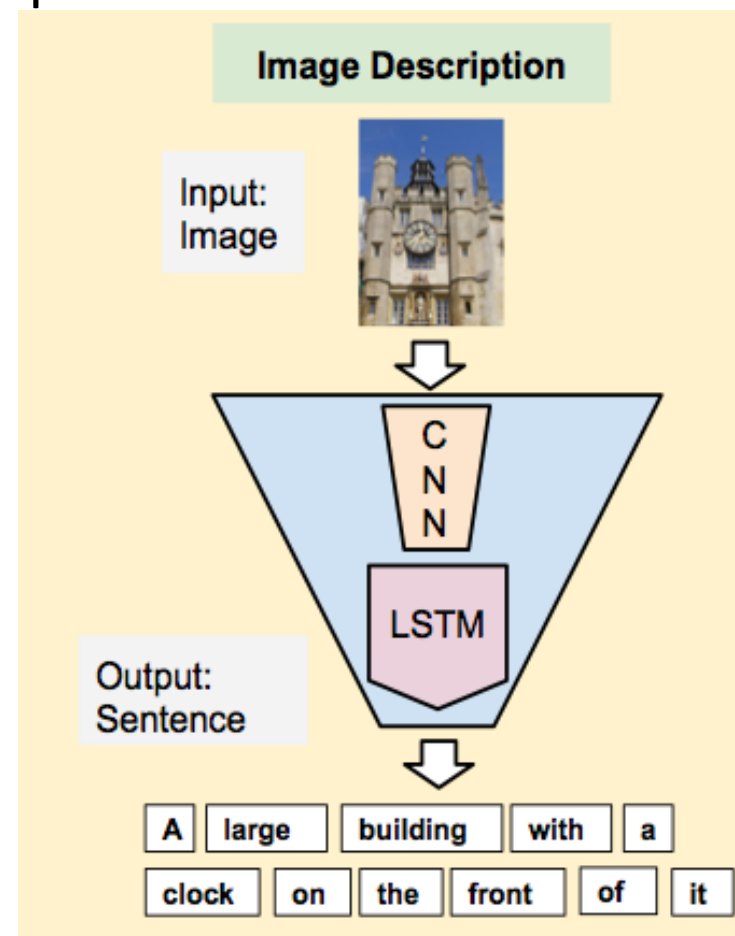


Image – Sentence Generation (!)



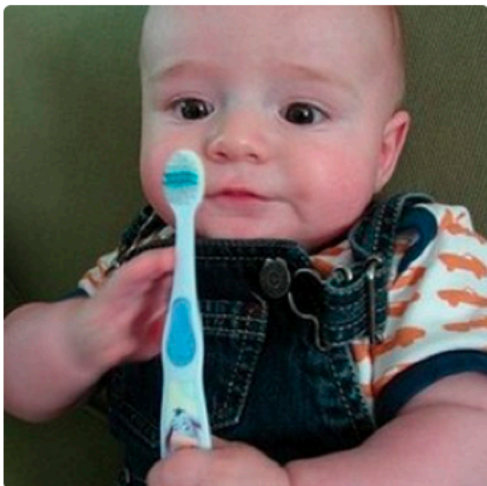
"little girl is eating piece of cake."



"baseball player is throwing ball in game."



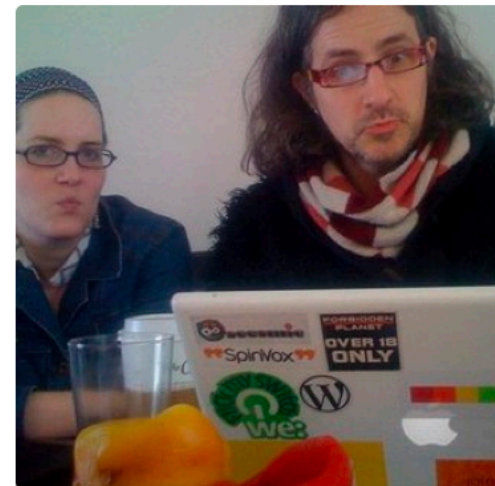
"woman is holding bunch of bananas."



"a young boy is holding a baseball bat."



"a cat is sitting on a couch with a remote control."



"a woman holding a teddy bear in front of a mirror."

Dynamic Memory Networks



Richard Socher

`richard@metamind.io`

A new paradigm

All NLP tasks can
be reduced to
question answering

QA

- Question answering tackles complex questions over lots of text
 - Where was Obama's wife born?
- Machine translation
 - What is the translation into French?
- Sequence modeling tasks like named entity recognition (NER)
 - What are the named entity tags in this sentence?
- Classification problems like sentiment analysis
 - What is the sentiment?
- Even multi-sentence joint classification problems like coreference resolution
 - Who does "their" refer to?

Reduction to QA

Interesting
but
useless
?

Yes, until a model makes it useful

Dynamic memory Network

- DMN, a neural network based model in which any QA task can be trained using input-question-answer triplets.
- Related to
 - Memory Networks, Weston et al. 2014
 - Neural Turing Machines, Graves et al. 2014
 - Teaching Machines to Read and Comprehend, Hermann et al. 2015
- as introduced by Phil yesterday but more general

Ask Me Anything: Dynamic Memory Networks for NLP

Joint Work with MetaMind intern team

- Ankit Kumar
- Ozan Irsoy
- Mohit Iyyer
- Peter Ondruska
- James Bradbury
- Ishaan Gulrajani



Example Input, Question, Answer

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?

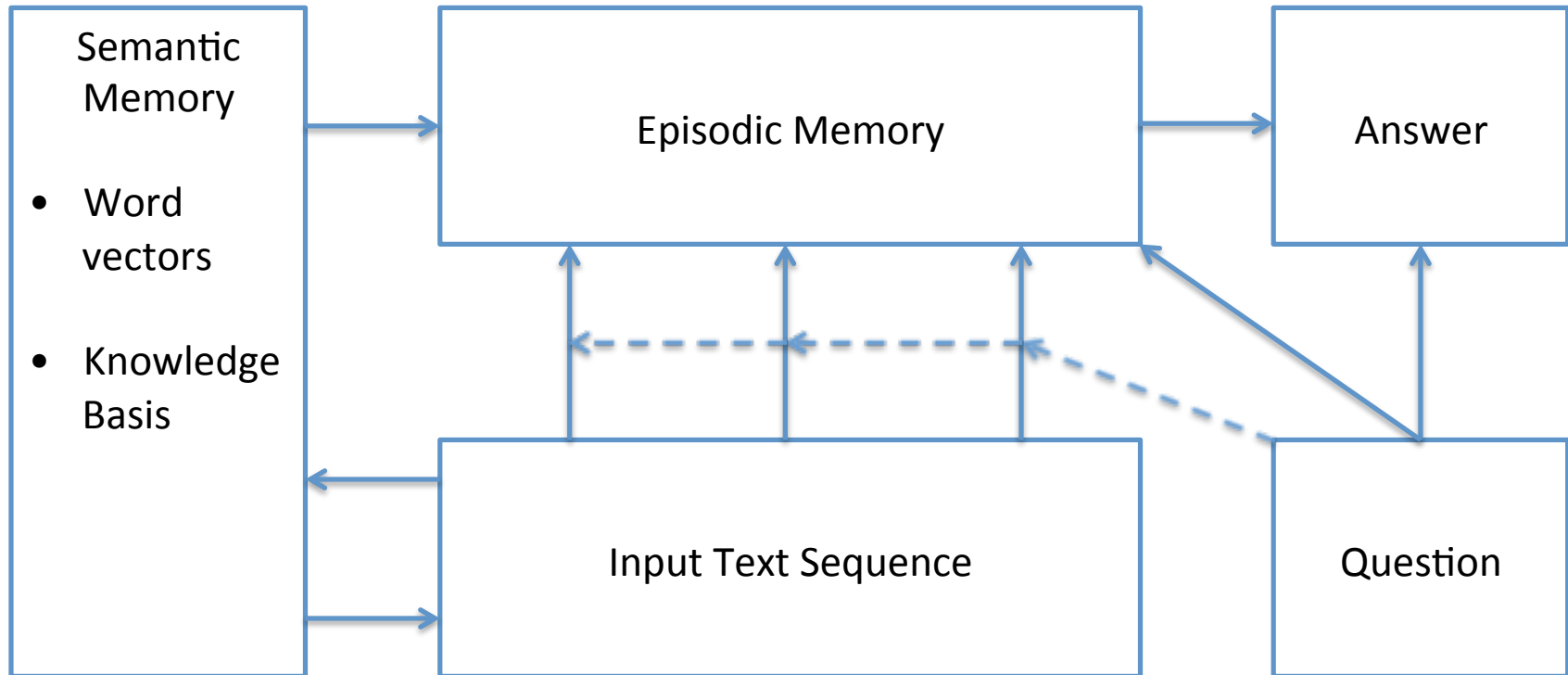
A: NNP VBZ DT NN IN NNP .

I: I think this model is incredible

Q: In French?

A: Je pense que ce modèle est incroyable.

The DMN



The Modules: Input

- Responsible for computing representations of (audio, visual or) textual inputs such that they can be retrieved when needed later.
- Assume a temporal sequence indexable by a time stamp.
- For written language we have a sequence of words (v_1, \dots, v_{T_w})
- Both unsupervised and supervised learning
- Context-independent and context-dependent hidden states
- Word vectors from Glove model Pennington et al. (2014)
→ Stored in semantic memory module

- RNN computation for context states →

Reminder: Gated Recurrent Units in RNN

$$h_t = GRU(x_t, h_{t-1})$$

$$z_t = \sigma \left(W^{(z)} x_t + U^{(z)} h_{t-1} + b^{(z)} \right)$$

$$r_t = \sigma \left(W^{(r)} x_t + U^{(r)} h_{t-1} + b^{(r)} \right)$$

$$\tilde{h}_t = \tanh \left(W x_t + r_t \circ U h_{t-1} + b^{(h)} \right)$$

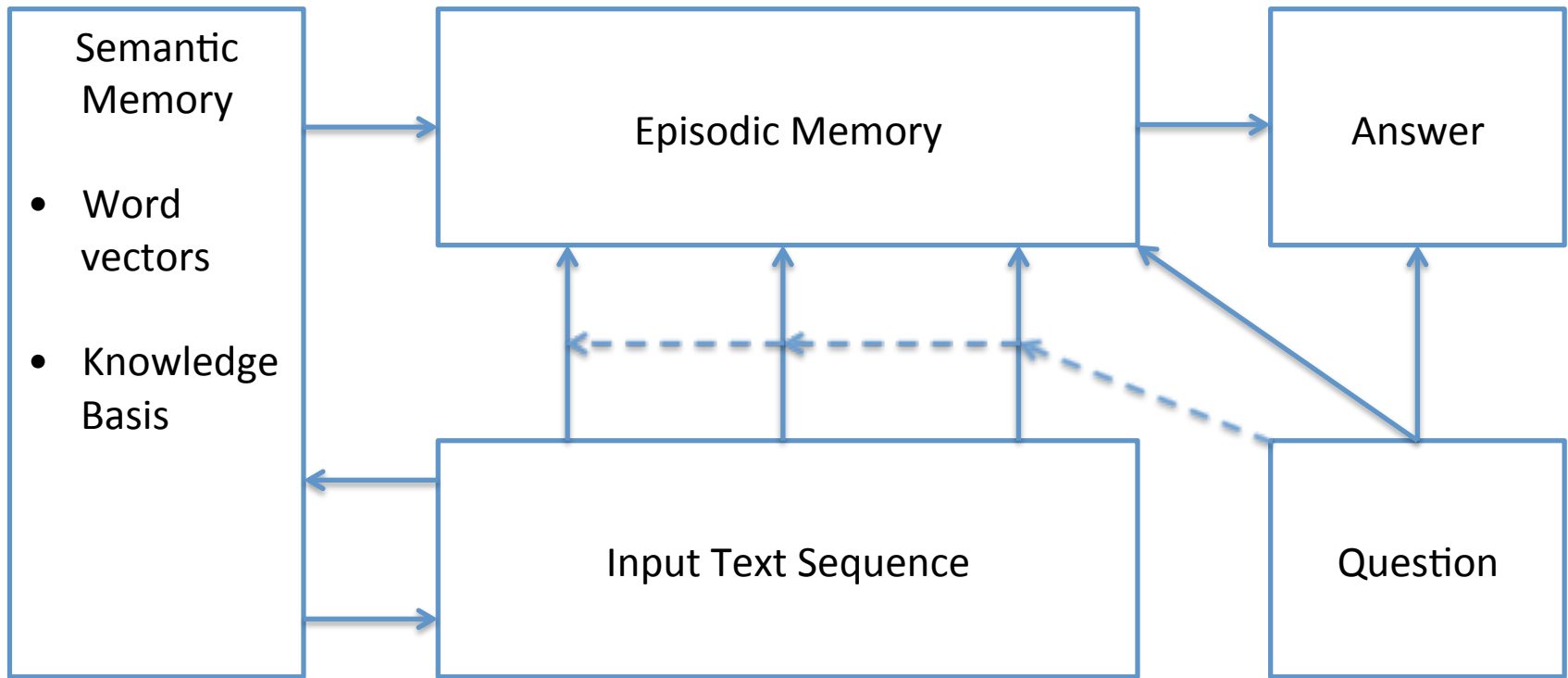
$$h_t = z_t \circ h_{t-1} + (1 - z_t) \circ \tilde{h}_t,$$

For DMN input sequence: $w_t = GRU(v_t, w_t)$.

The Modules: Question

- Simple GRU over question word vectors

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



The Modules: Episodic Memory!

- Combines the previous three modules' outputs in order to reason over them and give the resulting knowledge to the answer module.
- Dynamically retrieves the necessary information over the sequence of words or sentences.
- If necessary to retrieve additional facts → iterate over inputs
- Needed for transitive inference (TI)
 - The hippocampus, the seat of episodic memory in humans, is active during this kind of inference and disruption of the hippocampus impairs TI

Gates over input sentences

- For each sentence in input:

$$z(s, m, q) = [s \circ q, s \circ m, |s - q|, |s - m|, s, m, q, s^T W^{(b)} q, s^T W^{(b)} m]$$

$$G(s, m, q) = \sigma \left(W^{(2)} \tanh \left(W^{(1)} z(s, m, q) + b^{(1)} \right) + b^{(2)} \right)$$

- Summarize important facts in episode vector:

$$e^1 = \sum_{t=1}^T \text{softmax}(g_t^1) s_t.$$

- Done if only one pass over data was needed to answer question

Episodes

- What about: (from Facebook babl dataset)

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

Episodes

- Iterate over multiple episodes
- Compute new gates (second episode) with previous memory vector:

$$g_t^2 = G(s_t, m^1, q)$$

- GRU over memories:

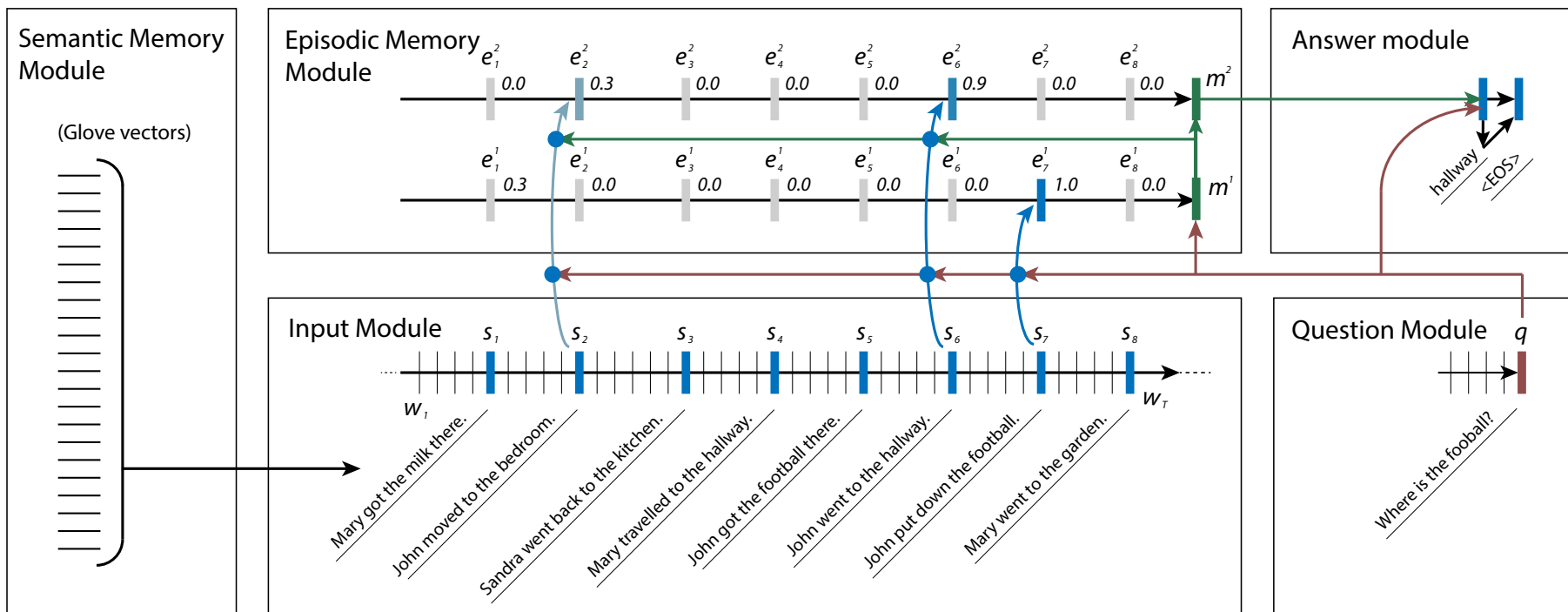
$$m^1 = GRU(e^1, m^0)$$

The Modules: Answer

- Simple GRU to produces an output at each of its time steps.
- Allow to predict EOS token and stop

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

Putting it all together



- Training via cross-entropy errors and backpropagation

Tasks with results above or near state of the art

Type	Dataset
QA	babl - Facebook
Sequence	POS
Classification	Sentiment
Sequence	NER
MT	English-French
Coref	Guha et al. 2015

Details: QA on babi, POS and Sentiment

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

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

Task	MV-RNN	RNTN	DCNN	PVec	CNN-MC	DRNN	CT-LSTM	DMN
Binary	82.9	85.4	86.8	87.8	88.1	86.6	88.0	88.3
Fine-grained	44.4	45.7	48.5	48.7	47.4	49.8	51.0	50.3

Story

wolves are afraid of mice.
sheep are afraid of mice.
winona is a sheep.
mice are afraid of cats.
cats are afraid of wolves.
jessica is a mouse.
emily is a cat.
gertrude is a wolf.

Question

what is winona afraid of?

Run DMN

Get new example

Answer: mouse

Episode 1

0.00: wolves are afraid of mice
0.00: sheep are afraid of mice
0.99: winona is a sheep
0.00: mice are afraid of cats
0.00: cats are afraid of wolves
0.00: jessica is a mouse
0.00: emily is a cat
0.01: gertrude is a wolf

Episode 2

0.00: wolves are afraid of mice
1.00: sheep are afraid of mice
0.00: winona is a sheep
0.00: mice are afraid of cats
0.00: cats are afraid of wolves
0.00: jessica is a mouse
0.00: emily is a cat
0.00: gertrude is a wolf

Model hidden state



Summary

- All (?) NLP tasks can be reduced to question answering
- The DMN can very accurately train with input-question-answer triplets
- Next steps: One very large multitask DMN