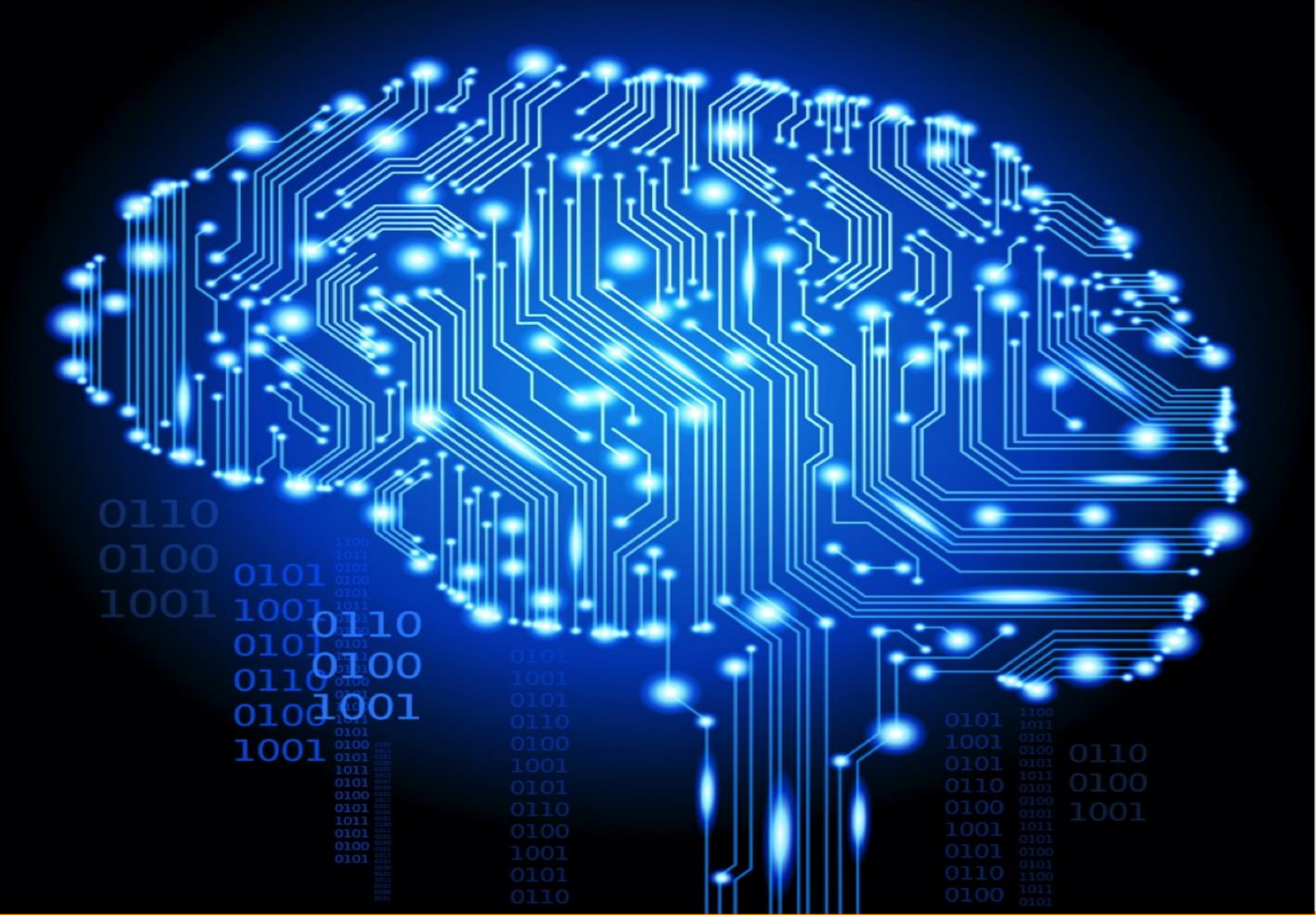


Lecture 10

Deep Learning@UvA

Previous Lecture

- Recurrent Neural Networks (RNN) for sequences
- Backpropagation Through Time
- RNNs using Long Short-Term Memory (LSTM)
- Applications of Recurrent Neural Networks

Lecture Overview

- Memory networks
- Recursive networks

Memory

Why memory? Example!

- ○ *“Bilbo travelled to the cave. Gollum dropped the ring there. Bilbo took the ring. This sentence is random noise for illustration purposes. Bilbo went back to the Shire. Bilbo left the ring there. Frodo got the ring. Frodo journeyed to Mount-Doom. Frodo dropped the ring there. Sauron died. Frodo went back to the Shire. Bilbo travelled to the Grey-havens. The End.”*

Why memory? Example!

- *“Bilbo travelled to the cave. Gollum dropped the ring there. Bilbo took the ring. This sentence is random noise for illustration purposes. Bilbo went back to the Shire. Bilbo left the ring there. Frodo got the ring. Frodo journeyed to Mount-Doom. Frodo dropped the ring there. Sauron died. Frodo went back to the Shire. Bilbo travelled to the Grey-havens. The End.”*
- *“Q: Where is the ring?”* → *“A: Mount-Doom”*

Why memory? Example!

- *"Bilbo travelled to the cave. Gollum dropped the ring there. Bilbo took the ring. This sentence is random noise for illustration purposes. Bilbo went back to the Shire. Bilbo left the ring there. Frodo got the ring. Frodo journeyed to Mount-Doom. Frodo dropped the ring there. Sauron died. Frodo went back to the Shire. Bilbo travelled to the Grey-havens. The End."*
- *"Q: Where is the ring?"* → *"A: Mount-Doom"*
- *"Q: Where is Bilbo now?"* → *"A: Grey-havens"*

Why memory? Example!

- *“Bilbo travelled to the cave. Gollum dropped the ring there. Bilbo took the ring. This sentence is random noise for illustration purposes. Bilbo went back to the Shire. Bilbo left the ring there. Frodo got the ring. Frodo journeyed to Mount-Doom. Frodo dropped the ring there. Sauron died. Frodo went back to the Shire. Bilbo travelled to the Grey-havens. The End.”*
- *“Q: Where is the ring?”* → *“A: Mount-Doom”*
- *“Q: Where is Bilbo now?”* → *“A: Grey-havens”*
- *“Q: Where is Frodo now?”* → *“A: Shire”*

Why memory? Example!

- *“Bilbo travelled to the cave. Gollum dropped the ring there. Bilbo took the ring. This sentence is random noise for illustration purposes. Bilbo went back to the Shire. Bilbo left the ring there. Frodo got the ring. Frodo journeyed to Mount-Doom. Frodo dropped the ring there. Sauron died. Frodo went back to the Shire. Bilbo travelled to the Grey-havens. The End.”*
- *“Q: Where is the ring?”* → *“A: Mount-Doom”*
- *“Q: Where is Bilbo now?”* → *“A: Grey-havens”*
- *“Q: Where is Frodo now?”* → *“A: Shire”*
- Can we design a network that answers such questions?

Memory networks

- Neural network models that
  - have large memory that can store many facts
  - have a learning component for how to read, store, forget and access these facts
- Intuitively, they should work like a “Neural RAM” or a “”Neural Wikipedia”
  - The network processes Wikipedia like information. It needs to store them appropriately for easy read/write/delete/access actions.
  - You make a question
  - The network should recognize the right types of memories
  - The network should reply the question with a meaningful (non trivial) answer.

What is difficult with memory?

- Some sentences are factual
  - “Frodo got the ring”, “Frodo went back to the Shire”
- Some sentences might be random noise
  - *“This sentence is random noice for illustration purposes”*
- To answer a question you might need to combine facts
  - “Where did Frodo get the ring?”
  - *“Bilbo went back to the Shire”* → *“Bilbo left the ring there.”* → *“Frodo got the ring.”*
  - To answer correctly all three sentences need to be carefully analyzed
- TOO MUCH INFORMATION within a single story!!!
  - Can a standard memory unit cope with that?

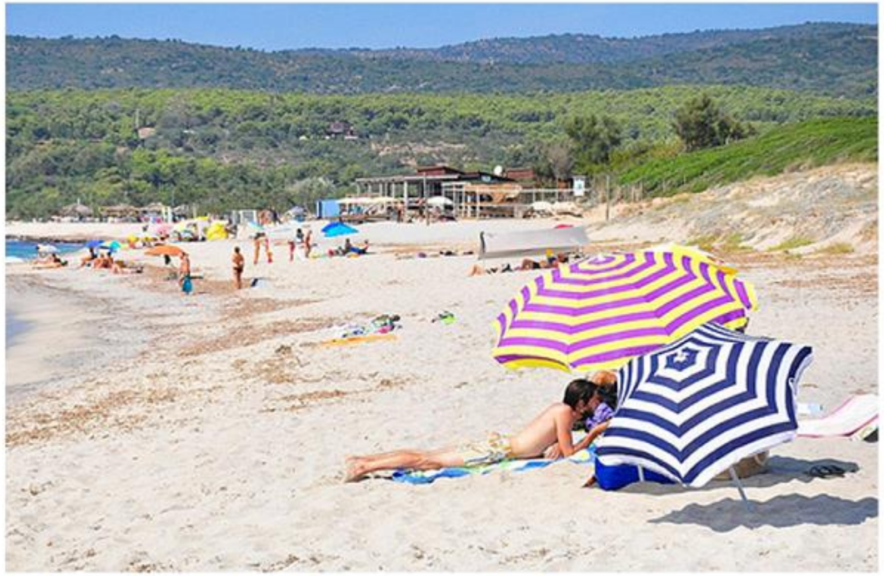
What is difficult with memory? *(2)*

- Each new story can be completely different
  - Very little data to actually train on
- If we use real data, we don’t (usually) have annotations
  - How to analyze mistakes?
- Solution for the last two problems: Make own dataset
  - Start from simple factual sentences and build artificial stories
- Example
  - “John is in the playground.
    Bob is in the office.
    John picked up the football.
    Bob went to the kitchen.”
    “Q: Where is the football?” → *“A:playground”*
    “Q: Where was Bob before the kitchen?” “A:office”

Why not simply LSTMs?

- Probably its memory is not large enough
- In latest experiments it seems that LSTMs are not flexible enough for these tasks
  - Although one could maybe create an LSTM-version more specific for the task
- At the end of the day this is still research of the last year
  - *“A research topic that has gained popularity within a small circle of deep learning researchers over the last few months is the combination of a deep neural net and short-term memory. Basically, the neural net acts as a "reasoning" engine that stores and retrieves data to be operated on from a separate memory.”*
  - https://www.facebook.com/FBAIResearch/posts/362517620591864 (Nov 3 2014)

Memory Networks



Attributes:

umbrella
beach
sunny
day
people
sand
laying
blue
green
mountain

Internal Textual Representation:

A group of people enjoying a <u>sunny</u> day at the <u>beach</u> with <u>umbrellas</u> in the sand.

External Knowledge:

An <u>umbrella</u> is a canopy designed to protect against rain or sunlight. Larger <u>umbrellas</u> are often used as points of <u>shade</u> on a <u>sunny beach</u>. A <u>beach</u> is a landform along the coast of an ocean. It usually consists of loose particles, such as <u>sand</u>....

Question Answering:

Q: Why do they have umbrellas? **A :** Shade.

Figure 1. A real case of question answering based on an internal textual representation and external knowledge. All of the attributes, textual representation, knowledge and answer are produced by our VQA model. Underlined words indicate the information required to answer the question.

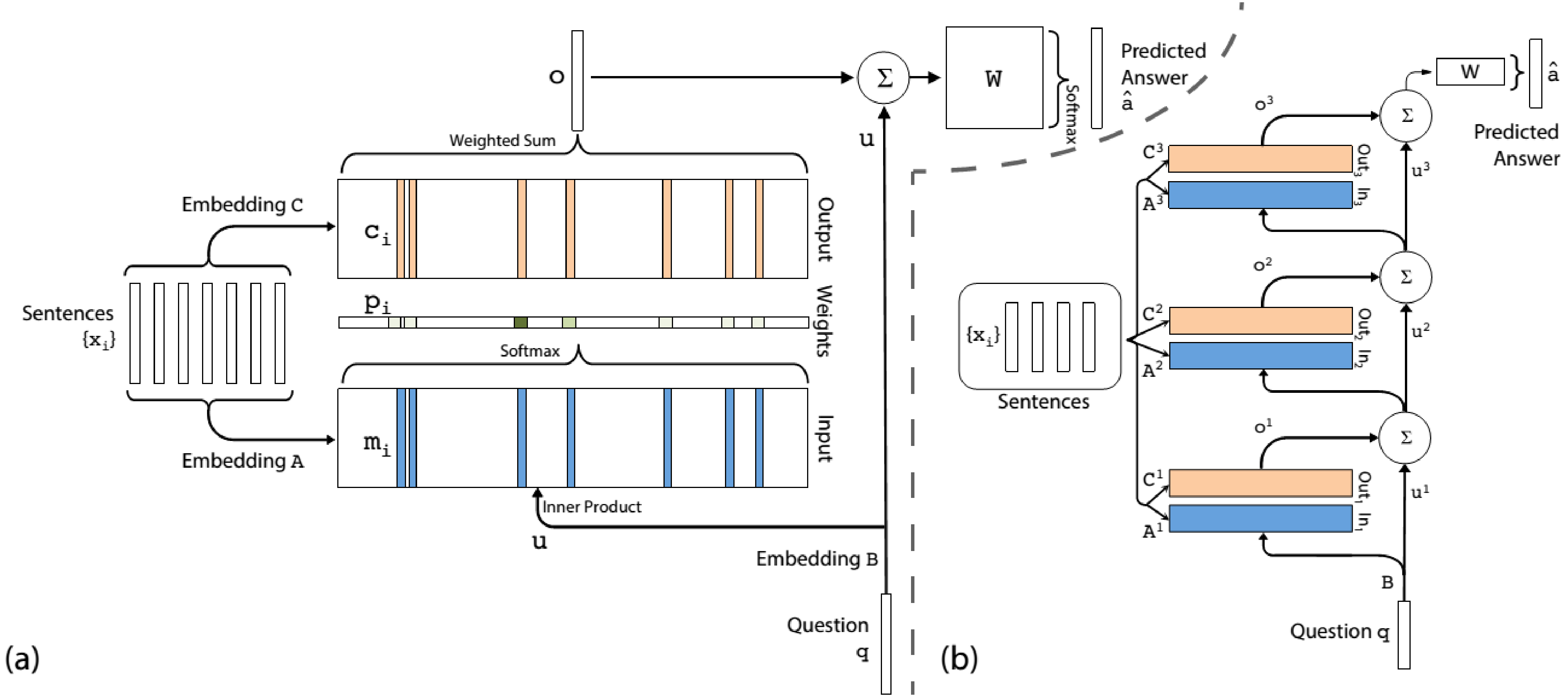
Papers in the literature

- *Neural Turing Machines*, A. Graves, G. Wayne, I. Danihelka, arXiv 2014
  - http://arxiv.org/abs/1410.5401
- *Memory Networks*, J. Weston, S. Chopra, A. Bordes, arXiv 2014
  - http://arxiv.org/abs/1410.3916
- *End-to-end Memory Networks*, S. Sukhbaatar, A. Szlam, J. Weston, R. Fergus, arXiv 2015
  - http://arxiv.org/abs/1503.08895
- *Ask Me Anything: Free-form Visual Question Answering Based on Knowledge from External Sources,* Q. Wu, P. Wang, C. Shen, A. van den Hengel, A. Dick, arXiv 2015
  - http://arxiv.org/abs/1511.06973

Papers in the literature

- *Neural Turing Machines*, A. Graves, G. Wayne, I. Danihelka, arXiv 2014
  - http://arxiv.org/abs/1410.5401
- *Memory Networks*, J. Weston, S. Chopra, A. Bordes, arXiv 2014
  - http://arxiv.org/abs/1410.3916
- *End-to-end Memory Networks*, S. Sukhbaatar, A. Szlam, J. Weston, R. Fergus, arXiv 2015
  - http://arxiv.org/abs/1503.08895
- *Ask Me Anything: Free-form Visual Question Answering Based on Knowledge from External Sources,* Q. Wu, P. Wang, C. Shen, A. van den Hengel, A. Dick, arXiv 2015
  - http://arxiv.org/abs/1511.06973

End-to-end Memory Networks



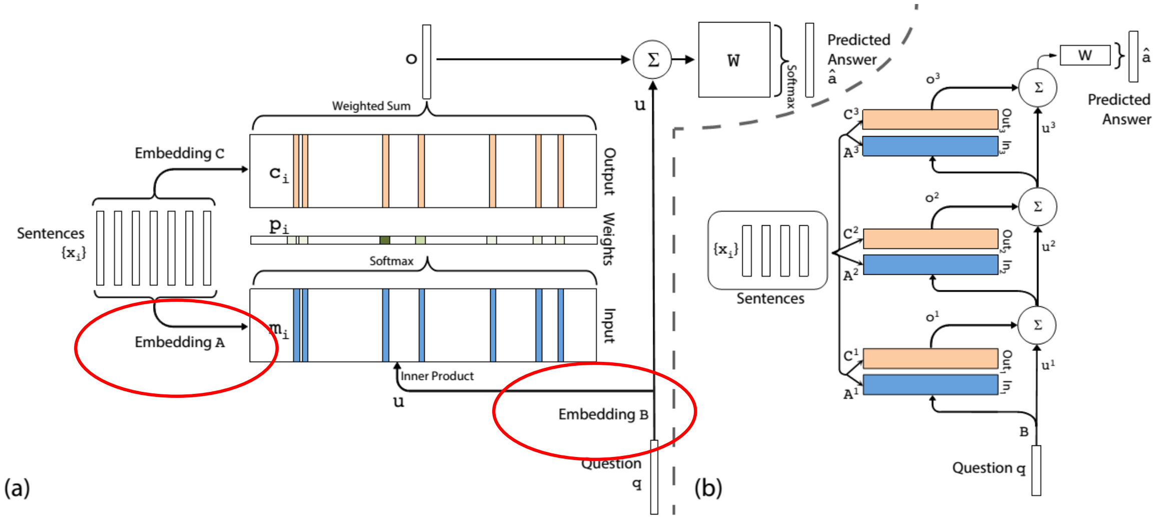
- Torch code available
  - https://github.com/facebook/MemNN/tree/master/MemN2N-lang-model

End-to-end Memory Network unit

- Input memory representation
  - Embeds incoming data to internal representation
- Generalization
  - Given a new input, this unit updates the network memories
- Output
  - Given the memories and given the input, this unit returns a new state variable in the internal representation space of the network
- Response
  - Given the output this unit returns a response recognizable by humans

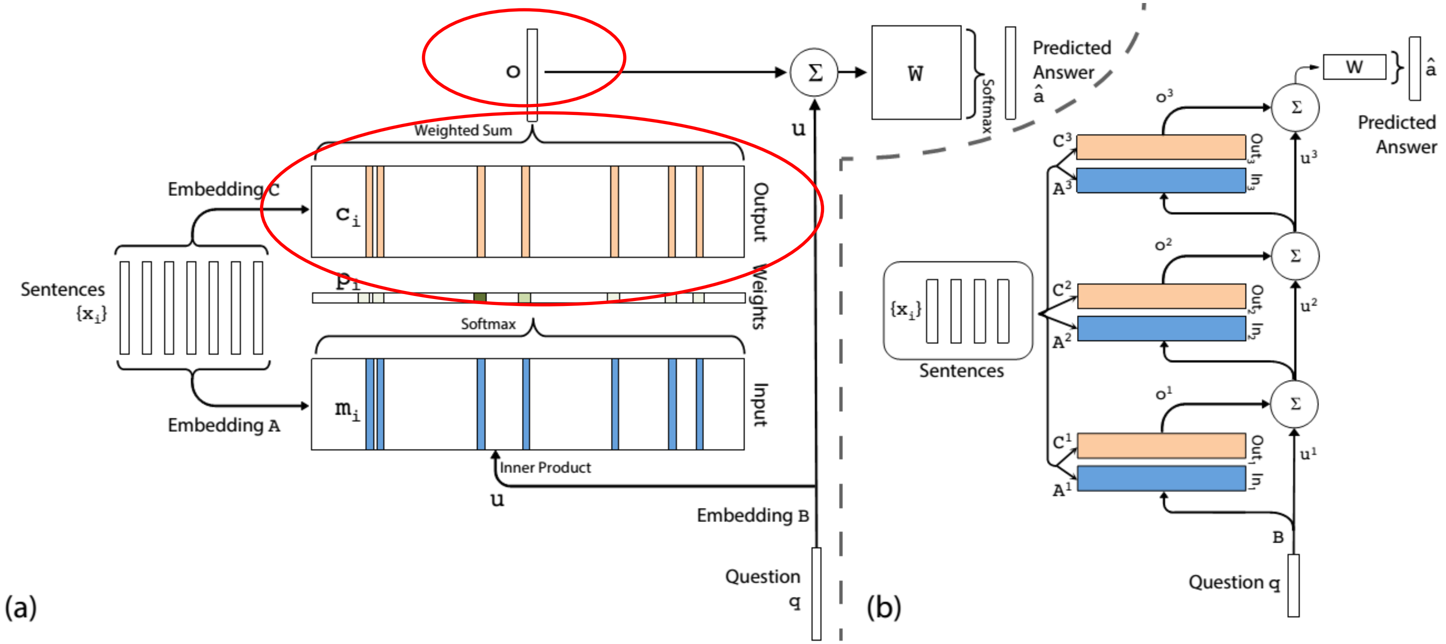
End-to-end Memory Networks: Step (1)

- Input memory representation
- Two embeddings A, B
  - A embeds stories into memory slots on an internal representation space → m_i
  - B embeds the question on the same internal representation space → u
  - To compare memories with questions → $p_i = \text{softmax}(u^T m_i)$



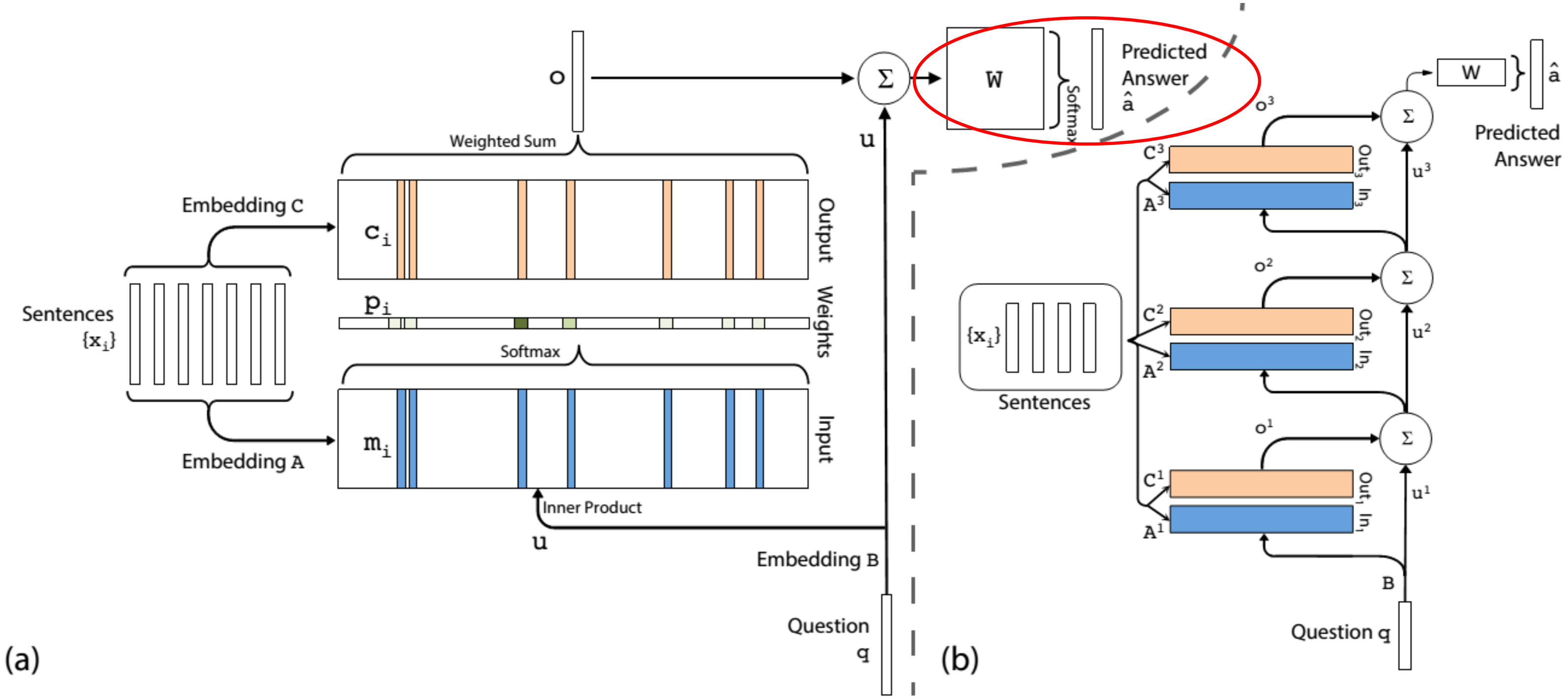
End-to-end Memory Networks: Step (2)

- Output memory representation
  - $o = \sum_i p_i c_i$, where $c_i = C x_i$
- The function that connects the output to the input is smooth
  - Easy to compute gradients for backpropagation



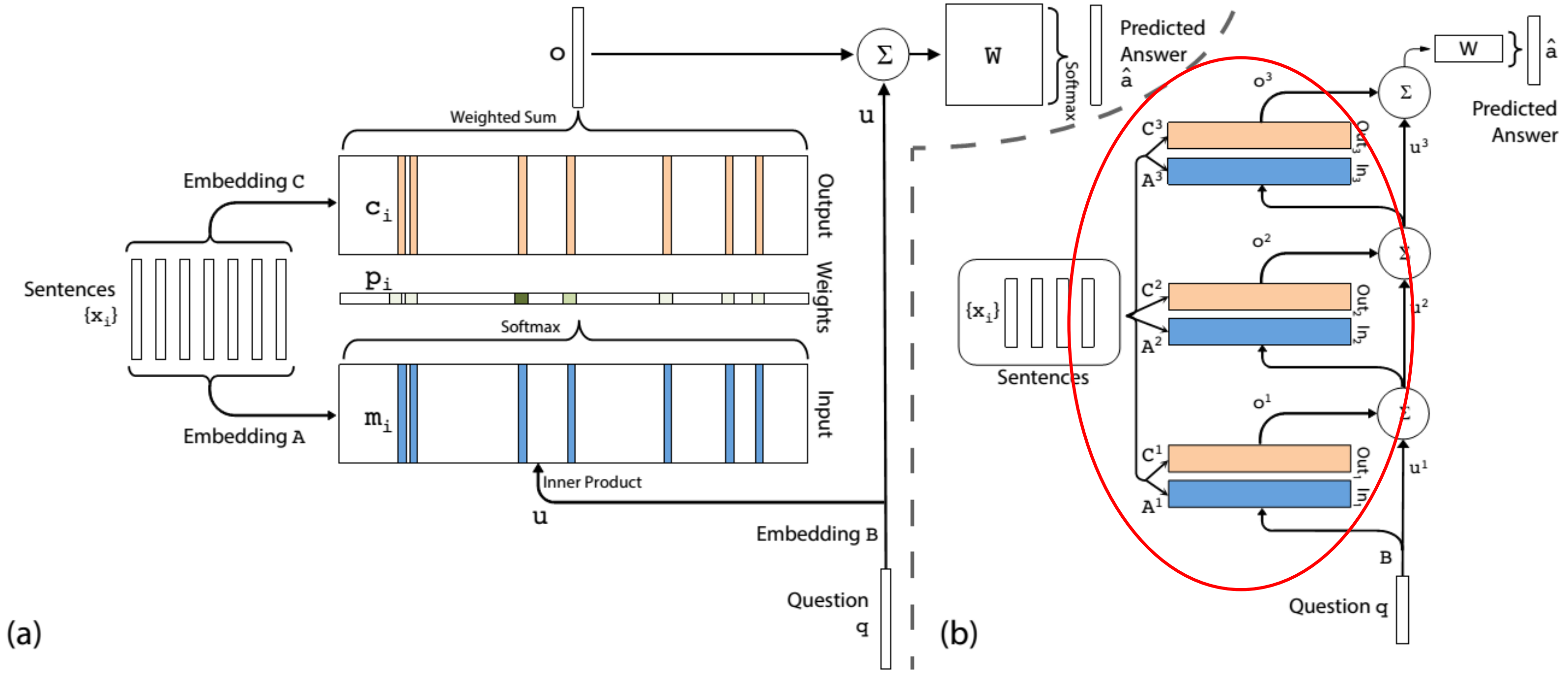
End-to-end Memory Networks: Step (3)

- Final prediction
  - Given the question embedding u and the generated output o
  - $\hat{a} = \text{softmax}(W(o+u))$



End-to-end Memory Networks: Step (4)

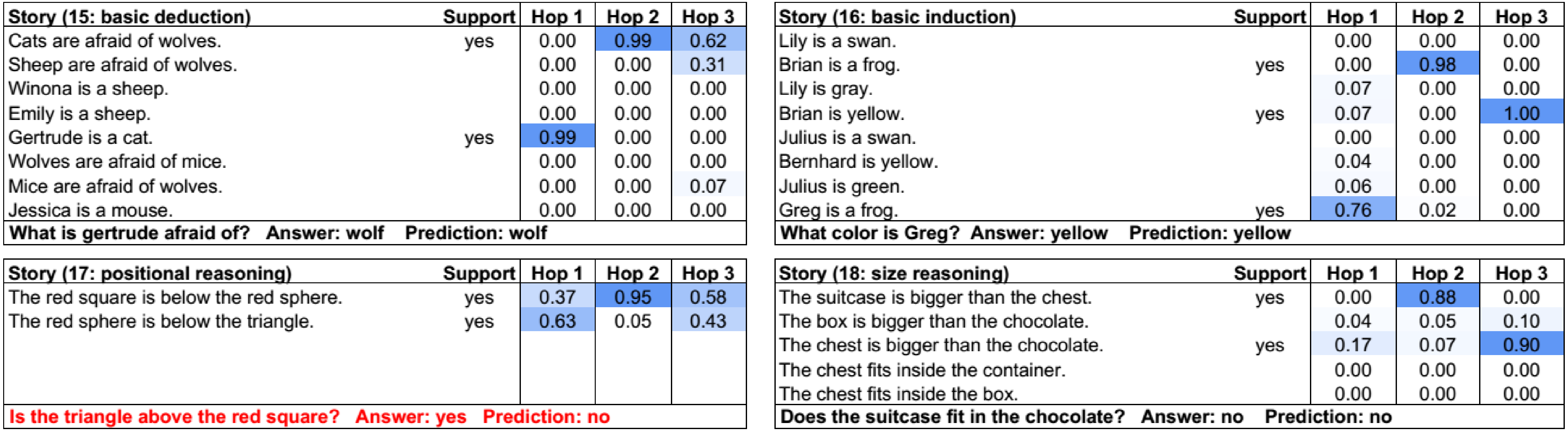
- Adding multiple memory layers
  - The input question for a layer is the output plus the question of the previous layer
  - $u^{k+1} = u^k + o^k$
- Each layer has its own embeddings, although they can be tied together



Additional tricks

- Use multiple hops/steps/layers of memories
  - Increases the memory depth of the network
- Use word embeddings and Bag-of-Words representations as inputs
- Use an RNN as response unit
- Add a forget mechanism for when memory is full
- Maybe go to lower level of tokenization
  - Words, letters, chunks
  - Put chunks into memory slots

Successes/failures



Story (15: basic deduction)	Support	Hop 1	Hop 2	Hop 3
Cats are afraid of wolves.	yes	0.00	0.99	0.62
Sheep are afraid of wolves.		0.00	0.00	0.31
Winona is a sheep.		0.00	0.00	0.00
Emily is a sheep.		0.00	0.00	0.00
Gertrude is a cat.	yes	0.99	0.00	0.00
Wolves are afraid of mice.		0.00	0.00	0.00
Mice are afraid of wolves.		0.00	0.00	0.07
Jessica is a mouse.		0.00	0.00	0.00

What is gertrude afraid of? Answer: wolf Prediction: wolf

Story (16: basic induction)	Support	Hop 1	Hop 2	Hop 3
Lily is a swan.		0.00	0.00	0.00
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 a swan.		0.00	0.00	0.00
Bernhard is yellow.		0.04	0.00	0.00
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 (17: positional reasoning)	Support	Hop 1	Hop 2	Hop 3
The red square is below the red sphere.	yes	0.37	0.95	0.58
The red sphere is below the triangle.	yes	0.63	0.05	0.43

Is the triangle above the red square? Answer: yes Prediction: no

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

What can't be done, what comes next?

- Current networks answer rather simple questions. Make questions harder
  - "Q: Who is teaching the Deep Learning Course?" → "A. Efstratios Gavves **and** Patrick Putzky"
- Use multiple supporting memories
- More extensive knowledge databases
- More realistic questions and answers
- Perhaps perform actions instead of answers

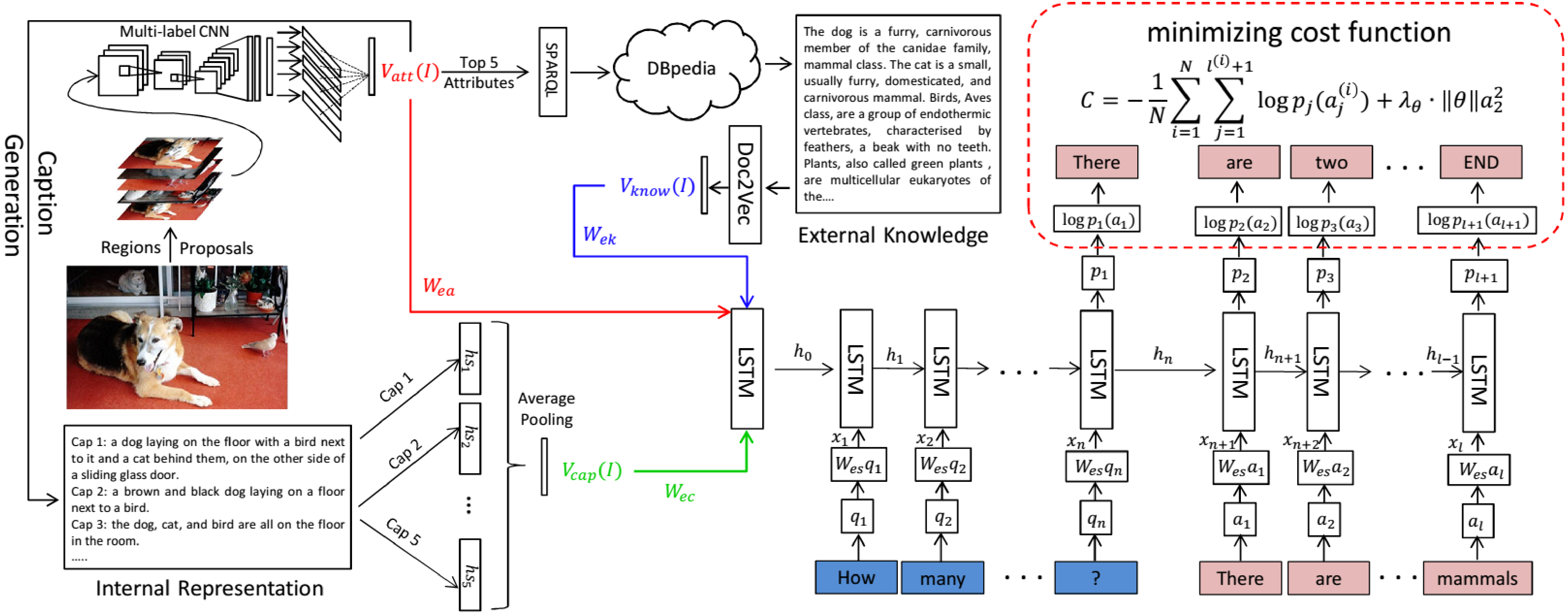
Papers in the literature

- *Neural Turing Machines*, A. Graves, G. Wayne, I. Danihelka, arXiv 2014
  - http://arxiv.org/abs/1410.5401
- *Memory Networks*, J. Weston, S. Chopra, A. Bordes, arXiv 2014
  - http://arxiv.org/abs/1410.3916
- *End-to-end Memory Networks*, S. Sukhbaatar, A. Szlam, J. Weston, R. Fergus, arXiv 2015
  - http://arxiv.org/abs/1503.08895
- *Ask Me Anything: Free-form Visual Question Answering Based on Knowledge from External Sources,* Q. Wu, P. Wang, C. Shen, A. van den Hengel, A. Dick, arXiv 2015
  - http://arxiv.org/abs/1511.06973

Papers in the literature

- *Neural Turing Machines*, A. Graves, G. Wayne, I. Danihelka, arXiv 2014
  - http://arxiv.org/abs/1410.5401
- *Memory Networks*, J. Weston, S. Chopra, A. Bordes, arXiv 2014
  - http://arxiv.org/abs/1410.3916
- *End-to-end Memory Networks*, S. Sukhbaatar, A. Szlam, J. Weston, R. Fergus, arXiv 2015
  - http://arxiv.org/abs/1503.08895
- *Ask Me Anything: Free-form Visual Question Answering Based on Knowledge from External Sources,* Q. Wu, P. Wang, C. Shen, A. van den Hengel, A. Dick, arXiv 2015
  - http://arxiv.org/abs/1511.06973

External databases for visual questions



The dog is a furry, carnivorous member of the canidae family, mammal class. The cat is a small, usually furry, domesticated, and carnivorous mammal. Birds, Aves class, are a group of endothermic vertebrates, characterised by feathers, a beak with no teeth. Plants, also called green plants , are multicellular eukaryotes of the....

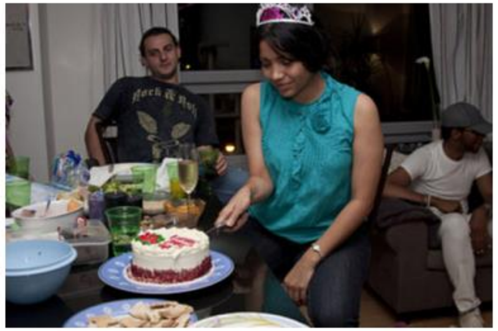
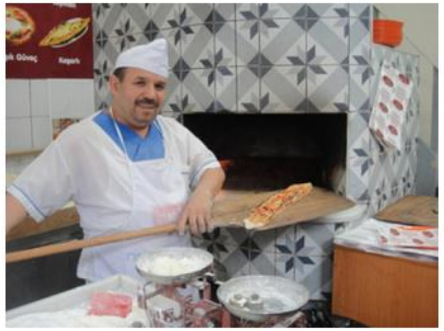
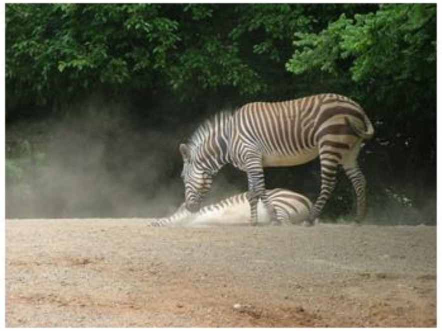
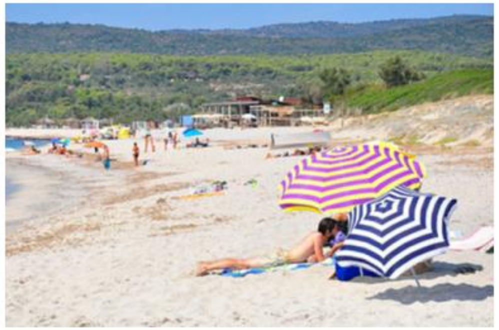
Cap 1: a dog laying on the floor with a bird next to it and a cat behind them, on the other side of a sliding glass door.
Cap 2: a brown and black dog laying on a floor next to a bird.
Cap 3: the dog, cat, and bird are all on the floor in the room.
.....

$$C = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{l^{(i)}+1}\log p_j(a_j^{(i)}) + \lambda_\theta \cdot \|\theta\|_2^2$$

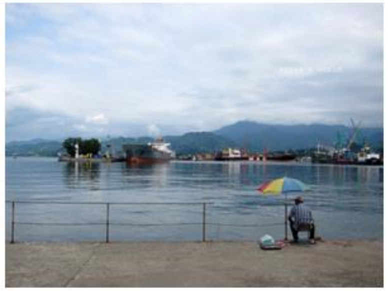
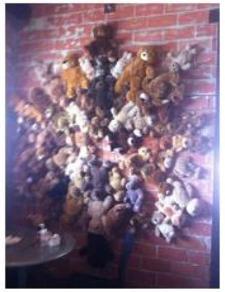
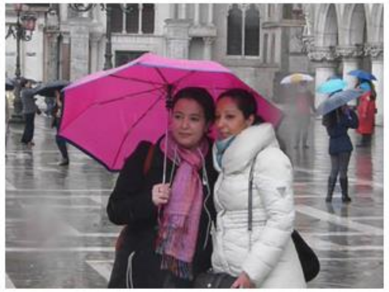
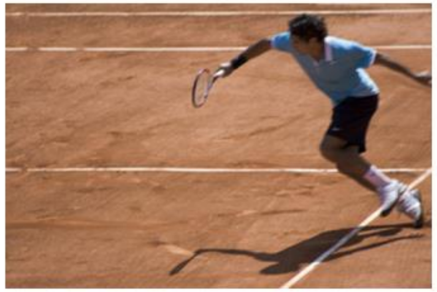
Method

- Classifying objects, attributes inside an image
  - Use a very deep, VGG-16 network fine-tuned on the MSCOCO image attributes
- Caption-based image representations
  - Use an LSTM
  - Start with an image
  - Generate a caption
  - Use the hidden state h_T of the final step as a representation
- Relate to external database
  - DBpedia
  - SQL-like queries using SPARQL
  - Represent returned text with Doc2Vec
- Combine everything and end-to-end learning
  - $x = [W_{ea} x_{att}(I),\ W_{ec} x_{cap}(I), W_{ek} x_{know}(I),]$

Results

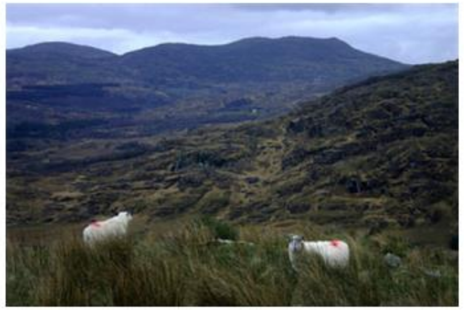
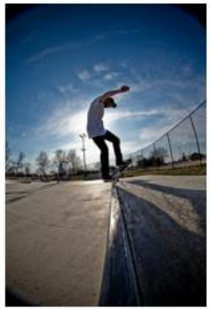
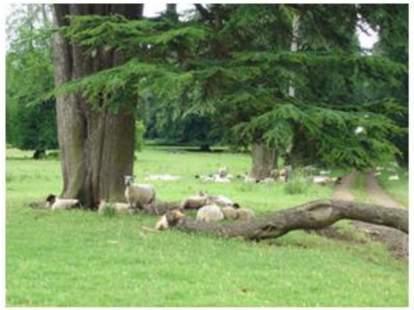
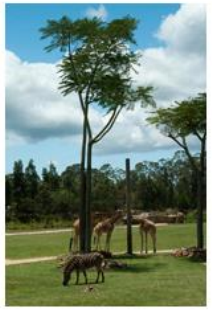





	Why is she wearing a crown?	Why is he smiling?	Why is the zebra on the ground?	Why do they have umbrellas?
Ours:	birthday	happy	resting	shade
Vgg+LSTM:	to eat	unknown	eat	raining
Ground Truth:	birthday	happy	resting	shade

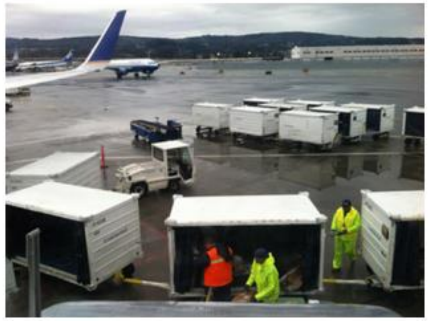
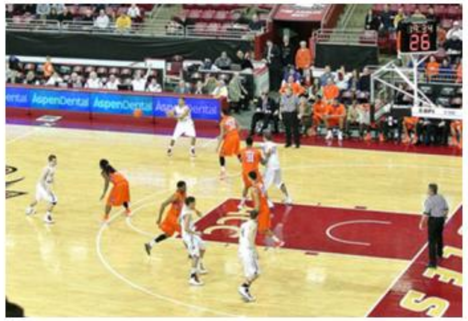
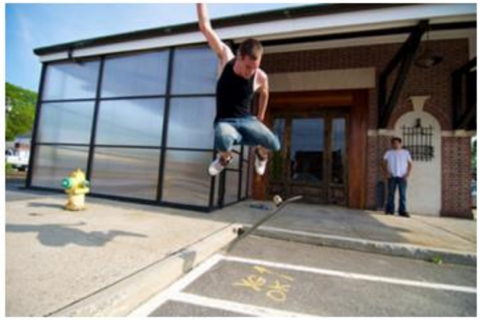
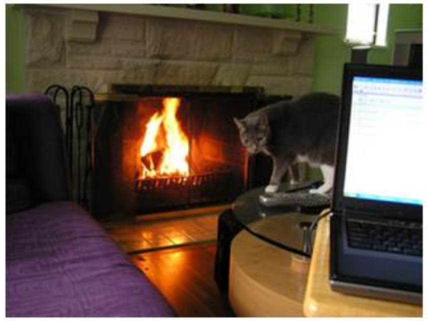





	Why is a man sitting under an umbrella?	Why are there animals pinned to the wall?	Why do they have umbrellas?	Why is he swinging backhand?
Ours:	shade	decoration	raining	to hit ball
Vgg+LSTM:	safety	teddy	yes	tennis ball
Ground Truth:	shade	decoration	raining	to hit ball

More results

	Why do these sheep have paint on them?	Why is his arm outflung?	Why are the animals laying here?	Why are all the giraffes gathered together?
Ours:	identification	balance	resting	eating
Vgg+LSTM:	to eat	to play	no	to play
Ground Truth:	identification	balance	resting	eating

	Why are they wearing such bright colors?	Why are the men wearing orange?	Why is the man jumping?	Why is this room warm?
Ours:	safety	team	skateboarding	fireplace
Vgg+LSTM:	yes	to	unknown	to sleep
Ground Truth:	safety	team	skateboarding	fireplace

Summary

- Memory networks
- Difficulties with modelling memory
- Memory networks for image-language reasoning

Next lecture

- Student presentations of Deep Learning papers