Deep Learning for Natural Language Processing

Question Answering

Karl Moritz Hermann

kmh@google.com

DeepMind

28 Feb 2017

Questions

Question
When were the first pyramids built?
Jean-Claude Juncker
How old is Keir Starmer?
What is the current price for AAPL?
What's the weather like in London?
Whom did Juncker meet with?
When did you get to this lecture?
Why do we yawn?

Questions

Question	Answer
When were the first pyramids built?	*2630 BC*
Jean-Claude Juncker	*Jean-Claude Juncker is a Luxembourgish politician. Since 2014, Juncker has been President of the European Commission.*
How old is Keir Starmer?	*54 years*
What is the current price for AAPL?	*136.50 USD*
What's the weather like in London?	*7 degrees Celsius. Clear with some clouds.*
Whom did Juncker meet with?	*The European Commission president was speaking after meeting with Irish Taoiseach Enda Kenny in Brussels.*
When did you get to this lecture?	*Five minutes after it started.*
Why do we yawn?	*When we're bored or tired we don't breathe as deeply as we normally do. This causes a drop in our blood-oxygen levels and yawning helps us counter-balance that.*

Why do we care about QA?

Because QA is awesome

1. QA is an AI-complete problem.
   If we solve QA, we have solved every other problem, too.
2. Many immediate and obvious applications
   Search, dialogue, information extraction, summarisation, ...
3. Some pretty nice results already
   IBM Watson and Jeopardy!, Siri, Google Search ...
4. Lots left to do!
   Plenty of interesting research and hard problems as well as low-hanging fruit.

Questions (again)

Question
When were the first pyramids built?
Jean-Claude Juncker
How old is Keir Starmer?
What is the current price for AAPL?
What's the weather like in London?
Whom did Juncker meet with?
When did you get to this lecture?
Why do we yawn?

Questions (again)

Question	Answer Source
When were the first pyramids built?	*Encyclopedia*
Jean-Claude Juncker	*Recent encyclopedia / Wikipedia*
How old is Keir Starmer?	*Very recent encyclopedia (i.e. Wikipedia). Extrapolate from date of birth.*
What is the current price for AAPL?	*NASDAQ Ticker*
What's the weather like in London?	*MET Office*
Whom did Juncker meet with?	*The Independent article "Jean-Claude Juncker doesn't want Northern Ireland and Republic to have post-Brexit hard border"*
When did you get to this lecture?	*Personal observation*
Why do we yawn?	*Various studies on the matter.*

Question Answering depends on three kinds of data

And this gives us a good system for thinking about various QA tasks

Question

Factual questions*

Complex/narrative questions

Information Retrieval

Library Reference

Context/Source

Sets of documents (corpus)

A single document

Knowledge Base

Non-linguistic types of data (GPS, images, sensors, ...)

Answer

A single fact

An explanation

A document

A sentence or paragraph extracted from somewhere

An image or other type of object

Another question

* *Factual question* $\sim$ *factoid question*, but that doesn't work in B.E.

Question Taxonomy

Many possible taxonomies for questions [1]

- Wh- words
- Subject of question
- The form of expected answers
- Types of sources from which answers may be drawn

For the purposes of building QA systems it is useful to start by considering the sources an answer may be drawn from.

Focus on the answer rather than the question.

[1] See e.g. Pomerantz (2005), A Linguistic Analysis of Question Taxonomies

QA Taxonomy Discovery

Three Questions for building a QA System

- What do the answers look like?
- Where can I get the answers from?
- What does my training data look like?

By the end of this lecture, being able to answer these questions should allow you to devise a QA system for any given task.

Areas in Question Answering

Reading Comprehension	• Answer based on a document • Context is a specific document
Semantic Parsing	• Answer is a logical form, possible executed against a KB • Context is a Knowledge Base
Visual QA	• Answer is simple and factual • Context is one/multiple image(s)
Information Retrieval	• Answer is a document/paragraph/sentence • Context is a corpus of documents
Library Reference	• Answer is another question • Context is the structured knowledge available in the library and the librarians view of it.

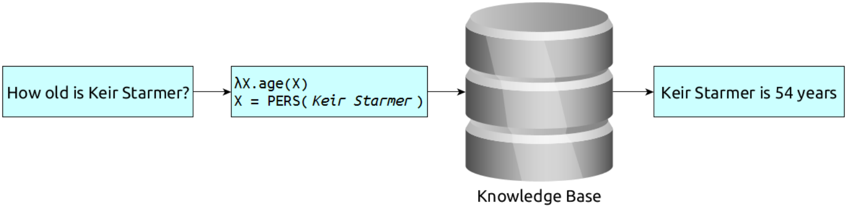
Remainder of this lecture

For the remainder of this lecture, we will cover four areas of question answering in more depth.

- Semantic Parsing
- Answer Sentence Selection
- Reading Comprehension
- Visual Question Answering

Semantic Parsing

Semantic Parsing is the process of mapping natural language into a formal representation of its meaning. Depending on the chosen formalism this logical representation can be used to query a structured knowledge base.



Question → Logical Form → KB Query → Answer

Semantic Parsing is Question→Logical Form.
We (often mistakenly) then assume that LF→Answer is trivial.

Knowledge Bases for QA with Semantic Parsing

Knowledge bases typically represent their data as triples
(married-to, Michelle Obama, Barack Obama)
(member-of, United Kingdom, European Union)
Generally: *(relation, entity1, entity2)*

There are several (large) databases freely available to use, e.g.:

Freebase 1.9 billion triples on general knowledge. Defunct as of 2016 and replaced by Google Knowledge Graph

WikiData Information on 25 million entities

OpenStreetMap 3 billion triples on geography

GeoQuery 700 facts about US geography. Tiny dataset, but frequently used in semantic parsing work.

KBs are cheap — Supervised Data is expensive!

Free917[2]	917 freebase annotated questions
GeoQuery[3]	880 questions on US geography
NLMaps[4]	2,380 natural language queries on the OSM data

These kinds of datasets are incredibly expensive to create as they require experts for the manual annotation process, who are trained in using a given database schema:

"Where are kindergartens in Hamburg?"

```
query(area(keyval(name,Hamburg)),
      nwr(keyval(amenity,kindergarten)),
      qtype(latlong))
```

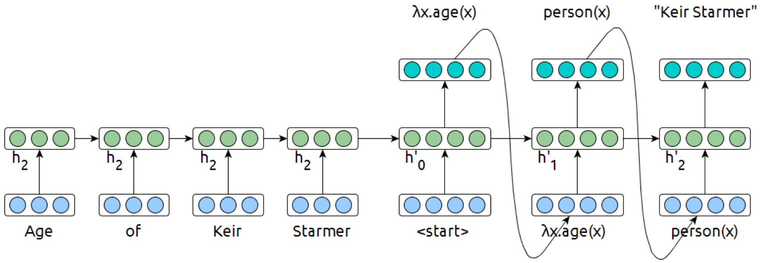
[2] Cai and Yates, 2013

[3] Zelle and Mooney, 1996

[4] Hass and Riezler, 2016

A Deep Learning Approach to Semantic Parsing

Semantic parsing can be viewed as a sequence to sequence model, not unlike machine translation.



Details

✔ Encode sentence with sequence models

✔ Decode with standard mechanisms from MT

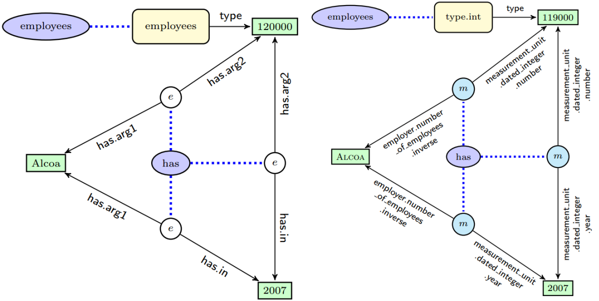
✘ Supervised training data hard to come by

✘ Depending on formalism used, highly complex target side

✘ How to deal with proper nouns and numbers?

One Solution to Sparsity: Avoid Logical Forms

Semantic parsing frequently reduce the reliance on supervised data (language-logical form) by exploiting other types of data such as question-answer pairs or corpora of questions only.



has.arg1(e, Alcoa) $\wedge$ has.arg2(e, 120000) $\wedge$ has.in(e, 2007) $\wedge$ employees(120000)

(a) Ungrounded Graph

employer.number_of_employees.inverse(m, ALCOA) $\wedge$ measurement_unit.dated_integer.number(m, 119000) $\wedge$ measurement_unit.dated_integer.year(m, 2007) $\wedge$ type.int(119000)

(b) Grounded Graph

Berant et al. (2013): Semantic Parsing on Freebase from QA Pairs
Reddy et al. (2014): Large-scale Semantic Parsing without QA Pairs

Improved Neural Semantic Parsing

We can apply the same idea to neural semantic parsing, and further take mechanisms from machine translation to improve performance and data efficiency:

- Like in MT, using attention can be helpful

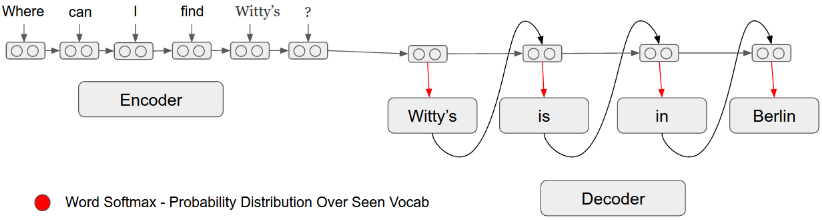
  Dong and Lapata (2016): Language to Logical Form with Neural Attention
- Exploit the highly rigid structure in the target side to constrain generation

  Liang et al. (2016): Neural Symbolic Machines

  Ling et al. (2016): Latent predictor networks for code generation
- Make use of semi-supervised training to counter sparsity

  Kocisky et al. (2016): Semantic Parsing with Semi-Supervised Sequential Autoencoders

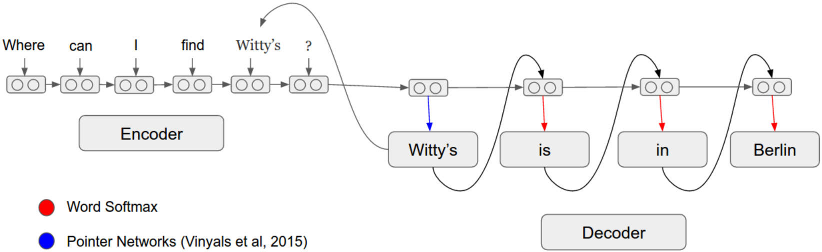
Generation with multiple sources



● Word Softmax - Probability Distribution Over Seen Vocab

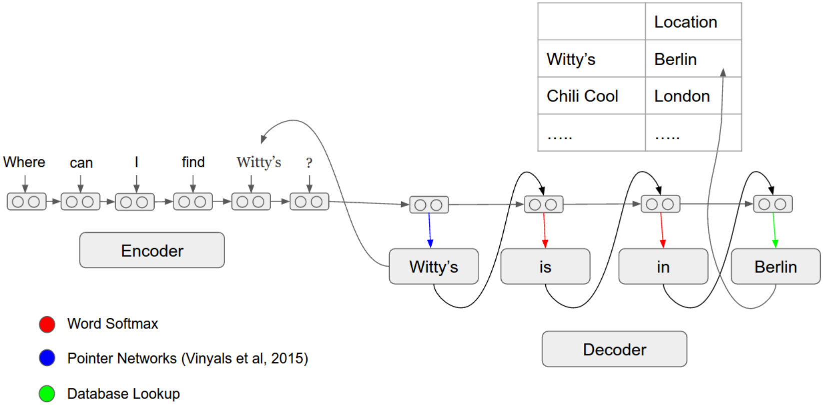
Ling et al. (2016): Latent predictor networks for code generation

Generation with multiple sources



Ling et al. (2016): Latent predictor networks for code generation

Generation with multiple sources



	Location
Witty's	Berlin
Chili Cool	London
…..	…..

🔴 Word Softmax

🔵 Pointer Networks (Vinyals et al, 2015)

🟢 Database Lookup

Ling et al. (2016): Latent predictor networks for code generation

Semantic Parsing Summary

✔ LF instead of answer makes system robust
✔ Answer independent of question and parsing mechanism
✔ Can deal with rapidly changing information

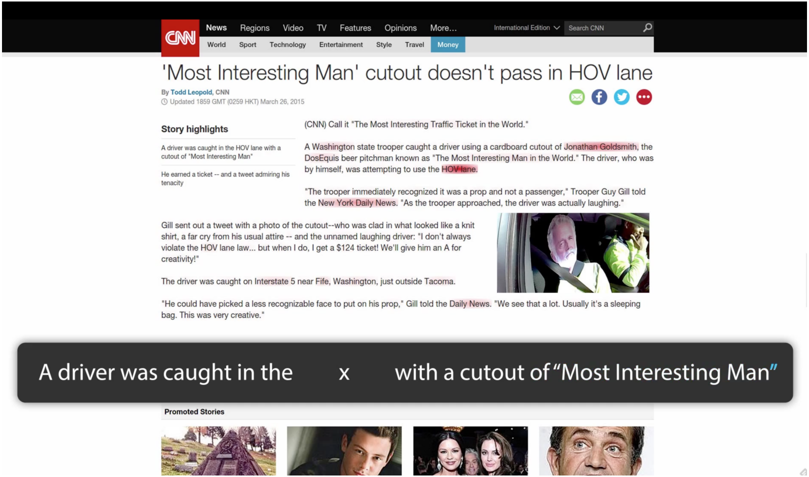
✘ Constrained to queriable questions in DB schema
✘ No database is large enough
✘ Training data hard to find

✔ When were the pyramids built?
? Jean-Claude Juncker
✔ How old is Keir Starmer?
✔ What is the price for AAPL?
✔ What's the weather in London?
✘ Whom did Juncker meet with?
✘ When did you get here?
✘ Why do we yawn?

Caveat: Each of these examples requires a different underlying KB!

Reading Comprehension

Answer a question related to a given document



'Most Interesting Man' cutout doesn't pass in HOV lane

By Todd Leopold, CNN
Updated 1859 GMT (0259 HKT) March 26, 2015

Story highlights

A driver was caught in the HOV lane with a cutout of "Most Interesting Man"

He earned a ticket -- and a tweet admiring his tenacity

(CNN) Call it "The Most Interesting Traffic Ticket in the World."

A Washington state trooper caught a driver using a cardboard cutout of Jonathan Goldsmith, the DosEquis beer pitchman known as "The Most Interesting Man in the World." The driver, who was by himself, was attempting to use the HOV lane.

"The trooper immediately recognized it was a prop and not a passenger," Trooper Guy Gill told the New York Daily News. "As the trooper approached, the driver was actually laughing."

Gill sent out a tweet with a photo of the cutout--who was clad in what looked like a knit shirt, a far cry from his usual attire -- and the unnamed laughing driver: "I don't always violate the HOV lane law... but when I do, I get a $124 ticket! We'll give him an A for creativity!"

The driver was caught on Interstate 5 near Fife, Washington, just outside Tacoma.

"He could have picked a less recognizable face to put on his prop," Gill told the Daily News. "We see that a lot. Usually it's a sleeping bag. This was very creative."

A driver was caught in the x with a cutout of "Most Interesting Man"

Corpora for Reading Comprehension

CNN/DailyMail[5] Over 1 million cloze form QA pairs with articles from CNN and Mail online for context. Pick an anonymised entity.

CBT[6] 700k QA pairs, children's books as context. Pick one of 10 candidates.

SQuAD[7] 100k manual QA pairs with 500 Wikipedia articles for context. Answer is a span.

Assumptions made in all of the above tasks

- Context is read on the fly and unknown during training phase
- Answer is contained in the context as a single word or span
- *This constraint does not hold for reading comprehension in general!*

[5] Hermann et al., 2015

[6] Hill et al., 2015

[7] Rajpurkar et al., 2016

CNN/DailyMail Dataset Example

CNN article

Document 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 broadcaster found he had subjected producer Oisin Tymon “to an unprovoked physical and verbal attack.” . . .

Query Producer **X** will not press charges against Jeremy Clarkson, his lawyer says.

Answer Oisin Tymon

We formulate *Cloze* style queries from the story paraphrases.

OOV and proper nouns are dealt with by replacing all entities with anonymised markers. This greatly reduces the vocabulary size.

CNN/DailyMail Dataset Example

CNN article

Document 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 *ent212* , his lawyer says .

Answer *ent193*

We formulate *Cloze* style queries from the story paraphrases.

OOV and proper nouns are dealt with by replacing all entities with anonymised markers. This greatly reduces the vocabulary size.

A Generic Neural Model for Reading Comprehension

Given context d and question q, the probability of an answer a can be represented as:

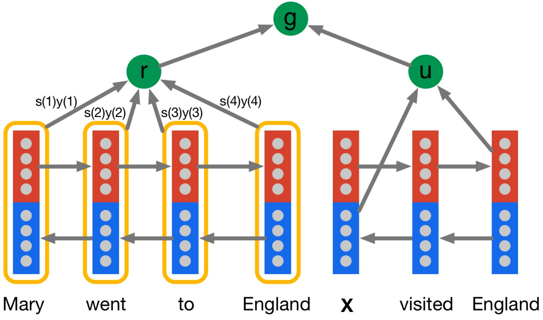
$$p(a|q,d) \propto \exp(W(a)g(q,d)), \quad \text{s.t.} a \in V$$

Fill in the details

✔ Encode question and context with sequence models
✔ Combine q and c with an MLP or attention
✔ Select answer from attention map, by using a classifier, or with generative setup

✘ How to deal with out of vocabulary (OOV) terms?
✘ How to deal with proper nouns and numbers?

Reading Comprehension with Attention



Read (encode) context document and question

Use question to attend to context

Use joint representation to generate answer

- Predict based on attention map
- Generate conditioned on joint representation
- Classify over set of candidate answers

Reading Comprehension with Attention

Denote the outputs of a bidirectional LSTM as $\overrightarrow{y}(t)$ and $\overleftarrow{y}(t)$. Form two encodings, one for the query and one for each token in the document,

$$u = \overrightarrow{y_q}(|q|) \mid\mid \overleftarrow{y_q}(1), \qquad y_d(t) = \overrightarrow{y_d}(t) \mid\mid \overleftarrow{y_d}(t).$$

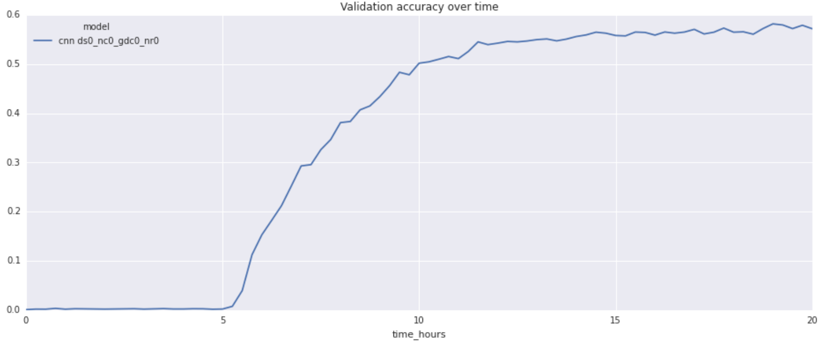
The representation r of the document d is formed by a weighted sum of the token vectors. The weights are interpreted as the model's attention,

$$\begin{aligned} m(t) &= \tanh\left(W_{ym} y_d(t) + W_{um} u\right), \\ s(t) &\propto \exp\left(\mathrm{w}_{ms}^{\intercal} m(t)\right), \\ r &= y_d s. \end{aligned}$$

Define the joint document and query embedding via a non-linear combination:

$$g^{\mathrm{AR}}(d, q) = \tanh\left(W_{rg} r + W_{ug} u\right).$$

Attentive Reader Training



Models were trained using asynchronous minibatch stochastic gradient descent (RMSProp) on approximately 25 GPUs.

Attention Sum Reader

The model can be modified to make use of the fact that the answer is a word from the context document. Now we calculate the probability of the answer being in position i of the context[8]:

$$p(i|q,d) \propto \exp(f_i(d) \cdot g(q))$$

Positional probabilities can then be summed to form token-based probabilities:

$$P(w|q,d) \propto \sum_{i(w,d)} P(i|q,d)$$

The rest of the model is equivalent to the attentive reader model presented before.

[8] Kadlec et al. (2016), Text Understanding with the Attention Sum Reader Network

Reading Comprehension Summary

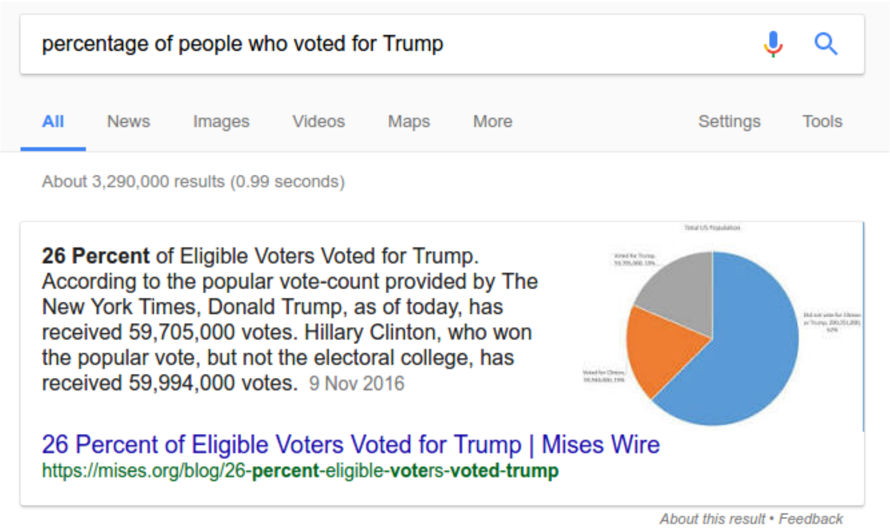
✔ Ask questions in context
✔ Easily used in discriminative and generative fashion
✔ Large datasets available

✘ Constraint on context often artificial
✘ Many types of questions unanswerable

? When were the pyramids built?
? Jean-Claude Juncker
? How old is Keir Starmer?
? What is the price for AAPL?
? What's the weather in London?
✔ Whom did Juncker meet with?
✘ When did you get here?
✘ Why do we yawn?

Caveat: Need context for any of these, and incredibly up-to-date context for some of these.

Answer Sentence Selection



percentage of people who voted for Trump
All
News
Images
Videos
Maps
More
Settings
Tools

About 3,290,000 results (0.99 seconds)

26 Percent of Eligible Voters Voted for Trump. According to the popular vote-count provided by The New York Times, Donald Trump, as of today, has received 59,705,000 votes. Hillary Clinton, who won the popular vote, but not the electoral college, has received 59,994,000 votes. 9 Nov 2016

Total US Population
Voted for Trump, 59,705,000, 19%
Did not vote for Clinton or Trump, 200,251,000, 62%
Voted for Clinton, 59,944,000, 19%

26 Percent of Eligible Voters Voted for Trump | Mises Wire

https://mises.org/blog/26-**percent**-eligible-**voters**-**voted**-**trump**

Answer Sentence Selection

Answer Sentence Selection describes the task of picking a suitable sentence from a corpus that can be used to answer a question.

Questions Factual questions, possibly with context

Data Source “The Web” or the output of some IR system

Answer One or several excerpts pertinent to the answer

The answer is guaranteed to be extracted, while in reading comprehension it could be either generated or extracted.

Corpora for Answer Sentence Selection

TREC QA track (8-13) Several hundred manually-annotated question answer pairs with around 20 candidates per instance.

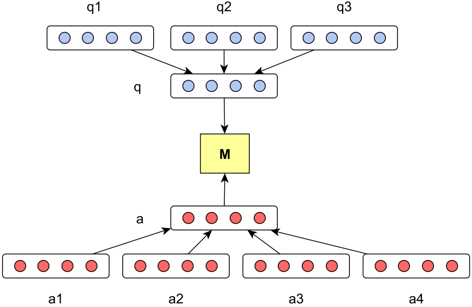
MS MARCO 100k question-answer pairs with 10 contextual passages each. Can also be used as a QA dataset for reading comprehension.

Likewise, answer sentence selection plays a role in any information retrieval setup, and datasets from IR and other QA tasks can easily be converted into answer selection style datasets.

A Neural Model for Answer Sentence Selection

We need to compute the probability of an answer candidate a and a question q matching. Note that this is different from the previous task as we now calculate that score independently of all other candidates:

$$p(y = 1|q, a) = \sigma(q^T Ma + b)$$



Yu et al., 2014

Evaluation

Unlike single entity style QA where we can use a simple accuracy measure, tasks such as answer sentence selection require more specialised metrics for evaluating model performance.

Measure	Description	Formula
Accuracy	Binary measure	$\#true/\#total$
Mean Reciprocal Rank	Measures position of first relevant document in return set.	$\frac{1}{\lvert Q\rvert}\sum_{i=1}^{\lvert Q\rvert}\frac{1}{\text{rank}_i}$
BLEU Score	Machine Translation measure for translation accuracy	*complicated*

Answer Selection Summary

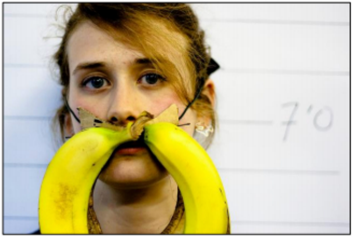
✔ Designed to deal with large amounts of context
✔ More robust than ‘true’ QA systems as it turns provides context with its answers
✔ Obvious pipeline step between IR and QA

✘ Does not provide answers, provides context only
✘ Real-world use depends on underlying IR pipeline

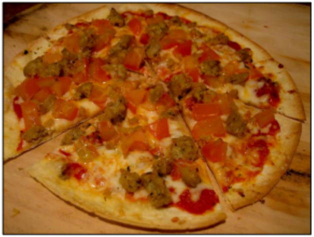
✔ When were the pyramids built?
✔ Jean-Claude Juncker
✘ How old is Keir Starmer?
✘ What is the price for AAPL?
✘ What’s the weather in London?
? Whom did Juncker meet with?
✘ When did you get here?
✘ Why do we yawn?

Note: Things like age or stock price may produce answers, but with no guarantee of accuracy (any mention of any AAPL price might be a good fit).

Visual Question Answering



What color are her eyes?
What is the mustache made of?



How many slices of pizza are there?
Is this a vegetarian pizza?

Sometimes questions require context outside of pure language.

Visual QA: Task and Corpora

In recent years a number of visual QA datasets have sprung up. Some of the more popular ones include:

VisualQA	Agrawal et al. (2015)
VQA 2.0	Goyal et al. (2016)
COCO-QA	Ren et al. (2015)

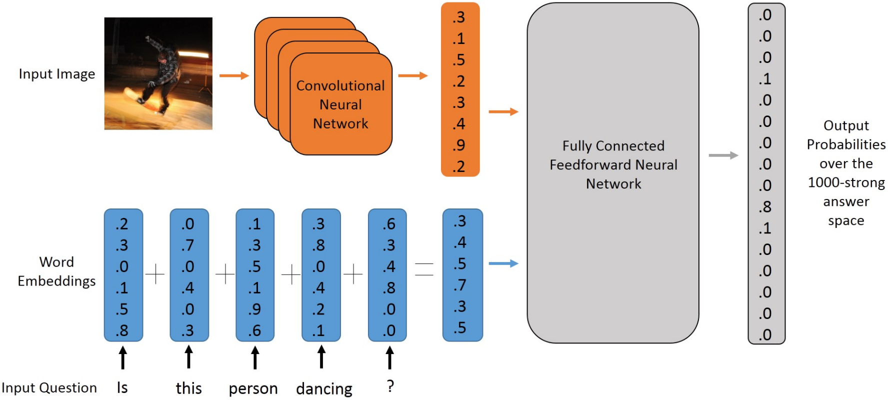
Details between these datasets vary, but the basic organisation remains the same of images paired with simple questions and answers (either free form or from a list of options).

All of these are reasonably large (100ks of images, over 1M questions).

Visual QA is quite straight-forward

- Question is language → some encoder
- Context is a single picture → convolutional network
- Answer is a single word → classifier function

We have covered all the components already:



Input Image

Convolutional Neural Network

Fully Connected Feedforward Neural Network

Output Probabilities over the 1000-strong answer space

Word Embeddings

Input Question

Is

this

person

dancing

?

Blind Model

Ignoring the images is a good baseline!

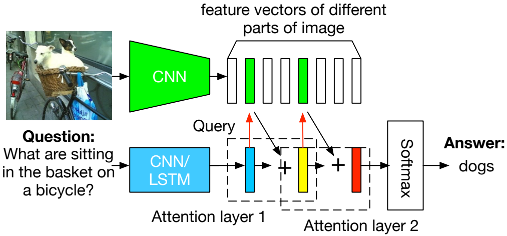
- What colour is the cat?
- How many chairs are around the table?
- What furniture is in the bedroom?
- Where is the person sleeping?

We can get reasonably good guesses in at many of these questions without seeing an image for context. See Goyal et al. (2016)

Attention Methods for Visual QA

Viewing VQA from the perspective of our default QA paradigm, there is significant overlap with reading comprehension style models. We use similar techniques to improve performance.

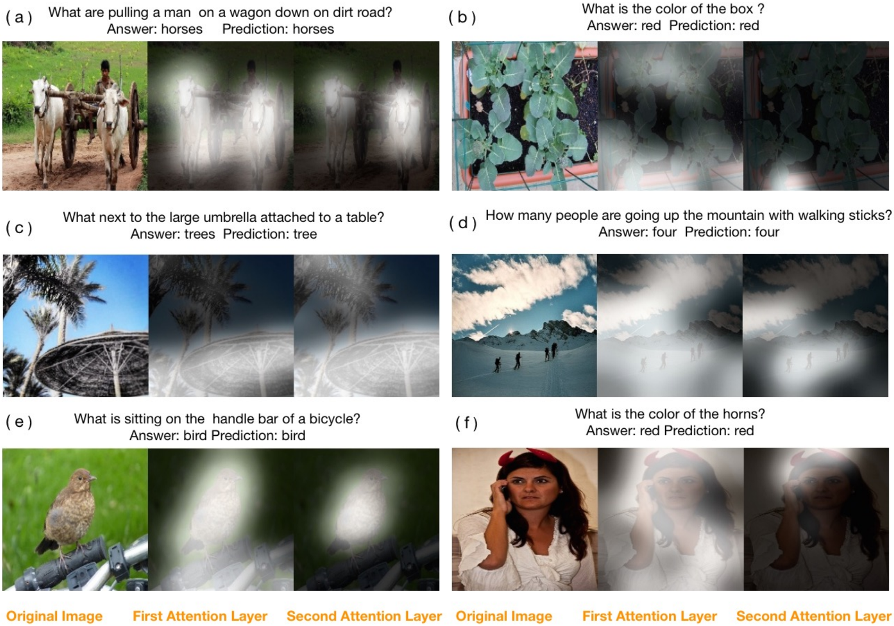
We can use attention on visual representations:



feature vectors of different parts of image
CNN
Query
Question:
What are sitting in the basket on a bicycle?
CNN/ LSTM
+
+
Softmax
Answer:
dogs
Attention layer 1
Attention layer 2

Yang et al. (2015): Stacked Attention Networks for Image Question Answering

Attention Methods for Visual QA



(a) What are pulling a man on a wagon down on dirt road? Answer: horses Prediction: horses

(b) What is the color of the box ? Answer: red Prediction: red

(c) What next to the large umbrella attached to a table? Answer: trees Prediction: tree

(d) How many people are going up the mountain with walking sticks? Answer: four Prediction: four

(e) What is sitting on the handle bar of a bicycle? Answer: bird Prediction: bird

(f) What is the color of the horns? Answer: red Prediction: red

VIisual Question Answering Summary

✔ Extra modality ‘for free’
✔ Plenty of training data available as of recently

✘ Currently quite gimmicky
✘ Still a long way to go

✘ When were the pyramids built?
✘ Jean-Claude Juncker
✘ How old is Keir Starmer?
✘ What is the price for AAPL?
? What’s the weather in London?
✘ Whom did Juncker meet with?
✘ When did you get here?
✘ Why do we yawn?

Summary (How to build your own QA system)

Build a QA model in seven questions

- What is the task?
- What do question, answer and context look like?
- Where does the data come from?
- Can you augment the data?
- How to encode question and context?
- How to combine question and context?
- How to predict or generate an answer?

There are plenty of open questions left in QA.
Just remember to start with the data!

End

Sources and Further Reading

Question Answering Theory and Datasets

Pomerantz (2005), A Linguistic Analysis of Question Taxonomies
Nguyen et al. (2016), MS MARCO: A Human Generated Machine Reading Comprehension Dataset
Haas and Riezler (2016), A Corpus and Semantic Parser for Multilingual Natural Language Querying of OpenStreetMap

Semantic Parsing

Artzi et al. (2013), Semantic Parsing with CCG
Berant et al. (2013), Semantic Parsing on Freebase from Question-Answer Pairs
`http://nlp.stanford.edu/software/sempre/`

Reading Comprehension

Hermann et al. (2015), Teaching Machines to Read and Comprehend
Kadlec et al. (2016), Text Understanding with the Attention Sum Reader Network

Visual QA

Yang et al. (2015), Stacked Attention Networks for Image Question Answering
Ren et al. (2015), Exploring Models and Data for Image Question Answering
Goyal et al. (2016), Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering.

`https://avisingh599.github.io/deeplearning/visual-qa/`