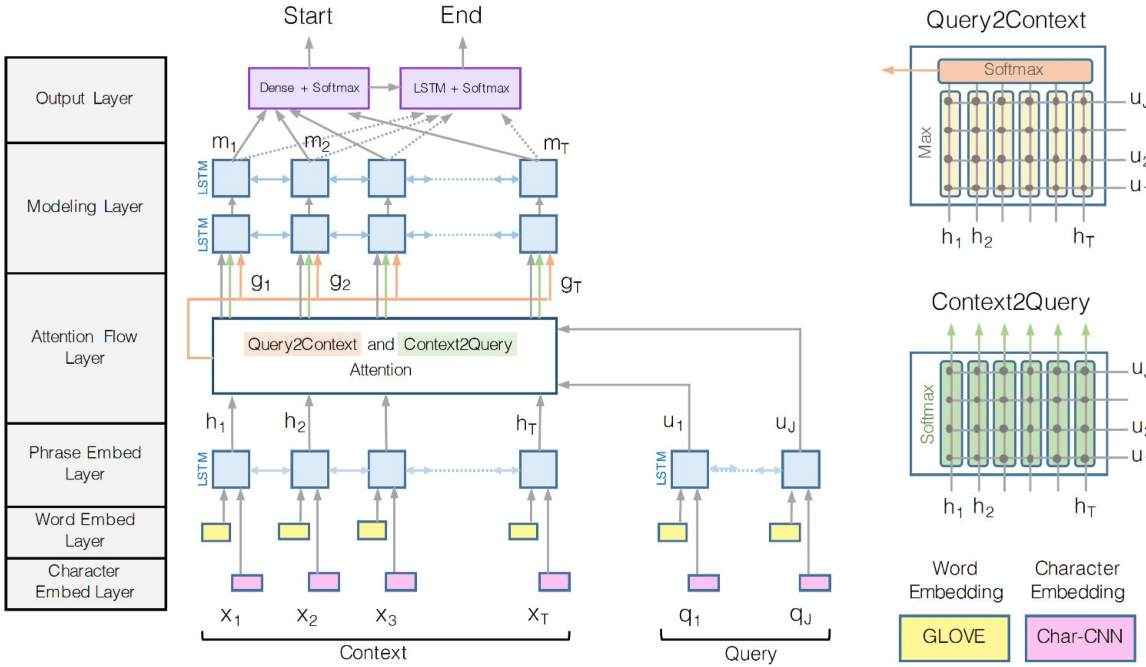
Machine comprehension with neural nets - chinese rooms?



Bi-directional attention flow for machine comprehension, M. Seo, A. Kembhavi, A. Farhadi, H. Hajishirzi, ICLR 2017

Leaderboard

Since the release of our dataset, the community has made rapid progress! Here are the ExactMatch (EM) and F1 scores of the best models evaluated on the test and development sets of v1.1. Will your model outperform humans on the QA task?

Rank	Model	EM	F1
1 Mar 2017	r-net (ensemble) *Microsoft Research Asia* http://aka.ms/rnet	76.922	84.006
2 Jun 2017	Interactive AoA Reader (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	76.492	83.745
3 May 2017	MEMEN (ensemble) *Eigen Technology & Zhejiang University*	75.370	82.658
4 Mar 2017	ReasoNet (ensemble) *MSR Redmond* https://arxiv.org/abs/1609.05284	75.034	82.552
5 May 2017	r-net (single model) *Microsoft Research Asia* http://aka.ms/rnet	74.614	82.458
6 May 2017	Mnemonic Reader (ensemble) *NUDT & Fudan University* https://arxiv.org/abs/1705.02798	73.754	81.863
7 Apr 2017	SEDT+BiDAF (ensemble) *CMU* https://arxiv.org/abs/1703.00572	73.723	81.530
7 Feb 2017	BiDAF (ensemble) *Allen Institute for AI & University of Washington* https://arxiv.org/abs/1611.01603	73.744	81.525
8 May 2017	jNet (ensemble) *USTC & National Research Council Canada & York University* https://arxiv.org/abs/1703.04617	73.010	81.517
8 Jan 2017	Multi-Perspective Matching (ensemble) *IBM Research* https://arxiv.org/abs/1612.04211	73.765	81.257
9 Apr 2017	T-gating (ensemble) *Peking University*	72.758	81.001

What is understanding?

SQuAD   Home   Explore

Tesla was born on 10 July [O.S. 28 June] 1856 into a Serb family in the village of Smiljan, Austrian Empire (modern-day Croatia). His father, Milutin Tesla, was a Serbian Orthodox priest. Tesla's mother, Đuka Tesla (née Mandić), whose father was also an Orthodox priest,:10 had a talent for making home craft tools, mechanical appliances, and the ability to memorize Serbian epic poems. Đuka had never received a formal education. Nikola credited his eidetic memory and creative abilities to his mother's genetics and influence. Tesla's progenitors were from western Serbia, near Montenegro.:12

What modern-day country was Tesla born in?
Ground Truth Answers: Croatia | Croatia | Croatia

What was the occupation of Tesla's father?
Ground Truth Answers: priest | priest | Serbian Orthodox priest

What was special about Tesla's memory?
Ground Truth Answers: eidetic | eidetic | eidetic

Who did Tesla credit for his abilities?
Ground Truth Answers: his mother's genetics | his mother | his mother

What was Tesla's fathers occupation?
Ground Truth Answers: priest | priest | Serbian Orthodox priest

What was Tesla's father's name?
Ground Truth Answers: Milutin Tesla | Milutin Tesla | Milutin Tesla

What steps you need to follow, to answer?

Do you need to understand?

Can you just point the answer? Is that understanding?

Does it matter if you know the language?

Or know how to read?

Does a “chinese room” understand, or not?

Any person in this room, could respond with chinese letters.

Based only on rules and data (lots of them).

The room as a whole acts as an intelligent chinese person.

Actually there is no understanding here, just a mechanistic symbol operation.

http://philosophy.hku.hk/joelau/media/chinese-room.jpg

Can chinese rooms perform any computation?

In theory, it is a device with von Neuman architecture.

It is Turing - complete.

Person = CPU
Book with instructions = program
File cabiner = memory
Pencil and eraser= write to mem

http://philosophy.hku.hk/joelau/media/chinese-room.jpg

Does a “chinese room” understand, or not?

The argument was first presented by philosopher John Searle in his paper, "Minds, Brains, and Programs", published in Behavioral and Brain Sciences in 1980.

- Wikipedia

John Searle

In essence, it is a critique of **“strong AI”** being possible, or that it it possible to attribute mental states, neurocorrelates of mind, to a program.

Also, that the **Turing test** is an incorrect test, because such a “chinese room” could trick the interviewer, provided with billions of conditional answers.

Very plausible, but met with extreme responses.

Actually, we still do not know how the brain works, or if humans “understand”, or are very complex, parallel neural processing units based on non-linear chemical “transistors”.

Critique of the chinese room argument

Lots of counter arguments have been made, each concerning certain parts of room, or allegedly a misundestanding of concepts on Searles's side.

Replies to Searle's argument may be classified according to what they claim to show:

Those which identify who speaks Chinese
Those which demonstrate how meaningless symbols can become meaningful
Those which suggest that the Chinese room should be redesigned in some way
Those which contend that Searle's argument is misleading
Those which argue that the argument makes false assumptions about subjective conscious experience and therefore proves nothing

Some of the arguments (robot and brain simulation, for example) fall into multiple categories.

Critique of the chinese room argument - from Turing

Turing actually thought about the same thing 30 years earlier. (called argument “from consiousness”)
Turing points out:

Chinese room argument = “problem of other minds” applied to machines

Nils Nilsson : “(...) For all I know, Seale may only be behaving as if he were thinking deeply.”

Turing: “people never consider the problem of other minds when dealing with each other”

“instead of arguing continually over this point it is usual to have the polite convention that everyone thinks.”

Stuart Russel, Peter Norvig : “Searle must be mistaken about the "knowability of the mental", and in his belief that there are "causal properties" in our neurons that give rise to the mind. They point out that, by Searle's own description, these causal properties can't be detected by anyone outside the mind, otherwise the Chinese Room couldn't pass the Turing test—the people outside would be able to tell there wasn't a Chinese speaker in the room by detecting their causal properties. Since they can't detect causal properties, they can't detect the existence of the mental. **In short, Searle's "causal properties" and consciousness itself is undetectable, and anything that cannot be detected either does not exist or does not matter**.

Are multi-stage hierarchical programs, made of neural nets, closer to humans, or to chinese rooms?

Neural nets are **black boxes**.

But just because we can not trace the millions of calculations, does not mean we can not understand them.

It's matrix multiplication.

What really is a "chinese room", is how does the network **do** all of this? **Why** would the learned representations **mean** anything? Probably they are meaningless, they just encode mappings between representations.

Experimentally picked "teachers", like loss functions and gradient descent, produce a system that seems to **think** but really does not.

How to make the computer “understand” characters, words, sentences?

Context

= article
= conversation
= document
= poetry

+

Query

= question
= summary task

=

Result

= answer

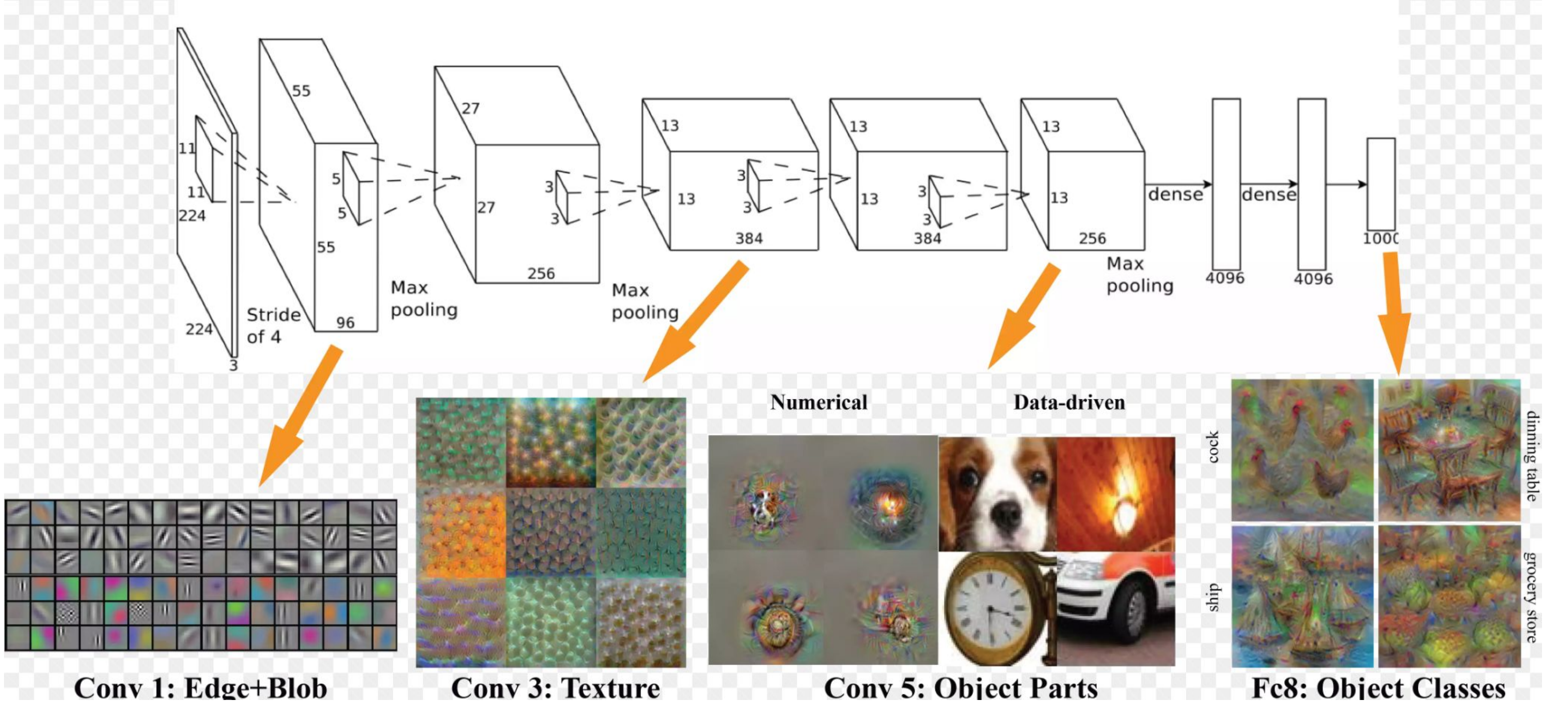
= index range - words from here, to here

= list of words - these words anser your question

Problems:

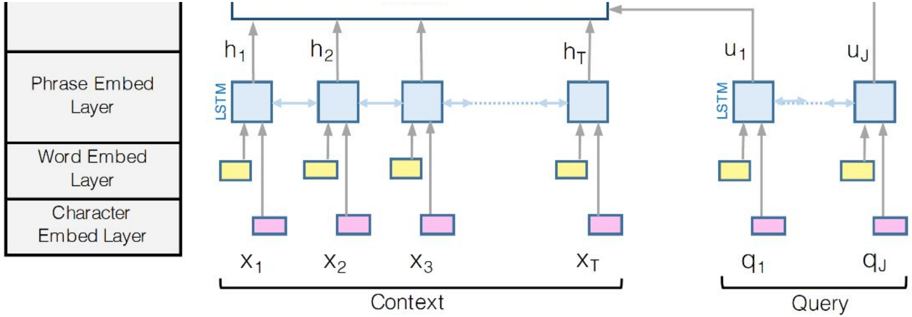
- context has “levels” and “granularity”
- you may need external knowledge, world history, sense of humour
- humans use metaphors, slogans, idioms, popular sayings
- what if none of the words in context can be picked as answer?

Hierarchical model comparison



https://www.slideshare.net/mobile/ckmarkohchang/applied-deep-learning-1103-convolutional-neural-networks

Hierarchical model comparison



whole context embedded with LSTM

phrase embedding

word embeddinngs

character embedding

. Đuka had never received a formal education. Nikola credited his eidetic memory and creative abilities to his mother's genetics and influence. Tesla's progenitors were from western Serbia, near Montenegro.:12z

Embedding - characters based - map each word to vector space

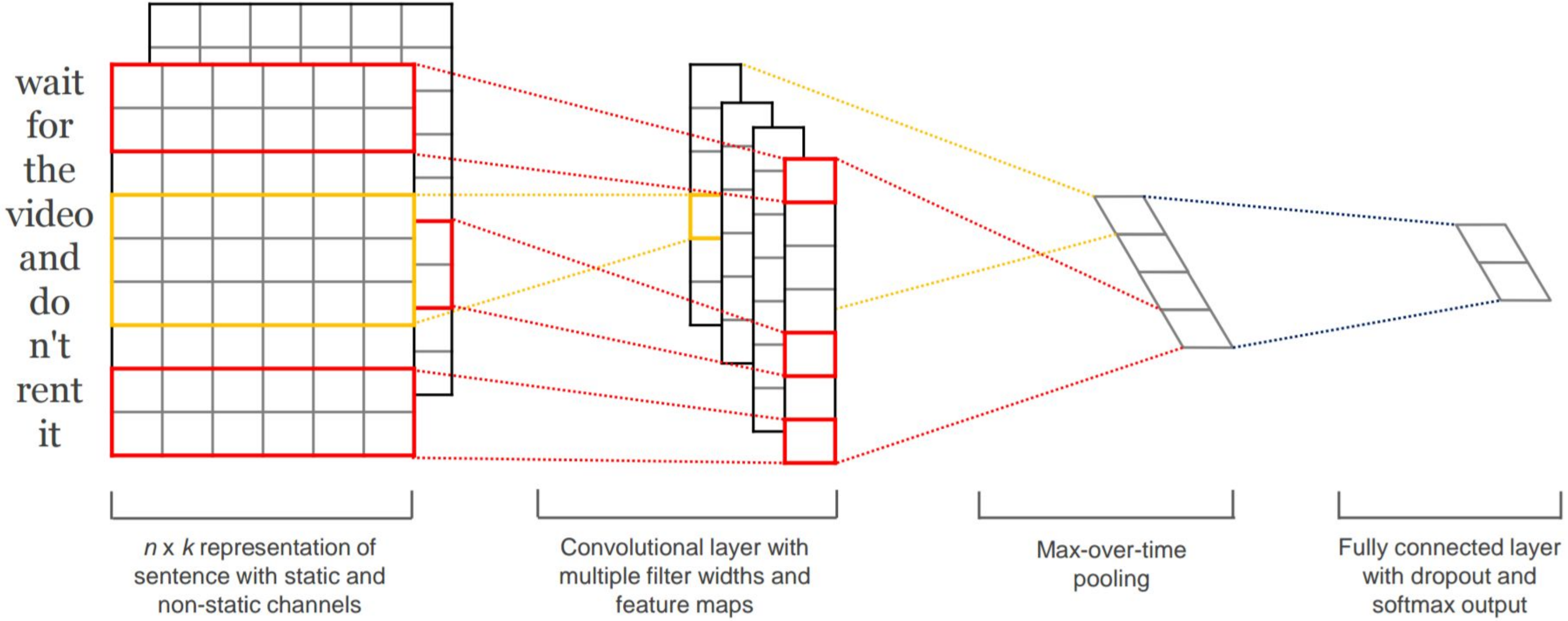


Figure 1: Model architecture with two channels for an example sentence.

Yoon Kim, Convolutional neural networks for sentence classification, EMNLP 2014

Embedding - word based - map each word to vector space

Global Vectors (GloVe) - unsupervised learning algorithm for obtaining representations for words.
Uses global (not local, like a sliding window) word-word co-occurence statistics so that the representations showcase linear substructures of word vector space.

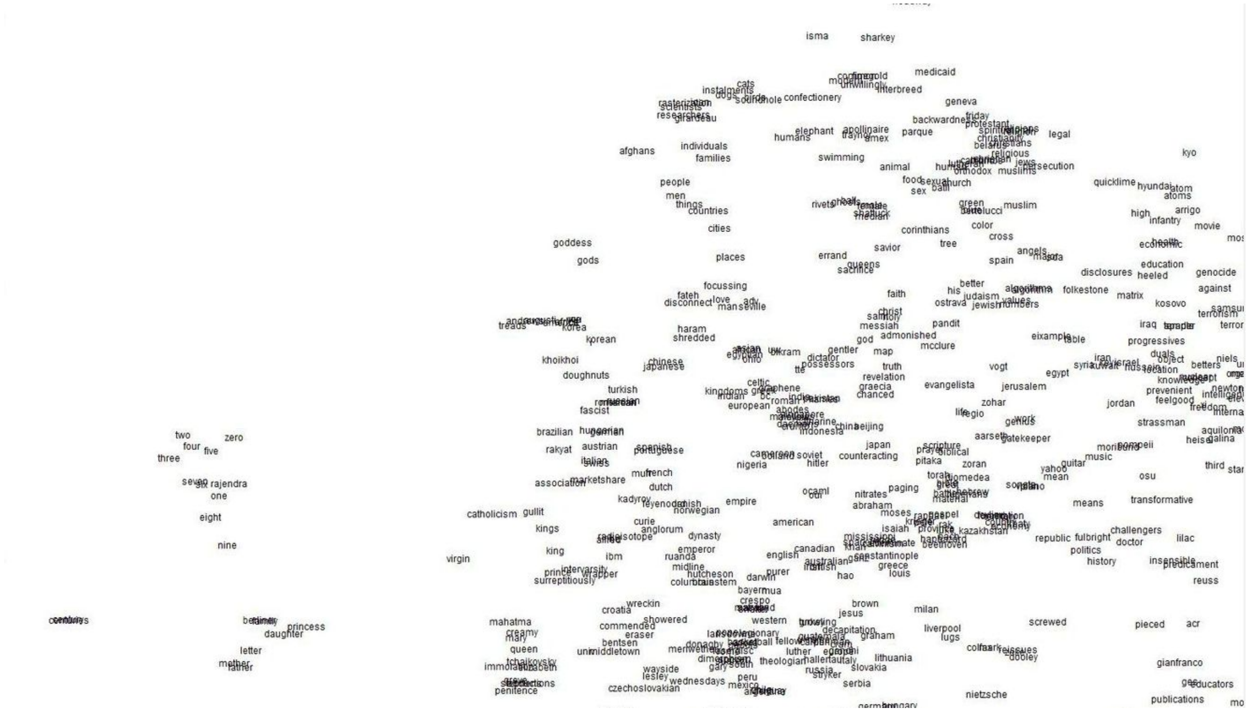
The training objective of GloVe is to learn word vectors such that their dot product equals the logarithm of the words' probability of co-occurrence.

GloVe is essentially a log-bilinear model with a weighted least-squares objective. The main intuition underlying the model is the simple observation that ratios of word-word co-occurrence probabilities have the potential for encoding some form of meaning.

In comparison,Word2Vec model is prediction based, not count based, and captures the relations between words in a window of for example 8 neighbours in sentence.

https://nlp.stanford.edu/projects/glove/

Embedding - word based - map each word to vector space



http://3.bp.blogspot.com/-4-bFATXGpBo/U6XYyCgrdCI/AAAAAAAAIGE/9ZPA3xpItLI/s1600/plottopleft.PNG

Embedding - contextual - model temporal interactions of words

For each word of context, use bi-directional LSTM through all sentences, to have a new representation of the whole context.

Concatenate outputs of LSTM in both directions.

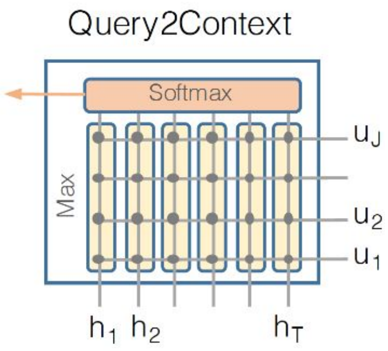
Do the same for query.

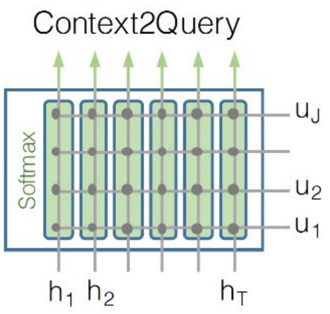
https://nlp.stanford.edu/projects/glove/

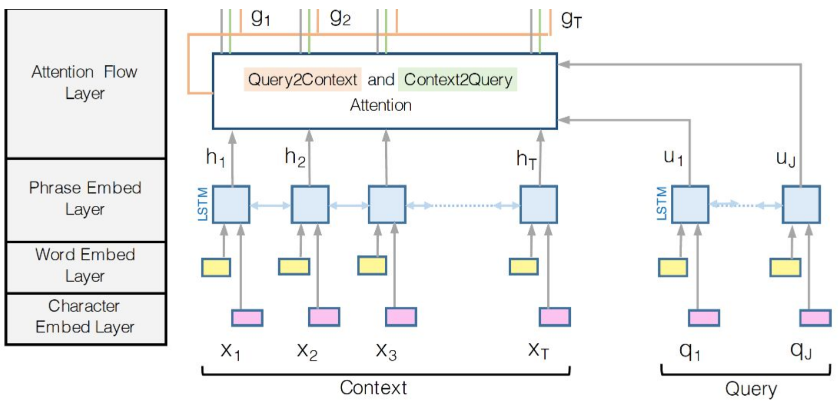
Attention flow layer

Attention flow layer is responsible for linking and fusing information from context and query.
At each time step, calculated attention vector along with embeddings from previous layers propagate to the modelling layer.
We define a shared similarity matrix between contextual embeddings of context and query.
Similarity itself is a scalar function with trainable paramers. 5 types were examined: dot product, linear weight matrix, bilinear weight matrix, linear mapping with multi-layer perceptron

Attention is calculated in two directions:
Context-to-query - which words in query are **most relevant** to each context word
Quer-to-context - which words in context have **the most similarity** to each query word and allow us to **answer**





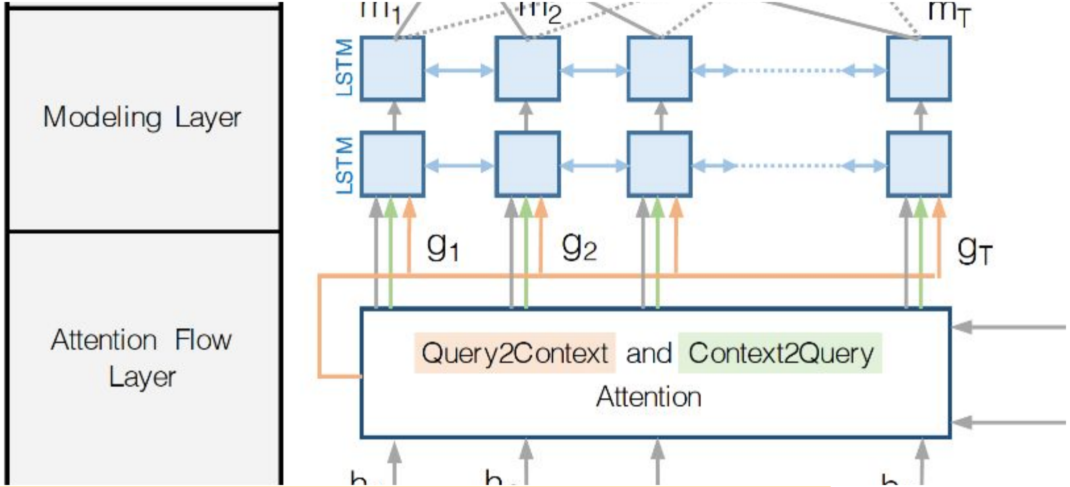


Modeling layer

Encodes **query-aware** representations of context words. Output of this layer captures interaction among context words **conditioned on** the query.

Similar trick to solution in contextual embed layer - use LSTM in both directions, but this time context vectors are dependant on query vectors.

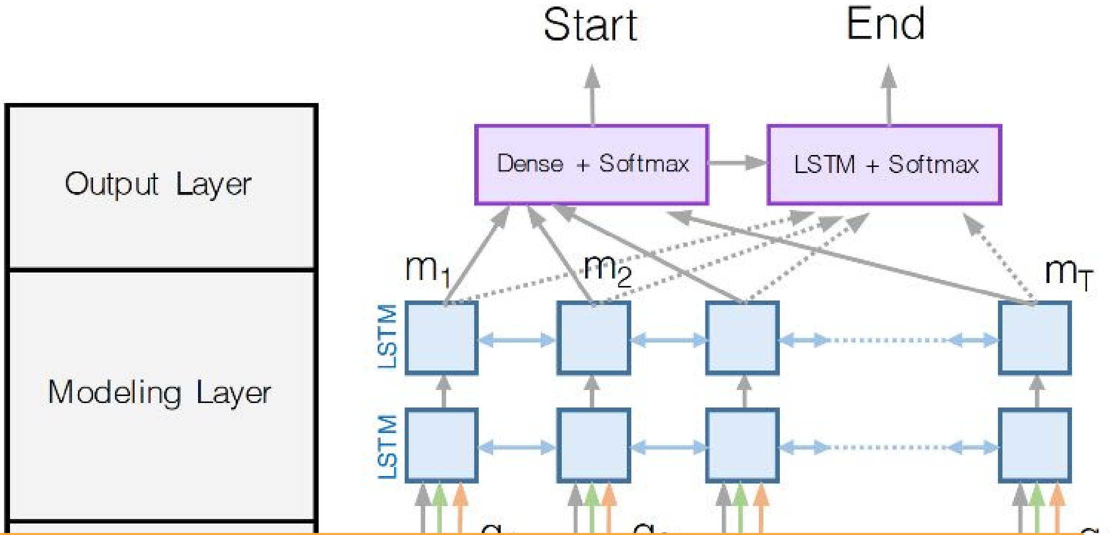
Each column vector of output is expected to contain contextual information about the word with respect to the entire context paragraph of query.



Output layer - application specific, “attachable head”

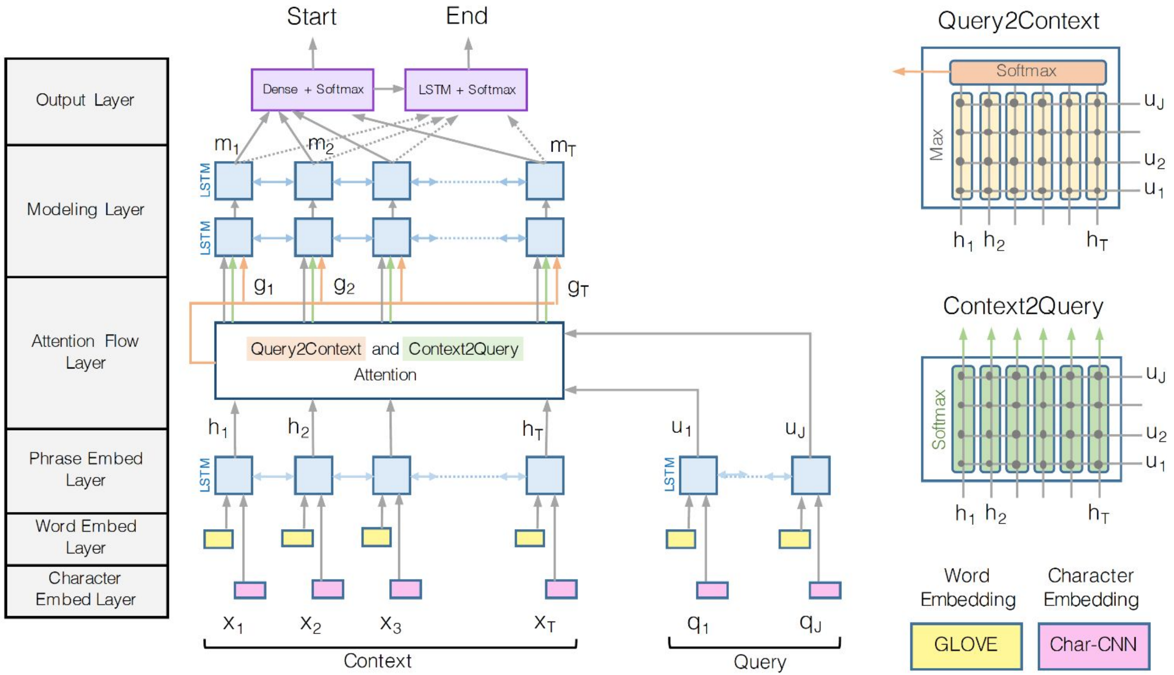
For **question answering (QA)** task, output layer is a soft-max layer - find the **subphrase** of context that answers the question.
So find two integer numbers - boundaries of subphrase.

After training, we obtain the probability distributions of start index over entire paragraph, and end index over entire paragraph. The biggest wins!



Mathematics and code implementation

See publiation on arXiv: 1611.01603v5, 24 Feb 2017
See github of Allen Institute for Artificial Intelligence: allenai.github.io/bi-att-flow/



Sources

Philosophy

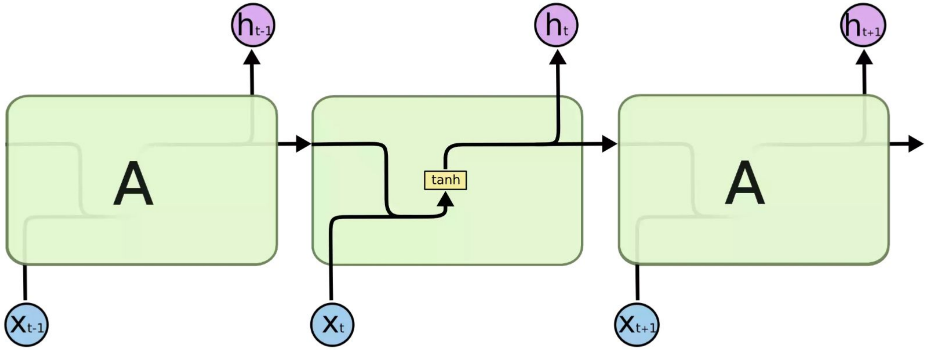
1. https://en.m.wikipedia.org/wiki/Chinese_room
2. http://philosophy.hku.hk/joelau/?n=Main.TheChineseRoomArgument

Machine reading/ comprehension/ question answering tasks

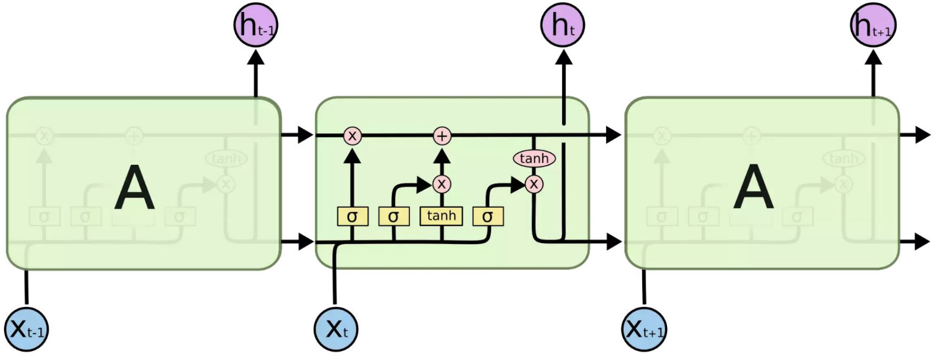
1. http://eric-yuan.me/compare-popular-mrc-datasets/
2. Machine Comprehension Using Match-lstm And Answer Pointer https://arxiv.org/abs/1608.07905 https://github.com/shuohangwang/SeqMatchSeq
3. **Bi-directional Attention Flow For Machine Comprehension** https://arxiv.org/abs/1611.01603 https://github.com/allenai/bi-att-flow/
4. **R-NET: Machine Reading Comprehension with Self-matching Networks** https://www.microsoft.com/en-us/research/publication/mrc/
5. Multi-Perspective Context Matching for Machine Comprehension https://arxiv.org/abs/1612.04211
6. Structural Embedding of Syntactic Trees for Machine Comprehension (SEST) - https://arxiv.org/abs/1703.00572

Bonus: Long Short Term Memory units

http://colah.github.io/posts/2015-08-Understanding-LSTMs/

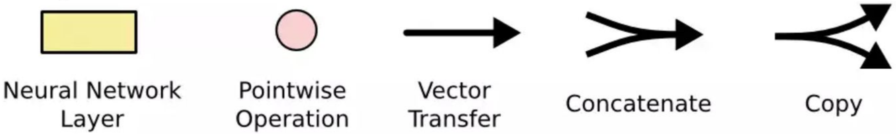


Regular recurrent neural network unit -
1 layer of tanh



The repeating module in an LSTM contains four interacting layers.

LSTM - three gates, logistic functions
Forget
Input
Output
And cell state



Bonus: Long Short Term Memory units

https://en.m.wikipedia.org/wiki/Long_short-term_memory