The Importance of Being Recurrent for Modeling Hierarchical Structure

ATOM, November 27 2018

Dr. Rachael Tatman, Kaggle

@rctatman

Outline

- How does this paper fit into the existing literature?
    - RNN’s & their drawbacks
    - Vaswani et al 2017
    - RNNs vs Transformers
- Tasks & results
    - Subject verb agreement
    - Logical inference
- Discussion

Outline

- How does this paper fit into the existing literature?
  - RNN’s & their drawbacks
  - Vaswani et al 2017
  - RNNs vs Transformers
- Tasks & results
  - Subject verb agreement
  - Logical inference
- Discussion

Outline

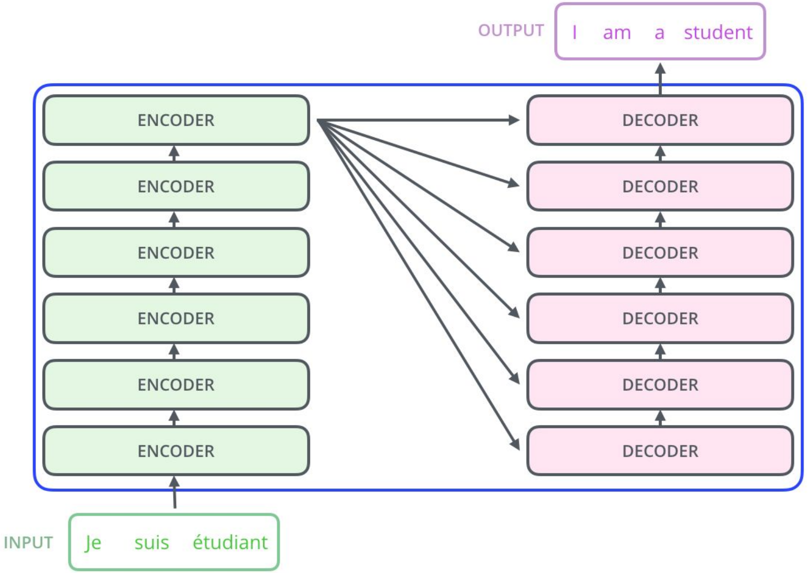
- **How does this paper fit into the existing literature?**
  - RNN’s & their drawbacks
  - Vaswani et al 2017
  - RNNs vs Transformers
- Tasks & results
  - Subject verb agreement
  - Logical inference
- Discussion

Drawbacks of RNN's

- For a good bit RNN's (2014ish - 2017ish) reigned supreme in NLP, in particular bi-LSTM with attention
  - e.g. Enhancing Sentence Embedding with Generalized Pooling (Chen & Ling 2018), which has the current SOTA for sentence-embedding based models on The Stanford Natural Language Inference (SNLI) Corpus
- But there are some problems with RNN's…
  - RNN's require a lot of memory bandwidth & you thus need a relatively small batch size (esp for inference)
  - Because of the sequential nature of training, it's hard to parallelize training over the entire sequence length (but see Parallelizing Linear Recurrent Neural Nets Over Sequence Length by Martin & Cundy, n.d.)
  - As a result, they've gained a bit of a reputation for being "inefficient and not scalable"

“Attention is all you need”

- Vaswani et al 2017, at NeurIPS, proposed a feed-forward self-attention model they called a transformer
- There are no sequential dependencies in training, so can be parallelized very efficiently
- Transformer models are currently state of the art for machine translation
    - Weighted Transformer Network for Machine Translation (Ahmen et al 2017)



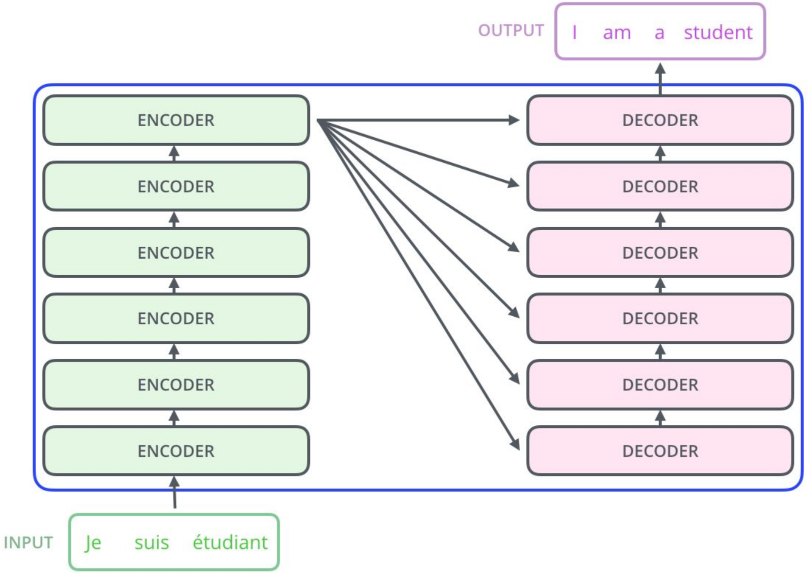
The Illustrated Transformer, Jay Alammar

@rctatman

“Attention is all you need”

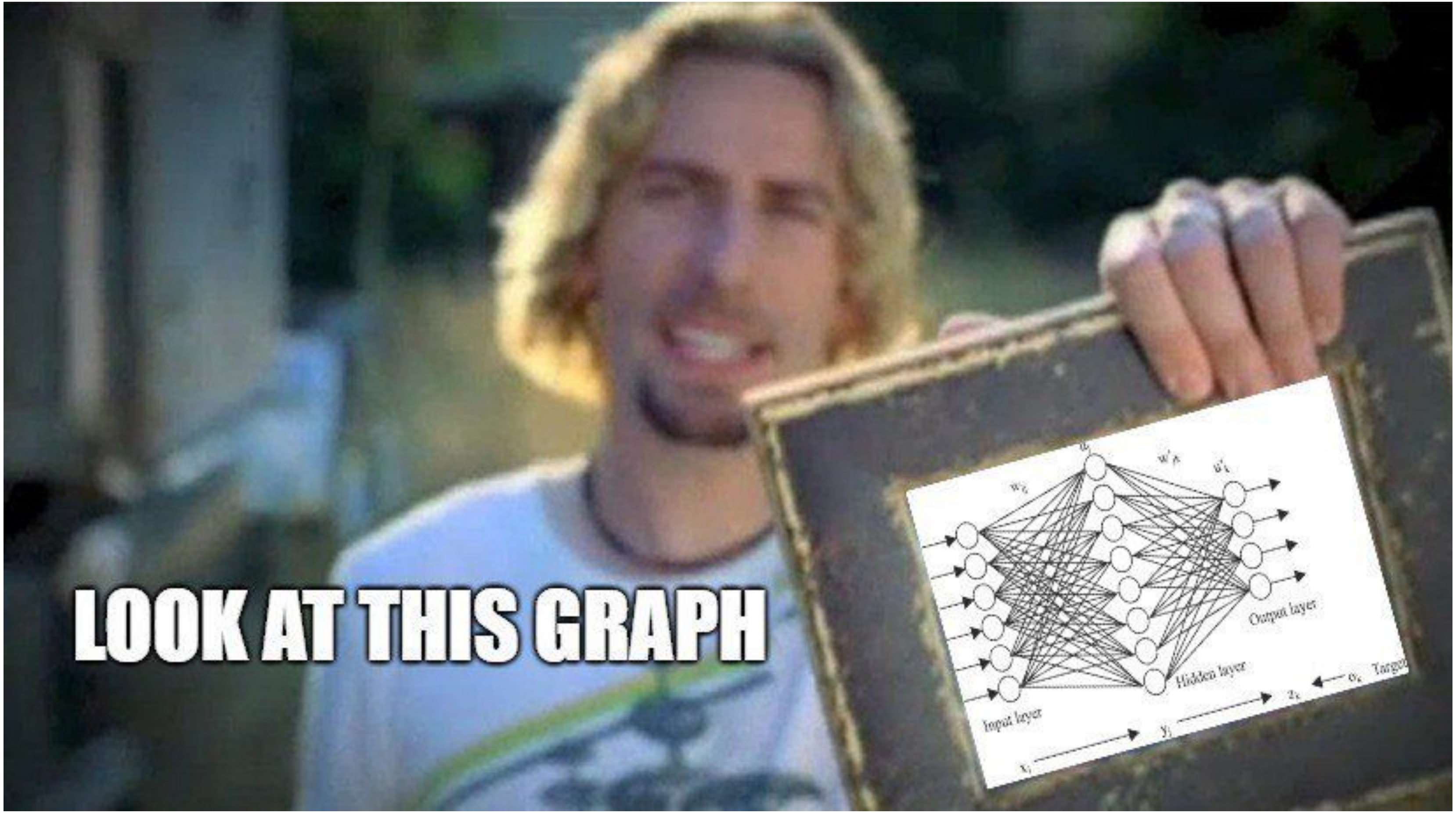
- Vaswani et al 2017, at NeurIPS, proposed a feed-forward self-attention model they called a transformer
- There are no sequential dependencies in training, so can be parallelized very efficiently
- Transformer models are currently state of the art [illegible] translation
  - ○ Weighted Tra[illegible] Machine Translation (Ahmed et al. 2017)

Based on BLEU which is a Bad Metric

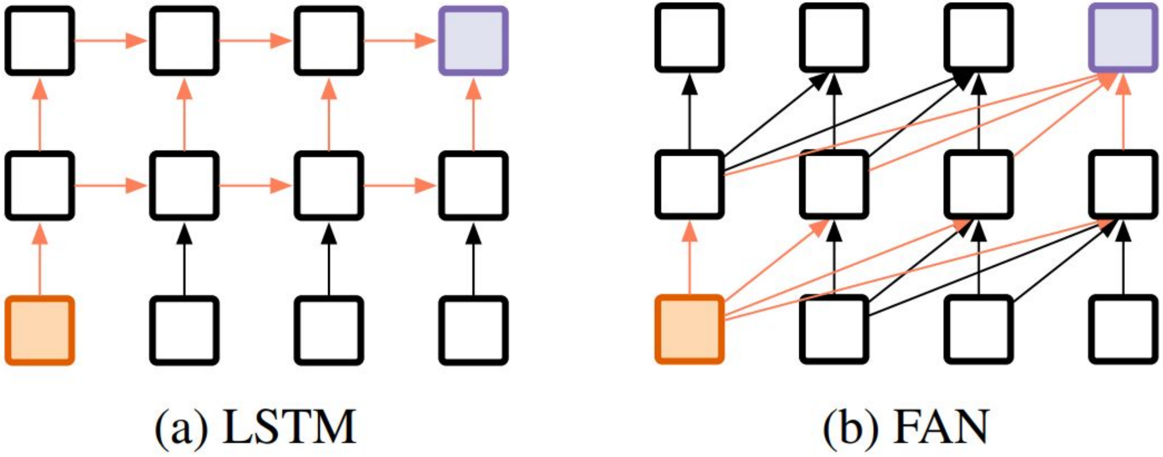


The Illustrated Transformer, Jay Alammar

@rctatman


LOOK AT THIS GRAPH
w_{ij}
u_j
w'_{jk}
u'_k
Output layer
Hidden layer
Targe
o_k
z_k
Input layer
y_j
x_i

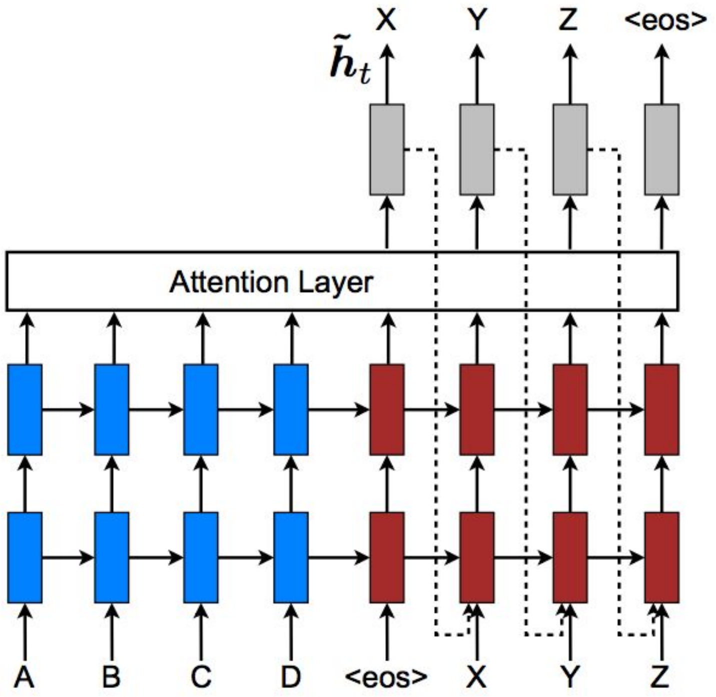
@rctatman



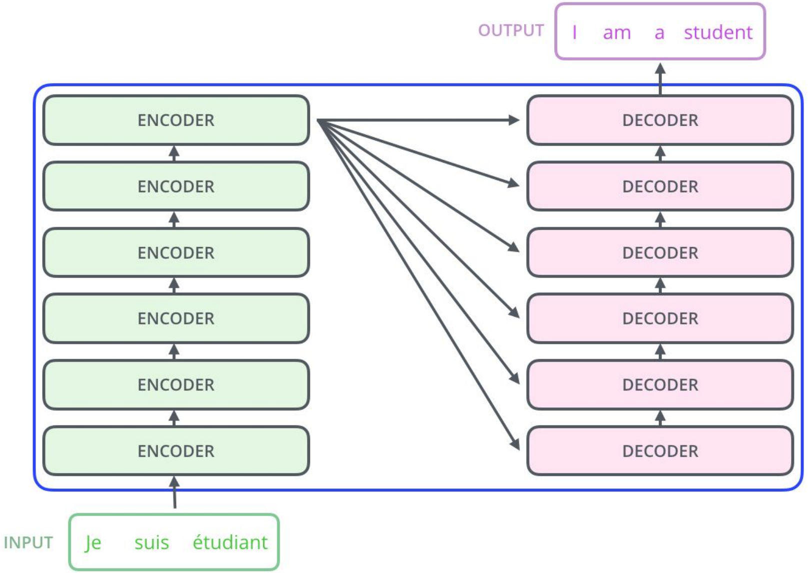
(a) LSTM

(b) FAN

Figure 1: Diagram showing the main difference between a LSTM and a FAN. Purple boxes indicate the summarized vector at current time step t which is used to make prediction. Orange arrows indicate the information flow from a previous input to that vector.



Effective Approaches to Attention-based Neural Machine Translation, Luong et al 2015



The Illustrated Transformer, Jay Alammar

@rctatman

Outline

- How does this paper fit into the existing literature?
    - RNN’s & their drawbacks
    - Vaswani et al 2017
    - RNNs vs Transformers
- **Tasks**
    - Subject verb agreement
    - Logical inference (entailment)
- Models
- Results
- Discussion

Outline

- How does this paper fit into the existing literature?
  - RNN’s & their drawbacks
  - Vaswani et al 2017
  - RNNs vs Transformers
- **Tasks & results**
  - Subject verb agreement
  - Logical inference
- Discussion

Hierarchical Structure

- All language (human or computer) contains hierarchical structure and recursion
  - “I saw the person with the binoculars”
  - “I put the keys on the stand, on the table, by the couch, next to the desk…”
- Linguists care about this A Lot & it’s also important for human-level performance in NLP tasks
- Traditionally included via explicit, often handbuilt representations (treebanks, dictionaries, etc.)

@rctatman

Subject-Verb Agreement

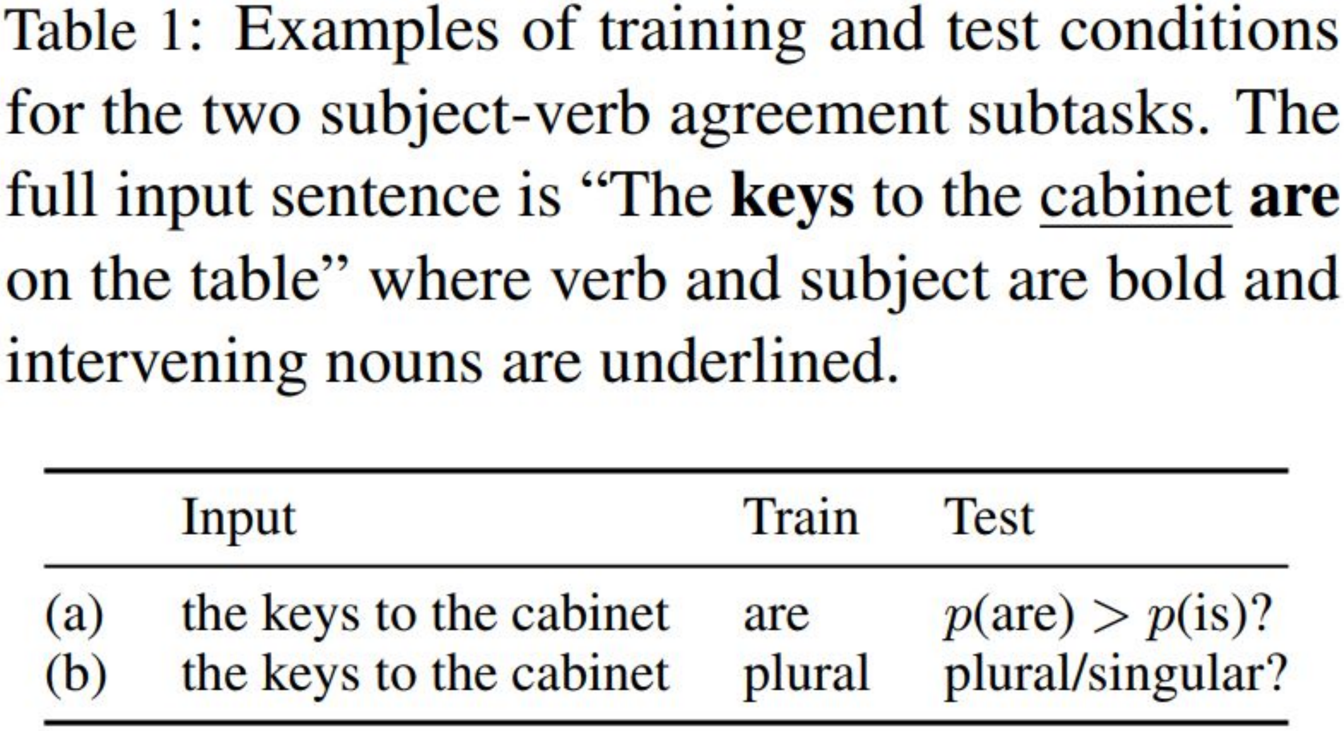


Table 1: Examples of training and test conditions for the two subject-verb agreement subtasks. The full input sentence is “The **keys** to the cabinet **are** on the table” where verb and subject are bold and intervening nouns are underlined.

	Input	Train	Test
(a)	the keys to the cabinet	are	$p(\text{are}) > p(\text{is})$?
(b)	the keys to the cabinet	plural	plural/singular?

@rctatman

Subject-Verb Agreement

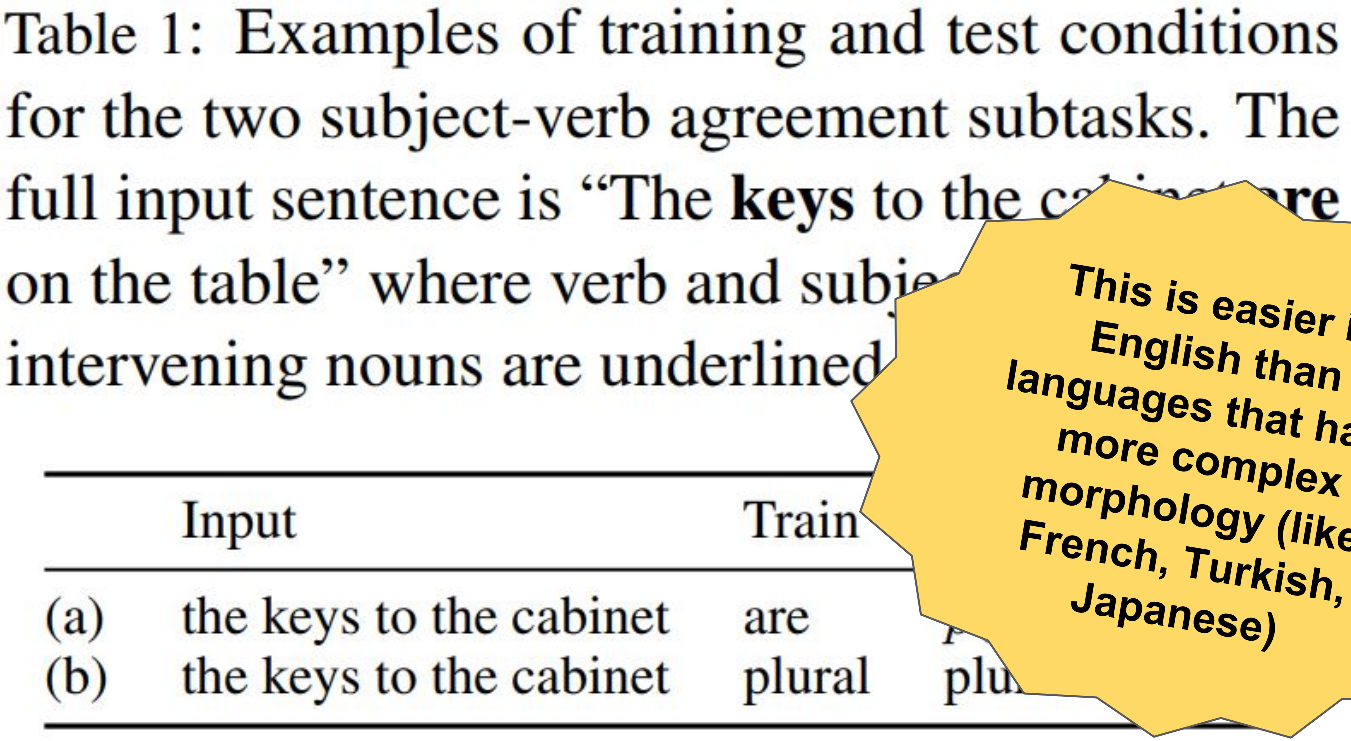


Table 1: Examples of training and test conditions for the two subject-verb agreement subtasks. The full input sentence is “The **keys** to the ca **re** on the table” where verb and subje intervening nouns are underlined

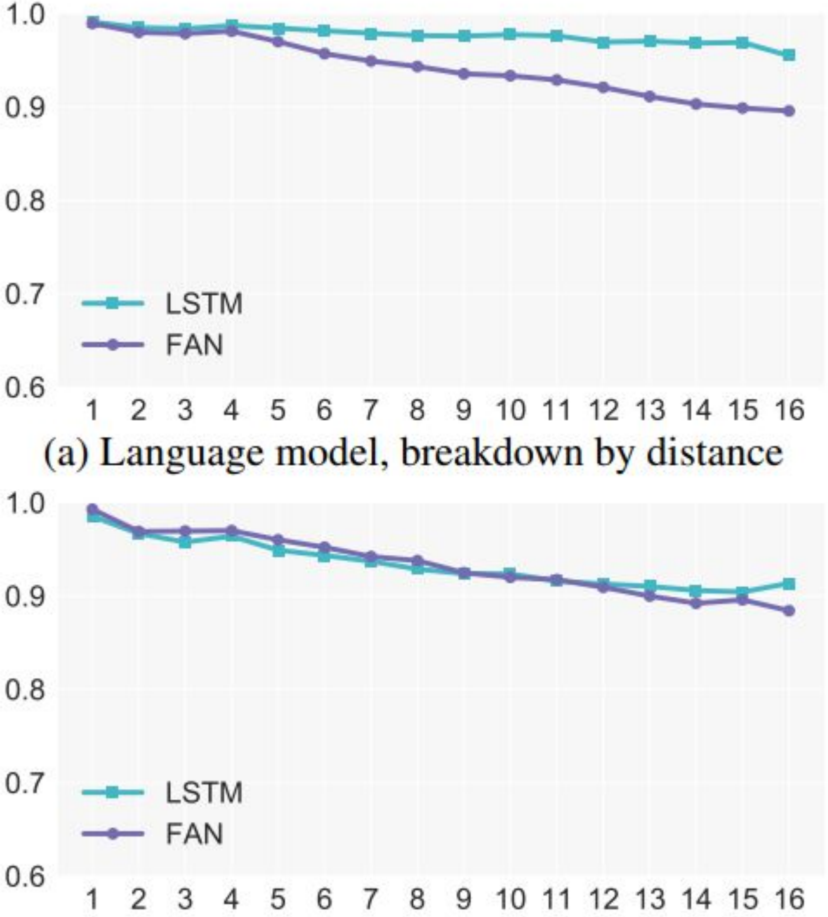
	Input	Train	
(a)	the keys to the cabinet	are	p
(b)	the keys to the cabinet	plural	plu

This is easier in English than languages that have more complex morphology (like French, Turkish, Japanese)

@rctatman

Hyperparameters: To allow for a fair comparison, we find the best configuration for each model by running a grid search over the following hyperparameters: number of layers in {2, 3, 4}, dropout rate in {0.2, 0.3, 0.5}, embedding size and number of hidden units in {128, 256, 512}, number of heads (for FAN) in {2, 4}, and learning rate in {0.00001, 0.0001, 0.001}. The weights of the word embeddings and output layer are shared (Inan et al., 2017; Press and Wolf, 2017). Models are optimized by Adam (Kingma and Ba, 2015).

This looks OK to me… but I haven't worked with transformers previously. Is this a fair comparison? Are there common training tricks omitted here?



(a) Language model, breakdown by distance

(c) Number prediction, breakdown by distance

The **cat is** in the kitchen.

Distance = 1

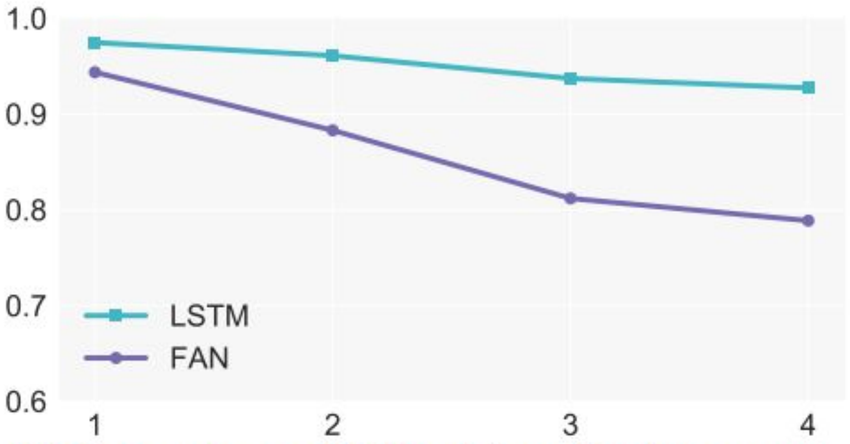
The **cat** that hasn’t been sleeping well this week **is** in the kitchen.

Distance = 8

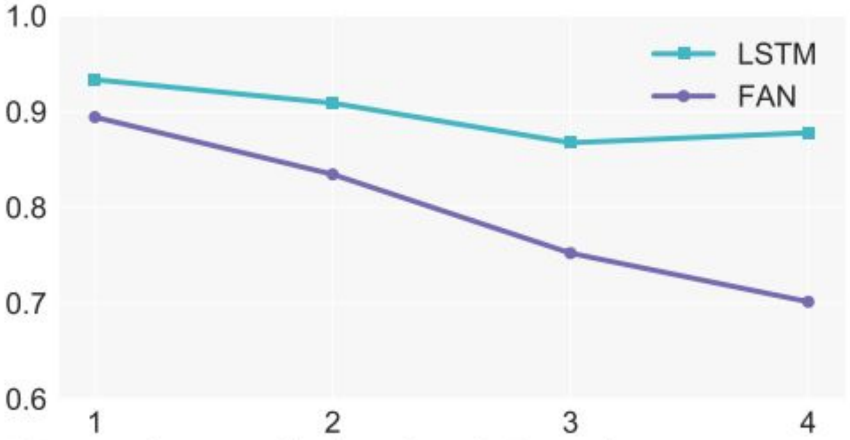
Points:

- The number prediction (singular vs. plural) tracks well between the models
- The LSTM language model is much better over longer distances (remember the embeddings are shared between models…)
- Authors: “[better language model results] may be due to better model optimization and to the embedding-output layer weight sharing”

@rctatman



(b) Language model, breakdown by # attractors



(d) Number prediction, breakdown by # attractors

The **bus** always **comes** late.

Attractors = 0

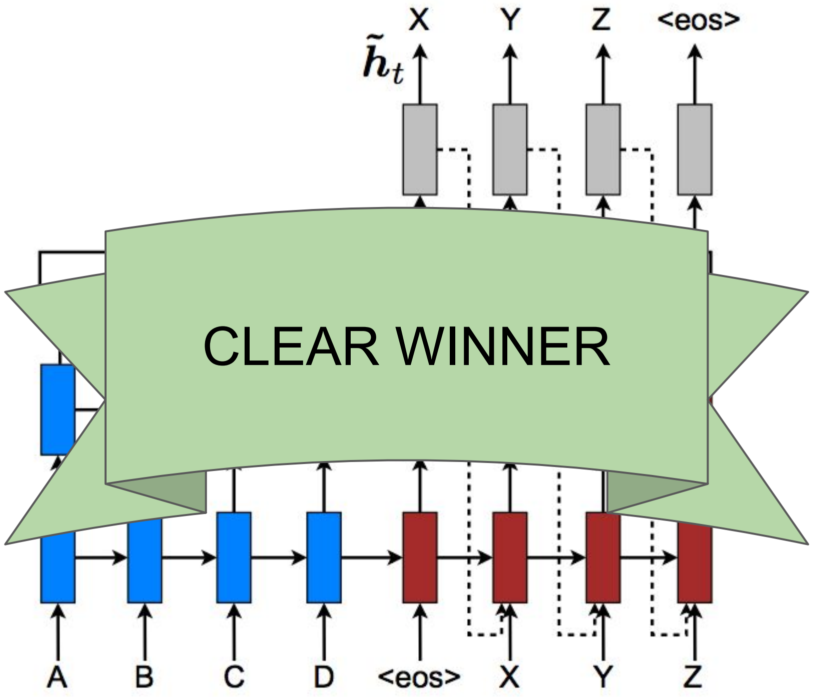
The **bus** that has a broken *windshield* always **comes** late.

Attractors = 1

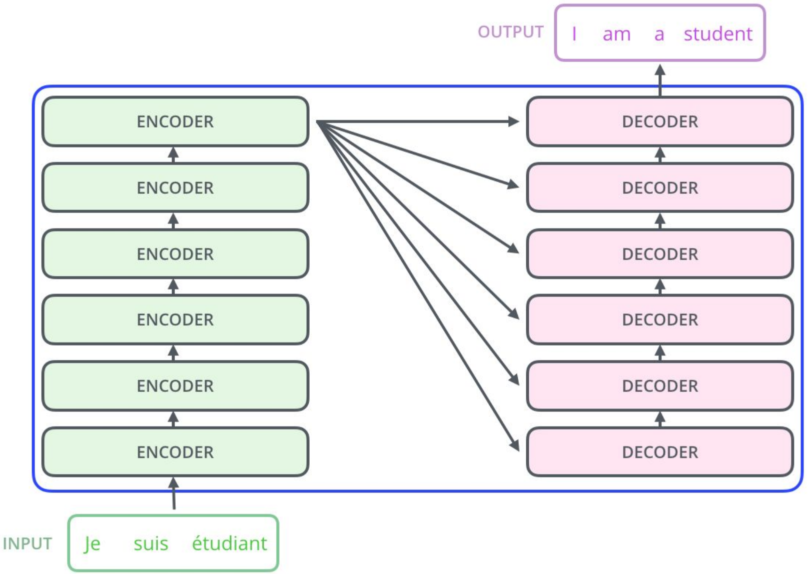
Points:

- LSTM strongly out performs the FAN here
- Not clear *a priori* why that should be
- Authors: "[it's possible] human memory limitations give rise to important characteristics of natural language, including its hierarchical structure. … by compressing the history into a single vector before making predictions, LSTMs are forced to better learn the input structure."

@rctatman

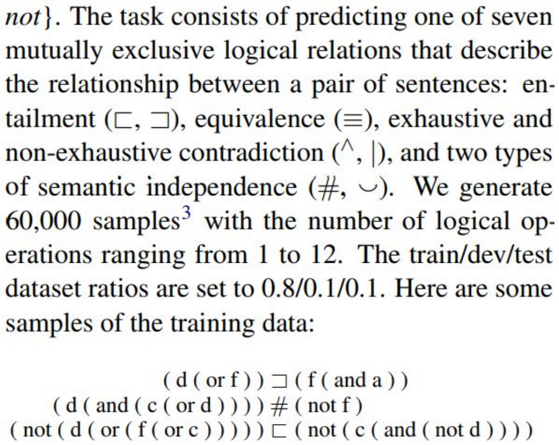


Effective Approaches to Attention-based Neural Machine Translation, Luong et al 2015



The Illustrated Transformer, Jay Alammar

Logical Inference



not}. The task consists of predicting one of seven mutually exclusive logical relations that describe the relationship between a pair of sentences: entailment ($\sqsubset$, $\sqsupset$), equivalence ($\equiv$), exhaustive and non-exhaustive contradiction ($^\wedge$, $|$), and two types of semantic independence ($\#$, $\smile$). We generate 60,000 samples[3] with the number of logical operations ranging from 1 to 12. The train/dev/test dataset ratios are set to 0.8/0.1/0.1. Here are some samples of the training data:

$$(\text{d} (\text{or f})) \sqsupset (\text{f} (\text{and a}))$$
$$(\text{d} (\text{and} (\text{c} (\text{or d})))) \# (\text{not f})$$
$$(\text{not} (\text{d} (\text{or} (\text{f} (\text{or c}))))) \sqsubset (\text{not} (\text{c} (\text{and} (\text{not d}))))$$

Similar Natural language examples:

Glen and Amber ate peaches CONTRADICTS Neither Glen nor Amber ate peaches

Glen and Amber ate peaches IS EQUIVALENT TO Amber and Glen ate peaches

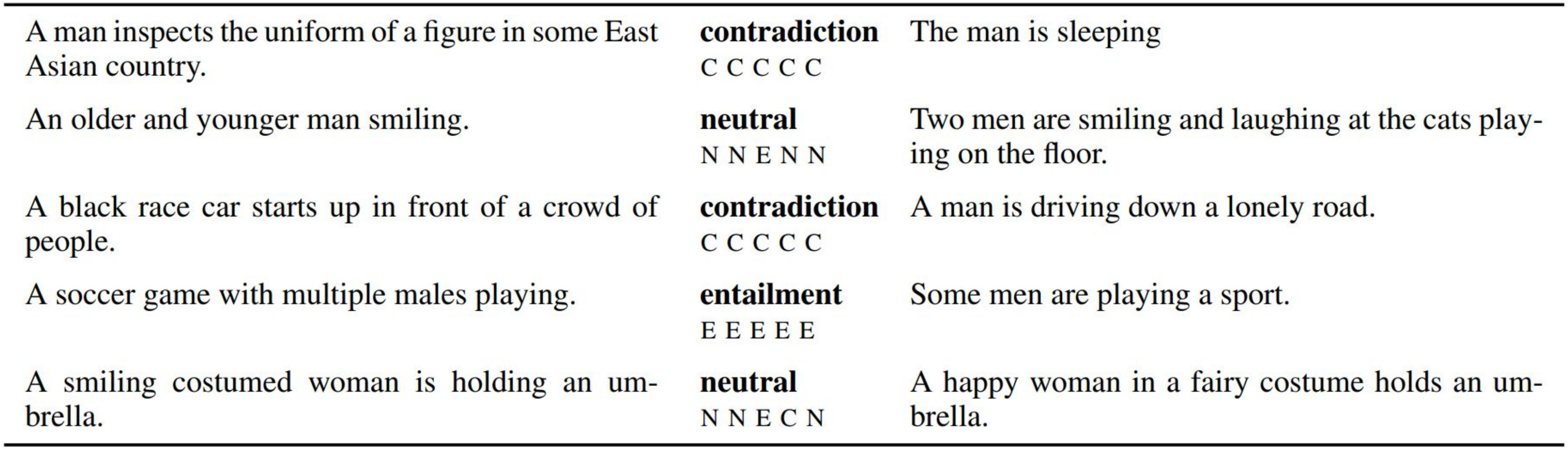
Glen and Amber ate peaches ENTAILS THAT Glen ate peaches

Amber ate a peach IS ENTAILED BY Glen and Amber ate peaches

Glen and Amber ate peaches IS INDEPENDENT OF WHETHER Ben is at the zoo

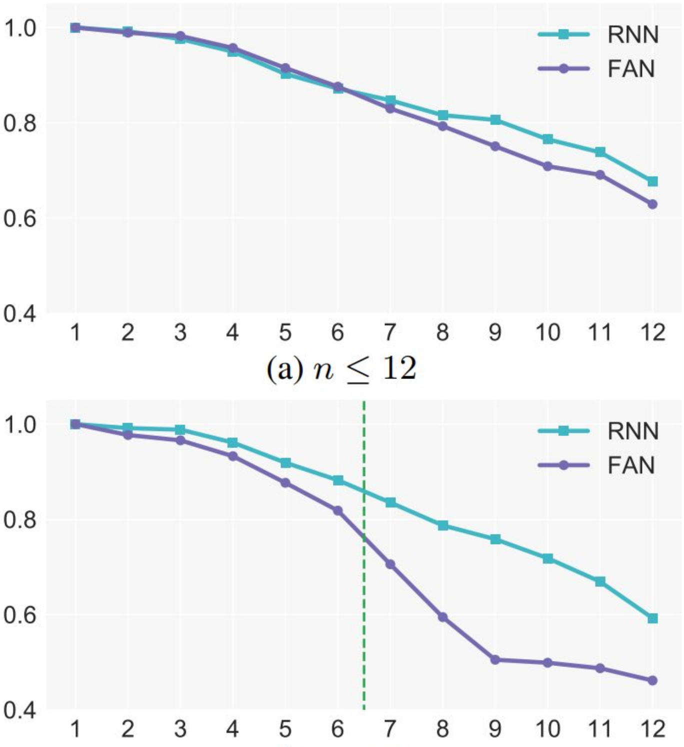
In the natural language example, the task would be to predict the UPPER CASE phrase given both sentences. (Note that only the symbolic artificial data was used in the experiment.)

More examples from Bowman et al 2015



A man inspects the uniform of a figure in some East Asian country.	**contradiction** C C C C C	The man is sleeping
An older and younger man smiling.	**neutral** N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	**contradiction** C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	**entailment** E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	**neutral** N N E C N	A happy woman in a fairy costume holds an umbrella.

This corpus (SNLI) based on human annotations.



(a) $n \leq 12$

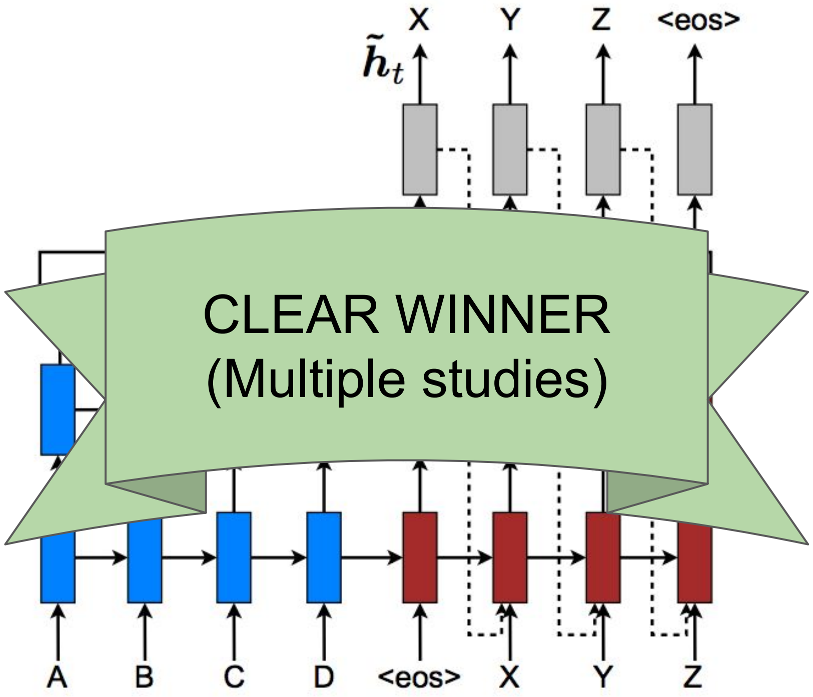
(b) $n \leq 6$

Figure 4: Results of logical inference when training on all data (a) or only on samples with at most n logical operators (b).

Number of logical operators:

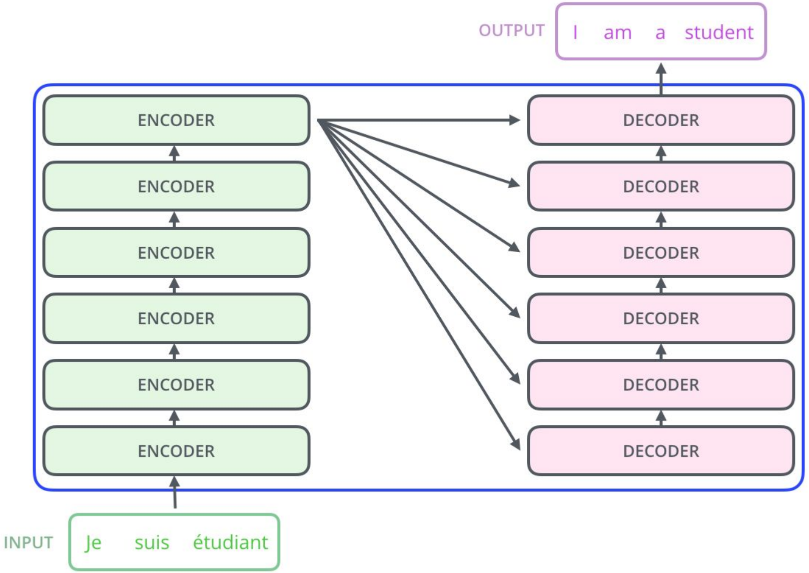
(d (**or** f)) ⊐ (f (**and** a)) = 2
(d (**and** (c (**or** d)))) # (**not** f) = 3

- Used same architecture as Bowman et al 2015, nothing really new there
- Similar accuracy when trained on all data, but RNN generalizes better
- Authors: “Concurrently to our work Evans et al. (2018) proposed an alternative data set for logical inference and also found that a FAN model underperformed various other architectures including LSTMs.”
- Why? ¯_(ツ)_/¯



CLEAR WINNER
(Multiple studies)

Effective Approaches to Attention-based Neural Machine Translation, Luong et al 2015



The Illustrated Transformer, Jay Alammar

@rctatman

Outline

- How does this paper fit into the existing literature?
  - RNN's & their drawbacks
  - Vaswani et al 2017
  - RNNs vs Transformers
- Tasks & results
  - Subject verb agreement
  - Logical inference
- **Discussion**

Discussion

- But WHY?
  - Possible that compression to a single vector compresses history in the same way that humans due (due to memory limitations)
  - I ❤ Empiricism… but I'd really, really like to see more theoretical results around deep learning (like "Neural Networks Should Be Wide Enough to Learn Disconnected Decision Regions" at ICML 2018)
- It's possible that LSTMs are doing better because they've been around longer & we've learned more tricks for getting them to work well…
  - Would a weighted transformer have done better?
  - Authors also didn't look at convolutional models (like ConvS2S)
- Your thoughts!

Thanks!
Other questions?

Slides:

@rctatman