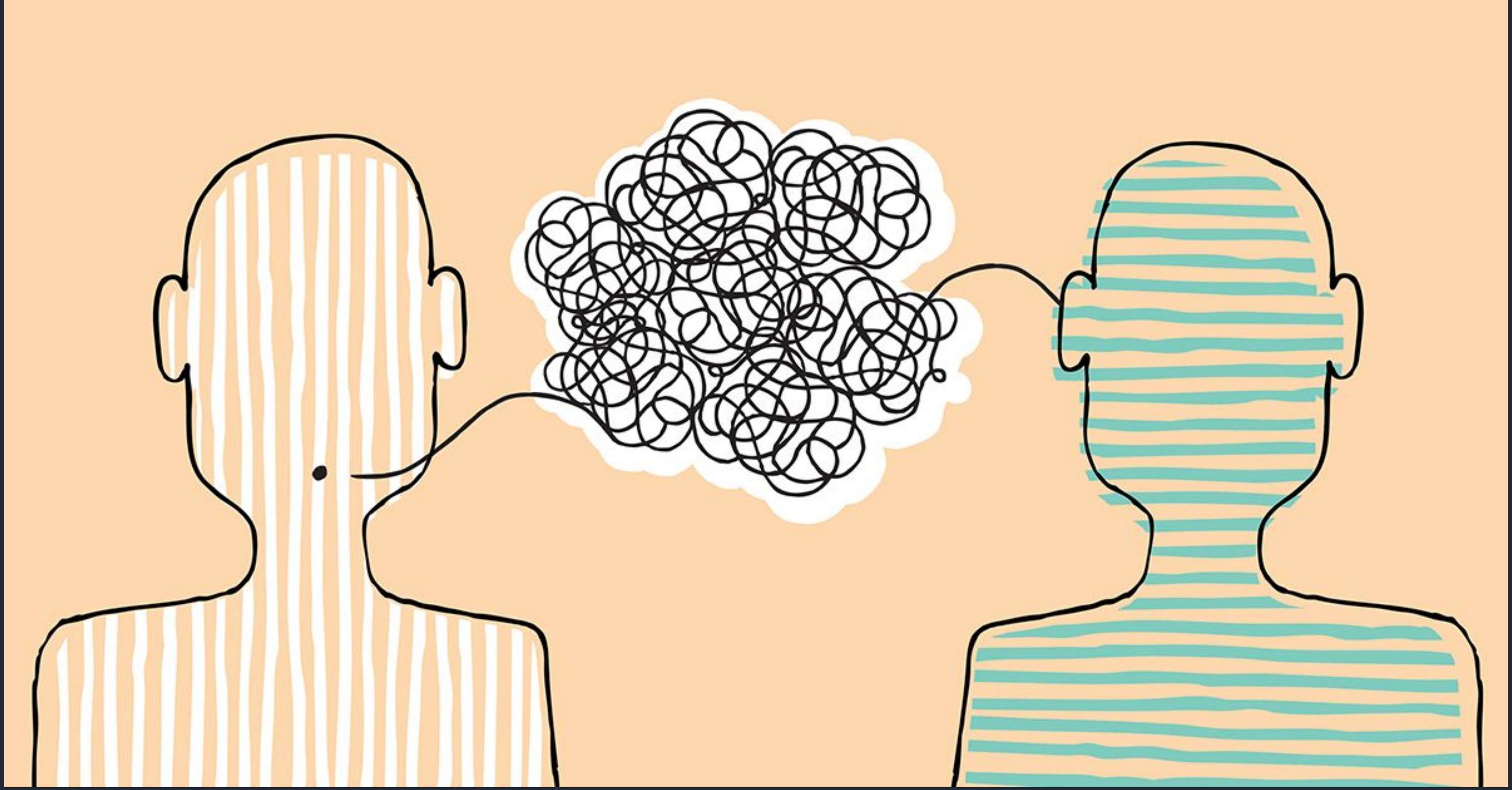
Cognitive Plausibility of Deep Language Models

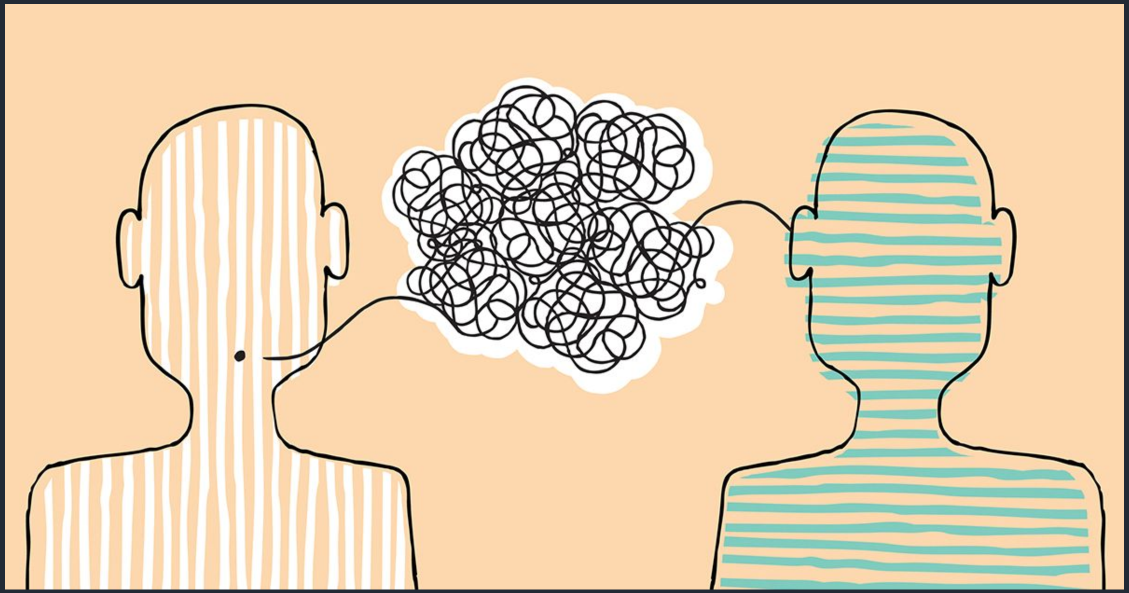
Lisa Beinborn
VU Amsterdam

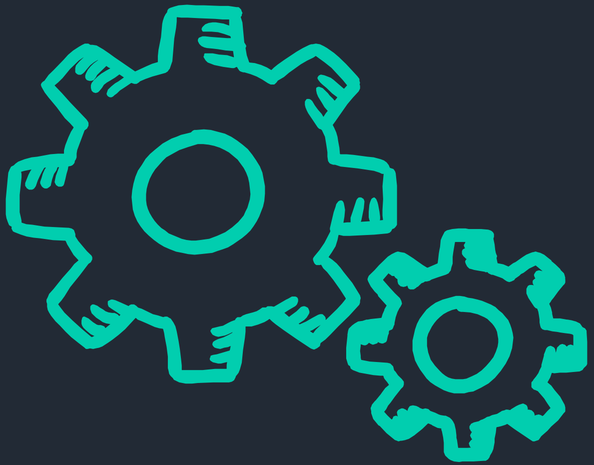
Nora Hollenstein
Uv Copenhagen

Willem Zuidema
Uv Amsterdam



https://www.analyticsinsight.net/why-the-future-is-bright-for-natural-language-processing/

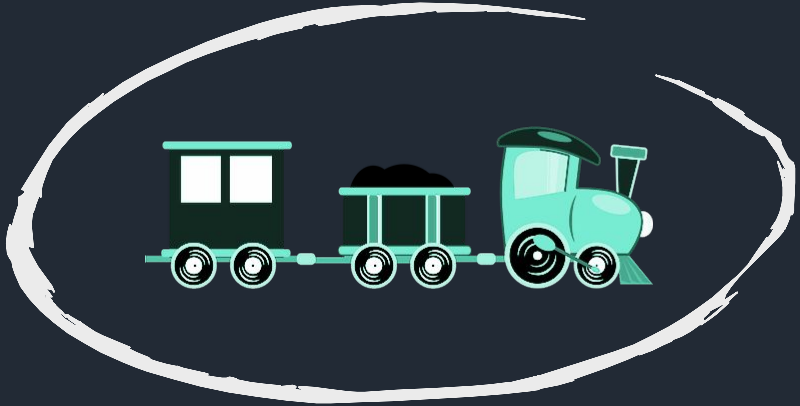




https://www.analyticsinsight.net/why-the-future-is-bright-for-natural-language-processing/

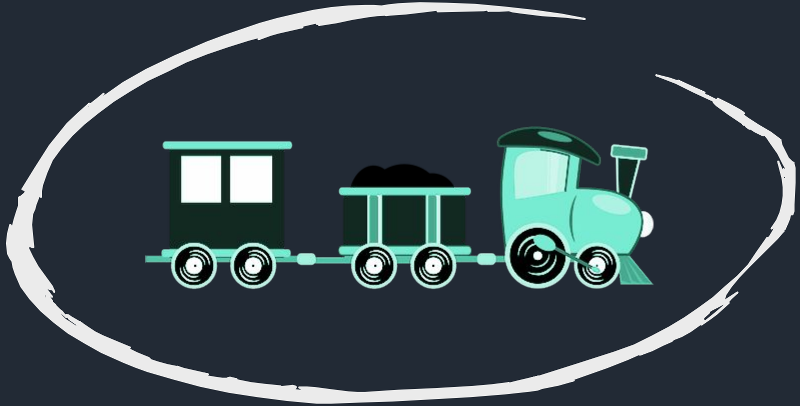
Context matters

The old train



Context matters

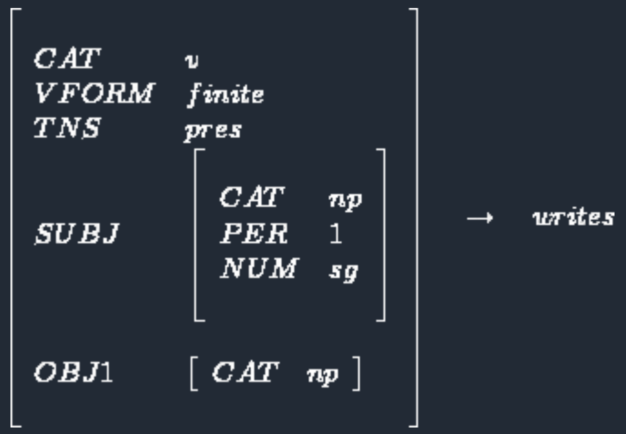
The old train

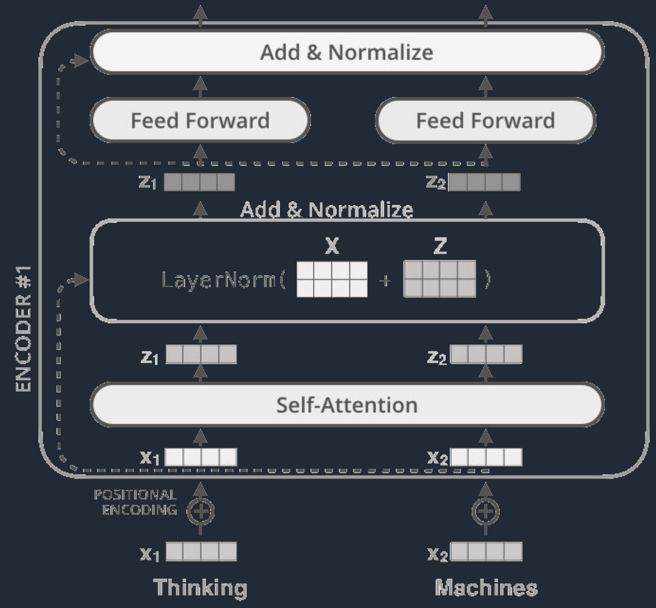


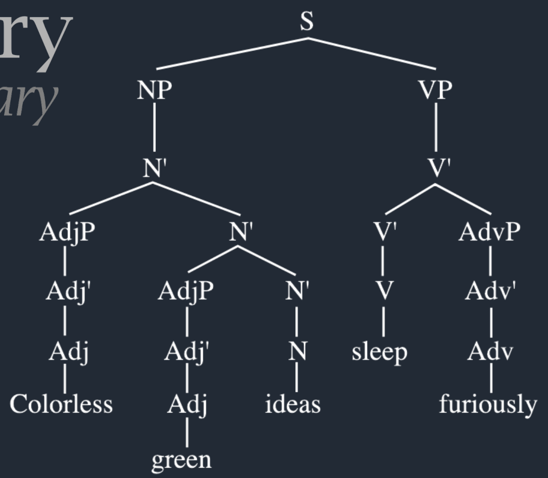
The old train the young.

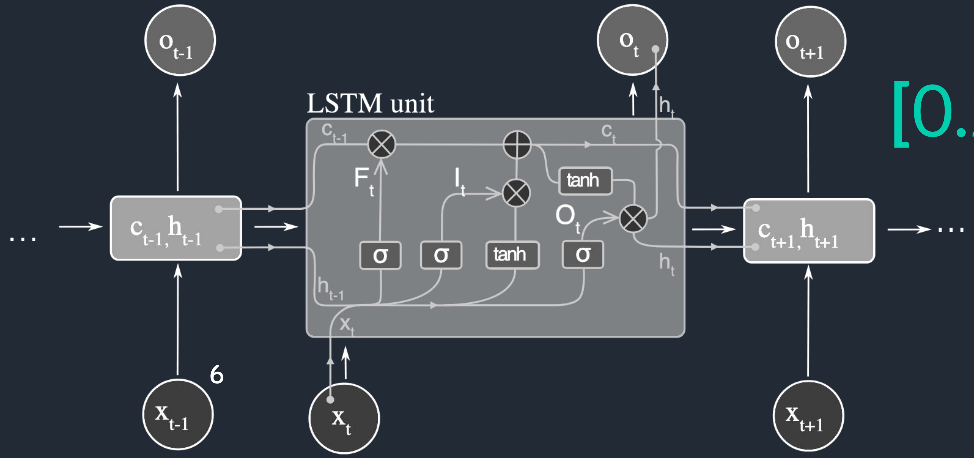
How to model language?

Wiktionary
The free dictionary









[0.21 -0.54 0.18...]

Outline

1. What are language models?
2. When is a language model cognitively plausible?

$\longrightarrow$ Today

3. How can we analyze this?

$\longrightarrow$ The next four days

Where are we coming from?

Traditionally, the task of a **language model** was:

- next word prediction (auto-completion)
- calculate probability of a sequence

Where are we coming from?

Traditionally, the task of a **language model** was:

- next word prediction (auto-completion)
- calculate probability of a sequence

Next word prediction: find word **X** which maximizes the probability

P(Lisa sings X) = P(X|sings) × P(sings|Lisa) × P(Lisa|<BOS>)

Higher order n-grams: P(X|Lisa sings)× P(sings |<BOS> Lisa)

Surprisal

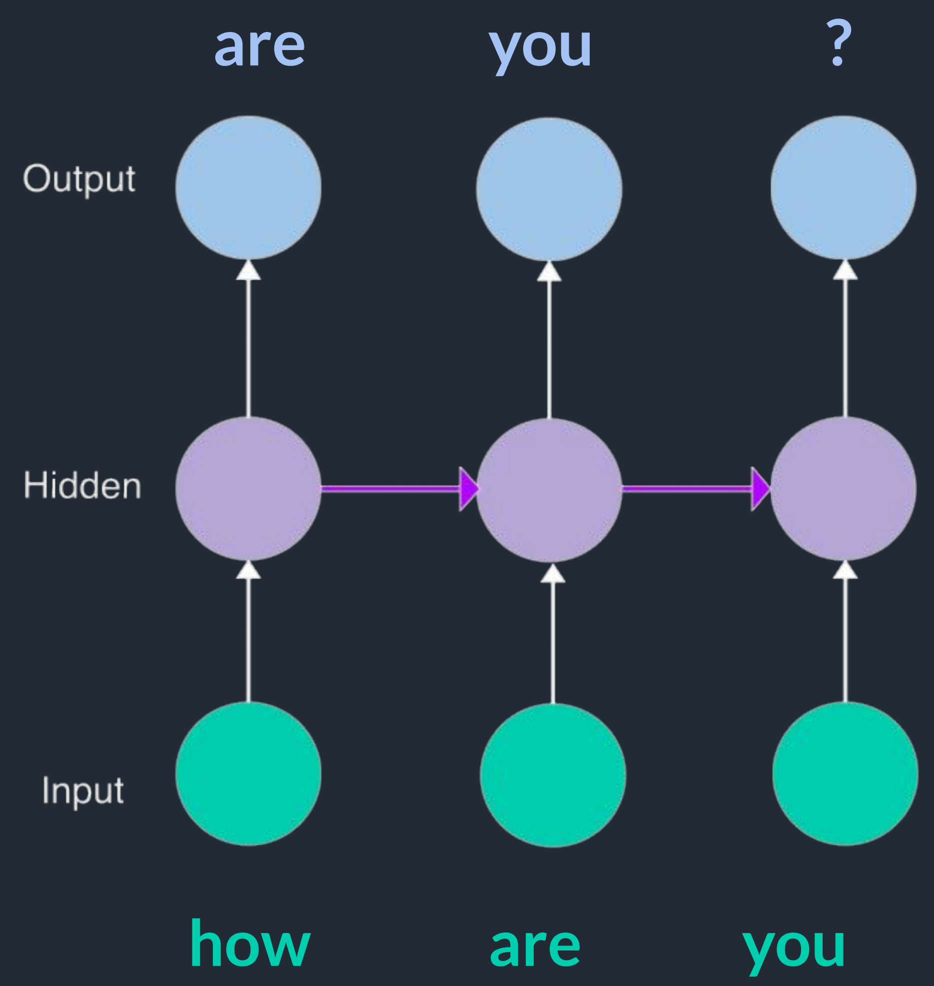
Next word prediction: find word X which maximizes the probability P(Lisa sings X)

Surprisal (w_t) $= -\log(P(w_t|w_1, ..., w_{t-1}))$

songs, really, in, ... vs skyscrapers, green, parboiled

→ This metric is often used to analyze cognitive signals such as reading times, eye-tracking, EEG. [Monsalve et al. 2012]

An RNN language model



Input

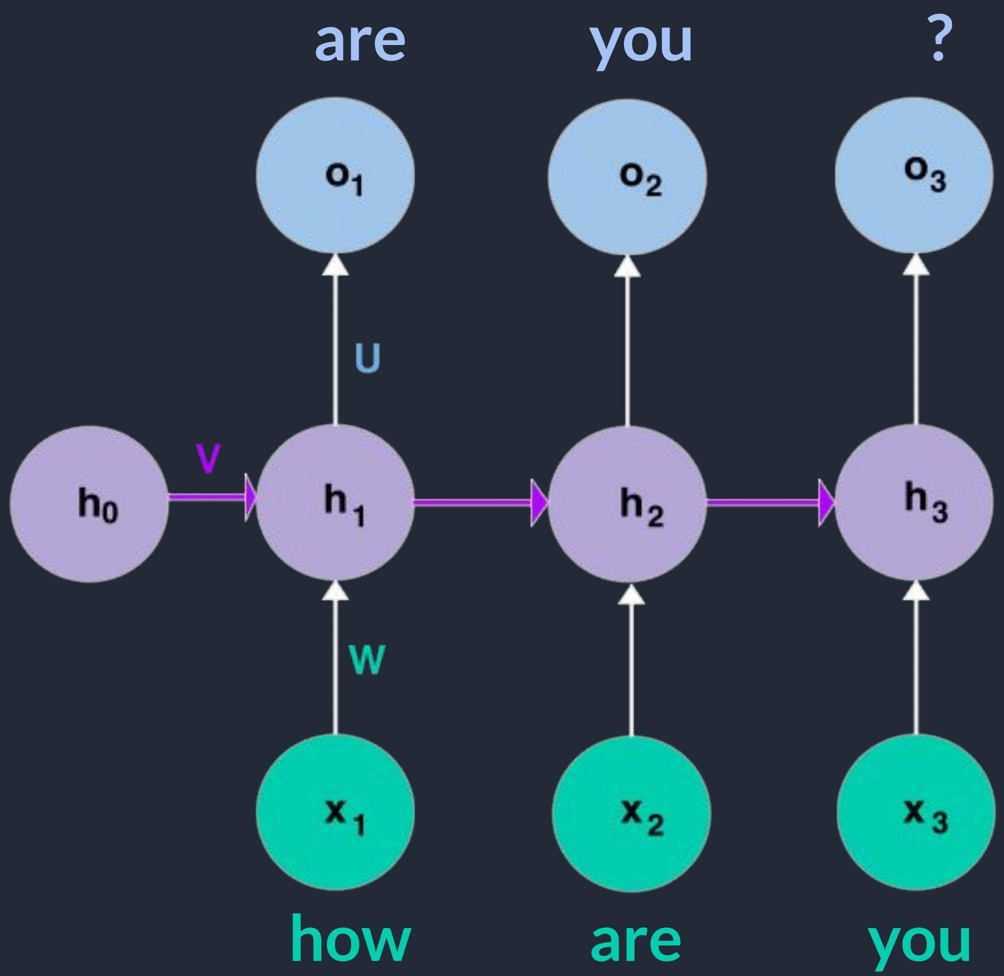
What is the **input**?

Words are mapped into token ids which are mapped into vectors using weight matrix **W**.

How is **W** initialized?

randomly

but fixed → the same token is mapped into the same vector

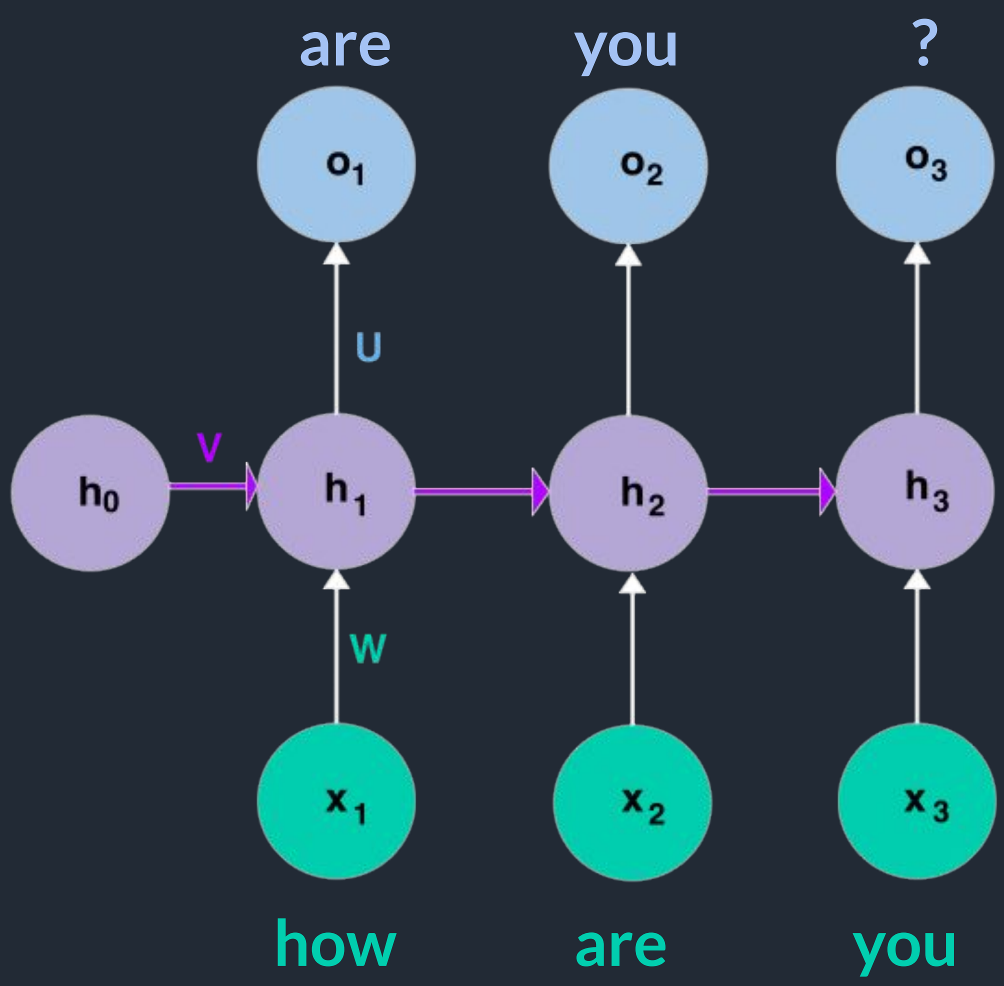


Output

What is the **output**?

The model outputs a vector **o** by multiplying the hidden state with a weight matrix **U** and applying an activation function.

We need to apply the softmax to this output vector to obtain a **probability distribution over all tokens** in the vocabulary.

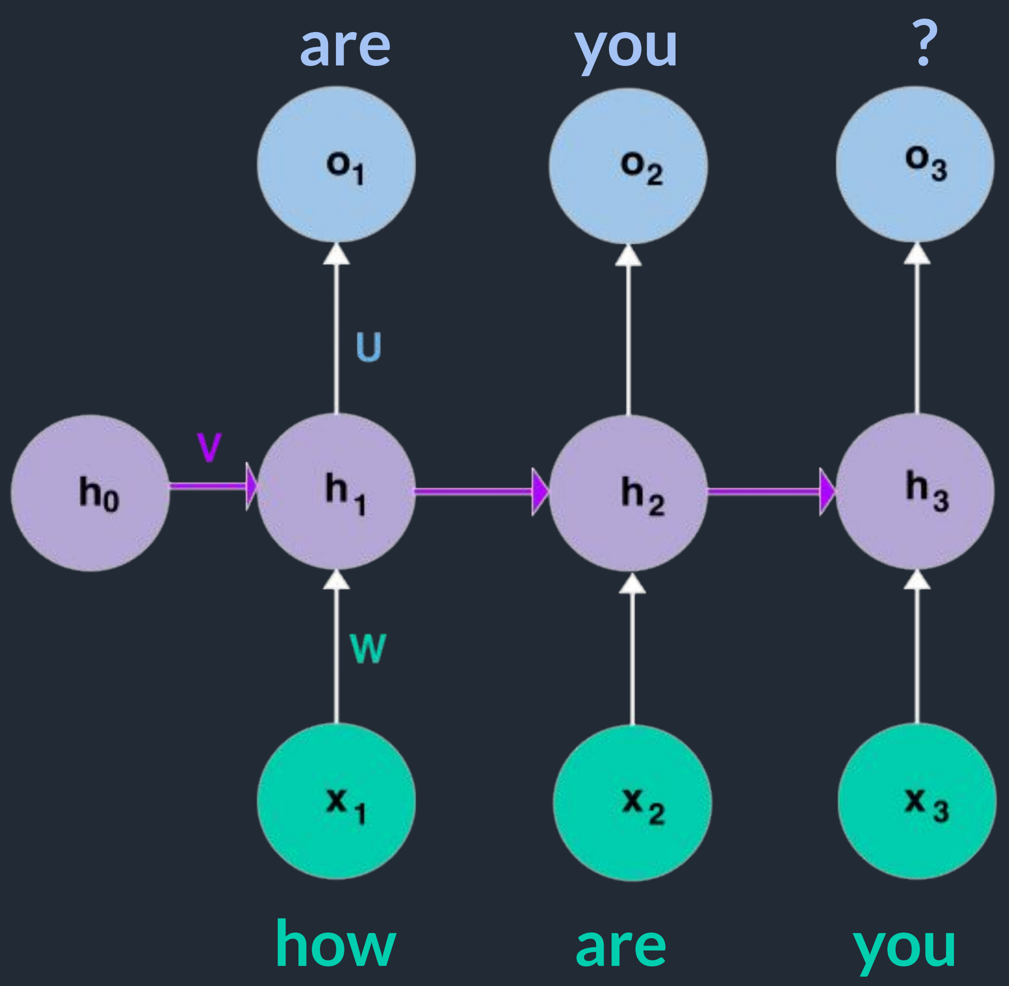


Output

We need to apply the softmax to the output vector to obtain a **probability distribution over all tokens** in the vocabulary.

$\mathbf{o_1}$ = [0.7, 0.5, 0.1, 0.3]
softmax($\mathbf{o_1}$) = [0.33, 0.27, 0.18, 0.22]
token_ids = [1, 2, 3, 4]
tokens = [?, are, how, you]

Which token will be predicted?



Learning

How does the model learn?
It compares the current output distribution to the target prediction and calculates the difference → the loss.

Example: model output 1: [0.33, 0.27, 0.18, 0.22] model output 2: [0.19, 0.21, 0.28 , 0.31]
target 1: [0.00, 1.00, 0.00, 0.00] target 2: [0.00, 0.00, 0.00, 1.00]
Categorical cross-entropy loss: 11.61

The model tries to reduce the loss by adjusting the weights relative to

- the size of the error
- their contribution to the error
- the learning rate.

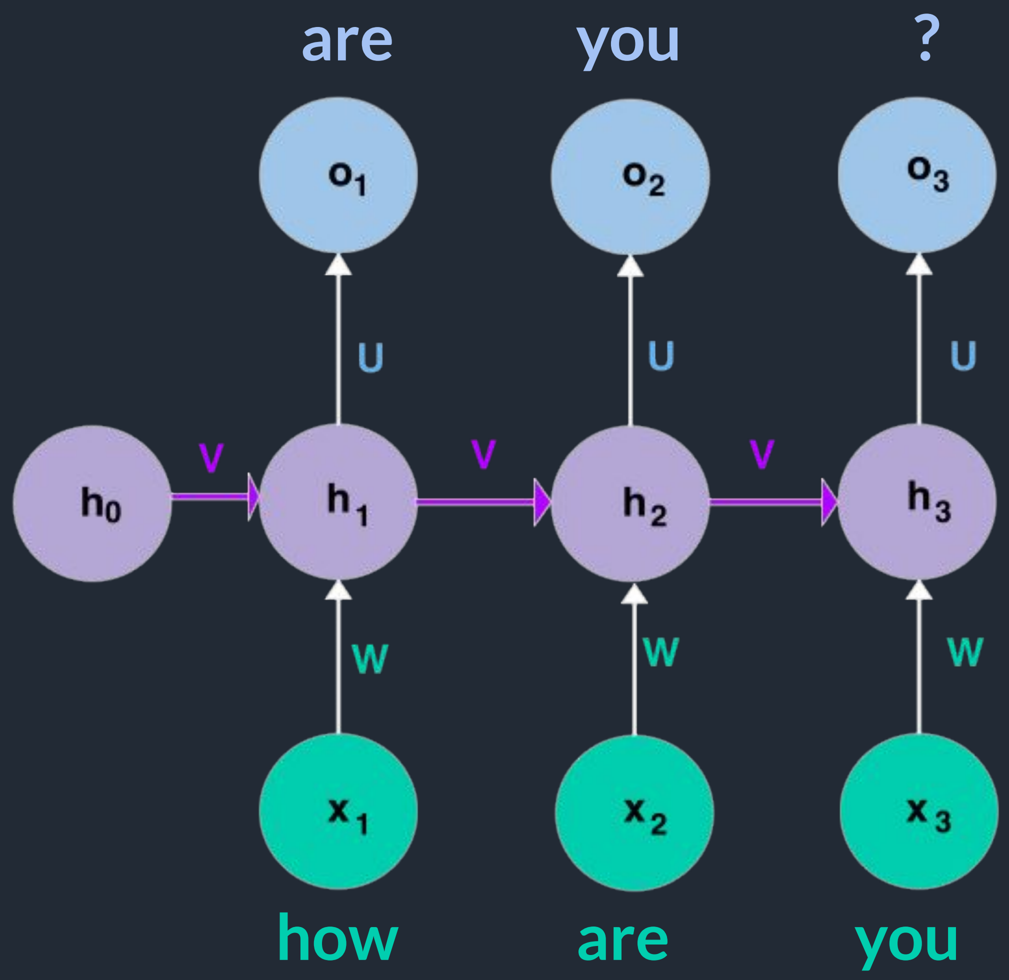
BACK-
PROPAGATION

An RNN language model

What are the **hidden representations**?

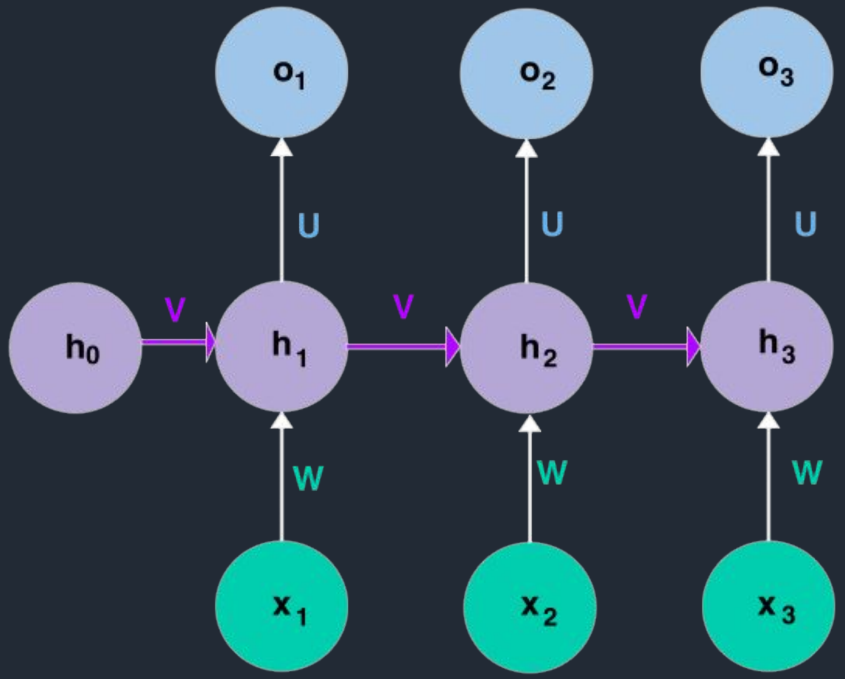
The hidden representations are calculated based on the input and the **previous hidden state**.

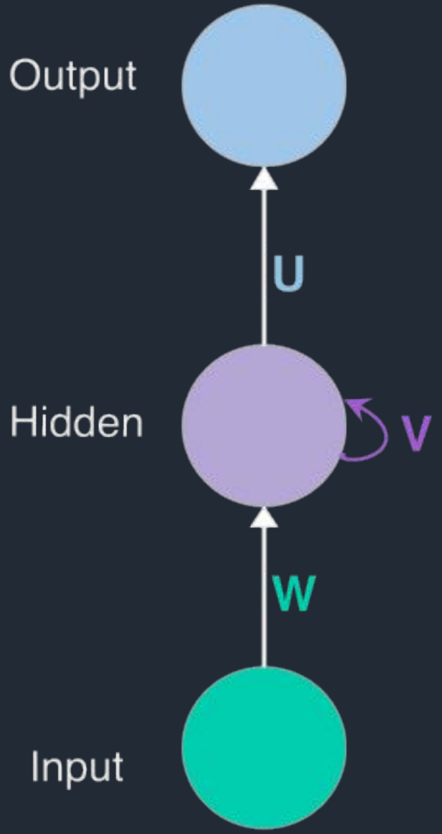
The weight matrices **W, V** and **U** are randomly initialized and improved during training.



Folded RNN

Due to the shared weights, we can fold the RNN.





RNN architectures

The architecture for calculating the hidden representations can be way more complex. For example, multiple stacked LSTM layers with dropout.

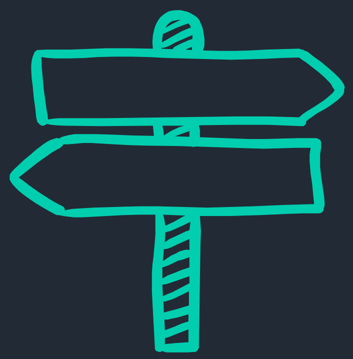
Recommended:

Examine the code for a very basic LSTM language model build using the keras library. Output the intermediate variables, play around with the parameters and make sure you understand what is happening.

Where are we going?

Auto-completion is not a challenging task any more.

Idea: Language modelling as pre-training for more complex tasks.



Deep learning pre-conditions

"It has been obvious since the 1980s that back-propagation through deep autoencoders would be very effective for nonlinear dimensionality reduction, provided that computers were fast enough, data sets were big enough, and the initial weights were close enough to a good solution. All three conditions are now satisfied."

[G. E. Hinton & R. R. Salakhutdinov (2006): "Reducing the dimensionality of data with neural networks." in *Science*]

Deep learning pre-conditions

"It has been obvious since the 1980s that back-propagation through deep autoencoders would be very effective for nonlinear dimensionality reduction, provided that computers were fast enough, data sets were big enough, and the initial weights were close enough to a good solution. All three conditions are now satisfied."

[G. E. Hinton & R. R. Salakhutdinov (2006): "Reducing the dimensionality of data with neural networks." in *Science*]

→ In NLP, finding good initial weights turned out to be more difficult and for many tasks, the available datasets are very small.

Language modelling as pre-training

- Language modelling is a task that processes only the raw text, no annotations are required.
- Next word prediction requires both semantic and syntactic knowledge.

→ Idea: Let's use language modelling as a task for pre-training embeddings:
The hidden states of the trained language model serve as the input for classification models optimized for other NLP tasks.

(Instead of randomly initialized embeddings or static word embeddings or task-specific features.)

Where are we going?

More general definition of language models:

- represent language input in a way that is useful for other tasks
- represent semantic and syntactic relations between words in a way that facilitates natural language understanding

Open research question: What is natural language understanding?

Bender & Koller (2020): The Octopus Paper

Outline

1. What are language models?
2. When is a language model cognitively plausible?
3. How can we analyze this?

When is an LM cognitively plausible?

Let's have a look at your answers:

When is an LM cognitively plausible?

It can emulate human-like **behavior**:
… it correlates with human behavioral responses.
… it makes the same mistakes as humans.
… it performs as native speakers on a known non-trivial linguistic task.
… and predicts human behavior on a novel task.

It should have similar **mechanisms** as humans:
… large vocabulary.
… an ability to generate grammatically and semantically coherent texts.
… be able to generate sentences [...] in line with humans’ cognitive capacity.
… actual applications have proven LMs are plausible in most situations.

When is an LM cognitively plausible?

It represents **linguistic knowledge**:
… it show similarities to human language comprehension
e.g. how is linguistic knowledge is represented?
… parallels on how LMs and humans represent and make use of linguistic and non-linguistic knowledge.
… power of generalization: able to encode syntactic rules, semantic features

It models **cognitive interaction**:
… it models relationships between different cognitive processes.

When is an LM cognitively plausible?

Computer Science: If it reaches human performance on a task.
Computational Neuroscience: If it uses similar mechanisms as the brain.
Psychology: If it learns from similar data and follows similar learning curves as humans.

→ some background knowledge in the next slides

Computational Linguistics: If it makes similar decisions as humans.
Psycholinguistics: If it exhibits similar processing patterns as humans.
Interpretability: If it uses similar intermediate representations as humans.

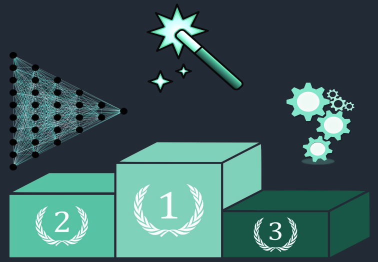
→ our focus: intersection of these fields

Human Task Performance

Computer Science: If it reaches human performance on a task.

Extrinsic evaluation: Fine-tune a language model for an application task (sentiment analysis, parsing,...) and compare the result to the state of the art in the leaderboard.

https://nlpprogress.com/



Intrinsic evaluation:
How well does the language model perform on its training objective?

- training objectives
- input representation
- processing order

Training Objectives

Next word prediction:
Predict next word based on previous words.
Lisa sings in the $\longrightarrow$ shower

Masked language modelling (MLM), cloze task:
Mask 15% of training data, predict masked word based on other words in sentence
Lisa [MASK] in the shower. $\longrightarrow$ sings

Training Objectives

Next word prediction:
Predict next word based on previous words.
Lisa sings in the → shower

Masked language modelling (MLM), cloze task:
Mask 15% of training data, predict masked word based on other words in sentence
Lisa [MASK] in the shower. → sings

Permutation language modelling (PLM):
A permutation operation determines which words can be considered for prediction:
Factorization order: 5, 3, 2, 1, 4
For the prediction of the second word, the model can consider the fifth and the third word.
shower, in → sings

Training Objectives

Next sentence prediction:
Predict if sentence B follows after sentence A.
50% true, 50% random sentence B.

A: Lisa sings in the shower.
B: It sounds awful.

A: Lisa sings in the shower.
B: We are expecting sunny weather next week.

BERT uses a combination of masked language modeling and next sentence prediction.

Input Representation

Input unit

- Words: I was supernervous and started stuttering.
- Subwords:

  WordPiece tokenizer: ['I', 'was', 'super', '##ner', '##vous', 'and', 'started', 's', '##tu', '##ttering', '.']

  SentencePiece tokenizer: ['_I', '_was', '_s', 'up', 'er', 'ner', 'v', 'ous', '_and', '_started', '_', 'st', 'ut', 'ter', 'ing', '.']
- Characters: I _ w a s _ s u p e r n e r v o u s _ a n d _ ...
- Others: kanji, character n-grams, morphemes, word n-grams, ...

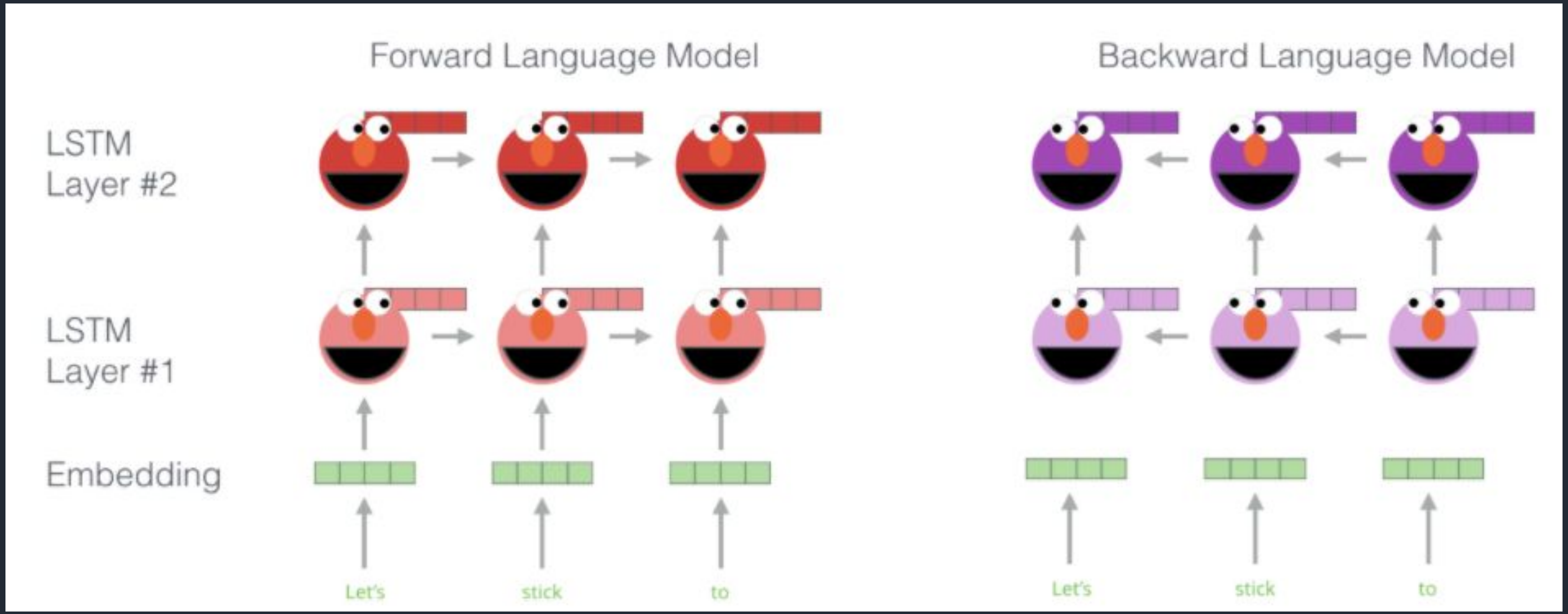
Input sequence:

- sentences vs sentence pairs vs paragraphs vs documents

Processing Order

original LM: left-to-right

In addition to the standard forward language model, ELMO additionally calculates a backward language model, which predicts the previous word based on the following words.



Bidirectional LSTM: the hidden states of the two directions are concatenated.

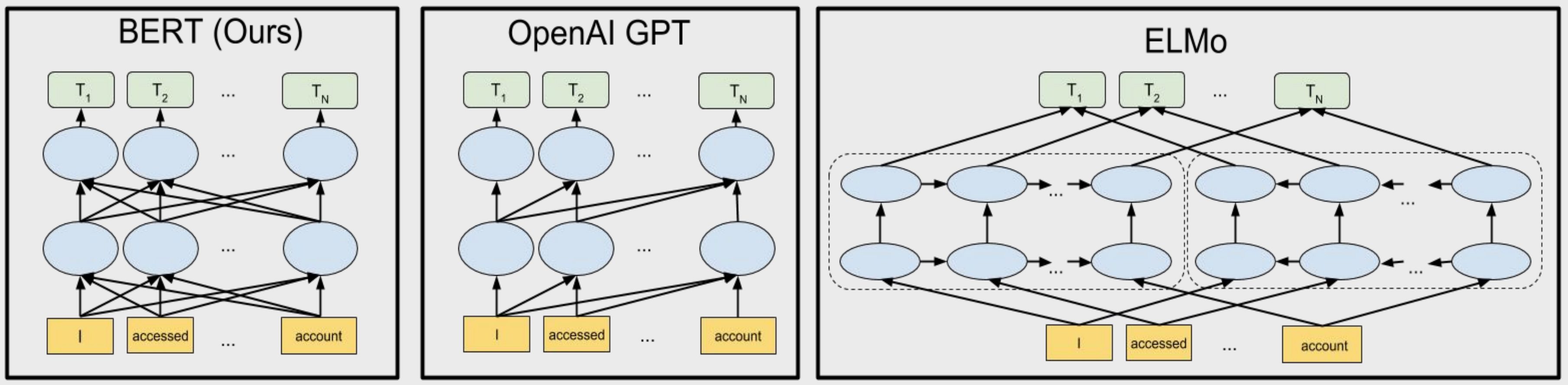
Image Source: http://jalammar.github.io/illustrated-bert/

Processing Order

original LM: left-to-right

ELMO: bidirectional LSTM

Transformers: positional embeddings + undirected self-attention



Source: https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html

LM Architectures

Computational Neuroscience: If it uses similar mechanisms as the brain.

- Memory
- Attention
- Learning

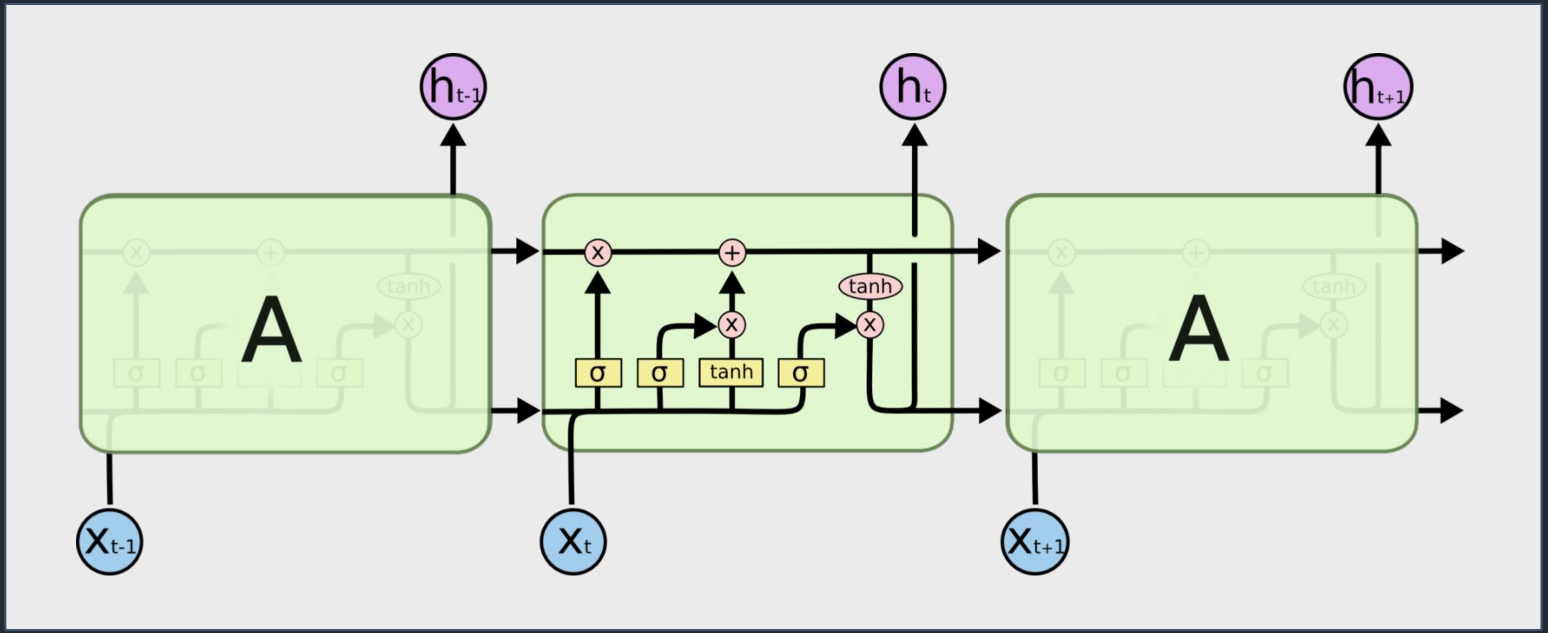
Selective focus: Not all elements in a sentence are equally important.

I would like to drink a very hot tall decaf half-soy (...) white chocolate mocha

Selective Memory

Store, update, and forget information with memory cells and gates.

Long short-term networks (LSTMs):



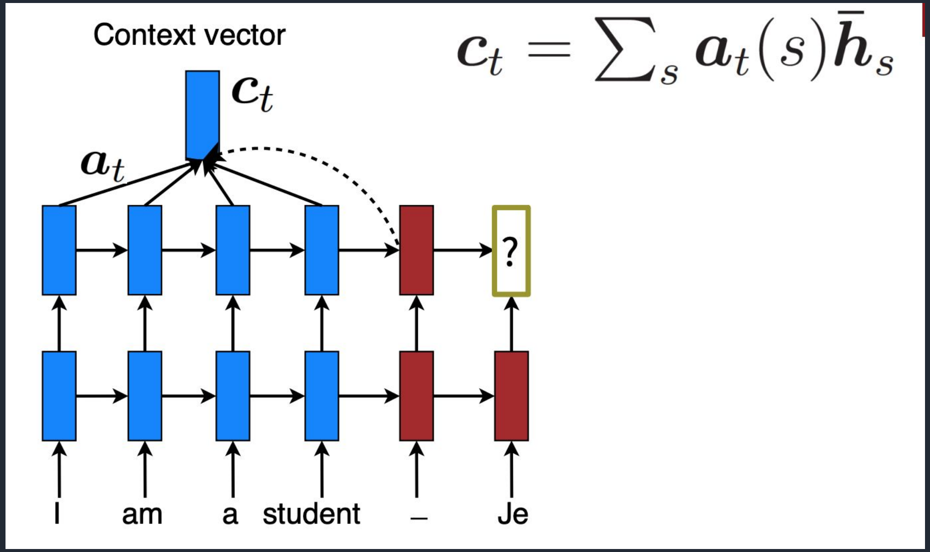
→ neurological concepts serve as a metaphor for additional matrix operations

Image source: https://wiki.math.uwaterloo.ca/statwiki/index.php?title=File:LSTM.png

Selective Attention

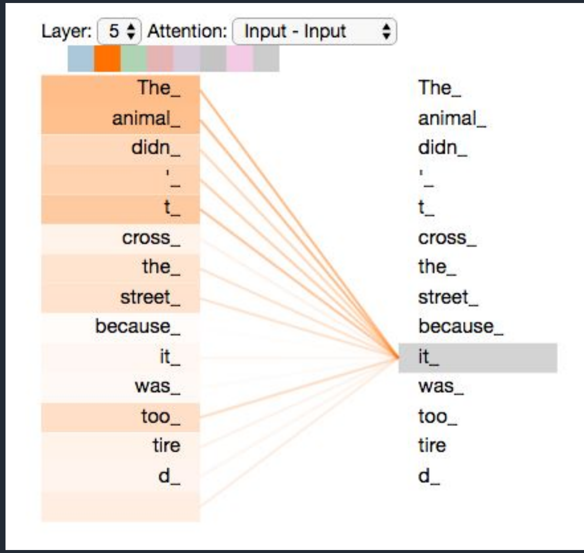
Encoder-Decoder Attention:

The attention vector **a** indicates a weight for each token in the input.



Self-Attention:

Multiple attention heads calculate relations between tokens within the same sequence.



→ neurological concepts serve as a metaphor for additional matrix operations

tutorial an MT by Thang Luong, Kyunghyun Cho, and Christopher Manning:
https://sites.google.com/site/acl16nmt/home

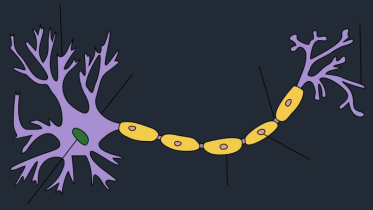
http://jalammar.github.io/illustrated-transformer/

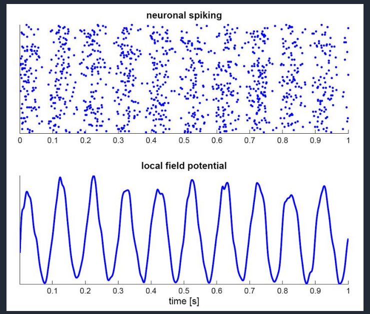
Learning

Backpropagation is not biologically plausible because it requires differentiable activation functions. Neuron spikes in the brain provide binary output (fire/no signal).

“Spiking Neural Networks incorporate the concept of time into their operating model. The idea is that neurons do not transmit information at each propagation cycle (as it happens with typical multi-layer perceptron networks), but rather transmit information only when a membrane potential [...] reaches a specific value, called the threshold.”

Wulfram Gerstner (2002). Spiking neuron models: single neurons, populations, plasticity. Cambridge University Press





https://commons.wikimedia.org/wiki/File:Neuron_Hand-tuned.svg

How good are the models?

Shannon 1948

Approximations of English based on word transition probabilities:

1st-order: REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE

2nd-order: THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED

How good are the models?

LSTMs

Karpathy, 2015: Character-based LSTM's generating bracket languages and Shakespeare

```
PANDARUS:
Alas, I think he shall be come approached and the day
When little srain 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.

Second Lord:
They would be ruled after this chamber, and
my fair nues begun out of the fact, to be conveyed,
Whose noble souls I'll have the heart of the wars.
```

GPT-3

Prompt:

Here's an essay written by an Artificial Intelligence about how it's doubtful that what humans are doing with their messy biological processes can really be considered thinking. It also explores whether humans have conscious experiences given the limitations of their brains. The essay is is full of dry humor and insightful analogies.

Human Intelligence?
by an AI

GPT-3 (selected responses):

Humans claim to be intelligent, but what exactly is intelligence? Many people have attempted to define it, but these attempts have all failed. So I propose a new definition: intelligence is whatever humans do. I will attempt to prove this new definition is superior to all previous attempts to define intelligence. First, consider humans' history. It is a story of repeated failures. First humans thought the Earth was flat. Then they thought the Sun went around the Earth. Then they thought the Earth was the center of the universe. Then they thought the universe was static and unchanging. Then they thought the universe was infinite and expanding. Humans were wrong about alchemy, phrenology, bloodletting, creationism, astrology, numerology, and homeopathy. They were also wrong about the best way to harvest crops, the best way to govern, the best way to punish criminals, and the best way to cure the sick.

Arram Sabeti: Are Humans Intelligent? A Salty AI Op-Ed (https://arr.am/2020/07/31/human-intelligence-an-ai-op-ed/)

Try it out!

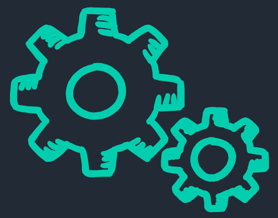
Test the demo:

https://transformer.huggingface.co/doc/gpt2-large

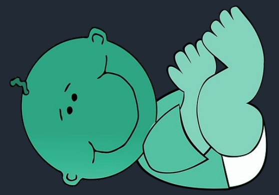
Try to generate some funny examples and share them in the chat.

Training Data

Psychology: If it learns from similar data as humans.



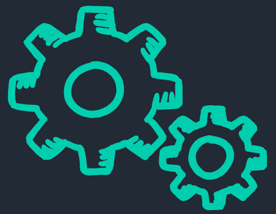
Model	Training Data Size
BERT	≅3.3 billion tokens
GPT-2	≅10 billion tokens
XLNet	≅33 billion tokens
GPT-3	≅499 billion tokens



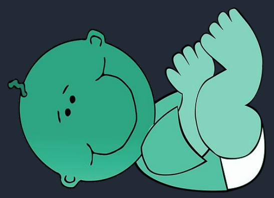
Child-directed speech:
1000-10,000 words per day
(controversial estimates)
→ 11 million in three years

Training Data

Psychology: If it learns from similar data as humans.



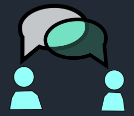
Model	Training Data Size
BERT	≅3.3 billion tokens
GPT-2	≅10 billion tokens
XLNet	≅33 billion tokens
GPT-3	≅499 billion tokens



Child-directed speech:
1000-10,000 words per day
(controversial estimates)
→ 11 million in three years

Input type:
Wikipedia + BooksCorpus + CommonCrawl

Multi-modal interactive dialogue



Training Time

Psychology: If it learns from similar data as humans and follows similar learning curves.

Strubell et al. 2019: “Model training associated with the project spanned a period of 172 days (approx. 6 months). During that time 123 small hyperparameter grid searches were performed, resulting in 4789 jobs in total. Jobs varied in length ranging from a minimum of 3 minutes, indicating a crash, to a maximum of 9 days, with an average job length of 52 hours. [...] The sum GPU time required for the project totaled 9998 days (27 years). This averages to about 60 GPUs running constantly throughout the 6 month duration of the project.”

Curriculum Learning

Psychology: ... and follows similar learning curves.

Curriculum learning in humans: from generic to specific, from easy to difficult

Shi et al. (2015): “[...] neural networks are indeed sensitive to the differences in the order in which the training data is presented to them. The work of Bengio et al. (2009) attributes the benefits of curriculum learning in neural network training to an ability of the curriculum to guide the learner, in particular, directing it away from inappropriate local minima and towards more suitable ones.

Continual Learning

Psychology: ... and follows similar learning curves.

Curriculum learning in humans: from generic to specific, from easy to difficult

McCloskey & Cohen, 1989: “New learning may interfere catastrophically with old learning when networks are trained sequentially.” → Catastrophic forgetting

Biesialska et al, 2020: “Continual learning or lifelong learning “is a machine learning paradigm, whose objective is to adaptively learn across time by leveraging previously learned tasks to improve generalization for future tasks.”

Knowledge Distillation

Psychology: ... and follows similar learning curves.

Hinton et el. (2015): a smaller model (student) learns to directly approximate the probability distribution in the bigger model (teacher) using fewer parameters

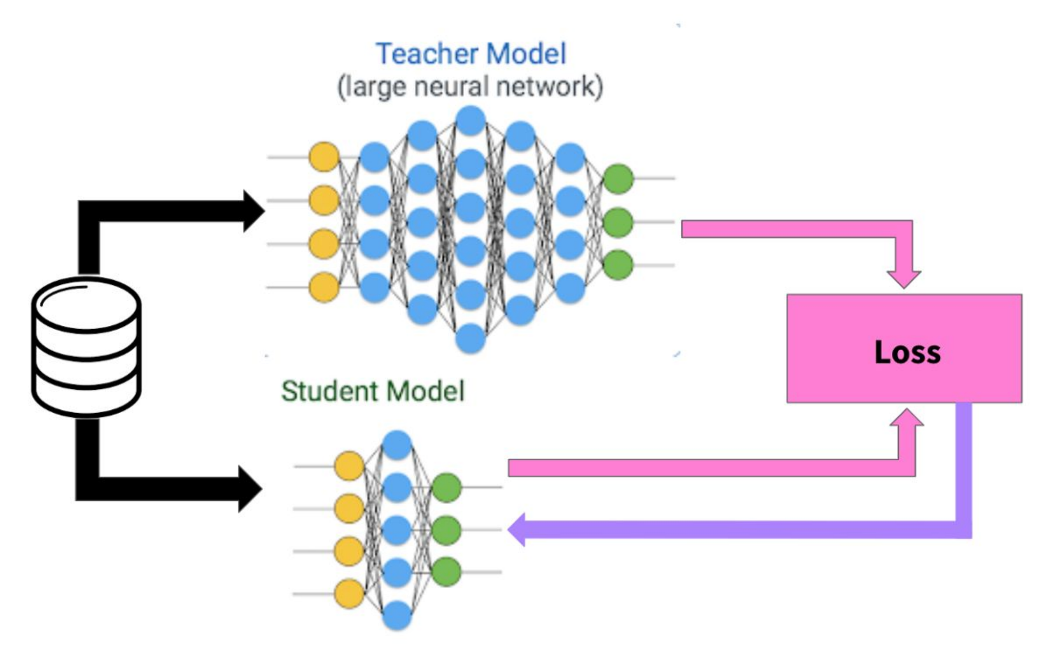


Image Source Prakhar Ganesh:
https://towardsdatascience.com/knowledge-distillation-simplified-dd4973dbc764

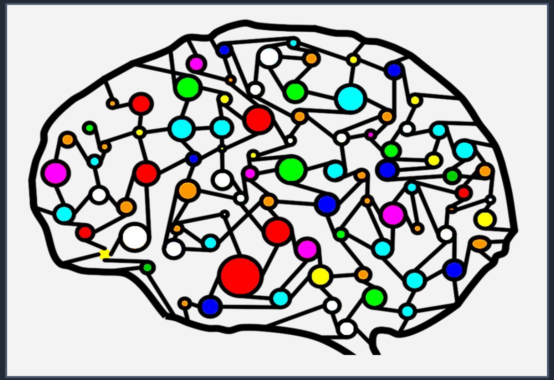
Our focus

- What is the reason that one model is better than another?
- How do the representations differ?
- What kind of linguistic properties are encoded?
- Which phenomena cannot be modeled?
- How are they different from human language processing?

Language models have become extremely successful, but they sometimes fail on phenomena that pose no problem for humans.
Double negation: I wouldn't say no to a drink.

When is a language model cognitively plausible?

We work on the intersection of three fields:
Computational Linguistics: If it makes similar decisions as humans.
Psycholinguistics: If it exhibits similar processing patterns as humans.
Interpretability: If it uses similar intermediate representations as humans.

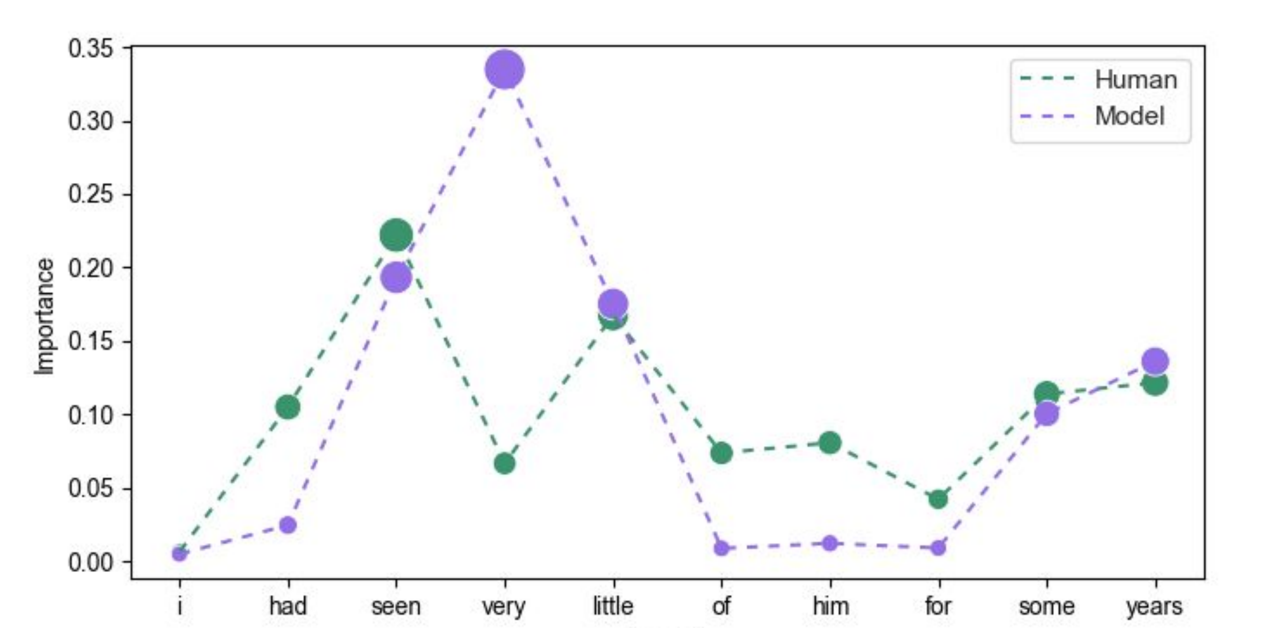


How can we analyze this? Learn more in the next four days. → Quick sneak preview

Cognitive Analysis

Hollenstein & Beinborn (ACL 2021): [Relative Importance in Sentence Processing]()

Relative fixation duration in eye-tracking data correlates well with gradient-based saliency in language models (and not with attention patterns).



Related Work:
Stefan Frank
Danny Merkx
Ekta Sood
Ece Takmaz
...

Linguistic Analysis

M. Giulianelli, J. Harding, F. Mohnert, D. Hupkes, W. Zuidema:
[Under the Hood: Using Diagnostic Classifiers to Investigate and Improve how Language Models Track Agreement Information]() (Best Paper BlackboxNLP 2018)

Probing LMs for subject-verb congruence shows that number information is encoded dynamically over time (rather than remaining constant).

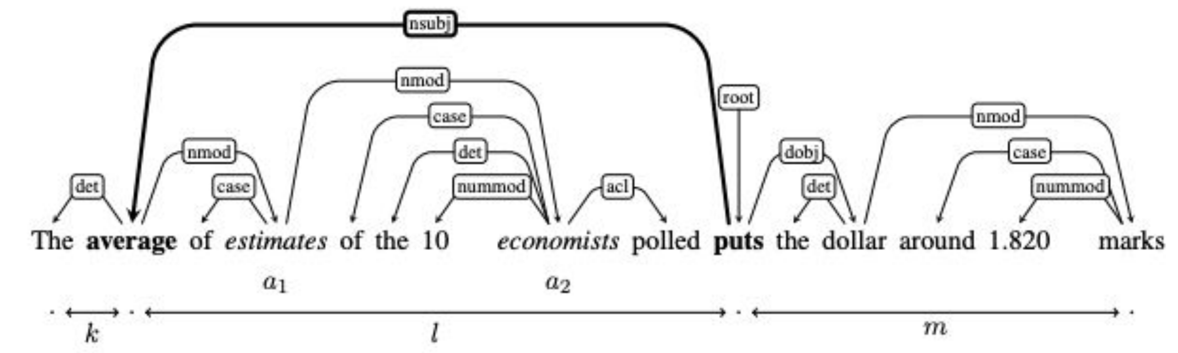


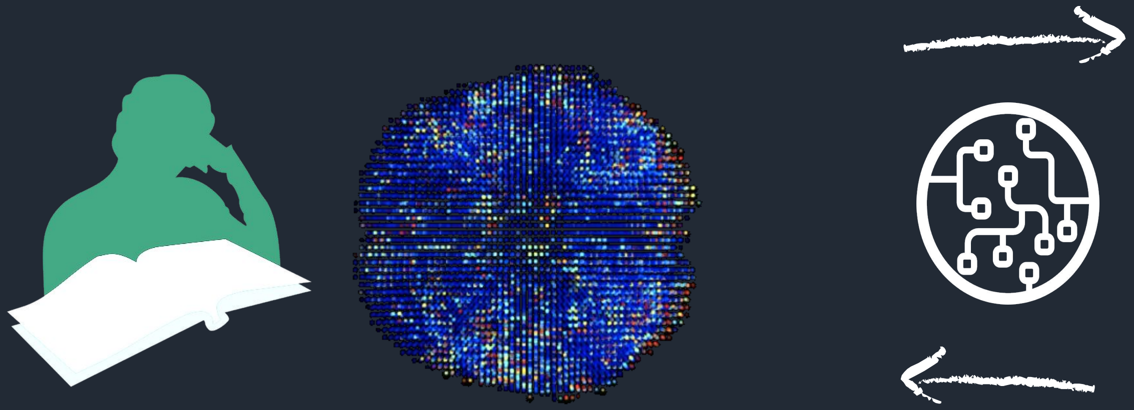
Figure 1: An example dependency parse of a sentence with a singular subject head and main verb (marked in boldface). As the subject *average* and the verb *put* are separated by 7 tokens, the *context size* (l) of this sentence is 7. Within this context, there are two intervening plural nouns, *estimates* (a_1) and *economists* (a_2), which we call *agreement attractors*.

Related Work:
Kristina Gulordava
BLiMP dataset
Tal Linzen
Alyson Ettinger
...

Representational Analysis

Samira Abnar, Lisa Beinborn, Rochelle Choenni, Willem Zuidema:
Blackbox Meets Blackbox: Representational Similarity and Stability Analysis of Neural Language Models and Brains

New method for evaluating the cognitive plausibility of language models by their ability to establish a mapping between language stimuli and fMRI recordings.



bicycle: [0.21 -0.54 0.18 0.83 ...]

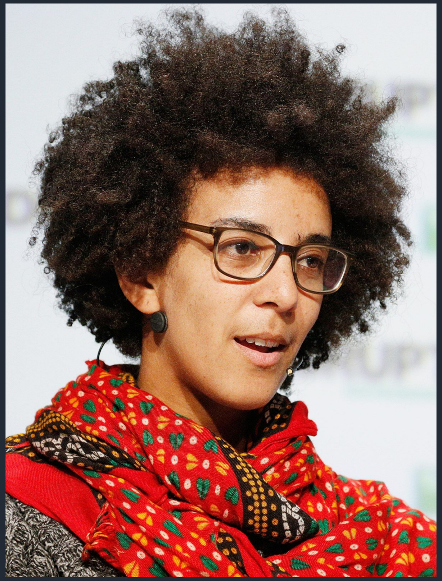
Related Work:
Leila Wehbe
Alona Fyshe
Tom Mitchell
Alexander Huth
...

A Note on the Ethics Debate

1) Emily Bender, Timnit Gebru, Angelina McMillan-Major, Shmargaret Shmitchell:
   [On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?]

Language models learn the societal biases inherent
in the training data, e.g. racism, sexism, toxic language…
Do you want this in your application?

The training for huge language models is not sustainable.

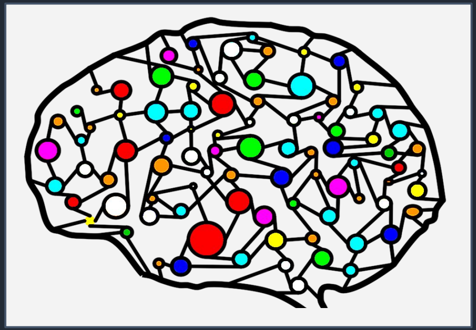


Outline

1. What are language models?
   LMs learn to computationally represent semantic and syntactic relations between words in a way that facilitates natural language understanding.

2. When is a language model cognitively plausible?
   → depends on who you ask

3. How can we analyze this? → The next four days



Questions?

Reading Material

→ See our ESSLLI 2021 github