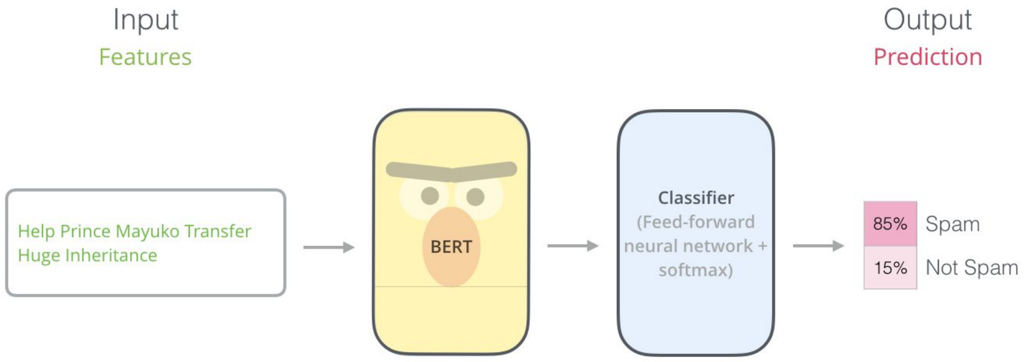
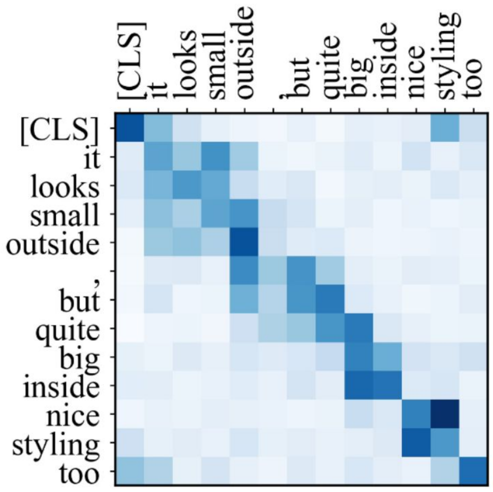
Interpretability

What happens inside a LM?

Interpretability of Deep Neural Networks

- Neural Networks as “black boxes”?
- Observing the operations inside a LM is easy: just print every step
- These prints are hard to *interpret*
- How does a NN make predictions?
- e.g. how is attention distributed across context words?

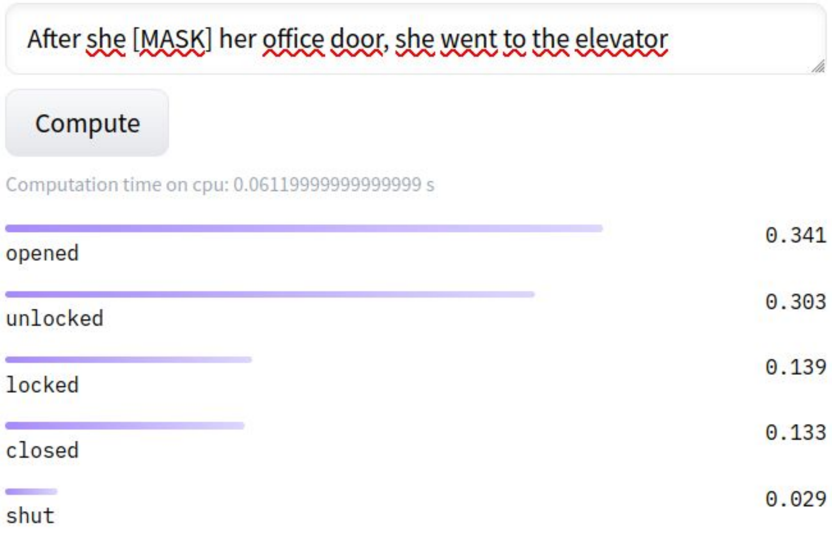




IMG 1: https://jalammar.github.io/illustrated-bert/ IMG 2: Cui et al. 2019 https://aclanthology.org/D19-1361.pdf

Central Questions

- Which kinds of knowledge are (not) used by the LM?
- Does a LM make similar linguistic generalizations as humans?
- (How) does the LM use
  - POS
  - syntactic structure?
  - semantic word fields?
- Can this knowledge be localized?
  - Layers/Neurons/Attention heads



Ex. from DistilBERT on https://huggingface.co/distilbert-base-uncased

Interpretability and Explainability in Machine Learning

- some use these terms interchangeably (e.g. Søgaard 2021)
- Clinciu and Hastie (2019): “interpretability as intersecting with explainability as some models may be interpretable without needing explanations”
- Some blog posts make similar distinctions and sometimes contradict each other
- So let’s not worry about the distinction here

Why should you care?

- Legal and fairness reasons
  - Applications of AI systems: EU guidelines include “right to explanation"
  - If you build/sell a product, you should know how it works to improve it
- Linguistic reasons
  - LMs are similar to linguistic theories: Both assign probabilities to text sequences
  - Big difference: LMs are based on much more data
  - Is there evidence for linguistic assumptions in the LM?

Methods for Explainable NLP

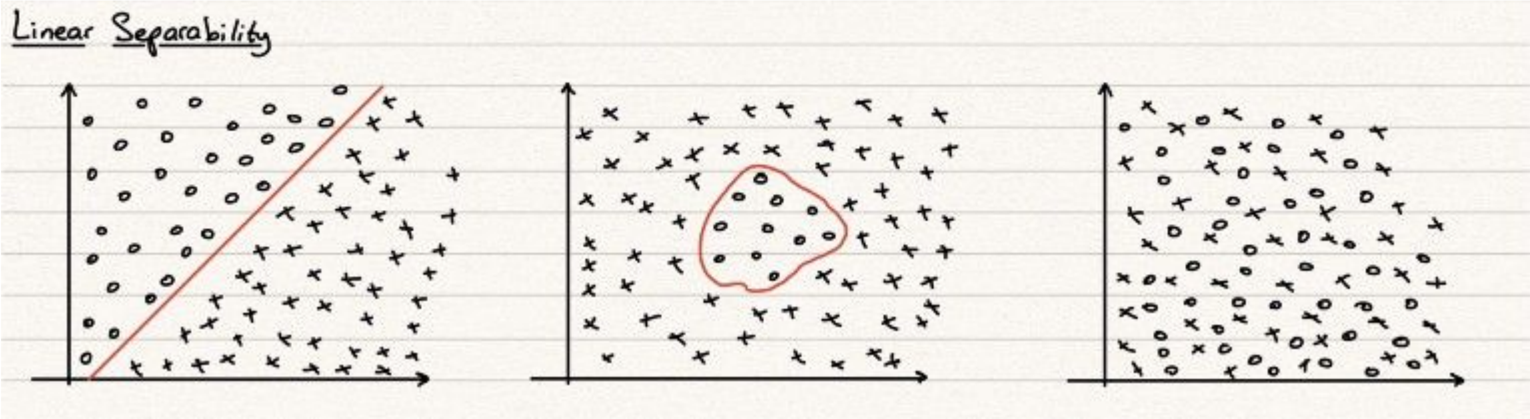
- There’s a giant, growing set of methods (Søgaard 2021 for an overview)
- Do you know some methods?

Methods for Explainable NLP

- There's a giant, growing set of methods (Søgaard 2021 for an overview)
- Do you know some methods?
- Based on forward pass:
    - Diagnostic classification: Train a simple model (POS-tagger, parser, etc.) on the LM representations and check performance
    - Observing attention patterns for individual sentences (see prev. slides)
- Based on backward pass:
    - Observing gradients, Layerwise relevance propagation: Which weights are most relevant for a prediction?
    - Weight pruning: Make the weight matrices sparser, remove some weights

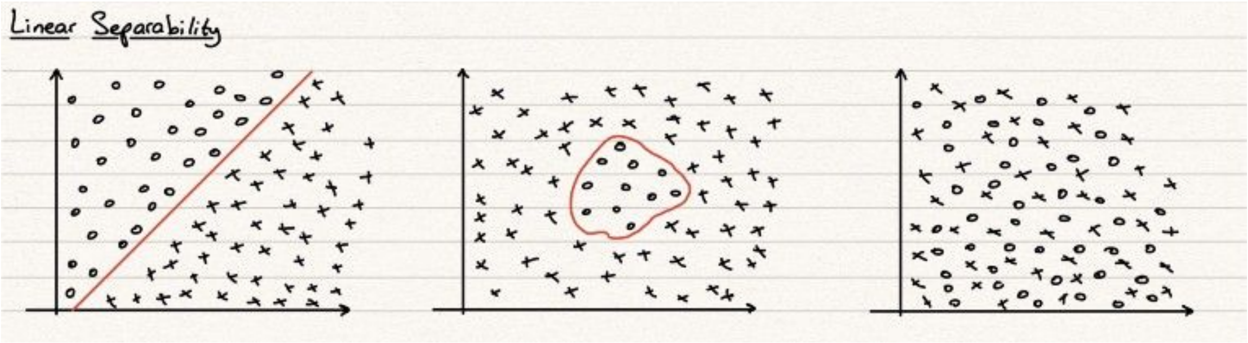
Example: Seyffarth et al. (2021): Causativity neurons

- Remember linear separability?



IMG: https://towardsdatascience.com/a-look-at-the-maths-behind-linear-classification-166e99a9e5fb

Example: Seyffarth et al. (2021): Causativity neurons



(4) a. This **affects** the calculation . *(caus)*

b. I **envy** you in that respect ! *(noncaus)*

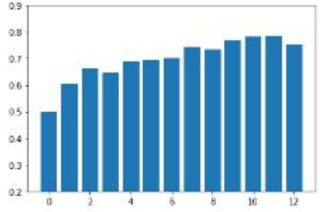
- Same idea, but with LM vectors
- o = causative, x = non-causative sentence
- 768 (BERT) instead of 2 dimensions

Example: Seyffarth et al. (2021): Causativity neurons

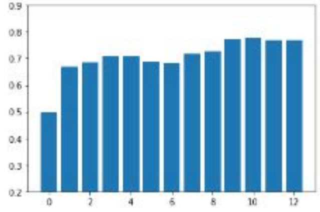
(4) a. This **affects** the calculation . *(caus)*

b. I **envy** you in that respect ! *(noncaus)*

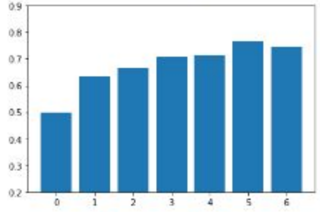
1. Collect LM representations for labeled dataset
2. Train linear classifier on LM representations
3. Good results -> LM “knows” what causativity is?
    - Are some layers more predictive than others for causativity?
    - Is there a small set of neurons that strongly correlates with causativity?



(c) D_{all}– BERT



(f) D_{all}– XLNet



(i) D_{all}– DistilBERT

NeuroX library (Dalvi et al. 2019)

- implements all steps in these experiments in python
- Used in a number of studies
  - Causativity: Seyffarth et al. 2021
  - POS-tagging, CCG supertagging, syntactic chunking, semantic tagging: Durrani et al. (2020)
- Let’s try it out!