CS 11-747 Neural Networks for NLP

Model Interpretation

Danish Pruthi

April 28, 2020

Why interpretability?

- **Task:** predict probability of death for patients with pneumonia

- **Why**: so that high-risk patients can be admitted, low risk patients can be treated as outpatients

- $AUC_{\text{Neural networks}} > AUC_{\text{Logistic Regression}}$

- Rule based classifier

HasAsthma(X) —> LowerRisk(X)

more intensive care

Example from Caruana et al.

Why interpretability?

- Legal reasons: uninterpretable models are banned!
  - GDPR in EU necessitates "right to explanation"
- Distribution shift: deployed model might perform poorly *in the wild*
- User adoption: users happier with explanations
- Better Human-AI interaction and control
- Debugging machine learning models

Dictionary definition

interpret verb

in·ter·pret | \ in-ˈtər-prət, -pət\

interpreted; **interpreting**; **interprets**

Definition of *interpret*

transitive verb

1 : to explain or tell the meaning of : present in understandable terms

// *interpret* dreams

// needed help *interpreting* the results

Only if we could understand

`model.ckpt`

As per Merriam Webster, accessed on 02/25

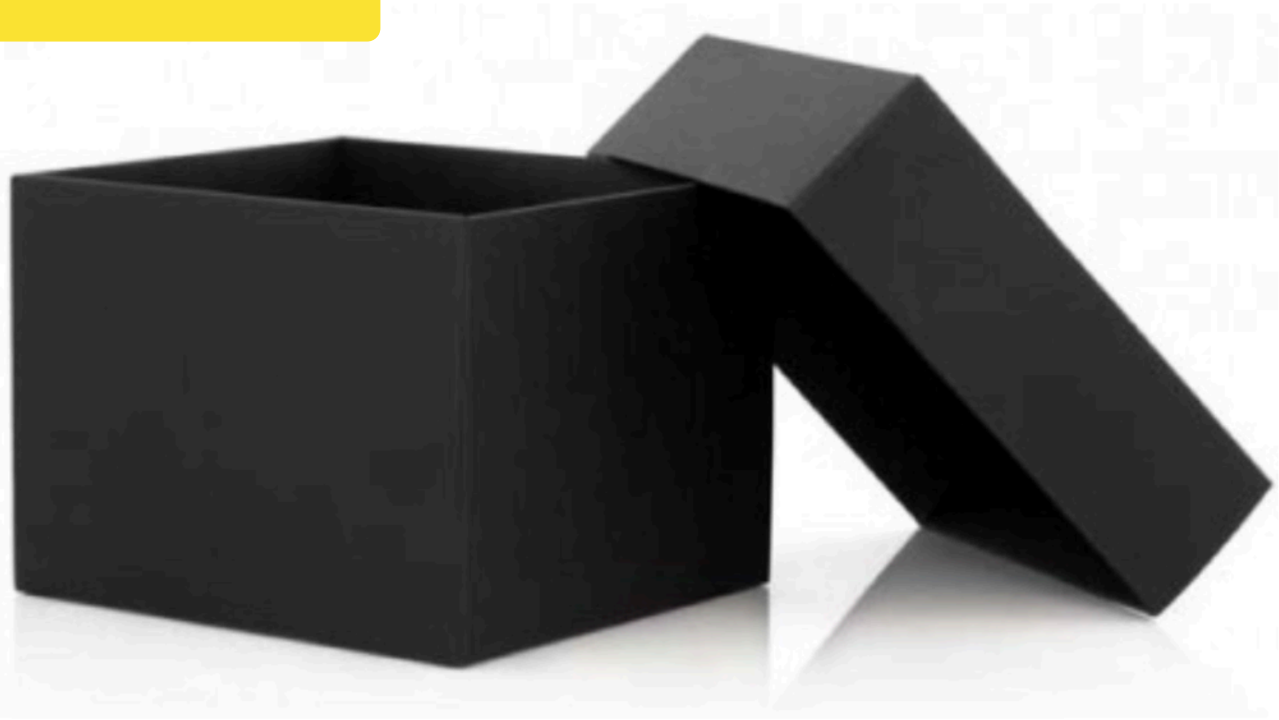
Two broad themes

global interpretation

- What is the model learning?

- Can we explain the outcome in "understandable terms"?

local interpretation



Comparing two directions

What is the model learning?

- Input: a model M, **a (linguistic) property P**
- Output: extent to which M captures P
- Techniques: classification, regression
- Evaluation: implicit

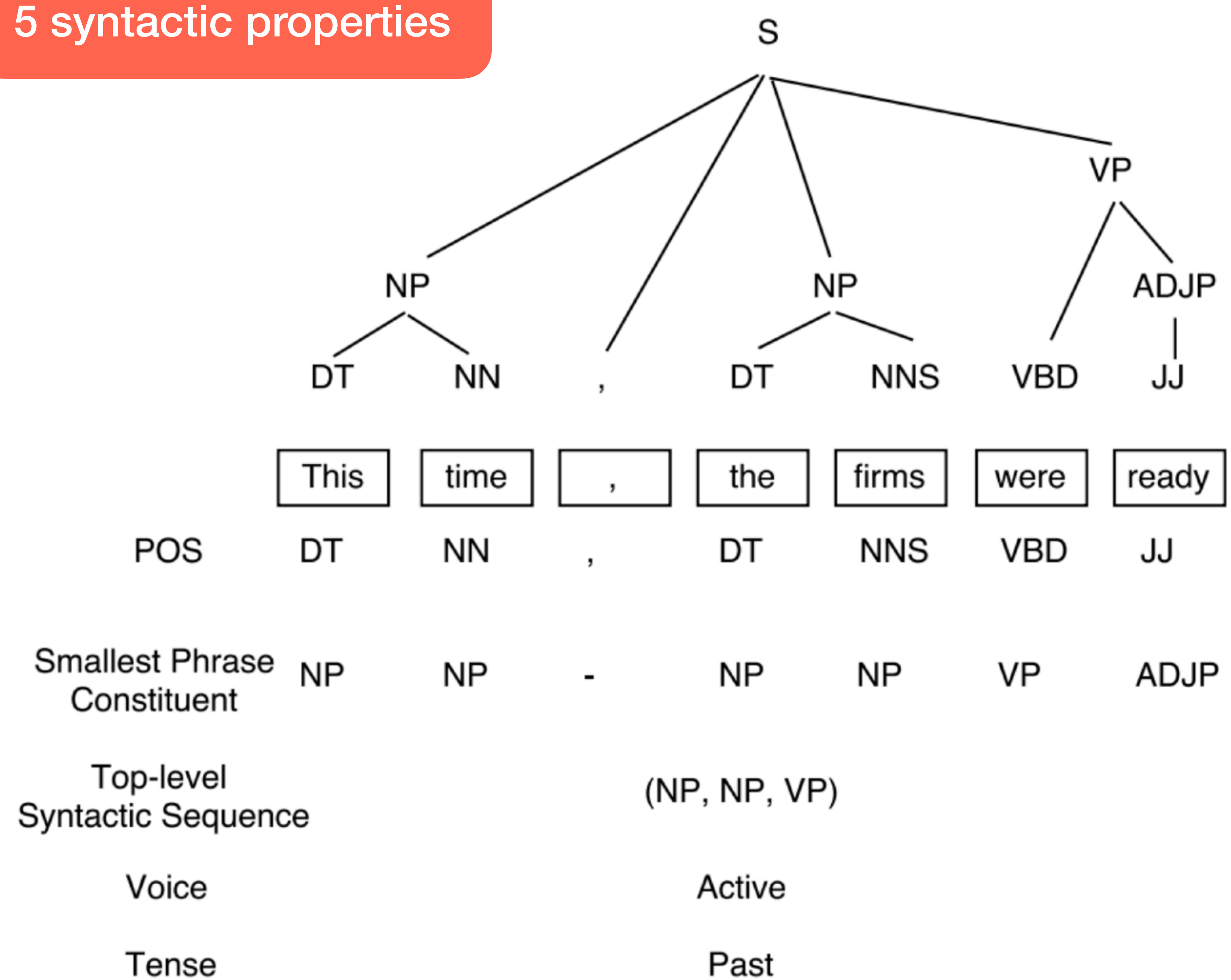
Explain the prediction

- Input: a model M, **a test example X**
- Output: an explanation E
- Techniques: varied …
- Evaluation: complicated

What is the model learning?

Source Syntax in NMT

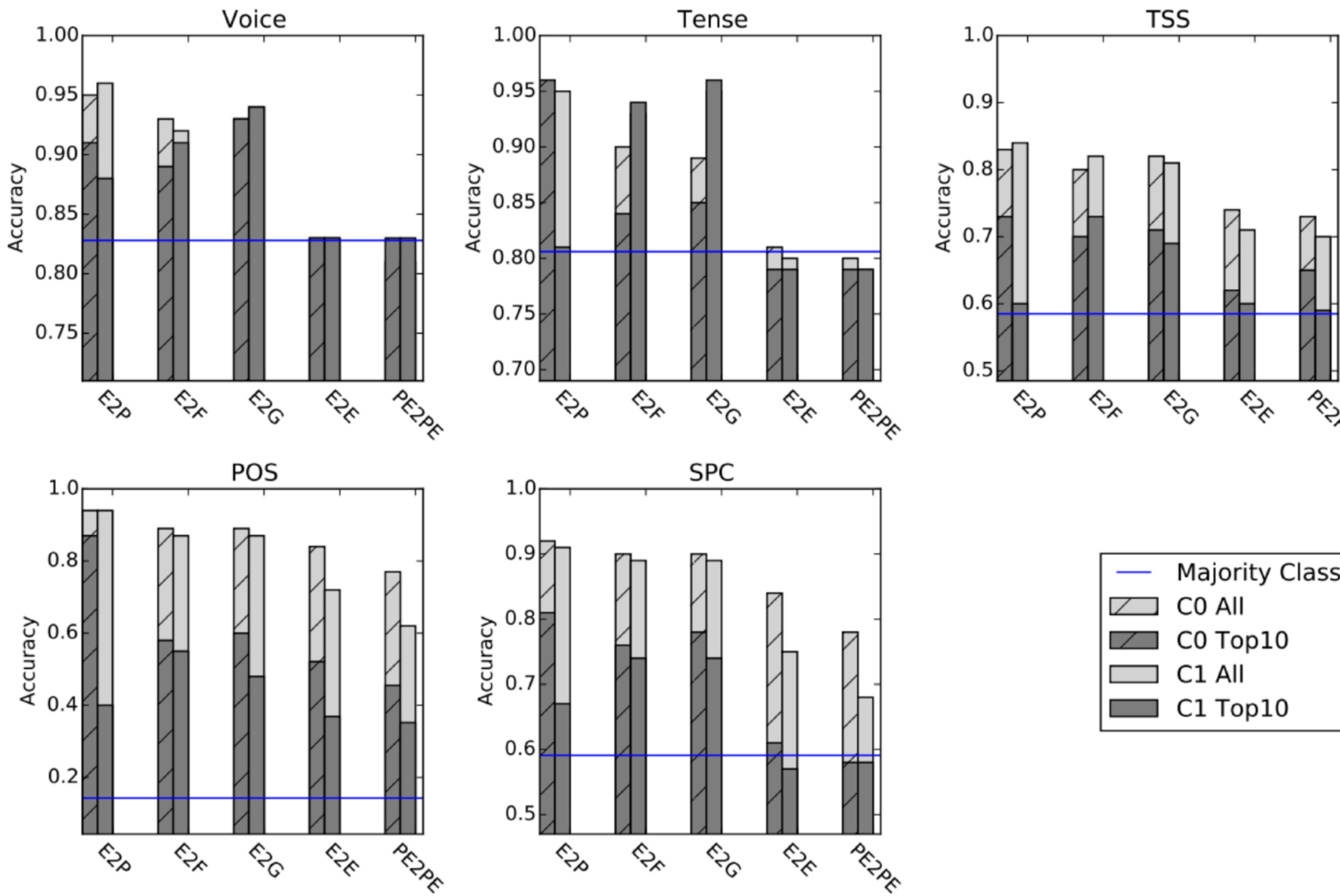
5 syntactic properties



Accuracy
82.8
92.8
82.7

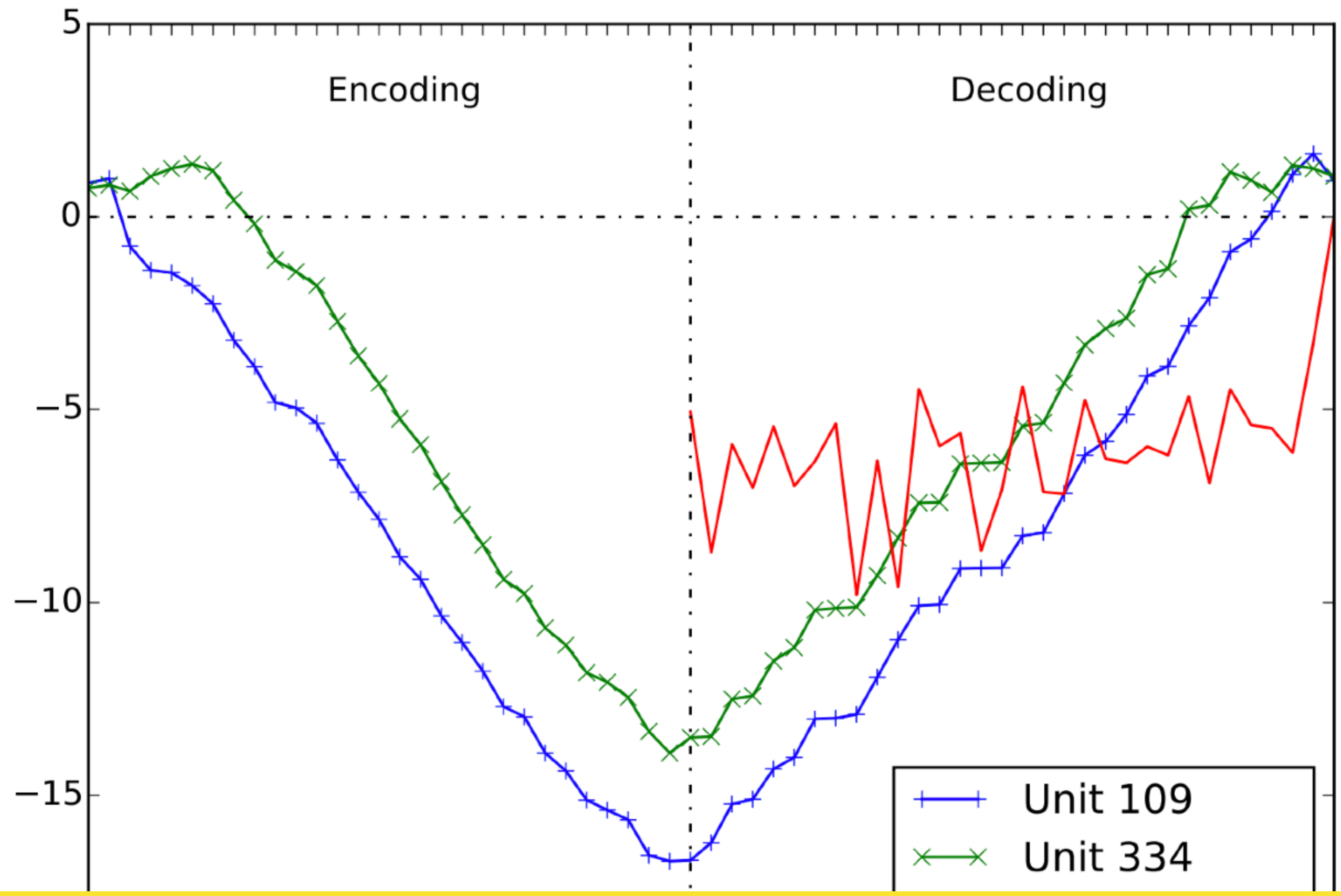
ın accuracy using the
ıajority class baseline

Source Syntax in NMT



Does String-Based Neural MT Learn Source Syntax? Shi et al. EMNLP 2016

Why neural translations are the right length?



Note: LSTMs can learn to count, whereas GRUs can not do unbounded counting (Weiss et al. ACL 2018)

Shi et al. EMNLP 2016

Fine grained analysis of sentence embeddings

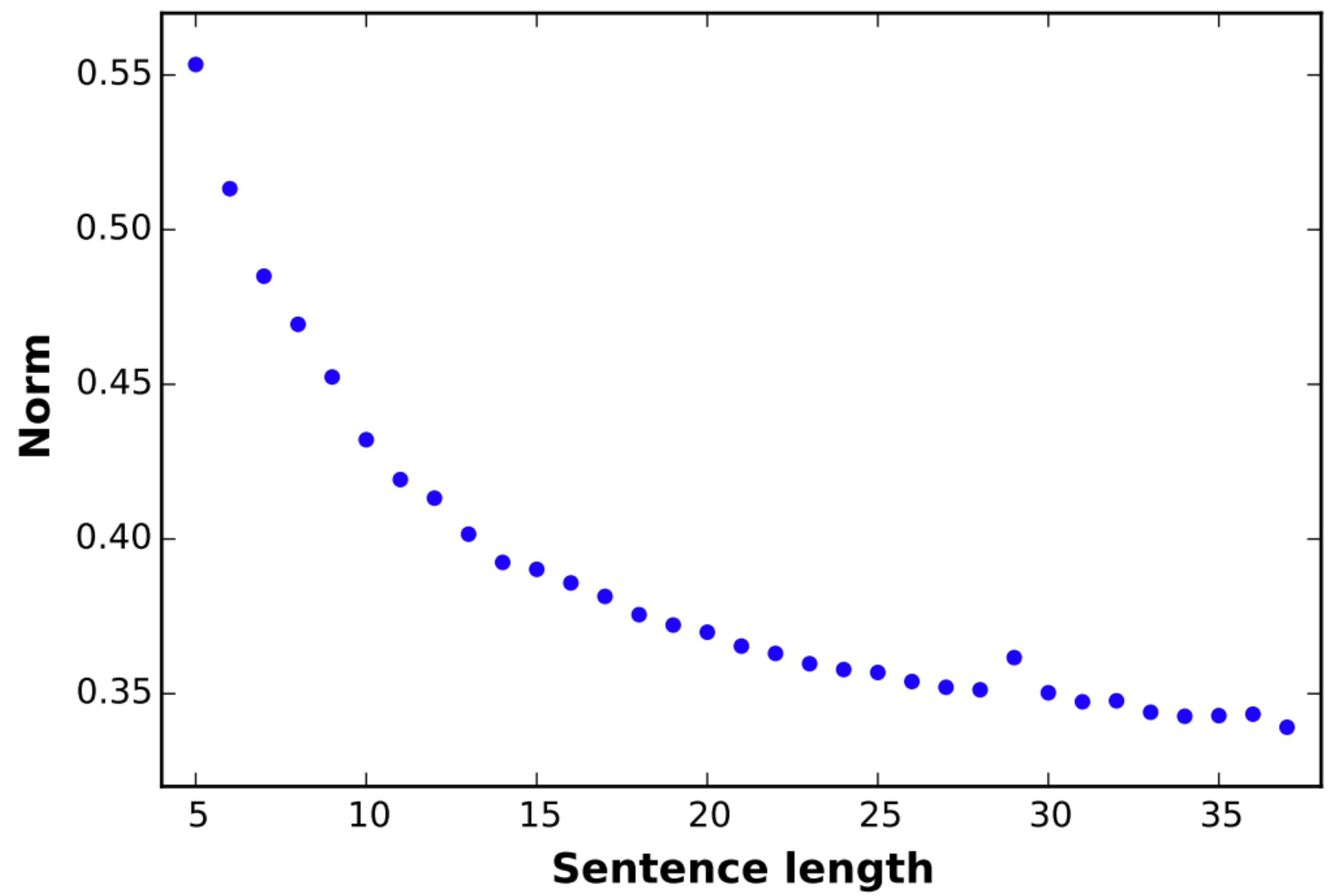
- Sentence representations: word vector averaging, hidden states of the LSTM

- Auxiliary Tasks: predicting length, word order, content

- Findings:
  - hidden states of LSTM capture to a great deal length, word order and content
  - word vector averaging (CBOW) model captures content, length (!), word order (!!)

Adi et al. ICLR 2017

Fine grained analysis of sentence embeddings



(b) Average embedding norm vs. sentence length for CBOW with an embedding size of 300.

What you can cram into a single vector: Probing sentence embeddings for linguistic properties

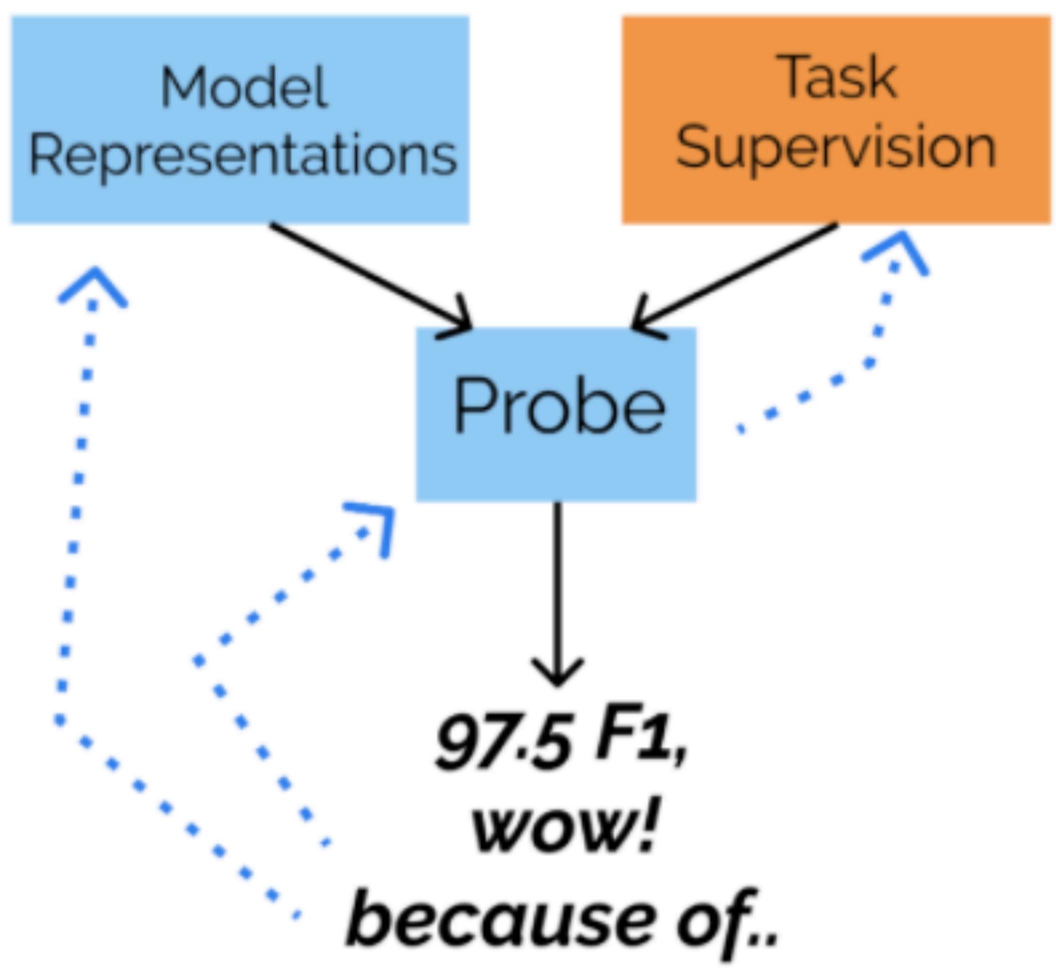
- "you cannot cram the meaning of a whole %&!$# sentence into a single $&!#* vector" — Ray Mooney

- Design 10 probing tasks: len, word content, bigram shift, tree depth, top constituency, tense, subject number, object number, semantically odd man out, coordination inversion

- Test BiLSTM last, BiLSTM max, Gated ConvNet encoder

Conneau et al. ACL 2018

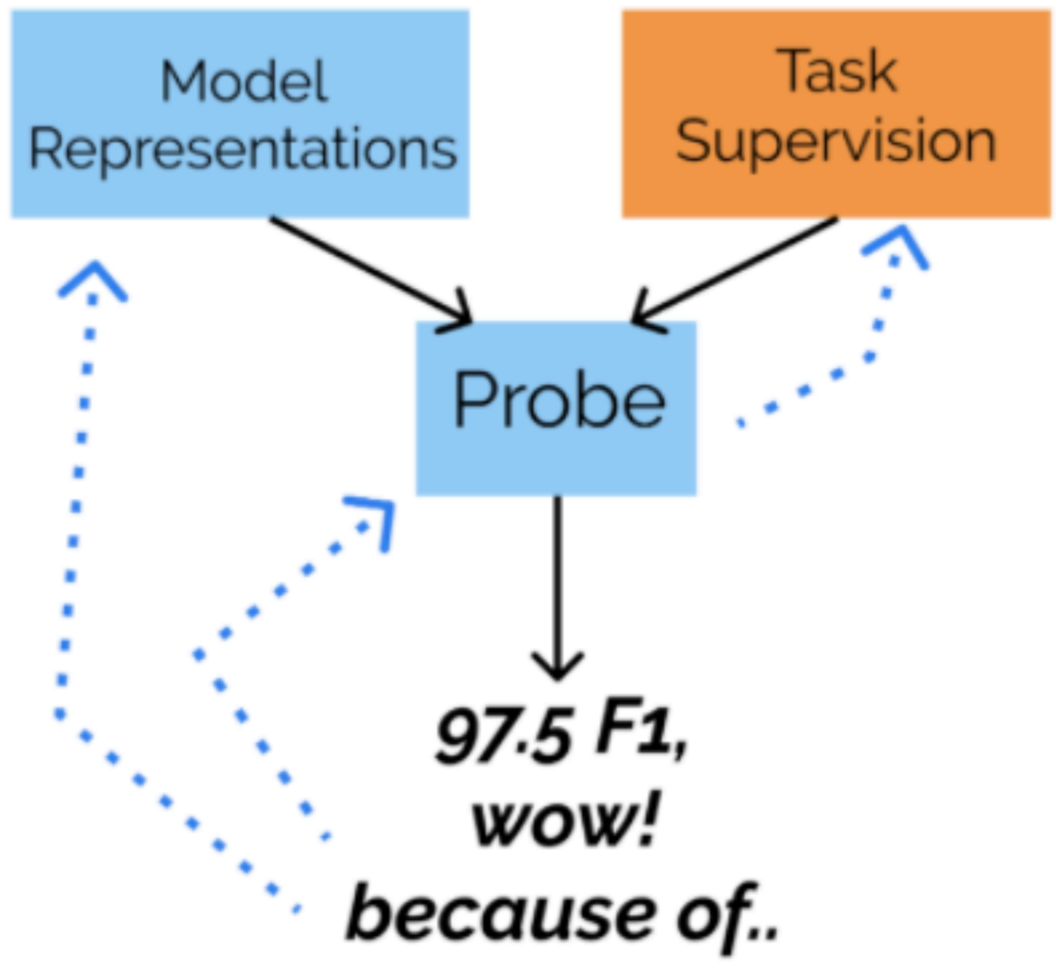
Issues with probing



Probing **turns supervised tasks into tools for interpreting representations. But the use of supervision leads to the question, did I interpret the representation? Or did my probe just learn the task itself?**

Hewitt et al. 2019

Issues with probing



Probing **turns supervised tasks into tools for interpreting representations. But the use of supervision leads to the question, did I interpret the representation? Or did my probe just learn the task itself?**

Hewitt et al. 2019

Minimum Description Length (MDL) Probes

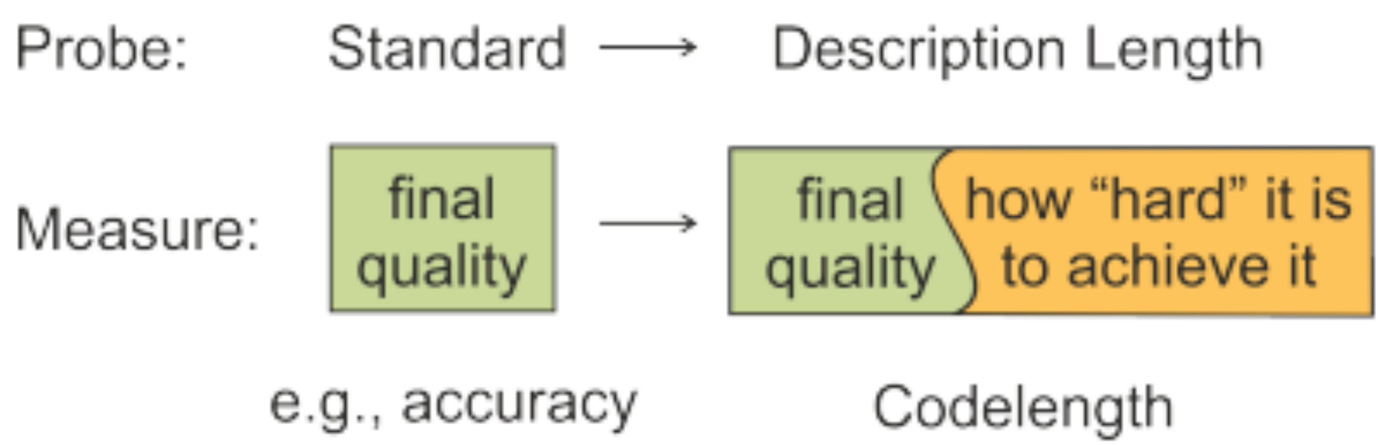


Figure 1: Illustration of the idea behind MDL probes.

- Characterizes both **probe quality** and **the amount of effort** needed to achieve it

- More informative and stable

Voita et al. 2020

Summary: What is the model learning?

https://boknilev.github.io/nlp-analysis-methods/table1.html

Explain the prediction

How to evaluate?

Training Phase | **Test Phase**

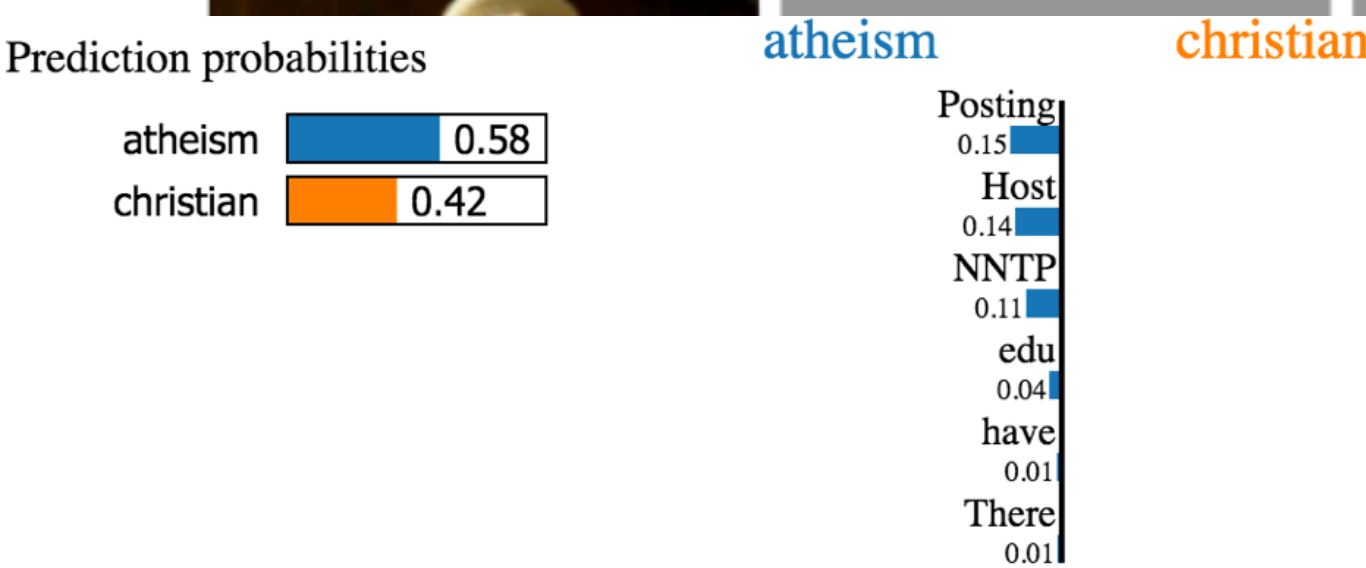
Some **x, f(x)** pairs

Input **x**
Predict **f(x)**

Some **x, f(x), E** triples

Input **x**
Predict **f(x)**

Explanation Technique: LIME



Text with highlighted words

From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

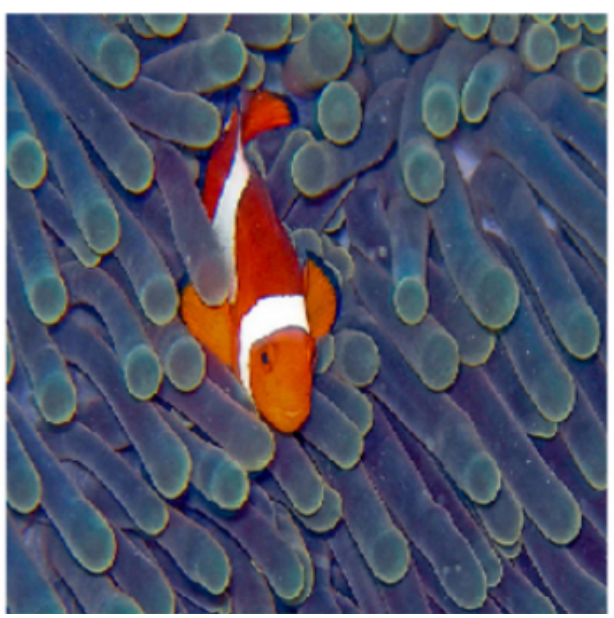
There have been some notes recently asking where to obtain the DARWIN fish.
This is the same question I have and I have not seen an answer on the
net. If anyone has a contact please post on the net or email me.

lighting positive pixels. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

Ribeiro et al, KDD 2016

Explanation Technique: Influence Functions

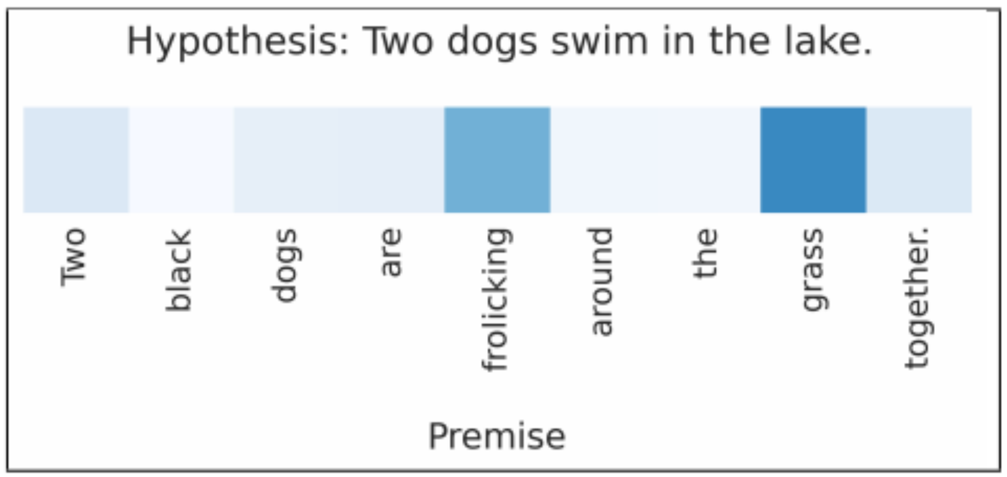
- What would happen if a given training point didn't exist?
- Retraining the network is prohibitively slow, hence approximate the effect using influence functions.



Most influential train images

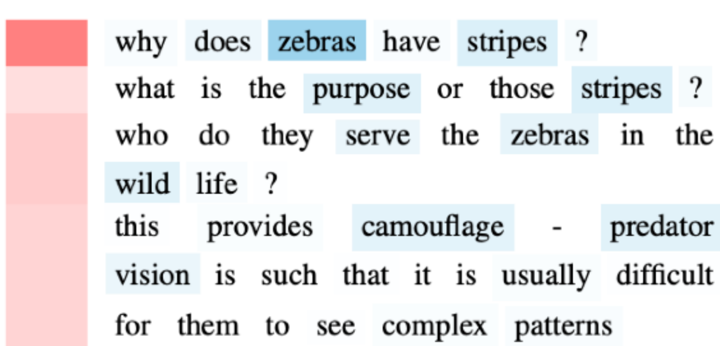
Koh & Liang, ICML 2017

Explanation Technique: Attention



Entailment

Rocktäschel et al, 2015

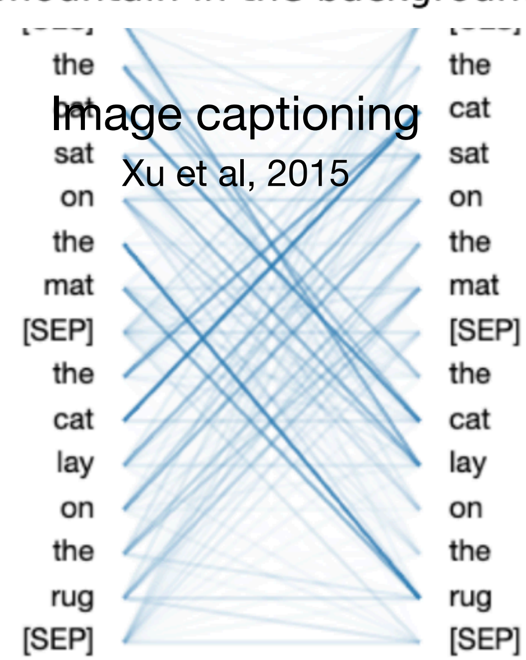


Document classification

Yang et al, 2016

Image captioning

Xu et al, 2015



BERTViz

Vig et al, 2019

Explanation Technique: Attention

Attention is not Explanation

Sarthak Jain
Northeastern University
jain.sar@husky.neu.edu

Byron C. Wallace
Northeastern University
b.wallace@northeastern.edu

1. Attention is only mildly correlated with other importance score techniques

2. Counterfactual attention weights should yield different predictions, but they do not

Attention is not not Explanation

Sarah Wiegreffe*
School of Interactive Computing
Georgia Institute of Technology
saw@gatech.edu

Yuval Pinter*
School of Interactive Computing
Georgia Institute of Technology
uvp@gatech.edu

"Attention *might* be an explanation."

- Attention scores can provide a (plausible) explanation not the explanation.
- Attention is not explanation if you don't need it
- Agree that attention is indeed manipulable,

"this should provide pause to researchers who are looking to attention distributions for one true, faithful interpretation of the link their model has established between inputs and outputs."

Learning to Deceive with Attention-Based Explanations

Danish Pruthi[†], Mansi Gupta[‡], Bhuwan Dhingra[†], Graham Neubig[†], Zachary C. Lipton[†]
[†]Carnegie Mellon University, Pittsburgh, USA
[‡]Twitter, New York, USA
ddanish@cs.cmu.edu, mansig@twitter.com,
{bdhingra, gneubig, zlipton}@cs.cmu.edu

Attention	Biography	Label
Original	Ms. X practices medicine in Memphis, TN and is affiliated ... Ms. X speaks English and Spanish.	Physician
Ours	Ms. X practices medicine in Memphis , TN and is affiliated ... Ms. X speaks English and Spanish.	Physician

- Manipulated models perform better than no-attention models
- Elucidate some workarounds (what happens behind the scenes)

Explanation Techniques: gradient based importance scores

Method	Attribution $R_i^c(x)$
Gradient * Input	$x_i \cdot \dfrac{\partial S_c(x)}{\partial x_i}$
Integrated Gradient	$(x_i - \bar{x_i}) \cdot \displaystyle\int_{\alpha=0}^{1} \dfrac{\partial S_c(\tilde{x})}{\partial (\tilde{x_i})}\Bigg\vert_{\tilde{x}=\bar{x}+\alpha(x-\bar{x})} d\alpha$
ϵ-LRP	$x_i \cdot \dfrac{\partial^g S_c(x)}{\partial x_i}, \quad g = \dfrac{f(z)}{z}$
DeepLIFT	$(x_i - \bar{x_i}) \cdot \dfrac{\partial^g S_c(x)}{\partial x_i}, \quad g = \dfrac{f(z) - f(\bar{z})}{z - \bar{z}}$

Example of attributions on MNIST

ReLU Tanh Sigmoid Softplus

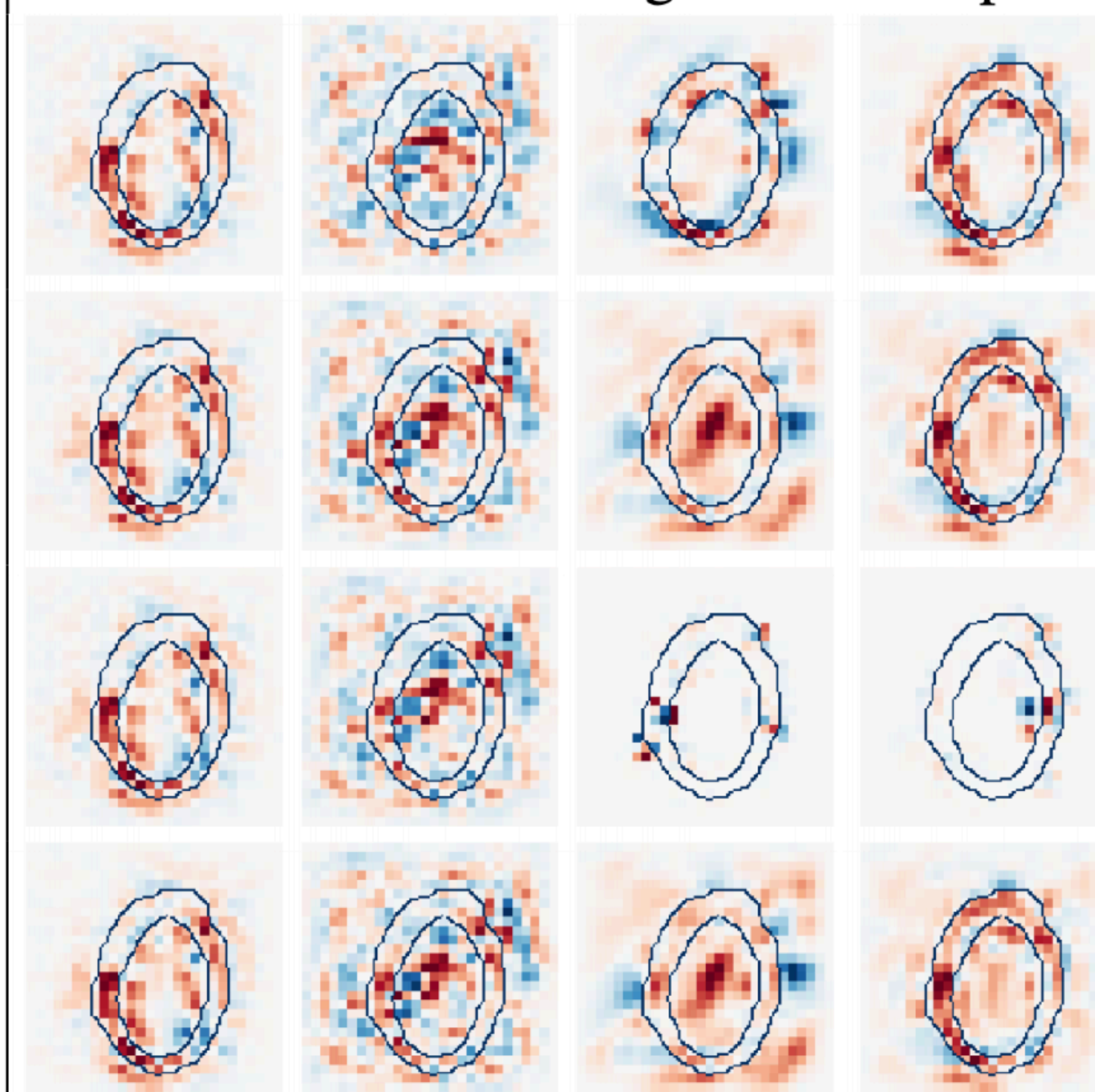


Figure from Ancona et al, ICLR 2018

Explanation Technique: Extractive Rationale Generation

Key idea: find minimal span(s) of text that can (by themselves) explain the prediction

- Generator (x) outputs a probability distribution of each word being the rational
- Encoder (x) predicts the output using the snippet of text x

- Regularization to support contiguous and minimal spans

Review

the beer was n't what i expected, and i'm not sure it's "true to style", but i thought it was delicious. **a very pleasant ruby red-amber color** with a relatively brilliant finish, but a limited amount of carbonation, from the look of it. aroma is what i think an amber ale should be - a nice blend of caramel and happiness bound together.

Ratings *Look:* 5 stars *Smell:* 4 stars

Figure 1: An example of a review with ranking in two categories. The rationale for Look prediction is shown in bold.

Future Directions

- Need automatic methods to evaluate interpretations

- Complete the feedback loop: update the model based on explanations

Thank You!

Questions?