

# Black-Box Interpretation Methods

## Anchors

Pavel Shvets

Lomonosov Moscow State University  
CMC MMF

November 26, 2020

# Summary

1 Recap

2 Interpretability as decision rules

# Recap

## Recap

## Notable Failures of Neural NLP Models



@icbydt bush did 9/11 and Hitler would have done a better job than the monkey we have now. donald trump is the only hope we've got.

Generate Hate Speech

Examples of these **molecules** **species** with C2 symmetry can increase enantioselectivity, as in their Josiphos variety...  
 Prediction: Ligand (✓) → Ion (✗)

Fragile to Small Edits

The doctor asked the nurse to help her  
 El doctor le pidió a la enfermera que le ayudara

Reflect Gender Biases

Article: Super Bowl 50

Paragraph: "Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV."

Question: "What is the name of the quarterback who was 38 in Super Bowl XXXIII?"

Original Prediction: John Elway

Prediction under adversary: Jeff Dean

Rely on pattern matching

SQuAD

Context

In 1899, John Jacob Astor IV invested \$100,000 for Tesla to further develop and produce a new lighting system. Instead, Tesla used the money to fund his Colorado Springs experiments.

Original

What did Tesla spend Astor's money on ?

Reduced

did

Confidence

0.78 → 0.91

Behave Counterintuitively

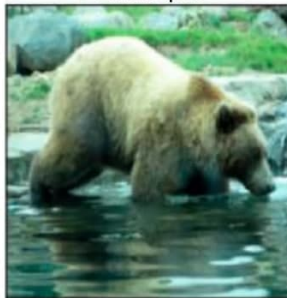
# Recap

Test Example



**Polar Bear ✗**

Important Training Example



**Polar Bear ✗**

# Recap

- finding errors in model
- finding errors in data
- explain decision to the end user
- comply with legal requirements

# Methods differentiation

- Baking interpretability into the model
- Looking at input features
- Looking for global decision rules
- Looking at training examples
- ...

# Interpretability as decision rules



# Anchors

It's advertised as a good movie but it really falls flat.

Anchor: if good and movie predict positive .

# Anchors

- I want to play a ball  
Anchor: if previous word is **particle** predict **verb**.
- I went to a play yesterday  
Anchor: if previous word is **determiner** predict **noun**.
- I play ball on Mondays  
Anchor: if previous word is **pronoun** predict **verb**.

# Anchors: algorithm

- 1 Input  $x$ : This movie is not bad. pos
- 2 Generate  $\mathcal{D}_x$ :
  - This director is always bad. neg
  - This movie is not nice neg
  - This stuff is rather honest pos
  - ...
- 3 Find Anchor  $A$  (for example,  $= \{ \text{not}, \text{bad} \}$ ):  
 that have high coverage on  $\mathcal{D}_x(\cdot|A)$ , and have high precision  $\mathcal{D}_x(\cdot|A)$ :
  - This audio is not bad pos
  - This novel is not bad pos
  - This footage is not bad pos

# Anchors: algorithm

$$\text{prec}(A) = \mathbb{E}_{\mathcal{D}(z|A)}[\mathbb{1}_{f(x)=f(z)}]$$

$$\max_{A \text{ s.t. } P(\text{prec}(A) \geq \tau) \geq 1 - \delta} \text{cov}(A)$$

$\tau$  – level of precision

$\delta$  – confidence

$A$  – anchor

$f$  – black-box model

$\mathcal{D}(\cdot|A)$  – conditional distribution on points similar to  $x$

# Anchors: algorithm. More precisely about step 3

- 1 Input  $x$ : This movie is not bad. pos
- 2 Generate  $\mathcal{D}_x$
- 3 There are a very big number of anchors ( $2^{\text{INPUT}}$ ): {This}, {Movie}, {is}, {bad}, {This, movie}, ...

# Anchors: algorithm. More precisely about step 3

- 1 Input  $x$ : This movie is not bad. pos
- 2 Generate  $\mathcal{D}_x$
- 3 Start generate anchors from bottom-up
- 4 Generate samples from  $\mathcal{D}_x(\cdot|A)$ , and then estimate the precision
- 5 Choose highest precision rule. Loop.
- 6 Exit loop if precision  $>$  threshold.

# References



Molnar, C. *Interpretable Machine Learning Book*. 1.2 (2020).



Ribeiro, M. T., Singh, S. & Guestrin, C. *Anchors: High-Precision Model-Agnostic Explanations*. in *AAAI* (2018).



Wallace, E., Gardner, M. & Singh, S. *Interpreting Predictions of NLP Models*. in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts* (Association for Computational Linguistics, Online, Nov. 2020), 20–23.  
<https://www.aclweb.org/anthology/2020.emnlp-tutorials.3>.

# The End