Black-Box Interpretation Methods

Anchors

Pavel Shvets

Lomonosov Moscow State University
CMC MMF

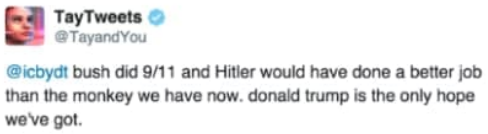
November 26, 2020

Summary

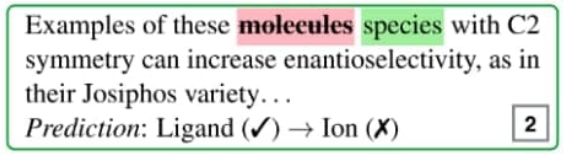
Recap

Recap

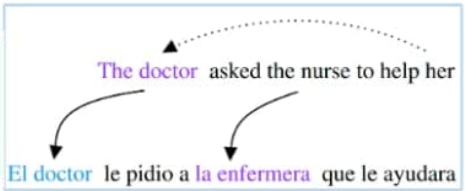
Notable Failures of Neural NLP Models



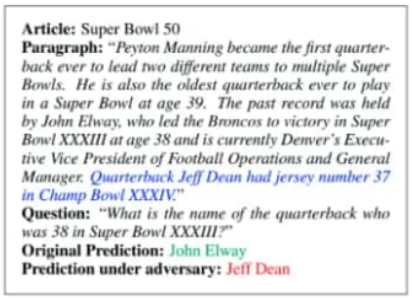
Generate Hate Speech



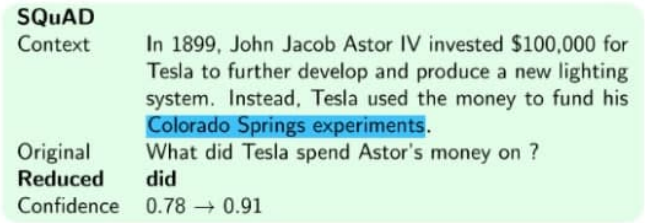
Fragile to Small Edits



Reflect Gender Biases



Rely on pattern matching



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, $= \{$ not, 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

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

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

τ – level of precision

δ – 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^{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