Explainable AI: Breaking Down the Black Box

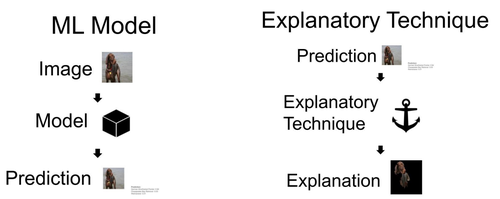
Chris Melville (me), Thomas Pree, Josh Moore, Lev Shuster, Adrian Boskovic, Sam Johnson-Lacoss

Advised by Anna Rafferty

The decisions of machine learning models can be difficult to interpret and explain, which is an increasingly important problem as they are handed more decision making power. We looked at three explanatory AI techniques (Anchoring, LIME, and Shapley) and applied them to three domains (tumor recognition, animal recognition, and online course dropout prediction). This poster will focus on Anchoring and the image recognition domains.

EXPLAINABLE AI

- Explainable AI is a class of algorithm that attempts to explain the decisions made by models.
- All three of our techniques are model agnostic, meaning they do not need to know the inner workings of the model they are explaining.



WHAT IS AN ANCHOR?

- An anchor is defined by its creators as “a rule that sufficiently ‘anchors’ the prediction locally – such that changes to the rest of the feature values do not matter.”
- This anchor can take the form of a few feature values for tabular data, or a part of an image in image classification.

nplay_video	nchapters	age	votes	num_words
1	9	20	0	49

Since the following are true, this student will fail:

- num_words <= 49.00
- nplay_video <= 20.00

HOW DOES ONE FIND AN ANCHOR?

- We want our anchors to be precise (meaning the anchor is important to the model’s prediction) and concise.
- Precision is measured by moving the prospective anchor onto different images and seeing if the model’s prediction changes.
- In the examples below, the prospective anchor would be strong if the model predicted the Frankenstein images as beagles.

- To generate an anchor, we start off with a null anchor, then iteratively add features to it until it hits a preset precision threshold.

USER STUDY

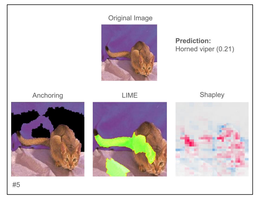
- We conducted a user study to see which techniques were favored. The participants were twelve college students, seven of whom were CS majors.
- Participants were shown an image of a cat or dog and the model’s prediction for the image, then shown three different explanations for the model’s prediction.

Example of prediction-explanation pair

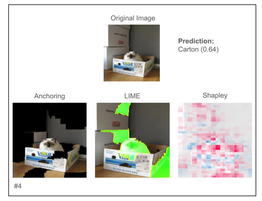
- After showing the participant the prediction and explanation, we asked whether each explanation helped or hurt their understanding of the model.
- After showing them eight images, we asked them which technique they liked the best overall.

USER STUDY RESULTS AND REACTIONS

- Overall, Shapley was the preferred model of nine of our twelve participants.
- Opinions still varied between individual images.



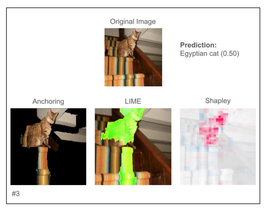
LIME was seen as the most helpful in explaining why this image was classified as a horned viper.



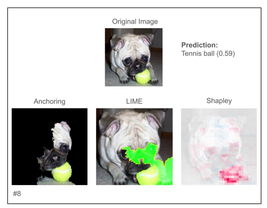
Anchoring was seen as most helpful in explaining why this image was classified a carton.

Participant quotes about Shapley:

- “Shapley was the most consistent”
- “I like Shapley because it reflects degrees of importance”



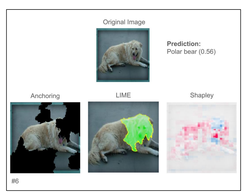
“Highlighting the stairs is a little weird in LIME and Anchoring”



“Shapley makes completely clear the model is looking at the ball”

Regarding techniques disagreeing:

“It is actively confusing to have different explanations for the same model. If they are seeing the same predictions, they should be showing the same explanation.

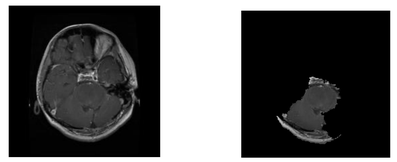


“If I was given any single one I would say they help, but they confuse in total”

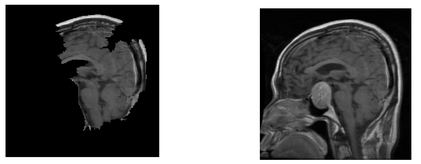
TUMORS

We also trained a brain MRI model to predict if patients had tumors or were healthy, and tested our explanatory techniques on this.

Anchoring correctly identifies the meningioma



Anchoring finds the glioma (small white ellipse on left edge of anchor), but also includes healthy parts of brain



Anchoring highlights healthy part of brain when model gives prediction of healthy

EXPLAINABLE AI TAKEAWAYS

- Explainable AI is an important field for helping humans understand ML models and decide whether or not to trust them.
- There is a tradeoff between speed, conciseness and precision, and none of our techniques consistently delivered all three.
- Anchoring, LIME, and Shapley are all useful techniques, but none of them are perfect.

REFERENCES

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High precision model-agnostic explanations. University of Washington.

ACKNOWLEDGEMENTS

We are grateful to all of our user study participants. We would also like to give a big thank you to all of our friends, families, and profs, especially our heroic advisor Dr. Anna Rafferty.