Week 6 Notes

By Eric Sullivan and Nina Scolieri

What is Interpretability?

- Interpretability is the degree to which a human can understand the cause of a decision
- Why is interpretability needed?
  - When problems or objectives are not completely specified
  - It allows humans to figure out if the ML model is meeting a specific set of criteria with its predictions
  - When there is an incompleteness in problem formalization

Why is Interpretability Desirable?

- Trust
- Causality
- Transferability
- Informativeness
- Fair and Ethical Decision Making

Trust

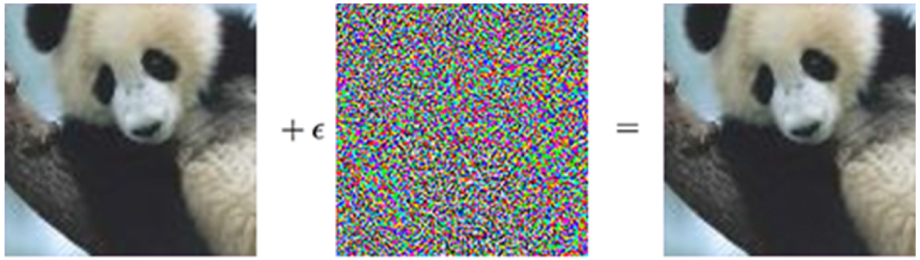
- People desire more than just good performance on standard metrics
- Being able to understand a model makes users feel more comfortable using certain systems
- People want a system that won’t discriminate based on race, gender sexual orientation, etc

Causality

- Most ML models are based off of correlations not causations
- People can draw from these correlations and try to interpret a causal relationship
- Correlations wouldn’t be understandable if you didn't have a interpretable model

Transferability

- ML models need to be able to do well on tasks it has already learned from and also transfer its knowledge to slightly different tasks
- A model trained to recognize pictures of a panda, noise is added to the model and it predicts that the next image it's fed is a "gibbon" not a panda



- Without interpretability, we wouldn't be able to understand how this prediction was made

Informativeness

- When you don’t intend to utilize the model for decision making
- Accuracy is not a factor because the model isn’t being used for decision making
- Without interpretability, you can’t have informativeness

Fair and Ethical Decision Making

- People want their models to be non-discriminatory
- An example would be the recidivism prediction: A model could be used to determine whether someone with a criminal record should be granted bail or sent back to prison
- Model should not make decisions solely based on race, sex, age, etc
- If this was a problem, we wouldn’t be able to understand why without interpretability

Arguments for Interpretability

- Important for instances where you might have causal associations
- Good for providing clarity in situations where completeness is lacking
- Pneumonia example
  - Predicted that people with asthma have a lower chance of getting pneumonia which isn't true
  - With interpretability, we'd be able to see people with asthma take more preventative measures so they don't contract pneumonia which is why their risk is lower
  - Positive correlation between “time to care” and “risk of dying from pneumonia”
- Causal Relationships
  - Deep learning models only look at correlations, not causations
  - Collecting the right data would not give the insight needed to uncover causal associations

Arguments Against Interpretability

- Don’t need to know the inner workings of everything
  - Example: Netflix movie recommendations
- The recommendations these algorithms produce have very low impact on our lives
- We may not fully understand why or how a system like Netflix recommends certain options for its users, but we have confidence that the system has been tested numerous times and makes accurate predictions most of the time