The Interpretation of Neural Networks

Special Topics, Spring 2025

Sike Ogieva ‘25
Advised by Dr. Haoze Wu

Overview and Assessment

This class covers methods for probing neural networks and understanding how they make predictions. Most of the class is modeled after the 2020 MIT IBM practicum: Structure and Interpretation of Deep Networks and its accompanying Github. In addition, we will examine the seminal papers that introduced Kolmogorov-Arnold Networks and Neurosymbolic AI.

Almost every week, two papers will be assigned for close reading. Some weeks have slides to study and assigned lab assignments in addition to these readings or instead of them. Each week, the student is required to turn in any assigned labs, the annotated copies of the assigned readings, and notes taken during background research on the readings. They are also to complete reviews that test their understanding of the papers.

These reviews involve answering four questions:

a. What is this paper about?
b. What are the strengths?
c. What are the weaknesses?
d. What are some significant follow up work from this paper? How do they differ from this paper?

Towards the end of the semester, the student will complete a reflection on the work covered so far. What they have learned, and what they are looking forward to learning going forward.

To write this syllabus, the student read Zhang, A Survey on Neural Network Interpretability (2021). All the material for this class is uploaded to this Github repository.

Textbook
https://christophm.github.io/interpretable-ml-book/shapley.html

Syllabus

Pre-Semester - Introduction to Interpretability
Practicum Slides: Why Care About Interpretability
Lipton, Mythos of Model Interpretability (2016)
Doshi-Velez, Kim, Towards A Rigorous Science of Interpretable Machine Learning (2017)
Ross, Right for the Right Reasons; (2017)
Belinkov, Analysis methods in NLP (2018)
01 Practicum Lab
Brian Christian, The Alignment Problem

Week One - Explaining Predictions - Vision
27-01 to 31-01
Practicum Slides: Saliency
Zhou: LIME: Learning Deep Features for Discriminative Localization (2015)
Smilkov, SmoothGrad: Removing Noise by Adding Noise; (2017)

Week Two - Explaining Predictions - Language
03-02 to 07-02
Li, Visualizing and Understanding Neural Models in NLP; (2016)
Ding, Saliency-driven Word Alignment Interpretation for Neural Machine Translation; (2019)
Mudrakarta, Did the Model Understand the Question?; (2018)

Week Three - Explaining Models
10-02 to 14-02
Practicum Slides - Models
Belinkov and Glass, Analysis Methods (section 2); (2019)

Week Four - Explaining Models
17-02 to 21-02
Bau, Network Dissection: Quantifying Interpretability of Deep Visual Representations (2017)
Bau, GAN Dissection: Visualizing And Understanding Generative Adversarial Networks (2018)

Week Five - Explaining Models
24-02 to 28-02
03 Lab Gan Dissection Exercise
03 Lab Probing Exercise

Week Six - Adversaries
03-03 to 07-03
Practicum Slides - Adversaries
Goodfellow, Explaining and Harnessing Adversarial Examples; (2015)
Smilkov, Simple gradient explanation with SmoothGrad (Revision);
04 Lab Adversaries

Week Seven - Bias
10-03 to 14-03
Practicum Slides - Bias and Fairness
Propublica, Machine Bias Risk in Criminal Sentencing
Buolamwini, Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification; (2018)
Chouldechova, Fair prediction with disparate impact: A study of bias in recidivism prediction instruments; (2016)

Week Eight - Spring Break
17-03 to 21-03

Week Nine - Bias
24-03 to 28-03
Ziad, Dissecting racial bias in an algorithm used to manage the health of populations (2019)
05 Lab Bias

Week Ten - Interaction
31-03 to 04-04
Practicum Slides - Interaction
Strobelt, Seq2Seq-Vis: A Visual Debugging Tool for Sequence-to-Sequence Models (2018)
Gehrmann, Visual Interaction with Deep Learning Models through Collaborative Semantic Inference; (2019)

Week Eleven - Interaction (and Complex Explanations)
07-04 to 11-04
06 GanPaintLab
Hendricks, Generating Visual Explanations; (2016)

Week Twelve - Complex Explanations
14-04 to 18-04
Hendricks, Grounding Visual Explanations; (2018)
Andreas, Neuralese: Analogs of Linguistic Structure in Deep Representations; (2017)

Week Thirteen
21-04 to 22-04 - April Break
23-04 to 25-04 - Reflection

Week Fourteen - More Topics
28-04 to 02-05
Sheth, Neurosymbolic AI - Why, What, and How (2023)
Liu, Kolmogorov-Arnold Networks (massive paper) (2024)

Week Fifteen
05-05 to 06-05
Reflection

Assignments

1. 50% → Reviews
2. 20% → Look-up Notes
3. 10% → Annotated slides, articles and papers
4. 10% → Labs
5. 10% → Reflection