

Visual Explainability in Machine Learning

Lecture 10: Conclusion



Ghassan AlRegib, PhD
Professor



Mohit Prabhushankar, PhD
Postdoctoral Fellow

Omni Lab for Intelligent Visual Engineering and Science (OLIVES)
School of Electrical and Computer Engineering
Georgia Institute of Technology
{alregib, mohit.p}@gatech.edu

Dec 7, 2023

Short Course Materials

Accessible Online



SCAN ME



Title: Visual Explainability in Machine Learning

Presented by: *Ghassan AlRegib, and Mohit Prabhushankar*

Omni Lab for Intelligent Visual Engineering and Science (OLIVES)

School of Electrical and Computer Engineering

Georgia Institute of Technology, Atlanta, USA

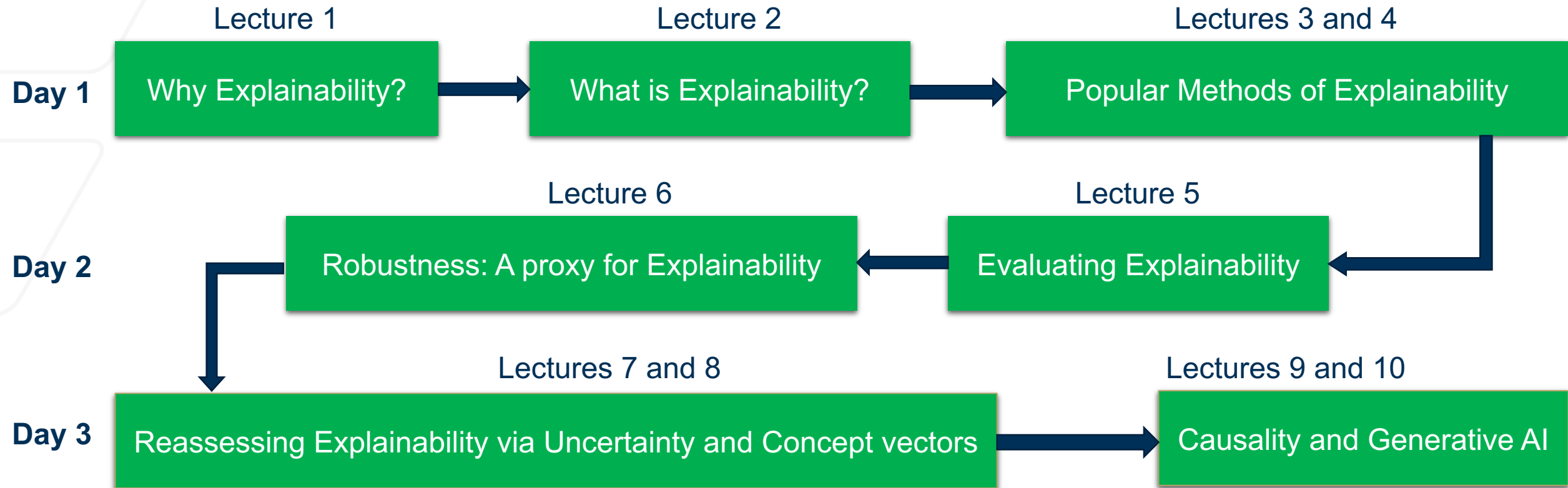
<https://alregib.ece.gatech.edu/>

<https://alregib.ece.gatech.edu/sps-education-short-course/>
{alregib, mohit.p}@gatech.edu

Short Course

Course Outline

Day 1: Define and Detail; Day 2: Evaluate; Day 3: Reassess



Generative AI and Explainability

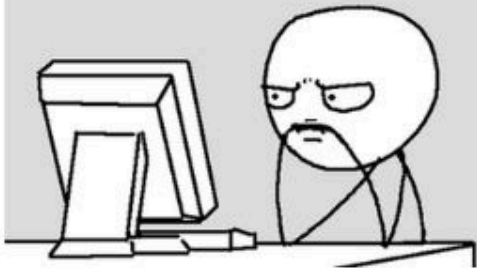
ChatGPT

Before Chat GPT

After Chat GPT

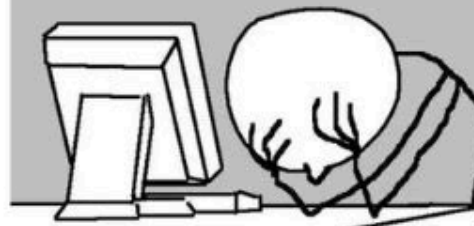
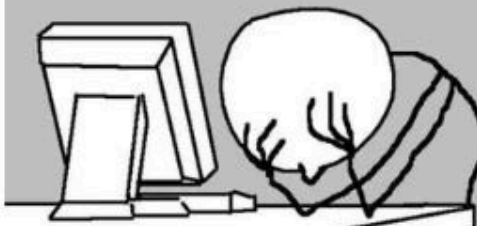
* Developer coding - 2 hours

* ChatGPT generating code - 5 min



* Developer debugging - 6 hours

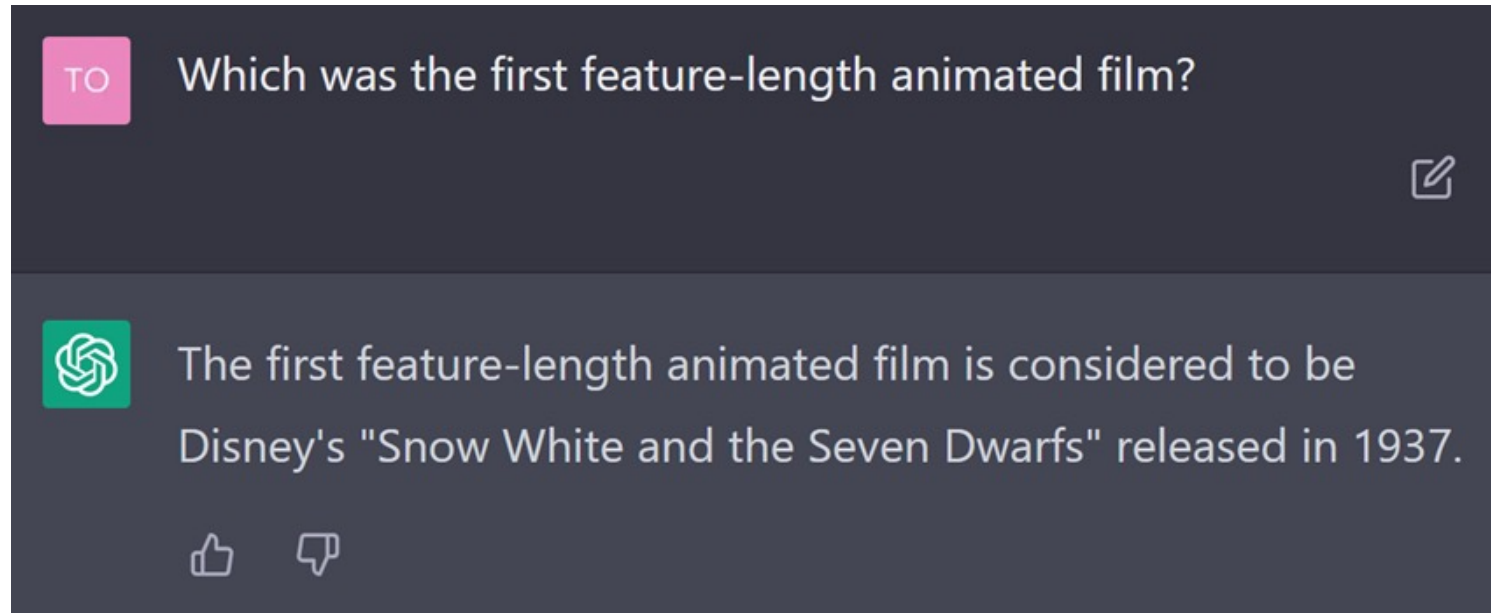
* Developer debugging - 24 hours



Generative AI and Explainability

ChatGPT: The Snow White Challenge

Deep Learning cannot reason effectively



Release year is correct but the **answer is incorrect**

True answer: Argentine film called El Apóstol—released in 1917,


Generative AI and Explainability

ChatGPT: The Snow White Challenge

Deep Learning cannot reason effectively

TO What do you know about the early El Apostol film?



 "El Apostol" was a silent animated film created in Argentina in 1917 by director Quirino Cristiani. It is considered to be the first animated feature film in Latin America and one of the earliest animated films in the world.

ChatGPT knows about El Apostol as well as its release date

Even with the correct facts, it cannot reason correctly. This is because, Snow White being the first animated movie is a common misconception and as such, has been used to train and teach ChatGPT.

And it feeds this back to us without considering the other facts it has learned.

Generative AI and Explainability

Explainability in NLP applications

Natural Language Interactivity allows direct and targeted explanations

Definition

| Methods | Indirect | Direct | Targeted |
|-------------------------------|----------|--------|----------|
| Deconvolution [21] | ✓ | — | — |
| Inverted Representations [22] | ✓ | — | — |
| Guided-Backpropagation [18] | — | ✓ | — |
| SmoothGrad [17] | — | ✓ | — |
| LIME [39] | — | ✓ | — |
| CAM [24] | — | ✓ | — |
| Graph-CNN [23] | ✓ | — | — |
| GradCAM [12] | — | — | ✓ |
| TCAV [40] | — | ✓ | — |
| GradCAM++ [16] | — | — | ✓ |
| RISE [35] | — | ✓ | — |
| Causal-CAM [15] | — | — | ✓ |
| Counterfactual-CAM [12] | — | — | ✓ |
| Goyal et al. [26] | — | — | ✓ |
| CEM [29] | — | — | ✓ |
| Contrast-CAM [13] | — | — | ✓ |
| Contrastive reasoning [14] | — | — | ✓ |

Indirect → Direct → Targeted

- Substantial analysis within indirect explanations helped vision applications
- Is there a case to be made for more indirect explanations with generative AI?

Short Course

Course Outline

Day 1: Define and Detail; Day 2: Evaluate; Day 3: Reassess



Takeaways

Day 1

Day 1: Define and Detail; Day 2: Evaluate; Day 3: Reassess

- **There is no “One explanation fits all”**
- Explainability techniques can be categorized based on their **method choices, properties, reasoning paradigms**
- Categorization based on the **knowledge of the humans** asking for the explanations provides an insight into the **evolution of explanatory research**
 - Indirect explanations are catered towards engineers and researchers
 - Direct explanations are directed towards general public and can be used in an educational setting
 - Targeted explanations allow interactivity between an AI system and a human
- **Gradients quantify the change** in models and assign an **importance score** to pixels
 - They can be directly used to explain decisions in the pixel space
 - They can act as weights on the activations to produce class-discriminative explanations

Takeaways

Day 2

Day 1: Define and Detail; Day 2: Evaluate; Day 3: Reassess

- **There is no “One evaluation fits all”**
- There are three large classes of evaluation
 - **Human evaluation:** Requires **direct comparison between explanations** from a large number of random subjects
 - **Application evaluation:** Requires choosing the right application where a human has already annotated the data and this annotation is leveraged for evaluating between explanations
 - **Network evaluation:** Requires interventions within data and evaluating the effect of Explainability as an intervention in the network
- In some cases, **explanation evaluation** turns into a **experimental design challenge**
- **Robustness of neural networks is seen as a proxy for Explainability**
 - More robust a network is, better its attribution of features correlated with required classes
 - **Gradient features, used to obtain explanations, can also provide robust predictions and detections**

Takeaways

Day 3

Day 1: Define and Detail; Day 2: Evaluate; Day 3: Reassess

- **Network evaluation only (partially) reduces predictive uncertainty**
- Explanations have some uncertainty associated with them
 - **This uncertainty is unique to the network, the data, and the explanatory method**
 - **This uncertainty exists because of all the residual explanations that *could have been***
- **Concepts** are abstract ideas that can be represented mathematically as a **collection of activation vectors**
- Sign of gradients is used to characterize concepts in test data against an additional trained model that maps concepts to classes
 - **This requires a large amount of labeled concepts** which is hard to come by
- Properties of concepts can be used to construct weakly-supervised pixel-wise explanations
- **Causal literature has played a pivotal role in constructing, defining, and evaluating explanations**
- Visual factor causal assessment is challenging in deep learning networks
 - **Disjoint features are not available**
 - **The neural network is not a structured causal model**

Takeaways

Gradients

Gradients are versatile

- **Robustness** under distributional shift in domains, environments, and adversaries are **challenges** for neural networks
 - **Gradients at Inference** provide a **holistic solution** to the above challenges
- **Gradients** can help **traverse** through a trained and unknown **manifold**
 - They approximate **Fisher Information** on the projection
 - They can be **manipulated** by providing **contrast** classes
 - They can be used to construct **localized contrastive** manifolds
 - They provide **implicit knowledge** about **all classes**, when only **one data** point is available at inference
- Gradients are useful in a number of **Image Understanding** applications
 - Highlighting features of the current prediction as well as **counterfactual** data and **contrastive** classes
 - Providing **directional information** in anomaly detection and testing via concept vectors
 - **Quantifying uncertainty** for out-of-distribution, corruption, and adversarial detection
 - Providing **expectancy mismatch** for human vision related applications