Information Visualization

eXplainable Artificial Intelligence

Dr. David Koop





Using Visualization in Al/Deep Learning

Visual Analytics in Deep Learning Interrogative Survey Overview

§4 WHY

Why would one want to use visualization in deep learning?

Interpretability & Explainability Debugging & Improving Models Comparing & Selecting Models Teaching Deep Learning Concepts

?

§6 WHAT

What data, features, and relationships in deep learning can be visualized? Computational Graph & Network Architecture Learned Model Parameters Individual Computational Units Neurons In High-dimensional Space Aggregated Information



Who would use and benefit from visualizing deep learning?

Model Developers & Builders Model Users Non-experts



How can we visualize deep learning data, features, and relationships?

Node-link Diagrams for Network Architecture **Dimensionality Reduction & Scatter Plots** Line Charts for Temporal Metrics Instance-based Analysis & Exploration Interactive Experimentation Algorithms for Attribution & Feature Visualization

D. Koop, CSCI 628, Fall 2021

WHEN **§8**

When in the deep learning process is visualization used? During Training After Training







Where has deep learning visualization been used?

Application Domains & Models A Vibrant Research Community





2



D. Koop, CSCI 628, Fall 2021







3

XAI Tradeoff







XAI Possibilities

Directly explainable model



- Linear model
- Decision tree
- Rule-based model

D. Koop, CSCI 628, Fall 2021



• Ensemble models









XAI Post-Hoc Techniques

Directly explainable model

> Explaining the model (global)

XAI









Feature Importance in a Decision

Customer: Jason Assets score: 88 No. Of satisfactory trades: 0 Mo. since account open: 3 No. of inquiries: 1 Debt percentage: 10%



Assets score
No. Of satisfactory trades
Mo. since account open
No. Of inquiries
Debt percentage
Why is Jason predicted of low risk?
Can I trust this prediction?

D. Koop, CSCI 628, Fall 2021



Risk of failing to repay: low





Counterfactual to Explain a Decision

Customer: Ana

Assets score: No. Of satisfactory trades: Mo. since account open: No. of inquiries: Debt percentage: **50%**

Sue

Assets score: 66 No. Of satisfactory trades: 1 Mo. since account open: 12 No. of inquiries: 3 Debt percentage: 28% Repaid on time



Bank customer

D. Koop, CSCI 628, Fall 2021



Risk of failing to repay: high

Why was my loan application rejected? How can I improve in the future?





8

XAI Evaluation

Inherent "goodness" metrics

- Fidelity/faithfulness
- Stability
- Compactness

User-dependent measures

- Comprehensibility
- Explanation satisfaction

. . .

. . .

D. Koop, CSCI 628, Fall 2021

Faithfulness

Correlation between the feature importance assigned by the interpretability algorithm and the effect of features on model accuracy.

Task oriented measures

- Task performance
- Impact on AI interaction
 - Trust (calibration) in model
- Task or AI system satisfaction









XAI as the Foundation for Responsible AI







Question-Driven XAI UX

Step 1

Identify user questions

Step 2

Analyze questions

Elicit user needs for XAI as questions

Also gather user intentions and expectations for asking the questions Cluster questions into categories and prioritize categories for the XAI UX to focus on

Summarize user intentions and expectations to identify key user requirements

Designers, users Designers, product team

D. Koop, CSCI 628, Fall 2021

Step 3

Map questions to modeling solutions

Step 4

Iteratively design and evaluate

Map prioritized question categories to candidate XAI techniques as a set of functional elements that the design should cover

A mapping guide for supervised ML is provided for refer<u>ence</u> Create a design including the candidate elements identified in step 3

Iteratively valuate the design with the user requirements identified in step 2 and fill the gaps

Designers, data scientists

Designers, data scientists, users





Projects

- Time to make progress on projects
- End of semester quickly approaching
- Questions?





Today's Paper

The Building Blocks of Interpretability

Interpretability techniques are normally studied in isolation. We explore the powerful interfaces that arise when you combine them and the rich structure of this combinatorial space.

CHOOSE AN INPUT IMAGE



For instance, by combining feature visualization (*what is a neuron looking for?*) with attribution (*how does it affect the output?*), we can explore how the network decides between labels like **Labrador retriever** and **tiger cat**.

CHANNELS THAT MOST SUPPORT ...

LABRADOR RETRIEVER

feature visualization of channel hover for attribution maps \rightarrow			
net evidence	1.63	1.51	
for "Labrador retriever"	1.22	1.24	
for "tiger cat"	-0.40	-0.27	

D. Koop, CSCI 628, Fall 2021



<u>Several floppy ear</u> detectors seem to be important when distinguishing dogs, whereas <u>pointy ears</u> are used to classify "tiger cat".



1.32 0.13





