Towards Automatic Concept-based Explanations

Radwa El Shawi
Machine learning Interpretability

“Interpretability is the degree to which an observer can understand the cause of a decision.”

~ Miller T., 2017, Explanation in AI: Insights from the Social Sciences

- Feature attribution
- Rules
- The most influential instances in the training data
- Counterfactual explanation
- High-level human understandable concepts
Why ML interpretability is challenging
A path the may take a step toward good explanation

1. Where are you going?

2. Take a human centric approach
Investigating post-training interpretability techniques

A trained machine learning model (e.g., neural network) Given a trained ML model, find the evidence of prediction.

Predictions & decisions
Saliency Maps

Were there more pixels on the ‘ATM’ than on the ‘person’?

Which concept mattered more for the prediction, ‘glass door’, ‘paper’?

Which concept mattered more for the prediction?

Would not be more useful if we can quantify the importance of each of user-defined concepts?

Goal of TCAV: Testing with Concept Activation Vectors

Quantitative explanation: how much a concept (e.g., stripes) was important for the prediction of specific class (e.g., zebra) in a trained model?

Goal of TCAV: Testing with Concept Activation Vectors

TCAV provides quantitative importance of a concept if and only if your network learned about it.

Trained NN model

P( )

TCAV score for concept Zebra

How important is concept stripes to the prediction of this image as zebra?

How to define concepts

1. Concept data set (stripes)
2. Class under test - zebra
3. Random images
4. Activation at layer $i$
5. \[
        \frac{\partial p(z)}{\partial v_C^i} = S_{C,k,l}(x)
    \]

Train a linear classifier to separate activations. CAV ($v_c'$) is the vector orthogonal to the decision boundary.
TCAV: Testing with Concept Activation Vectors

How important is concept striped to the prediction of this image as zebra?

1. Learning CAVs

2. Getting TCAV score

$$TCAV_{C,k,l} = \frac{|\{x \in X_k : S_{C,k,l}(x) > 0\}|}{|X_k|}$$
Can we do better?

1. Where are you going?

2. Take a human centric approach

3. Can we do better?
TCAV: Testing with Concept Activation Vectors


Would not be more useful if we can quantify the importance of automatically extracted concepts?
Automated Concept-based Decision Tree Explanations for CNNs ACDTE

ACDTE Stage1: Concept extraction

(a) Extract a set of similar images to the image to be explained either from the main task dataset or related dataset. Each image in the selected images is segmented.

(b) Segments are clustered in the activation space and outliers are removed to form coherent clusters that represent concepts.
ACDTE Stage 2: Learning interpretable concept and extracting concept data

(c) Learning interpretable concept models on clusters segments
(d) Extracting concept data from images similar to the image being explained

(c) Training a linear model for each concept to act as a concept detector. (d) For each image in the activation space, use concepts detectors to form a binary feature vector.

ACDTE Stage 3: Building explanation decision tree

(e) Build decision tree using extracted concept data along their predication from the black-box model

(e) Feature vectors along with the prediction of the target network are used to train a shallow decision tree. The decision tree provides a natural explanation for the contributing concepts for the prediction, in addition to counterfactual explanation.
Human Evaluation of the Visual Explanations

Which group of images is more meaningful to you?

- group a
- group b
Human Evaluation of the Visual Explanations (Cont.)

Which of the images below highly contribute to the prediction of the image above as a street?

Which of the images below highly contribute to the prediction of the image above as a park?
Open Challenges

• No single explanation can fit all users
• Rigorous, agreed upon evaluation protocols
• Human involvement (e.g. better interactive, “social” explanations)
Thank You