ShapeShop: Towards Understanding Deep Learning Representations via Interactive Experimentation

Fred Hohman
fredhohman@gatech.edu

Nathan Hodas
nathan.hodas@pnnl.gov

Polo Chau
polo@gatech.edu

Summary
ShapeShop is an interactive system for visualizing and understanding what representations a neural network model has learned. Built using standard web technologies, ShapeShop allows users to experiment with and compare deep learning models to help explore and understand the relationship between data and its learned representations.

ShapeShop: Visualizing Deep Learning Representations in Simple Shapes via Interactive Experimentation

Classifier Model Builder
1. Select Training Data

2. Select Model
   - MLP: multilayer perceptron
   - CNN: convolutional network

3. Select Hyperparameters
   - Initial Image
   - Step-size: 0.005, 0.01
   - Epoch Count: 50, 100, 150, 200, 250

4. Train and Visualize
   - Train classifier and visualize its learned representations by performing class activation maximization on each class.

Experiment Results
[1]: Model: MLP, Step-size: 0.005, Epoch Count: 100
   - Yellow = ✓
   - Red = ✗

[2]: Model: CNN, Step-size: 0.005, Epoch Count: 100
   - Yellow = ✓
   - Red = ✗

Example
Addition of diverse data produces more human recognizable representations.

1. Original binary classifier
   - Correlation coefficient improves to 0.52 after adding both circular classes

2. Correlation coefficient improves to 0.59 after adding both noise classes

Ongoing Work
Incorporating human feedback to improve interpretability via semantic highlighting.

Yellow = ✓
Red = ✗

This work supported in part by grants NSF IIS-1563816 and NIH BD2K U54EB020404.

ShapeShop: Towards Understanding Deep Learning Representations via Interactive Experimentation.
Fred Hohman, Nathan Hodas, Duen Horng (Polo) Chau.