Neural Network Interpretability
A Study Based on the "Zoom In" Article

This document outlines a series of experiments designed to explore and validate the concepts presented in the "Zoom In: An Introduction to Circuits" article on Distill.pub.
https://distill.pub/2020/circuits/zoom-in/

The experiments focus on:

- **Feature visualization,**
- **Argument validation,**
- **Universality**

A detailed description of the experiments is shared below

Feature Visualization

Objective

This experiment will provide visual evidence of the different types of low-level features that CNNs can detect and learn.

To visualize various low-level features such as *curves, edges, colors, shapes, textures, gradients, and high-low frequencies* detected by convolutional neural networks (CNNs).

Methodology

1. Select a pre-trained CNN model and a set of input images.
2. For each layer in the model, extract the feature maps.
3. Visualize the feature maps to identify the presence of different low-level features.
4. Compare/Validate expected outcomes

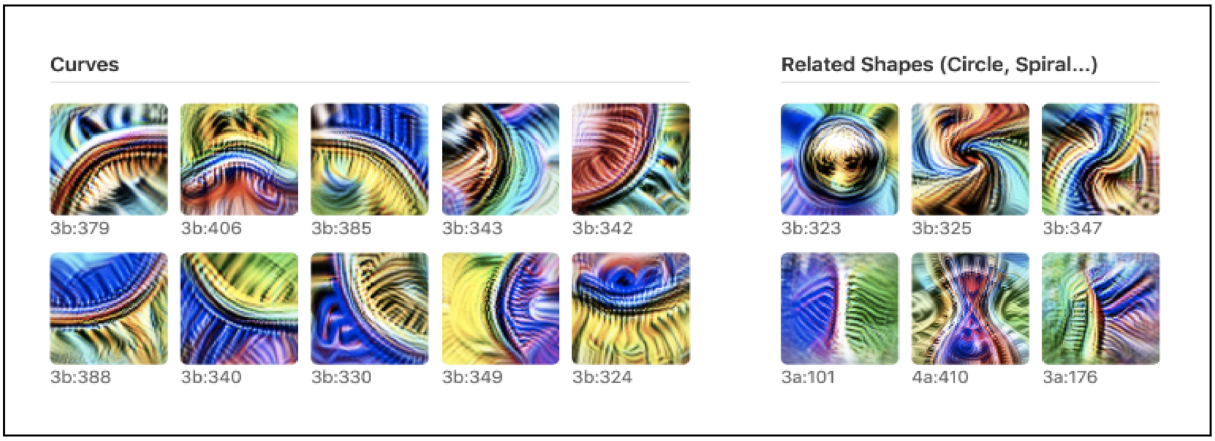


Figure 1: Feature visualization for curve detectors

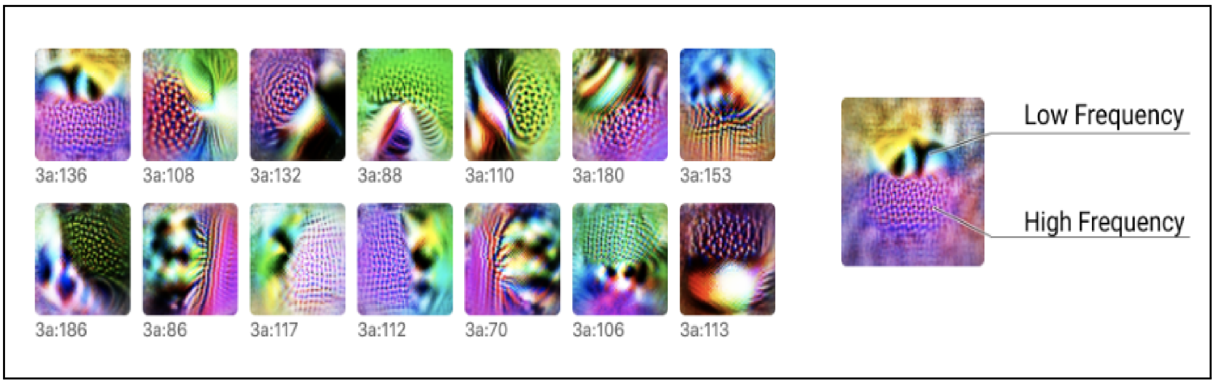


Figure 2: Feature visualization for High-low frequency detectors

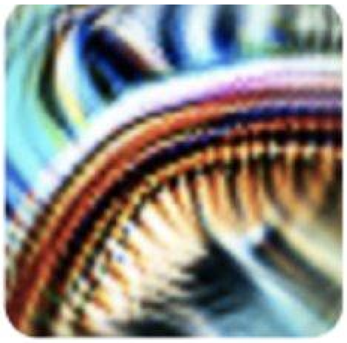
Arguments Validation

Objective

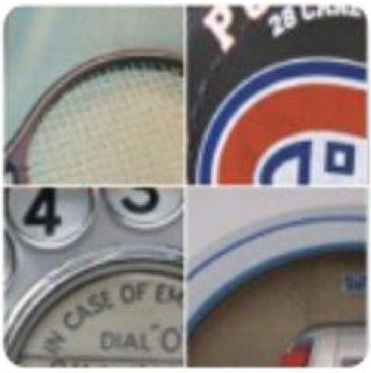
To validate the first four arguments presented in the "Zoom In" article.

Methodology

- Select a pre-trained CNN model and a set of input images.
- For each argument, design an experiment or analysis that can provide evidence for or against the argument.
- For Argument 1 (Feature Visualization):
  You can work on activation maximization of neurons to observe what features they react to.

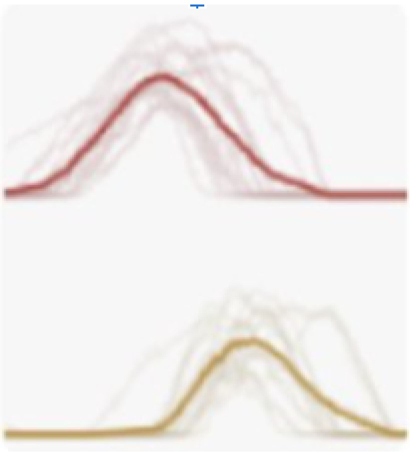


- For Argument 2 (Dataset examples):
  Check if the real images from the dataset that activate the selected neuron contain the expected feature (e.g., curves of a specific orientation).



- For Argument 3(Synthetic Examples):
  Verify if the selected neuron responds as expected to artificially created images with varying features (e.g., curves of different orientations and curvatures).

- For Argument 4 (Joint Tuning):
  Monitor the activation of the selected neuron as the input images are rotated, and observe if neurons detecting the same feature in different orientations activate in sequence.



- Conduct the experiments or analyses and record the results.

Universality

Objective

To test the universality of the findings from the "Zoom In" article across different models, datasets, and hyperparameters.

Methodology

- Select a combination from a range of pre-trained CNN models, datasets, and hyperparameters.
  You can even select a custom model and custom dataset for this experiment.
- For each combination of model, dataset, and hyperparameters, conduct the feature visualization and argument validation experiments.
- Compare the results across different combinations to assess the universality of the findings.

Keep in mind that you have total freedom to go beyond these experiments and get interpretability insights and some fun visualizations.

You are free to choose any library and framework you are comfortable with.