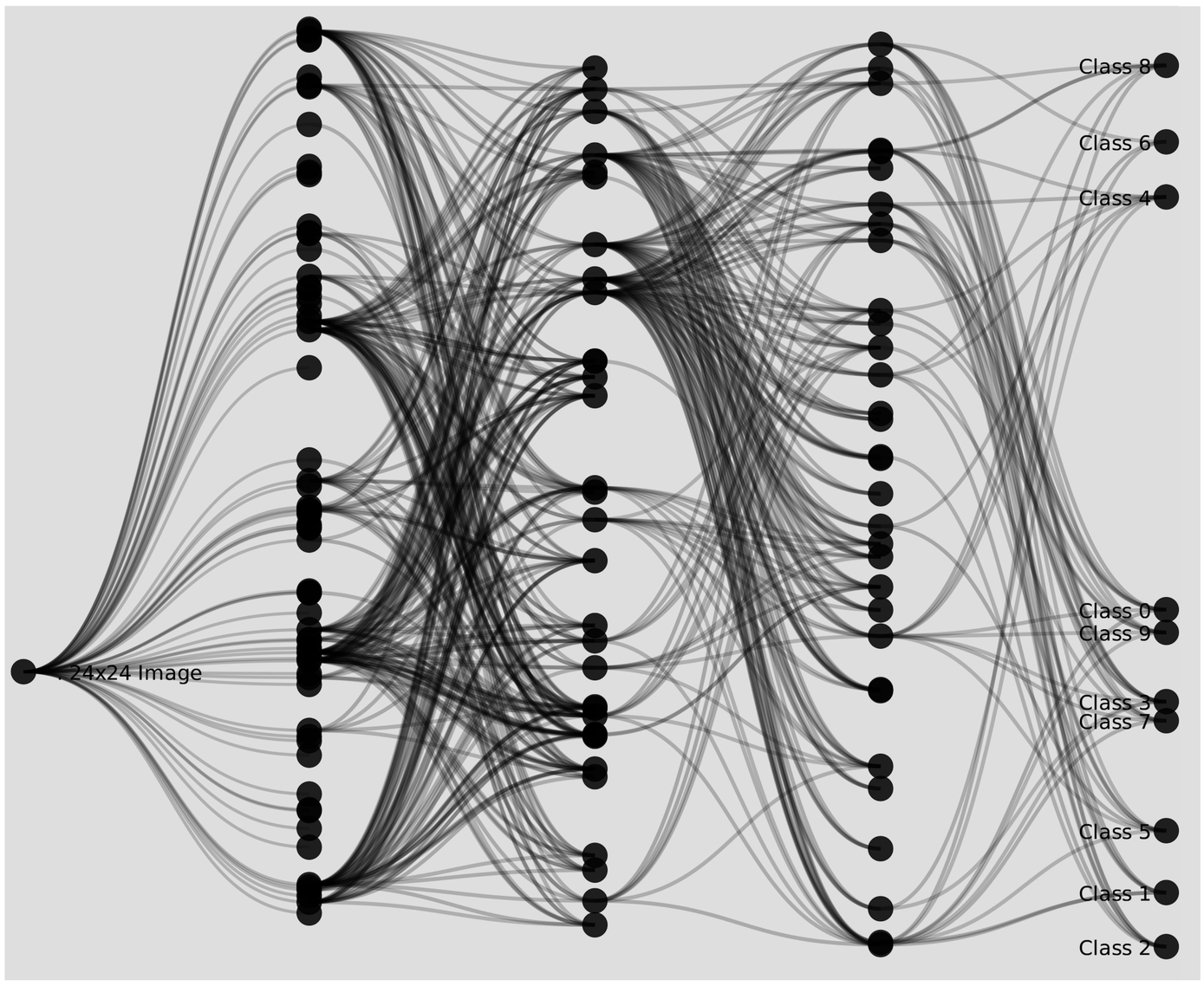
UW CSE 512::S2015

Neural Network Behavior

Kendall Lowrey & Tam Nguyen



Visualization of 3-layer (64x32x32) network for recognizing hand written numbers from the MNIST dataset.

A Black Box?

Artificial Neural Networks have been traditionally treated as black boxes. Trained on tremendous amounts of data to robustly perform tasks in computer vision, robotic control, function approximation, and others, their inner workings are often overlooked.

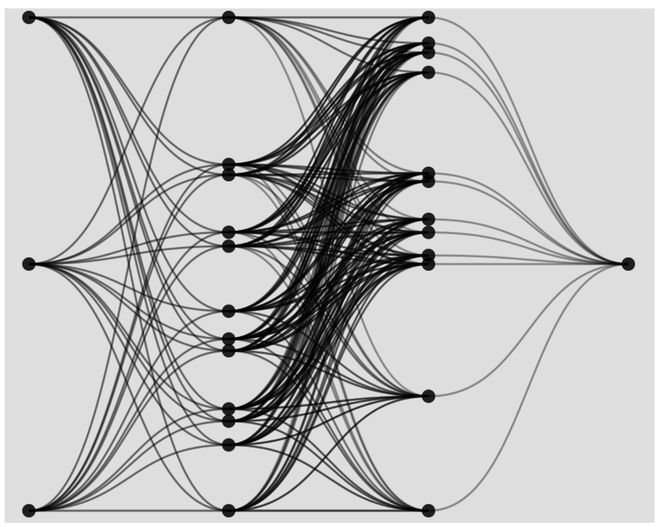
Some visualizations are application specific (such as those in computer vision), or only examined with respect to some input data. Capturing the network in just one instance in time fails to understand its holistic operation. We present a biologically inspired approach to visualizing the behavior of neural networks through discovery of their structure of activations.

Instead of glimpses, we attempt to open the box to see the inside.

How to Peer Inside

Simply put, neural networks utilize layers of individual neurons that respond to their inputs from previous layers. As each layer transforms its inputs, the neurons activate according to their trained weights. Our method builds a data set from a network during run time by recording each neuron's activations due to a test set of network inputs.

Using the tSNE algorithm we find which neurons co-activate together which provides a natural grouping by proximity for neurons of a layer. For neurons between layers, we show co-activation as stronger flow lines much like a Sankey diagram. The similarity to traditional neural network graphs was intentional: the network inputs start from the left and flow to the outputs on the right through the layers.

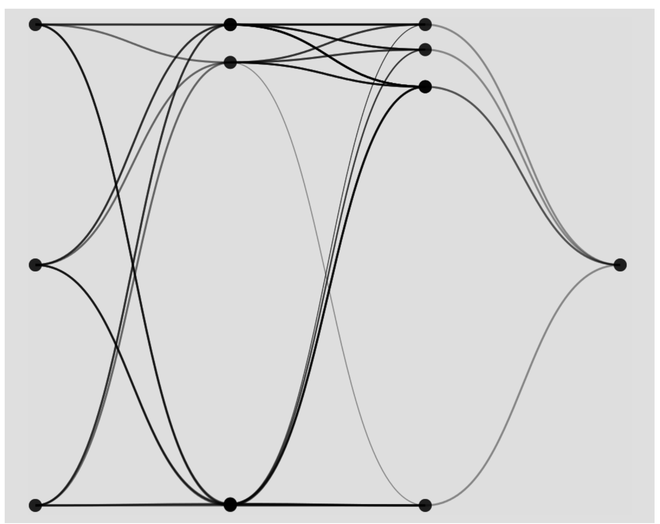


A network trained to recognize 3-way exclusive OR (XOR) with two 16-neuron layers. Visualizing its structure suggests a smaller network with 6 and 5 neuron layers.

Results and Future Work

The goal of visualizing neural network behavior is meant to discover interesting features of a network and how it is used. Through this method, we can see that certain co-activating groups of neurons can be replaced with just one, simplifying the network. Additionally, we can examine these network's behavior for confirmation that their inner workings reflect the desired function -- humanoid robot walking control network splitting in two could indicate left and right leg control.

Future work includes further refining of the rendering algorithm to more explicitly highlight the structure in the presence of different classes of inputs, and additionally how the structure changes during training of the neural network itself.



Both networks have similar performance with the smaller network using less than half the neurons.