TITLE: Neural Networks Explanation via Visualization

Authors: Juan Puebla, Vicente Corral Arreola, Edwin Tomy George, Sophia Montenegro, Martine Ceberio

ABSTRACT:

A growing number of applications rely on Neural Networks for making decisions. Yet, we still do have a good understanding about how decisions are made and whether they are correct. This work focuses on neural networks with the goal of helping users to make more sense about the algorithms subsumed by the network. Our research is meant for Software Engineers (SWEs) who use Neural Networks and are looking to better understand them through visualization. SWEs would be able to see how the data travels through the NN and see what sections of the NN are the most important or doing the most work.

The weight values and output of nodes of a fully connected NN can be obtained but it is difficult to analyze that amount of data. By visualizing the NN as a graph with activated nodes we can quickly find what nodes in the NN are most important. If a specific node is never activated, it could mean that the NN could function without this specific node. Thus, the NN could be optimized, resulting in reduced memory usage and faster NN. Additionally, with visualization, non-technical users can see what's happening in the network without having to understand the math behind it. An example of a current solution to this problem is, e.g., the CNN explainer. This tool focuses on visualizing convolution neural networks (CNNs) as a learning tool. It is described as an interactive visualization tool designed for non-experts to learn and examine CNNs. It helps users to understand the underlying components of CNNs and examine the interactions between low-level mathematical operations and high-level model structures. CNN explainer is meant to teach: it isn't meant for further evaluation of neural networks or to find opportunities for optimization. The tool is limited to classifying images on a pre-defined pre-trained CNN model, which is reasonable because the tool is meant to teach.

In our work, specifically, we propose visualizing a NN as a graph where the activated nodes are highlighted. We provide a user interface that allows customization of FFNN (define layers and nodes) and activation threshold options. The goal is to provide the user with a toolbox that will show how data travels through a NN and shows if optimization of the NN is possible. In our research, we have trained various versions of fully connected neural network models on the MNIST dataset (70,000 images of handwritten digits). We are able to obtain the value of the weights after each epoch: this shows us how the weights change during training. After the model is trained, we can obtain the node output values when a test example is run through the model. The node output values show which nodes are activated. Currently, we are visualizing the layers and nodes on the command line with 'X' representing an active node and 'O' representing an inactive node. We are experimenting with different threshold options, such as the mean of the layer's node output values. The results show the same activation map for images of the same digit class and show unique activation maps for different digit classes.