Juan Puebla
Advisor: Dr. Ceberio

Explaining Neural Networks with Visualization

There is a growing number of applications using neural networks to make decisions. They are used for making critical decisions such as early detection of cancer, autonomous driving, and credit scoring. Given the importance and impact that neural networks have on society and individual's lives, it is crucial that their decisions be validated. Neural networks, because of their sizes, are black-box in nature, which results in a general lack of understanding of how they work and the reasoning behind the decisions they yield. This machine learning model lacks transparency and can be incomprehensible, leading to a lack of trust in Artificial Intelligence.

Explainable Artificial Intelligence (XAI) encompasses methods developed to make the decision-making processes and outputs of artificially intelligent programs more understandable to humans. The work in this research is a contribution to XAI, and focuses on the visualization of neural networks to explain their decisions. The goal of this research is to help people make sense of the decision algorithms subsumed by the network. Our research specifically deals with the visualization of network node activations and weights to understand how data travels through the network and why decisions are made.

We posit that seeing how information travels through the neural network will allow us to gain a better understanding of how neural networks make decisions. We test our ideas using the MNIST dataset. We created images containing individual features of a digit and fed the images to a trained neural network. We observed the activation of nodes as the data travels through the network. Based on the node's weights, we can extract the characteristics that a specific node is looking for. We concluded that it is possible to predict a desired class by activating specific nodes in the first hidden layer. For instance, in our presentation, we will showcase examples where, if we trigger an activation on a node in the first hidden layer, this activation will cause a chain reaction through the network and output the desired class to some probability.