Feature Analysis and Activation Function Study of Image Reconstruction Using Artificial Neural Networks

Shahd Hekal, 2027

The purpose of this study is to elucidate the "black-box" nature of artificial intelligence through explainable AI. Throughout the summer, we worked on studying different artificial neural network structures for the purpose of image reconstruction and exploring how these architectures differ in terms of feature maps or neuron activation relative to their outputs. We essentially studied how, in loose terms, different artificial neural networks "see" or "think". Our main catalyst for this study was the proposal of the Sinusoidal Implicit Neural Network (SIREN) by Sitzmann et. al. in their 2020 paper that compares this specific architecture with other artificial neural networks, and showcases the significant increase in performance when using an optimized sin function compared to previously available activation functions.

As part of our methodology, we worked with a multilayer perceptron (MLP) network to reconstruct an image by creating a grid where the network predicts the pixel color values after a few cycles of training, and comparing the predicted pixel colors from the network with the ground truth or original image. Using this general model as a basis, we tested and compared multiple parameters that affect predicted image quality, including the number of layers in the network, the number of neurons in each layer, and the rate at which the network learns, as well as the functions used to process the given data between layers.

Each neuron in the network can be visualized when activated as a feature map, meaning that we can see exactly what shapes are "lighting up" for each neuron or what each neuron is learning about the image. We analysed the feature maps for a variety of neuron numbers across different architectures and settled on using and analyzing 256 neurons per layer across different activation functions as per the Sitzmann paper. We found that across different activation functions (functions that process data between network layers), the output of the network may seem identical to the user, but how the network learns and how neurons activate wildly differs between activation functions. Piecewise linear functions and their variants create feature maps with sharp edges and shapes, while oscillatory functions cause repeated circular patterns and halos. We categorized different activation functions into groups according to observed similar feature map patterns, and noticed that in some cases, the learning patterns differ significantly between function variants that use sharp vs smooth transitions, even though the function variants are mathematically similar.

We also ran a number of experiments on novel activation functions that we designed from the results of our analyses. We compared these functions to existing functions in terms of the results of network performance and efficiency in using architectural resources (networks that use or activate all their neurons, where all internal layers are critical to performance vs networks that kill or don't activate a lot of their neurons). Additionally, we ran some experiments using inpainting to study how models process noise at different levels and how feature maps or learnt features might differ from their base-line with the introduction of noise. We also ran some limited supplementary experiments on classification tasks to test if our results have correlations between different deep learning task types or if they're constricted to regression only. We added some metrics as well to test the segmentation potential of some of our novel activation functions. This summer research is co-authored by Joram Kim, class of 2027.

Faculty Mentor: Jeova Farias.

Funded by the Bowdoin Research Award from the Office of Student Fellowships and Research.

References: Sitzmann, V., Martel, J. N. P., Bergman, A. W., Lindell, D. B., & Wetzstein, G. (2020). Implicit neural representations with periodic activation functions. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*.