Feature Analysis and Activation Function Study of Image Reconstruction Using Artificial Neural Networks
Joram Kim, Class of 2027

This project was conducted in collaboration with Shahd Hekal, Class of 2027. Our primary goal was to explore **explainable artificial intelligence** by investigating how activation functions influence the internal behavior of artificial neural networks. We used image reconstruction tasks as a lens into how networks "think, " focusing on how they activate, learn, and represent information. To do this we visualized spatial representation of neurons' output when reconstructing images at each layer. These maps reveal what features of an image each individual neuron is sensitive to, offering a window into how networks decompose and reconstruct visual data. By analyzing these feature maps and evaluating architectural efficiency, we aimed to illuminate the **black box** of learning dynamics and assess how specific activation functions affect both model performance and neuron and pixel level interpretability.

To better understand per-neuron behavior, we applied K-means clustering to the feature maps generated by neurons, allowing us to identify patterns in how different activation functions learn and compare how neurons group similar pixel regions. Furthermore, we performed per-layer feature selection by identifying the most prominent or active neurons in each layer providing further insight into the learning capacity and selectivity of different activation functions. These results prompted us to investigate a key limitation of

Rectified Linear Units (ReLU) having the tendency to produce "dead" neurons. ReLU, while widely used due to its simplicity and computational efficiency, suffers from a high percentage of inactive or dead neurons of up to 90% per layer in MLP architectures. This limits effective neuron utilization and can lead to inefficient learning. Our work compared ReLU with both bounded and oscillatory alternatives, including bounded, sinusoidal activations such as SIREN (Sitzmann et al., 2020) and bounded functions like a modified version of mReLU (Zhao & Griffin, 2016), focusing on their impact on model performance and internal weight behavior. We found that networks that use sinusoidal oscillations tend to perform better in image reconstruction than ReLU. Furthermore, we also found that networks composed of bounded activations tend to have more impactful learning at each active neuron while also keeping the weights that connect each layer of neurons more controlled.

From these conclusions we proposed a novel, bounded, jagged, sin-like activation function (JaSIN) that combines the computation efficiency of ReLU with the controlled weights and reconstruction performance of bounded and oscillatory functions. Through experiments, we evaluated each activation using Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), effective ranking, and average weight magnitude per layer.

Our results showed that bounded activation functions consistently produced lower mean weights per layer, higher effective rankings, and better PSNR/SSIM scores than the unbounded counterparts. These trends support the hypothesis that bounded activations promote implicit weight regularization and improved architectural efficiency. In contrast, unbounded functions showed higher weight magnitudes and more neuron inactivity, suggesting less efficient feature utilization.

Our main findings were complemented by two exploratory side investigations aimed at further opening up the black box of deep learning. In the first, we introduced varying levels of noise to input images and examined how this affected the proportion of "dead" neurons. In the second, we trained many networks on the MNIST dataset and then attempted to classify digits using only the learned weights, independent of input data. Both efforts reinforced our core aim: to make neural computation more transparent and better understood.

Faculty Mentor: Jeová Farias
Funded by the Freedman Summer Research Fellowship in Computer Science
References: Sitzmann, V ., Martel, J. N. P., Bergman, A. W., Lindell, D. B., & Wetzstein, G. (2020). *Implicit neural representations with periodic activation functions*. Advances in Neural Information Processing Systems, 33, 7462–7473.
Zhao, Q., & Griffin, L. D. (2016). *Suppressing the unusual: Towards robust CNNs using symmetric activation functions*. arXiv preprint arXiv:1603.05145. https://arxiv.org/abs/1603.05145