High-speed Training Using Binary Neural Networks

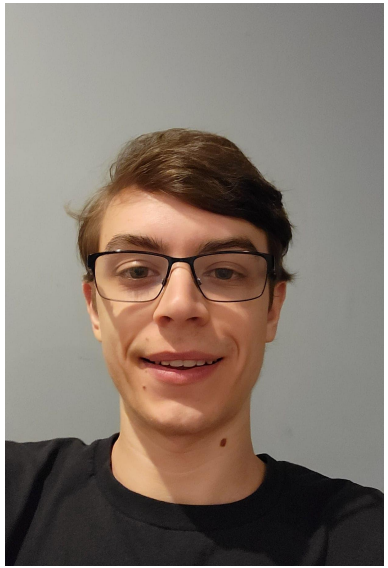
Project Mentor: Dr. Richard Martin

Team Members

Sachin Mathew ‘22

Daniel Maevsky ‘24

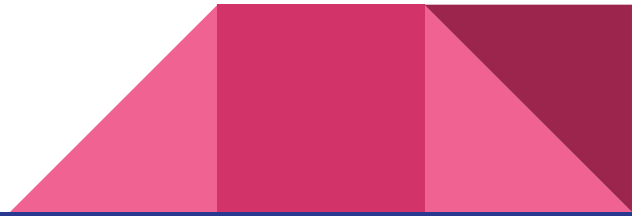
Daniel Chen ‘24



Tommy Forzani ‘24

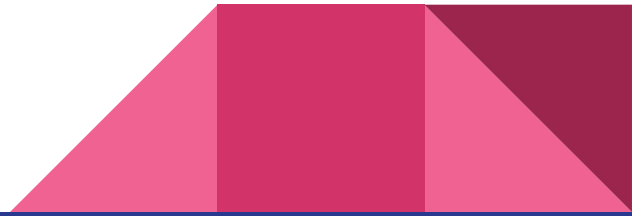
Our Goals

- Training machine learning systems is currently very slow and chip space intensive.
- Recent work has shown promise by using simpler representations of numbers than the commonly used floating point ones.
- In this project, we will create and measure neural networks which use only binary or fixed point numbers for both training and inference to test their theoretical space/energy efficiency.

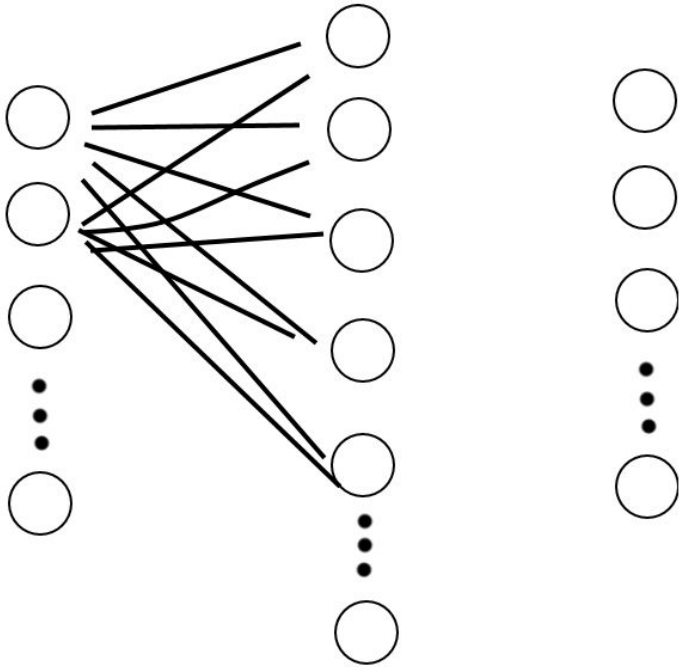


Weekly Progress

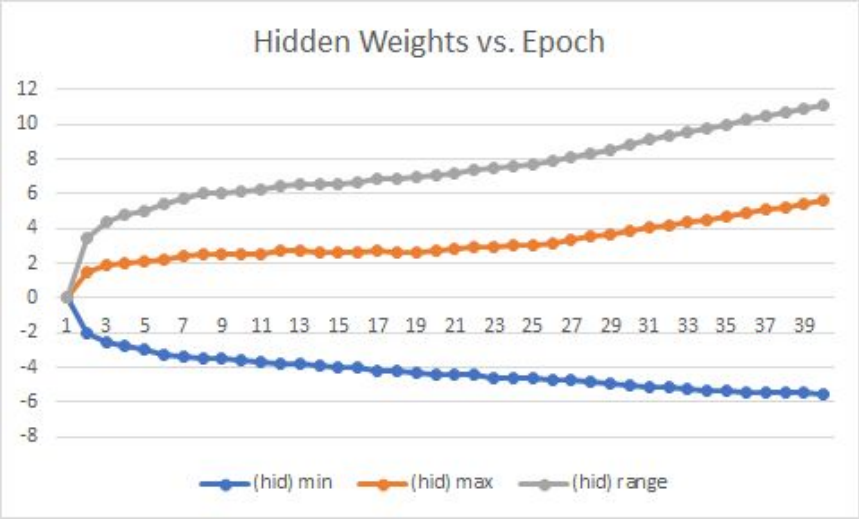
- Began checking intermittent accuracy for GONN model using validation set
- Plotted absolute max, min, and range alongside accuracy data for the model
- Found accuracy plateau to help find fixed point range



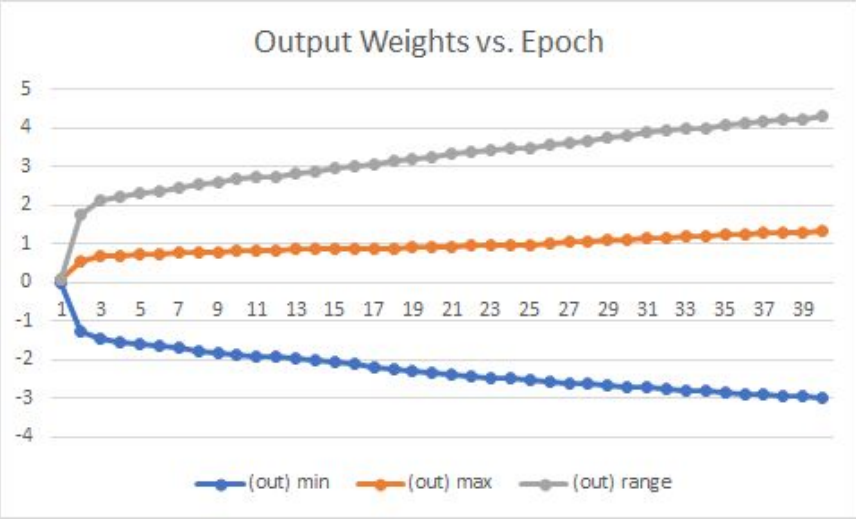
Three Layers in our MNIST network



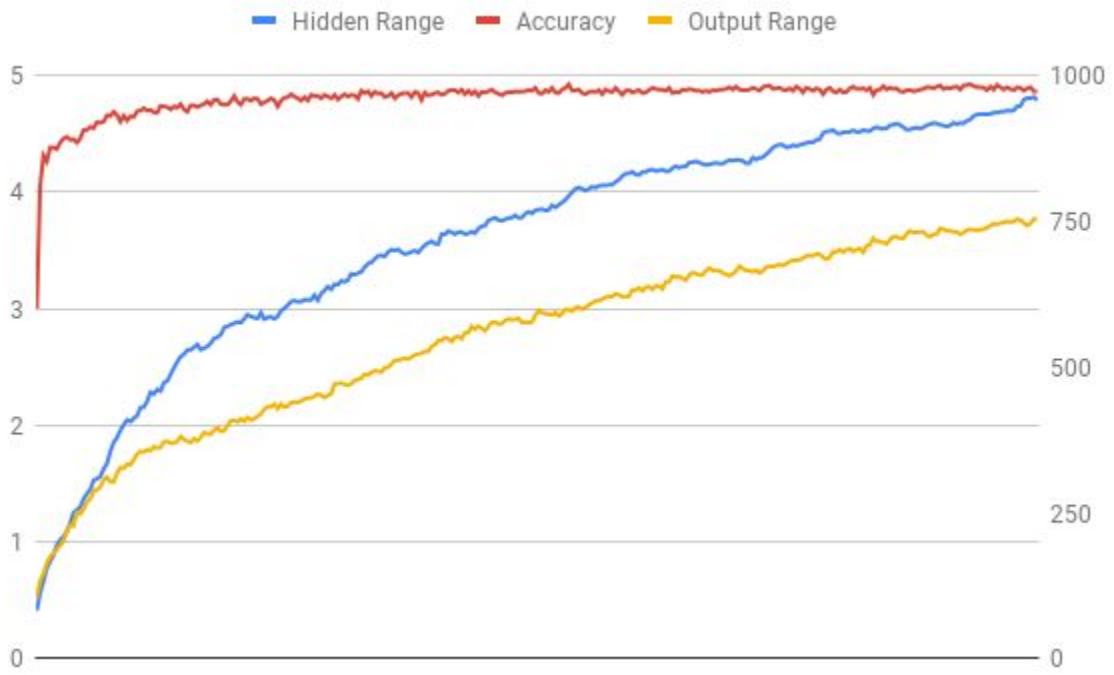
Input Layer to Hidden Layer Weights



Hidden Layer to Output Layer Weights



Weight Value Range and Accuracy vs Training Time



Next Steps

- Run more experiments to confirm the instability of the dynamic range, with more epochs (wary of over-fitting)
- Determine whether or not the range, though it might be unstable, is still usable somehow
- Discuss implementation of fixed-point numbers and how we will move forward with the research