Project 3 Neural Network - Final Milestone

Karaen Senthilkumar (CNET ID: karaen)

March 12, 2025

1 Performance Metrics

Version	Processor or GPU	Accuracy	Grind Rate	Training Time	TPB or CPU Cores
GPU native	A100	97.549	543,436	4.6003	256
GPU native	V100	97.549	521,468	4.7941	256
GPU cuBLAS	A100	97.549	480,517	5.2027	-
CPU native	Caslake (Intel)	97.750	36,824	67.8887	32
CPU BLAS	Caslake (Intel)	97.439	225,475	11.0877	32

Table 1: Performance metrics obtained for the following: learning rate $\alpha = 0.1$, batch size $nb = 500$, $epochs = 50$, using $50K$ training samples and $10K$ validations samples.

2 Shortcomings and Observations

- I tried implementing tiled matrix multiplication using shared memory for the native version of the GPU, but the correctness was a bit off. Although a little slower, the current implementation has correct loss curves in terms of training and validation. So I finalized the same.
- For the GPU implementations, I have all the matrix operations (add, subtract, multiply, and copy) in the form of kernels. So, if the total number of threads (blocks x ntpb) is less than the entries handled in the kernel, the results are unpredictable. Since the goal was to extract maximum performance and correctness, I had to sacrifice generality.
- I expected the cuBLAS version to perform better than the GPU native version, but for me, it was a little slower. I guess it's got to do with the organization of my arrays, which could have resulted in more cache misses when cuBLAS tries to work in column major.
- Before running the program, please set appropriate path for the MNIST datasets to be loaded in `mnist.h` file under the *final* directory.

See next page for loss curves.

3 Loss curves

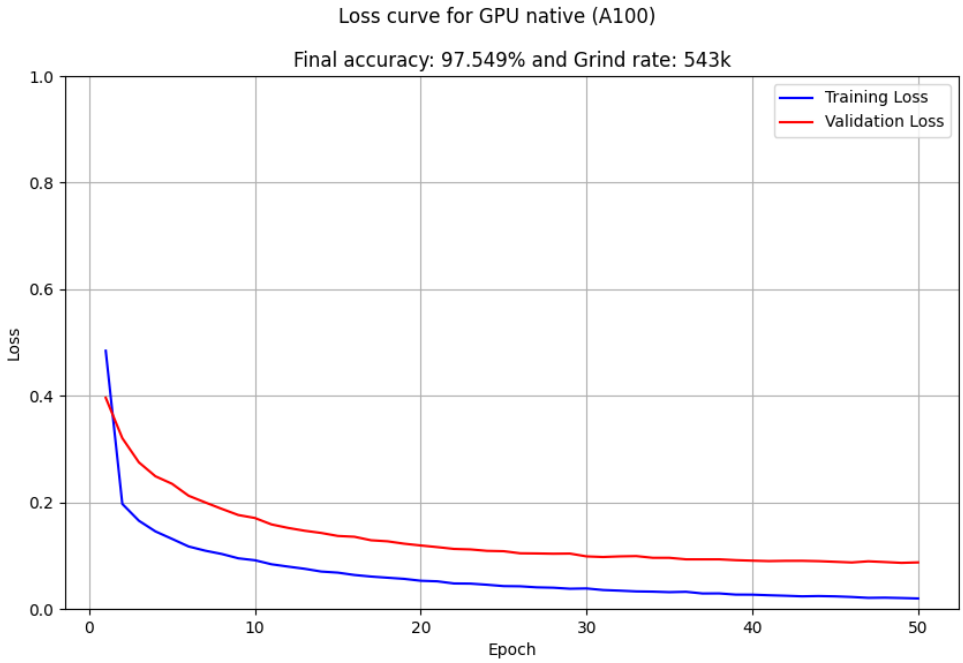

Figure 1: GPU native run on A100 (nvcc) for $\alpha = 0.1$, $nb = 500$, $epochs = 50$

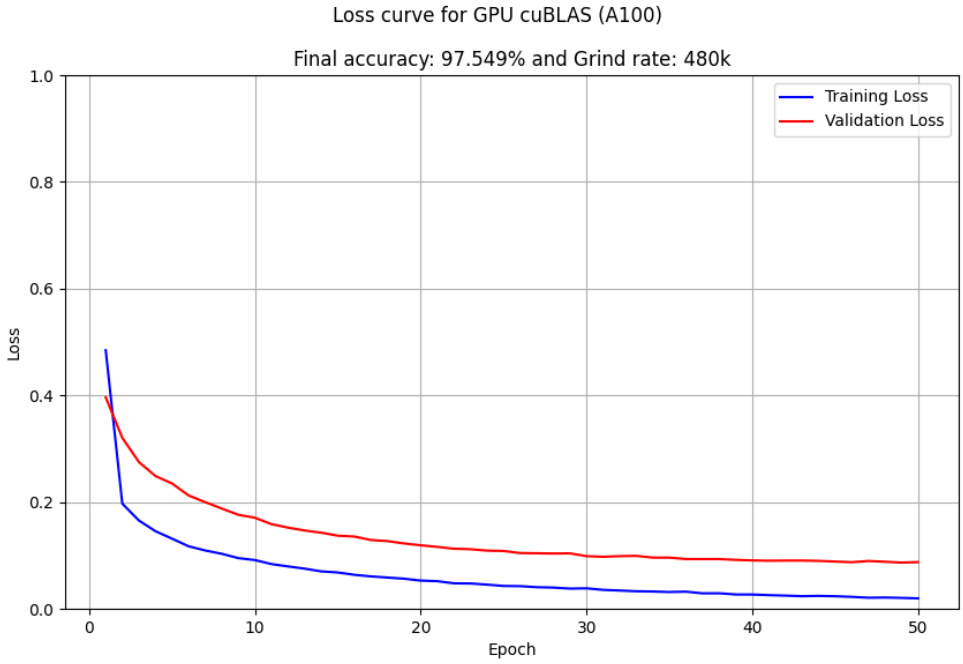

Figure 2: GPU cuBLAS run on A100 (nvcc) for $\alpha = 0.1$, $nb = 500$, $epochs = 50$

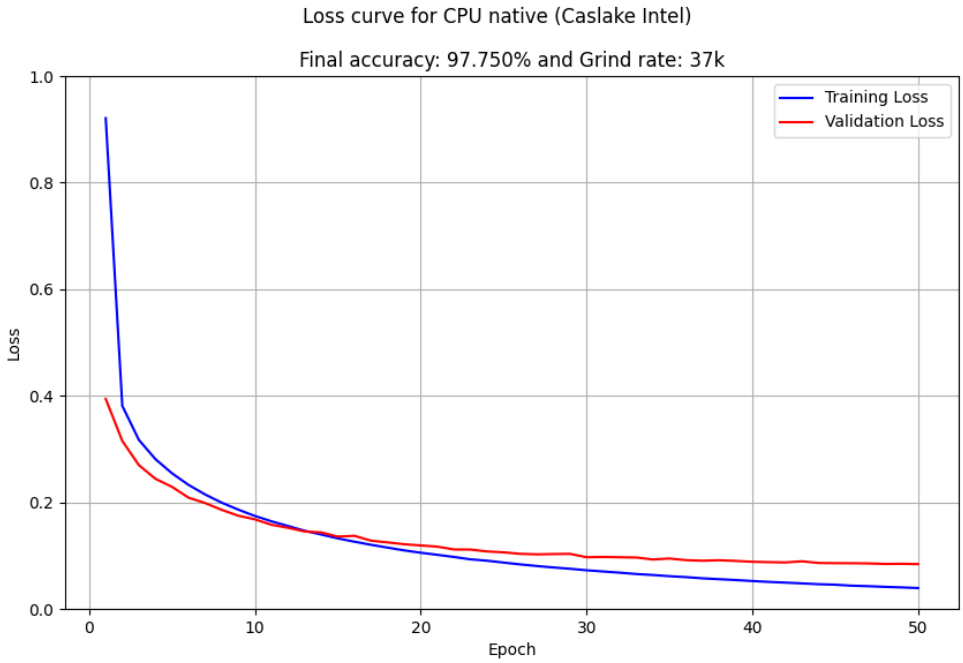

Figure 3: CPU native run on Caslake (icx) for $\alpha = 0.1$, $nb = 500$, $epochs = 50$

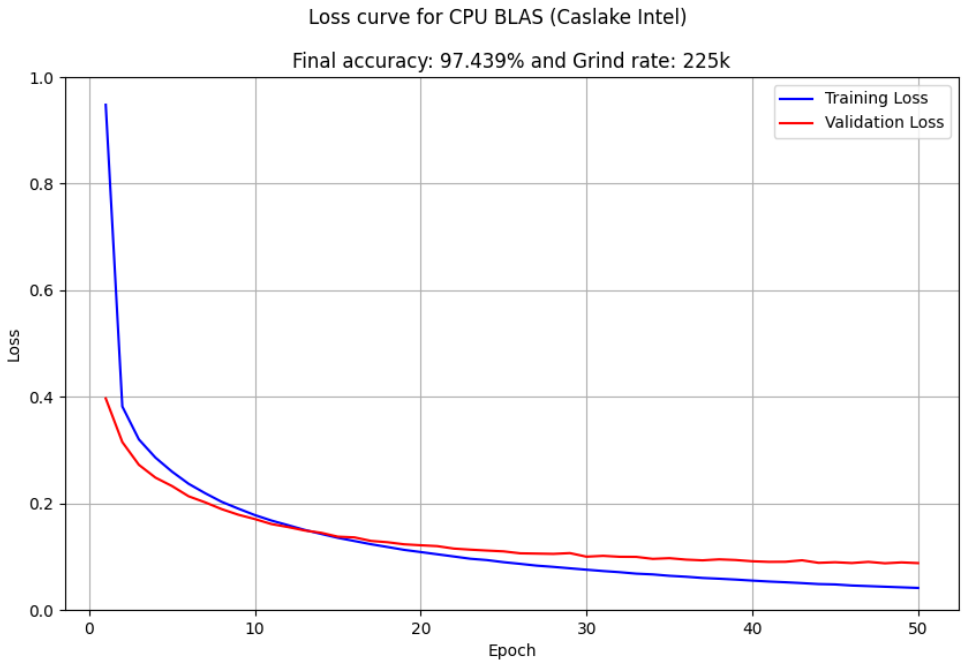

Figure 4: CPU BLAS run on Caslake (icx) for $\alpha = 0.1$, $nb = 500$, $epochs = 50$