1. Analyze your results, when does it make sense to use the various approaches?

Use the CPU approach when the matrix size is small (smaller than 50\*50 approximately). For larger matrix sizes, it would make sense to first implement the GPU Tiled approach and switch to using the cuBLAS approach if the size exceeds around 500\*500.

1. How did your speed compare with cuBLAS?

When the matrix size is small (10\*10), all the approaches demonstrate a faster runtime when compared to cuBLAS. As the size increases, CPU no longer runs as fast as the GPU approaches, including cuBLAS. However, when the matrix size is 100\*100, GPU and GPU Tiled still run faster than cuBLAS. cuBLAS proves to be the fastest approach only when the matrix size reaches 1000\*1000.

1. What went well with this assignment?

Understanding the requirements and writing the code for the CPU and GPU approach went pretty well.

1. What was difficult?

Understanding the cuBLAS code and implementing our own version was challenging. In addition, finding a way to compare the runtime results of different approaches took some thinking.

1. How would you approach differently?

Every time: start earlier.