OpenMP

kevin Lopez Chavez

March 10, 2025

1 Report deliverables

- The file is named `report.pdf`.
- An explanation of why I think this application is a good choice for OpenMP (Include the link to the exact location where the original program is referenced from) in Sections **??**.
- An explanation of the code, the program's functionality, flow, and OpenMP acceleration in Section **??**.
- Explained the estimated speed up in Section **??**.
- Proof of achieved speedup with expectations in Section **??**.

2 Explanation why data normalization is a good choice of program for OpenMP

OpenMP can achieve a high speed over independent for-loop interactions, and it's the idea of why it can achieve a high speedup on z-score normalization (or any normalization strategy, also *batch normalization*). In the case of z-score normalization, the total(sum) must be calculated, which can be parallelized. Calculate the difference to the mean, which can also be parallelized, and update all the current data, which can also be normalized. With that being said, it's possible to parallelize all the different operations in z-score normalization.

2.1 Links to Exact Locations

I referred to the following articles to understand the implementation and usage of Z-Score Normalization.

Explanation and example usage of data normalization:

- `https://developers.google.com/machine-learning/crash-course/numerical-data/normalization#z-score_scaling`

The code implementation of the Z-score normalization (Scikit-learn repository):

- `https://github.com/scikit-learn/scikit-learn/blob/main/sklearn/preprocessing/_data.py`

3 Code Explanation

Here, I will explain the original code implementation in Section **??** and the OpenMP code implementation in Section **??**.

3.1 Original Algorithm

The original algorithm (not optimized) uses the basic functions of mean and standard deviation to update each value individually and one by one. Since this needs to be run a single thread or process at a time, it takes n times to finish executing the algorithm, where n is the number of data points. I have made functions to find the mean and standard deviation. And used in the main normalization function. Each function is defined as follows:

- The following formula is used to calculate the *Z-score normalization*: $Z = \frac{\vec{X}-\mu}{\sigma}$. where: $\vec{X}$ is the original input, μ is the mean of $\vec{X}$, σ is the standard deviation of $\vec{X}$.
- The mean is calculated by $\mu = \frac{\sum_{i=1}^{n} x_i}{n}$; where: x_i is the ith data point, and n is the total points.
- The standard deviation is calculated by: $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}}$.

3.2 OpenMP

It is possible to use OpenMP to handle multiple data points simultaneously. In the case of parallelizing z-score-normalization, it is possible to parallelize a lot of the work done on the data. Below are the different loops that can be parallelized and provide a high speed-up.

- Calculating the sum (needed for the mean).
- Calculating the sum of squared deviations (needed for the standard deviation)
- Z-score Normalizing the actual data.

3.2.1 Matrix Normalization

The following formula is used to calculate the *Z-score normalization*: $Z = \frac{\vec{X}-\mu}{\sigma}$. where: $\vec{X}$ is the original input; μ is the mean of $\vec{X}$; σ is the standard deviation of $\vec{X}$; In my case of normalizing data with *Z-score normalization*, I need to calculate the mean (Section **??**)and Standard Deviation (Section **??**), which is parallelized using OpenMP.

Once I have the mean and standard deviation, we must update every value based on the formula. To speed up the process of updating every value, it is also possible to use OpenMP, handling n data points at a time. Where n is the number of cores allowed to run parallel done by the function`omp_set_num_threads(n)`. It is possible to get the best speed up when using the maximum number of cores in the PC by using the function `omp_get_max_threads()`

To actually normalize the data, it is possible to use the basic formula from the original function. Since this function does not have any dependencies, it is possible to optimize it using OpenMP basic for loop `#pragma omp parallel for`. Using this OpenMP macro allows the compiler to distribute the work between the different processors configured and achieve higher speed up compared to the single-threaded program.

3.2.2 Mean

The mean is calculated by $\mu = \frac{\sum_{i=1}^{n} x_i}{n}$; where: x_i is the ith data point, and n is the total number of data points.

To speed up the mean calculation, it is possible to calculate the sum in parallel; the only issue is that the sum variable is a value shared across all the running processes. OpenMP does allow to speed this up even with the dependency by using the **reduction** keyword along with the variable shared. The complete macro to speed up the sum calculation is `#pragma omp parallel for reduction(+:sum)` where the sum is the shared variable(each process would have its own copy of it where they accumulate the sum). In the end, OpenMP reduces all the sum variables copied into the processes into one global variable.

3.2.3 Standard Deviation

The standard deviation is calculated by: $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}}$; where: x_i is the ith data point, μ is the mean, n is the total points.

With the mean of the whole data, it is possible to get the deviation of every data point and sum to the variable *sum*. Again, since it is necessary to write to the same variable name, this would cause race conditions; in this case, we again use the same method **reduction** from opemMP to handle it. The full line to optimize all the operations is `#pragma omp parallel for reduction(+:sum)` where *sum* variable is the sum of squared deviations. In the end, to get the variance, we just divide it by the number of elements.

3.3 Compilation Steps and Flags

I used the following commands to compile the code

- **Original version:**
  - `$ g++ ./z_score_norm_original.cpp -o original.exe`
- **OpenMP version:**
  - `$ g++ -fopenmp ./z_score_norm_OpenMP.cpp -o OpenMP.exe`

Where the flags mean the following:

- `g++`: Compiler used to compile the C++ code.
- `-fopenmp`: This flag enables OpenMP.
- `./z_score_norm_original.cpp ./z_score_norm_OpenMP.cpp`: The source file.
- `-o original.exe`: The `-o`: name of the output executable file.

to run the code and get meaningful results the code was run 20 times in a row, for this I used the bash script below.

4 Estimated Speedup

Looking down from the big picture, we might assume that this should get a speed up of n where n is the number of processors, but there is one thing that we still need to take into account, and there are still some portions of the code that need to be handled sequentially. In the **reduce** method of OpenMP, the multiple processors need to create private locations of the sum variable that allow it to accumulate the sum value across different processors, and it would need to reduce it back to one single sum. This would take some time with the addition of the rest of the code that needs to run in a single core. For example, entering functions, making the last set of operations(divisions), and other operations.

5 Proof of Achieved Speedup

To prove the speed-up achieved, I have attached the logs and Figures **??**,**??**. I posted the logs (run the programs 20 times) for the original version in section **??** and the oepnMP version in section **??**.

The programs were executed 20 times, the numbers are shown in table **??**:

$$\text{Speed-up} = \frac{\text{Time for Original Version}}{\text{Time for OpenMP Version}} = \frac{41.79}{6.48} \approx 6.45$$

One can assume that the speed-up would be close to the number of parallel processes running. When looking in, we can see that there is more to why it won't run at the expected theoretical max. There are portions of the code that need to run in a single process, and also, the OpenMP instructions can add some overhead(The reduce method would need to make copies of the sum variable in each process and then reduce it back to one). Therefore, getting a speedup of ≈ 6.45 is a valid number for this application.

Table 1: OpenMP and Original Comparison

Run_number	original_time	OpenMP_time
run #1	52.54	6.04
run #2	39.52	5.84
run #3	36.36	6.01
run #4	36.04	6.46
run #5	33.96	6.37
run #6	52.62	5.22
run #7	52.31	6.12
run #8	31.13	6.70
run #9	31.12	5.68
run #10	30.87	5.74
run #11	36.23	5.19
run #12	60.24	4.43
run #13	36.28	5.55
run #14	37.14	6.05
run #15	33.95	8.20
run #16	36.74	8.13
run #17	60.59	8.15
run #18	63.19	7.24
run #19	36.95	8.33
run #20	38.10	8.09
average	**41.79**	**6.48**
standard_deviation	10.71	1.16
speed_up	1.00	**6.45**

Figure 1: Proof of speed up(Original)

Figure 2: Proof of speed up (OpenMP)

5.1 Logs for Original File Execution

5.2 Logs for OpenMP File Execution