Katholieke
Universiteit
Leuven

**Department of
Computer Science**

BIG DATA ANALYTICS PROGRAMMING

Parallel Computing and Data Processing

Cas Coopman (r0996992)

Academic year 2023–2024

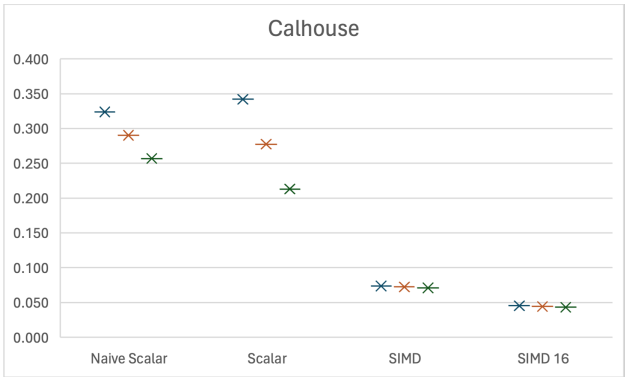

Figure 1: The box plots for the different evaluation algorithms on the Calhouse dataset.

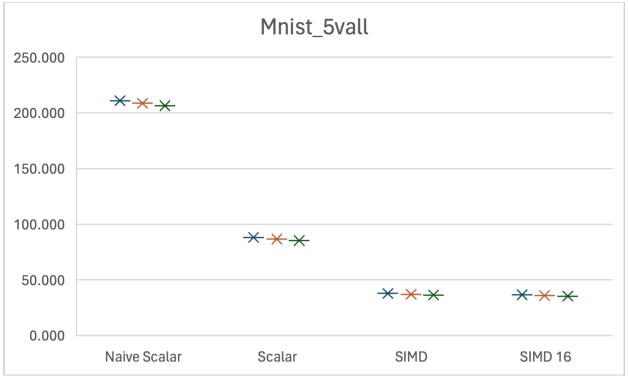

Figure 2: The box plots for the different evaluation algorithms on the Mnist_5vall dataset.

1 SIMD

The SIMD algorithm for this assignment was made on a MacBook M2. The device's CPU makes use of the Neon instruction set.

The average run time and the standard deviation of the algorithms can be seen on the next page, in figure 3. These calculations were made by running every algorithm for 100 iterations through each dataset. The **average speedup** of the SIMD implementation **is 2.8**. The highest observable speedup of the SIMD implementation over the scalar one, is 3.8 times. This can be seen on the Calhouse dataset.

Aside from the requested SIMD and Scalar algorithms, a Naive Scalar and SIMD 16 are also added for demonstration purposes. The Naive scalar, acting as a baseline, indicates the performance of a non-optimized implementation on this MacBook M2. The SIMD 16 algorithm, having the lowest and most consistent running time, achieves significant performance boosts on the Calhouse and Cpusmall dataset. However, for the Diamonds and Cpusmall datatset, some calculations were off. I was not able to pinpoint the error in the implementation.

The optimized Scalar algorithm unrolls the loops for efficiency gains. The difference between SIMD and SIMD 16, is that the latter reuses the coefficient vector for different rows. By adding this vertical stride, over the rows, a significant performance boost was achieved. In particular, on the Calhouse dataset a maximum speedup of 6,3 over the (optimized) Scalar version is achieved. Increasing the horizontal stride, from the basic four number vectorization, did not increase the performance on this device however. Such a implementation can be seen in evaluate_simd_wider. This is unexpected as a MacBook M2 is supposed to have multiple usable registers.

For easier visual comparison, box plots of the running times for each dataset can be seen in figure 1 and figure 2.

	Calhouse				
	Running Time		Prediction error		
	Average Time	*Standard Deviation*	*Average Error*	*Variance*	*SqrdError*
Naive Scalar	0.290	0.033	-0.014%	0.116	2386
Scalar	0.278	0.064	-0.014%	0.116	2386
SIMD	0.072	0.001	-0.014%	0.116	2386
SIMD 16	0.044	0.001	-0.014%	0.118	2436

	Allstate				
	Running Time		Prediction error		
	Average Time	Standard Deviation	Average Error	Variance	SqrdError
Naive Scalar	78.938	2.911	0.003%	0.338	63669
Scalar	30.442	2.881	0.003%	0.338	63669
SIMD	13.918	0.807	0.003%	0.338	63669
SIMD 16	13.360	0.933	0.006%	0.380	71491

	Diamonds				
	Running Time		Prediction error		
	Average Time	Standard Deviation	Average Error	Variance	SqrdError
Naive Scalar	2.934	0.065	-0.006%	0.031	1662
Scalar	1.402	0.518	-0.006%	0.031	1662
SIMD	0.678	0.025	-0.006%	0.031	1662
SIMD 16	0.510	0.029	0.242%	10191	549710000

	Cpusmall				
	Running Time		Prediction error		
	Average Time	Standard Deviation	Average Error	Variance	SqrdError
Naive Scalar	0.224	0.024	-0.042%	96.425	789915
Scalar	0.166	0.038	-0.042%	96.425	789915
SIMD	0.045	0.001	-0.042%	96.425	789915
SIMD 16	0.028	0.001	-0.124%	142.88	1170510

	Mnist_5vall				
	Running Time		Prediction error		
	Average Time	Standard Deviation	Average Error	Variance	SqrdError
Naive Scalar	208.666	2.298	-0.134%	0.018	1290
Scalar	86.712	1.427	-0.134%	0.018	1290
SIMD	37.054	0.715	-0.134%	0.018	1290
SIMD 16	35.976	0.660	-0.459%	0.026	1842

Figure 3: The table with the average running and prediction statistics for the different algorithms and datasets.

2 Distributed Computing

The **heatmap for Stephen Curry can be seen in figure 4 below**. Blue indicates a higher hit rate. We conclude that he scores many of the shots from around the hoop. The shots from further are, as expected, more frequently missed. Finally, we can clearly distinct the three-point line. For illustration purposes, an official basketball court has been overlaid on this heatmap in figure 5.

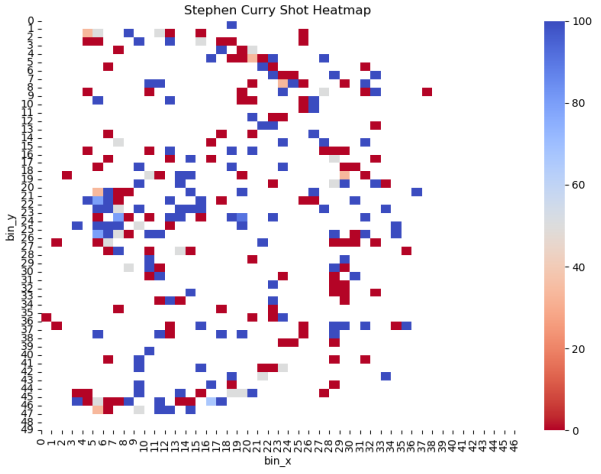

Figure 4: The shot heat map for Stephen Curry. Blue indicates more shots made, red indicates more shots missed.

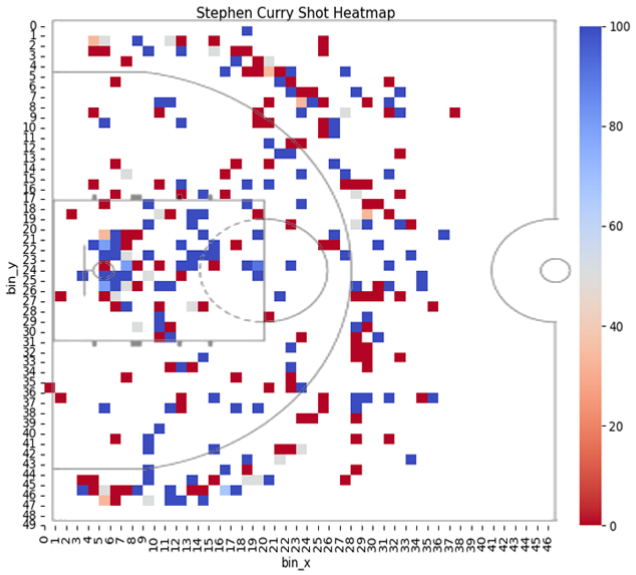

Figure 5: The same shot heat map as figure 4 but situated on an official basketball court.