A1 - Benchmarking Crimes

Linzh (Zhongyu Zhao)

1. What is benchmarking (0.25 point)?

In many situations, people always wonder how good some products are. However, if different products use different methods to evaluate their performance, the customers will never know whether product A is better than product B.

Therefore, people proposed benchmarking methods to address this problem. Benchmarking is the method by which different products are run on the same public dataset or task, and the results are compared to determine their performance. The important point of benchmarking is that it must be run on the same task/dataset, otherwise, the result will be unfair.

For example, to measure your gaming PC, we could use 3DMark, which will give you a score for your CPU, GPU, etc. In the field of neural network, especially, object detection or semantic segmentation in computer vision, we always use PASCAL VOC data to evaluate the performance of the model we proposed.

2. Why is benchmarking important? (0.25 points)

As we discussed above, the performance is hard to describe without a unify standard. Moreover, some merchant also use the dataset that reflect the best performance. Thus, we can only get the true performance by using the same dataset. Namely, benchmarking enables the comparison between methods and products. In this way, the performance or other result/analysis could be derived fairly.

3. How would you represent the results described in this comic strip? (0.5 points) https://imgs.xkcd.com/comics/percentage_points.png

Garyton had a 20% polling rate before his speech. However, due to the promise of a tax break and UAV authorization, his polling rate is now about 16.2%, which means 19% of his supporters changed their minds.

4. What can you infer from Table 1 and Table 2? (1 point).

In the Table 1, we can see that the author set the performance of processor R as the standard to compare three processors.

The performances of process M and Z varies among tasks. Specifically, the processor M have much better performance on task E, f, I, K, and not good enough to handle task H. Similarly, processor Z is not good at task H, I. However, a processor do not good at some tasks does not mean that the processor is not good. The mean value is meaningless. It seems that the processor R has the best performance. However, the table 2 give us a different result.

If we centered the value to processor M, we will find that the result is completely different based on the average. The processor M becomes the best processor, and the processor R is much more worse than M.

The conflict of results show that this method is unstable and unreliable. For the multiple benchmark, we should not use arithmetic mean to represent the actual performance.

5. How would you use the data in Table 1 and Table 2 to compare the performance of the processors (1 point)

First, I will analyze the differences and similarity among those benchmarks and processor architecture to figure out why different processors are good at different benchmarks. For example, if a processor support AVX-512 (Advanced Vector Extensions 512), this processor will have better performance on neural network training and inferring. However, if a processor support AVX-512, the power consumption will also increase significantly.

With more information and analyze, we will get more information related to the strength and weakness of the processor. Then we will give the my conclusion of the comparing.

Moreover, if I must give a summary without any future information, I will say like: about benchmark E, processor M and Z are much better than processor R, and Z is the best processor for this task. In benchmark H, processor R has the best performance, and there is no huge difference between M and Z.