Background:

The research and development within the field of Artificial Intelligence (AI) has extensively increased over the last decade which focuses on capabilities such as Object and Speech Recognition, Game Playing, and machine translation. This is made possible by the increasingly large computationally intensive deep-learning machine. Within 6 years, the training cost for the most advanced AI machine has increased by 300,000 times with a trend of doubling its cost every month. It is also worth noting that although the rise of computational cost increase significantly, the accuracy performance of the AI does not linearly correlate with the cost. Hence, the presence of diminishing returns of increased cost. We refer to AI that focuses on only accuracy performance with such characteristics as RED AI.

Problem Discuss:

The estimated carbon footprint of RED AI is both environmentally unfriendly and expensive. The article introduces GREEN AI – an alternative AI research focus on efficiency to reduce waste is discussed as a solution that would also allow people to do AI research with fewer resources.

Within the AI community, researchers only focus on the accuracy of their machines, not efficiency in reducing waste, time, and resources. This is shown in papers from AI conferences which the majority target the resulting accuracy of their machine. As mentioned before, AI research based on efficiency would find the optimal way to decrease the resources needed and its carbon waste. There are many criteria we could do to determine the efficiency capabilities of machines such as Carbon emission, Electricity usage, Elapsed real-time, Number of parameters, and FPO.

With carbon emission, we would like to minimize the quantity of carbon waste we are producing with our AI research. However, it would be hard to evaluate the exact value and it would differ from location to location since of the difference in energy infrastructure.

Electricity usage and carbon emission have a positive correlation. Nevertheless, the GPUs use within the research would provide an exact numerical value of electricity consumed by the machine. However, this is hardware-dependent. Therefore, it would be difficult to compare it with other machines.

Elapsed real time would naturally measure the efficiency of the model. It would also correlate with the amount of work. A faster model would do less computational work. However, the time needed is heavily affected by many factors such as the type of hardware and the amount of hardware. This would hinder the comparison with other machines with different hardware usage structures.

The number of parameters correlates with the amount of work. However, unlike the previous method, it is not affected by hardware but is affected by the amount of memory usage within its computational process. Hence, different machines with the same number of parameters but with different algorithms would perform a different amount of works.

FPO or floating-point operations are discussed as the optimal measure of efficiency in AI research.

“FPO provides an estimate of the amount of work performed by a computational process. It is computed analytically by defining a cost to two base operations, ADD and MUL. Based on these operations, the FPO cost of any machine learning abstract operations can be computed as a recursive function of these two operations.” FPO directly computes the amount of work done by the running machine when doing a specific instance of a model and is thus tied to the amount of energy consumed. FPO is not hardware-dependent. FPO also considers the amount of work done at each step.

To promote the research, development, and usage of GREEN AI, reporting the budget vs performance curves is essential to provide information about the difference in size, algorithm, and hardware that would affect the machine working cost.