Analysis: Computational Limits of Deep Learning

This article purports the computational demands of deep learning applications in five prominent application areas: Image classification, object detection, question answering, named entity recognition, and machine translation. Deep learning has been one of the recent achievements, defeating humans in performance in image and voice recognition, translation, and other tasks. But this progress has been achieved by using enormous computing power. This paper shows that progress in five aforementioned areas is strongly reliant on increases in computing power. Extrapolating forward, this reliance reveals that progress along current lines is rapidly becoming economically, technically, and environmentally unsustainable. Continued progress will require dramatically more computationally efficient methods, which either will have to come from changes to deep learning itself, or from moving to other machine learning method.

Conclusion

The explosion in computing power used for deep learning models has set new benchmarks for computer performance on a wide range of tasks. However, deep learning's prodigious appetite for computing power imposes a limit on how far it can improve performance in its current form, particularly in an era when improvements in hardware performance are slowing. This paper shows that the computational limits of deep learning will soon be constraining for a range of applications, making the achievement of important benchmark milestones impossible if current trajectories hold. Finally, we have discussed the likely impact of these computational limits: Forcing deep learning toward less computationally intensive methods of improvement and pushing machine learning toward techniques that are more computationally efficient than deep learning.

Likes:

1. The team examined more than 1000 research papers in image classification, object detection, machine translation and other areas, looking at the computational requirements of the tasks.
2. Deep learning being forced towards less computationally intensive methods of improvement, or else machine learning being pushed towards techniques that are more computationally efficient than deep learning.
3. I think this was one of the best papers to read. Usually I need some more research to get the gist of any article. This paper is now one of my favorites.
4. Their ideas were interesting, they say that most of the computations today involve GPU and TPU and with the performance diminishing they propose moving towards Quantum Computing and Computational complexity.

Dislikes:

1. From the graphs, I think that the predictions made are for the average models and not for the best models.
2. I agree with them on moving towards techniques that are more computationally efficient but I can not disagree to Moore's law either, that states that we can expect the speed and capability of our computers to increase every couple of years, and we will pay less for them. Another tenet of Moore's Law asserts that this growth is exponential.

It still needs me a lot of paper reading and knowledge to criticize any work, but I was researching on the internet about this paper and a couple of interesting comments I found were:

1. Their datasets are all top-right censored: if you assume that over time, model performance conditioned on compute budget goes up, then they are missing a bunch of future points in the top-right of each plot.
2. It's amazing that they wrote an entire paper trying to estimate performance scaling with compute, and ignored what looks like the entire literature doing actual controlled highly-precise experiments on scaling up fixed architectures (no citations to any of them that I could see) in favor of grabbing random datapoints from the overall literature.

I could not formulate any opinion on these points but since I know that you acquainted with a lot of literature, I just wanted to ask what do you think about these two points?