Phase 4: Summary of Research Paper

Dietrich Nigh

Paper: Transformers for Image Recognition at Scale (https://arxiv.org/abs/2010.11929)

I. What was the original context of this paper when it was written?

Transformers are presently quite popular thanks to ChatGTP and its '**G**eneric **P**re-trained **T**ransformer' backbone. Transformers are, and were, very popular and very flexible; however, they are also quite resource intensive and require a large amount of training data to be useful. They are typically used with natural language processing or time-series problems. They were not used for image classification problems.

Convolutional neural networks (CNNs) consistently outperformed these models in image recognition tasks. As the internet and data science have matured, progressively larger datasets have been produced. This prompted Google to reassess a transformer's image classing potential in the era of Big Data.

I. Summary of the paper findings/outcomes

Google was able to marginally alter its current vision transformer to produce impressive results after training on larger datasets. The team began comparing their ViT, Vision Transformer, against state-of-the-art CNN, Big Transfer or BiT.

Smaller datasets favored BiT, but, as soon as more training data was provided, ViT consistently surpassed or approximated BiT's performance. At 1 million training images, BiT was far and away the model of choice, scoring almost 10% higher accuracy on the test set. The story quickly changes. At 14 million images, the accuracy of the models is almost equivalent. At 300 million, ViT now surpasses BiT. This is true even when they created a ensemble model composed of several previous state-of-the-art CNNs. While the margin by which ViT surpasses BiT is small, ViT uses a quarter of the computational resources, making it a marked improvement.

I. How can this paper inform your work as a junior data scientist?

> I am going into an industry where real-time image classification is paramount. For the longest time, machine learning engineers have been steered toward CNNs. The battle for many revolves around single stage vs two stage models, e.g. YOLO vs Faster-RCNN. As of the 2020s, single stage models have taken the lead and are now the current state-of-the-art with no clear path for two-stage models to catch up. As a student of science all of my life, I have internalized that knowledge is not stagnant. There will soon be better models that can help our mission of saving lives.

As our mission is so critical, I need to be aware of any developments in the industry. Transformers have already been shown to be unbelievably powerful. ChatGTP writes code and songs, does math, and tells jokes. ViT and models

like it attempt to bring that power to image classing. Once optimized for speed and small image recognition improves, these models have the potential to become the new standard as single-stage CNNs have done. If I would like to have the best models possible, I will need learn more about transformers, their limitations and what they excel at, as they develop.

I. Why is this paper important/why does it matter to a non-technical business stakeholder?

   Abstract concepts like transformer-based image classification do not inspire the interest of the common layperson. However, they are important. Many are now questioning the ethics behind an application that can replicate someone's likeness or take an MCAT. These are serious questions that require serious debate. Moreover, the stakes continue to get higher.

This debate, however, cannot occur if people do not know enough to talk intelligently on the subject. Whether willful or involuntary, ignorance of technological developments blindsides the populus and often leads to backlash. Radio, television, the internet itself were all met with angst. These technologies did not just one day appear. They were the products of tedious development and took time to implement. Yet, very few people were aware of them until they appeared in a neighbor's living room. By then, conversations surrounding the ethics were too late; the cat was out of the bag.

The same thing is now happening with ChatGTP. The transformer is impressive but also terrifying. Schools, governments, parents, and more are struggling with the impacts of a technology that seemingly fell from the sky. Now Pandora's box is open and there is not putting the chaos back. The race to fix the mess could have been avoided with regulation based on ethical discussions of experts (would be preferable to have an educated populous).

As I mentioned before, development does not rest. It also doesn't wait for consensus if its right or wrong. Businesses and other non-technical individuals cannot afford to be reactive. A clear understanding of what is coming down the pipeline will allow for proper ethics debates and for preparation to occur prior to its implementation. This is will prevent them from feeling blindsided, and, without this terror, the technology may be implemented to its fullest from its advent.