Paper 5 – An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale

This presentation was about the use of a transformer, the Vision Transformer (ViT), that could demonstrate its capabilities of image recognition. The fascinating aspect of this model is that instead of relying on CNNs, the transformer is directly applied to images to perform classification. This is done by converting the images to sequences and processing them similarly to tokens. Nevertheless, the process is similar to that of CNNs. They both utilize training in an attempt to achieve high accuracies out of their model, however it requires fewer resources to run leading to its superior performance. Additionally, the ViT also has less inductive bias. This is especially important because inductive bias could lead to future inaccurate information. This should be caught early on, as it could create a snowball effect of misinformation.

Although this transformer can perform efficiently regularly, there are some limitations that must be considered when using this model. Firstly, this model relies heavily on training. Due to the fact that the ViT does not using anything complex, such as neural networks in CNNs, its efficiency is inferior to that of CNNs with small datasets. This also affects its ability to handle inductive biases. Since it relies on data to learn, there is a higher possibility to encounter either inductive biases or inaccurate information. Additionally, his model’s abilities are limited to image recognition. In trade for efficiency, the ViT’s capabilities fail to exceed anything further than recognition. This includes inability to evaluate images, object detection, segmentation, etc.

I think that this model is a great sort of “pocket-tool” to use. If I were ever to run across a task where I would need to classify images of some similar nature, I would use it. Due to the fact that this model requires intensive training to use, I think it’s viability is limited to images of the same category. For example, in my project I used potato chip images using CNNs for classification. If I were needing to classify potato chips based on different qualities, such as color, crispiness, saltiness, etc., I would use this model as opposed to binary classification. However, if I had images of multiple categories, I would use CNNs as it would not require extensive training like the ViT would. This is to allow me to save time and resources as well as be able to reuse the model for future use. Additionally, the python library support on image classification allows me to also retrieve more information about my dataset using CNNs, such as loss. In conclusion, I would prefer CNNs over this transformer model.