Identifying Limitations of Vision Transformers in Structured Image Recognition

Shivi Jindal | Lab Mentor: Mariam Avagyan | Faculty Mentor: John Wright
Carnegie Mellon University, Columbia University

Introduction

Vision Transformers (ViTs) have emerged as a breakthrough architecture in the field of computer vision, altering image classification tasks by surpassing the performance of Convolutional Neural Networks (CNNs) on very large datasets. Unlike CNNs, which rely on convolutional layers, ViTs employ self-attention mechanisms to process images as sequences of patches, allowing the model to capture interdependence among patches and images. Despite their accomplishments in image classification tasks, the ViTs that underlie popular image generation models such as Stable Diffusion face challenges in spatial recognition and generating structured images. This issue manifests itself in distorted representations of structure–hands with too many fingers and checkerboards with irregular square patterns. As a result, we ask: Which classes of structured image recognition tasks are particularly challenging for Vision Transformers, and how can we determine this? What factors contribute to their failures on these tasks?

Methods

In order to capture the limitations of the ViT, I built a "baby" Vision Transformer with a multihead self-attention mechanism and Multilayer Perceptron (MLP) block. It is trained on the CIFAR-10 dataset (60,000 32x32 images categorized into 10 classes) on two Nvidia RTX A5000 GPUs. After performing hyperparameter tuning on variables such as attention heads, patch size, etc., I generated smaller datasets (ranging from 250 to 1000 images in size) to determine classes of tasks that are particularly difficult for ViTs. Using these smaller datasets, I finetuned the classifier head of the model and tested its ability to 1) count separated objects and 2) count connected components in the image foreground.

Results

ViT Pretraining

After experimenting with various model hyperparameters, the following hyperparameters were used: Dropout = 0.1, Embed_dims = 252, hidden_dims = 504, num_heads = 12, num_layers = 6, patch_size = 4, num_patches = 64, learning_rate = 0.003, max_epochs = 200. This yielded a 77.6% validation accuracy and 77.3% testing accuracy on the CIFAR-10 dataset.

ViT Finetuning

Task	Dataset Size	Epochs	Learning Rate	Train Accuracy	Test Accuracy
Dot Counting	250	10	0.001	70.4%	55.0%
Dot Counting	500	20	0.001	74.0%	65.0%
Dot Counting	1000	20	0.001	80.3%	71.3%
Connected Objects	250	20	0.001	32.5%	18.4%
Connected Objects	250	20	0.002	34.0%	16.5%

Conclusion

The ViT's attention mechanism computes pairwise inner products and does not store information about the number of objects in an image. As a result, I find that ViT has moderate difficulty with counting separate objects, but this can be somewhat overcome and learned by training on larger datasets. The model has very high difficulty with counting the number of connected components in the image foreground, which is not easily overcome with larger datasets or hyperparameter tuning.

References

1. Ballal, A. (n.d.). Akshay's personal website. https://www.akshaymakes.com/blogs/vision-transformer (Accessed July 26, 2023)
2. Dosovitskiy A, et al. ICLR. 2021.
3. Tutorial 15: Vision Transformers — UvA DL Notebooks v1.2 documentation. https://uvadlc-notebooks.readthedocs.io/en/latest/tutorial_notebooks/tutorial15/Vision_Transformer.html (Accessed July 26, 2023)
4. Vaswani A, et al. NeurIPS. 2017; 30: 5998–6008.

Acknowledgements

I would like to thank Dr. John Wright, Mariam Avagyan, the rest of the Wright Lab, and the Amazon-Columbia SURE program for their support throughout my research project.