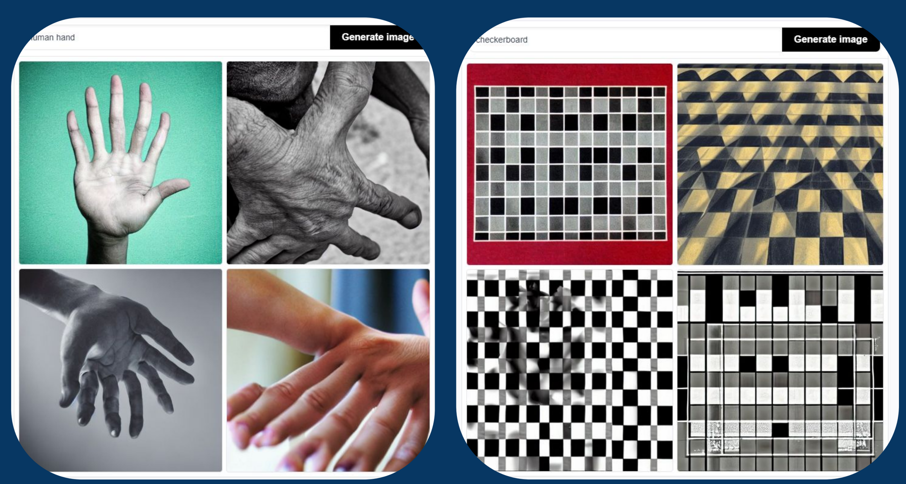
Identifying Limitations of Vision Transformers in Structured Image Recognition

By: Shivi Jindal | Lab Mentor: Mariam Avagyan | Faculty Mentor: John Wright

Department of Electrical Engineering, Columbia University

Introduction

- **Vision Transformers (ViTs)** are a relatively new architecture that surpass the performance of convolution-based architectures on image classification tasks with very large datasets.
- **Problem**: ViTs that underlie generative image models like Stable Diffusion face challenges in spatial recognition and producing structured images.



Distorted Hands and Checkerboards Produced by Stable Diffusion

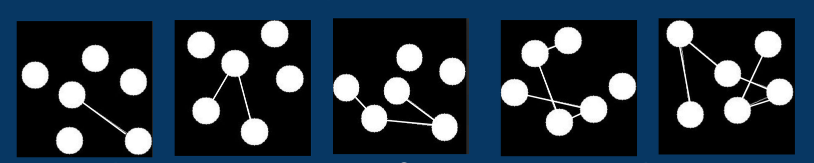
Question: Which classes of structured image recognition tasks are challenging for ViTs, and how can we determine these classes? What contributes to their failures on these tasks?

Method

1. **Built a Vision Transformer with the following components:**
   - Image Embedding: Divide images into patches and append texture and position encodings as learnable parameters onto flattened patch vector.
   - Multihead Self-Attention: Each attention head computes relationships between inputs patches via pairwise inner products and concatenates resulting vectors into a matrix.
   - Multilayer Perceptron Block (MLP): Two fully connected layers and GELU activation function
   - Classification: Extract class token value from tensor
2. **Pretrained ViT on CIFAR-10 image dataset**
   - 60,000 32x32 images divided evenly into 10 classes
   - Trained on two Nvidia RTX A5000 GPUs
3. **Created codebase to generate datasets for fine-tuning**

3 Datasets of 250 images, 500 images, 1000 images

Sample Data Images for Dot-Counting Task

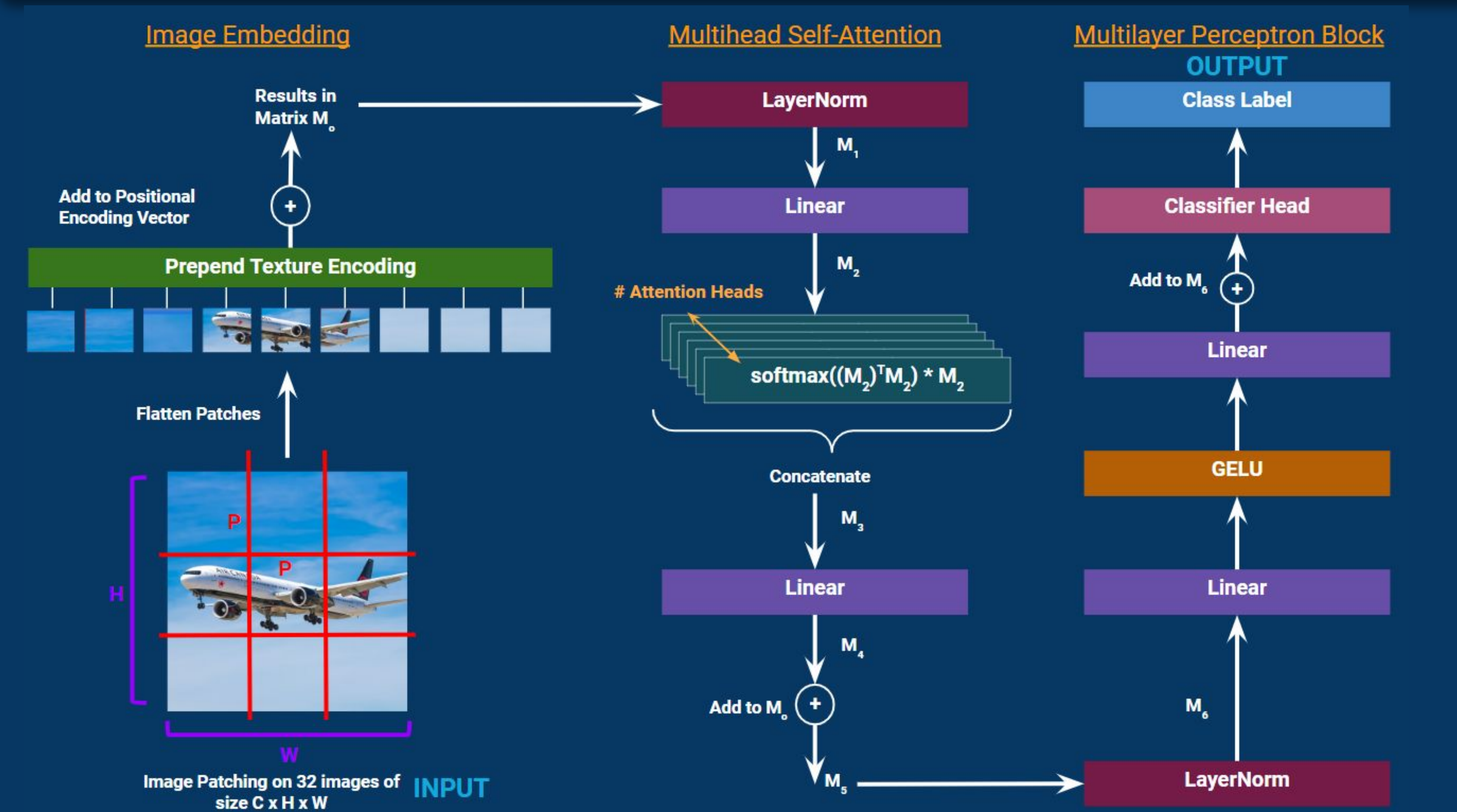


Dataset of 250 images

Sample Data Images for Connected Component Counting Task

4. **Transfer learning: Finetune ViT for testing on smaller tasks such as those created in Step 3**

Model Architecture

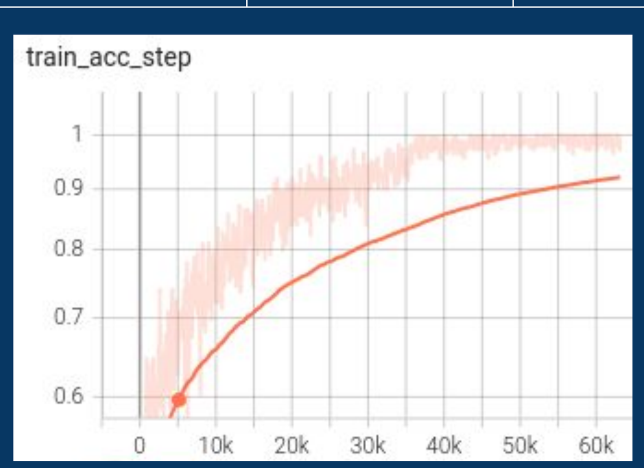


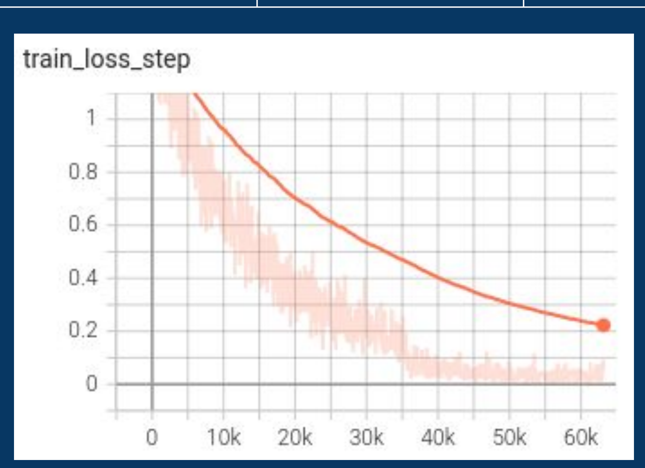
Experiments & Results

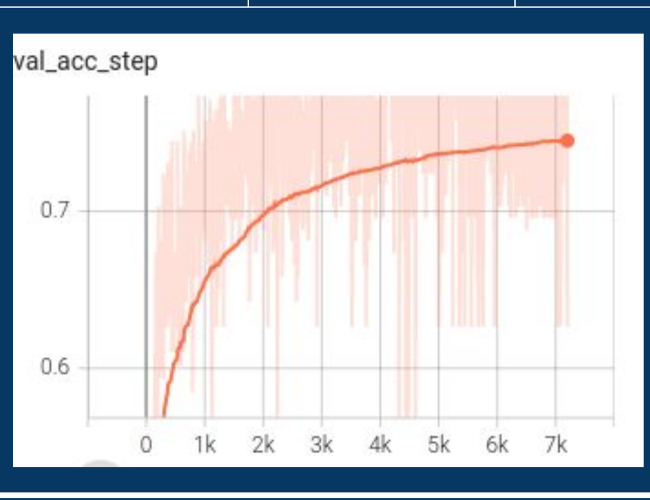
Model Pretraining Hyperparameters and Performance on CIFAR-10

Validation Accuracy: 77.6% Testing Accuracy: 77.3%

Dropout	Embed_dims	Hidden_dims	Num_heads	Num_layers	Patch Size	Num_patch	Max_epochs	Learning Rate
0.1	252	504	12	6	4	64	200	0.0003







Fine-tuning Experiments

Task	Dataset Size	Num_epochs	Learning Rate	Training 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 Components	250	20	0.001	32.5%	18.4%
Connected Components	250	20	0.002	34.0%	16.5%

- ViT has moderate difficulty with counting separate objects
  - Can somewhat be overcome and learned by training on larger datasets
- ViT has very high difficulty with counting number of connected components in image foreground
  - Attention is computed by pairwise inner products and does not utilize information about the number of objects in an image

Future Goals

- Generate datasets to test the model's ability to perform other classes of tasks, such as:
  - Pattern detection
  - 2D/3D Object Detection
- Rigorously investigate lower bounds on hyperparameters such as number of attention heads, layers, etc. needed for model's ability to perform above tasks
- Study the mathematical underpinnings of model's failures on such tasks to introduce modifications that will help it perform better
- Develop a decoder block to accompany the current encoder (ViT) for image generation

References

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

Acknowledgements

I would like to thank Dr. John Wright, Mariam Avagyan, and the rest of the Wright Lab for their assistance and support throughout my project. I would also like to thank the Amazon-Columbia SURE program for providing me with the opportunity to conduct research this summer.