Model scaling with switch transformer methodology for transformer-based image classification models

Team:

Raj Ghodasara - rg4357
Anand Vishwakarma - asv8775

Proposed Objective

- Advances in transformer based CV models have surpassed CNNs in Top-1 Accuracy
- Problem: Computationally expensive
- Scale transformers to increase performance, without a significant increase in computational requirements

Summary - MoE and Switch Transformer

Task - Use different parameters for each incoming sample

Predecessor - Mixture of Experts(MoE)

- Pass tokens to different multiple experts with high probability.
- Cons - complexity, communication costs, and training instability

Switch Transformer

- Pass token through only 1 expert with highest probability (simplified MOE routing algorithm)
- Aims to reduce the complexity by introducing sparsity and adaptivity in the attention mechanism.

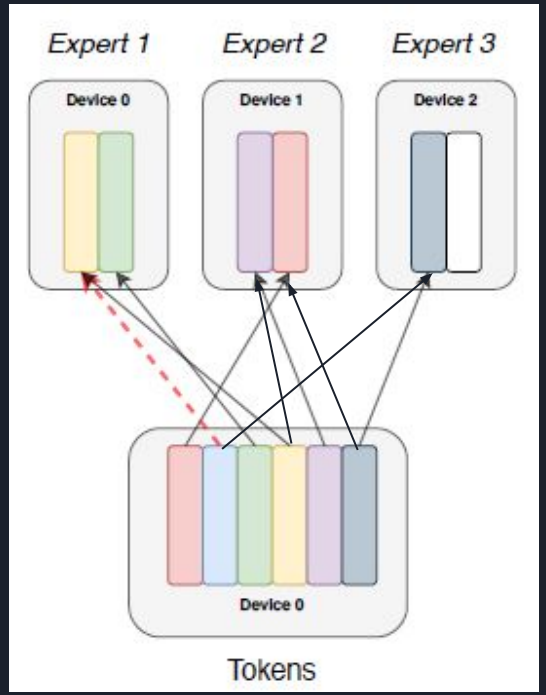
Summary - Vision Transformer (ViT)

- Image classification model that replaces convolutional layers with a transformer encoder, which was originally designed for natural language processing.
- The transformer encoder processes the image as a sequence of patches, where each patch is treated as a token and fed into the transformer encoder.
- ViT also includes a classification head that takes the final hidden state of the transformer encoder and maps it to a prediction over the output classes.
- One key advantage of ViT is that it can handle input images of arbitrary size without requiring additional modifications to the model architecture.

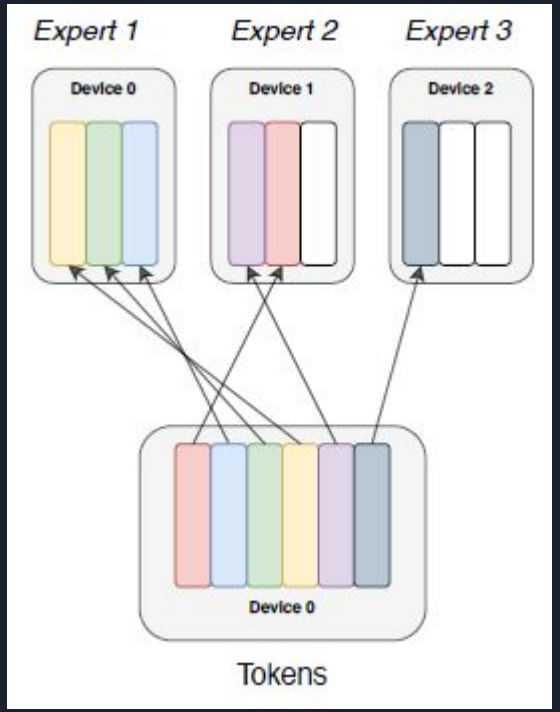
Drawback:

- Requires significant computational resources and is more computationally expensive than traditional convolutional neural networks.

MoE vs Switch

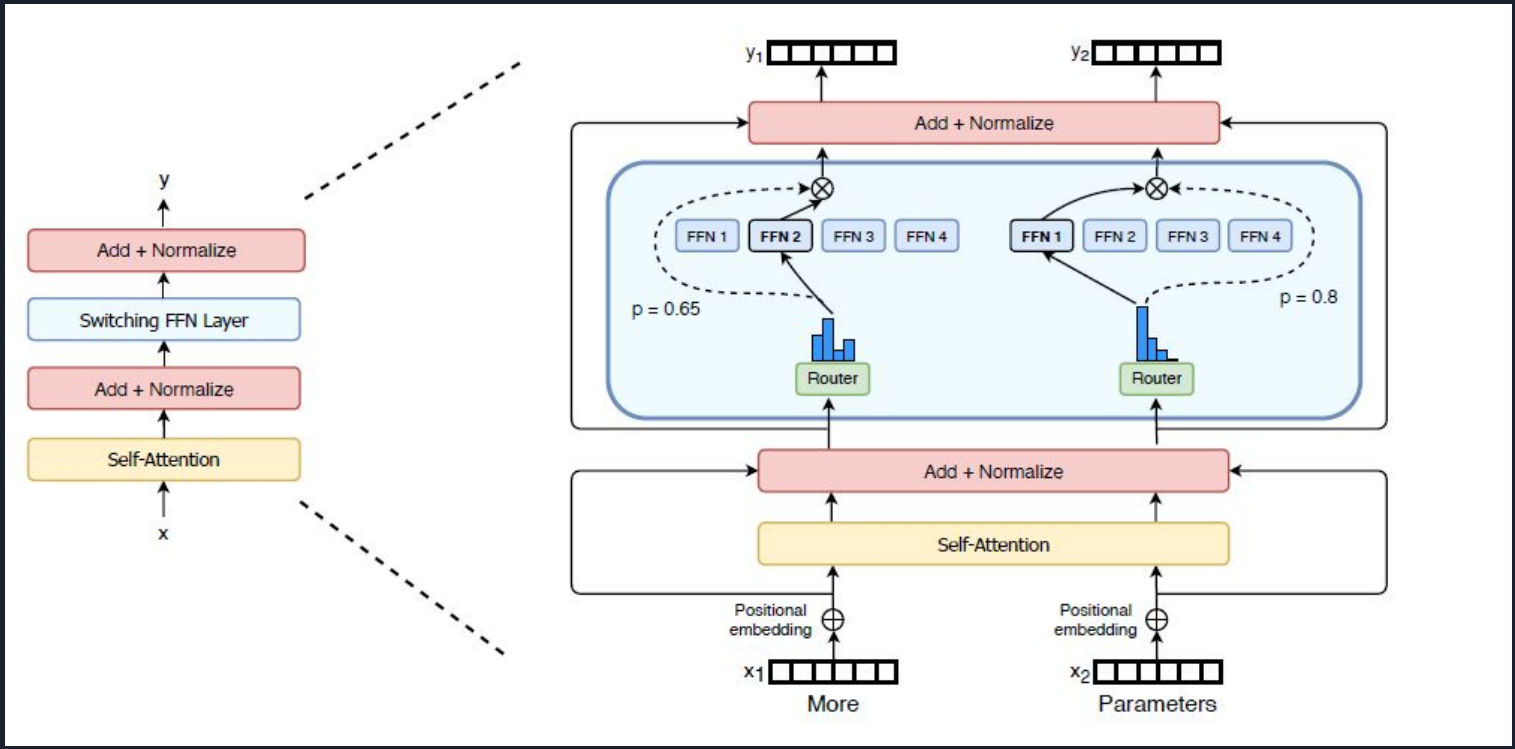


MoE



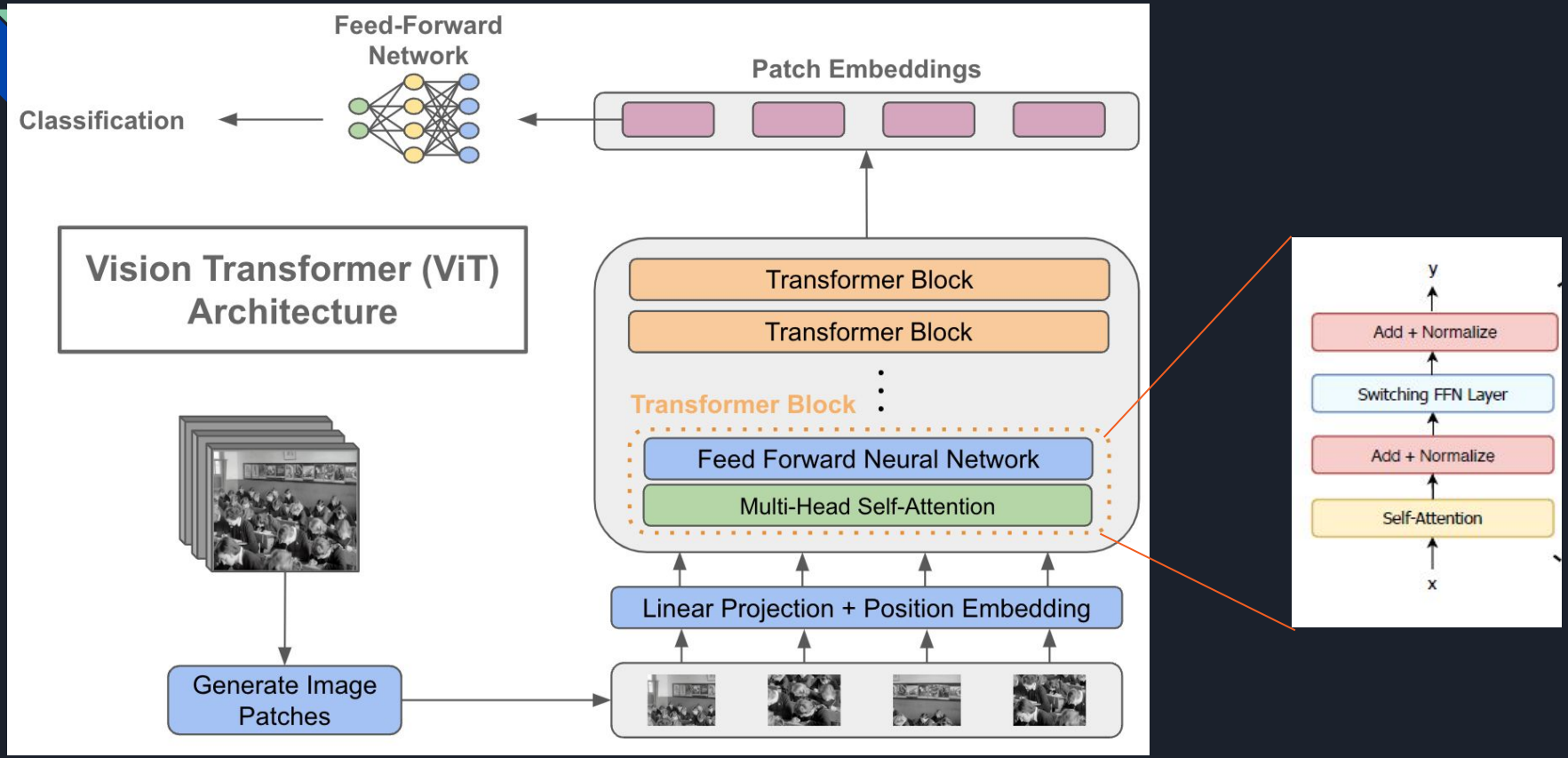
Switch

Switch routing



Switch in ViT

- Changing the Transformer Layer of Vision transformer with Switch Transformer layer
- New Switch ViT routing of tokens
  - Patch embeddings and positional embeddings are same
  - Self-attention of multi-head attention (mha) layer of switch vit is also kept same
  - Feed-forward of mha layer is **replaced** with Switch-FeedForward layer.
  - Classification layer is kept same as original ViT.
- Switch-FeedForward layer
  - Routing layer to calculate best expert (Feed-forward) for every token.
  - Route the individual tokens to their respective expert.
  - Scale the output according to the probability from routing layer
  - Forward it to the next layer.

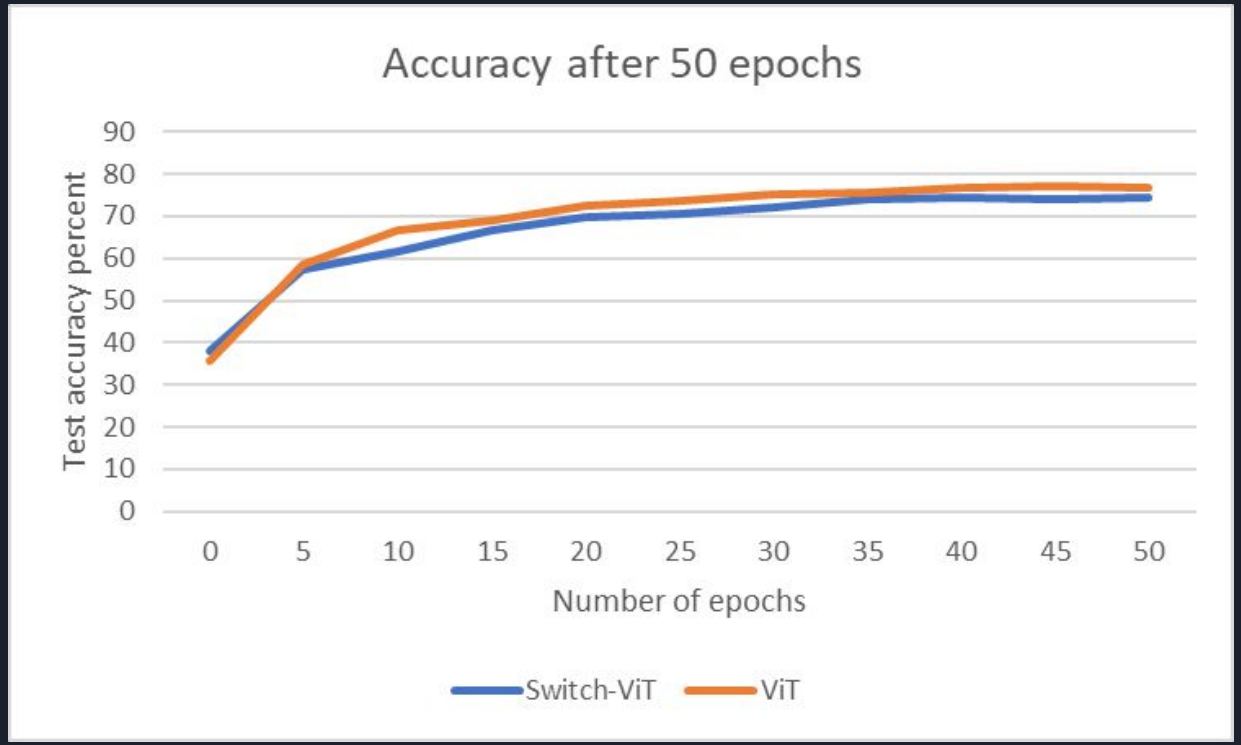

Feed-Forward Network
Classification
Patch Embeddings
Vision Transformer (ViT) Architecture
Transformer Block
Transformer Block
Transformer Block
Feed Forward Neural Network
Multi-Head Self-Attention
Linear Projection + Position Embedding
Generate Image Patches
y
Add + Normalize
Switching FFN Layer
Add + Normalize
Self-Attention
x

Experimental results:
Training and inference of ViT and Switch-ViT

Training on 4 RTX8000 for 350 epochs with 256 effective batch size

Model	Parameters	Scale	Test Accuracy	Training Time (60,000 images)	Inference Time (10,000 images)
ViT	12,798,490	1	81.36	11.8 secs	4.6 secs
Switch-ViT (32 experts)	238,732,426	18.6	79.04	68.86 secs	4.98 secs
Switch-ViT (64 experts)	464,669,962	36.3	78.8	130.52 secs	5.18 secs

Test Accuracy comparison over 50 epochs



Details of experiments

Experiments were conducted with the following variations:

- Number of experts: 12, 32, 64
- Size of patches: 2, 4 (on 32 x 32 CIFAR 10 Data)
- Effective Batch size: 64, 128, 256
- No. of transformer layers kept same across all experiments: 12

Technical Challenges

Data:

- Due to the scale of Switch-ViT network, we need a large amount of data to notice significant jumps in performance (even regular ViT requires dataset of 14M+ to work well)

Algorithmic (implemented):

- We employed Distributed Data Parallel (DDP) framework for parallel training on 2 and 4 GPUs.
- Pipeline parallelism using GPipe.

Algorithmic (not implemented):

- Switch Transformer used selective precision training for better stability in performance.
- Expert parallelism over multiple threads.

System issues:

- Limited hardware resources, leading to training on a smaller dataset (CIFAR10)
- Limited with hardware availability to perform multiple experiments.

Conclusions/Observations

- Hypothesis of increasing performance not proven; however, we have multiple future improvements as to how this can be improved:
  - Size of dataset:
    - ViT paper mentioned that training on a considerably larger dataset such as ImageNet-1K, then fine tuning it on smaller datasets showed better results.
    - With Switch-ViT we increased the number of experts (Feed-Forward network) in the transformer layer, requiring much larger number of samples to generalize compared to ViT.
  - Selective Precision Training: Switch Transformer had better results and performance when trained with float16 (half precision), only keeping the routing algorithm in float32(single precision).
  - Currently, according to switch architecture different patches of image sample (tokens) is routed through different experts, we hypothesize that selecting experts based on whole image patch (all tokens of images) rather than selecting based on individual tokens will give better results.
  - Expert parallelism will help improve training time significantly.

References

1. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity) (https://arxiv.org/pdf/2101.03961.pdf)
2. Residual Mixture of Experts (https://arxiv.org/pdf/2204.09636v1.pdf)
3. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale (https://arxiv.org/pdf/2010.11929.pdf)

Github Link :

https://github.com/anand-vishwakarma-1/switch-vit