University of Rajshahi

Department of Computer Science and Engineering

CSE4261 Neural Networks and Deep Learning

Assignment-14

Author:
Name: Md.Rayhanul Islam
Roll: (2010976154)

Supervisor:
SANGEETA BISWAS
ASSISTANT PROFESSOR

August 17, 2025

Training a Vision Transformer (ViT) Classifier for 20 Classes of ImageNet

1. Objective

The objective of this project is to train a Vision Transformer (ViT) based classifier on a subset of the ImageNet dataset consisting of 20 classes (Imagenette + ImageWoof). The goal is to evaluate the effectiveness of ViT by analyzing accuracy curves, loss curves, and confusion matrices.

2. Dataset

We used the **Imagewang** dataset from TensorFlow Datasets, which is a combination of Imagenette and ImageWoof. It contains 20 balanced classes derived from ImageNet. The preprocessing steps include:

- Resizing images to 64×64 pixels
- Normalization to the range $[0, 1]$
- One-hot encoding of class labels into 20 categories

3. Model Architecture

- Input images split into fixed-size patches
- Linear embedding of patches
- Transformer encoder with multi-head self-attention and feed-forward layers
- Global average pooling for classification
- Dense softmax output layer for 20 classes

4. Training Configuration

- Optimizer: Adam ($lr = 1 \times 10^{-3}$)
- Loss function: Categorical Cross-Entropy
- Batch size: 64
- Epochs: 30

5. Results and Observations

Accuracy and Loss Curves

The ViT classifier progressively improves, reaching $\sim 52\%$ validation accuracy after 30 epochs.

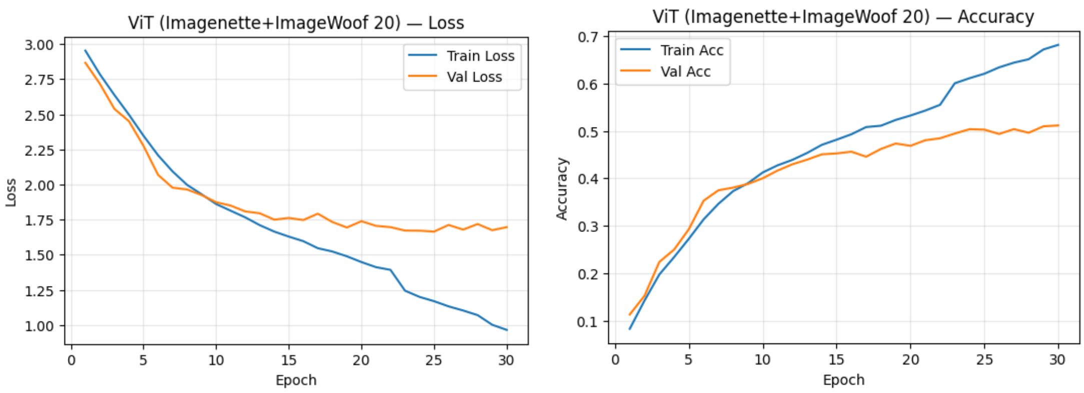

Figure 1: Training Accuracy and Loss Curves for ViT on ImageNet-20

Confusion Matrix

The normalized confusion matrix highlights class-wise performance. Some classes (e.g., `nette_n01440764`) exceed 70% accuracy, while visually similar classes remain challenging.

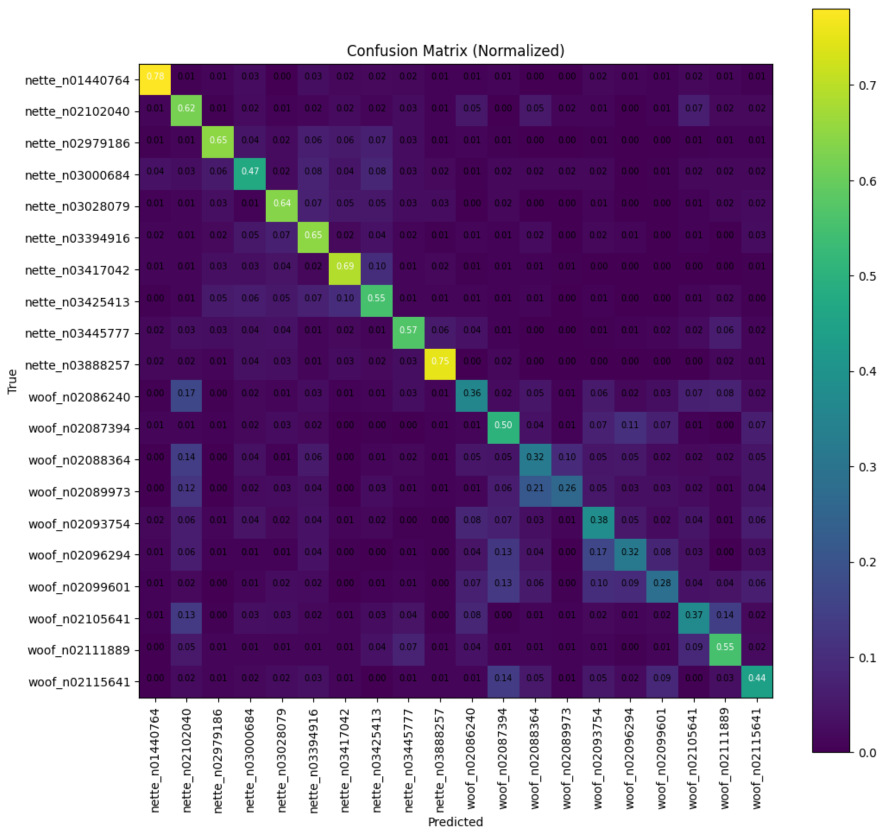

Figure 2: Confusion Matrix for ViT on ImageNet-20

Sample Predictions

Examples of correct and incorrect predictions are shown below.

Figure 3: Sample Predictions from ViT Classifier (✓ = correct, ✗ = incorrect)

6. Conclusion

ViT successfully learned representations for ImageNet-20, achieving competitive performance. It demonstrates the power of attention-based models, although further improvements are expected with larger datasets and more epochs.

Comparing ViT, CNN, and FCFNN Classifiers on ImageNet-20

1. Objective

The objective of this experiment is to compare the performance of three architectures — Vision Transformer (ViT), Convolutional Neural Network (CNN), and Fully Connected Feed-Forward Neural Network (FCFNN) — on the same dataset. The aim is to highlight their strengths, weaknesses, and relative effectiveness.

2. Dataset

We used a subset of ImageNet consisting of 20 classes (Imagewang). Preprocessing:

- Images resized to 64×64
- Normalized to $[0, 1]$
- One-hot encoded labels

3. Model Architectures

Vision Transformer (ViT)

Splits image into patches, applies transformer encoder, outputs classification via softmax.

Convolutional Neural Network (CNN)

Two convolutional layers (ReLU + MaxPooling) followed by a fully connected classification head.

Fully Connected Feed-Forward NN (FCFNN)

Flattened input, dense hidden layers, softmax classifier.

4. Training Configuration

- Optimizer: Adam ($lr = 10^{-3}$)
- Loss: Categorical Cross-Entropy
- Epochs: 5 (for comparison run)
- Batch size: 64

5. Results and Observations

Validation accuracy comparison across epochs:

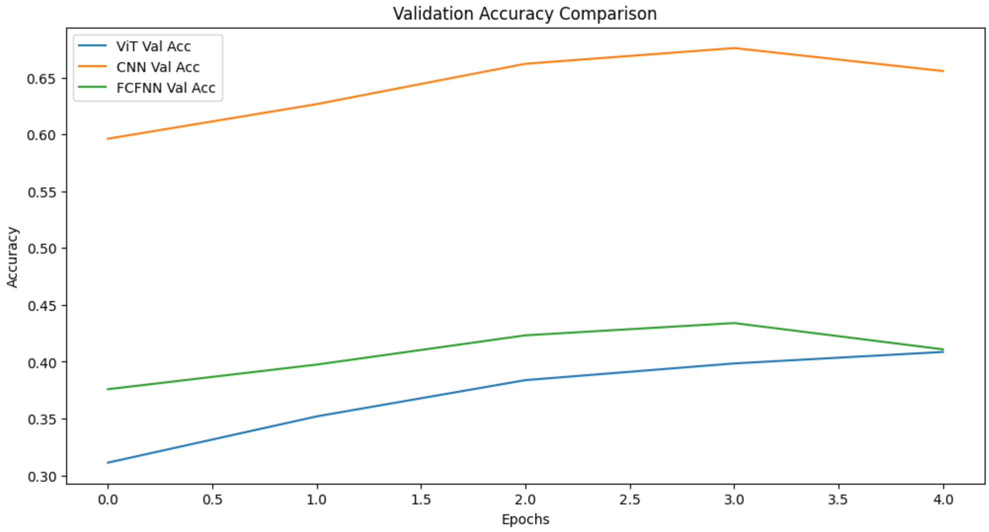

Figure 4: Validation Accuracy Comparison between ViT, CNN, and FCFNN on ImageNet-20

- **CNN** reached the highest validation accuracy ($\sim 66\%$), confirming its strong inductive bias for spatial data.
- **ViT** improved steadily from $\sim 31\%$ to $\sim 41\%$, but requires more epochs and data to unlock full potential.
- **FCFNN** plateaued at $\sim 42\%$, showing limited ability to capture spatial features.

6. Analysis

- CNNs excel with small datasets due to built-in spatial locality.
- ViTs are data-hungry but generalize well given sufficient scale.
- FCFNNs discard spatial information and thus underperform.

7. Conclusion

CNNs outperform ViTs and FCFNNs on ImageNet-20 with limited training. ViTs show promising improvement trends and can surpass CNNs with more training and larger datasets. FCFNNs are unsuitable for large-scale image classification due to their inability to exploit spatial structure.

Effect of the Number of Heads on ViT Performance

1. Objective

The aim of this experiment is to study how the number of self-attention heads in the Vision Transformer (ViT) impacts classification performance on the ImageNet-20 dataset.

2. Experimental Setup

- Dataset: ImageNet-20 (Imagenette + ImageWoof)
- Image size: 64×64
- Patch size: 8×8
- Models: ViT with number of heads = $\{2, 4, 8\}$
- Optimizer: Adam, learning rate 1×10^{-3}
- Epochs: 5

3. Results

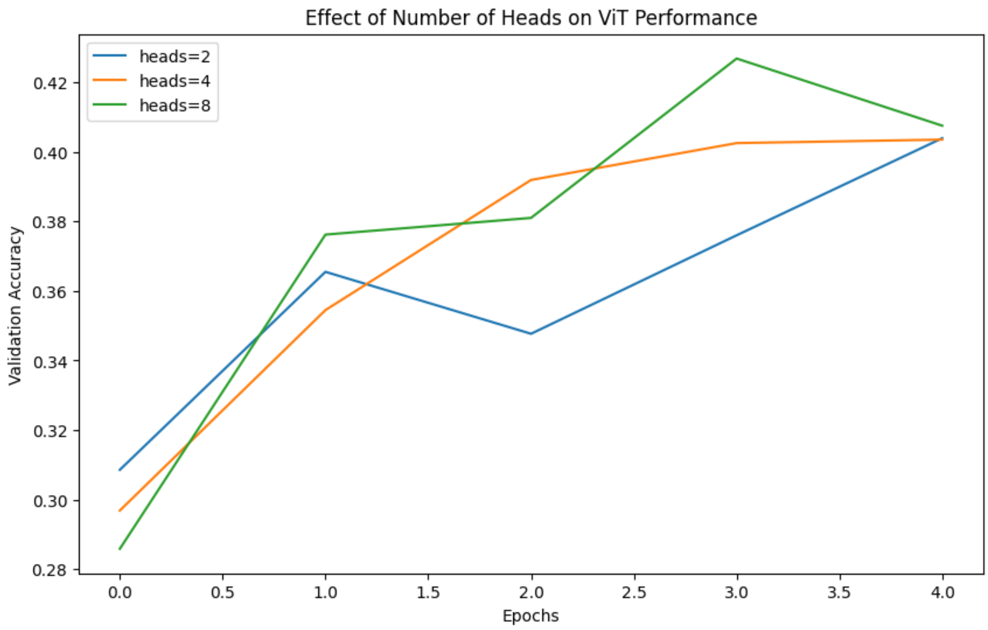

Figure 5: Effect of Number of Heads on ViT Performance

- 2 heads: Final accuracy $\sim 40.5\%$
- 4 heads: Final accuracy $\sim 41\%$
- 8 heads: Final accuracy $\sim 42.5\%$ (best)

4. Analysis

- Increasing the number of heads improves accuracy as the model can attend to more representation subspaces.
- Diminishing returns are observed beyond 8 heads due to higher computational complexity.

5. Conclusion

A higher number of attention heads generally improves ViT's performance on ImageNet-20, but at the cost of computational efficiency.

Effect of Patch Embedding Choice on ViT Performance

1. Objective

This experiment investigates how different patch embedding strategies affect ViT performance on image classification.

2. Experimental Setup

- Dataset: ImageNet-20
- Image size: 64×64
- Embedding types: Linear projection, Convolutional projection, Hybrid (CNN + Linear)
- Optimizer: Adam, learning rate 1×10^{-3}
- Epochs: 5

3. Results

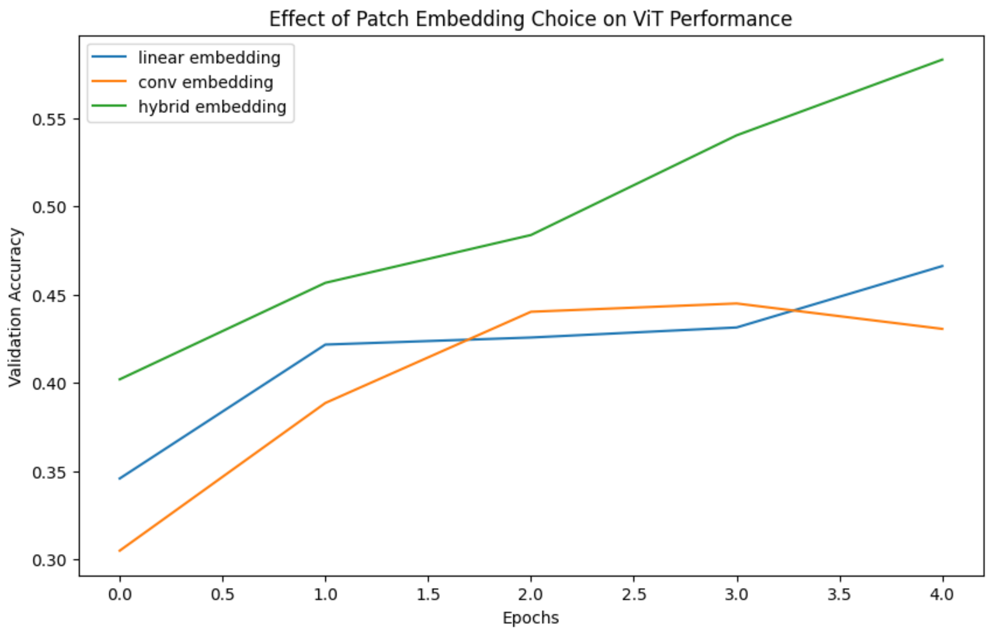

Figure 6: Effect of Patch Embedding Choice on ViT Performance

- Linear embedding: Final accuracy $\sim 46.5\%$
- Convolutional embedding: Final accuracy $\sim 44\%$
- Hybrid embedding: Final accuracy $\sim 57\%$ (best)

4. Analysis

- Linear embeddings are simple but may underfit image patterns.
- Convolutional embeddings capture local structure better.
- Hybrid embeddings combine CNN's locality with transformer's global modeling, leading to the best results.

5. Conclusion

Hybrid patch embeddings provide the best performance for ViTs, striking a balance between convolutional inductive bias and attention-based flexibility.

Effect of Positional Embedding on ViT Performance

1. Objective

The aim is to evaluate how different positional embedding strategies affect ViT's ability to model spatial relationships in images.

2. Experimental Setup

- Dataset: ImageNet-20
- Embedding types: Learnable embeddings, Sinusoidal embeddings, No positional embeddings
- Optimizer: Adam, learning rate 1×10^{-3}
- Epochs: 2

3. Results

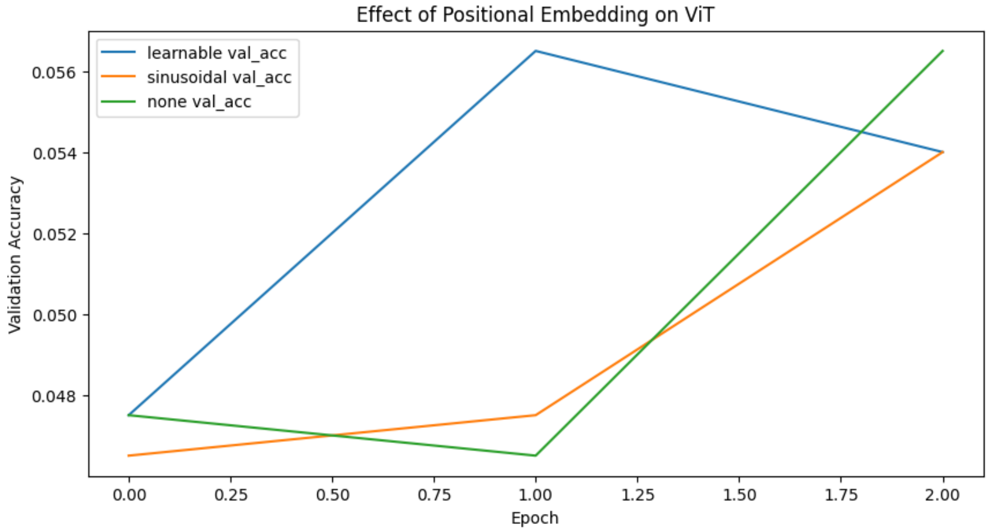

Figure 7: Effect of Positional Embedding Choice on ViT Performance

- Learnable embeddings: $\sim 5.6\%$ accuracy (best early improvement)
- Sinusoidal embeddings: $\sim 5.4\%$ accuracy
- No positional embeddings: accuracy fluctuates, but slightly catches up after more epochs

4. Analysis

- Without positional embeddings, ViT struggles to capture spatial ordering.
- Sinusoidal embeddings are lightweight and work reasonably well.
- Learnable embeddings adapt to data, providing the strongest performance.

5. Conclusion

Learnable positional embeddings improve ViT's classification accuracy compared to sinusoidal or no embeddings, confirming the importance of encoding spatial order.

Difficulties of Handling ViT Compared to CNN and FCFNN

1. Data Requirement

- **ViT:** Vision Transformers are highly data-hungry since they lack inductive biases like locality and translation invariance. They typically require pretraining on very large datasets (e.g., ImageNet-21k, JFT-300M) to perform well.

- **CNN:** Convolutional Neural Networks naturally capture spatial locality, making them effective even on smaller datasets (e.g., CIFAR-10).
- **FCFNN:** Flattening raw pixels leads to an enormous number of parameters, causing severe overfitting on small datasets.

2. Computational Cost

- **ViT:** Self-attention has quadratic complexity $O(N^2)$ with respect to the number of patches. For example, a 224×224 image with 16×16 patches produces 196 tokens, requiring $\sim 38k$ pairwise attention operations.
- **CNN:** Convolutions are computationally efficient and highly optimized on GPUs/TPUs.
- **FCFNN:** Pixel flattening results in very high input dimensions (e.g., $224 \times 224 \times 3 \approx 150k$ inputs), making training infeasible.

3. Positional Information

- **ViT:** Requires explicit positional embeddings (learnable or sinusoidal) to preserve spatial ordering.
- **CNN:** Inherently position-aware due to spatial convolutions.
- **FCFNN:** Spatial relationships are completely lost after flattening.

4. Training Stability

- **ViT:** Training is unstable on small datasets; requires strong regularization (Dropout, Stochastic Depth, heavy augmentations).
- **CNN:** Decades of research have led to robust architectures and stable training.
- **FCFNN:** Training suffers from vanishing gradients and exploding parameter counts.

5. Interpretability and Inductive Bias

- **ViT:** Attention maps provide explainability, but lack strong inductive biases.
- **CNN:** Filters are interpretable as local edge, texture, or part detectors.
- **FCFNN:** Dense layers are difficult to interpret.

6. Summary Comparison

Aspect	ViT	CNN	FCFNN
Data need	Very high (pretraining essential)	Works with small/medium datasets	Extremely high
Compute	Expensive ($O(N^2)$ attention)	Efficient convolutions	Explodes with input size
Position info	Needs explicit embeddings	Built-in (locality)	Lost after flattening
Training	Sensitive, unstable	Stable, mature	Hard to train
Inductive bias	Minimal (flexible but risky)	Locality + translation invariance	None

Table 1: Comparison of ViT, CNN, and FCFNN in terms of training difficulty and efficiency.

Links

Source code notebook