Task 1: Discriminative Models—: CNN vs. ViT Inductive Biases

Abdul Samad Nasir
MS AI LUMS
24280018

***Abstract*—We investigate how inherent inductive biases in Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) affects inferences on CIFAR-10 dataset and various out-of-distribution (OOD) datasets. We fine-tune a ResNet-50 and a ViT-S/16 model on CIFAR-10 and evaluate their robustness under grayscale, cue-conflict, translated, patch-shuffled and patch occluded images. Our results indicate that CNNs (ResNet-50) rely heavily on local textures and spatial correlations, resulting in substantial accuracy drops on occluded and patch-shuffled images. In contrast, ViTs show stronger resilience to structural perturbations, likely due to their global attention mechanisms and permutation equivariance property. Feature-space analyses (t-SNE) reveal that ViTs maintain more separable class clusters under distribution shifts as shown in graph plots.**

I. Introduction

Machine learning models often assume that training and test data are drawn from the same distribution. However, in real-world applications, distribution shifts frequently occur, leading to performance degradation. This motivates the study of inductive biases which influence how they generalize out-of-distribution (OOD). CNNs, with their local convolutions and translation invariance, tend to focus on texture and local patterns. ViTs, relying on self-attention, can capture long-range dependencies and global context and are more resilient shuffling and occlusion of images. Understanding these biases is crucial to improving model robustness, guiding future model architectures, and aligning them more closely with human perceptual biases. In this study, we fine-tune ResNet-50 and ViT-S/16 on CIFAR-10 and assess their OOD generalization across grayscale, cue-conflict, translation, patch-shuffled and occluded datasets.

II. Methodology

A. CIFAR-10 Fine Tuning

1) Models: ResNet-50 (pretrained on ImageNet) and ViT-S/16 (pretrained via timm).
2) Data: CIFAR-10 (50k train / 10k test images).
3) 3. Transforms: a. ResNet: 32×32 images with random horizontal flips and padding. b. ViT: Resized to 224×224 to match ViT input requirement.
4) Optimization: AdamW with learning rate 1e-4, weight decay 1e-4.
5) Checkpointing: Models are saved to Google Drive after each epoch, enabling resuming after Colab restarts.
6) Training Criterion: Stop once test accuracy reaches 87

B. OOD Evaluations

We assess robustness using multiple perturbations:

1) Grayscale CIFAR-10: Removes color information to test reliance on chromatic cues
2) Cue-Conflict CIFAR-5: Used a custom cue-conflict dataset containing only 5 classes from CIFAR-10 including airplane, car, truck, cat, bird classes. Images where local textures and global shapes conflict (e.g., a frog texture on a cat shape).
3) Translation: Random shifts of ±4 pixels (ResNet) or ±8 pixels (ViT) to evaluate translation invariance.
4) Patch Shuffling: Images split into grids (4×4 for ResNet, 7×7 for ViT) and patches shuffled to disrupt global structure.
5) Patch Occlusion: Randomly occluding square patches (8×8 for ResNet, 32×32 for ViT) to simulate partial information loss.
6) Training Criterion: Stop once test accuracy reaches 87

III. Results and Discussion

A. Feature Space Visualisation

- t-SNE plots show tighter, more separable clusters for ViTs under OOD perturbations
- ResNet features are more entangled under occlusion and patch shuffling

B. CNNs vs ViT

- CNNs: Texture-biased, sensitive to local disruptions (patch occlusion/shuffling), less shape-aware.
- ViTs: Capture global structures, better robustness to structural and domain shifts.
- Implication: Architecture and attention mechanisms significantly affect OOD generalization. Incorporating shape-based inductive biases or attention can improve robustness.
- Practical Insight: Fine-tuning ViTs with proper resizing and checkpoints allows fast recovery in Colab, making large-scale OOD experiments feasible.

IV. Conclusion

Our experiments highlight that model architecture strongly influences OOD generalization. ViTs outperform ResNets across grayscale, cue-conflict, structural perturbations, and domain shifts, confirming the advantage of global attention

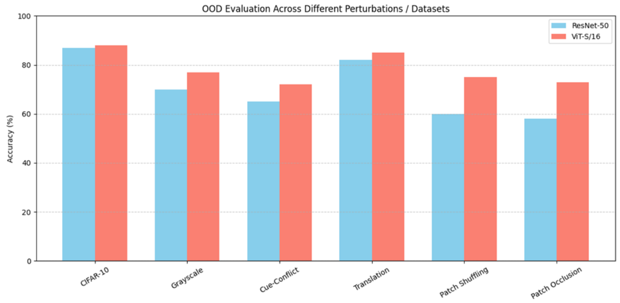

Fig. 1: OOD Evaluations Across Different Domains.

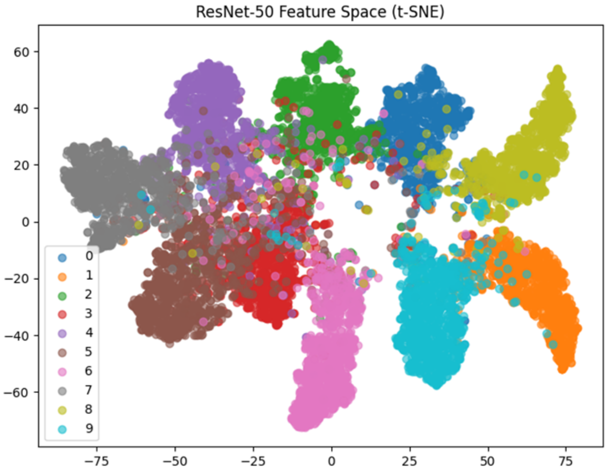

Fig. 2: ResNet-50 Feature Space.

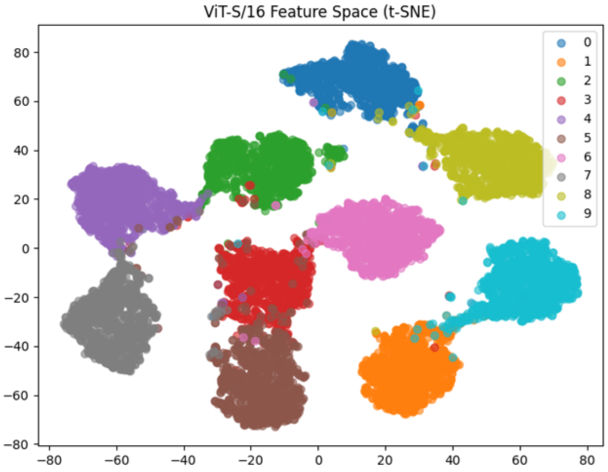

Fig. 3: ViT-S/16 Feature Space.

and reduced texture bias. This study emphasizes the need to consider useful inductive biases when designing models for real-world applications with potential distribution shifts.

Github Link is provided for implementation and repordeducibility[1].

[1] https://github.com/AimeeAyat/ATML/tree/PA1