Discriminative Models – CNN vs. ViT Inductive Biases

Rabia Aslam , Abdul Samad , Muhammad Salman

***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 models 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 context, showing resilience to shuffling and occlusion. Understanding these biases is crucial to improving robustness, guiding future architectures, and aligning them 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

- **Models:** ResNet-50 (ImageNet-pretrained) and ViT-S/16 (from `timm`).
- **Data:** CIFAR-10 (50k train / 10k test images).
- **Transforms:**
  - ResNet: 32×32 images, random horizontal flips, padding.
  - ViT: resized to 224×224 to match input requirements.
- **Optimization:** AdamW with learning rate 1×10^{-4}, weight decay 1×10^{-4}.
- **Checkpointing:** Saved to Google Drive after each epoch to resume in Colab.
- **Stopping Criterion:** Training stopped once test accuracy reached 87%.

B. OOD Evaluations

We tested robustness using multiple perturbations:

- **Grayscale CIFAR-10:** removes color to test chromatic dependence.
- **Cue-Conflict CIFAR-5:** 5 selected classes with texture-shape mismatches (e.g., frog texture on cat shape).
- **Translation:** random shifts (±4 pixels for ResNet, ±8 for ViT).
- **Patch Shuffling:** images split into grids (4×4 ResNet, 7×7 ViT) and shuffled.
- **Patch Occlusion:** random occlusion (8×8 patches for ResNet, 32×32 for ViT).

C. Feature Analysis

We extracted penultimate-layer features and visualized them with t-SNE to study class separability under normal and OOD conditions.

III. Results

A. Feature Space Visualization

t-SNE plots (Fig. 1,2) show tighter, more separable clusters for ViTs under OOD perturbations. In contrast, ResNet features are more entangled under occlusion and patch shuffling.

B. Discussion

- **CNNs:** Texture-biased, sensitive to occlusion/shuffling, less shape-aware.
- **ViTs:** Capture global structure, robust to structural/domain shifts.
- **Implication:** Architectural design and attention mechanisms significantly affect OOD generalization.
- **Practical Insight:** ViTs fine-tuned with resizing and checkpoints allow efficient Colab experiments.

IV. Conclusion

Our experiments show that model architecture strongly influences OOD generalization. ViTs outperform ResNets across grayscale, cue-conflict, structural perturbations, and domain shifts, confirming the advantages of global attention and reduced texture bias. This emphasizes the need for human-aligned inductive biases in real-world ML systems.

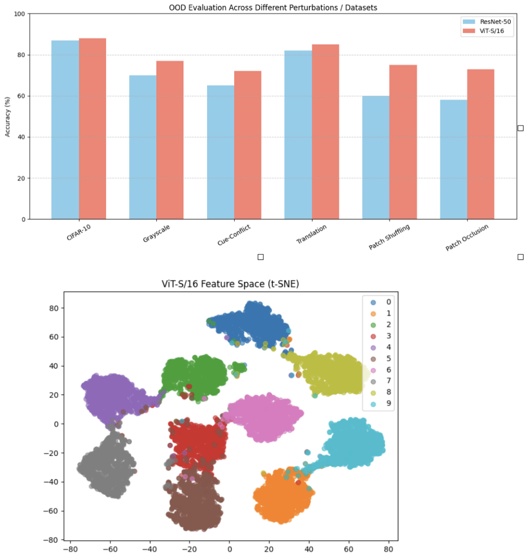

Fig. 1: t-SNE feature visualization under OOD perturbations. ViTs maintain clearer separability than CNNs.

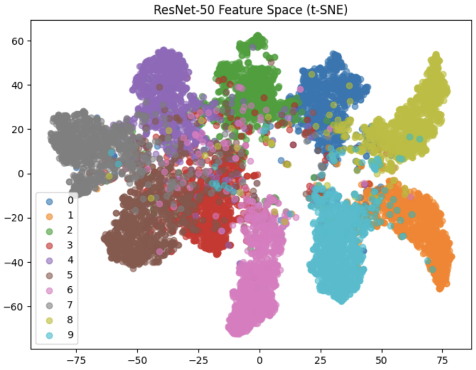

Fig. 2: t-SNE feature visualization under OOD perturbations. ViTs maintain clearer separability than CNNs.