Shortcut Learning Susceptibility in Vision Classifiers(Student Abstract)

Pirzada Suhail, Vrinda Goel, Amit Sethi

IIT Bombay
psuhail@iitb.ac.in

Abstract

Shortcut learning, where machine learning models exploit spurious correlations in data instead of capturing meaningful features, poses a significant challenge to building generalizable models. Vision classifiers based on Convolutional Neural Networks (CNNs), Multi-Layer Perceptrons (MLPs), and Vision Transformers (ViTs) leverage distinct architectural principles to process spatial and structural information, making them differently susceptible to shortcut learning. In this study, we systematically evaluate these architectures by introducing deliberate shortcuts into the dataset that are correlated with class labels both positionally and via intensity, creating a controlled setup to assess whether models rely on these artificial cues or learn actual distinguishing features. We perform both quantitative evaluation by training on the shortcut-modified dataset and testing on two different test sets—one containing the same shortcuts and another without them—to determine the extent of reliance on shortcuts. Additionally, qualitative evaluation is performed using network inversion-based reconstruction techniques to analyze what the models internalize in their weights, aiming to reconstruct the training data as perceived by the classifiers. Further, we evaluate susceptibility to shortcut learning across different learning rates. Our analysis reveals that CNNs at lower learning rates tend to be more reserved against entirely picking up shortcut features, while ViTs, particularly those without positional encodings, almost entirely ignore the distinctive image features in the presence of shortcuts.

Introduction

Machine learning models are expected to learn meaningful patterns from data to make accurate predictions and generalize well across different domains. However, in many cases, models do not learn the intended task-relevant features but instead rely on shortcut learning, where they exploit spurious correlations in the training data that happen to be predictive of the labels. While shortcut learning may improve performance on in-distribution test data, it significantly degrades model robustness when evaluated on out-of-distribution samples in a real-world settings where these spurious correlations no longer hold.

In computer vision, different classification architectures process input data in fundamentally distinct ways, which may lead to varying tendencies toward shortcut learning. Convolutional Neural Networks (CNNs) operate through convolutional filters that extract spatial patterns, Multi-Layer Perceptrons (MLPs) learn feature representations in a fully connected manner without spatial biases, and Vision Transformers (ViTs) leverage self-attention mechanisms to model long-range dependencies. Since each of these architectures encodes spatial and structural information differently, their responses to shortcuts may vary.

Shortcut learning has been identified as a fundamental limitation of deep learning models, leading to poor robustness and transferability (Geirhos et al. 2020). The study in (Hermann et al. 2024) examines shortcut learning from a theoretical perspective by investigating the factors that influence whether a model will rely on a shortcut. In (Brown et al. 2023), the authors propose a method to detect shortcut learning in clinical machine learning models by applying multitask learning to identify improper correlations that may cause biased predictions. This phenomenon has also been studied in medical image segmentation, where zero-padding and center-cropping, introduce unintended shortcuts that influence segmentation accuracy(Lin et al. 2024). The paper (Bleeker et al. 2024) investigates shortcut learning in vision-language models and evaluates how contrastive learning-based models tend to latch onto unintended patterns in multi-caption training scenarios. The work in (Ma et al. 2024) explores how ViTs might be particularly prone to shortcut learning due to their reliance on self-attention mechanisms.

Methodology

Our approach to analyzing shortcut learning susceptibility in different vision classifier architectures involves introducing deliberate shortcuts into the dataset including (1) **Positional shortcuts**, where specific pixel regions are modified such that their locations deterministically correlate with the class labels, and (2) **Intensity-based shortcuts**, where the pixel intensity in a certain region of the image is altered.

The models are trained on the shortcut-modified dataset, and their generalization capabilities are evaluated by comparing their performance on two test sets: (1) **one with the same shortcut modifications**, and (2) **another without shortcuts**. A model that generalizes well and does not rely on shortcuts should perform comparably on both

test sets, whereas a model that latches onto shortcuts will show significant degradation in performance on the later. To quantitatively evaluate shortcut susceptibility, we define **Accuracy Difference** as the absolute difference between the model's accuracy on the shortcut test set and the normal test set. Further, qualitative evaluation is performed using **network inversion-based reconstruction** method proposed in (Suhail and Sethi 2024a,b) to understand the internal representations of neural networks and reconstruct the training data as perceived by different architectures, allowing us to systematically compare their susceptibility to shortcuts.

Quantitative Results

We evaluate the shortcut learning susceptibility by introducing Positional shortcuts, where a 4×4 white patch is inserted at different spatial locations for different classes, and Intensity-based shortcuts, where the pixel intensity in a specific region of the image is altered in a way that correlates with the class labels. Subsequently, the trained models are assessed for their reliance on shortcuts by evaluating their performance on both the shortcut-embedded and standard test sets. Higher values for Accuracy Difference indicates greater susceptibility to shortcut learning, and a lower value suggests better generalization across the two test conditions.

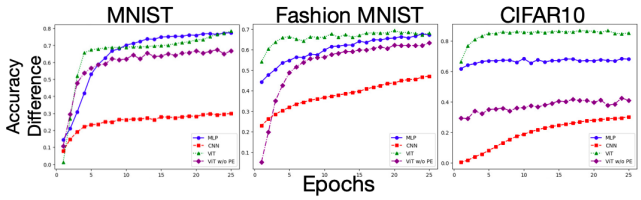

Figure 1: Accuracy Difference for positional shortcuts.

In Figure 1, we observe similar trend across all datasets for positional shortcuts. While accuracy on the shortcut test set is nearly perfect, accuracy on the normal test set is significantly degraded, highlighting the extent to which models exploit shortcut cues. Among the architectures, ViTs with positional encodings exhibit the highest susceptibility to shortcut learning, with the largest values for both Accuracy Difference. This may be attributed to the presence of positional encodings in ViTs, which inherently reveal the spatial locations of the introduced shortcuts, making them easily learnable by the model. However, the results on ViTs without positional encodings are comparably better, with performance much closer to CNNs. On the other hand, CNNs demonstrate the best resistance to shortcuts, showing the smallest differences in accuracy and loss across the two test sets. MLPs display intermediate levels of shortcut susceptibility, performing better than ViTs but worse than CNNs, likely due to their lack of strong spatial priors.

Further, when trained with a large learning rate, models converge rapidly, often at the expense of learning generalizable features. This fast convergence drives models toward optimizing for easily learnable patterns—including shortcuts—rather than extracting meaningful class-discriminative

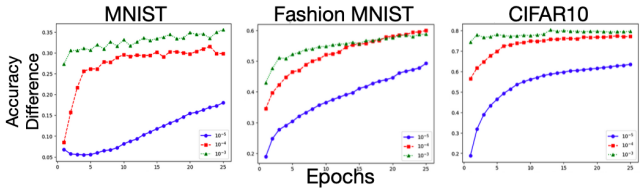

Figure 2: Accuracy Difference across learning rates.

features. As a result, models trained with high learning rates exhibit significantly larger Accuracy Difference as shown in Figure 2 indicating stronger reliance on shortcut cues. In contrast, models trained with a smaller learning rate undergo a more gradual learning process allowing the network to explore a wider range of features before settling into shortcut-based patterns. Sustained exposure to a dataset with shortcuts over prolonged training periods tends to reinforce shortcut reliance making it difficult for the model to recover.

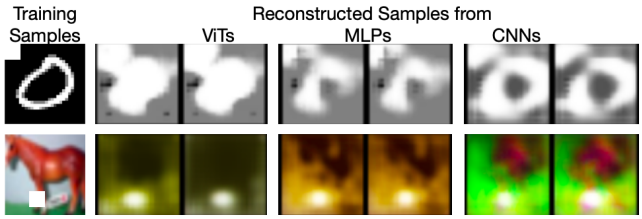

Figure 3: Reconstructed samples from the classifiers.

Qualitative Results

Qualitative assessment of shortcut susceptibility for positional shortcuts is performed by reconstructing the training data as perceived by each classifier using (Suhail and Sethi 2024a) as shown in Figure 3. The first column contains actual training samples with embedded shortcuts, while the subsequent columns display the reconstructed images from ViTs, MLPs, and CNNs, respectively. We observe that CNNs, which exhibited the best generalization performance, also retain some aspects of the actual image features rather than solely relying on shortcuts. In contrast, ViTs display the highest shortcut reliance, completely ignoring the intrinsic image features and almost entirely memorizing the shortcut patterns. MLPs exhibit intermediate levels of shortcut susceptibility, consistent with the trends observed in the Accuracy Difference curves.

Conclusion

In this paper, we evaluated the shortcut learning susceptibility of three different vision classifier architectures across four benchmark datasets using both direct performance comparisons and network inversion-based reconstruction techniques. The results consistently demonstrated that ViTs with positional encodings exhibited the highest reliance on shortcut features, whereas CNNs showed the strongest resistance, with MLPs displaying intermediate behavior.

References

Bleeker, M.; Hendriksen, M.; Yates, A.; and de Rijke, M. 2024. Demonstrating and Reducing Shortcuts in Vision-Language Representation Learning. *Transactions on Machine Learning Research*.

Brown, A.; Tomasev, N.; Freyberg, J.; Liu, Y.; Karthikesalingam, A.; and Schrouff, J. 2023. Detecting shortcut learning for fair medical AI using shortcut testing. *Nature Communications*, 14(1).

Geirhos, R.; Jacobsen, J.-H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; and Wichmann, F. A. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673.

Hermann, K.; Mobahi, H.; FEL, T.; and Mozer, M. C. 2024. On the Foundations of Shortcut Learning. In *The Twelfth International Conference on Learning Representations*.

Lin, M.; Weng, N.; Mikolaj, K.; Bashir, Z.; Svendsen, M. B. S.; Tolsgaard, M. G.; Christensen, A. N.; and Feragen, A. 2024. Shortcut Learning in Medical Image Segmentation. In Linguraru, M. G.; Dou, Q.; Feragen, A.; Giannarou, S.; Glocker, B.; Lekadir, K.; and Schnabel, J. A., eds., *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, 623–633. Cham: Springer Nature Switzerland. ISBN 978-3-031-72111-3.

Ma, C.; Zhao, L.; Chen, Y.; Guo, L.; Zhang, T.; Hu, X.; Shen, D.; Jiang, X.; and Liu, T. 2024. Rectify ViT Shortcut Learning by Visual Saliency. *IEEE Transactions on Neural Networks and Learning Systems*, 35(12): 18013–18025.

Suhail, P.; and Sethi, A. 2024a. Network Inversion for Training-Like Data Reconstruction. In *Neurips Safe Generative AI Workshop 2024*.

Suhail, P.; and Sethi, A. 2024b. Network Inversion of Convolutional Neural Nets. In *Muslims in ML Workshop co-located with NeurIPS 2024*.