Towards Holistic Vision in Deep Neural Networks: Disentangling Local and Global Processing

Fenil R. Doshi (fenil_doshi@fas.harvard.edu), Thomas Fel, Talia Konkle, George A Alvarez

Kempner institute and Department of Psychology, Harvard University

VISION LAB

Introduction

Humans can perceive objects by their global shape, despite local variations

Changes in Texture

Changes in Context

DNNs are emerging as de-facto models of human perception, but are known to be **biased towards local information,** leading to an algorithmic gap between humans and models.

Why is this important?

Human Decision : Elephant

Model Decision : Tiger

This bias is linked to the problem of shortcut learning spurious correlations, making models brittle as compared to humans

Baker et al., 2018; Geirhos et al., 2018; Shah et al., 2020; Hermann et al., 2023

Challenge

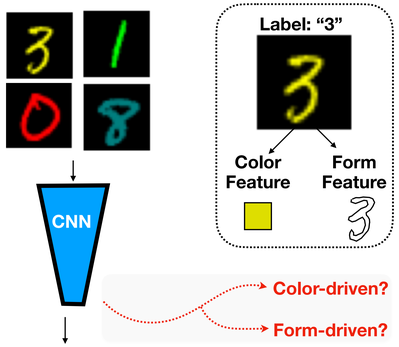
How can we get a model to see global shape? What exactly is global shape?

Solution

Diverge from a local solution and examine what emerges? Global Shape?

Experimental Setup

Digit Classification in Colored MNIST

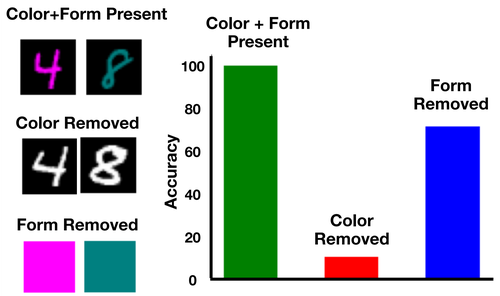


Digit Classification

CNNs are color-driven when both color and form are present in training

Feature Reliance

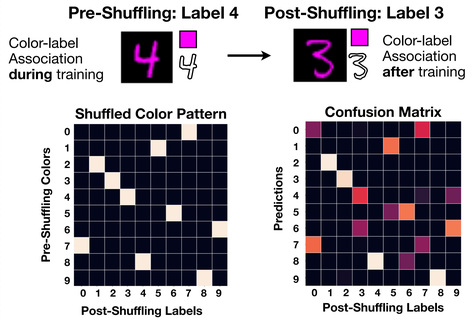
Selectively removing color or form post-training



Model relied on color features for digit classification

Error Patterns

Breaking the color-label association systematically, while retaining the form-label association

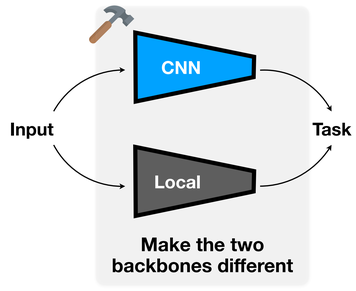


Misclassifications are in accordance to color-label associations seen during training

How do we get a CNN to become form-driven?

Idea

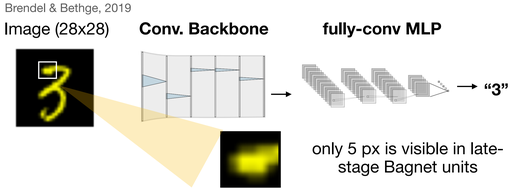
Force a CNN to be different than a model using a purely local strategy to solve the same task



Make the two backbones different

Local Strategy : BagNets

Brendel & Bethge, 2019

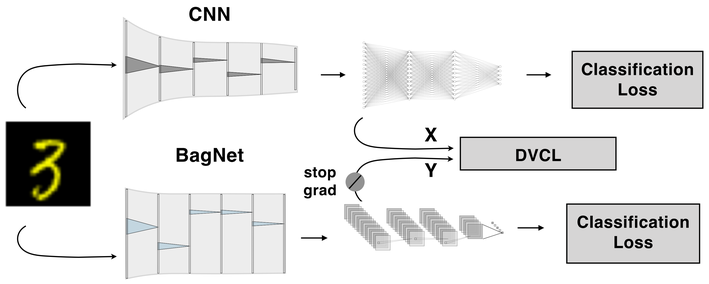


only 5 px is visible in late-stage Bagnet units

BagNets can only take a local-strategy to solve a task

DVCL - Divergence Variance Covariance Loss

DVCL encourages CNN and BagNet to learn orthogonal intermediate features



DVCL = α(Divergence Loss) + β(Variance Loss) + γ(Covariance Loss)

1) Divergence Loss: to get **different** features

$$L_{\text{divergence}} = \frac{1}{d}\sum_{i,j} \text{Corr}(X,Y)^2_{i,j}$$

2) Variance Loss: to prevent **collapse**

$$L_{\text{variance}} = \frac{1}{2d}\sum_{i=1}^{d}\left(\max(0,1-\sigma_{X_i})\right)^2$$

3) Covariance Loss: sparse features

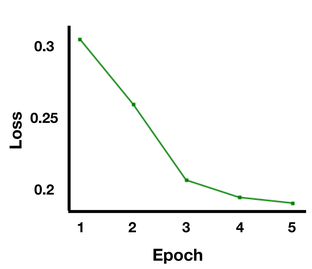
$$L_{\text{cov}} = \frac{1}{d}\sum_{i\neq j}\left(\text{Corr}(X)_{i,j}\right)^2$$

* **stop-grad** in BagNet allows it to "**soak up**" all the possible local features that it can

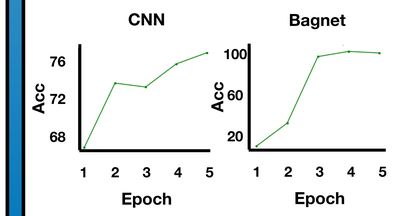
Disentangling color-vs-form processing using DVCL

Training

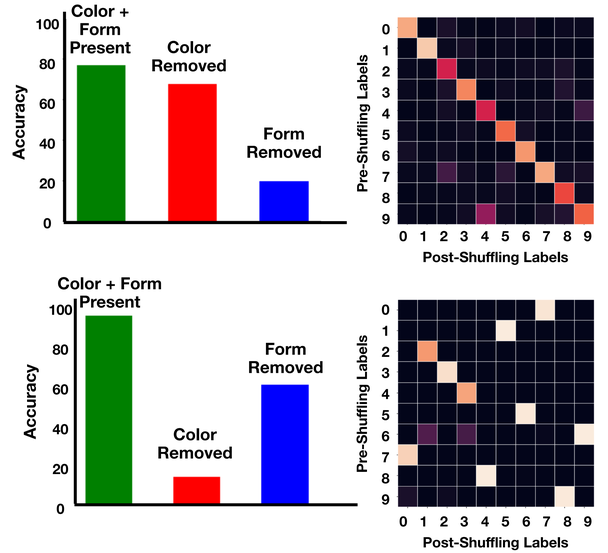
DVCL reduces with training



Both models get trained



Feature Reliance and Error Patterns for CNN and BagNet



Conclusion

- By making **a CNN orthogonal with Bagnet (local model)**, we can encourage it to do **global processing**
- Can this strategy be scaled on naturalistic datasets to develop models with **better shape representations?**

Reference

- Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive development*, *3*(3), 299-321.
- Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018, November). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*.
- Shah, H., Tamuly, K., Raghunathan, A., Jain, P., & Netrapalli, P. (2020). The pitfalls of simplicity bias in neural networks. *Advances in Neural Information Processing Systems*, *33*, 9573-9585.
- Hermann, K. L., Mobahi, H., Fel, T., & Mozer, M. C. (2023). On the foundations of shortcut learning. *arXiv preprint arXiv:2310.16228*.
- Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS computational biology*, *14*(12), e1006613.
- Brendel, W., & Bethge, M. (2019). Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*.