On the importance of single directions for generalization

Ari S. Morcos, David G.T. Barrett, Neil C. Rabinowitz, and Matthew Botvinick

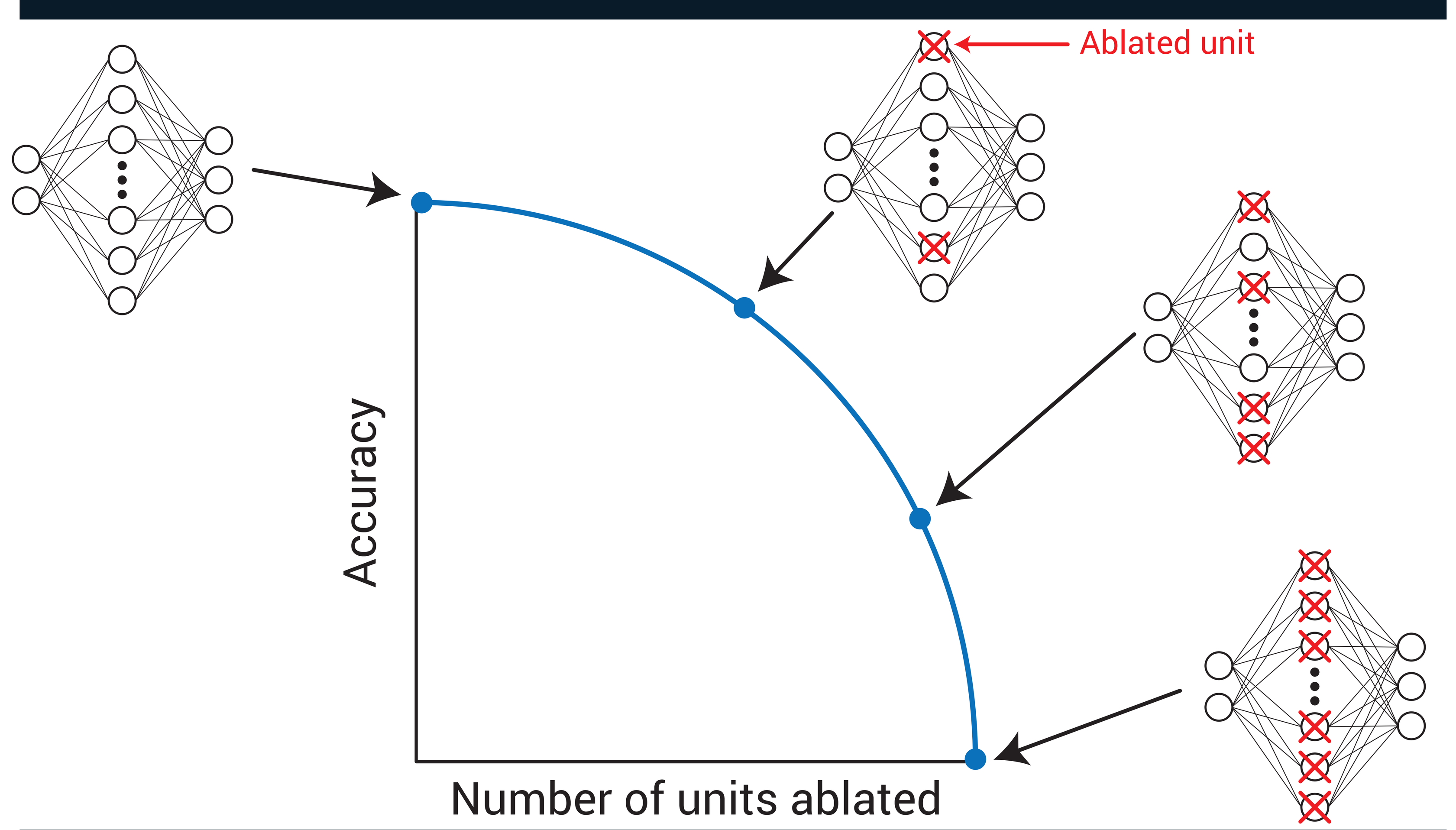
DeepMind

Introduction

Recent work has demonstrated that deep neural networks (DNNs) are capable of memorizing extremely large datasets such as ImageNet (Zhang et al., 2017). Despite this capability, DNNs in practice achieve low generalization error on tasks ranging from image classification (He et al., 2015) to language translation (Wu et al., 2016). These observations raise a key question: why do some networks generalize while others do not?

Here, we demonstrate that a network's reliance on single directions in activation space is a good predictor of its generalization performance, across networks trained on datasets with different fractions of corrupted labels, across ensembles of networks trained on datasets with unmodified labels, and over the course of training. While dropout only regularizes this quantity up to a point, batch normalization implicitly discourages single direction reliance, in part by decreasing the class selectivity of individual units. Finally, we find that class selectivity is a poor predictor of task importance, suggesting not only that networks which generalize well minimize their dependence on individual units by reducing their selectivity, but also that individually selective units may not be necessary for strong network performance.

Experimental design



Memorizing networks are more sensitive to cumulative ablations

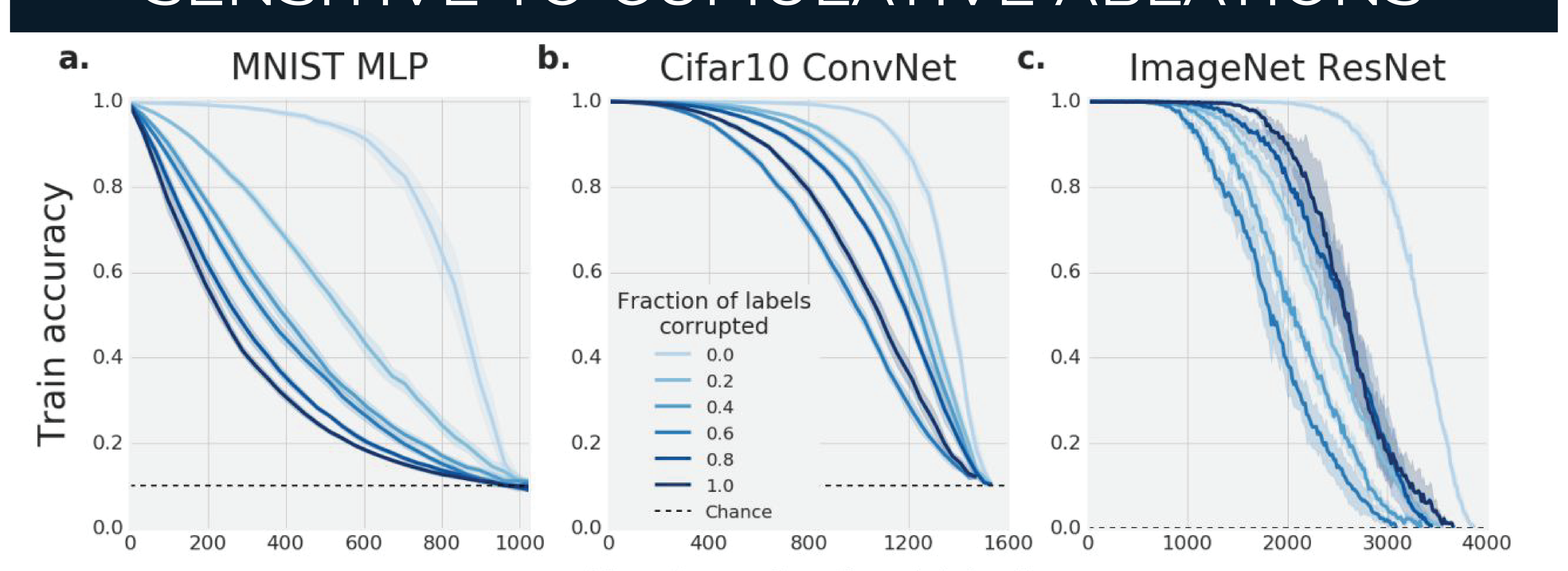


Figure 2: Networks were trained on MNIST (2-hidden layer MLP, **a**), CIFAR-10 (11-layer convolutional network, **b**), and ImageNet (50-layer ResNet, **c**) with various fractions of corrupted labels. In a, all units in all layers were ablated, while in **b** and **c**, only feature maps in the last three layers were ablated. Error bars represent standard deviation across 10 random orderings of units to ablate.

Networks which generalize poorly are more reliant on single directions

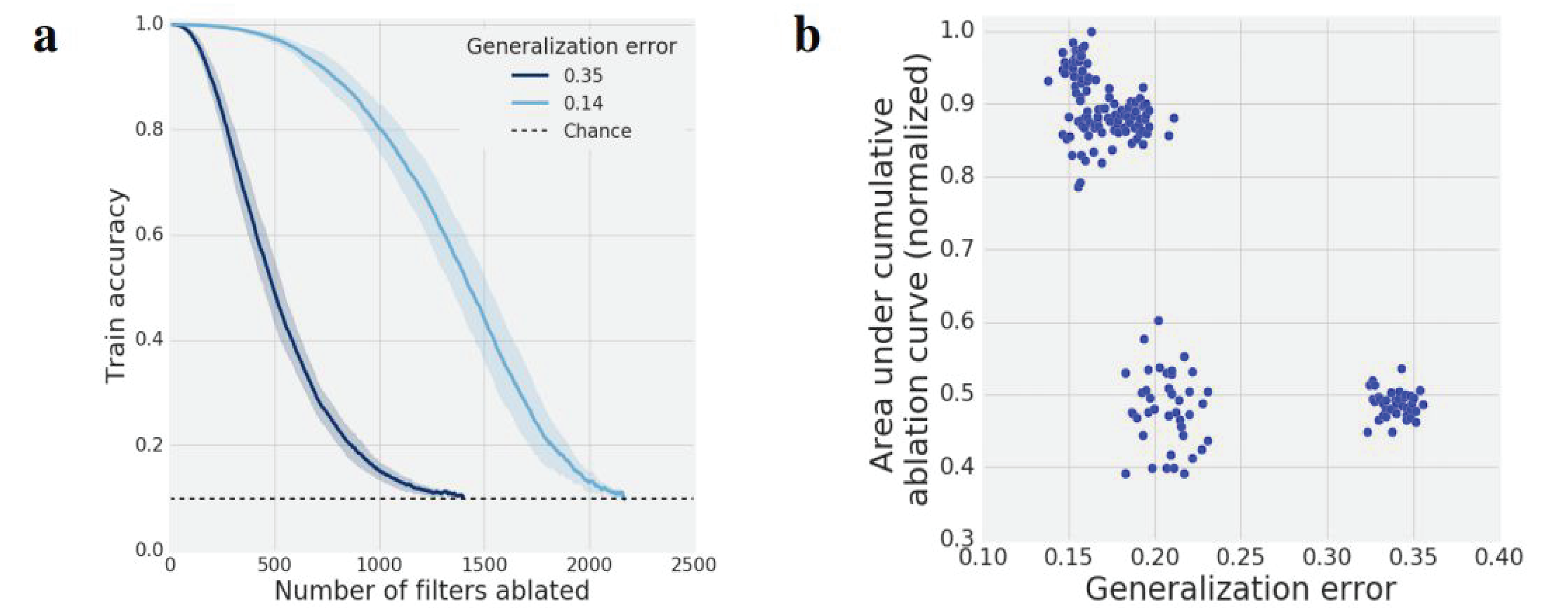


Figure 3: 200 networks with identical topology were trained on unmodified CIFAR-10. **a,** Cumulative ablation curves for the best and worst 5 networks by generalization error. Error bars represent standard deviation across 5 models and 10 random orderings of feature maps per model. **b,** Area under cumulative ablation curve (normalized) as a function of generalization

Single direction reliance as a signal for hyperparameter selection and early stopping

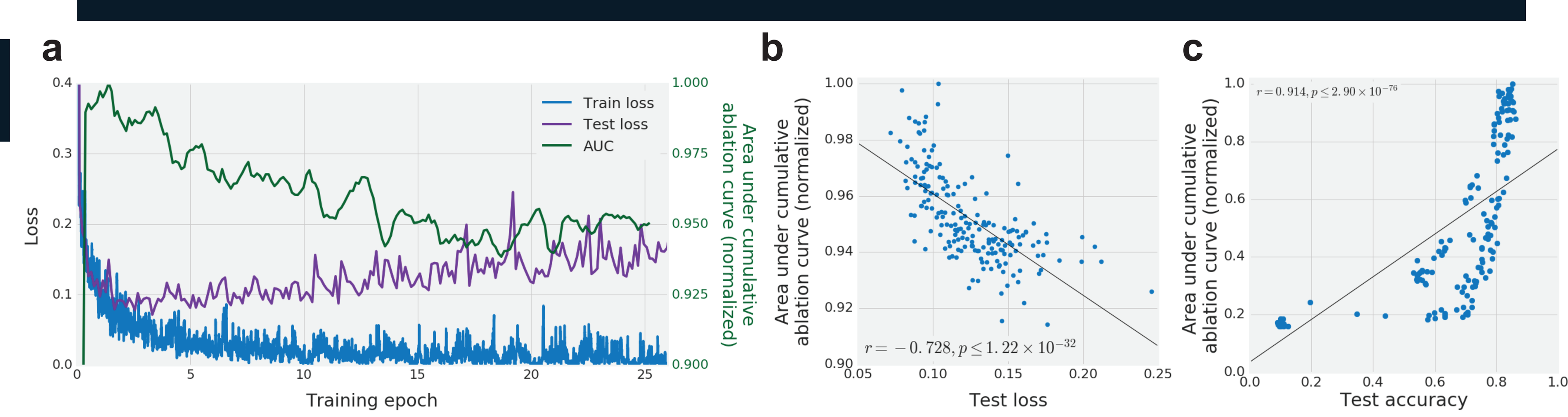


Figure 4: a, Train (blue) and test (purple) loss, along with the normalized area under the cumulative ablation curve (AUC; green) over the course of training for an MNIST MLP. Loss y-axis has been cropped to make train/test divergence visible. **b,** AUC and test loss are negatively correlated over training. **c,** AUC and test accuracy are positively correlated across a hyperparameter sweep of CIFAR-10 models (96 hyperparameters with 2 repeats each). AUC selected the top 1, 5, and 10 settings 13%, 83%, and 98% of the time, repsspectively with an average difference between the best model selected by AUC and the optimal model of only 1 ± 1.1% (mean ± std).

Impact of regularizers on networks' reliance on single directions

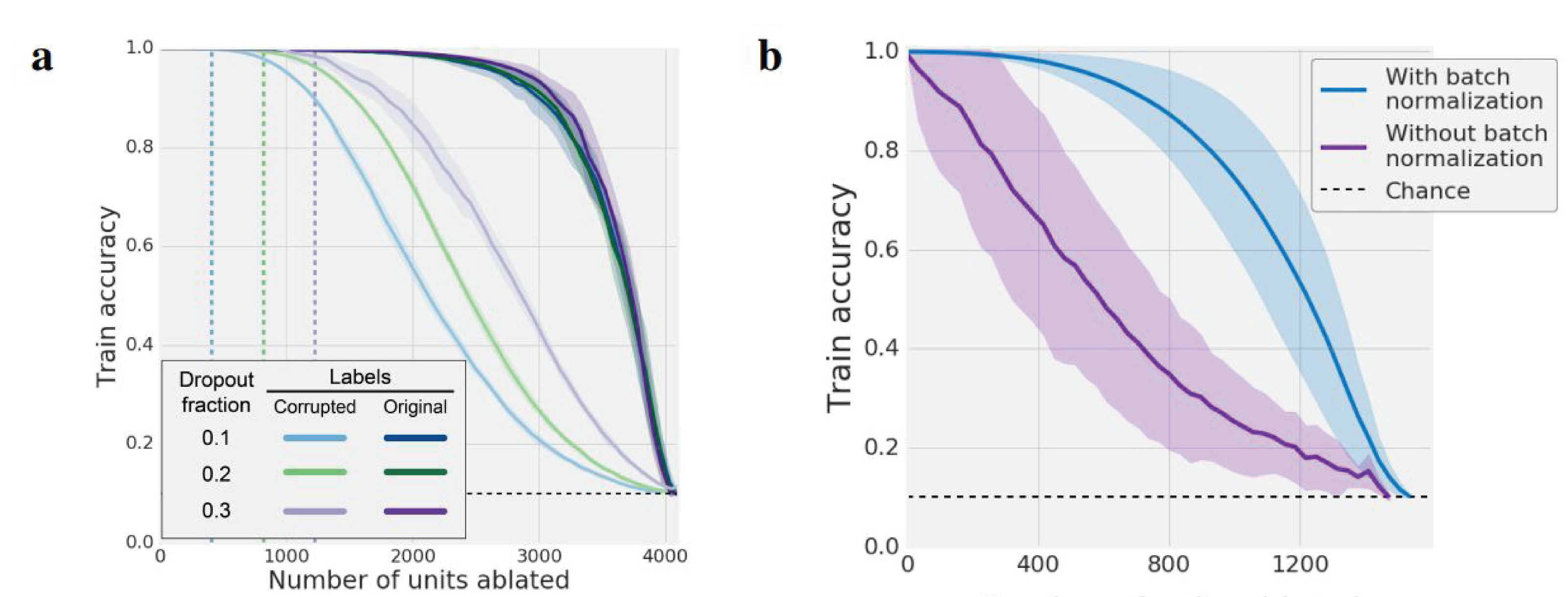


Figure 5: a, Cumulative ablation curves for MLPs trained on unmodified and fully corrupted MNIST with dropout fractions ∈ {0.1, 0.2, 0.3}. Colored dashed lines indicate number of units ablated equivalent to the dropout fraction used in training. Note that curves for networks trained on corrupted MNIST begin to drop soon past the dropout fraction with which they were trained. **b,** Cumulative ablation curves for networks trained on CIFAR-10 with and without batch normalization. Error bars represent standard deviation across 4 model instances and 10 random orderings of feature maps per model.

Batch normalization decreases class selectivity and increases mutual information

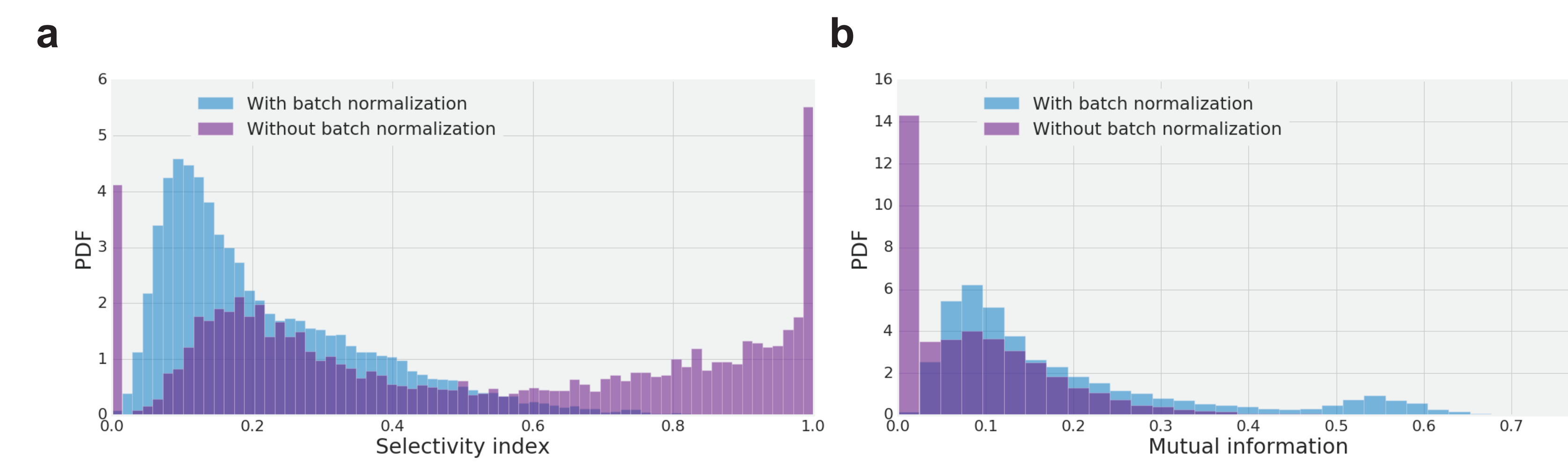


Figure 6: Distributions of class selectivity (**a**) and mutual information (**b**) for networks trained with (blue) and without batch normalization (purple). Each distribution comprises 4 model instances trained on uncorrupted CIFAR-10.

Selective and non-selective directions are similarly important

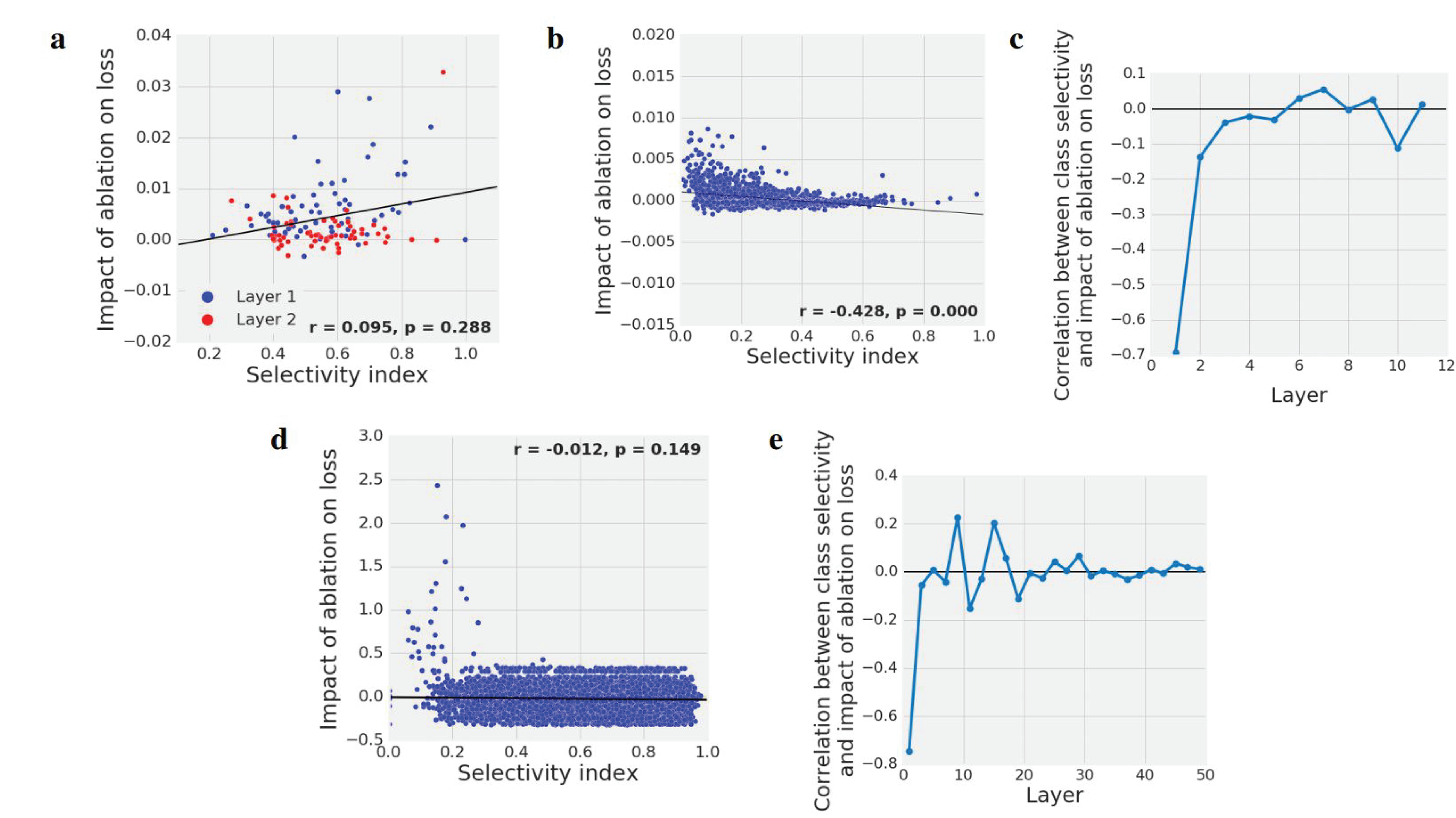


Figure 7: Impact of ablation as a function of class selectivity for MNIST MLP (**a**), CIFAR-10 convolutional network (**b-c**), and ImageNet ResNet (**d-e**). **c** and **e** show regression lines for each layer separately.

Summary

- Networks which generalize well are less reliant on single directions than those which generalize poorly
- While dropout only regularizes single direction reliance up to the dropout fraction, batch normalization implicitly regularizes reliance on single directions
  - It may do this by discouraging sparse representations in which information is focused in single units
- The selectivity of single units is a poor predictor of that unit's importance to the network output

Acknowledgements

We would like to thank Chiyuan Zhang, Ben Poole, Sam Ritter, Avraham Ruderman, and Adam Santoro for critical feedback and helpful discussions.