On the Importance of Single Directions for Generalization

A.S. Morcos, D.G.T. Barrett, N.C. Rabinowitz, M. Botvinick

DeepMind

arXiv: 1803.06959

Reviewed by : Bill Zhang
University of Virginia
https://qdata.github.io/deep2Read/

Outline

Introduction

Basic Premise and Motivation

- Research has suggested that DNNs can memorize entire datasets such as ImageNet
- Despite this, DNNs still generalize well; why do some networks generalize better than others?
- Other work: flatness of minima and PAC-bayes bounds, information content stored in network weights, SGD encourages generalization,...
- Focus on ablation analyses to measure reliance of network on single directions
- Define single direction in activation space as activation of single unit or feature map or some linear comb. of units in response to some input

Approach

Models and Datasets

- Three models: 2-hidden layer MLP trained on MNIST, 11-layer CNN trained on CIFAR-10, 50-layer residual network trained on ImageNet
- ReLU non-linearities applied to all layers but output
- Batch normalization was used for all networks
- Partially Corrupted Labels: used datasets with differing fractions of randomized labels to control degree of memorization: distribution of labels maintained, but any patterns were broken

Approach

Perturbation Analyses: Ablations

- Measured importance of single direction to a network by seeing how performance degrades once direction influence was removed
- To remove coordinate-aligned single direction, clamped activity in the direction to fixed value
- Ablations were performed on single units in MLPs or entire feature maps in CNNs and performed in activation space, not weight space
- See how network performance degrades as increasing subsets of single directions are ablated
- Decided to clamp to 0

Approach

Perturbation Analyses: Noise

- To test network dependence on random single directions (as opposed to coordinate-aligned), add Gaussian noise to all units with zero mean and progressively increasing variance
- Normalize variance by empirical variance of unit's activations across training set

Approach

Quantifying Class Selectivity of Individual Units

- Used metric inspired by selectivity indices used in systems neuroscience

$$selectivity = \frac{\mu_{max} - \mu_{-max}}{\mu_{max} + \mu_{-max}}$$

  where μ_{max} is highest class-conditional mean activity and μ_{-max} is mean activity for all other classes
- Metric ranges from 0 (unit's average activity same for all classes) to 1 (unit only active for inputs of single class)

Approach

Quantifying Class Selectivity of Individual Units

- Imperfect measure of selectivity: unit with little information about every class would have low index, but would measure discriminability of classes
- Replicate results using mutual information which highlights units with information about multiple classes

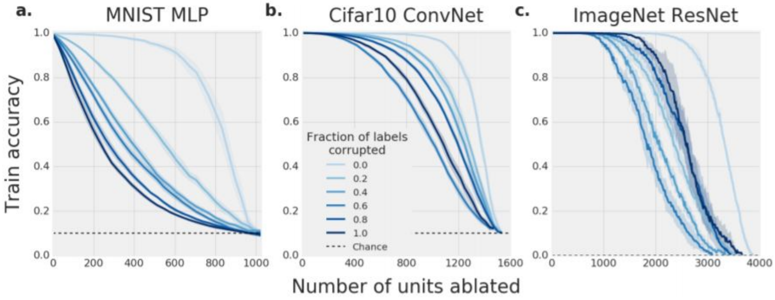
Experiments
Intuition

- Consider two large networks: one which memorizes the dataset, one which learns the structure and thus generalizes well
- Memorizing network should have larger minimal description length than generalizing network
- Therefore, memorizing network should use more capacity, and by extension, more single directions

Experiments

Effect of Memorization on Single Direction Dependence

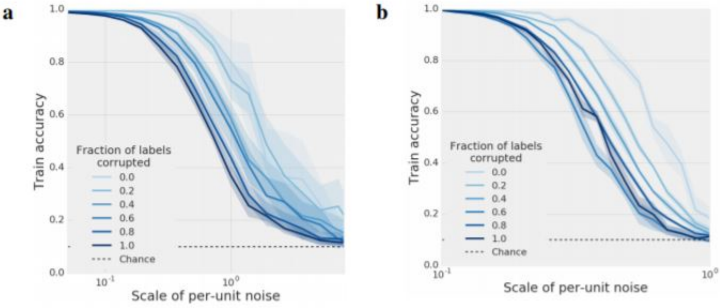
- Test whether memorization leads to greater dependence on single directions: train variety of networks on datasets with differing amounts of random labels and evaluate performance as more single directions were ablated
- More corrupted labels increased sensitivity to ablations



Experiments

Effect of Memorization on Random Single Directions

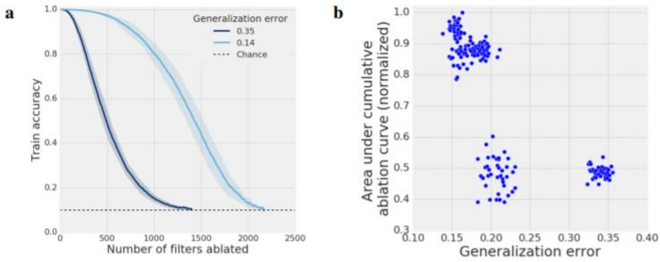
- Repeated similar experiment with random noise perturbation; similar findings
- Graphs show MLP on MNIST (a) and CNN on CIFAR-10 (b)



Experiments

Networks Trained on Same Data

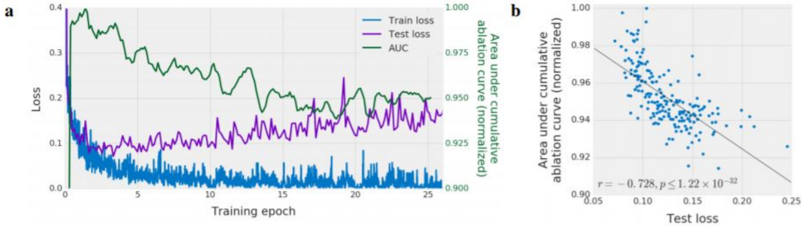
- Also want to see if conclusions apply to networks which are not forced to memorize set (i.e. trained with uncorrupted data)
- Trained 200 networks on CIFAR-10 with different initializations and training data order
- Compared 5 networks with best generalization and 5 networks with worst; similar findings as before
- Plot area under cumulative ablation curve for all 200 networks



Experiments

Single Directions as Signal for Model Selection

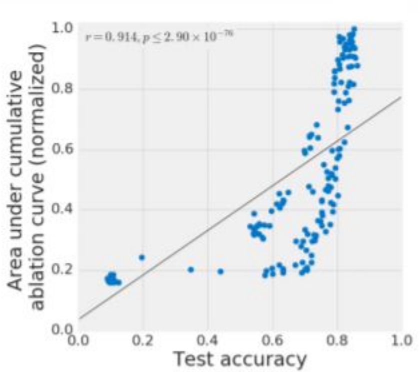
- Can single direction reliance be used to estimate generalization performance without need for held-out test set?
- Trained MLP on MNIST and measured area under cumulative ablation curve (AUC) over course of training
- AUC starts to drop when test and train accuracies start to diverge, AUC and test loss negatively correlated



Experiments

Single Directions as Signal for Model Selection

- Can single direction reliance be used for hyperparameter selection?
- Trained 192 CIFAR-10 models with different hyperparameters
- AUC and test accuracy highly correlated in hyperparameter sweep



Experiments

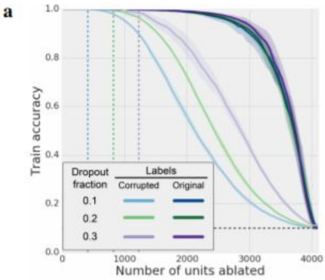
Relation to Dropout

- Similar to using dropout at training time; seems to discourage reliance on single directions
- However, network is only robust to ablations up to dropout fraction
- With enough capacity, a network can guard against dropout by making multiple copies of a single direction; however, network will only make minimum number of copies
- Network robust to dropout as long as all redundant single directions were not removed at the same time

Experiments

Relation to Dropout

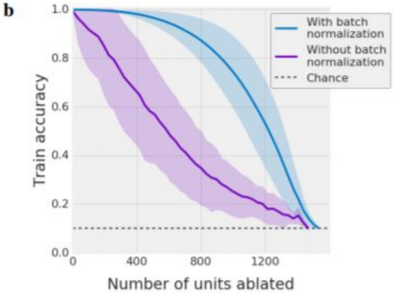
- Trained MLPs on MNIST with dropout probabilities of 0.1, 0.2, and 0.3 on both corrupted and unmodified labels
- Took longer to converge and converged to worse solutions; implies that memorization is discouraged
- However, past dropout, networks much more sensitive to ablations; suggests dropout is an effective regularizer, but only until dropout fraction



Experiments

Relation to Batch Normalization

- Batch normalization does appear to discourage reliance on single directions
- Trained CNNs on CIFAR-10 with and without and measured robustness to ablation



Experiments

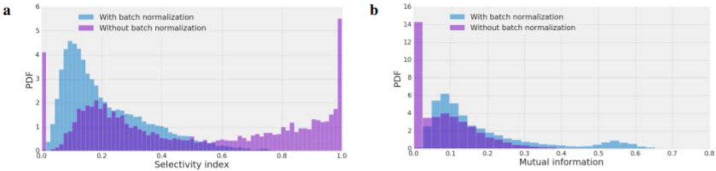
Class Selectivity and Importance

- Results suggest that networks less reliant on single directions generalize better
- Counter-intuitive to past work in neuroscience and deep learning which highlight important of single units/feature maps which are selective for particular features of classes
- Test whether class-selectivity of single directions affects importance of directions to a network's output

Experiments

Class Selectivity and Importance

- Test if batch normalization influences distribution of information about class across single directions
- Use selectivity index from before, trained 4 uncorrupted models on CIFAR-10
- Batch normalization actually discourages presence of feature maps with concentrated class information; raises question of whether highly selective feature maps are beneficial



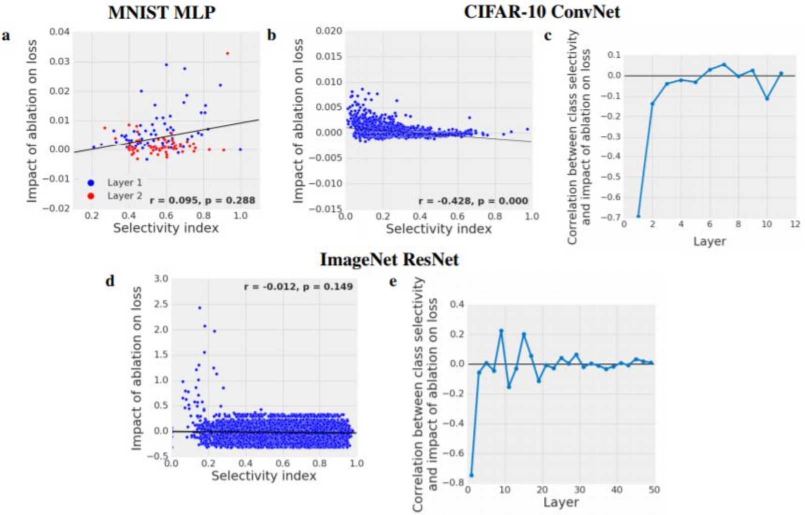
Experiments

Class Selectivity and Importance

- Determine if selectivity of unit affects impact of ablating said unit
- For MLPs trained on MNIST, very small correlation (Spearman's: 0.095)
- Many highly-selective units had minimal impact when ablated
- Similar results for CNNs on CIFAR-10
- Actually, CIFAR-10 had negative correlation; found to be driven by early network layers
- In all 3 networks, earlier ablations more impactful
- Repeated with mutual information and got similar results
- Overall: selective and non-selective units are similarly important

Experiments

Class Selectivity and Importance



Experiments

Selectivity and L1-norm

- Compared class selectivity to L1-norm of filter weights, a metric which is a good predictor of feature map importance
- Found to be largely unrelated (if not negatively related)
- Suggests class selectivity may in fact by detrimental to network perform; more research needs to be done

Related Work

- Direct inspiration from Zhang et al. (2017); replicated results using partially corrupted labels and answer the posed question: is there an empirical difference between networks which memorize and those which generalize?
- Linking generalization to sharpness of minima
- Contextualizing generalization in information theory
- Analysis on properties of models trained on corrupted labels
- Perturbation analyses: model pruning, finding maximally important direction, highlighting single selective units,...
- Concept selectivity metric

Discussion and Future Work

Conclusion

- Taken an empirical approach to comparing memorizing and generalizing networks
- Found the generalizing ability is related to reliance on single directions in models trained on both corrupted and uncorrupted data, and also over the course of training for a single network
- Showed that batch normalization discourages dependence on single directions
- Class selectivity largely uncorrelated to importance of unit to output; batch normalization actually decreases selectivity, which suggests that class selectivity may harm output

Discussion and Future Work

Future Work

- Construct regularizer which penalizes dependence on single directions
- Could assess generalization performance without sacrificing training data to be used as validation set
- Could use single direction reliance as a signal for early-stopping or hyperparameter searching
- Find extent to which train and test set overlap affects single direction dependence

References

- https://arxiv.org/pdf/1803.06959.pdf