ON THE IMPORTANCE OF SINGLE DIRECTIONS FOR GENERALIZATION

Ari S. Morcos & David G.T. Barrett & Neil C. Rabinowitz & Matt Botvinick
DeepMind
London, UK
{arimorcos,barrettdavid,ncr,botvinick}@google.com

ABSTRACT

Despite their ability to memorize large datasets, deep neural networks often achieve good generalization performance. However, the differences between the learned solutions of networks which generalize and those which do not remain unclear. 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. Finally, we find that while dropout only regularizes this quantity up to a point, batch normalization implicitly discourages single direction reliance.

1 INTRODUCTION

Recent work has demonstrated that deep neural networks (DNNs) are capable of memorizing extremely large datasets such as ImageNet [16]. Despite this capability, DNNs in practice achieve low generalization error on tasks ranging from image classification [6] to language translation [15]. These observations raise a key question: why do some networks generalize while others do not?

Answers to these questions have taken a variety of forms. A variety of studies have related generalization performance to the flatness of minima and PAC-Bayes bounds [7, 9, 10, 5], though recent work has demonstrated that sharp minima can also generalize [4]. Others have focused on the information content stored in network weights [1], while still others have demonstrated that stochastic gradient descent itself encourages generalization [3, 12, 14].

Here, we use ablation analyses to measure the reliance of trained networks on single directions in activation space. We define a single direction in activation space as the activation of a single unit or feature map in response to some input. We find that networks which memorize the training set are substantially more dependent on single directions than those which do not, and that this difference is preserved even across sets of networks with identical topology trained on identical data, but with different generalization performance. Moreover, we found that as networks begin to overfit, they become more reliant on single directions, suggesting that this metric could be used as a signal for early stopping. Finally, we show that batch normalization implicitly regularizes reliance on single directions.

2 APPROACH

Ablations We measured the importance of a single direction to the network's computation by asking how the network's performance degrades once the influence of that direction was removed. To remove a coordinate-aligned single direction , we clamped the activity of that direction to a fixed value (i.e., ablating the direction). Ablations were performed either on single units in MLPs or an entire feature map in convolutional networks. For brevity, we will refer to both of these as 'units.' Critically, all ablations were performed in activation space, rather than weight space.

More generally, to evaluate a network's reliance upon sets of single directions, we asked how the network's performance degrades as the influence of increasing subsets of single directions was removed by clamping them to a fixed value (analogous to removing increasingly large subspaces within activation space). This analysis generates curves of accuracy as a function of the number of directions ablated: the more reliant a network is on low-dimensional activation subspaces, the more quickly the accuracy will drop as single directions are ablated.

Summary of models analyzed We analyzed three models: a 2-hidden layer MLP trained on MNIST, an 11-layer convolutional network trained on CIFAR-10, and a 50-layer residual net trained on ImageNet. In all experiments, ReLU nonlinearities were applied to all layers but the output. Unless otherwise noted, batch normalization was used for all convolutional networks [8]. For the ImageNet ResNet, top-5 accuracy was used in all cases.

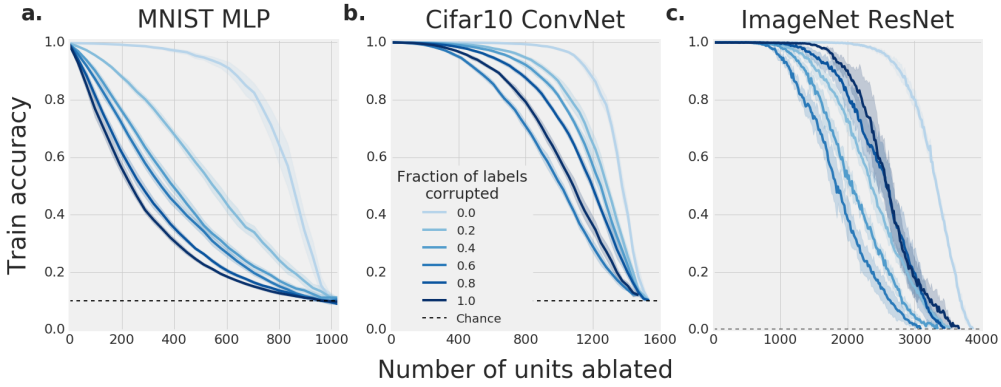

Figure 1: Memorizing networks are more sensitive to cumulative ablations. Networks were trained on MNIST (2-hidden layer MLP, **a**), CIFAR-10 (11-layer convolutional network, **b**), and ImageNet (50-layer ResNet, **c**). In **a**, all units in all layers were ablated, while in **b** and **c**, only feature maps in the last three layers were ablated. Error bars represent standard deviation across 10 random orderings of units to ablate.

Partially corrupted labels As in [16], we used datasets with differing fractions of randomized labels to ensure varying degrees of memorization. To create these datasets, a given fraction of labels was randomly shuffled and assigned to images, such that the distribution of labels was maintained, but any true patterns were broken.

3 Experiments

3.1 Generalization

Here, we provide a rough intuition for why a network's reliance upon single directions might be related to generalization performance. Consider two networks trained on a large, labeled dataset with some underlying structure. One of the networks simply memorizes the labels for each input example and will, by definition, generalize poorly ('memorizing network') while the other learns the structure present in the data and generalizes well ('structure-finding network'). The minimal description length of the model should be larger for the memorizing network than for the structure-finding network. As a result, the memorizing network should use more of its capacity than the structure-finding network, and by extension, more single directions. Therefore, if a random single direction is perturbed, the probability that this perturbation will interfere with the representation of the data should be higher for the memorizing network than for the structure-finding network[1].

To test whether memorization leads to greater reliance on single directions, we trained a variety of network types on datasets with differing fractions of randomized labels and evaluated their performance as progressively larger fractions of units were ablated[2]. As many of the models were trained on datasets with corrupted labels and, by definition, cannot generalize, training accuracy was used to evaluate model performance. Consistent with our intuition, we found that networks trained on varying fractions of corrupted labels were significantly more sensitive to cumulative ablations than those trained on datasets comprised of true labels, though curves were not always perfectly ordered by the fraction of corrupted labels (Fig. 1).

The above results apply to networks which are forced to memorize at least a portion of the training set – there is no other way to solve the task. However, it is unclear whether these results would apply to networks trained on uncorrupted data. In other words, do the solutions found by networks with the same topology and data, but different generalization performance exhibit differing reliance upon single directions? To test this, we trained 200 networks on CIFAR-10, and evaluated their generalization error (train accuracy - test accuracy) and reliance on single directions[3]. We found that 5 networks with the best generalization performance were more robust to the ablation of single directions than the 5 networks with the worst generalization performance (Fig. 2a). To quantify this further, we measured the area

[1] Assuming that the memorizing network uses a non-negligible fraction of its capacity.

[2] By definition, these curves must begin at the network's training accuracy (approximately 1 for all networks tested) and fall to chance levels when all directions have been ablated.

[3] All networks had the same topology and were trained on the same dataset (unmodified CIFAR-10). Individual networks only differed in their random initialization (drawn from identical distributions) and the data order used during training.

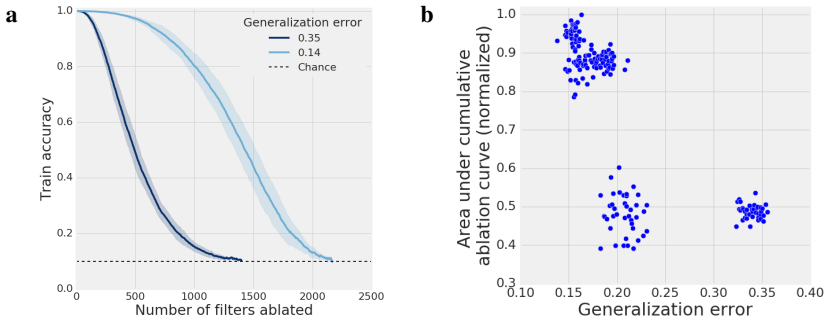

Figure 2: Networks which generalize poorly are more reliant on single directions. 200 networks with identical topology were trained on unmodified CIFAR-10. **a**, Cumulative ablation curves for the best and worst 5 networks by generalization error. Error bars represent standard deviation across 5 models and 10 random orderings of feature map ablations per model. **b**, Area under cumulative ablation curve (normalized) as a function of generalization error.

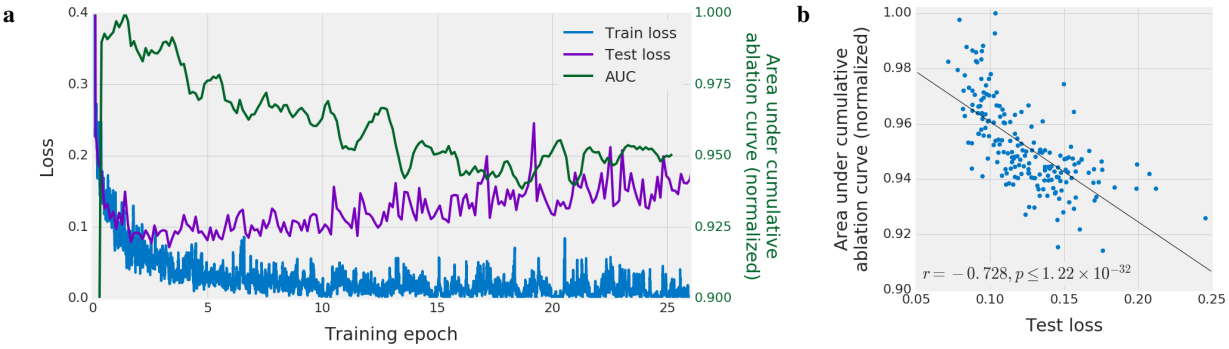

Figure 3: Single direction reliance as a signal for early stopping. a, Train (blue) and test (purple) loss, along with the normalized area under the cumulative ablation curve (AUC; green) over the course of training for an MNIST MLP. Loss y-axis has been cropped to make train/test divergence visible. **b,** AUC and test loss are negatively correlated over the course of training.

under the ablation curve for each of the 200 networks and plotted it as a function of generalization error[4] (Fig. 2b). These results demonstrate that the relationship between generalization performance and single direction reliance is not merely a side-effect of training with corrupted labels, but is instead present even among sets networks with identical training data.

3.2 Reliance on single directions over the course of training

This relationship raises an intriguing question: can single direction reliance be used to estimate generalization performance without the need for a held-out test set? And if so, might it be used as a signal for early stopping? As a proof-of-principle experiment, we trained an MLP on MNIST and measured the area under the cumulative ablation curve (AUC) over the course of training along with the train and test loss. Interestingly, we found that the point in training at which the AUC began to drop was the same point that the train and test loss started to diverge (Fig. 3a). Furthermore, we found that AUC and test loss were negatively correlated (Spearman's correlation: -0.728; Fig. 3b). These results suggest that single direction reliance may serve as a good proxy for early stopping, but further work will be necessary to evaluate whether these results hold in more complicated datasets.

3.3 Relationship to dropout and batch normalization

Dropout Our experiments are reminiscent of using dropout at training time, and upon first inspection, dropout may appear to discourage networks' reliance on single directions [13]. However, while dropout encourages networks to be robust to cumulative ablations up until the dropout fraction used in training, it should not discourage reliance on single directions past that point, as a memorizing network could merely copy the information stored in a given direction to several other directions to guard against the training fraction. In such a case, the network would be robust to dropout

[4]Interestingly, networks appeared to undergo a discrete regime shift in their reliance upon single directions (though there was also a negative correlation present within clusters); however, this effect might have been caused by degeneracy in the set of solutions found by the optimization procedure.

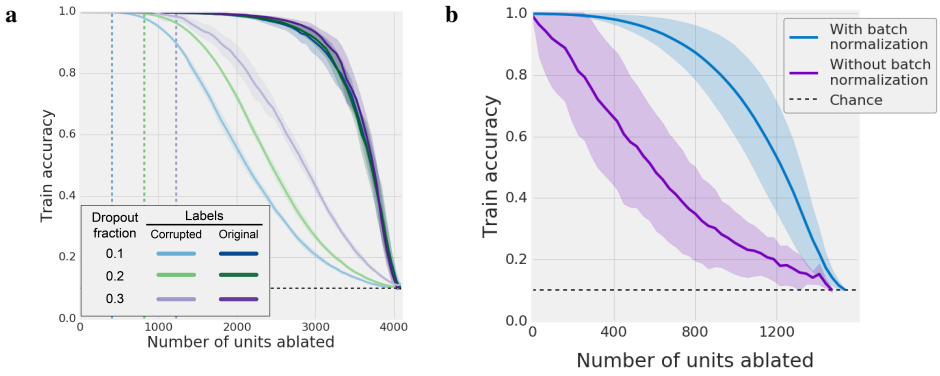

Figure 4: Impact of regularizers on networks' reliance upon single directions. a, Cumulative ablation curves for MLPs trained on unmodified and fully corrupted MNIST with dropout fractions $\in \{0.1, 0.2, 0.3\}$. Colored dashed lines indicate number of units ablated equivalent to the dropout fraction 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.

so long as all redundant directions were not simultaneously removed, yet still be highly reliant on single directions past the dropout fraction used in training.

To test whether this intuition holds, we trained MLPs on MNIST with dropout probabilities $\in \{0.1, 0.2, 0.3\}$ on both corrupted and unmodified labels. While networks trained on both corrupted and unmodified labels exhibited minimal loss in training accuracy as single directions were removed up to the dropout fraction used in training, past this point, networks trained on randomized labels were much more sensitive to cumulative ablations than those trained on unmodified labels (Fig. 4a). Interestingly, networks trained on unmodified labels with different dropout fractions were all similarly robust to cumulative ablations. These results suggest that while dropout may serve as an effective regularizer to prevent memorization of randomized labels, it does not prevent over-reliance on single directions past the dropout fraction used in training.

Batch normalization In contrast to dropout, batch normalization does appear to discourage reliance upon single directions. To test this, we trained convolutional networks on CIFAR-10 with and without batch normalization and measured their robustness to cumulative ablation of single directions. Networks trained with batch normalization were consistently and substantially more robust to these ablations than those trained without batch normalization (Fig. 4b). This result suggests that in addition to reducing covariate shift, as has been proposed previously [8], batch normalization also implicitly discourages reliance upon single directions.

4 Related work

Much of this work was directly inspired by [16], and we replicate their results using partially corrupted labels on CIFAR-10 and ImageNet. By demonstrating that memorizing networks are more reliant on single directions, we also provide an answer to one of the questions they posed: is there an empirical difference between networks which memorize and those which generalize?

Our work is also related to work linking generalization and the sharpness of minima [7, 9, 10]. These studies argue that flat minima generalize better than sharp minima (though [4] recently found that sharp minima can also generalize well). This is consistent with our work, as flat minima should correspond to solutions in which perturbations along single directions have little impact on the network output.

Another approach to generalization has been to contextualize it in information theory. For example, [1] demonstrated that networks trained on randomized labels store more information in their weights than those trained on unmodfied labels. This notion is also related to [11], which argues that during training, networks proceed first through a loss minimization phase followed by a compression phase. Here again, our work is consistent, as networks with more information stored in their weights (i.e., less compressed networks) should be more reliant upon single directions than compressed networks.

More recently, [2] analyzed a variety of properties of networks trained on partially corrupted labels, relating performance and time-to-convergence to capacity. They also demonstrated that dropout, when properly tuned, can serve as an

effective regularizer to prevent memorization. However, we found that while dropout may discourage memorization, it does not discourage reliance on single directions past the dropout probability.

5 Conclusion

In this work, we have taken an empirical approach to understand what differentiates neural networks which generalize from those which do not. Our experiments demonstrate that generalization capability is related to a network's reliance on single directions, both in networks trained on corrupted and uncorrupted data, and over the course of training for a single network. They also show that batch normalization, a highly successful regularizer, seems to implicitly discourage reliance on single directions. Finally, we use a proof-of-principle experiment to show that this metric could be used as a proxy signal for early stopping without the need for a held-out test set.

References

[1] Alessandro Achille and Stefano Soatto. On the Emergence of Invariance and Disentangling in Deep Representations. pp. 1–17, 2017. URL http://arxiv.org/abs/1706.01350.

[2] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A Closer Look at Memorization in Deep Networks. 2017. ISSN 1938-7228. URL http://arxiv.org/abs/1706.05394.

[3] Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of machine learning research: JMLR*, 2(Mar):499–526, 2002. ISSN 1532-4435. URL http://www.jmlr.org/papers/volume2/bousquet02a/bousquet02a.pdf.

[4] Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp Minima Can Generalize For Deep Nets. 2017.

[5] Gintare Karolina Dziugaite and Daniel M. Roy. Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data. 2017. URL http://arxiv.org/abs/1703.11008.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. *Arxiv.Org*, 7(3):171–180, 2015. ISSN 1664-1078. doi: 10.3389/fpsyg.2013.00124.

[7] Sepp Hochreiter and Jürgen Schmidhuber. Flat Minima. *Neural Comput*, 9(1), 1997.

[8] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Arxiv*, 2015. URL http://arxiv.org/abs/1502.03167.

[9] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikahail Smelyanskiy, and Ping Tak Peter Tang. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. In *ICLR*, pp. 1–16, 2017.

[10] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. Exploring Generalization in Deep Learning. 2017. URL https://arxiv.org/abs/1706.08947.

[11] Ravid Shwartz-Ziv and Naftali Tishby. Opening the Black Box of Deep Neural Networks via Information. *arXiv*, pp. 1–19, 2017. URL http://arxiv.org/abs/1703.00810.

[12] Samuel L. Smith and Quoc V. Le. Understanding Generalization and Stochastic Gradient Descent. pp. 1–11, 2017. URL http://arxiv.org/abs/1710.06451.

[13] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout : A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research (JMLR)*, 15: 1929–1958, 2014. ISSN 15337928.

[14] Ashia C. Wilson, Rebecca Roelofs, Mitchell Stern, Nathan Srebro, and Benjamin Recht. The Marginal Value of Adaptive Gradient Methods in Machine Learning. pp. 1–14, 2017. URL http://arxiv.org/abs/1705.08292.

[15] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[16] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. 2017. URL http://arxiv.org/abs/1611.03530.