CS 7545 Final Project - Neural Network Complexity and Generalizability

Sai Pradyumna Chermala, Chaewon Park, Jeffrey Chang
Georgia Institute of Technology

May 5, 2023

Introduction

Machine learning and deep learning have transformed the way we approach many complex problems in artificial intelligence. These powerful techniques can learn complex patterns from data by themselves, allowing them to make predictions and decisions with reasonably high accuracy. Despite their enormous complexity, and difficult optimization process amidst several possible global minima, these models generalize surprisingly well to unseen data. Generalization error of a model refers to the difference between the its performance on the test data and that on the training data. Generalization error gauges the model's ability to make accurate predictions on new data.

Generalization is a fundamental challenge in machine learning that a number of researchers have extensively studied over the past few decades. Understanding generalization is critical for advancing the field of ML and DL since the ultimate goal is to build stable and reliable models that can perform well even on new, unseen data, but this is challenging due to several factors. Highly complex models, as is the case with most modern Deep Learning methods, are prone to overfitting. A model that overfits the training data and fails to generalize well is of little use in practice. Therefore, understanding the theoretical underpinnings of generalization and improving model performance on unseen data is crucial for building more robust and reliable ML and DL models.

The three papers, "Understanding Deep Learning Requires Rethinking Generalization (Zhang et al., 2016)," "Exploring Generalization in Deep Learning (Neyshabur et al., 2017))," and "Spectrally-normalized margin bounds for neural networks (Bartlett et al., 2017)," all tackle the challenge of generalization by investigating the theoretical foundations of generalization in standard neural network models with benchmark datasets. These papers provide insights into why and how neural network models can generalize well, and take strides at providing new generalization bounds that are more relevant to these models and are tighter than popular traditional measures. By summarizing and connecting these papers, we aim to provide a comprehensive overview of the state-of-the-art research in the field of generalization in machine learning.

Paper Summaries

Understanding Deep Learning Requires Rethinking Generalization

(Zhang et al., 2016) questions on what distinguishes neural networks that generalize well from those generalize poorly. The authors stress that answer to such a question is fundamental to building a reliable architecture. Nonetheless, they point out that none of the popular traditional theories are capable of distinguishing networks with varying generalization performance well enough. These include complexity measures of model families such as Rademacher complexity, VC dimension, uniform stability, or the use of regularization techniques during training. They cast doubts on existing complexity measures' ability by conducting several experiments with standard image classification network architectures and benchmark datasets.

Randomization

Inspired by non-parametric randomization tests, the authors perform a series of experiments to derive surprising insights. They perform two type of randomization tests - label randomization and input randomization.

- **Label Randomization.** Authors discover from their experiments that deep neural networks can easily fit randomized labels and still achieve zero training error. And optimization of such a model isn't a challenge either, with convergence proceeding similar to the true label training, albeit slower. In fact, the hyperparameter choices remained identical to the ones in true label training. The paper also presents partial randomization tests with varying proportions of labels randomized. All such trainings converge to zero training error (on CIFAR10 dataset) but the generalization error increased with increasing randomization. Since this randomization doesn't change any of the model parameters or the optimizer but changes the generalization performance, measures like VC-dimension, Rademacher complexity which depend on the properties of model family fail to explain these results. Further, the ability to fit any random labeling leads to trivial bounds with these measures. On the other hand, uniform stability measures consider the properties of training algorithm, but by ignoring the data and label distributions, they cannot explain the results of these tests either.
- **Input Randomization.** Keeping the labels unchanged, the authors also analyze the results of training models on randomized input. They perform tests by randomly shuffling the pixels of images and by even replacing the pixels altogether by random sampling from a Gaussian distributions. The findings remain largely the same as the previous test and they note a correlation between the noise level and generalization error. By destroying the structure of the input data, they rule out the possibility of the network architecture itself acting as a regularizer.

Finite Sample Expressivity

Contrary to the popular work, the authors argue that finite sample expressivity is more important in practice than population level understanding. Using uniform convergence theorems, one can convert population level results to finite sample results but the resulting bounds are impractical due to strong dependence on sample size and network depth. Instead, they analyze the finite-sample expressivity directly and theoretically show that a simple two-layer ReLU neural network can express any labeling of training data when the number of trainable parameters $p = 2n + d$ where n is sample size and d is the dimensions. This finding explains that even a network with a small depth is enough to represent any labeling in contrast to the existing beliefs that a network needs far more parameters ($O(dn)$) to achieve such performance.

Explicit Regularization

Authors argue that explicit regularization (e.g., weight decay, data augmentation, dropout, etc.) does not control generalization errors aside from the fact that it may help improve generalization performance. Firstly, models trained on CIFAR10 and ImageNet datasets generalize well even without any explicit regularization (although the error is higher than those trained with explicit regularization). In addition, these techniques cannot prevent near-perfect fitting on randomized labels. While the traditional notion is that explicit regularization is a must in overparameterized settings, the paper notes that it plays a different role in deep learning.

Implicit Regularization

Implicit regularization techniques like early stopping and batch normalization, either do not offer any benefit or help improve the generalization by a small margin but do not explain it fully. The authors then shift their focus to the training algorithm and study the solutions of Stochastic Gradient Descent (SGD) on linear models. They find that even without any other regularization, SGD converges to minimum norm solution under overparameterized settings. They derive a closed form solution for linear models trained with SGD and use it on MNIST, and CIFAR10 datasets to achieve low error rates. While low norm solutions do not necessarily mean better generalization always, the authors suggest that SGD acts as an implicit regularizer and that a deeper understanding is required of the algorithm and the models it outputs.

Exploring Generalization in Deep Learning

Neyshabur et al. (2017) is one of the many works attempting to understand the reason behind the generalization of deep neural networks. The authors begin by pondering over the effectiveness of these methods despite being complex, non-convex to optimize, and heavily overparameterized. The paper delves into the existing explanations for generalization, their shortcomings and the need for scale normalization. In addition, connections are made between sharpness measure and PAC-Bayes theory.

The paper notes that complexity measures which depend solely on architecture ignoring the algorithmic choices like initialization, optimization algorithm, number of hidden units etc., do not sufficiently explain the generalization. This is because for the same architecture, the choice training on actual labels or randomized labels and the choice of optimization algorithm could produce different generalization results. A good complexity measure should be able to distinguish these factors in line with empirical evidence.

- **VC dimension.** Existing works have derived tight VC-dimension bounds in terms of the number of parameters in the network. For a feed forward ReLU network this bound is $\tilde{O}(d \cdot \dim(w))$ where d is the depth of the network and $\dim(w)$ is the number of parameters. However, these do not explain the generalization in over-parameterized settings and the fact that generalization could improve with more parameters.

- **Norm.** Similar to linear classifiers, norm based bounds have been derived for feed forward networks as well. These use l_1/l_∞, $l_{p,q}$ group norms, and l_p path norms. However, the paper points two issues with norm based capacity bounds.

    1. Zero/one losses are independent of the scale of the norms. Different versions of the same network with arbitrarily scaled weights would result in the same loss but their norms and correspondingly, (unnormalized) norm based capacities would be different.

    2. Scale sensitive losses like cross-entropy force the weight norms to go towards infinity as the training progresses. Comparing norms in such case wouldn't be useful as they all go towards infinity.

  Instead, they should be scaled for a meaningful comparison. One such scaling factor could be the "margin". The authors examine multiple such margin-scaled norm bounds using a soft margin formulation. They provide empirical evidence that these measures match the expectations in that:

    1. Models trained on random labels show higher complexity than those trained on actual labels, and

    2. Increasing the training set size only marginally increases the complexity for models trained on true labels but considerably for those trained on random labels.

- **Lipschitz Continuity.** Since norm based measures also constrain the Lipschitz constant of these networks, the authors examine whether the latter is the true measure. However, they point out that bounding Lipschitz based measures still scales the previous measures (norm based bounds) exponentially in both the input dimensions and the network depth.

- **Sharpness.** Sharpness as introduced in McAllester (2003) gives lower bounds for networks trained on random labels than those trained on actual labels (for small networks). Instead, the authors propose an alternative view by linking sharpness with PAC-Bayes framework. They show that sharpness is only one of the terms in the framework and does not explain the other term - KL divergence of "posterior" to the "prior". As the bound from the PAC-Bayes framework is valid for any distribution, the paper proposes a simple bound by considering spherical Gaussians for both the prior and the posterior. The resulting bound has two terms - expected sharpness and norm both related by the variance σ^2 of the Gaussians. Increasing σ decreases the norm term but increases the sharpness term and vice versa, thus establishing generalization as a balance between norm and sharpness.

Empirical evidence

The paper then analyzes the measures discussed so far on two tests.

- **Different global minima.** By training on a varying size of "confusion" sets in addition to the training set, the authors force the network into global minima with varying generalization errors. Here, norm based measures and PAC-Bayes interpretation of sharpness explain the practical observations but sharpness alone doesn't. This means that norm based measures and PAC-Bayes were able to distinguish generalizable models from those with low generalizability through comparing these measures.
- **Network size.** None of the measures discussed could fully explain the improved generalization with increasing hidden units. As the size of the network increased, our complexity measures greatly increased, even though there was no associated loss in generalization ability.

Bounding sharpness. Finally, the authors propose a set of conditions to guide the optimization towards low sharpness solutions. They note that sharpness could be high if there are weak interactions among the layers, if the changes in number of activations is exponential even for small perturbations, or if an unactivated node becomes active due to the perturbations. They propose conditions to prevent each of them from happening and derive a generalization bound under these conditions which only scales linearly in depth.

Spectrally-normalized margin bounds for neural networks

Gauging how well a neural network model generalizes, or in other word how it can fit any unseen dataset, has been a popular topic in machine learning for decades. (Bartlett et al., 2017) criticizes traditional complexity measures such as VC-dimension and further examines other magnitude-sensitive complexity measures including Rademacher, covering numbers, and Lipschitz constant. Moving forward, they prove a generalization bound for neural networks based on the Lipschitz constant, but normalized with the margin of predictor in their paper.

The authors start their paper with empirically examining the Lipschitz constant of networks chosen by SGD. Excess risk, which is the difference between test and train error, is closely correlated with the Lipschitz constant when modeling both original and random labels. Yet, Lipschitz constant itself is not sufficient to represent the complexity of the model as it continuously grows, whereas the excess risk curve flattens as the number of epochs grows. To address such an issue, the authors introduced margin to Lipschitz constant. Margin here measures the distance between the output for the correct label and other labels. When normalized by the margin, networks with increasing Lipschitz constants showed a non-increasing or even a decaying curve. Proposed generalization bounds that scales with the Lipschitz constant divided by the margin have following properties.

1. Bound has no dependence on combinatorial parameters outside of log factors.
2. Bound has no explicit dependence on the number of classes in a multi-class dataset.
3. Bound measures complexity against a reference network.

Authors started with a theorem that provides a generalization bound for neural networks that involve fixed non-linearities and weight matrices with bounded spectral complexity $R_{\mathcal{A}}$ which is defined as follows.

$$R_{\mathcal{A}} = \Big(\Pi_{i=1}^{L} \rho_i \|A_i\|_\sigma\Big)\Big(\sum_{i=1}^{L} \frac{\|A_i^T - M_i^T\|_{2,1}^{\frac{2}{3}}}{\|A_i\|_\sigma^{\frac{2}{3}}}\Big)^{\frac{3}{2}} \tag{1}$$

The theorem is then proved by using traditional complexity measures, covering numbers and Rademacher complexity.

- Upper bounds of covering number complexity

  The theorem is proved by controlling the Rademacher complexity for networks with bounded $R_{\mathcal{A}}$ via covering numbers. First, given a matrix covering number bound for one layer, it induces a covering number bound for the entire network. Then, the resulting covering number bound along with the standard Dudley entropy integral upper bound are used to derive the proposed bound.

- Lower bounds of Rademacher complexity

  The proposed bound is for neural-networks with fixed non-linearities. However, authors suggested using reduction to eliminate non-linearities in order to present new lower bounds on the Rademacher complexity that scale with $\Pi_{i=1}^{L}\|A_i\|$.

The bound is then used to illustrate the margin distribution - distribution of margins of all data points. However, margin distribution at the end of training doesn't imply much information if not normalized - meaning that there isn't much of a difference with and without random labels. With properly normalized margin distribution, authors further studied the generalization behavior of neural networks.

- **Comparing datasets.** Authors empirically demonstrated the practical effectiveness of their proposed method by comparing complexities of cifar10, cifar100, mnist dataset. They investigated the normalized margin distribution plot, and found that the dataset whose margin distribution is further to the right of another dataset is considered to be "easy" to learn.
- **Convergence of margins.** Margin distributions converge during training even though the weight matrices continue to grow.
- **Regularization.** It has been empirically proved that regularization improves test error by a little to no degree.

Synthesis

We will start off by synthesizing the results of the papers, and explaining how they relate to each other, as they discuss many of the same topics. All three papers were submitted in 2016-2017 within a few months of each other. As such, they indicate that this period of time had relatively significant results in this area. Each of these papers contributed to the overall understanding of generalization and why machine learning neural network models are generalizable. When we look at them as a collective, we get a much clearer understanding of generalizability. Throughout this section, we will refer to those three papers as the first, second, and third paper in the order it's introduced in Paper Summaries section.

Randomization Tests

The first paper published was Zhang et al. (2016), which very notably used randomized tests on labels and inputs to show that generalization was not just happening because of the specific model acting as a regularizer, as these changes clearly force a lack of generalizability with no corresponding changes to the model. This same idea is further used in Neyshabur et al. (2017) to evaluate margin-scaled bounds, and whether the complexity matches the less accurate results produced by randomly labeled data.

Using randomization tests is a valuable tool in understanding which models are able to generalize. This is because real world data has some level of connections between its label and the data given. However, when given randomized training labels, there is no way for the model to give real results, as it is just guessing randomly on test data. By studying the behavior of models on data that is randomly changed, we can observe how the behavior changes as we become less able to generalize to real data. For the first paper, this is utilized to prove that generalization isn't caused by an invariant to this change and that the models can brute force training error. On the other hand, the second paper uses this as a valuable tool to force a model into different global minima and to judge whether complexity measures distinguish the different minima each of which exhibit a different generalization error.

Complexity Measures

Another major common theme throughout these papers is the rejection of Rademacher complexity, VC-dimension, and regularization as useful ways to interpret generalization ability. Each of the papers mentions the inability of these measures to explain some aspect of generalizability, and proves that we need new ways to measure a model to show whether we believe it to be accurate to outside untrained data. Both the first and the second papers also state that complexity measures that depend on architecture alone ignoring the other factors like data and label distribution, choice of optimization algorithm, initialization etc. cannot explain the observed generalization.

The first paper looks at all three of these topics. Regarding the Rademacher complexity, it explains that since the randomization tests found that neural networks were able to perfectly match the random labels, the Rademacher complexity for the model class was just approximated as 1, which is a trivial upper bound that gives us no useful information regarding generalization in real world examples. For VC-dimension, it is explained how the bounds on the fat-shattering dimension are irrelevant because a completely accurate model has no useful bounds that have been discovered.

The other two papers cover the same few topics in varying levels of depth. The third paper gives a more positive outlook on the potential of Rademacher complexity. Although it also reiterates the issues mentioned in the other papers, it finds a new generalization bound for neural networks $R_{\mathcal{A}}$ that uses Rademacher complexity as part of it. By controlling for the Rademacher complexity using covering numbers, the bounds on this new complexity measure was found to have successful applications to measuring generalization behavior. Thus, it seems possible that with more research, Rademacher complexity may have some applications to the subject.

The second paper shows more rigorously how the bounds on VC-dimension directly lead to an inability to explain generalization errors. For a feedforward network with ReLU activations, the VC-dimension bound in terms of parameters is $\tilde{O}(d \cdot \dim(w))$ where d is the depth of the network and $\dim(w)$ is the number of parameters. However, in many deep learning models, this is completely irrelevant, as the network may have significantly more parameters than samples, and can perfectly fit random labels. Furthermore, an increase in parameters can oftentimes reduce generalization error as the number of hidden units are increased in a model. The third paper also briefly mentions the issues with VC-dimension. However, it also talks about how future work in developing complexity analyses could use VC-dimension ideas as a basis for how to understand the bounds on generalizability.

The second and the third paper both discuss Lipschitz complexity in a good amount of detail. While the second paper seems to take the viewpoint that the Lipschitz constant has significant issues, the third paper looks at it far more favorably and focuses on it as a potential method to measure model complexity. The second paper mentions that certain norm measures correlate well with generalization test performance and that bounding norms also bounds Lipschitz constant of a network. But they conclude that the generalization is not a consequence of simply bounding latter. This is because bounding Lipschitz constant will still scale the capacity exponentially in input dimensions and network depth, which is the opposite of what we desire for a generalizability complexity measure.

On the other hand, the third paper (Bartlett et al. (2017)) provides a generalization bound for neural networks based on the Lipschitz constant, but normalizes it with the margin of predictor. Their results seem to show that Lipschitz constants are closely correlated with the excess risk. To account for the growth of the function with the progress of training, the authors introduced the concept of "margin". Once normalized, the results seemed to indicate that this is a successful area for potential future research.

It is worth noting that the second paper also independently introduced "margin" normalized bounds - claiming norm measures can be scaled arbitrarily without changing the generalization performance. Both these papers motivate margin based normalization from the observation that measures like norm, Lipschitz constant continue to grow with training and unless normalized, the comparisons would not be useful. While the second paper favors norm based bounds dissing Lipschitz measures, the third paper does the opposite.

However, the latter does not provide any tests with normalized norm measures.

It was very interesting that these papers seemed to have opposite viewpoints on the issue - they were published only a single day apart so they definitely were not able to consult the other paper.

Regularization

Regarding regularization, all three papers are in agreement. They all mention how explicit regularization is very little of the reason why we find success in generalization in neural networks. In the first paper, the authors challenge the role of explicit regularization by performing a series of experiments. When empirically testing models on the CIFAR10 and ImageNet datasets, it was found that generalization happens at a very similar rate, regardless of the explicit regularization techniques used. They were all able to reach 100% or very close accuracy on the training data. Even without any of these techniques, the models generalized extremely well - far above the performance of the random labels. These empirical tests from the first paper are even referenced in the third as evidence, showing a level of collaboration between the two. The second paper explains how their model uses no explicit regularization, and thus believes that their successful result regarding their test results were likely due to some aspect of implicit regularization.

When discussing implicit regularization, it seems that there is far more potential seen by these papers. The first paper references how SGD has a lower complexity as measured by norms as a result of how the algorithm converges, which acts like a sort of implicit regularization, and seems to have good results. The second paper mentions how implicit regularization was likely the cause of their success in training a feedforward network on the MNIST dataset. The third paper also mentions the success of SGD, and explains how the algorithm itself may lead to refined generalization bounds, even when large margin predictors are considered.

Overall, it seems like the most insightful areas for finding good methods to analyze neural network complexity are through implicit regularization, normalized Lipschitz and norm measures, and PAC-Bayes.

Conclusion

Over the past few years, machine learning and deep learning have greatly improved because of the development of new technologies and upgrades in the abilities of older ones. In particular, we find ourselves ever increasingly more reliant on artificial intelligence in our daily lives, with tools like ChatGPT and similar chat bots becoming a regular feature of many popular websites. However, it is not well understood why many of the neural network algorithms work as well as they do - why they are able to generalize so well to data beyond their training data.

Our paper looked at a few important papers in this subject, and synthesized their results together. Through the amalgamation of these papers, we believe that we have found valuable connections and gained insight into the current body of research on why generalization works for neural networks.

In particular, there seemed to be significant agreement about the challenges Rademacher complexity, VC-dimension, and explicit regularization have in explaining generalization. On the other hand, it seemed like implicit generalization had significant relationships to generalization, although the mechanism for this is still unknown. Some additional candidates with potential for analyzing generalizability of ML models are the normalized Lipschitz and norm measures, and PAC-Bayes. This will serve as a useful basis for understanding the current status, and proposing future directions for research.

References

P. L. Bartlett, D. J. Foster, and M. J. Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

D. McAllester. Simplified pac-bayesian margin bounds. In B. Schölkopf and M. K. Warmuth, editors, *Learning Theory and Kernel Machines*, pages 203–215, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg. ISBN 978-3-540-45167-9.

B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro. Exploring generalization in deep learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 5949–5958, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.