Understanding Deep Learning Requires Rethinking Generalization - An Overview

Universidad Nacional de Colombia, Sede Bogotá
Mathematics for Machine Learning

June, 2023

Paulina Castillo
ancastillov@unal.edu.co

Ricardo Marino
rmarino@unal.edu.co

Sebastian Molina
smolinad@unal.edu.co

1 Introduction

Deep artificial neural networks can generalize well despite having a large number of trainable parameters compared to the available training data. This observation raises the question of what factors contribute to the generalization performance of neural networks and how to design more principled and reliable model architectures. Statistical learning theory offers complexity measures such as VC dimension, Rademacher complexity, and uniform stability to control generalization error. However, this work challenges the traditional view of generalization by demonstrating that these measures fail to distinguish between neural networks with significantly different generalization performances.

Through a variant of randomization tests inspired by non-parametric statistics, it is shown that deep neural networks can fit random labels perfectly, indicating that the effective capacity of neural networks is sufficient to memorize the entire dataset. Even when trained on random labels, neural networks optimize efficiently, requiring only a small increase in training time compared to using true labels. The randomization of labels is found to be a transformative data operation, leaving other properties of the learning problem unchanged. Additionally, experiments with convolutional neural networks reveal their ability to fit random noise, deteriorating generalization as noise level increases.

Explicit forms of regularization, such as weight decay, dropout, and data augmentation, are investigated but found to be inadequate in explaining generalization error. While explicit regularization can improve generalization, its absence does not necessarily lead to poor performance. This contrast with classical convex empirical risk minimization, where regularization is essential to prevent trivial solutions, highlights the distinct role of regularization in deep learning.

The concept of finite sample expressivity is introduced, demonstrating that large neural networks can express any labeling of the training data. A simple theoretical construction of a two-layer ReLU network with parameter size $p = 2n+d$ showcases its ability to represent any labeling of a sample size n in d dimensions. This emphasizes the focus on neural network expressivity concerning finite samples rather than the entire function space.

Implicit regularization is explored by analyzing how stochastic gradient descent (SGD) acts as an implicit regularizer, implicitly constraining the norm of solutions. The effectiveness of SGD as a regularizer is demonstrated in linear models, suggesting the need for further investigation into the properties inherited by models trained using SGD.

2 Effective capacity of neural networks

The goal of this study is to explore the effective model capacity of feed-forward neural networks. Inspired by non-parametric randomization tests, the authors adopt a methodology to investigate the behavior of neural networks trained on datasets with random or corrupted labels. Surprisingly, the training process for various standard architectures remains largely unaffected by this label transformation, challenging previous assumptions about generalization. The authors experiment with different levels of randomization, including label and input randomization, using image classification datasets. They observe that even with random labels or shuffled pixels, the networks can fit the data well. The learning curves demonstrate that the networks converge quickly and achieve perfect fitting of the training set. The study also examines the impact of label corruption on convergence time and generalization error. These findings raise conceptual challenges for traditional approaches to reasoning about generalization, such as Rademacher complexity, VC dimension, and uniform stability. The results suggest that existing complexity measures and stability notions may not explain the observed generalization behavior in neural networks. Particularly, the complexity of Rademacher is explained below since it was a question that arose in the preparation of this summary and the team considered it important

2.1 Explanation: Rademacher Complexity

Consider training error for a learning model, *i.e.*,

$$\begin{aligned} E_{\text{in}}(h) &= \frac{1}{N}\sum_{i=1}^{N}[h(x_i) \neq f(x_i)] \\ &= \frac{1}{N}\sum_{i=1}^{N}\frac{1 - f(x_i)h(x_i)}{2} \\ &= \frac{1}{2} - \frac{1}{2N}\sum_{i=1}^{N} f(x_i)h(x_i). \end{aligned}$$

Observe that rewriting training error as before, minimizing training error corresponds to maximizing the correlation between the labels, $-f(x_n)-$, and the predictions, $-h(x_n)-$, *i.e.*, finding

$$\arg\max_{h} \frac{1}{N}\sum_{i=1}^{N} f(x_i)h(x_i).$$

Replacing the labels with Rademacher random variables $\sigma_n \in \{\pm 1\}$ gives us the Rademacher complexity, *i.e.*,

$$\widehat{\mathfrak{R}}_n(\mathcal{H}) = \mathbb{E}_\sigma\left[\sup_{h\in\mathcal{H}} \frac{1}{N}\sum_{i=1}^{N}\sigma_i h\left(x_i\right)\right]$$

As a trivial bound for this complexity is 1, and for the experiments this bound is approximately 1, this complexity measure does not give relevant information about generalization. Similar results come from considering VC dimension and uniform stability.

3 The role of regularization

This study aims to investigate the role of regularizers in deep learning and their impact on generalization performance. Regularizers are commonly used techniques to prevent overfitting by reducing the complexity of the hypothesis space. However, in deep learning, explicit regularization seems to have a different effect. The study compares the behavior of deep neural networks with and without regularizers by examining several commonly used network architectures.

The regularizers covered in the study include data augmentation, weight decay, and dropout. Data augmentation involves applying domain-specific transformations to the training set. On the other hand, weight decay aims to penalize solutions with a high l_2-norm. Lastly, dropout technique ignores certain randomly chosen layer outputs making the training process noisy. The experiments show that even with all regularizers turned off, the models still generalize well, indicating that other hidden factors play a significant role in generalization.

The results demonstrate that data augmentation, which leverages known symmetries in the data, has a more substantial impact on improving performance than weight decay, showing low training error. The study also explores implicit regularizations such as early stopping and batch normalization. Early stopping shows potential benefits in some cases, while batch normalization helps stabilize learning dynamics but has a limited impact on generalization performance.

4 Finite-sample expressivity

This study focuses on the expressive power of neural networks in the context of finite samples. While previous research has primarily examined the population-level expressivity of neural networks, this study argues that understanding their performance on finite samples is more relevant in practical applications.

To establish finite sample results, uniform convergence theorems are typically employed to transfer population-level findings. However, these theorems often require sample sizes that are polynomially large in the input dimension and exponentially large in the network's depth, which is unrealistic in practice.

Instead, this study directly analyzes the finite-sample expressivity of neural networks and finds that the picture becomes significantly simpler. Specifically, a simple two-layer neural network with ReLU activations can represent any function of the input sample. In other words, for every sample of size n in d dimensions and every function defined on that sample, there exists a neuronal network setting of weights that perfectly matches the function on the sample. The proof of this theorem is found within the paper, particularly in Appendix C.

5 Implicit regularization: an appeal to linear models

In this section, the authors explore the generalization capability of linear models and draw parallels to understanding the generalization of neural networks. They begin by considering the empirical risk minimization (ERM) problem for linear models, where the goal is to minimize a nonnegative loss function over a set of distinct data points. If the number of features is greater than or equal to the number of data points, it is possible to fit any labeling. However, the authors question whether such rich models can generalize well without explicit regularization.

To investigate the quality of global minima and their generalization performance, the authors examine the curvature of the loss function. In the case of linear models, all optimal solutions have the same curvature. They demonstrate this by showing the form of the Hessian matrix and its degeneracy at global optimal solutions.

Instead of relying on curvature, the authors suggest looking at the behavior of stochastic gradient descent (SGD) to understand which solution it converges to. By analyzing the update rule of SGD, they show that the solution lies in the span of the data points when the initial weight guess is set to zero. If perfect interpolation of labels is achieved, this reduces to a single equation involving the dot products between data points. This result corresponds to the *kernel trick* and shows that any set of labels can be perfectly fitted using the Gram matrix and solving a linear system.

The authors demonstrate the effectiveness of this approach on benchmark datasets such as MNIST and CIFAR10. Solving the linear system using the Gram matrix, achieves excellent performance without any preprocessing. The authors also observe that adding regularization does not improve the model's performance in these cases.

The authors highlight the interpretation of the kernel solution in terms of implicit regularization. They show that it corresponds to the minimum l_2-norm solution of the linear system. However, they caution that the minimum-norm solution is not predictive of generalization performance. They provide examples where solutions with larger norms than the one obtained from the *kernel trick*, generalized even better, indicating that the minimum-norm intuition is only a small part of the overall generalization story.

6 Conclusion

This work introduces an experimental framework to investigate the effective capacity of machine learning models. The experiments demonstrate that several successful neural network architectures possess a high effective capacity, enabling them to perfectly fit or shatter the training data. As a result, these models have the potential to memorize the entire training dataset. This poses a challenge to statistical learning theory, as conventional measures of model complexity struggle to explain the generalization ability of large neural networks.

The authors contend that a precise formal measure to characterize the simplicity of these extensive models has yet to be discovered. Furthermore, the experiments reveal an interesting observation: despite the models' failure to generalize, the optimization process remains empirically easy. This suggests that the reasons behind the ease of optimization differ from the fundamental cause of generalization.

References

[1] C. Zhang et al. "Understanding deep learning requires rethinking generalization". In: *International Conference on Learning Representations*. 2017. URL: `https://openreview.net/forum?id=Sy8gdB9xx`.