EECS 6322 Week 11, Paper 2
Understanding Deep Learning Requires Rethinking Generalization

Shivani Sheth (Student #: 218011783)
shivs29@yorku.ca

I. Summary and contributions

The paper studies the reasons behind the generalization obtained by large neural networks in practice. In contrast to the conventional belief that neural networks generalize well due to the regularization techniques or the properties of their model families, the paper proposes that a simple depth two neural network already has the perfect finite sample expressivity when the number of parameters increase than the number of data points. This generalization is quantitatively unaffected by explicit regularization and is present also if the true images are replaced by completely unstructured random noise. Finally, the experiments are also compared with traditional models to verify the findings.

The generalization error is usually seen as the difference between the training error and the test error. To control the generalization error, conventional theories based on statistical learning proposed different measures such as Rademacher complexity, uniform stability, and VC dimension. The theory also suggested that some form of regularization in terms of implicit or explicit regularization was required to reduce the generalization error. However, the randomization experiments performed by the authors prove that the traditional views of generalization were incapable of differentiating the neural networks with different generalization performance.

The randomization test basically suggests that deep neural networks easily fit random labels, hence achieving a high training accuracy in both true labels and random labels. The experiment was performed with several standard architectures on datasets such as CIFAR10 and ImageNet. The data was divided into two different sets where one set of data had true labels and the second set of data was a copy of the same dataset with randomized labels. The models when trained on the randomized labels achieved 0 training error but a very high test error. Hence, the overall generalization error of the model was high on randomized labels without changing the model, its size, hyperparameters, or the optimizer, which contradicts the conventional belief of generalization. Thus, the experiment concludes that the effective capacity of neural networks is sufficient for memorizing the entire data set and optimization on the random labels is easy. When the ratio of randomization was smoothly varied between no noise and complete noise, it was found that the generalization error also increased smoothly as the noise increased. The authors also prove that although explicit regularization improves performance, it is neither necessary nor sufficient to control the generalization error, and that a very simple two-layer ReLU network with $p = 2n + d$ parameters can express a given labeling of any sample of size n in d dimensions.

To further understand the effect of randomized labels on generalization, the authors performed a few experiments on the CIFAR10 and ImageNet datasets with the Inception V3 architecture on ImageNet and Alexnet/ MLPs on CIFAR10. A few modification on the labels and input images were applied that included partially corrected labels where the label of each image was corrupted as a uniform random class, random labels where all true labels were replaced with random labels, shuffled pixels where a random permutation of the pixels was chosen and then the same permutation was applied to all the images in both training and test set, random pixels where a different random permutation was applied to each image independently, and a Gaussian distribution which was used to generate random pixels for each image. The optimization results when stochastic gradient descent with unchanged hyperparameter settings was applied on each of these conditions showed that the weights fit to the random labels perfectly, even without any relationship between the images and labels. This suggests that the model is able to over fit the training data perfectly and achieve 0 error on the training data. On the CIFAR10 dataset, Alexnet and MLPs converged to zero loss on the training set and on the ImageNet dataset, the models were also able to achieve high training accuracies. However, with the increase in random noise, the generalization error for data with high noise, which is equal to the test error in this case since the training errors are 0, reached up to 90%, which is equal to random guessing.

The randomization experiment implies that the traditional approaches that provided measures to estimate generalization fail to explain the results in the randomized data experiments conducted by the authors. The Rademacher complexity and VC-dimension both provide a generalization bound that do not hold true for the randomized labels experiments and in general do not lead to useful generalization bounds in realistic settings. Additionally, the uniform stability method considers properties of the algorithm used for training but however, does not take into account specifics of the data or the distribution of the labels. Thus, this conventional method also fails to provide a useful generalization bound in realistic settings.

The authors also tested the role of regularization on the Inception V3 architecture on ImageNet and Alexnet/ MLPs on CIFAR10. In theory and practice, regularizers are a tool to mitigate overfitting in the regime when there are more parameters than data points. Their idea is that although the original hypothesis is too large to generalize well, regularizers help confine learning to a subset of the hypothesis space with manageable complexity. The experiments were conducted with the commonly-used regularizers in practice such as data augmentation, weight decay, and dropout. Data augmentation augments the training set via domain-specific transformations such as random cropping, random perturbation of brightness, saturation, hue and contrast in images. Weight decay is basically a l_2 regularizer on the weights, and dropout masks out each element of a layer output randomly with a given dropout probability. The results show that even with dropout and weight decay, InceptionV3 is able to fit the random training set very well on ImageNet, and on CIFAR10, both Inception and MLPs perfectly fit the random training set with weight decay turned on. However, AlexNet with weight decay turned on fails to converge on random labels. Thus, while regularization is important, they do not primarily contribute to improving the generalization ability of the models. Similarly, experiments with implicit regularization show that methods such as early stopping and batch normalization could potentially improve the generalization performance but they are not the fundamental reasons for generalization.

Finally, the authors conduct experiments to measure the expressive power of neural networks on a finite sample of size n. The experiments analyze the finite-sample expressivity of neural networks that prove that as soon as the number of parameters p of a network is greater than n, even simple two-layer neural networks with ReLU activations and $2n + d$ weights can represent any function of the input sample. In addition, the authors also propose a linear equation to reduce the generalization error in linear models given by $XX^T \alpha = y$ which resulted in very low generalization error when tested on CIFAR10 and ImageNet.

II. Strengths

The paper provides strong empirical results as to why the conventional methods used to measure the generalization power of neural networks fail to match the actual generalization error of neural networks used in practice. Through a series of experiments conducted on randomized labels in the CIFAR10 and ImageNet datasets, the authors also prove that contrary to the popular belief, regularizers help in improving the generalization error in a few cases, but are not the fundamental reasons for generalization. In addition, the authors also prove that simple two-layer neural networks with ReLU activations and $2n + d$ weights can represent any function of the input sample.

III. Weaknesses

Although the paper provides strong empirical results on why the existing methods to measure generalization fail to capture the actual generalization error of neural networks in practice, it does not provide alternative methods to accurately measure the generalization power of non-linear models such as neural networks.

IV. Correctness

The claims and empirical methodology of the work proposed by the authors is correct and is supported by thorough experiments on randomized labels in two datasets namely CIFAR10 and ImageNet. The experiments have also been conducted with the same architectures with various data preprocessing methods, all of which support the claims made by the authors. The authors also provide strong theoretical grounding to characterize the expressivity of neural networks.

V. Clarity

The paper is very written which thoroughly includes all the important details of the experiments, along with the illustrations and tables of the results reproduced. The conventional methods are also covered orderly and the effect of each method is clearly explained with respect to each dataset. Overall, the paper has good clarity and is easy to understand.

VI. Relation to prior work

Prior work in the representational power of neural networks include universal approximation theorems for multi-layer perceptrons that characterize which mathematical functions can be expressed by certain families of neural networks over the entire domain. On the contrary, the proposed work studies the representational power of neural networks for a finite sample of size n, which leads to a simple proof stating that even O(n) sized two-layer perceptrons have universal finite-sample expressivity. A few other works give an upper bound on the generalization error of a model trained with stochastic gradient descent in terms of the number of steps taken by gradient descent that are independent of the labeling of the training data. These methods fail to distinguish between models that have a low generalization error with true labels and the same models that have a high generalization error with randomized labels.

VII. Reproducibility

Sufficient details in terms of the experiments conducted, the architectures/ datasets used, and the preprocessing of the data required to conduct the experiments are provided by the authors which can supplement in reproducing the major details of the work.

VIII. Additional comments

An extended study to improve generalization from the linear models to the non-linear models would supplement the work proposed.