An analysis of the paper - Understanding Deep Learning Requires Rethinking Generalization by Zhang et al

Parth Shah

June 27, 2020

In 2012, AlexNet was able to beat all of its competitors by a 10 percent margin in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). AlexNet wrote a remarkable story. It combined all the pre-existing pieces and won over the trust and hopes of researchers on behalf of Deep Learning. It wasn't long before newer and better deep learning architectures like Inception V3 blew past AlexNet's level on the ImageNet benchmark. Deep learning is now used in many real-world applications that are seeing more human interactions every day. But as more and more people use these applications, many of its shortcomings have also surfaced. Most recently, these imperfections took the internet by a storm when a Deep Generative Model reconstructed a blurry picture of Barack Obama as a Caucasian male. In the backdrop of the pandemic, we have become oddly familiar with the failures of Deep Learning and AI systems in general.

Deep artificial neural networks(DNN) perform strikingly well on unseen data. Even with their vast depths, they demonstrate minuscule generalization errors, i.e., the difference between test error and training error. This generalization is essential to transfer their knowledge to newer, similar situations after being deployed in the wild. Overfitting (or memorization) would yield in the failure of these devices, which could lead to further complications down the pipeline. Hence, researchers and developers equipped with a better understanding of generalization could avoid any or all mishaps altogether. The paper titled "Understanding Deep Learning Requires Rethinking Generalization" by Zhang et al. is a critical analysis of traditional views on generalization.

A vital argument that DNNs must defend against is the memorization of data by these models. Low levels of the generalization error are often the primary justification provided in support of these models. The paper challenges this rhetoric of Deep Neural Networks having exceptional generalization capability through the conduction of randomization tests. While training on data with randomly assigned labels, DNNs surprisingly achieve zero training error and, understandably, a high test error. The authors claim that the DNNs easily fit random labels, implying the structures' memorization of data. Randomization also doesn't seem to affect the difficulty of optimization since the time taken for the architectures to attain these levels of error remains comparable to when training with true labels. With label randomization just being a data transformation, truly generalizing models (or algorithms for training) must fail to converge. However, the "success" of these models is indicative of their memorization and refutes their claims on generalization.

To ensure high generalization capacity, the importance of regularization increases with an increase in the parameter size. Implicit regularizations, like early stopping, and explicit regularizations, like dropout, weight-decay, are helpful techniques for enhancing generalization. However, the authors dispute the necessity and the sufficiency of these methods for non-convex optimizations, the ones that are generally tackled by the use of DNN. An extension of the randomization tests performed with and without regularizations still has a low training error, exposing that these regularizations do not force generalization. Further, as reported by Krizhevsky et al., weight decay sometimes even helps optimization. Another technique called BatchNorm, although not designed for regularization, is found to aid in generalization capacity. Architectures trained with BatchNorm are also able to attain low training errors.

Some measures proposed by statistical theory, like the Rademacher Complexity, VC dimension, and uniform stability, provide some reasons for generalization. However, these measures offer a trivial upper bound on the generalization of the occurrence of this phenomenon.

The paper further proves that a simple two-layer (or above) ReLU network with 2n+d parameters can express any labeling of n - d dimensional Random Samples. This goes to show the effective capacity of neural networks and their memorization capabilities.

Overall, this paper proves that Deep Neural Networks are capable of memorization of a dataset. Therefore, it demonstrates that DNNs learn random datasets by memorizing them, hence prompting the question of how Deep Neural Networks learn non-random semantic datasets. Further, time taken during randomization tests is more than that of training with accurate labels, indicating that these architectures can exploit semantic regularities, when present, and opt for memorization in the lack thereof. Krueger et al. build on these observations to claim that DNNs do not memorize; rather, they learn an available simple hypothesis that fits the finite random sample.