New insights into training dynamics of deep classifiers

March 8 2023

An MIT study takes a first theoretical analysis inside a convolutional neural network and provides new insights into how properties emerge during the network's training. Credit: Kris Brewer/CBMM

A new study from researchers at MIT and Brown University

characterizes several properties that emerge during the training of deep classifiers, a type of artificial neural network commonly used for classification tasks such as image classification, speech recognition, and natural language processing.

The paper, "Dynamics in Deep Classifiers trained with the Square Loss: Normalization, Low Rank, Neural Collapse and Generalization Bounds," published today in the journal *Research*, is the first of its kind to theoretically explore the dynamics of training deep classifiers with the square loss and how properties such as rank minimization, neural collapse, and dualities between the activation of neurons and the weights of the layers are intertwined.

In the study, the authors focused on two types of deep classifiers: fully connected deep networks and convolutional neural networks (CNNs).

A previous study had examined the structural properties that develop in large neural networks at the final stages of training. That study focused on the last layer of the network and found that deep networks trained to fit a training dataset will eventually reach a state known as "neural collapse." When neural collapse occurs, the network maps multiple examples of a particular class (such as images of cats) to a single template of that class. Ideally, the templates for each class should be as far apart from each other as possible, allowing the network to accurately classify new examples.

An MIT group based at the MIT Center for Brains, Minds and Machines studied the conditions under which networks can achieve neural collapse. Deep networks that have the three ingredients of stochastic gradient descent (SGD), weight decay regularization (WD), and weight normalization (WN) will display neural collapse if they are trained to fit their training data. The MIT group has taken a theoretical approach—as compared to the empirical approach of the earlier study—proving that

neural collapse emerges from the minimization of the square loss using SGD, WD, and WN.

Co-author and MIT McGovern Institute postdoc Akshay Rangamani states, "Our analysis shows that neural collapse emerges from the minimization of the square loss with highly expressive deep neural networks. It also highlights the key roles played by weight decay regularization and stochastic gradient descent in driving solutions towards neural collapse."

Weight decay is a regularization technique that prevents the network from over-fitting the training data by reducing the magnitude of the weights. Weight normalization scales the weight matrices of a network so that they have a similar scale. Low rank refers to a property of a matrix where it has a small number of non-zero singular values. Generalization bounds offer guarantees about the ability of a network to accurately predict new examples that it has not seen during training.

The authors found that the same theoretical observation that predicts a low-rank bias also predicts the existence of an intrinsic SGD noise in the weight matrices and in the output of the network. This noise is not generated by the randomness of the SGD algorithm but by an interesting dynamic trade-off between rank minimization and fitting of the data, which provides an intrinsic source of noise similar to what happens in dynamic systems in the chaotic regime. Such a random-like search may be beneficial for generalization because it may prevent over-fitting.

"Interestingly, this result validates the classical theory of generalization showing that traditional bounds are meaningful. It also provides a theoretical explanation for the superior performance in many tasks of sparse networks, such as CNNs, with respect to dense networks," comments co-author and MIT McGovern Institute postdoc Tomer Galanti. In fact, the authors prove new norm-based generalization

bounds for CNNs with localized kernels, that is a network with sparse connectivity in their weight matrices.

In this case, generalization can be orders of magnitude better than densely connected networks. This result validates the classical theory of generalization, showing that its bounds are meaningful, and goes against a number of recent papers expressing doubts about past approaches to generalization. It also provides a theoretical explanation for the superior performance of sparse networks, such as CNNs, with respect to dense networks. Thus far, the fact that CNNs and not dense networks represent the success story of deep networks has been almost completely ignored by machine learning theory. Instead, the theory presented here suggests that this is an important insight in why deep networks work as well as they do.

"This study provides one of the first theoretical analyses covering optimization, generalization, and approximation in deep networks and offers new insights into the properties that emerge during training," says co-author Tomaso Poggio, the Eugene McDermott Professor at the Department of Brain and Cognitive Sciences at MIT and co-director of the Center for Brains, Minds and Machines. "Our results have the potential to advance our understanding of why deep learning works as well as it does."

More information: Mengjia Xu et al, Dynamics in Deep Classifiers Trained with the Square Loss: Normalization, Low Rank, Neural Collapse, and Generalization Bounds, *Research* (2023). DOI: 10.34133/research.0024