Abstract

Machine learning using deep neural networks (DNNs) has been used in vast areas such as image recognition, object detection, and natural language processing. The success of DNNs has been brought by its ability to represent arbitrarily complex functions when a sufficiently large number of parameters are used and to extract essential features from large data sets. In most cases, however, DNNs are like a black box, and the reason for their high generalization performance is not fully understood. Recently, a statistical mechanical analysis based on the spin glass theory was performed to understand what is going on inside DNNs. By regarding the DNN as a spin system and examining the spin glass order parameters in each layer of the DNN, it was shown that the layers near the input and output are spin glassy (or solid like), while the central layers remain paramagnetic (or liquid like). Such a characteristic phase structure appears to play an important role in machine learning using DNNs, but further research is required to clarify the internal structure of DNNs in actual learning, since it is in general different from the equilibrium state in the statistical mechanical analysis.

In this thesis, we study the internal structures of DNNs based on the spin glass theory, by performing actual learning in a feed-forward neural network with back propagation. Specifically, we introduce "temperature" in the activation function in the DNNs. We first set the temperature to be uniform in all layers and tune its value. We find that the DNNs maximizes its performance in an intermediate temperature range, indicating that the temperature is a key quantity to enhance the network performance. We next allow the temperature to take different values for different layers and optimize them in the learning process. We find that the DNNs can achieve higher performance than the uniform temperature. In particular, the DNN shows the best performance when we optimize the temperature with an algorithm called label smoothing. We find that the distribution of temperature after learning is highly asymmetric with respect to the distances from the input and output layers; temperatures are relatively low in some layers near the output layer, while those in the other layers are close to the optimum value which maximizes the network performance in the case with the uniform temperature. In addition, by analyzing the spin glass order parameters, we find that the internal structure of DNNs is asymmetric because of the back propagation in the learning process, in contrast to the previous study. Our findings would pave the way for clarification of the mechanism of deep learning and a better design of DNNs.