CSIAM Trans. Appl. Math.
doi: 10.4208/csiam-am.SO-2024-0011

Vol. **6**, No. 2, pp. 272-319
June 2025

Understanding the Initial Condensation of Convolutional Neural Networks

Zhangchen Zhou[1,2], Hanxu Zhou[1,2], Yuqing Li[1,3,*] and Zhi-Qin John Xu[1,2,4,*]

[1] *School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai 200240, P.R. China.*
[2] *Institute of Natural Sciences, MOE-LSC, Shanghai Jiao Tong University, Shanghai 200240, P.R. China.*
[3] *CMA-Shanghai, Shanghai Jiao Tong University, Shanghai 200240, P.R. China.*
[4] *Key Laboratory of Marine Intelligent Equipment and System, Ministry of Education, Shanghai 200240, P.R. China.*

Received 2 April 2024; Accepted 10 January 2025

Abstract. Previous research has shown that fully-connected neural networks with small initialization and gradient-based training methods exhibit a phenomenon known as condensation [T. Luo *et al.*, J. Mach. Learn. Res., 22(1), 2021]. Condensation is a phenomenon wherein the weight vectors of neural networks concentrate on isolated orientations during the training process, and it is a feature in the non-linear learning process that enables neural networks to possess better generalization abilities. However, the impact of neural network architecture on this phenomenon remains a topic of inquiry. In this study, we turn our focus towards convolutional neural networks (CNNs) to investigate how their structural characteristics, in contrast to fully-connected networks, exert influence on the condensation phenomenon. We first demonstrate in theory that under gradient descent and the small initialization scheme, the convolutional kernels of a two-layer CNN condense towards a specific direction determined by the training samples within a given time period. Subsequently, we conduct systematic empirical investigations to substantiate our theory. Moreover, our empirical study showcases the persistence of condensation under broader conditions than those imposed in our theory. These insights collectively contribute to advancing our comprehension of the non-linear training dynamics inherent in CNNs.

AMS subject classifications: 68U99, 90C26, 34A45

Key words: Convolutional neural network, dynamical regime, condensation.

*Corresponding author. *Email addresses:* zczhou1115@sjtu.edu.cn (Z. Zhou), zhouhanxu@sjtu.edu.cn (H. Zhou), liyuqing_551@sjtu.edu.cn (Y. Li), xuzhiqin@sjtu.edu.cn (Z. Xu)

 ©2025 Global-Science Press

1 Introduction

As large neural networks continue to demonstrate impressive performance in numerous practical tasks, a key challenge has come to understand the reasons behind the strong generalization capabilities exhibited by overparameterized neural networks [4, 37]. Indeed, in overparameterized settings, there are many solutions that perform well on the training data, but most of them do not generalize well. Surprisingly, it seems that gradient based algorithms gravitate towards solutions that manifest superior generalization even in the absence of explicit regularization terms [31, 36], thereby challenging the attribution of the success of deep learning to explicit regularization. Consequently, it is believed that gradient-based algorithms induce an implicit bias [24] which prefers solutions that generalize well, and characterizing such bias has been a subject of extensive research. For instance, it has been established in [28, 34, 35] that neural networks (NNs) tend to fit target functions from low to high frequencies known as the frequency principle or spectral bias. Moreover, Luo *et al.* [20] observed a phenomenon wherein the input weights of hidden neurons in the two-layer rectified linear unit (ReLU) neural networks condense into isolated orientations during training with small initialization. Fig. 1 visually exemplifies this phenomenon, illustrating the reduction of a large condensed network to an effectively smaller network comprising only two neurons. As the generalization error can be bounded in terms of complexity [3], NNs with condensed parameters tend to possess better generalization abilities. This observation potentially provides valuable insights into the mechanisms through which overparameterized neural networks achieve good generalization performance in practice. In a more practical setting, [39] shows that

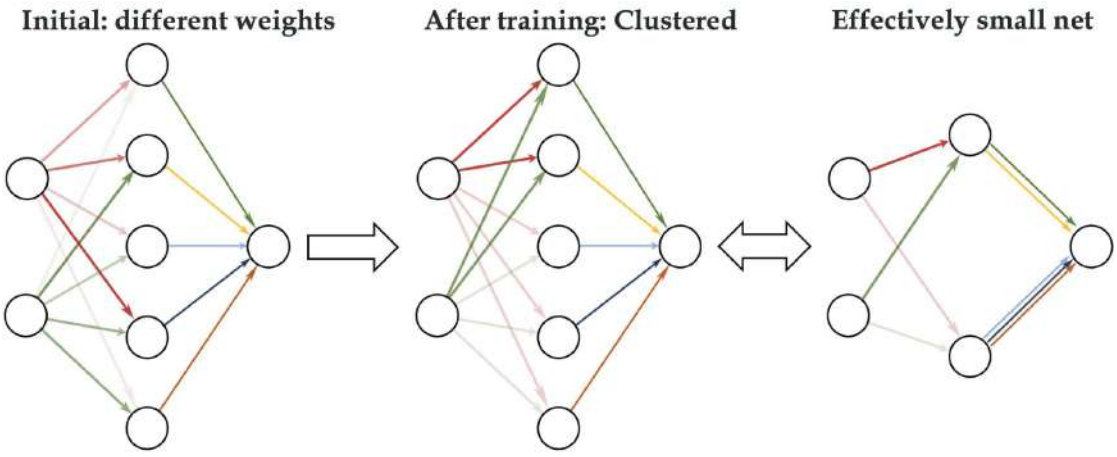

Figure 1: Illustration of condensation. The color and the intensity of a line indicate the strength of the weight. Initially, weights are random. Soon after training, the weights from an input node to all hidden neurons are clustered into two groups, i.e. condensation. Multiple hidden neurons can be replaced by an effective neuron with low complexity, which has the same input weight as the original hidden neurons the same output weight as the summation of all output weights of the original hidden neurons.