SCIENCE CHINA
Information Sciences

• PERSPECTIVE •

January 2021, Vol. 64 116101:1–116101:3
https://doi.org/10.1007/s11432-020-2885-6

Why over-parameterization of deep neural networks does not overfit?

Zhi-Hua ZHOU

National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023*, China*

Received 11 April 2020/Accepted 15 April 2020/Published online 14 September 2020

Citation Zhou Z-H. Why over-parameterization of deep neural networks does not overfit? Sci China Inf Sci, 2021, 64(1): 116101, https://doi.org/10.1007/s11432-020-2885-6

Deep neural networks often come with a huge number of parameters, even larger than the number of training examples, but it seems that these over-parameterized models have not suffered from overfitting. This is quite strange and *why over-parameterization does not overfit*? poses a fundamental question concerning the mysteries behind the success of deep neural networks.

In conventional machine learning theory, let $\mathcal{H}$ denote the hypothesis space, m is the size of a training set with i.i.d. samples, then the gap between the generalization error and empirical error is often bounded by $O(\sqrt{|\mathcal{H}|/m})$ where $|\mathcal{H}|$ is about the hypothesis space complexity. If the whole hypothesis space represented by a deep neural network is considered, then the numerator grows with the parameter count (depth $\times$ width), which can be even larger than the denominator, leading to vacuous bounds. Thus, many studies resorted to consider *relevant* subset of hypothesis space, e.g., by introducing implicit bias depending on specific algorithms such as the norms controlled by stochastic gradient descent (SGD) [1, 2]. The results, however, were not that satisfactory and recently there were even claims that conventional learning theory could not be used to explain generalization of deep neural networks even if the implicit bias of specific algorithms had been taken into account to the fullest extent possible [3].

Although many arguments may have their own groundings, we feel that an important fact should be noticed; that is, conventional learning theory concerns mostly about the training of a learner, or more specifically, a classifier in classification tasks, from a feature space, but concerns little about the construction of the feature space itself. Therefore, conventional learning theory can be exploited to understand the behavior of generalization, but one must be careful when it is applied to *representation learning*.

It is well-known that deep neural networks accomplish end-to-end learning through integrating feature learning with classifier training. As illustrated in Figure 1(a), a deep neural network can be decomposed into two parts, where the first part devotes to *feature space transformation*, i.e., converting the original feature space represented by the input layer to the final feature space represented by the final representation layer, in which a classifier is constructed.

First, let's focus on the classifier construction (CC) part in Figure 1(a), where the number of parameters depends on the number of units in the final representation layer. It is well-known that there occurs overfitting if the number of parameters is more than needed [4, 5]. For example, Figure 1(b) depicts a typical training-testing performance plot based on results presented in [4], which shows that the testing performance degrades although the training performance increases as the number of parameters becomes too large. This exhibits clearly that for the CC part in Figure 1(a), over-parameterization can lead to overfitting; this confirms with what conventional learning theory tells.

Next, let's focus on the feature space transformation (FST) part in Figure 1(a). Most doubts about the incapability of conventional learning theory on deep neural networks actually come from the fact that there seems no overfitting even when the parameter count (depth $\times$ width) is very large [1–3]. Here, we want to point out that the parameters of the FST part should not be simply regarded as parameters of the hypothesis space concerned by conventional learning theory. In fact, when we say that conventional learning theory tells us over-parameterization will lead to overfitting, the parameters refer to those concerning the hypothesis space, such as the parameters of the CC part; as for parameters for feature space transformation, conventional learning theory does not claim anything. Indeed, the connection between overfitting and the parameterization of feature space transformation has rarely been theoretically studied before, and thus, there is no theory concludes that over-parameterization of feature space transformation will lead to overfitting; this applies to not only deep neural networks but also other feature space transformation techniques.

Take distance metric learning [6] for example. It is able to transform the original feature space to a "better" feature space in which a relatively simple classifier can solve a problem that can be hard in the original feature space, just like the FST part in deep neural networks transforming the feature space such that an originally complicated task can be addressed by a simple fully-connected linear layer.

Email: zhouzh@nju.edu.cn

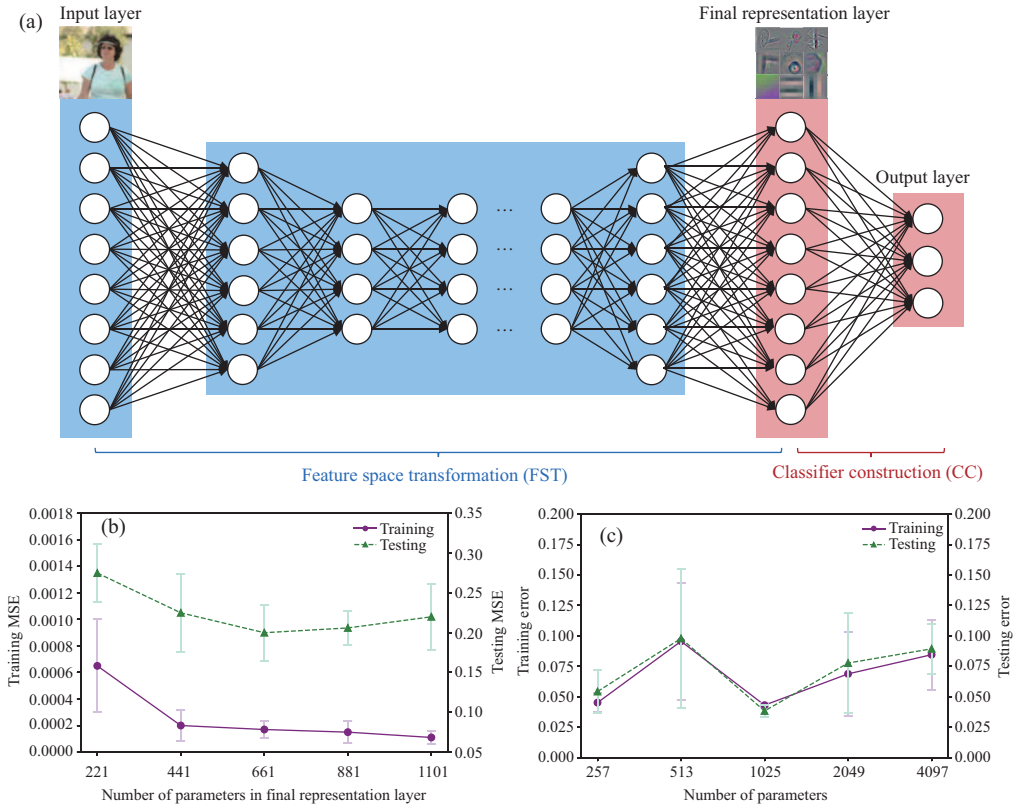

Figure 1 (Color online) (a) A decompositional view of deep neural networks; (b) a typical performance plot showing that over-parameterization of the CC part can lead to overfitting (replot based on experimental results presented in [4]); (c) a typical performance plot which shows that over-parameterization of FST does not necessarily lead to overfitting.

Given a training set $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)\}$, let $\mathcal{M}$ and $\mathcal{C}$ denote the must-link and cannot-link constraints, respectively. These constraints can be extracted from the data; for example, a pair of examples falling into the same class leads to the extraction of a must-link constraint, and otherwise a cannot-link constraint. The well-known distance metric learning algorithm [7] attempts to solve the following problem:

$$\begin{aligned} \min_{\boldsymbol{M}} \quad & d_{\mathrm{LD}}(\boldsymbol{M}, \boldsymbol{I}) \\ \text{s.t.} \quad & \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_{\boldsymbol{M}}^2 \leqslant u, \quad \forall (\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{M}, \qquad (1) \\ & \|\boldsymbol{x}_i - \boldsymbol{x}_k\|_{\boldsymbol{M}}^2 \geqslant l, \quad \forall (\boldsymbol{x}_i, \boldsymbol{x}_k) \in \mathcal{C}, \qquad (2) \\ & \boldsymbol{M} \succeq 0, \end{aligned}$$

where $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_{\boldsymbol{M}}^2$ is the Mahalanobis distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ based on the positive semi-definite metric matrix $\boldsymbol{M}$; $d_{\mathrm{LD}}(\boldsymbol{M}, \boldsymbol{I})$ is the log-det divergence between $\boldsymbol{M}$ and identity matrix $\boldsymbol{I}$; Eq. (1) demands the must-link examples to be close, with pairwise distances smaller than u; Eq. (2) demands the cannot-link examples to be faraway, with pairwise distances larger than l. Assuming that strong duality holds and considering the Lagrange dual form, the solution is

$$\begin{aligned} \boldsymbol{M}^{-1} = & \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{M}} \lambda_{ij} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} \\ & - \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_k) \in \mathcal{C}} \mu_{ik} (\boldsymbol{x}_i - \boldsymbol{x}_k)(\boldsymbol{x}_i - \boldsymbol{x}_k)^{\mathrm{T}} \\ & + (1 - \alpha) \boldsymbol{I}, \end{aligned}$$

where $\alpha \in \mathbb{R}$ is the dual parameter corresponding to the positive semi-definite constraint, $\lambda_{ij} \geqslant 0$ and $\mu_{ik} \geqslant 0$ are parameters in the dual form corresponding to the must-link and cannot-link constraints, respectively. It is evident that the number of parameters can be large, even larger than the number of training examples, by simply extracting more constraints from the training data. Using a linear classifier in the transformed feature space, Figure 1(c) presents a typical performance plot which shows that the influence of the number of parameters on the testing performance is non-monotonic. More importantly, this breast-cancer data set has 450 training examples (and 233 testing examples) whereas the numbers of parameters showing in Figure 1(c) are much larger; it is observable that the training and testing curves are quite consistent, implying that overfitting does not occur even when the number of parameters is larger than the number of training examples.

In summary, we want to indicate that when conventional learning theory concludes that over-parameterization leads to overfitting, the parameters concerned are about hypothesis space from which the classifiers are constructed; in deep neural networks such parameters are those of the CC Part in Figure 1(a). As for FST parameters, there was no such claim; this applies to not only deep neural networks but also other feature space transformation techniques. Thus, an important future direction is to rigorously study the influence of the number of parameters on the performance of feature space transformation, ideally by establishing learning theory about feature space transformation, and this may shed light on further understanding of mysteries behind deep neural networks.

Acknowledgements This work was supported by National Natural Science Foundation of China (NSFC) (Grant Nos. 61751306, 61921006). The author wants to thank Shen-Huan LYU and Zhi-Hao TAN for discussion and help in figures.

References

1 Neyshabur B, Tomioka R, Srebro N. Norm-based capacity control in neural networks. In: Proceedings of the 28th Conference on Learing Theory, Paris, 2015. 1376–1401

2 Zhang C Y, Bengio S, Hardt M, et al. Understanding deep learning requires rethinking generalization. In: Proceedings of the 5th International Conference on Learning Representation, Toulon, 2017

3 Nagarajan V, Kolter J Z. Uniform convergence may be unable to explain generalization in deep learning. In: Proceedins of Advances in Neural Information Processing Systems, 2019. 11615–11626

4 Lawrence S, Giles C L, Tsoi A C. Lessons in neural network training: overfitting may be harder than expected. In: Proceedings of the 14th National Conference on Artificial Intelligence, Providence, 1997. 540–545

5 Liu Y Y, Starzyk J A, Zhu Z. Optimized approximation algorithm in neural networks without overfitting. IEEE Trans Neural Netw, 2008, 19: 983–995

6 Kulis B. Metric learning: a survey. Found Trends Mach Learn, 2013, 5: 287–363

7 Davis J V, Kulis B, Jain P, et al. Information-theoretic metric learning. In: Proceedings of the 24th International Conference on Machine Learning, Corvalis, 2007. 209–216