【Grant-in-Aid for Scientific Research (S)】

Theory and Practice of Deep Learning Based on Fisher Information Matrix and MDL Principle

Principal Investigator	Kyushu University, Faculty of Information Science and Electrical Engineering, Professor Takeuchi Jun'ichi Researcher Number:80432871
Project Information	Project Number : 23H05492 Project Period (FY) : 2023-2027 Keywords : deep learning, neural networks, AI, MDL, data science

Purpose and Background of the Research

●Outline of the Research

We perform a fundamental research for rapidly developed artificial intelligence (AI). Nowadays most technologies called AI are based on deep learning, whose performance is superior to the other machine learning techniques, but its nature has many unclear parts. Hence, its development mainly relies on empirical approaches, and it is hard to guarantee the reliability of AI. In this research, we focus on the over fitting problem described below, aiming to establish the reliability of deep learning. Deep learning is a kind of machine learning using neural networks, which are a mathematical model of brains. Brains have learning ability and can obtain various functions. Deep neural networks, which are used in deep learning, have many layers and very much complex. (Fig. 1) There are various models for machine learning other than neural networks, and many achievements have been obtained since 1980's. Among such studies, over-fitting had been one of main subjects by the rising of deep learning around 2015. Over-fitting is a situation that a learning agent cannot handle new input data because the training is too adjusted to the training data, It is likely to occur if the model is too complex, but if the model's complexity is low, then it cannot bring out the information in the data. Hence, we should choose the model with appropriate complexity. (See Fig. 2.) Here, the complexity of the model is measured with the number of parameters.

The first quantitative solution to this problem is Akaike's information criterion (AIC; 1973), which gave great influence to statistics and machine learning. The minimum description (MDL) principle is an information criterion influenced by AIC. These theories can be though of as quantitative implementation of Occam's razor, which says `simple explanations of a given phenomenon are to be preferred over complex ones.' The way of this thought, which had been a common sense of machine learning, does not work now. In fact, deep neural networks are very much complex in the viewpoint of information criteria. On the contrary, it is common to say, the more complex the model is, the better it is, This situation means that the theory of machine learning faces the problem why such complex models show high performance. The primal purpose of this study is to solve the problem.

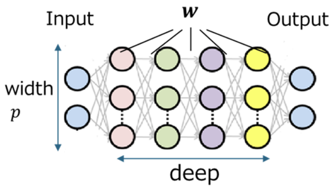

Fig. 1. Deep Neural Network

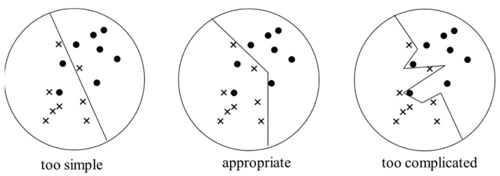

Fig. 2. Concept of Model Selection

[1] Y. Takeishi, M. Iida, and J. Takeuchi, "Approximate spectral decomposition of Fisher information matrix for simple ReLU networks," *Neural Networks*, Volume 164, July 2023, Pages 691-706.

●Approximate Eigenvalue Decomposition of Fisher information matrix

We investigated the Fisher information of two-layer neural networks, which are the minimum unit of deep neural networks. We can know important directions in the parameter space to learn by eigenvalue decomposition of Fisher information matrix. We obtained a form of approximate eigenvalue decomposition, which shows that a small number of directions ($O(d^2)$. The d is in Fig.3) are important. (Fig. 4)

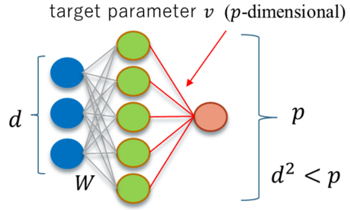

Fig. 3. Two-layer Neural Networks

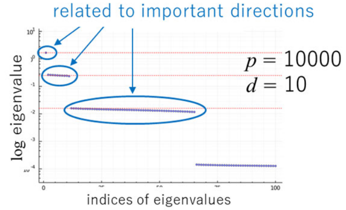

Fig. 4. Approximate Eigenvalue Decomposition

Expected Research Achievements

●Road Map to the Goal（Fig. 5)

If it is sufficient to learn only $O(d^2)$ directions, the neural network is much simpler than it seems, independently of size of p. Our first goal is to guarantee the generalization error depending only on d^2. (Task 1)

For the two-layer case, since the Fisher information is constant, the problem is simple. It does not hold for the general case, but it approximately holds, owing to the theory of neural tangent kernel. Armed with this fact, we aim to extend the performance guarantee results for two-layer neural networks, and to show the lower bounds of generalization error (Task 4). Besides these theoretical issues, we will cope with the tasks shown in Fig. 6.

(Task 1) Analysis of 2-layer neural networks
(Task 2) Intuitive understanding of double decent phenomina
Based on approximate spectral decomposition
(Task 3) Study of relation with natural gradient method
Interpretation of biased eigenvalue distribution
(Task 4) Extenstion ot general deep neural networks
(Task 5) Design of learning methods
(Task 6) Applications to real problems
- Data analysis for cyber security
- Image reconstruction for MRI
- Design of decoder for sparse superposition code

Fig. 6. List of individual tasks

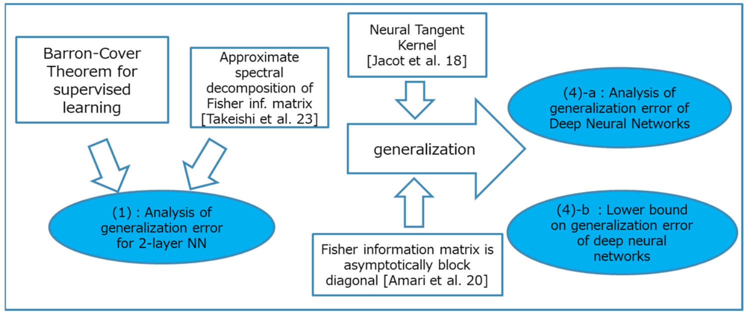

Fig. 5. Road map to the solution to the unsolved problems for deep learning

generalization error: error for the data not used in the training
Barron-Cover Theorem: giving performance guarantee for MDL estimators
double decent: The phenomenon that the generalization error decreases again after the peak caused by over-fitting, when increasing the mode complexity.

Homepage Address, etc.	http://www.me.inf.kyushu-u.ac.jp/%7etak/homepage.en.html