Deep Learning Model Complexity: Concepts and Approaches

Xia Hu* Lingyang Chu† Jian Pei* Jiang Bian‡ Weiqing Liu‡

Abstract

In this tutorial on deep learning model complexity, we will provide an overview of the model complexity problem including motivation, historical notes, technical challenges, and fundamental properties. We will discuss two fundamental questions: model expressive capacity, and effective model complexity. We will connect model complexity with other important problems (e.g., generalization) to illustrate how model complexity can help tackle these problems. We will discuss some interesting and promising future directions for model complexity.

Keywords: Model complexity, Expressive capacity, Deep neural network, Deep learning

1 Introduction

Deep learning is disruptive in many applications mainly due to its superior performance. At the same time, many fundamental questions about deep learning remain unanswered. Model complexity of deep neural networks is one of them. Model complexity is concerned about how complicated a problem that a deep model can express and how nonlinear and complex the function of a model with given parameters can be.

In machine learning, data mining and deep learning, model complexity is always an important fundamental problem. Model complexity affects learnability of models on specific problems and data, as well as generalization ability of the model on unseen data. Moreover, the complexity of a learned model is affected not only by the model architecture itself, but also by the data distribution, data complexity, and information volume. In recent years, model complexity has become a more and more active direction, and has developed theoretical guiding significance in many areas, such as model architecture searching, graph representation, generalization study and model compression.

We propose this tutorial to overview the state-of-the-art research on deep learning model complexity. We summarize the model complexity studies into two directions: model expressive capacity and effective model complexity, and review the latest progress on these two directions.

- **Model expressive capacity.** Expressive capacity captures the capacity of deep models in expressing complex problems. Approaches for classical model expressive capacity study (e.g., VC dimension) are with limited applicability on deep learning models due to the complicated structure of deep models. Recent works explore this problem from the perspective of the effectiveness of network depth [2, 12, 13], the effectiveness of network width [11], and the expressible functional space of deep models [5, 8].
- **Effective model complexity.** Effective complexity reflects the practical, effective complexity of the functions of deep models with given parameters. Exploring effective model complexity calls for feasible effective complexity measures [7, 16]. Besides, a series of studies find that even with overparameterized architecture and high expressive capacity, the effective complexity of a learned neural network may still be much lower than the expressive capacity [6].

We will discuss the applications of model complexity in generalization, optimization and others to demonstrate the usefulness of model complexity. We will conclude this tutorial and discuss several interesting and promising future directions.

2 Tutorial Outline

This tutorial will be organized into five sessions. Please visit our tutorial website for detailed tutorial materials: `http://sfu.ca/~huxiah/sdm21_tutorial`.

In the first section, we will provide an overview of the model complexity problem including motivation, historical notes, technical challenges and fundamental properties.

In the second section, we will discuss the expressive capacity of deep architectures. That is, what functions and problems can be expressed, and how architectures affect the expressive capacity.

In the third section, we will discuss the effective model complexity. That is, what complexity of a model with given parameters can be.

*Simon Fraser University (huxiah@sfu.ca, jpei@cs.sfu.ca)

†McMaster University (chul9@mcmaster.ca)

‡Microsoft Research ({Jiang.Bian, Weiqing.Liu}@microsoft.com)

In the fourth section, we will connect model complexity with several important issues: generalization capability, optimization, model selection, to show how model complexity helps tackle these problems.

In the last section, we will summarize and discuss some open problems and potential future directions.

The tentative outline and schedule of the 2-hour tutorial are shown below.

1. Introduction and Overview (10 mins)
1.1 Motivation
1.2 Model complexity
1.3 Historical notes on model complexity
2. Expressive Capacity (30 mins)
2.1 Effectiveness of model depth
2.2 Effectiveness of model width
2.3 Expressible functional spaces
3. Effective Model Complexity (30 mins)
3.1 General measures of effective complexity
3.2 High-capacity low-reality phenomenon
4. Application Examples of Deep Learning Model Complexity (30 mins)
4.1 Model complexity in understanding generalization
4.2 Model complexity in optimization
4.3 Model complexity in model selection and design
5. Summary and future directions (20 mins)

3 Target Audience

The tutorial will be largely self-contained. We only assume that the audience understands the basic concepts of neural networks. We will introduce the advanced concepts and use concrete examples to explain them.

This tutorial mainly targets at three groups of audience. First, the researchers and graduate students who are interested in understanding deep learning models and model complexity will learn the frontier of this new direction. Second, our tutorial is also attractive to the researchers who are interested in applying machine learning theory and techniques to optimize data analysis and processing, and provides them a new angle from model complexity. Last, the industry practitioners and data scientists who are interested in the general intuition and ideas about model complexity will find practical guidelines from this tutorial on model design and selection from the model complexity perspective. Through the tutorial, the audience can quickly understand the fundamental ideas, the latest progress, the major challenges and the research opportunities about model complexity.

4 Tutors' Biographies

Xia Hu is currently a Ph.D. Candidate at the School of Computing Science, Simon Fraser University, Canada. Her research interests lie in deep learning, machine learning and data mining, with an emphasis on the interpretability, explainability and model complexity. She is currently focusing on exploring the explanatory representation of complex models, as well as the interpretation and analysis of internal mechanisms of network models (e.g., model complexity). She is also interested in various applications of deep learning in solving data mining problems.

Lingyang Chu is an Assistant Professor at the Department of Computing and Software at McMaster University. Before joining McMaster University, he was a principal researcher at Huawei Technologies Canada, where he led a research team focused on deep neural network interpretation, AI fairness, federated learning and big data analytics. During 2015-2018, he worked as a postdoctoral fellow at Simon Fraser University. He received the doctoral degree from the University of Chinese Academy of Sciences in 2015. His research works in large scale graph mining and interpretable AI have been published in top-tier venues. One of his works on interpretable AI was reported by a mainstream web portal of AI research in China in 2018.

Jian Pei is a Professor in the School of Computing Science and an associate member of the Department of Statistics and Actuarial Science at Simon Fraser University, Canada. His general areas include data science, big data, data mining, and database systems. His expertise is in developing effective and efficient data analysis techniques for novel data intensive applications. He is recognized as a Fellow of the Royal Society of Canada (i.e., the national academy of Canada), the Canadian Academy of Engineering, ACM and IEEE.

Jian Pei is a productive and influential author in data mining, database systems, and information retrieval. Since 2000, he has published one textbook, two monographs, and over 200 research papers in refereed journals and conferences, which have been cited over 100,000 times in literature, and over 41,000 times in the last 5 years. His research has generated remarkable impacts substantially beyond academia. His algorithms have been adopted by industry in production and popular open-source software suites. He is responsible for several commercial systems of unprecedentedly large scale. He received many prestigious awards, such as the 2017 ACM SIGKDD Innovation Award, the 2015 ACM SIGKDD Service Award, the 2014 IEEE ICDM Research Contributions Award, a KDD Best Application Paper Award (2008), and an IEEE ICDE Influential Paper Award (2019).

Jiang Bian is a Principal Researcher and Research

Manager at Microsoft Research with research interests in AI for finance, AI for logistics, business AI, deep learning, multi-agent reinforcement learning, computational advertising, and a variety of machine learning applications. Prior to that, he was a senior scientist, leading the recommendation and search modeling, at Yidian Inc., a startup on content-oriented content delivery platform. He also used to work at Yahoo! Labs as a Scientist and did a lot of studies on content optimization and personalization for the Yahoo! key content modules. He has authored tens of research papers published at several well-recognized AI-related conferences with thousands of citations, such as KDD, ICDE, AAAI, NIPS and ICML. He has been served as Program Committee Member/Peer Reviewer of many influential academic conferences and journals.

Weiqing Liu is a Senior Researcher at Microsoft Research. He holds a Ph.D. degree in the Department of Computer Science from the University of Science and Technology of China. His research interests focus on data mining and machine learning. He is actively transferring research to significant real-world applications, especially to finance scenarios. Currently, one of his research focuses is on the common critical challenge of applying AI into the finance area, especially the interpretability issue of machine learning models in application scenarios. His work has led to tens of research papers in prestigious conferences, such as KDD, WWW, WSDM, AAAI and IJCAI.

References

[1] R. Arora, A. Basu, P. Mianjy, and A. Mukherjee, *Understanding deep neural networks with rectified linear units*, arXiv preprint arXiv:1611.01491, (2016).

[2] Y. Bengio and O. Delalleau, *On the expressive power of deep architectures*, in International Conference on Algorithmic Learning Theory, Springer, 2011, pp. 18–36.

[3] M. Bianchini and F. Scarselli, *On the complexity of neural network classifiers: A comparison between shallow and deep architectures*, IEEE transactions on neural networks and learning systems, 25 (2014), pp. 1553–1565.

[4] O. Delalleau and Y. Bengio, *Shallow vs. deep sum-product networks*, in Advances in Neural Information Processing Systems, 2011, pp. 666–674.

[5] I. Gühring, G. Kutyniok, and P. Petersen, *Complexity bounds for approximations with deep relu neural networks in sobolev norms*, (2019).

[6] B. Hanin and D. Rolnick, *Complexity of linear regions in deep networks*, arXiv preprint arXiv:1901.09021, (2019).

[7] X. Hu, W. Liu, J. Bian, and J. Pei, *Measuring model complexity of neural networks with curve activation functions*, in Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 1521–1531.

[8] V. Khrulkov, A. Novikov, and I. Oseledets, *Expressive power of recurrent neural networks*, arXiv preprint arXiv:1711.00811, (2017).

[9] J. Kileel, M. Trager, and J. Bruna, *On the expressive power of deep polynomial neural networks*, in Advances in Neural Information Processing Systems, 2019, pp. 10310–10319.

[10] T. Liang, T. Poggio, A. Rakhlin, and J. Stokes, *Fisher-rao metric, geometry, and complexity of neural networks*, arXiv preprint arXiv:1711.01530, (2017).

[11] Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang, *The expressive power of neural networks: A view from the width*, in Advances in neural information processing systems, 2017, pp. 6231–6239.

[12] H. Mhaskar, Q. Liao, and T. Poggio, *Learning functions: when is deep better than shallow*, arXiv preprint arXiv:1603.00988, (2016).

[13] G. F. Montufar, R. Pascanu, K. Cho, and Y. Bengio, *On the number of linear regions of deep neural networks*, in Advances in neural information processing systems, 2014, pp. 2924–2932.

[14] B. Neyshabur, S. Bhojanapalli, D. McAllester, and N. Srebro, *Exploring generalization in deep learning*, in Advances in Neural Information Processing Systems, 2017, pp. 5947–5956.

[15] R. Novak, Y. Bahri, D. A. Abolafia, J. Pennington, and J. Sohl-Dickstein, *Sensitivity and generalization in neural networks: an empirical study*, in International Conference on Learning Representations, 2018.

[16] M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. S. Dickstein, *On the expressive power of deep neural networks*, in Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR, 2017, pp. 2847–2854.