A Conversation with Tengyu Ma: Developing mathematical tools to understand deep learning systems

INNOVATION LEADERS

by Bushra Anjum

Editor's Introduction

In this interview, Ubiquity's senior editor Dr. Bushra Anjum chats with Dr. Tengyu Ma, an assistant professor of Computer Science and Statistics at Stanford University. They discuss Dr. Ma's research that aims to bridge the gap between theory and practice in deep learning by developing novel mathematical tools to understand complex and mysterious deep learning systems.

A Conversation with Tengyu Ma: Developing mathematical tools to understand deep learning systems

INNOVATION LEADERS

by Bushra Anjum

Tengyu Ma is an assistant professor of computer science and statistics at Stanford University. He received his Ph.D. from Princeton University and B.E. from Tsinghua University. His research interests include topics in machine learning and algorithms, such as deep learning and its theory, non-convex optimization, deep reinforcement learning, representation learning, and high-dimensional statistics. He is a recipient of NIPS'16 best student paper award, COLT'18 best paper award, ACM Doctoral Dissertation Award Honorable Mention, Sloan Fellowship, and NSF CAREER Award.

What is your big concern about the future of computing to which you are dedicating yourself?

Deep learning, the subarea of machine learning on neural networks, is the backbone of many extraordinary recent advances in artificial intelligence. However, we still lack sufficient foundational understanding of how it works that can guide new practical methods. Past breakthroughs in deep learning have largely relied on large-scale experimental explorations. Mathematical insights are limited, and mysteries are abundant. Classical machine learning theory oftentimes can neither explain those surprising phenomena observed in deep learning nor provide sufficient inspiration for practical improvements.

I believe that theory should go hand in hand with experimentation to boost and sustain the rapid progress of deep learning. My long-term career goal is to contribute in fundamental ways to the advancement of deep learning and other areas of artificial intelligence. I aim to develop theories that can inspire efficient and reliable algorithms.

Optimization and generalization, the two pillars of machine learning, both posit theoretical challenges in the deep learning era. First, how do we design and analyze optimization algorithms for the loss functions in deep learning? Past optimization theory oftentimes focuses on

minimizing convex functions efficiently, but deep learning uses loss functions that are not convex. Second, when, and why do neural networks generalize to unseen test examples, that is, when they can reliably work on data that are not present in the training dataset?

Surprisingly, empirical findings have shown that these questions are intertwined: the generalization power of neural networks likely stems from the particular properties of the optimizers that are used to train them in the first place. However, rigorous understanding is still largely missing. Addressing these questions theoretically will make neural networks more reliable and interpretable.

Besides these two core questions, I'm also broadly interested in the interaction of neural networks with other areas of machine learning, such as reinforcement learning and unsupervised learning. On a more mathematical level, the common barrier to these questions is the notorious nonlinearities in neural networks. I dedicate myself to developing new mathematical tools for analyzing nonlinearities.

How did you become interested in developing tools for analyzing optimization in deep learning?

I started my Ph.D. in 2012 at Princeton University under the supervision of Prof. Sanjeev Arora. Sanjeev's research agenda deeply influenced me—he was thinking to design and analyze algorithms for realistic problem instances, instead of for worst-case instances. Focusing on realistic instances allows us to go beyond the intractability results and to achieve customized but stronger performance guarantees.

I worked with Sanjeev on the analysis of machine learning, which is an almost perfect area for instantiating Sanjeev's agenda—the computational problems in machine learning are often NP-hard in the worst-case, but abundant structures in the data exist and can be leveraged by the algorithms.

Towards developing tools for analyzing optimization in deep learning, I started to work on nonconvex optimization for classical machine learning problems such as sparse coding, matrix completion, and tensor decomposition. With these works and many other works around the same time, the community started to form the following hypothesis for explaining the success of optimizing nonconvex functions. Even though the practically used objective functions are not convex, we hypothesize that most local minima are approximately global minima. Therefore, gradient descent can find a global minimum even for these nonconvex problems. The community

was able to prove the hypothesis for many questions in classical machine learning. This early success motivated my future studies of nonconvex optimization in deep learning.

During my Ph.D., deep learning started to take off, with many empirical breakthroughs, and gradually I started to work more broadly on deep learning theory beyond the optimization aspect, including the generalization theory of neural networks, the theory for unsupervised deep learning, etc. I paid close attention to the empirical progress of deep learning so that I can find high-impact theoretical questions for my research.

What are some of your recent research projects focused on developing better understanding of deep learning systems?

With my group members at Stanford University, I am working on a project which aims to build a comprehensive generalization theory for deep neural networks. The project consists of three components that are increasingly more challenging. First, we aim to develop core mathematical tools to characterize the generalization of neural nets in the standard supervised setting where the training and test scenarios are the same. Second, we aim to address the so-called out-of-domain setting where the test scenarios are different from the idealized training environments. The lack of out-of-domain generalization has become a bottleneck for deploying machine learning systems in the real world. Third, the project aims to develop methods for numerically estimating the generalization performance accurately. In many risk-sensitive ML applications such as healthcare, we are in dire need of such numerical bounds before deploying the model in a new environment, e.g., deploying an ML model for medical diagnosis in a new hospital requires an estimate of the false positive and false negative rates.

Recently, machine learning is also undergoing a paradigm shift with the rise of large-scale models such as BERT and GPT-3. They are trained on large-scale unlabeled dataset and can be adapted to a wide range of downstream tasks. This is stronger than the generalization to an unseen data distribution because the models can be adapted to even different tasks with different goals. My research group has been working on extending the deep learning theory to this new paradigm as well.

Another thrust of my group's research is the theoretical study of deep reinforcement learning, a promising approach with wide applications in games, robotics, and healthcare. My group members and I designed novel algorithms with improved empirical performance and provided theoretical insights and guarantees. A central goal of the theory is to use statistical tools to design efficient algorithms that require fewer trial-and-error in the real world.

Biography

Bushra Anjum, Ph.D., is a health IT data specialist currently working as the Senior Analytics Manager at the San Francisco based health tech firm Doximity. Aimed at creating HIPAA secure tools for clinicians, she leads a team of analysts, scientists, and engineers working on product and client-facing analytics. Formerly a Fulbright scholar from Pakistan, Dr. Anjum served in academia (both in Pakistan and the USA) for many years before joining the tech industry. A keen enthusiast of promoting diversity in the STEM fields, her volunteer activities, among others, involve being a senior editor for ACM Ubiquity and the Standing Committee's Chair for ACM-W global leadership. She can be contacted via the contact page bushraanjum.info/contact or via Twitter [@DrBushraAnjum](https://twitter.com/DrBushraAnjum).

DOI: 10.1145/3486624