On Exact Computation with an Infinitely Wide Neural Net

Sanjeev Arora[1,2] Simon S. Du[2] Wei Hu[1]
Zhiyuan Li[1] Ruslan Salakhutdinov[3] Ruosong Wang[3]

[1]Princeton University
[2]Institute for Advanced Study
[3]Carnegie Mellon University

Introduction

- Recent papers [Li and Liang, Du et al., Allen-Zhu et al., Zou et al.] suggested that NNs with sufficiently large width can achieve 0 training error via gradient descent.

- [Jacot et al.] showed that as one increases the width to infinity, a certain limiting behavior, called neural tangent kernel (NTK), can emerge.

Questions we studied

- 1. Can we formally show that the prediction of NNs is equivalent to that of NTKs when width is sufficiently large?

- 2. How does NTK perform?

Theoretical Contribution

Theorem (Arora, Du, Hu, Li, Salakhutdinov, Wang, NeurIPS 2019): When width is sufficiently large (polynomial in number of data, depth and the inverse of target accuracy ε), the predictor learned by applying gradient descent on a neural network is ε-close to the kernel regression predictor of the corresponding neural tangent kernel.

Experimental Contribution

Dynamic programming-based algorithms for calculating NTKs for CNNs (CNTKs) + efficient GPU implementations.

Depth	CNN-V	CNTK-V	CNN-GAP	CNTK-GAP
3	59.97%	64.47%	63.81%	70.47%
4	60.20%	65.52%	80.93%	75.93%
6	64.11%	66.03%	83.75%	76.73%
11	69.48%	65.90%	82.92%	77.43%
21	75.57%	64.09%	83.30%	77.08%

Future Directions

- Understand the design of neural network architectures and common techniques in deep learning, e.g., batch normalization and residual layers, from the lens of neural tangent kernel.

- Combine NTK with other techniques in kernel methods to further improve the performance.

Thanks!

- Full paper: https://arxiv.org/abs/1904.11955
- Code: https://github.com/ruosongwang/CNTK