Backward Feature Correction: How Deep Learning Performs Deep (Hierarchical) Learning

Zeyuan Allen-Zhu ZEYUANALLENZHU@META.COM
Meta FAIR Labs
Yuanzhi Li YUANZHI.LI@MBZUAI.AC.AE
Mohamed bin Zayed University of AI

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

Deep learning is also known as hierarchical learning,[1] where the learner *learns* to represent a complicated target function by decomposing it into a sequence of simpler functions to reduce sample and time complexity. This paper formally analyzes how multi-layer neural networks can perform such hierarchical learning *efficiently* and *automatically* by applying stochastic gradient descent (SGD) or its variants on the training objective.

On the conceptual side, we present a theoretical characterizations of how certain types of deep (i.e. super-constantly many layers) neural networks can still be sample and time efficiently trained on some hierarchical learning tasks, when no existing algorithm (including layerwise training, kernel method, etc) is known to be efficient. We establish a new principle called "backward feature correction", where *the errors in the lower-level features can be automatically corrected when training together with the higher-level layers*. We believe this is a key behind how deep learning is performing deep (hierarchical) learning, as opposed to layerwise learning or simulating some known non-hierarchical method.

On the technical side, we show for every input dimension $d > 0$, there is a concept class of degree $\omega(1)$ multi-variate polynomials so that, using $\omega(1)$-layer neural networks as learners, a variant of SGD can learn any function from this class in $\mathsf{poly}(d)$ time to any $\frac{1}{\mathsf{poly}(d)}$ error, through learning to represent it as a composition of $\omega(1)$ layers of quadratic functions using "backward feature correction". In contrast, we do not know any other simpler algorithm (including layerwise training, applying kernel method sequentially, training a two-layer network, etc) that can learn this concept class in $\mathsf{poly}(d)$ time even to any $d^{-0.01}$ error. As a side result, we prove $d^{\omega(1)}$ lower bounds for several non-hierarchical learners, including any kernel methods, neural tangent or neural compositional kernels.

[2]

References

Zeyuan Allen-Zhu and Yuanzhi Li. Backward feature correction: How deep learning performs deep (hierarchical) learning. *ArXiv e-prints*, abs/2001.04413, 2021. Full version available at http://arxiv.org/abs/2001.04413.

Yoshua Bengio. *Learning deep architectures for AI*. Now Publishers Inc, 2009.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[1] Quoting Bengio (2009), "deep learning methods aim at *learning feature hierarchies* with features from higher levels of the hierarchy formed by the composition of lower level features." Quoting Goodfellow et al. (2016) "the hierarchy of concepts allows the computer to learn complicated concepts by *building them out of simpler ones*."

[2] Extended abstract. Full version appears as Allen-Zhu and Li (2021)[v6]. Most of the work was done when Z.A. was at Microsoft Research Redmond.