The Expressive Power of Tuning Only the Normalization Layers

Angeliki Giannou[*] GIANNOU@WISC.EDU
Shashank Rajput[*] RAJPUT.SHASHANK11@GMAIL.COM
Dimitris Papailiopoulos DIMITRIS@PAPAIL.IO
University of Wisconsin-Madison

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

Feature normalization transforms such as Batch and Layer-Normalization have become indispensable ingredients of state-of-the-art deep neural networks. Recent studies on fine-tuning large pretrained models indicate that just tuning the parameters of these affine transforms can achieve high accuracy for downstream tasks. These findings open the questions about the expressive power of tuning the normalization layers of frozen networks. In this work, we take the first step towards this question and show that for random ReLU networks, fine-tuning only its normalization layers can reconstruct any target network that is $O(\sqrt{\text{width}})$ times smaller. We show that this holds even for randomly sparsified networks, under sufficient overparameterization, in agreement with prior empirical work.

In particular, we prove that any given neural network can be perfectly reconstructed by only tuning the normalization layers of a wider, or deeper random network that contains only a factor of $\widetilde{\mathcal{O}}(\sqrt{d})$ more parameters (including both trainable and random).

Theorem 1. [*Informal*] *Let g be any fully connected neural network with l layers and width d. Then, any randomly initialized fully connected network f with l' layers and width d', with normalization layers can exactly recover the network g functionally, by tuning only the normalization layer parameters, as long as $d'l' \geq 2d^2l$, $d' \geq d$ and $l' \geq 2l$. Further, if f has sparse weight matrices, then the total number of parameters (trainable and random) only needs to be a factor of $\widetilde{\mathcal{O}}(\sqrt{d})$ larger than the target network.*

Our results show that the expressive power of scaling and shifting transformations of random features is indeed non-trivial. More specifically, our results can be decomposed into three cases for reconstructing any target network of width d and depth L:
Case 1. Reconstruction by a random network of width d^2 and depth $2L$.
Case 2. Reconstruction by a random network with skip connections of width d^2/k and depth $2kL$ for any $k = 1, \dots, d$.
Case 3. Reconstruction by a random network, sparsified with probability $\Theta(\sqrt{\log d/d})$, of width d^2 and depth $poly(d)$.
The first two results hold with probability 1, while the last one with probability at least $1 - 1/d$. Our findings are aligned to previous experimental study from Frankle et al. (2020).

The proofs of the theorems for each case rely crucially on the invertibility of Khatri-Rao products of random (possibly sparse) matrices. Due to the complex structure of the Khatri-Rao product, refined probabilistic arguments are required in the case of reconstruction by a deep neural network or a sparsified one.

Keywords: Normalization Layers, Khatri-Rao Product, Parameter Efficient Fine-tuning

[*] Equal contribution, listed alphabetically.
[1] Extended abstract. Full version appears as [arxiv:2302.07937].

Acknowledgements

DP acknowledges the support of an NSF CAREER Award #1844951, a Sony Faculty Innovation Award, an AFOSR & AFRL Center of Excellence Award FA9550-18-1-0166, an NSF TRIPODS Award #1740707, and an ONR Grant No. N00014- 21-1-2806. SR acknowledges the support of a Google PhD Fellowship award.

References

Jonathan Frankle, David J Schwab, and Ari S Morcos. Training batchnorm and only batchnorm: On the expressive power of random features in cnns. *arXiv preprint arXiv:2003.00152*, 2020.