Funneled Energy Landscapes in Deep Neural Networks

ABSTRACT

Abstract

ACM Reference Format:

. 2020. Funneled Energy Landscapes in Deep Neural Networks. In *Proceedings of ACM Conference (Conference'17)*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Over the past decade, Deep Neural Networks (DNNs) have proven remarkably effective on a wide range of computer vision (CV), natual language processing (NLP), and other domains. Moreover, larger and deeper DNN models, with hundreds to thousands of layers, perform tasks seemingly impossible just a few years ago. For example, the CV architecture ResNet has been successfuly trained with over 1000 layers, showing excellent generqlization performance on a wide range of data sets (CIFAR10, CIFAR100, SVHN, ImagaeNet, etc. Most recently, openAI released the NLP Language model GPT3, which has been trained on nearly a half trillion words, using 175 billion parameters, and achieving state-of-the-art (SOTA) performance on several NLP benchmarks.

The incredible size and depth of these models poses a new and deep theoretical challenges. [blah blah blah] Discuss Energy Landscape and ruggedly convexity

We do have some insight into how the Energy Landscape behaves by visualizing 2-dimensional cross-sections of small models during training, such as ResNet25. -summarize findings

Norm-based metrics such as WeightWatcher

Cross-Section is not a generalization metric, is not global

In order to characterize the Energy Landscape, traditional approaches attempt to count the number of local minima (i.e the complexity). And while this is well for theoretical analysis (such as spin glass theory, random matric theory, etc), numerically this is quite hard. Especially for the massive production size DNNs in use today.

Here, we suggest an new, alternative approach–to study the Empirical Spectral Density (ESD) of the data-dependent Jacobian, which is readily calculated with a single epoch of Backprop using any off-the-shelf toolkit such as TensorFlow, PyTorch, etc.

Similar to the weightwatcher studies..

Show picture: compare relatively random / flat vs a deeply funneled convex Landscape ESDs

random: real world data, randomly labeled

Here is a summary of our main results:

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00

https://doi.org/10.1145/nnnnnnn.nnnnnnn

-
-
-

Organization of this paper.

2 CONCLUSION

CONCLUSION

A APPENDIX

In this appendix, we provide more details on several issues that are important for the reproducibility of our results.