Loss Landscape Visualization

Report for the Seminar on Optimization and Neural Architecture Search

28. Januar 2021

Abstract

Neural networks have emerged as powerful function approximators with large parameter sets. These parameters are optimized according to a loss function. Many assumptions have been made about the shape of the resulting loss landscape. However, only recently qualitative and empricial studies have been conducted. Here, we give an overview for recent advances of this field. We will explore different methods in detail and discuss their results and impact.

Introduction

When creating neural networks for a given task, practicioners or have to take many decisions. From architecture design, over optimizers, and schedules to hyperparameter choice there are many options that will play a key role in the eventual performance of the model. Anecdotal knowledge, experience and luck often lead to the final configuration and there is little empirical knowledge of what is actually effective. Loss landscape visualization can play a large role in guiding the community towards a more empirically foundation and maybe help build the groundwork for theoretical approaches.
Several recent works have shown that a better knowledge about the loss landscape can help build methods grounded in theory [Mutschler and Zell, 2020, Chaudhari et al., 2019]. Moreover, loss landscape visualization also helped to explain the effectiveness of existing methods such as stochastic gradient descent (SGD) [Robbins and Monro, 1951, Xing et al., 2018] and architectures with residual connections [He et al., 2016, Li et al., 2017].
As the space of parameters is too large to visualize in any meaningful way, methods of loss landscape visualization have to compromise. They might just consider small, visualizable subspace of a model's parameters. Results, thus, should be considered with caution and any conclusions should be drawn only carefully.

Background

Methods

Linear Interpolation

Filter Normalization

[Xing et al., 2018]

Results

Conclusions

Literatur

[Chaudhari et al., 2019] Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L., and Zecchina, R. (2019). Entropy-sgd: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124018.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[Li et al., 2017] Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. (2017). Visualizing the loss landscape of neural nets. *arXiv preprint arXiv:1712.09913*.

[Mutschler and Zell, 2020] Mutschler, M. and Zell, A. (2020). Parabolic approximation line search for dnns. *arXiv preprint arXiv:1903.11991*.

[Robbins and Monro, 1951] Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

[Xing et al., 2018] Xing, C., Arpit, D., Tsirigotis, C., and Bengio, Y. (2018). A walk with sgd. *arXiv preprint arXiv:1802.08770*.