Paper Review

Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs

Daniel May

daniel.may.20@ucl.ac.uk

UCL

March 11, 2021

1 Overview

Training neural networks involves minimizing a loss function that is complex and highly multi-modal. The loss surfaces are influenced by factors including the choice of architecture, optimizer, and hyper-parameters such as batch size or learning rate, and remain the subject of active research. Their structure has implications for the generalization of neural network models, and can hint at more powerful methods for training them.

Prior study of loss surfaces can be divided into two areas. The first concerns their global structure; for example, Goodfellow et al. (2015) provide evidence that the trajectories of neural networks trained with stochastic gradient descent (SGD) do not seem to be hindered by the many local optima or saddle points on their training path. The second area concentrates on the local structure of the loss surface at minima found by SGD; e.g. Keskar et al. (2017) argue that small-batch converges to a flatter region of the loss than large-batch SGD, which has been linked with better test dataset generalization.

We review Garipov et al. (2018), which focuses on the latter area. As loss along a line segment connecting optima tends to increase greatly, intuition suggested that local optima found by training the same DNN architecture multiple times would be isolated in the loss surface, as in the left panel of Figure 1. In fact, they show that it is possible to find paths connecting them, with only a single bend, along which the loss is near-constant. Freeman and Bruna (2017) had showed that for a simple neural network, there existed a connecting curve along which the loss can be upper-bounded, but the result did not extend to the multi-layer case. They introduced a dynamic programming algorithm to find such curves, but their test accuracy was relatively low for tasks such as CIFAR-10 image classification.

In contrast, this paper proposes a training procedure for finding a simple path connecting two local optima that has high test accuracy, using state-of-the-art architectures. Beginning with two sets of parameters w_1 and w_2 found by independently training a DNN, we can parameterize a path $\phi_\theta(t)$, such as a quadratic Bezier curve or polygonal chain, mapping from $[0, 1]$ to the parameter space, with $\phi_\theta(0) = w_1$ and $\phi_\theta(1) = w_2$. To find a path between the optima, we can minimize the expectation of the loss along the path over a uniform distribution on $t \in [0, 1]$, where $L(\phi_\theta(t))$ is the loss function used for training the networks:

$$l(\theta) = \int_0^1 L(\phi_\theta(t))\, dt = E_{t\sim U(0,1)} L(\phi_\theta(t)).$$

The loss $l(\theta)$ can be minimized using SGD by sampling $\tilde{t}$ uniformly from $[0, 1]$, computing the gradient of $L(\phi_\theta(\tilde{t}))$ with respect to θ, and repeating until convergence. Figure 1 (middle and right) exemplifies the mode connectivity of a loss surface of ResNet-164 on CIFAR-100.

They demonstrate that points along the path correspond to meaningfully different parameterizations: in an experiment using ResNet-164 for CIFAR-100, they take an ensemble of two networks parameterized by $\phi_\theta(0) = w_1$ and points $\phi_\theta(t)$, and find that test error started to decrease around $t \approx 0.1$, and was as low as an ensemble of w_1 and w_2 by $t \geq 0.4$, so that $\phi_\theta(t)$ must correspond to a distinct representation, after only a small step in parameter space.

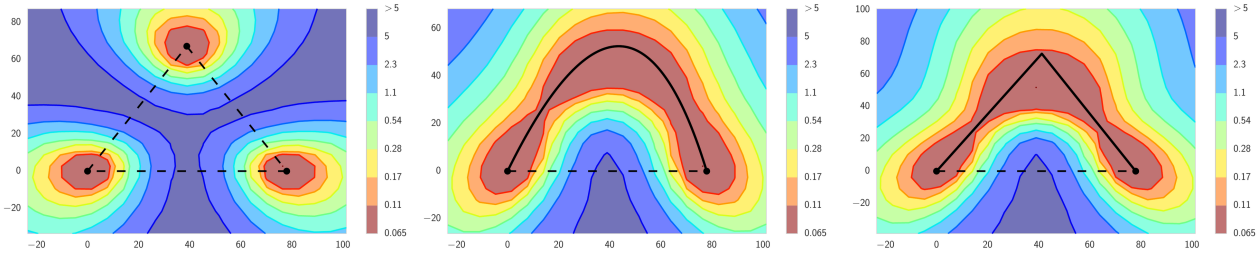

Figure 1: Images from Garipov et al. (2018) showing the loss surface of ResNet-164 on CIFAR-100. **Left**: Isolated optima of independently trained networks; **Middle** and **Right**: Two optima connected by a Bezier curve (middle) and polygonal chain (right).

Inspired by their finding, they introduce a technique called *Fast Geometric Ensembling* (FGE), which can discover many meaningfully distinct models in the time taken to train a single network. Starting with one set of weights $\hat{w}$ from a near-fully trained network, FGE begins a cyclic learning rate schedule with a small cycle length, which allows it to explore and exploit the parameter space, with relatively small steps, and accumulate a number meaningfully distinct models by storing the parameters when the learning rate is at its minimum.

On CIFAR-100, FGE consistently outperformed SnapShot Ensembles (SSE), a state-of-the-art ensembling technique by Huang et al. (2017), and ensembles of independently trained networks, for the same training budget. FGE is similar to SSE, using a cyclic learning weight to collect diverse models, but SSE collects models from the beginning of training, and uses a cycle length of 20-40 epochs, while FGE uses 2-4, based on their finding that it is possible to encounter distinct models by a short step in parameter space.

2 Strengths and Weaknesses

The paper is strong in a number of areas, while perhaps limited or incomplete in others. In terms of strengths: firstly, mode connectivity seems surprising, and verifies that there exist relatively flat regions of the loss which contain multiple distinct solutions, which may help to explain why they may generalize better, as per Keskar et al. (2017), and their mathematical procedure and visualizations (as in Figure 1) are intuitive. Secondly, FGEs demonstrate that studying loss surfaces is practically important, as they consistently outperform SSEs and ensembles of independently trained networks for the same computational budget on a number of image classification tasks and a next word prediction task, using recent DNN architectures. Thirdly, the paper provides a number of clear and exciting avenues for future work, particularly within Bayesian deep learning, and using deep ensembles that explore multiple modes as an alternative to explicitly Bayesian approaches, as in Fort et al. (2019).

There are a few areas in which the paper is not entirely clear. The first concerns mode connectivity, which is perhaps a surprising property. As such, more discussion about *why* high-accuracy paths between modes exist would be especially interesting. They recognize Li et al. (2018), which finds that residual networks are more regular than classic architec-

tures, which may also go some way to explaining the existence of these paths, and it has been proposed that flatter loss regions generalize better, and the idea that these regions contain diverse solutions could help explain this. Secondly, some questions remain about these paths: is there always a path between any two modes? If not, what does it depend on (e.g. choice of architecture)? How many such paths are there? Can we expect FGE to traverse paths such as those found by the connection procedure? Finally, the experiments are largely limited to image classification tasks and a small selection of architectures, and it would be great to see how FGE performs in other scenarios, including whether there are settings where it consistently performs worse for the same training budget.

3 Follow ups

As discussed in Section 2, the paper leaves open a number of questions about mode connectivity, such as whether it is always possible to low loss paths with only a single bend between any two modes, how many such paths might exist, and whether FGE will traverse these paths. These could be addressed empirically, for example by running multiple curve finding experiments with the same two modes and checking whether they are equivalent, or by demonstrating counterexamples where it is not possible to find a low loss path. It would also be interesting to investigate whether mode connectivity applies more generally to a wider range of architectures and tasks, and how these affect the loss of the paths that are found. Gotmare et al. (2018) evaluate the robustness of mode connectivity, showing that the curve finding procedure can connect modes trained with different initializations, optimizers, and hyper-parameters including learning rate schemes, batch sizes, and regularization.

Wilson (2020) uses mode connectivity to motivate a view of deep ensembles as performing approximate Bayesian inference; by collecting a diversity of models through exploring multiple modes, they may form a Bayesian model average which performs better than recent explicitly Bayesian approaches which focus on a single mode. It would be interesting to research new ensembling methods, or to extend current Bayesian approaches to encourage exploring multiple modes, by traversing these flat loss regions. For example, Zhang et al. (2020) propose a cyclical stepsize schedule within stochastic gradient MCMC for this purpose, and Maddox et al. (2019) uses a high constant learning rate schedule to form a Gaussian distribution over DNN weights. Finally, the paper proposes adapting the curve finding procedure to find curves with particularly notable characteristics, perhaps by experimenting with different loss functions, which may lead to additional new training methods.

4 Summary

Garipov et al. (2018) contribute the idea of mode connectivity to our understanding of loss surfaces, and introduce a practical DNN ensembling method which is fast and performant, based on their insight. Their account of mode connectivity is not fully comprehensive, but is an exciting area for future research, especially in assessing its robustness to a range of tasks and architectures, and to motivate other powerful ensembling and Bayesian methods.

References

Fort, S., Hu, H., and Lakshminarayanan, B. (2019). Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*. pages 3

Freeman, C. D. and Bruna, J. (2017). Topology and geometry of half-rectified network optimization. pages 2

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D., and Wilson, A. G. (2018). Loss surfaces, mode connectivity, and fast ensembling of dnns. *arXiv preprint arXiv:1802.10026*. pages 2, 3, 4

Goodfellow, I. J., Vinyals, O., and Saxe, A. M. (2015). Qualitatively characterizing neural network optimization problems. pages 2

Gotmare, A., Keskar, N. S., Xiong, C., and Socher, R. (2018). A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation. *arXiv preprint arXiv:1810.13243*. pages 4

Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J. E., and Weinberger, K. Q. (2017). Snapshot ensembles: Train 1, get m for free. pages 3

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. (2017). On large-batch training for deep learning: Generalization gap and sharp minima. pages 2, 3

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. (2018). Visualizing the loss landscape of neural nets. pages 3

Maddox, W., Garipov, T., Izmailov, P., Vetrov, D., and Wilson, A. G. (2019). A simple baseline for bayesian uncertainty in deep learning. pages 4

Wilson, A. G. (2020). The case for bayesian deep learning. pages 4

Zhang, R., Li, C., Zhang, J., Chen, C., and Wilson, A. G. (2020). Cyclical stochastic gradient mcmc for bayesian deep learning. pages 4