Large Scale Structure of Neural Network Loss Landscapes

Stanislav Fort*
Google Research, Zurich
stanislav.fort@gmail.com

Stanisław Jastrzębski+
New York University, USA
staszek.jastrzebski@gmail.com

arXiv 1906.04724

*This work has been done as a part of the Google AI Residency

+A part of this work was done while an intern in Google Zurich

Introduction

Optimization of deep neural networks is still relatively poorly understood. **Despite the high dimension** of the weight space loss landscape, the optimization behavior shows many **surprisingly simple features**.

Surprising observations

1) No significant obstacles along the way.
[1] show that there are no significant obstacles along a path from initialization to an optimum.
→ **long directions**

2) Solutions everywhere and dense.
[2] and [3] show that constraining optimization to random, low-dimensional sections of the weight space is sufficient for good optimization.
→ **solution manifold is high-dimensional**

3) Optima are connected by tunnels.
[4] and [5] show that independently optimized optima are connected by low-loss tunnels. Those are virtually impossible to find at random.
→ **the solution manifold is sponge-like**

Our contribution

1) Integrating these observations into a **unified phenomenological model of the neural network loss landscape**

2) Constructing the model *explicitly* in TensorFlow and **replicating all experiments** on it.

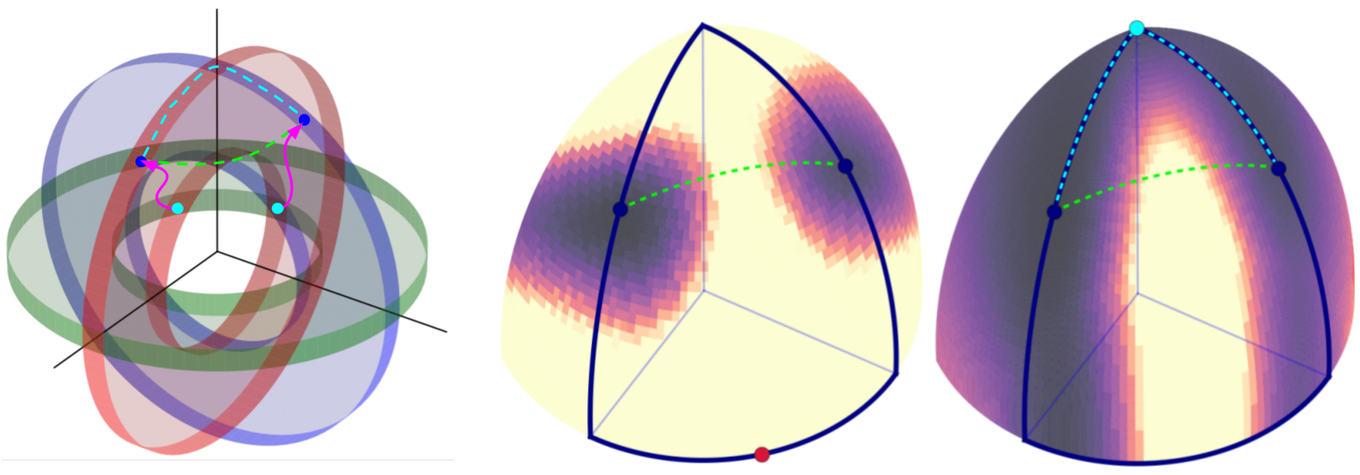
3) Extending the notion of a tunnel between 2 optima to **m-tunnels between m optima**.

4) Explicit **connection to the intrinsic dimension** of the loss landscape.

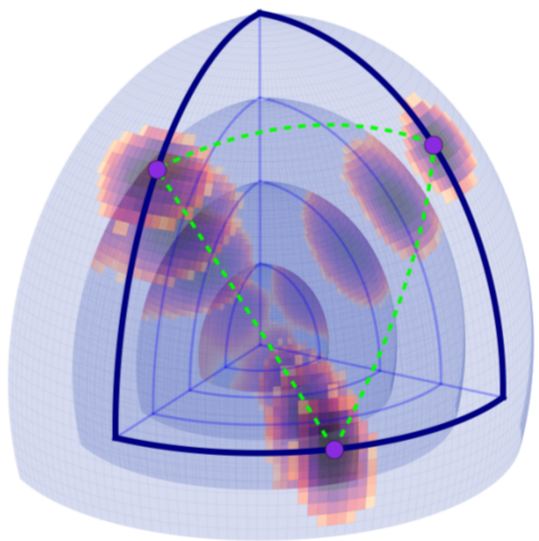
5) We investigate **consequences for modern ensembling schemes**.

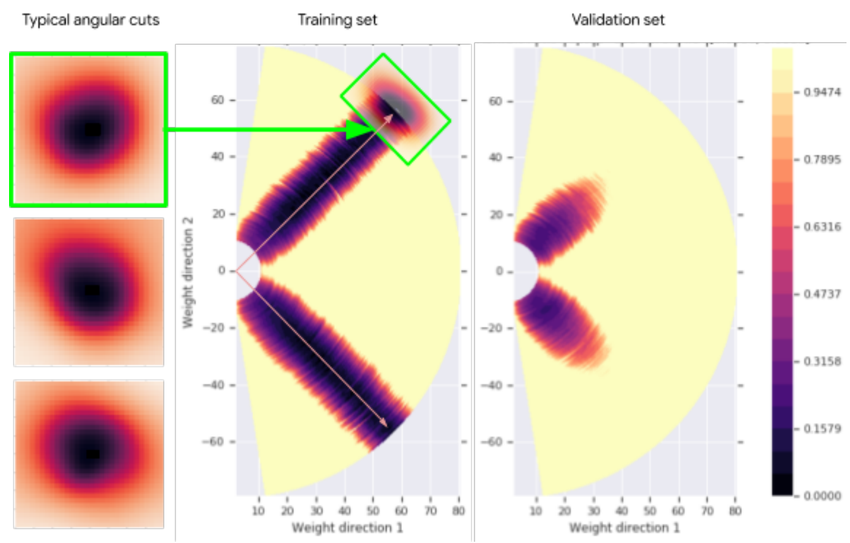
Our landscape model

The low-loss manifold of the **D**-dimensional weight space is made of high-dimensional **n-wedges**, where **n** is the number of their **long directions** and D ≅ n. The number of their short directions **s = D – n**.



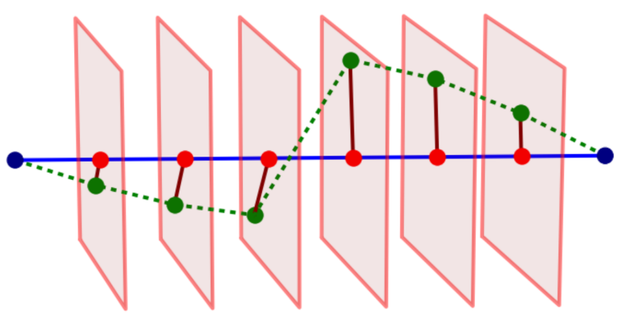
When observed along random directions, the **n-wedges appear like radial tunnels** due to the high dimension of the space. In fact, their **true nature is *extremely difficult* to show on any random landscape sections**.

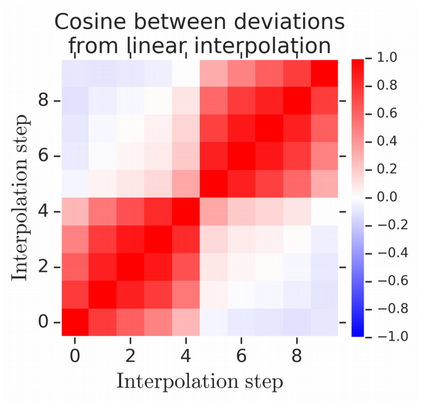


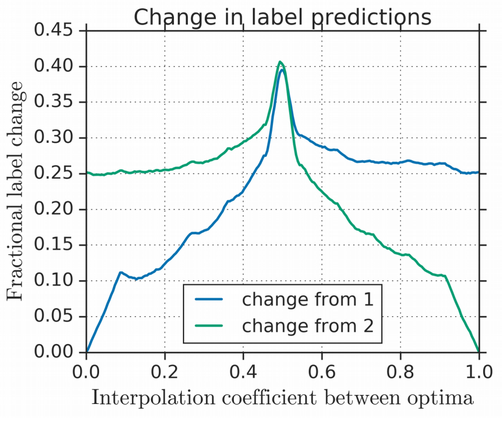


Tunnels and generalized *m*-tunnels

We generalize the notion of a low-loss tunnel between 2 optima to an ***m*-tunnel between m optima at once.** E.g. 3-tunnel connects 3 optima by a deformed sheet.

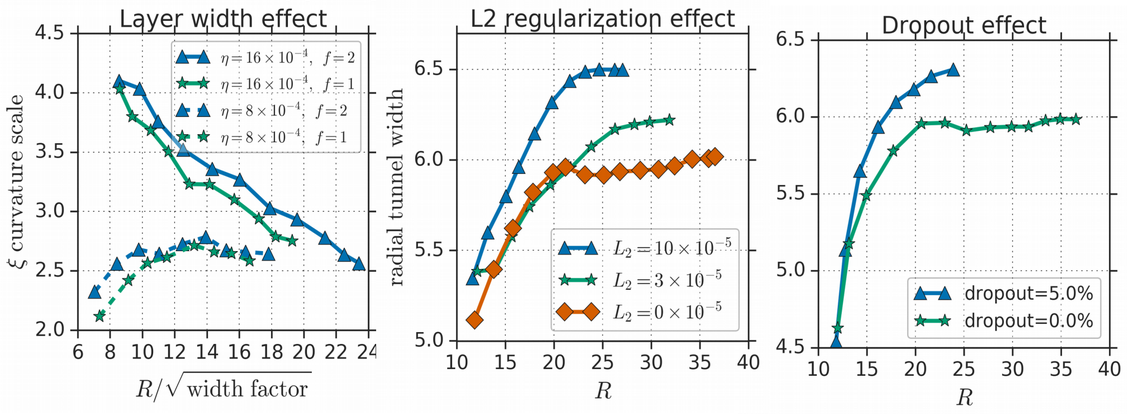




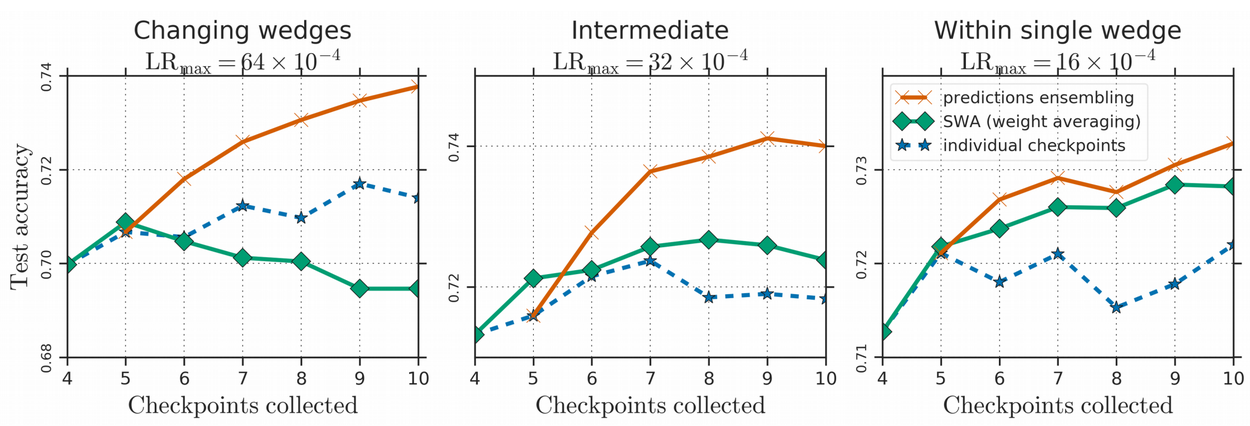


Experiments

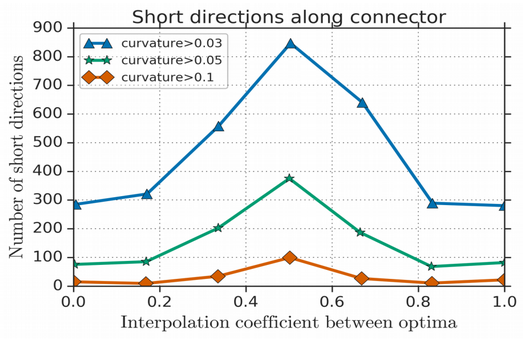
The effect of hyperparameters on the radial tunnel width.

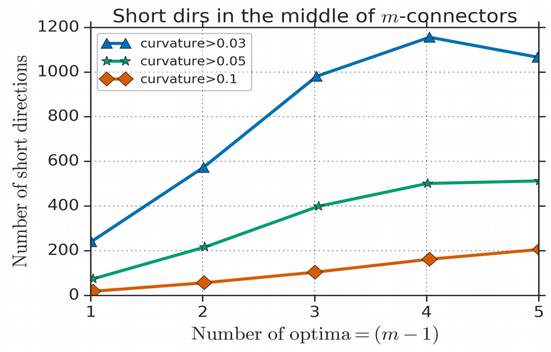


The effect of the landscape on ensembling. Stochastic Weight Ensembling (SWA) works well only if solutions do not change n-wedges.



Constructing *m*-tunnels and verifying the increasing number of short directions with *m*.





Conclusion

We built a phenomenological model of the low-loss manifold of the neural network loss landscape. We integrate previously observed phenomena of 1) no significant obstacles along optimization trajectory, 2) dense and distributed nature of the solution manifold, and 3) the connectedness of independent optima. Based on our model, we make further predictions and verify them in real networks.

References:
[1] Ian J. Goodfellow, O Vinyals, and A M Saxe. Qualitatively characterizing neural network optimization problems, 2014.
[2] Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes, 2018
[3] S Fort and A Scherlis. The Goldilocks zone: Towards better understanding of neural network loss landscapes
[4] F Draxler, K Veschgini, M Salmhofer, F A Hamprecht. Essentially no barriers in neural network energy landscape
[5] Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns.