Visualizing Loss landscape of LLMs

Aayush Agarwal
aka7919

Saksham Bassi
sb7787

Rahul Sankar
rrs6684

Abstract

Trained deep learning models performing poorly on Out-Of-Distribution (OOD) datasets have motivated researchers to study how can models generalize. Observing objective metrics on held-out test data (OOD) is a good first measure to analyze if a model generalizes well. It is argued that flatter minima in loss landscape lead to a generalized model. We experiment by empirically visualizing the loss landscape of Large Language Models to observe if smoother, flatter minima correspond to a generalized model. We notice that model trained with SAM (Foret et al., 2021), a fine-tuning objective improves generalization and has wider optima in the loss landspace.

1 Introduction

Initial work on observing and visualizing one-dimensional loss landscape in parameter space by Goodfellow et al. (2015) showed that gradient descent algorithms try to find flat optima. It deploys a technique for traversing in the one-dimensional space from initial weight parameters to final weight parameters to study the nature of models. Li et al. (2018) extends the work on visualizing loss landscape by working in two-dimensional parameter space. The work introduces the "filter normalization" method to visualize the loss landscape in two dimensions - which allowed for a clearer empirical analysis of minima flatness. Recent work by Hao et al. (2019) studied the effectiveness of the BERT model using visualizing the loss landscape on a variety of language datasets. They observe that compared to training models from scratch, BERT initializing and finetuning results in a wider optima. Recent optimization objectives like Sharpness-aware Minimization (SAM Foret et al. (2021)) proposed an objective function similar to gradient descent, wherein the model tries to find weight parameters that minimize the loss value of an entire "neighborhood", as opposed to a single point.

2 Experiments

We visualize the loss of fine-tuned language models on the COLA task of GLUE benchmark (Wang et al., 2018). We compare findings on two models - the baseline model - BERT and the BERT model fine-tuned with the SAM objective. The setup involves loading model weights from the pre-trained BERT model on HuggingFace and fine-tuning both models for 10 epochs. To plot loss landscapes, we use Bernardi's library - "loss-landscapes".

3 Preliminary Results

We can see from Table 1 that BERT fine-tuned with SAM performs better than the Baseline model (BERT fine-tuned without SAM) on the validation dataset. BERT with SAM achieves a lower validation loss of 0.745 and Matthew's Correlation score of 0.617.

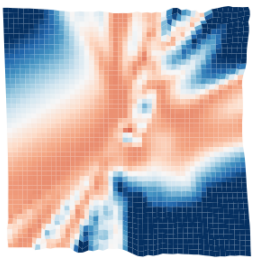
Model	Validation Loss ↓	Matthew's Correlation ↑
Baseline	0.940	0.581
BERT With SAM	**0.745**	**0.617**

Table 1: Results on Validation Dataset of COLA. The Baseline model is BERT without SAM.

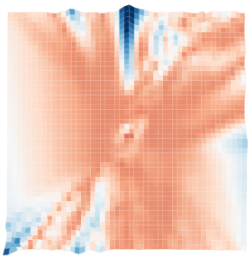
Our results are further corroborated by Figure 1, where we can observe that the model fine-tuned without SAM has an unstable neighborhood with sharper minima, leading to poor generalization. BERT with SAM has a relatively flatter minima, which is substantiated by the model's performance on the unseen validation set. This empirical analysis shows that the SAM objective in fact finds a low-loss neighborhood, or put differently has a flatter optima. The top view figures in Figure 1 show clearly that the SAM objective has a low minimum loss in the neighborhood around the trained model parameters (center of the figure).

4 Further work

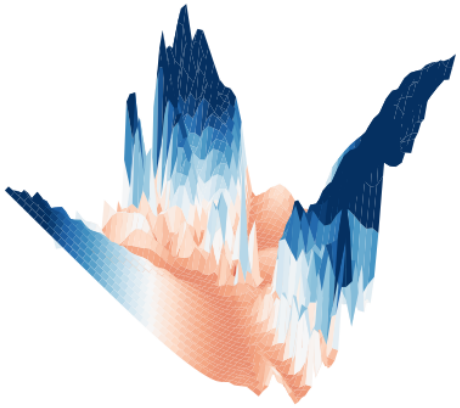
We plan to extend our experiments on the MNLI, RTE, SST-2, and MRPC tasks from the GLUE benchmark, and visualize their loss landscapes. Furthermore, we plan to quantify the flatness of the optima of each of these models and corroborate our visualizations by computing the sharpness-based complexity measure shown in Jiang et al. (2020).



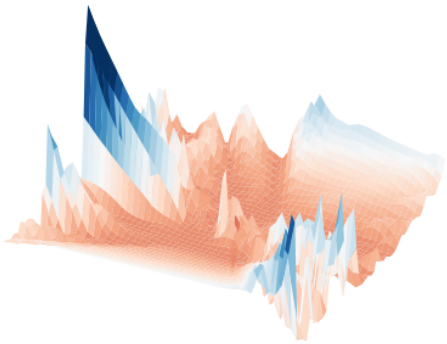
(a) Top-view of Loss landscape on Baseline model



(b) Top-view of Loss landscape on BERT model with SAM



(c) Loss landscape on Baseline model



(d) Loss landscape on BERT model with SAM

Figure 1: Loss landscape comparison of baseline model and BERT model with SAM objective. The loss landscapes plots are generated using the "filter normalization" technique using Bernardi (2019). Here, the trained model parameters are present at the exact center of the figure.

5 Conclusion

Our primary objective behind this work was to study the hypothesis that flat loss structures are related to higher performance. We study this hypothesis in the space of natural language processing. We show in the preliminary results that the hypothesis is indeed true- the model fine-tuned with SAM objective tries to find flatter optima and performs better on one of the tasks in the GLUE benchmark.

References

Marcello De Bernardi. 2019. loss-landscapes. https://github.com/marcellodebernardi/loss-landscapes.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. 2021. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*.

Ian Goodfellow, Oriol Vinyals, and Andrew Saxe. 2015. Qualitatively characterizing neural network optimization problems. In *International Conference on Learning Representations*.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2019. Visualizing and understanding the effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4143–4152, Hong Kong, China. Association for Computational Linguistics.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. 2020. Fantastic generalization measures and where to find them. *International Conference on Learning Representations*.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. 2018. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.