Visualizing Loss landscape of LLMs

Aayush Agrawal
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 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 the model trained with SAM (Foret et al., 2021), a fine-tuning objective improves generalization and has wider optima in the loss landspace on 4/5 GLUE tasks.

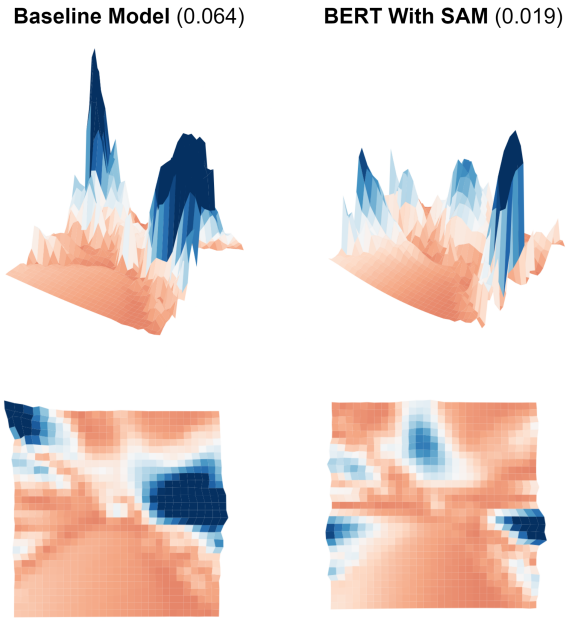


Figure 1: Loss landscapes of baseline model and baseline + SAM model on MRPC dataset. The loss landscape of the baseline+SAM model is flatter with an average slope to the center equal to 0.019 (flatter slope) vs 0.064 of the baseline model. Refer to Figure 3 to see results on some more of the GLUE tasks.

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.

Bahri et al. (2022) trained T5 (Raffel et al., 2022) and mT5 (Xue et al., 2021) with and without the SAM objective, on the SuperGLUE (Wang et al., 2019), GLUE (Wang et al., 2018), Web Questions (Berant et al., 2013), Natural Questions (Kwiatkowski et al., 2019), Trivia QA (Joshi et al., 2017) and TyDiQA (Clark et al., 2020) tasks. They observed that SAM boosted the performance across the board, achieving significant improvements over baseline models.

In this work, we visualize the loss landscapes of the vanilla BERT model and BERT + SAM model on a variety of language datasets. This work enables us to test the hypothesis - flat minima (i.e. flat loss landscapes around the optimum parameters) correlate to better generalization. We experiment with the above 2 models on GLUE benchmark tasks which we discuss in Section 3. The loss landscapes plots are generated using Li et al.'s "filter normalization" technique (see Figures 1, 3). In addition

to visualizing, we came up with a simple formula to quantify the flatness of sharpness using the average slope of lines from centers to corners of the landscape, which we discuss in detail in section 2.2.

2 Data & Methodology

2.1 Data

We implement our experiments on some of the GLUE benchmark tasks (Wang et al., 2018). GLUE benchmark consists of various datasets for natural language tasks that are used to train and evaluate models. The five tasks that we experiment on are:

- COLA: It consists of English sentences and serves to perform a binary classification of whether the sentence is grammatically correct or not.
- MNLI: It consists of pair of English sentences with labels inferring textual entailment. The first sentence of the pair is a premise and the second acts as its hypothesis. This serves to perform a ternary classification such that the hypothesis either entails, is neutral, or contradicts the premise.
- MRPC consists of pair of English sentences extracted from news sources and serves to perform a binary classification of whether the sentences are semantically equivalent or not.
- RTE is structurally similar to MNLI, however, the neutral and contradiction labels are collapsed together as not-entailment, hence, it is a binary classification dataset.
- SST2 consists of movie review sentences with labels being positive or negative.

2.2 Methodology

BERT (Devlin et al., 2019) is a machine learning model based on transformers architecture. Because of the positional encoding and the attention mechanism in the architecture, the model is not constrained by any particular sequence of tokens. The pre-training of BERT involves masked language modeling, where a proportion of input tokens are masked, and then predicted by the model. This enables the model to learn the context in both the forward as well as backward directions. This is useful for providing contextualized embedding for downstream tasks through pretraining. In our experiments, we use the pre-trained BERT model for all the tasks and fine-tune it for our use-cases.

Loss landscape is a three-dimensional space where model parameters are on the x-y plane and corresponding loss values are on the z-axis. It is computed by perturbing the trained model parameters in two orthogonal directions and computing the loss as shown in Eq. 1

$$f(\alpha, \beta) = L(\theta + \alpha\delta + \beta\eta) \tag{1}$$

where L represents the loss function, and θ represents the trained model parameters. Moreover, δ and η represent the two orthogonal unit vectors, which are sampled from a Gaussian distribution and are filter-wise normalized as per Li et al. (2018).

SAM (Foret et al. (2021), Bahri et al. (2022)) objective works on the principle to minimize the sharpness in the loss landscape along with minimizing the loss itself. It aims to find a parameter w such that the overall "neighborhood" around it has a low training loss. This is achieved by optimizing a minimax function given by:

$$\min_{w} \max_{||\epsilon||_2 \leq \rho} L_{train}(w + \epsilon) \tag{2}$$

Here, the function looks for the maximum loss around a parameter w, at a distance ϵ away, bounded by some ρ. This way it selects the sharpest loss in the neighborhood. Eventually, we minimize the sharpest loss possible by globally searching through parameters w.

Sharpness Measure is calculated using the average gradient of the lines generated using the loss values at the center (3x3 center matrix) to the edges in the loss landscape. We fit line equations on the points on these projections from each center point in the 3×3 center matrix (see Figure 2) and then calculate the average slope of all these lines ($3 \times 3 \times 8 = 72$ lines).

3 Experiments

We visualize the loss of fine-tuned language models on the COLA, MNLI, MRPC, RTE, and SST2 tasks of the GLUE benchmark (Wang et al., 2018). We compare findings on two models - the baseline model - BERT and the BERT fine-tuned with SAM objective. The setup involves loading model weights from the pre-trained BERT model (`bert-base-uncased`) on HuggingFace and fine-tuning both models for 10 epochs. We train

GLUE Tasks	Baseline Model			BERT With SAM		
	Metric ↑	Validation Loss ↓	Sharpness Measure ↓	Metric ↑	Validation Loss ↓	Sharpness Measure ↓
COLA (Matthew's Correlation)	0.592	0.804	0.055	**0.596**	**0.488**	**0.020**
MNLI (Accuracy)	**0.840**	**0.826**	**0.037**	0.839	0.989	0.072
MRPC (Accuracy)	0.821	0.726	0.064	**0.838**	**0.522**	**0.019**
RTE (Accuracy)	0.686	0.905	0.033	**0.700**	**0.839**	**0.023**
SST2 (Accuracy)	0.930	0.327	0.048	**0.935**	**0.217**	**0.019**

Table 1: Results on Validation Datasets for COLA, MNLI, MRPC, RTE and SST2. The baseline model is BERT with AdamW (without SAM).

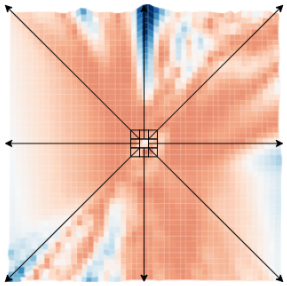


Figure 2: An example showing the lines considered to calculate sharpness from one center point. This step is repeated for the rest of the points around the center (that touch the center point directly). The center points are shown in the 3×3 matrix drawn.

each model with an initial learning rate of $2e^{-5}$ and a weight decay of 0.01 with the AdamW (Loshchilov and Hutter, 2019) optimizer. The hyperparameter ρ is set to 0.01 for the setup with SAM objective. To plot loss landscapes, we use Bernardi's library - "loss-landscapes". For each model, we plot the loss-landscape for 30 "steps", i.e. with the trained model at the origin, the parameters are perturbed in two randomly chosen orthogonal directions for 30 steps, to obtain a plane of $30 \times 30 = 900$ points.

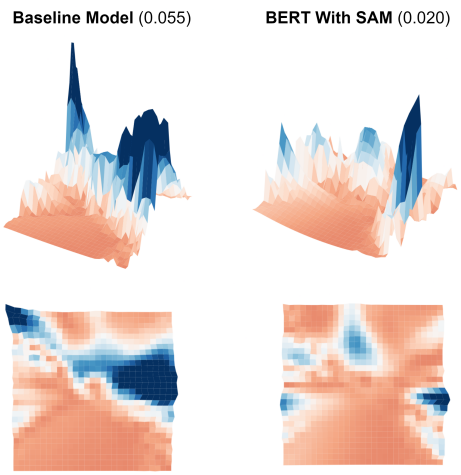
4 Results

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 datasets for COLA, MRPC, RTE and SST2 tasks. We further observe that the sharpness measure corroborates our results as well, i.e. for tasks that achieved a lower validation loss and a higher metric score, the sharpness measure was lower compared to the baseline model. However, we see that the baseline model performs better in the case of MNLI where the amount of data is larger as seen in Table 2. We hypothesize that SAM performs better in situations where the size of data available is relatively small. Additionally, Bahri et al. (2022) observed that SAM tends to work best when there is a scarcity of training data, which is what we observe in our experiments as well.

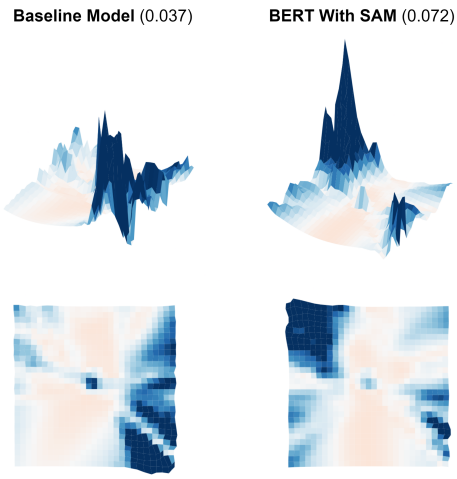
Our results are further verified by Figures 1, 3 where, for the COLA, MRPC, RTE, and SST2 tasks, we can observe that the model fine-tuned without SAM has an unstable neighborhood with sharper minima, leading to poor generalization. BERT with SAM has a relatively flatter minima, substantiated by the model's performance on the unseen validation set for these tasks. 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 Figures 1, 3 show clearly that the SAM objective has a low minimum loss in the neighborhood around the trained model parameters (center of the figure).

GLUE dataset	Train size	Test size
COLA	8.5k	1k
MNLI	393k	20k
MRPC	3.7k	1.7k
RTE	2.5k	3k
SST2	67k	1.8k

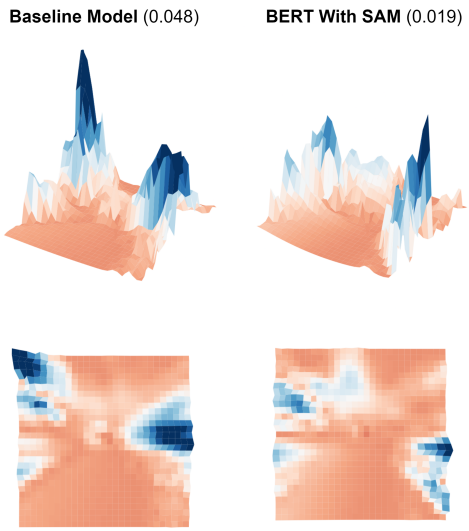
Table 2: Size of GLUE datasets used (Wang et al., 2018)



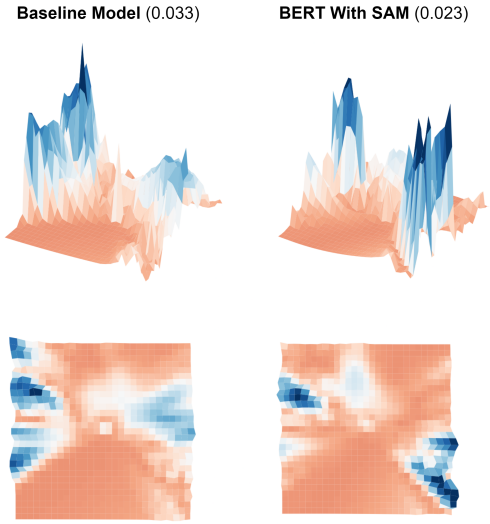
(a) COLA



(b) MNLI



(c) SST2



(d) RTE

Figure 3: Loss landscape comparison of baseline model and BERT model with SAM objective trained on GLUE tasks. For each figure, the sharpness measure is given in parentheses and the most optimal model parameters are present at the exact center of the figure. The orange colors represent lower loss areas and the blue color areas represent a higher loss.

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 showed that the hypothesis is indeed true for the majority of tasks - model fine-tuned with SAM objective performs better on a majority of GLUE tasks. We further showed that the resulting loss landscapes corroborate this fact, i.e. models trained with SAM tend to have a flatter neighborhood compared to the baseline model, and are likely to generalize better on unseen data.

6 Collaboration Statement

This work was equally divided between all the collaborators. The literature study, ideation, and write-up were done by everyone on the team. For experimentation, Aayush set up the skeleton code for language modeling. Saksham was responsible to incorporate loss-landscape visualization for different models and calculating landscape sharpness. Rahul was responsible for fine-tuning Language models using HuggingFace.

References

Dara Bahri, Hossein Mobahi, and Yi Tay. 2022. Sharpness-aware minimization improves language model generalization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7360–7371, Dublin, Ireland. Association for Computational Linguistics.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.

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

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

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.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2022. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. 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.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.