Visualizing Loss Landscape of LLMs

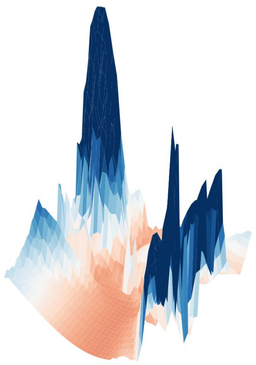
Aayush Agrawal*, Saksham Bassi*, Rahul Sankar*

*Equal contribution

What is a loss landscape ?

One can visualize the neighborhood of the model optima by perturbing the model parameters along arbitrary directions and observing how the loss changes.

Hypothesis: Flatter neighborhood of model optima = better generalization



Methodology

- Use **S**harpness **A**ware **M**inimization to find a minima with flat neighborhood
- Visualize the loss landscape against baseline model.
- Compute sharpness measure and see if they agree with the loss landscape.

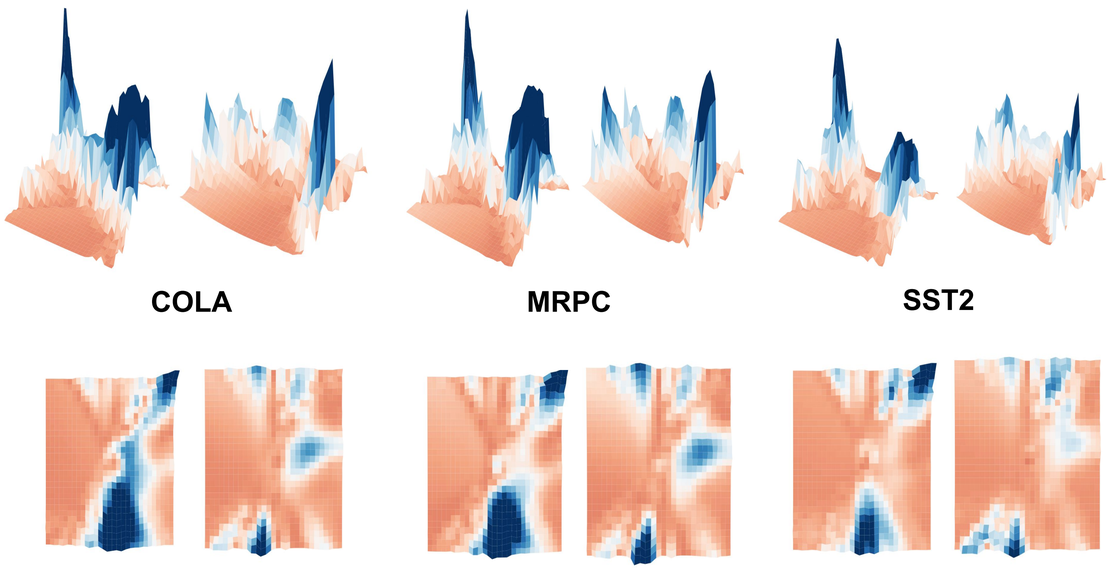
Sharpness measure corroborates the neighborhood seen in loss landscapes

Do LLMs trained with SAM objective generalize better?

We show in results that the hypothesis is true for ⅘ tasks we tested on - the model fine-tuned with SAM objective tries to find flatter optima and leads to better performance on unseen data.

We visualize the loss landscape of fine-tuned language models on various tasks of GLUE benchmark

Loss Landscape for GLUE tasks without (left) vs. with SAM (right)



SAM objective:

$$\min_{w} \max_{||\epsilon||_2<\rho} L(w+\epsilon)$$

Sharpness measure:

Average gradient of diagonal lines from the trained model parameters (3x3 center matrix) to the edges.

Tasks	Without SAM		With SAM	
	Metric ↑	Sharpness ↓	Metric ↑	Sharpness ↓
COLA (Matthew's Correlation)	0.592	0.055	**0.596**	**0.020**
MNLI (Accuracy)	**0.840**	**0.037**	0.839	0.072
MRPC (Accuracy)	0.821	0.064	**0.838**	**0.019**
RTE (Accuracy)	0.686	0.033	**0.700**	**0.023**
SST2 (Accuracy)	0.930	0.048	**0.935**	**0.019**