Convergence of Loss Function Surface in Transformers

Egor Petrov

Moscow Institute of Physics and Technology
Course: My first scientific paper
Consultant: Nikita Kiselev, BSc
Expert: Andrey Grabovoy, PhD

2025

Loss Function Landscape Convergence for transformers

Training a neural network involves searching for the minimum point of the loss function, which defines the surface in the space of model parameters.

Goal

Investigation of the Loss Function's Landscape for Transformer's architecture to find the minimal dataset size

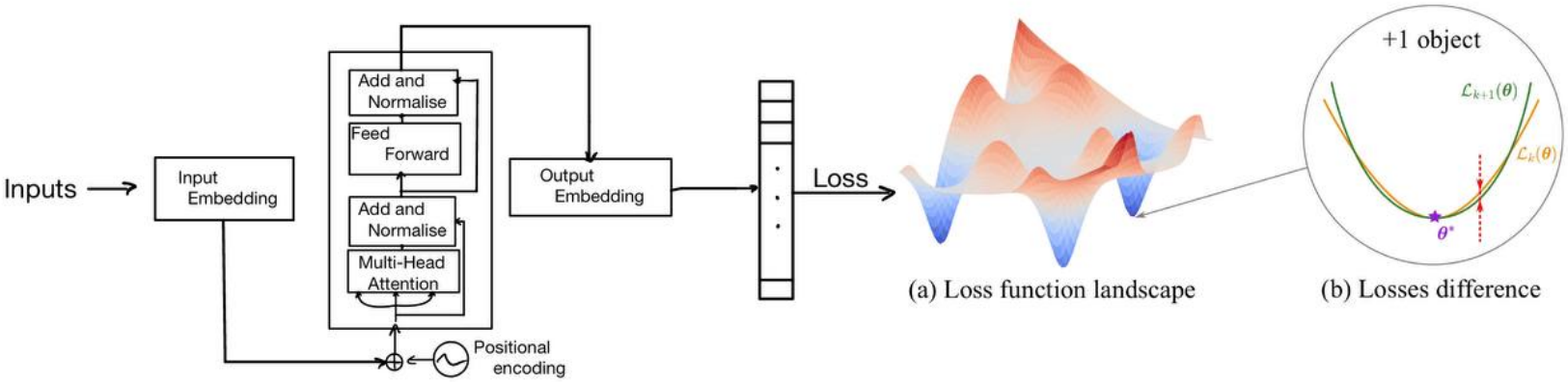
Problem

Determine the minimal dataset size k^* for loss function surface convergence in transformers within a predefined error threshold.

Solution

1) Hessian-based approach to find the critical sufficient dataset size for transformer architectures.
2) Empirical studies on the task of image classification using ViT's
3) Reduce of computational resources with the minimal data size

Loss function Landscape Convergence for a Transformer Block



1. In the neighborhood of a local minimum, the loss function can be approximated by a quadratic form
2. When incrementally adding samples to the training set, we observe convergence in the optimization landscape

$$\mathcal{L}_k(\boldsymbol{w}) \approx \mathcal{L}_k(\boldsymbol{w}^*) + \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}^*)^\top \mathbf{H}^{(k)}(\boldsymbol{w}^*)(\boldsymbol{w} - \boldsymbol{w}^*)$$

The described convergence will provide estimates on the minimum data size for efficient training.

Problem Statement

Objective

Determine the minimal dataset size k^* for loss function surface convergence in transformers within a predefined error threshold.

Challenges

- Analyze Hessian $\mathbf{H}_k(\mathbf{w})$ to quantify landscape evolution with k.
- Derive bounds for $\mathcal{L}_{k+1}(\mathbf{w}) - \mathcal{L}_k(\mathbf{w})$.
- Validate empirically for transformers (e.g., ViTs).

Motivation

Efficient training in data-scarce domains (e.g., medical imaging) with limited resources

Solution

Approach

- Decompose Hessian: $\mathbf{H}_k = \mathbf{H}_o + \mathbf{H}_f$.
- Use Taylor approximation at $\mathbf{w}^*$:

$$\mathcal{L}_k(\mathbf{w}) \approx \mathcal{L}_k(\mathbf{w}^*) + \frac{1}{2}(\mathbf{w} - \mathbf{w}^*)^\top \mathbf{H}_k(\mathbf{w}^*)(\mathbf{w} - \mathbf{w}^*).$$

Bound

$$|\mathcal{L}_{k+1} - \mathcal{L}_k| \leq \frac{1}{k+1} |l_{k+1} - \mathcal{L}_k| + \frac{\|\mathbf{w} - \mathbf{w}^*\|_2^2}{2(k+1)} \left\| \mathbf{H}_{k+1} - \frac{1}{k} \sum \mathbf{H}_i \right\|_2.$$

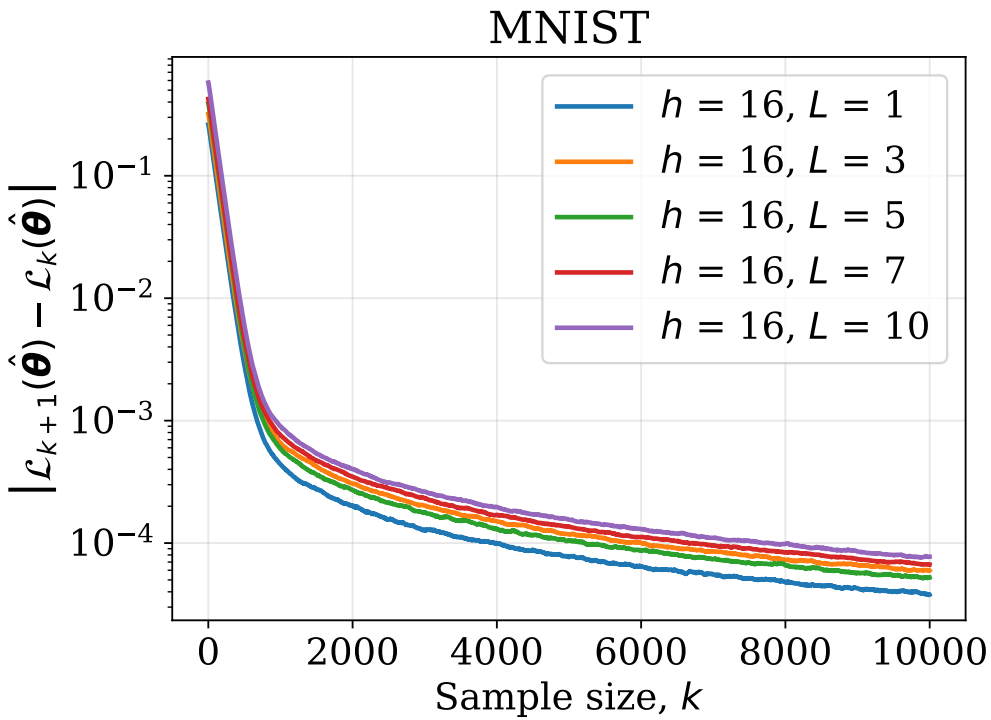
Outcome

Estimate k^* for minimal dataset size, reducing computational costs.

Experiments

Transformer Experiment

- Fine-tuned ViT on small image datasets (LoRA, unfreezing layers).
- Monitored accuracy and $|\mathcal{L}_{k+1} - \mathcal{L}_k|$.



Conclusion

Summary

- Hessian-based framework for k^* estimation in transformers.
- Theoretical bounds validated via ViT experiments.
- Practical for resource-efficient training.

Future Work

- Extend to multi-layer transformers.
- Apply to specific tasks (e.g., medical imaging).