Convergence of the loss function surface in transformer neural network architectures

Egor Petrov
MIPT
Moscow, Russia
petrov.egor.d@phystech.edu

Nikita Kiselev
MIPT
Moscow, Russia
kiselev.ns@phystech.edu

Vladislav Meshkov
MIPT
Moscow, Russia
email@phystech.edu

Andrey Grabovoy
MIPT
Moscow, Russia
email@phystech.edu

Abstract

Training a neural network involves searching for the minimum point of the loss function, which defines the surface in the space of model parameters. The properties of this surface are determined by the chosen architecture, the loss function, and the training data. Existing studies show that as the number of objects in the sample increases, the surface of the loss function ceases to change significantly. The paper obtains an estimate for the convergence of the surface of the loss function for the transformer architecture of a neural network with attention layers, as well as conducts computational experiments that confirm the obtained theoretical results. In this paper, we propose a theoretical estimate for the minimum sample size required to train a model with any predetermined acceptable error, providing experiments that prove the theoretical boundaries.

Keywords: Neural networks, Transformer, Loss landscape, Hessian, Dataset size threshold.

1 Introduction

TODO

Contributions. Our contributions can be summarized as follows:

- We present...
- We demonstrate the validity of our theoretical results through empirical studies...
- We highlight the implications of our findings for...

Outline. The rest of the paper is organized as follows...

2 Related Work

Topic #1. TODO

Topic #2. TODO

Preprint. Under review.

3 Preliminaries

3.1 General notation

In this section, we introduce the general notation used in the rest of the paper and the basic assumptions.

3.2 Assumptions

TODO

4 Method

5 Experiments

To verify the theoretical estimates obtained, we conducted a detailed empirical study...

6 Discussion

TODO

7 Conclusion

TODO

References

A Appendix / supplemental material

A.1 Additional experiments / Proofs of Theorems

TODO