Convergence of the loss function surface in transformer neural network architectures

Egor Petrov, Nikita Kiselev, Vladislav Meshkov, Andrey Grabovoy

2025

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

References