Why Transformers are good?

Narybaev Nursultan

nnurseultan07@gmail.com

Abstract. Transformers have demonstrated state-of-the-art performance across a wide range of tasks, largely attributed to their ability to map input data into a high-dimensional polynomial feature space via the attention mechanism. However, their computational cost and memory footprint make them prohibitive for many real-world applications, especially in resource-constrained environments. In this work, we investigate alternative architectures that retain the polynomial feature representation property of Transformers while significantly reducing computational overhead. We explore three primary approaches: (1) ReLU squared activation functions to introduce quadratic interactions in neural representations, (2) Bilinear layers with identical inputs to capture multiplicative interactions efficiently, and (3) Linear layers applied to squared inputs, mimicking the polynomial expansion implicit in attention. We analyze these methods in terms of theoretical expressiveness and empirical performance on benchmark tasks, demonstrating that computationally lighter architectures can achieve comparable results to Transformers while improving efficiency.

Keywords: ReLU squared · BiLinear layer · Linear squared layer.

1 Introduction and Motivation

1.1 Introduction

The Transformer architecture has revolutionized deep learning by achieving remarkable performance in natural language processing, computer vision, and other domains. A key factor behind its success is the self-attention mechanism, which effectively captures long-range dependencies and enables rich feature representations. However, the high computational cost and memory consumption of Transformers, primarily due to the quadratic complexity of self-attention, pose significant challenges for scaling and deployment on edge devices.

Sample Heading (Third Level) Only A crucial insight into the effectiveness of Transformers lies in their ability to project inputs into a polynomial feature space, where interactions between tokens are captured through multiplicative terms. This perspective suggests that the power of attention may not be unique to its mechanism but rather to its capacity to generate high-order

feature interactions. Thus, an important research direction is identifying alternative, computationally efficient layers that can approximate this polynomial expansion without the heavy costs of self-attention.

In this work, we explore three alternative approaches that introduce polynomial feature interactions while maintaining a lower computational burden. First, we leverage the squared ReLU activation function, which inherently introduces quadratic terms and has been shown to improve model expressiveness. Second, we investigate bilinear layers with the same input for both terms, providing an explicit second-order expansion. Lastly, we propose a simple linear transformation applied to squared inputs, mimicking the polynomial behavior of attention without requiring explicit pairwise interactions.

Our study aims to bridge the gap between efficiency and expressiveness by demonstrating that replacing Transformers with these computationally lighter layers can retain competitive performance. Through theoretical analysis and empirical validation, we assess whether these alternative architectures can serve as viable replacements for attention mechanisms in various deep learning applications.