<Title of your paper>

<Subtitle of your paper>

Lauri Keskull[1]

Supervisor(s): Prof. Maliheh Izadi[1]**, Aral De Moor**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
May 23, 2024

Name of the student: Lauri Keskull
Final project course: CSE3000 Research Project
Thesis committee: Prof. Maliheh Izadi, Aral De Moor, <Examiner>

Abstract

Increasing the length of the context window in transformers is a sought after performance enhancement. One of the latest and most promising solutions to the quadratic cost of increasing the context window is Infini-Attention, which promises a linear increase in cost. In this paper we explore to what extent their proposed method can pushed.

1 Introduction

Transformers have revolutionized natural language processing, yet they grapple with limitations due to fixed-size context windows, restricting their capacity to process lengthy documents and grasp intricate dependencies outside these windows. Enhanced models such as Transformer-XL and BigBird have addressed these issues through sparse attention mechanisms and cacheing Key-Value matrices, yet challenges persist in terms of scalability and computational efficiency.

In response to these ongoing challenges, Munkhdalai et al. introduced the concept of Infini-attention, an innovative approach allowing efficient management of infinitely long contexts **?**. Infini-attention integrates a compressive memory into the transformer architecture, allowing for a dynamic and scalable handling of extended contexts without a proportional increase in computational demands. This mechanism not only enhances memory efficiency but also fosters continuous contextual understanding over extended sequences.

While Infini-attention has demonstrated promising results in fine-tuning scenarios, this study proposes to explore its integration at an earlier stage—during model pre-training. This approach hypothesizes that initiating Infini-attention in the pre-training phase could leverage its full potential, thereby embedding a deeper contextual understanding from the outset. Specifically, this research seeks to (1) evaluate the impact of Infini-attention when introduced in pre-training versus during fine-tuning, and (2) examine the relationship between the context length a transformer can handle and the capacity of its compressive infini-attention memory.

The insights gained from this study could significantly influence future strategies for deploying transformer models across various real-world applications,

Furthermore, this research will also shed light on the extent to which the compressive memory component of Infini-attention can compensate for shorter local context lengths. By determining how effectively compressive memory can stand in for direct local context interactions, the findings may offer valuable guidelines for optimizing transformer configurations, particularly in applications where extending the context window is constrained by computational or memory resources. This would potentially lead to more robust, context-aware systems capable of handling complex tasks such as document summarization and legal analysis more effectively.