INFORMATICS INSTITUTE OF TECHNOLOGY

INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER

"CompressO"

Compression of Transformers using Pruning

A Dissertation by

Ms. Weerasinghe Madhushani

Supervised by

Ms. Sapna Kumarapathirage

Submitted in partial fulfilment of the requirements for the BEng (Hons) Software Engineering degree at the University of Westminster.

July 2022

Abstract

Providing meaningful representations of words has been considered one of the fundamental goals of Natural Language Processing (NLP) since its inception. With the advancement of technology, researchers have experimented with different approaches to achieve this goal. The introduction of word embeddings can be considered a significant improvement in this field of study which provides a vector representation of words.

Models based on the Transformer architecture have reached state-of-the-art performance for numerous NLP tasks. However, due to the high dimensionality of these models making use of them has become computationally intensive. The focus of this research is to address the above issue by reducing the model size by compressing the Transformers architecture.

Among various techniques used for Transformer model compression, Global Unstructured Pruning has been used in this project.

Keywords: Transformer Model Compression, Transformer architecture-based models, Neural network compression, Global Unstructured Pruning for Transformers