PRUNELM: COMPRESSING BI-DIRECTIONAL LONG SHORT-TERM MEMORY ARCHITECTURE BASED PRE-TRAINED CONTEXTUALIZED WORD EMBEDDINGS

Sahan Dilshan

A dissertation submitted in partial fulfilment of the requirement for Bachelor of Engineering (Honours) degree in Software Engineering

Department of Computing
Informatics Institute of Technology, Sri Lanka
in collaboration with
University of Westminster, UK

2021

Abstract

Deep Learning has gained popularity in recent years due to its high accuracy in fields such as pattern recognition and computer vision. As a result, almost every task in Natural Language Processing, such as sentiment analysis, named entity recognition, language translation, and text classification, now uses deep learning architectures and methods to achieve their goals. There's a special layer called the word embedding layer when it comes to Natural Language Processing with Deep Learning. The purpose of the word embedding layer is to give a vectorize representation to a word so that computers can do calculations on words as the human brain does.

Due to highly dimensionality of these embedding layers, they require high memory requirements, hence these models are unable to use with memory constraint devices. To address this issue various research have been conducted on word embedding compression. After going through these research and identifying some valid gaps, *PruneLM* is focused to compress Bi-LSTM based contextualized word embeddings.

Unlike other compression systems, PruneLM provides a novel approach to compression. Pruning has been used as the compression method in the PruneLM. PruneLM provides two different compressions for a given model. A given Bi-LSTM based model can be compressed with or without retraining with the PruneLM and with the provided statistics dashboards, users can compare and view the performance of the compressed models.

Keywords: Language Model Compression, Bi-LSTM based contextualized Word Embedding, Word Embedding Compression, Model Pruning, Bi-LSTM Language Model