Project Formulation

Comparing Various Dimensionality Reduction Approaches For Embedding Vectors

Group 14

June 1, 2023

1 Motivation

In recent years the trend was to develop more powerful language models which often correlates to an increase in size. As different language models usually contain high dimensional embeddings this entails a number of challenges. Storing a high number of embeddings requires quite a lot of memory, while working with them requires significant amounts of time, energy and computational resource, thus limiting the accessibility of this technology. To ease this impediment, it would be necessary to create embeddings with drastically less dimensions, that still provide (most of) the capabilities of the original large embedding. Besides enhancing established kinds of usages, this also could clear the way for new adaptations and use-cases in the fields of Internet of Things (IoT) and mobile, where time, memory and energy constraints are particularly tight.

2 Dataset

In order to compare the results of the language models with compressed embeddings we train them to perform a downstream task such as classification on different datasets. For this we will be using the following three classification datasets to compare the results:

1. **Dataset: CheckThat Lab - Subjectivity Detection:** The task is to predict whether a given sentence from a news article is subjective or objective.
   The train set contains 830 sentences
   The dev set contains 219 sentences
   The test set contains 244 sentences
   The data contains binary labels SUBJ or OBJ

2. **Dataset: Aspect Sentiment Classification Dataset:** The aim is to classify the sentiment towards each sentence as positive, negative or neutral given the aspect and the review sentence.
   The dataset is divided into train test and dev sets. Each entity has a classification label, 'polarity' having the values positive, negative or neutral.

3. **Dataset: FEVER (Fact Extraction and VERification):** The task is to predict whether a given claim is supported or refuted by verifying the facts or whether notEnoughInfo is present to make a decision.
   The dataset consists of 185,445 sentences.
   The claims are classified as Supported, Refuted or NotEnoughInfo.

3 Methodology

In the beginning, statistical properties such as the proportion of each class, the distribution of document lengths, and word frequencies would be examined. The document lengths can be utilized as an outlier detection method that aids in selecting a representative sample of each dataset that is balanced (with regard to classes). The project then involves selecting various dimensionality reduction techniques. Statistical methods including Principal Component Analysis (PCA) and Linear Discriminant Analysis

(LDA), as well as various Autoencoder architectures such as Vanilla and Variational Autoencoders, are used in these methodologies. The SBERT model would be employed to calculate the vector embeddings since it is well-known, simple to use, and allows the embeddings' dimensions to be set at various numbers that can be compared to the dimensionality reduction methods. In order to compare the performances, the outputs of several approaches would then be evaluated on a downstream job, which is the classification using the k-nearest neighbors algorithm.

4 Expected Results

The proposed solution aims to assess the impact of using different compression techniques to reduce high dimensional embeddings in language models while producing similar results as the original models. We expect to find a compression technique that demonstrates that compressed models can produce similar scores as the original models while being computationally inexpensive.

1. **Preservation of semantic information:** We expect that the models will retain their ability to capture and represent the underlying semantic information that is present in the original models even while working with embeddings with reduced dimensions.
2. **Computational efficiency:** A primary goal of this research is to identify a compression technique not only produces comparable results to the original models but also reduces computational complexity in terms of memory and time.
3. **Robustness to noise:** Language models often encounter noisy or perturbed inputs in real-world scenarios. We anticipate that the compressed models exhibit resilience against such variations and maintain their performance.

By achieving these expected results, this project will contribute to the development of practical and computationally efficient solutions that can be used to deploy powerful language models in resource-constrained environments without compromising performance.

5 Evaluation Metrics

The performance of the various approaches will be compared in terms of memory consumption, embedding generation time, runtime of downstream task using different embeddings, and classification task performance. Precision, Recall, F-score, and NDCG@K would be computed and compared to evaluate the performances.

6 Limitations & challenges

We anticipate facing various challenges and dealing with various constraints throughout the project. The following is a brief description of the potential challenges and constraints:

- **Computing resources:** we believe that our biggest challenge will be limited access to computer resources. Recent advances in natural language processing (NLP) are mainly due to a large number of training parameters and a huge amount of available data. To keep up with them, one needs access to a large amount of memory and computational resources, such as a Graphical Processing Unit (GPU), which is out of reach for us in this project. Moreover, for the same reasons, we may not be able to experiment with embeddings with very high dimensions, which restricts our experience.
- **Unforeseeable outcome:** as with almost all other research projects, there is a possibility that we may not end up with the desired outcome. In other words, in the end, we might conclude that the proposed method is not practical in the real world and does not make a significant difference by being applied on top of the state-of-the-art (SOTA) models.

- **Model selection:** since there are not many open source models available, it might be difficult to find one that is not only performing promising but also is open source that we can/are allowed to use in this project.
- **Generalization:** followed by the latter two items, because we cannot experiment with enough language/compression models and datasets, we may not be able to generalize the obtained results to all NLP models.

7 Task Assignment:

As it is a team project, the tasks will be distributed equally among the members upon completing each milestone.

- Reading up on dimensionality reduction techniques, their characteristics, and their applications in natural language processing is the first step in the project.
- The next step would be to implement the SBERT model and these methods to calculte different embeddings.
- Poster Presentation (July 9^{th})
- In the following stage, these embeddings would be given to a k-nearest neighbor method and the outcomes would be compared using different evaluation metrics.
- Final Report (August 7^{th})