VIETNAM GENERAL CONFEDERATION OF LABOUR

TON DUC THANG UNIVERSITY
FACULTY OF INFORMATION TECHNOLOGY

Final Assignment

Deep Learning

Instructor: **Le Anh Cuong PhD.**
Student Name: **Pham Long Duy Tien**
Student ID: **520K0220**
Group: **09**
Class: **20K50301**
Course: **503077**

Ho Chi Minh City, May 6, 2023

Contents

List of Figures

1 Introduction

This paper is a report of a final assignment. Detailing a history, inner workings of DistilBERT, and a demonstration and calculations to use in an emotional analysis application.

2 DistilBERT

2.1 Context

2.1.1 History

Before DistilBERT, there are other ways of doing natural language processing (NLP) tasks, with varied results, specialized tasks or other factors.

2.1.2 Long Short-Term Memory

Long Short-Term Memory (LSTM) models have been widely utilized in various natural language processing tasks. However, despite their success, they are not without their limitations. Firstly, the model processes every word sequentially, resulting in slower processing times for longer inputs. Secondly, the use of simple concatenation restricts their ability to be bidirectional, and as a consequence, they may fail to capture the meaning of some sentences based on the input. Although there have been attempts to address these shortcomings, they remain prevalent.

2.1.3 Transformers

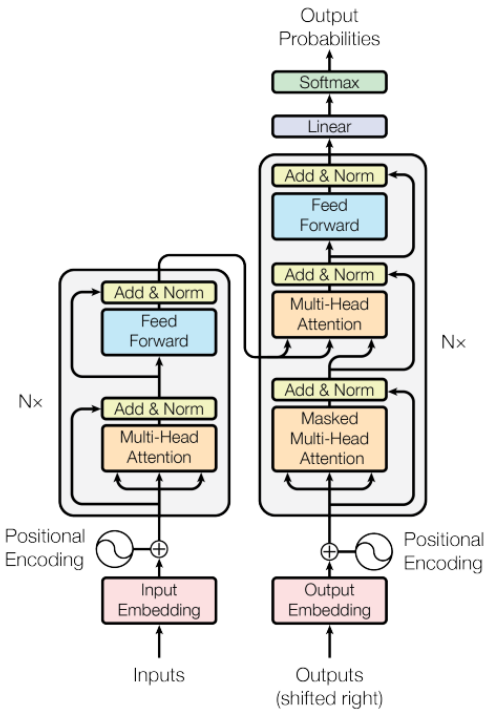

Figure 1: A typical Transformer model architecture

In 2017[1] introduced the Transformer model as an alternative to LSTM-based models, which had encountered problems with processing large sequences due to their sequential nature. The Transformer model addresses this issue by allowing for parallel processing of words and utilizing a true bidirectional approach.

The Transformer architecture consists of two main components: the Encoder and the Decoder. The Encoder processes the input sequence of words in parallel and generates embeddings as vectors. The Decoder takes the information from the Encoder and uses it to generate the output sequence. This architecture is unique and differs from the traditional sequence-to-sequence models.

Due to this, it allowed a degree of modularity for different models when spilt apart between the encoder and decoder. For example: GPT focused on decoders, while BERT focused on encoders.

2.1.4 DistilBERT

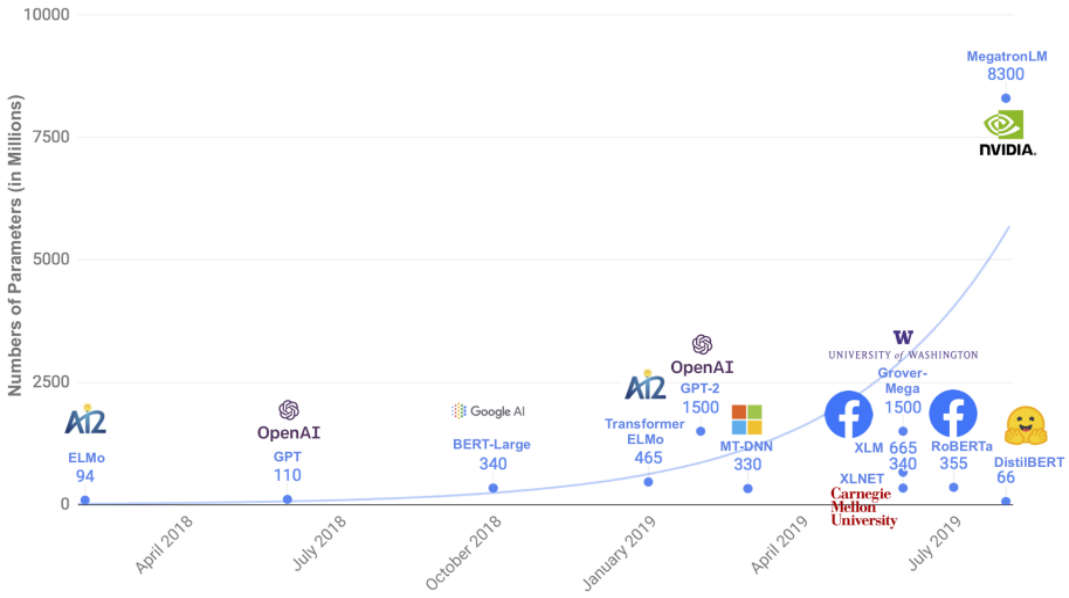

Figure 2: A graph on the growth rate of Transformer-based models

Due to the growing size of Transformer-based models is that the parameters keeps growing as time goes on, resulting in more advanced models, needs more advanced computing power, and therefore more cost and more time, as a result making it more expensive.

In 2019, HuggingFace released a paper for DistilBERT[2], a Transformer-based model using knowledge distillation, using BERT as a base, various loss functions, softmax temperature and other methods to bring down the complexity and training time of other models, while still being as accurate as BERT.

2.2 Structure

2.2.1 Teacher-Student Learning

The concept of teacher-student learning has gained prominence in the field of deep learning, with its proposal and utilization dating back to 2015 [3][4]. This approach entails training a new model, referred to as the student, to reproduce the behavior of a much larger base model, known as the teacher.

Notably, the DistilBERT model has achieved significant success in employing this technique. In the typical implementation of this learning approach, the student is assigned two objectives: (1) minimizing the cross-entropy between the student's predictions and the one-hot empirical distribution of the training labels, and (2) minimizing the Kullback–Leibler divergence between the student's and teacher's predictions.

The teacher's ability to assign small probabilities to incorrect answers is a manifestation of its ability to generalize based on its training. Consequently, the student can learn all the probabilities from the teacher.

However, the minimization of Kullback–Leibler divergence can be problematic as the student may not learn much from incorrect answers. This can result in a sharp distribution on a correct answer with high probability, making the rest improbable, rendering the distillation loss ineffective, especially when working with a well-trained teacher.

2.2.2 Softmax Temperature

One of these solutions for this is to use a softmax-temperature approach.

$$\frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

Where:

- $T = 1$: Standard softmax formula.
- $T > 1$: Probability distribution is softer, giving more weights to incorrect answers.
- $T = 1$ is set at inference time for the student to produce the standard softmax outputs.

This is the same temperature T to the student and teacher at training time. The teacher (trained and frozen) produced adjusted probability estimates which the student should learn.

2.2.3 Architecture

DistilBERT, as mentioned in the original paper, is a smaller and faster version of BERT. This was achieved by reducing the number of encoder blocks by a factor of two, which decreased the original 12 layers of BERT down to 6 in DistilBERT. The weights from BERT were reused to initialize DistilBERT without any modifications, which helped to preserve speed.

Ablation tests have shown that the removal of 768-dimensional embedding vectors does not have a significant impact on the performance of DistilBERT. This removal enabled the use of cosine-distance loss between DistilBERT and BERT, which aligns the directions of the hidden vectors of the student (DistilBERT) and the teacher (BERT). By keeping the same dimensionality, this alignment was made possible, which is crucial for transferring knowledge from BERT to DistilBERT.

2.3 Training and Evaluation Methodologies

According to the original paper, HuggingFace utilized English Wikipedia and Toronto Book Corpus as the datasets for training their DistilBERT model, along with 8 Nvidia V100 GPUs that have a memory capacity of 16GB each. The training process lasted for 90 hours. For comparison, Meta's RoBERTa model was trained on 1024 Nvidia V100 GPUs with a memory capacity of 32GB for a day.

Following the training process, the performance of DistilBERT was found to be comparable to that of BERT.

Table 1: **DistilBERT retains 97% of BERT performance.** Comparison on the dev sets of the GLUE benchmark. ELMo results as reported by the authors. BERT and DistilBERT results are the medians of 5 runs with different seeds.

Model	**Score**	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST-2	STS-B	WNLI
ELMo	68.7	44.1	68.6	76.6	71.1	86.2	53.4	91.5	70.4	56.3
BERT-base	79.5	56.3	86.7	88.6	91.8	89.6	69.3	92.7	89.0	53.5
DistilBERT	77.0	51.3	82.2	87.5	89.2	88.5	59.9	91.3	86.9	56.3

Figure 3: A comparison of DistilBERT (highlighted) with other models

3 Emotion Analysis

3.1 Introduction

Emotion Analysis (henceforth **EA**) is a natural language processing technique to identify, extract and quantity emotions in data such as video, speech and text.

The goal of emotion analysis is to automatically classify the emotions expressed in a piece of text or speech, such as joy, sadness, anger, or fear.

EA has many practical applications, such as predicting consumer behavior, detecting sentiment trends in social media, identifying customer satisfaction levels, and providing personalized medical support.

Despite its many applications and potential benefits, emotion analysis is still a challenging task due to the complex nature of human emotions and the nuances of language, such as irony.

3.2 Process

3.2.1 Data Collection

Data collection involves gathering text data from various sources, such as social media platforms, customer reviews, or customer service chat logs. The data collected should be relevant to the application and should represent the target audience.

3.2.2 Preprocessing

To provide optimal formatting of the dataset to the model, various functions employed to clean the data, removing noise, and normalizing. Common preprocessing techniques include tokenization, stopword removal, stemming, and lemmatization.

3.2.3 Feature Extraction

By converting the preprocessed text, we can make the data into a set of numerical features that can be used as input to machine learning algorithms. Common feature extraction techniques include bag-of-words, n-grams, and word embeddings.

3.2.4 Evaluation

This involves training a machine learning algorithm on the extracted features to predict the emotion or sentiment of the text data. There are many of these that can do this, some can include Support Vector Machines (SVM), Naive Bayes, Decision Trees, and various Neural Networks.

4 Emotion Analysis with DistilBERT

4.1 Methodology

4.1.1 Data and Compute Power

Using a single Nvidia T4 GPU for computational power, we have conducted training for ten epochs with no further adjustments, utilizing the `sem_eval_2018_task_1` and `subtask5.english` datasets. To facilitate comparison with DistilBERT, we have chosen to employ BERT[5] and RoBERTa[6].

4.1.2 Training

We configured the model with a batch size of 8 for both training and evaluation, a weight decay of 0.01, and a learning rate of 2×10^{-5} for 10 epochs. This same configuration was employed for both BERT and RoBERTa models.

The training times for BERT and RoBERTa were 30 minutes each, while DistilBERT required only 15 minutes for training.

4.2 Results

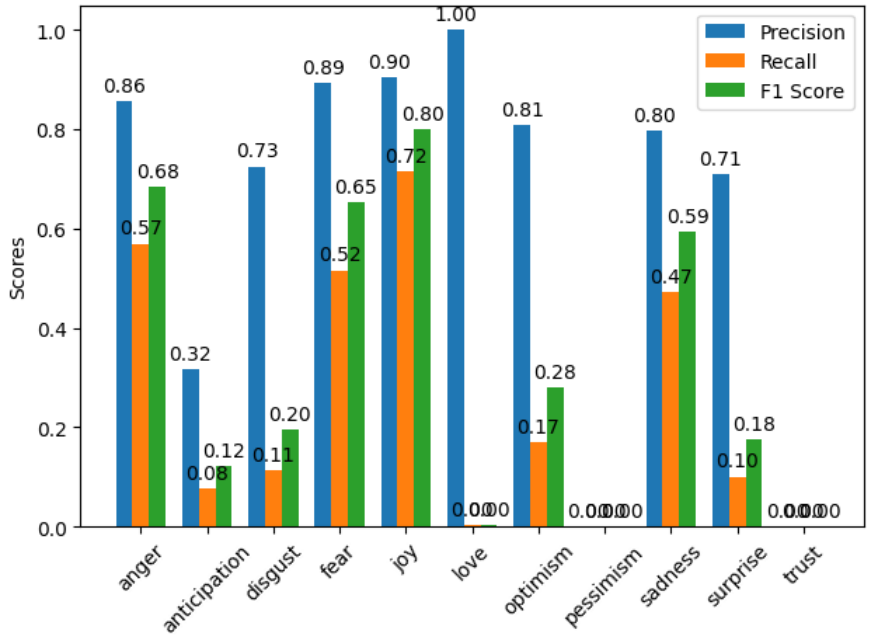

Figure 4: Results of DistilBERT

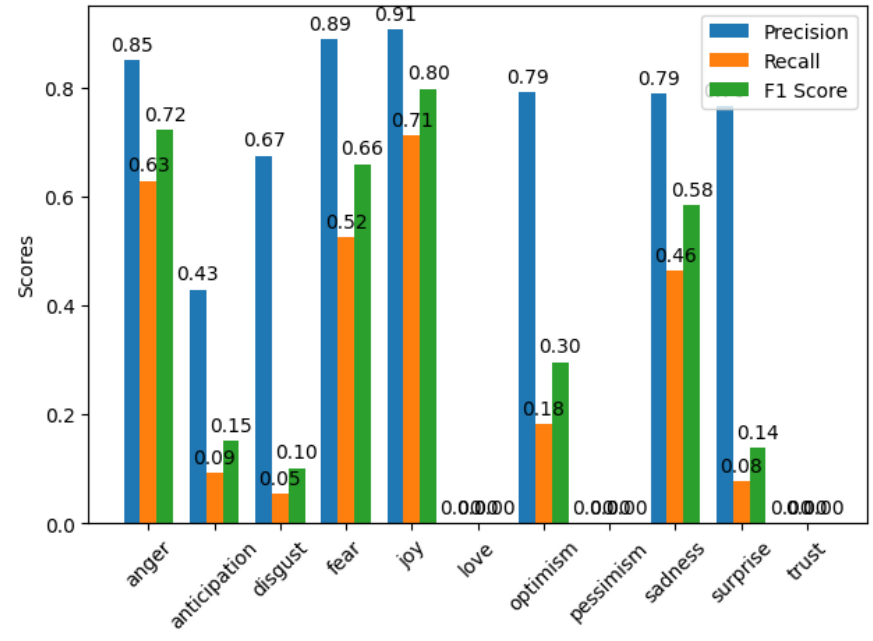

Figure 5: Results of BERT

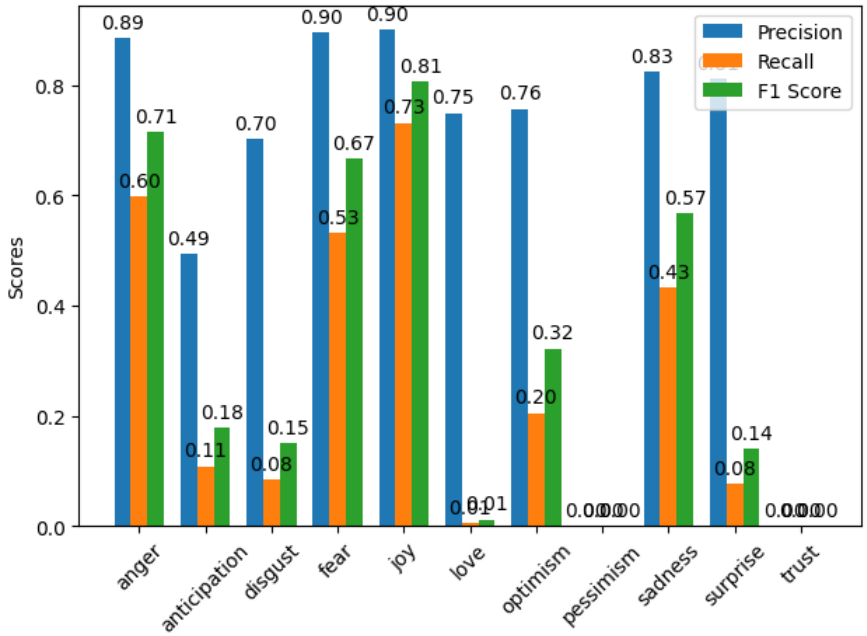

Figure 6: Results of RoBERTa

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[2] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," 2020.

[3] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015.

[4] J. Yim, D. Joo, J. Bae, and J. Kim, "A gift from knowledge distillation: Fast optimization, network minimization and transfer learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[6] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.

End of paper.