# Deep Learning

## Final Assignment: Emotion Analysis with DistilBERT

P.L.D. Tien (520K0220)

Ton Duc Thang University

May 6, 2023

# Outline

Introduction

Context

DistilBERT

Emotion Analysis with DistilBERT

End

# What is Emotion Analysis

- A more comprehensive version of *Sentiment Analysis*.
- Instead of neutral, negative and positive, it focuses on emotions.
- These two sentences are considered positive for sentiment analysis
  - "This milkshake is good."
  - "I am loving this milkshake already."
- An emotion analysis model will say that the second sentence has a stronger positive emotion than the first.

# What is DistilBERT

- *DisbilBERT* is a state of the art language model based on transformers.
- Smaller and faster than *BERT*
  - As it's pretrained on massive amounts of data
- Commonly used in sentiment analysis via fine-tuning, such as sentiment prediction.
- With an appropriate setup, it is able to detect nuances.
  - Irony, sarcasm, slang, etc.

# Context: LTSM

- Before Transformers, we have *LTSM*, but it suffers from two problems.
  1. The model sequentially looks at every word on the input.
     - Passed in and generated sequentially.
     - Long words, longer processing.
  2. Not truly bidirectional.
     - Uses simple concatenation and analyses words separately.
     - True meaning of words lost slightly.

# Context: Transformers

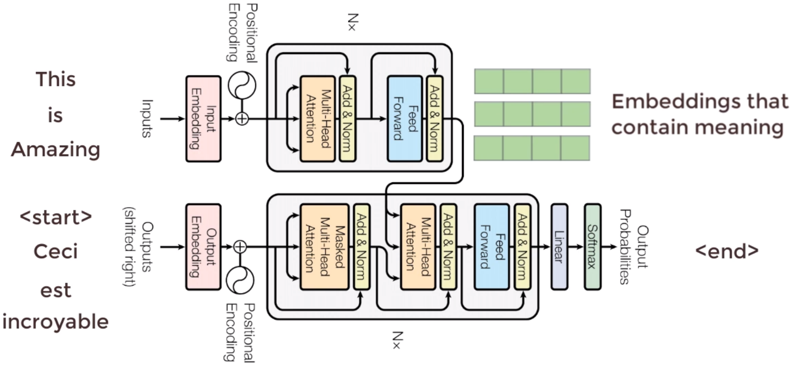
- *Transformers* solve both issues.
  1. Process words simultaneously
     - Is therefore faster
  2. Deeply bidirectional.
     - Able to learn from both directions simultaneously

# Context: Transformers

- *Transformers* have two components:
  1. Encoder
     - Takes each word simultaneously.
     - Generates embeddings (the meanings) simultaneously, being vectors.
  2. Decoder
     - Takes what the encoder generated, and uses to generate text (translation, for instance).

# Context: Transformers

A basic transformer model for English-French translation.[1]




[1] https://youtu.be/xI0HHN5XKDo?t=120

# Context: Transformers

- The benefits of this is that we can actually see a separation in tasks.
  1. Encoder
     - For the translation example, this one knows what is English and context.
  2. Decoder
     - Knows how English words relate to French words.
- Both of them know some bit of language.
- Due to this, we can modify each side.
  - Focus on decoders, we get *GPT*
  - Focus on encoders, we get *BERT*

# What is DistilBERT

- *DistilBERT* is a variant of BERT, focusing on memory-efficiency and speed, while maintaining the accuracy of BERT (within a margin of error). Released in 2019 by HuggingFace.

# Why DistilBERT

- The foundational problems regarding Transformer-based models is that it keeps growing larger.

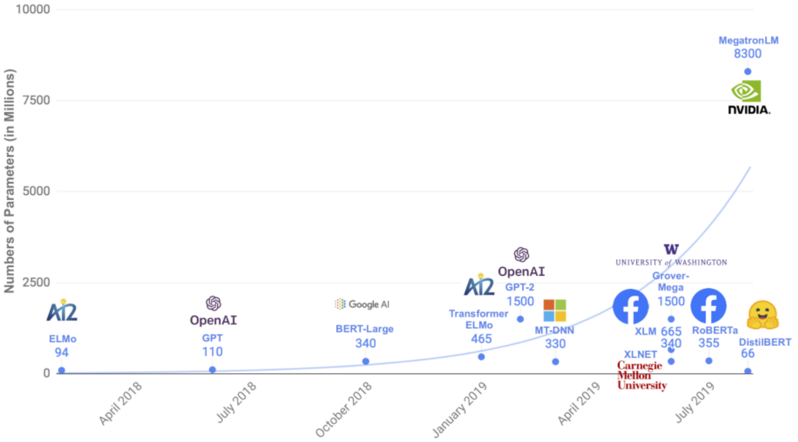



Figure 1: **Parameter counts of several recently released pretrained language models.** [1]

[1] https://arxiv.org/abs/1910.01108, page 1

# How did DistilBERT work

- How it did it was:
  1. Knowledge distillation was utilized, based on the original BERT model.
  2. Using triple loss, able to make a 40% smaller Transformer, pre-trained through distillation. While being 60% faster at inference time.
     - Cross-Entropy
     - Distillation Loss
     - Cosine Embedding Loss

# Knowledge Distillation[1]

- A compression technique introduced in 2006 by [2] in which a small model is trained to reproduce the behavior of a larger model (or models).
- This is one of the main components of *DistilBERT*.

[1] Often called "teacher-student" learning, we'll call it like so.
[2] Caruana et al. In 2015, Geoffrey et al. utilized this in deep learning

# Knowledge Distillation (cont.)

- In DistilBERT, the original BERT is seen as the teacher.
- The student has two objectives:
  1. Minimize cross entropy between the student's prediction and the one-hot empirical distribution of the training labels.
  2. Minimize Kullback–Leibler divergence between the student's and teacher's prediction.

# Knowledge Distillation (cont.)

- The teacher might assign small probabilities to incorrect answers, manifesting how the teacher generalizes.

From Hinton's paper[1]:

An image of a BMW, for example, may only have a very small chance of being mistaken for a garbage truck, but that mistake is still many times more probable than mistaking it for a carrot.

[1] https://arxiv.org/abs/1503.02531

# Knowledge Distillation (cont.)

- Because of this, the student can learn all the probabilities from the teacher.
- But minimizing the Kullback–Leibler divergence has a problem: The student doesn't learn much from the incorrect answers.
- Meaning, a well-trained teacher produces a very sharp distribution with high probability on a correct answer, and the rest are improbable.
- This will render distillation loss useless.

# Knowledge Distillation (cont.)

- One of the solutions is to use softmax-temperature.

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

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

# Other tricks

- In BERT, there are 12 encoder blocks that alternate multi-head self-attention and feed-forward layers.
- DistilBERT cuts that in half, making 6.
- BERT's weights are reused to initialize DistilBERT, providing a massive speed benefit.

# Other tricks

- The 768-dimensional embedding vectors are kept since removing it doesn't do that much of a difference.
- The same dimensionality allowed the use of *cosine-distance loss between DistilBERT and BERT*.
  - To align directions of the the hidden vectors of the student and teacher.

# The training

- Within the same corpus of BERT[1].
- DistilBERT is trained on 8 16GB NVIDIA V100 GPUs for 90 hours.
- For comparison, Facebook/Meta's RoBERTa model took one day on **1024 32GB V100s**.

[1] English Wikipedia and Toronto Book Corpus

# The results

- DistilBERT was able to keep ahead with the original 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 |

[1]

- DistilBERT was also faster for on device computation, 71% faster on an iPhone 7 Plus, running a question-answering model.

[1] https://arxiv.org/abs/1910.01108, page 3

# The results

- An ablation study[1] showed that the Masked Language Modeling had little impact.

Table 4: **Ablation study.** Variations are relative to the model trained with triple loss and teacher weights initialization.

| Ablation | **Variation on GLUE macro-score** |
|---|---|
| $\emptyset$ - $L_{cos}$ - $L_{mlm}$ | -2.96 |
| $L_{ce}$ - $\emptyset$ - $L_{mlm}$ | -1.46 |
| $L_{ce}$ - $L_{cos}$ - $\emptyset$ | -0.31 |
| Triple loss + random weights initialization | -3.69 |

[2]

- Meaning that the student learns from the teacher than the training data.

[1] Remove a part of the model to see what contributes the most
[2] `https://arxiv.org/abs/1910.01108`, page 4

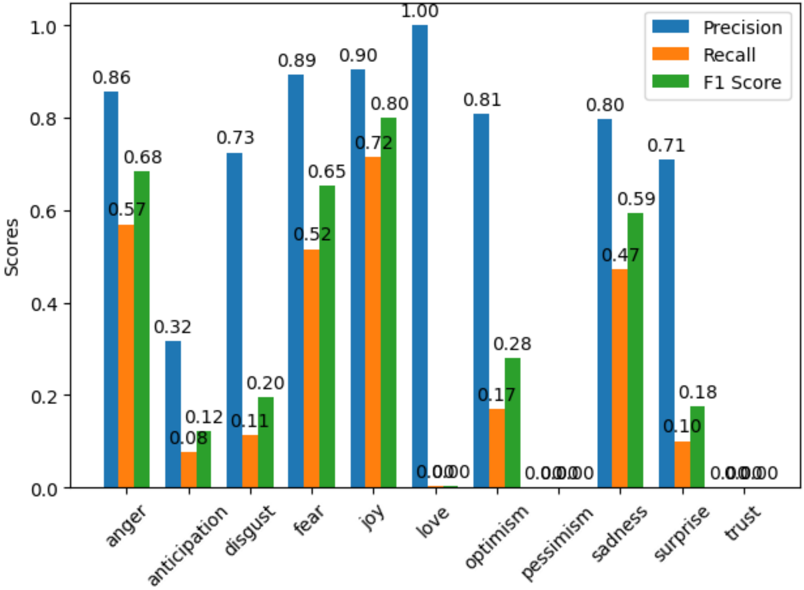
# Methodology

- Datasets used are `sem_eval_2018_task_1` and `subtask5.english`.
- A single Nvidia T4 GPU.
- BERT and RoBERTa for comparison.

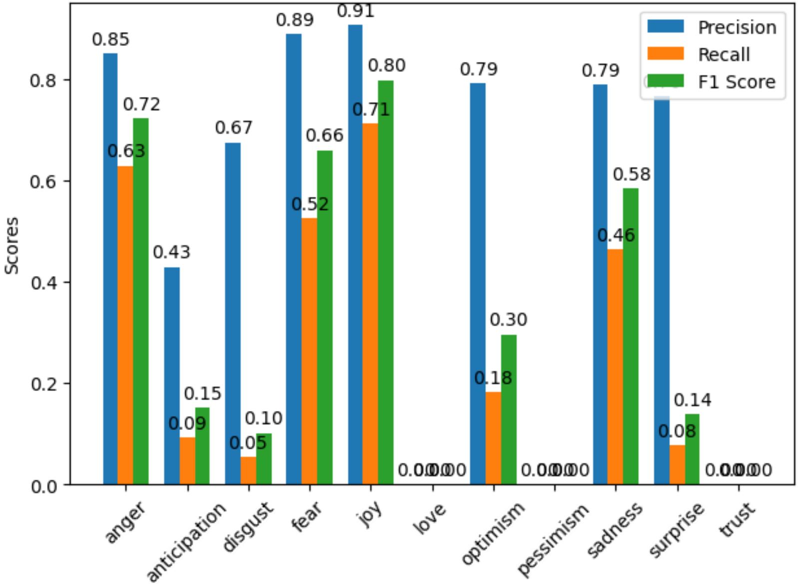
# Training

- Datasets used are `sem_eval_2018_task_1` and `subtask5.english`.
- Parameters were:
  - Batch size of 8
  - Weight decay of 0.01
  - Learning rate of $2 \times 10^{-5}$
  - 10 epochs.
- BERT and RoBERTa for comparison.
- DistilBERT took 15 minutes to train; BERT, RoBERTa 30.

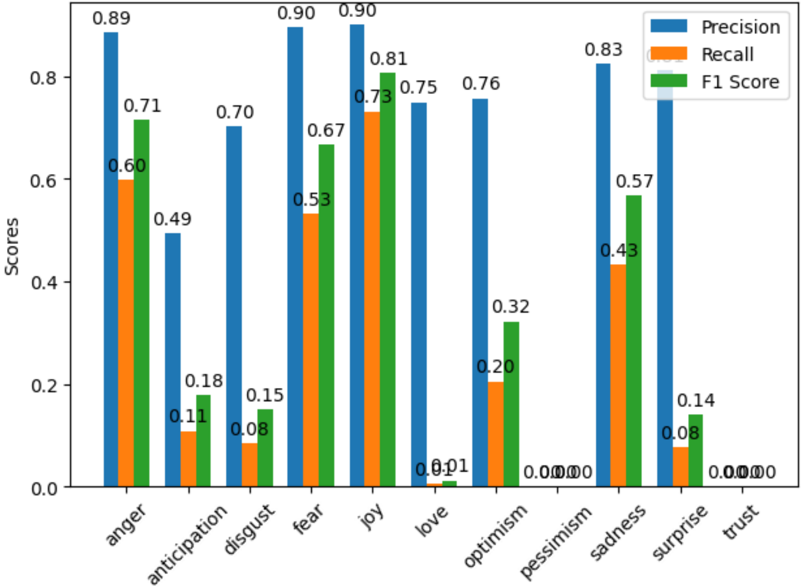
# Results: DistilBERT



# Results: BERT



# Results: RoBERTa



Anything to talk about?