Deep Learning Project Final Submission

Group 7:

Dhruv Bhutani	2019A7PS0080G
Kushal Joseph	2019A7PS0135G
Kaushal Khator	2019A7PS0180G

We present a brief report of our work based on the paper DistilBERT, a distilled version of BERT. Please note that all the working, explanation and diagrams, details and more have been included in the submitted python notebook.

Contributions

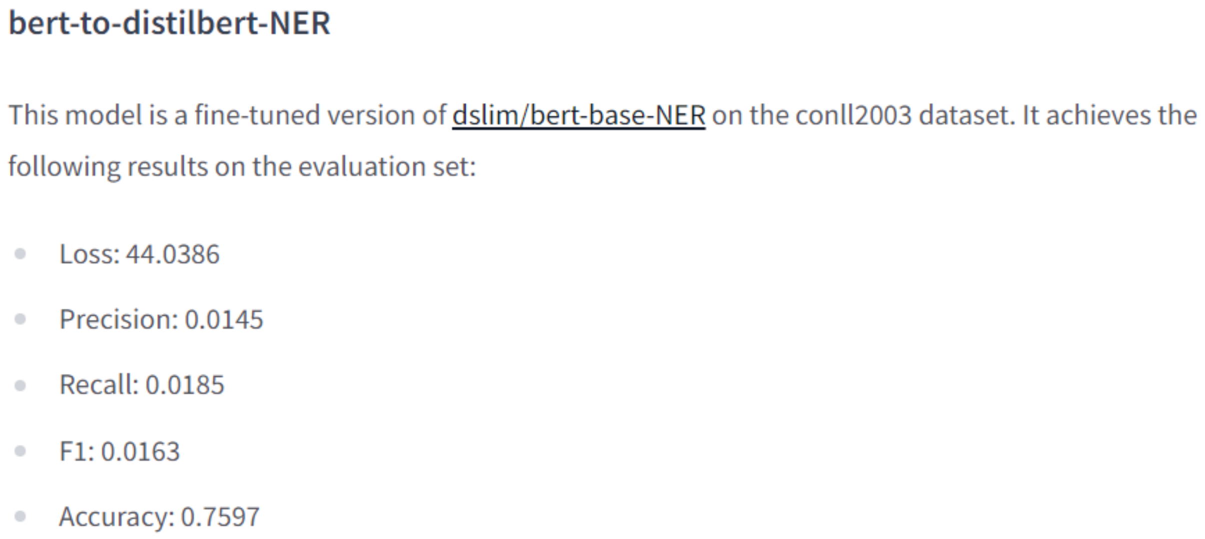
1. In the original paper - DistilBERT, the authors did not fine tune their model for the NER (Named Entity Recognition) task. For this purpose, we used a pretrained BERT model that was specifically trained (fine-tuned) for NER and distilled it to a distilBERT type model.
2. We performed hyperparameter tuning optimization for the model, in search for the best hyperparameters, including distillation hyperparameters, “alpha”, and “temperature”. We achieved an accuracy of >90%, but the model did not generalize well on basic input, we believe due to overfitting. So our final saved and uploaded model is the one with basic default values for hyperparameters. We have kept our hyperparameter tuning code in our notebook which can be run again (however, takes 1-2 hours for it)
3. We used a different loss function in our paper than mentioned in the paper. We used a KL-divergence loss function in the model, which is found in python from torch.nn.KLDivLoss. It is different, but the idea, as mentioned in the paper (of finding the similarity between the student and teacher’s output logits to a given input) is similar.
4. We have uploaded our model to HuggingFace, to allow anybody to use it. Anybody can run sample input online to test our model (on the HuggingFace interface), or you can just run a few lines of code to download the model, and test/try on an ipython notebook (Refer to transformers documentation for downloading pretrained models).

Results

BERT - Bidirectional Encoder Representations is a transformer-based machine learning technique for natural language processing. However, it’s large size may lead to computational inefficiencies on end systems.

We worked on distillation of BERT, that is, the process of transferring knowledge from a large model to a smaller one. We successfully **reduced the size of the model by approximately 50%**, and found that the model retains its capacity to generalize well on standard inputs. We got an **accuracy of approximately 76%** on the dataset, CONLL2003.

Here is a screenshot of our uploaded transformer model’s card:



bert-to-distilbert-NER

This model is a fine-tuned version of dslim/bert-base-NER on the conll2003 dataset. It achieves the following results on the evaluation set:

- Loss: 44.0386
- Precision: 0.0145
- Recall: 0.0185
- F1: 0.0163
- Accuracy: 0.7597

Features of our Work

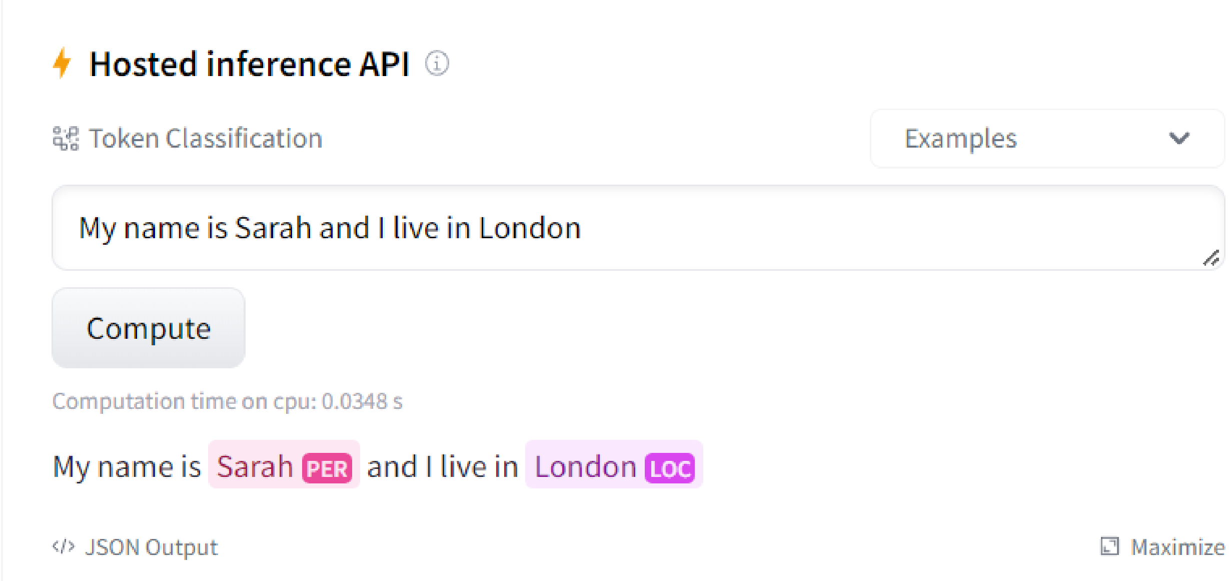
We used the technique of neural network distillation (as inspired from our paper, DistilBERT), to create a compact student model, by distilling from a pretrained (larger) model of BERT, which was fine tuned for Named Entity Recognition.

We created our **custom subclasses** of the transformers “Trainer” class for our training purposes, because we wanted to specifically train the student with the teacher in context (this is obviously not provided by default), so we had to include our distillation parameters.

As mentioned below, we have also uploaded our model online for anyone to see or use

HuggingFace API:
You can directly try out your own input sentence using the HuggingFace API on the following link: (A sample output has been shown for you)

[importsmart/bert-to-distilbert-NER · Hugging Face]()



My name is Sarah and I live in London

Computation time on cpu: 0.0348 s

My name is Sarah PER and I live in London LOC