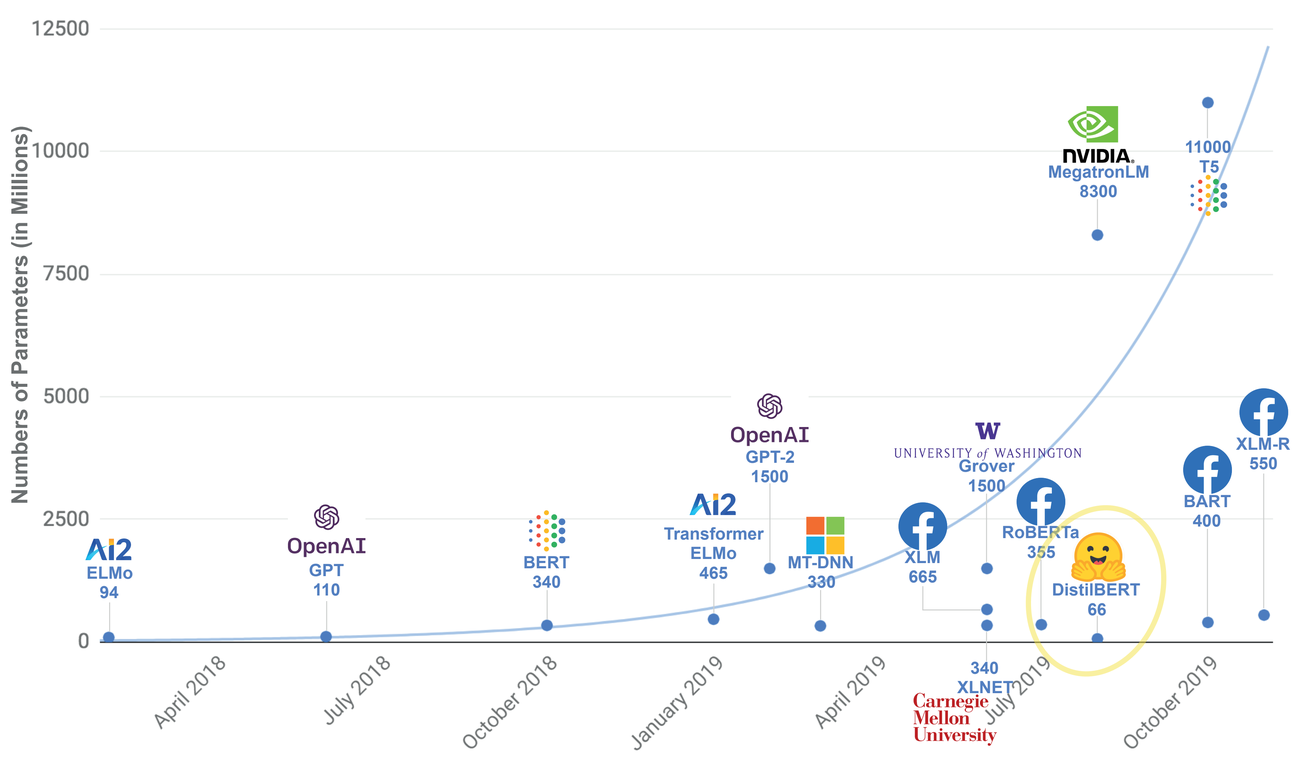
DistilBERT, a distilled version of BERT: smaller, faster cheaper and lighter

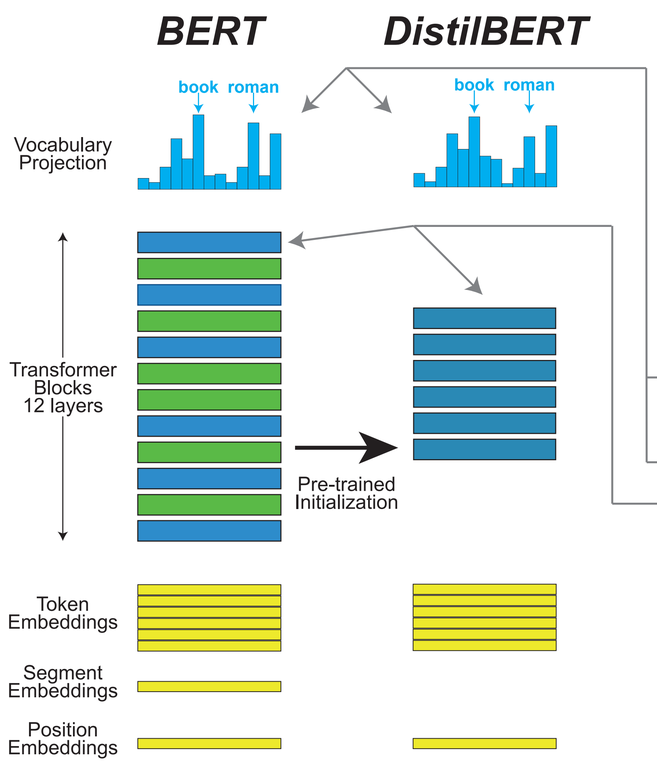
Victor SANH, Lysandre DEBUT, Julien CHAUMOND, Thomas WOLF
{victor, lysandre, julien, thomas}@huggingface.co

Large scale pre-trained Language Models are at the basis of many state-of-the-art approaches in NLP. The size of these models (**~ several hundred million parameters**) makes it difficult to use them in production or on device (**large memory footprint and slow inference**). Moreover, pre-training these large models have **a huge environmental cost**.

We show that it is possible to **reach similar performances** on downstream-tasks with **much smaller Language Models pre-trained with Knowledge Distillation**. Trained with a triple loss, **DistilBERT** is **40% smaller and 60% faster than BERT**, while reaching **97% of its original performance and being cheaper to train**, making DistilBERT a competitive option for on-the-edge applications.



Architecture and Initialization



- In most modern frameworks, matrix multiplications are highly optimized and the hidden size has a small impact on computation efficiency. We **reduce the number of Transformer layers by 2** for fast inference.
- We remove the **segment embeddings** and the **pooler**.
- We leverage the common dimensionality between teacher and student networks and **initialize the student from the teacher**.

Training: Knowledge Distillation Triple Loss

We train the student using a linear combination of 3 losses:

- **Knowledge Distillation Loss:** we used *Softmax-temperature* to reveal the *dark knowledge*.
- **Masked Language Modeling Loss:** the teacher's initial training loss.
- **Cosine Loss:** it aligns the directions of the hidden state vectors of the student and teacher.

Data and Compute

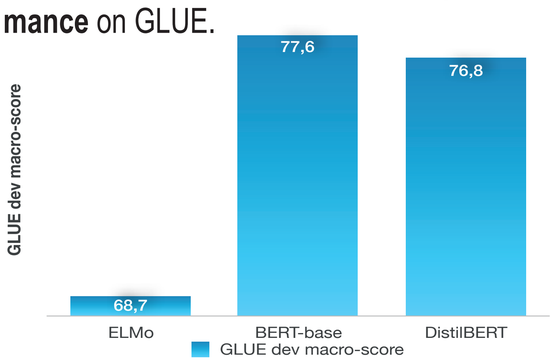
- We pre-train **DistilBERT** on the **same corpus as the original BERT** model: a concatenation of English Wikipedia and Toronto Book Corpus.
- DistilBERT was trained on **8 16GB V100 GPUs for ~90 hours**.

Paper, Code and Pre-trained Weights

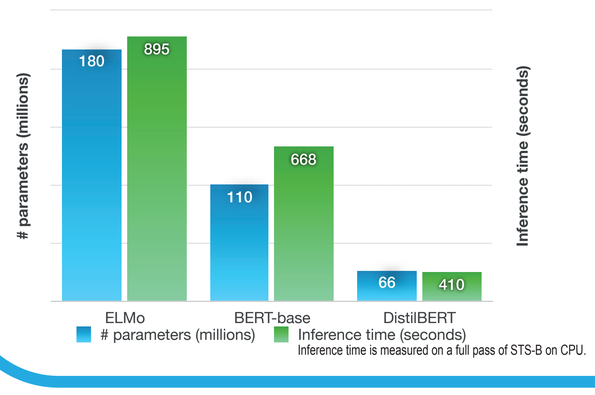
https://github.com/huggingface/transformers

Experiments

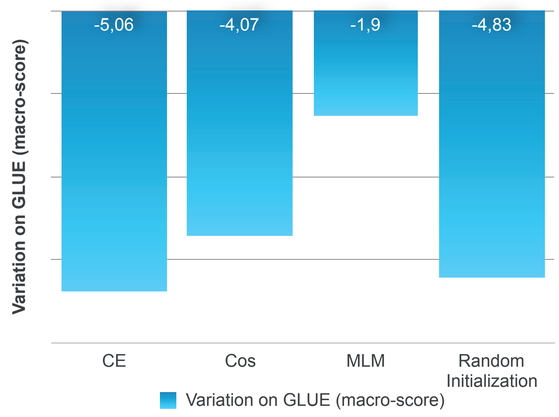
DistilBERT reaches **97% of BERT's performance** on GLUE.



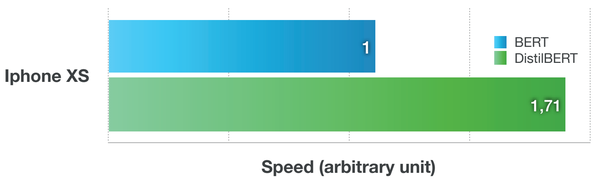
DistilBERT is **40% smaller and 60% faster** than BERT.



The **distillation losses** account for a large portion of the performance. **Initialization is key.**

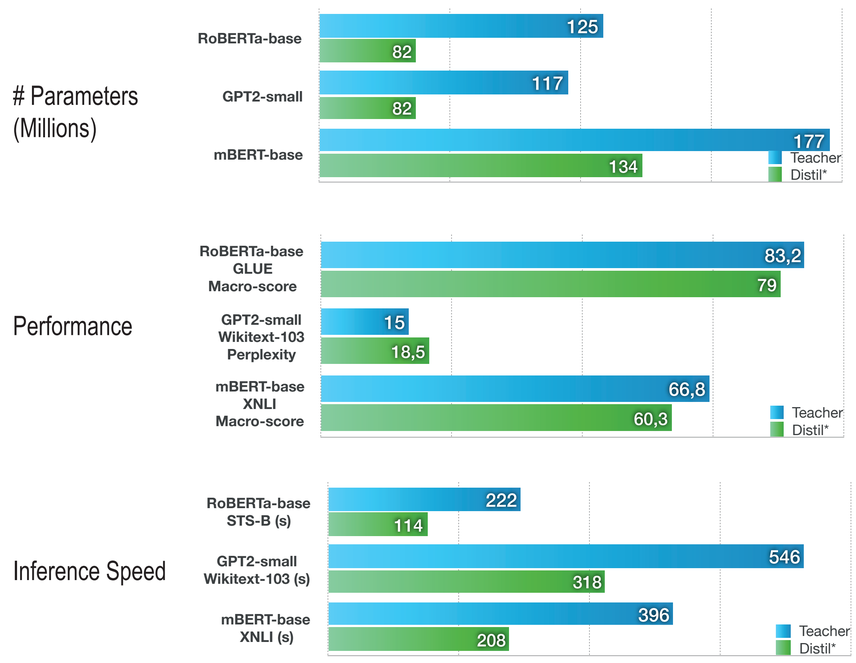


DistilBERT is a compelling alternative for **on-device applications**: light and fast. Further gains could be obtained with **quantization** techniques.



Additional Results

Pre-training by knowledge distillation is a poweful **general method** that can be applied to **a range of different models** (including models for Language Understanding, Language Generation or multi-linguality).



The 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing @ NeurIPS 2019
December 2019