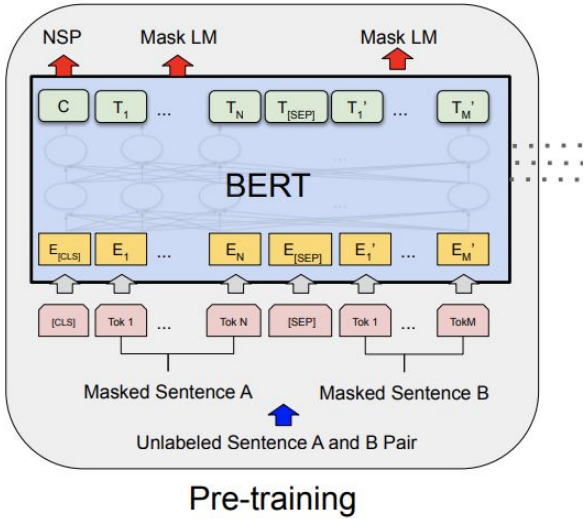
DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter

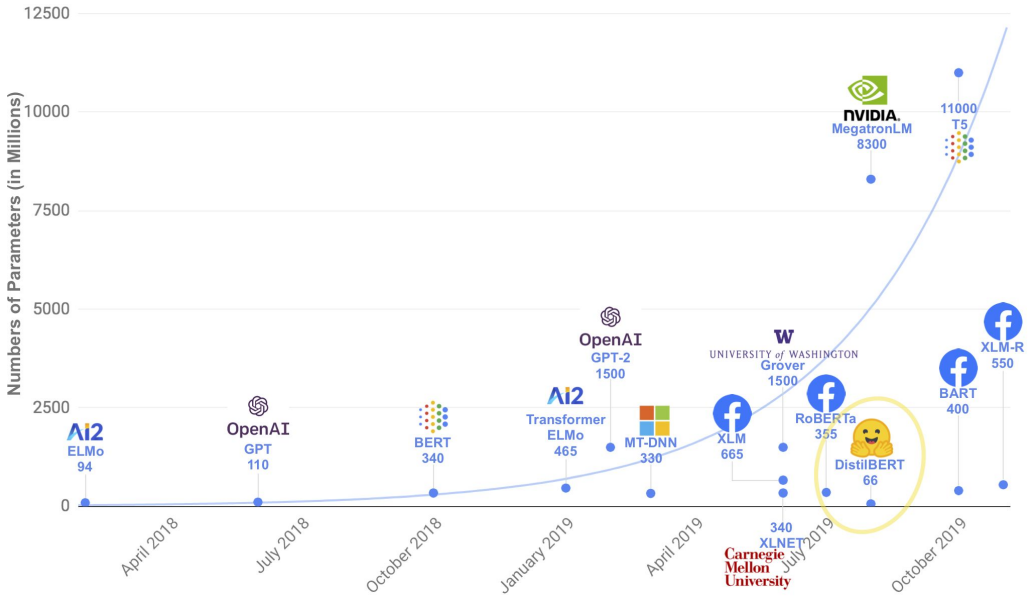
Victor Sanh, *Lysandre Debut, Julien Chaumond and Thomas Wolf*
Hugging Face

EMC2 Workshop @ NeurIPS 2019
December 2019

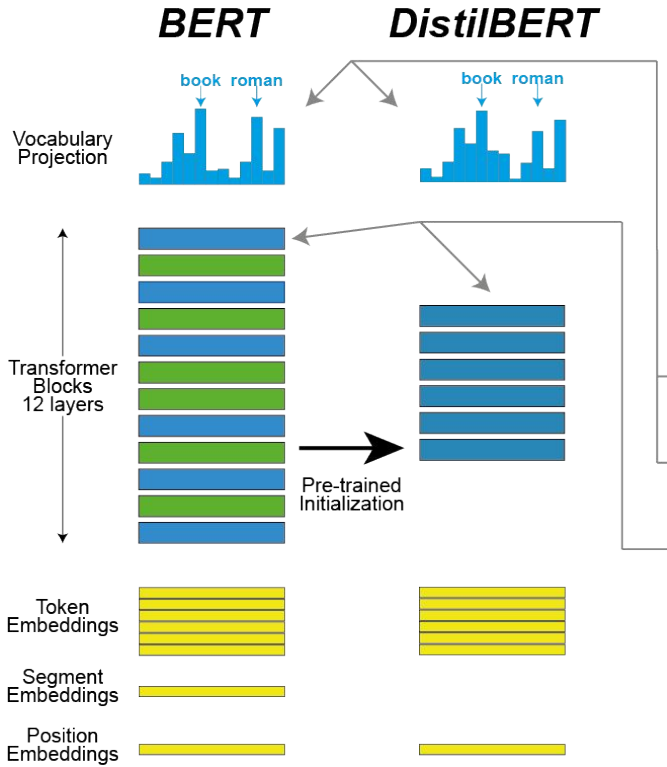
Large Transformer Language Models Pre-Training



Source: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova.In NACCL, 2019.



Architecture and Triple Loss



Triple Loss:

- Knowledge Distillation Loss
- Masked Language Modeling Loss
- Cosine Loss

Results: Sneak Peak

- **97%** of BERT's performance on GLUE
- **40% smaller** than BERT
- **CPU: 60% faster** than BERT in CPU
- **Device: 70% faster** than BERT (Iphone XS)

- **Distillation losses** drive the performance.
- A **general method** than can be applied to other models:
  - GPT2
  - RoBERTa
  - Multi-lingual BERT

Code, Pre-trained Weights and Paper

https://github.com/huggingface/transformers