



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

Paper By: Victor SANH, Lysandre DEBUT, Julien CHAUMOND, Thomas WOLF (Hugging Face)

Presented by: Deva Dharshni Rajarajeswari
Priyanka Singh



Challenges : Large-Scale Pre-Trained Language Models

- ❖ Operations of large model on edge applications
- ❖ Computational Constraints during training.
- ❖ Several Hundred million parameters are used.
- ❖ High Inferences Time
- ❖ Finally growing computational and memory requirements of these models may hamper wide adoption.



In Proposed System - DistilBERT

1. **DistilBERT is a smaller general-purpose language representation model.**
2. Later fine-tuned on wide range of tasks.
3. Aim is to reduce the size of a BERT model by 40%, while retaining 97% of its language understanding capabilities and being 60% faster.
4. Compressed models are small enough to run on the edge, e.g. on mobile devices.

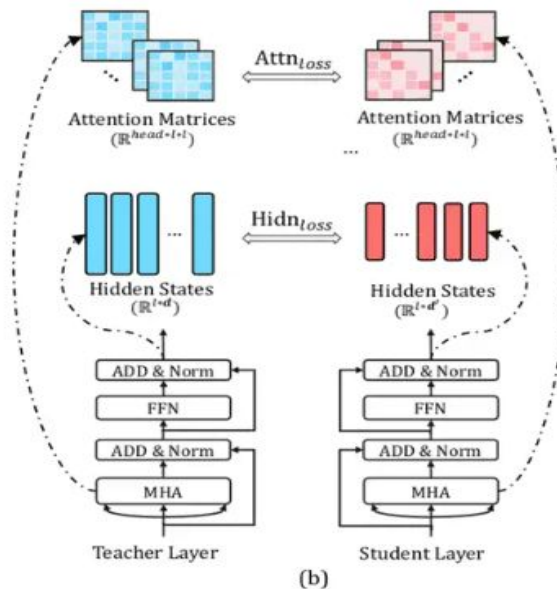
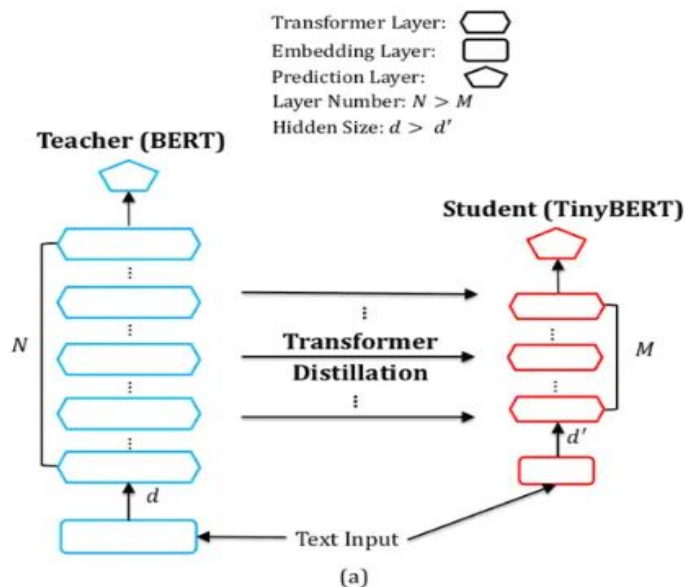





DistilBERT: a distilled version of BERT

- ❖ **Knowledge distillation** is a compression technique in which a compact model - the student - is trained to reproduce the behaviour of a larger model - the teacher - or an ensemble of models.
- ❖ **Triple Loss** is a linear combination of Distilled Loss ($L_{ce} = \sum_i p_i t_i \log(s_i)$ where t_i (resp. s_i)), masked language modeling loss (L_{mlm}) and cosine embedding loss (L_{cos}), which will tend to align the directions of the student and teacher hidden states vectors.

DistilBERT vs BERT



- 
- ❖ **Architecture Similarity:** DistilBERT maintains the basic architecture of BERT but with some modifications.
 - ❖ **Layer Reduction:** One of the primary changes in DistilBERT is the reduction in the number of layers by a factor of 2 compared to the original BERT model. This reduction helps in decreasing computational complexity and memory requirements while sacrificing some depth in the network.
 - ❖ **Removal of Token-Type Embeddings and Pooler:** DistilBERT removes the token-type embeddings and the pooler. Their removal in DistilBERT contributes to the model's simplification and reduction in computational overhead.
 - ❖ **Student initialization :** We initialize the student sub-network from teacher because of the common dimensionality.



Best practices for training DistilBERT

- ❖ **Distillation on Large Batches:** DistilBERT is distilled, or trained, using very large batches of data, which can improve computational efficiency and potentially enhance the model's generalization capabilities.
- ❖ **Gradient Accumulation:** Gradient accumulation involves accumulating gradients over multiple batches before updating the model's weights.
- ❖ **Dynamic Masking:** Dynamic masking involves applying masks to parts of the input data dynamically during training. Masking certain tokens allows the model to learn bidirectional representations by predicting masked tokens from the surrounding context.
- ❖ **Removal of Next Sentence Prediction Objective:** DistilBERT skips this next sentence prediction objective during training. This simplification might contribute to the efficiency of the training process.

Comparison	BERT October 11, 2018	RoBERTa July 26, 2019	DistilBERT October 2, 2019	ALBERT September 26, 2019
Parameters	Base: 110M Large: 340M	Base: 125 Large: 355	Base: 66	Base: 12M Large: 18M
Layers / Hidden Dimensions / Self-Attention Heads	Base: 12 / 768 / 12 Large: 24 / 1024 / 16	Base: 12 / 768 / 12 Large: 24 / 1024 / 16	Base: 6 / 768 / 12	Base: 12 / 768 / 12 Large: 24 / 1024 / 16
Training Time	Base: 8 x V100 x 12d Large: 280 x V100 x 1d	1024 x V100 x 1 day (4-5x more than BERT)	Base: 8 x V100 x 3.5d (4 times less than BERT)	[not given] Large: 1.7x faster
Performance	Outperforming SOTA in Oct 2018	88.5 on GLUE	97% of BERT-base's performance on GLUE	89.4 on GLUE
Pre-Training Data	BooksCorpus + English Wikipedia = 16 GB	BERT + CCNews + OpenWebText + Stories = 160 GB	BooksCorpus + English Wikipedia = 16 GB	BooksCorpus + English Wikipedia = 16 GB
Method	Bidirectional Transformer, MLM & NSP	BERT without NSP, Using Dynamic Masking	BERT Distillation	BERT with reduced parameters & SOP (not NSP)



Results

Evaluation of the language understanding and generalization capabilities of DistilBERT on the General Language Understanding Evaluation (GLUE) benchmark.

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

Table 2: **DistilBERT yields to comparable performance on downstream tasks.** Comparison on downstream tasks: IMDb (test accuracy) and SQuAD 1.1 (EM/F1 on dev set). D: with a second step of distillation during fine-tuning.

Model	IMDb (acc.)	SQuAD (EM/F1)
BERT-base	93.46	81.2/88.5
DistilBERT	92.82	77.7/85.8
DistilBERT (D)	-	79.1/86.9

Table 3: **DistilBERT is significantly smaller while being constantly faster.** Inference time of a full pass of GLUE task STS-B (sentiment analysis) on CPU with a batch size of 1.

Model	# param. (Millions)	Inf. time (seconds)
ELMo	180	895
BERT-base	110	668
DistilBERT	66	410



Ablation Study

- ❖ Here they investigate the influence of various components of the triple loss and the student initialization on the performances of the distilled model.
- ❖ Removal of Masked Language Modeling loss has little impact while the two distillation losses account for a large portion of the performance.

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



On Device Computation

- ❖ DistilBERT could be used for on-the-edge applications by building a mobile application for question answering.
- ❖ DistilBERT is 71% faster than BERT, and the whole model weighs 207 MB (which could be further reduced with quantization) on a smartphone (iPhone 7 Plus) against our previously trained question answering model based on BERT-base.
- ❖ Code is Available in link : <https://github.com/huggingface/swift-coreml-transformers>



Conclusion

- ❖ DistilBERT, a general-purpose pre-trained version of BERT, 40% smaller, 60% faster, that retains 97% of the language understanding capabilities.
- ❖ A general-purpose language model can be successfully trained with distillation and analyzed the various components with an ablation study.
- ❖ DistilBERT is a compelling option for edge applications.



Thank You