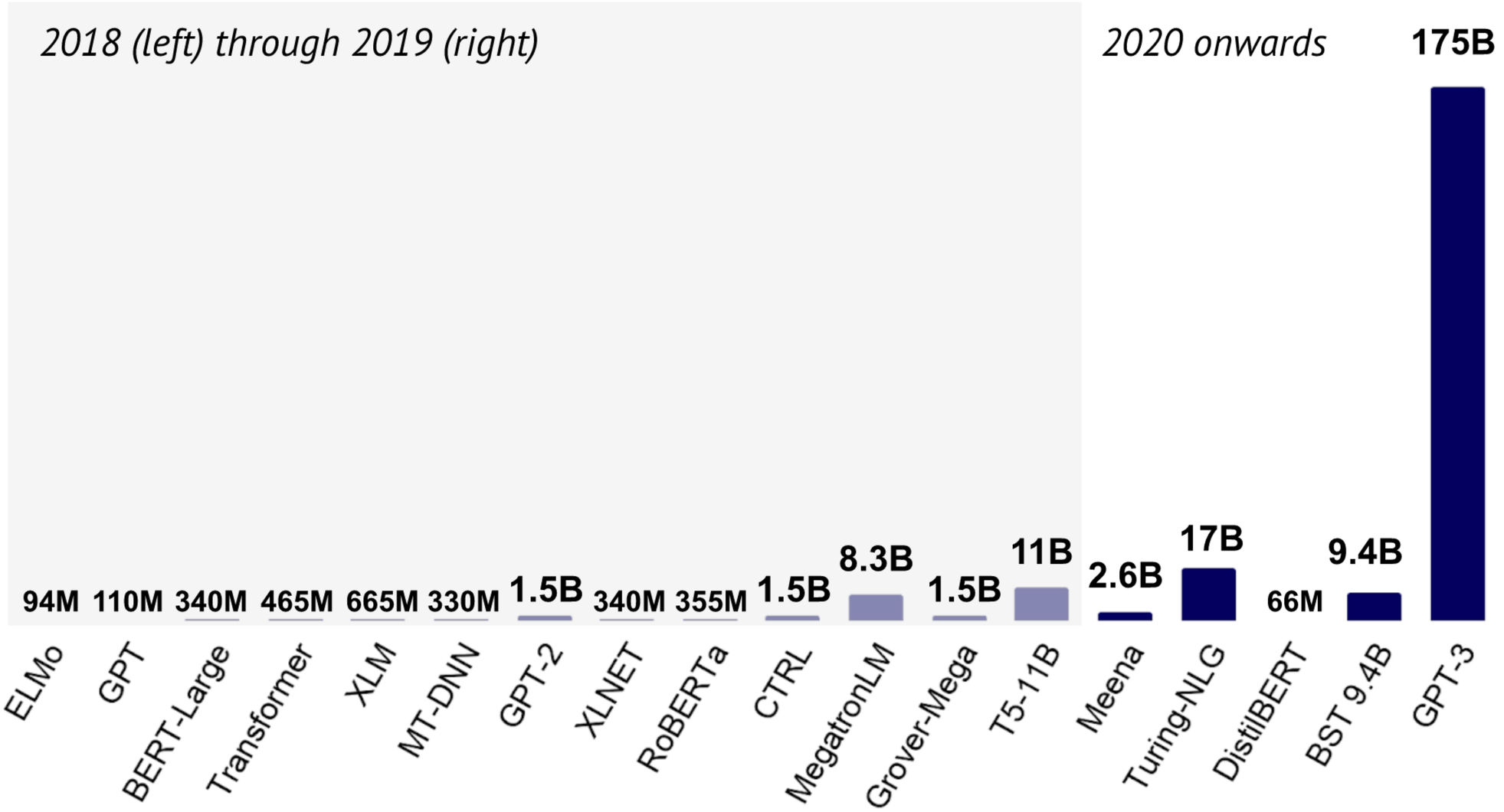
DistilBERT,
a distilled version of BERT:
smaller, faster, cheaper, and lighter

QRAFT | AXE
GUIJIN.SON

2021.07.28

Recent NLP Trends

DistilBERT : a distilled version of BERT



Wider Adoption of NLP

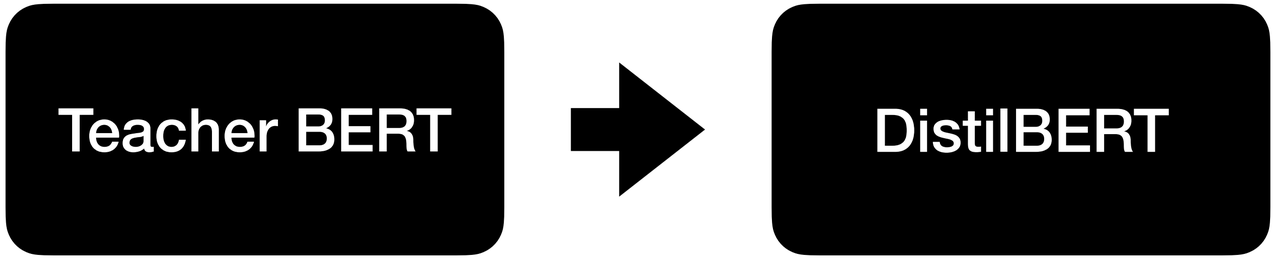
DistilBERT : a distilled version of BERT

“ while operating these models on-device in real-time has the potential to enable novel and interesting language processing applications, the growing computational and memory requirements of these models may hamper wide adoption."

Model Architecture

DistilBERT : a distilled version of BERT

- Student Architecture
  - Changes
    - 1/2 Number of Layers
  - Removed
    - Token-Type Embedding
    - Pooler Output
- Student Initialization
  - Brought from the Teacher Model



Training Loss

DistilBERT : a distilled version of BERT

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

- **Triple Loss:**
  - **MLM Loss**
  - **Distillation Loss**
  - **Cosine Similarity Loss**

Performance

DistilBERT : a distilled version of 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

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