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

김재화

Abstract

- **General purpose language representation model**, called DistilBERT,
- While most prior work investigated the use of distillation for building task-specific models,
- we leverage **knowledge distillation during the pre-training phase** and show that it is possible to reduce the **size of a BERT model by 40%**, while **retaining 97% of its language understanding capabilities** and being **60% faster.**

Introduction

- **large-scale pre-trained language models becoming a basic tool** in many NLP tasks

- they often have **several hundred million parameters** and current research on pre-trained models indicates that **training even larger models still leads to better performances** on downstream tasks.

- several concerns
  - the **environmental cost** of exponentially scaling these models’ computational requirements
  - while operating **these models on-device in real-time** … <u>the growing computational and memory requirements</u> of these models may <u>hamper wide adoption</u>.

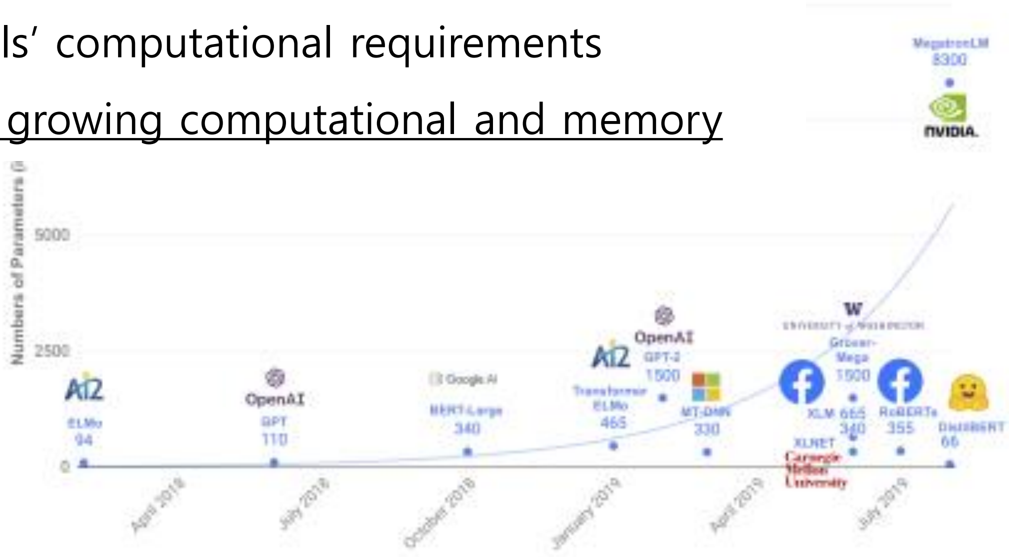

Figure 1: **Parameter counts of several recently released pretrained language models.**

Architecture

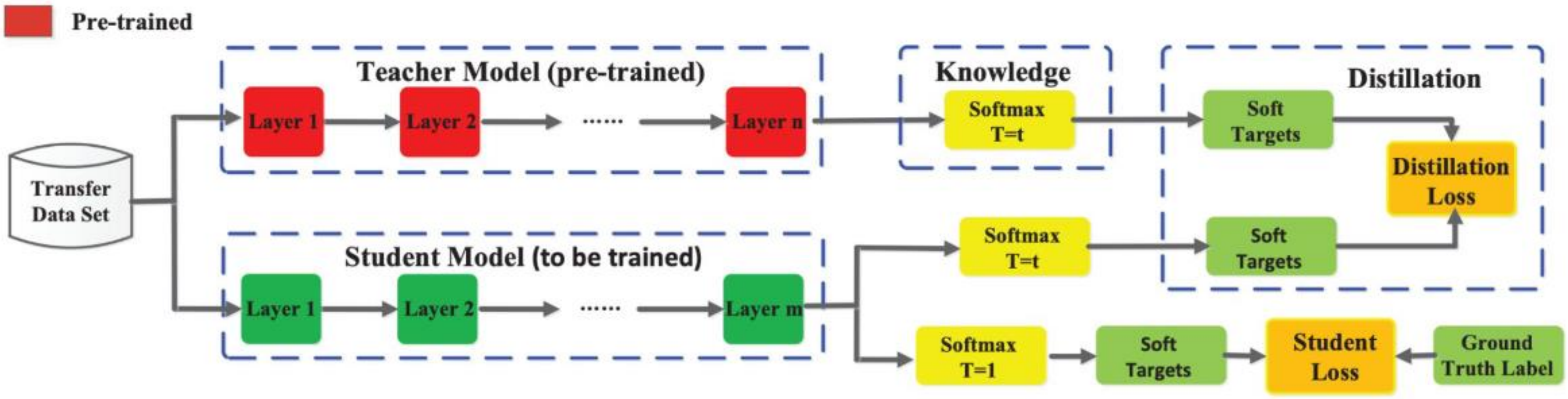

Fig. 5 The specific architecture of the benchmark knowledge distillation (Hinton et al., 2015).

$$q_i = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)} \tag{1}$$

https://arxiv.org/pdf/2006.05525.pdf

Training loss

- The student is trained with a **distillation loss** over the soft target probabilities of the teacher.

$$L_{ce} = \sum_i t_i * \log(s_i)$$

- The supervised training loss, in our case the **masked language modeling loss** (L_{mlm})

- We found it beneficial to add a **cosine embedding loss** (L_{cos}) which will tend to align the <u>directions of the student and teacher hidden states vectors.</u>

DistilBERT: a distilled version of BERT

Student architecture

- the same general architecture as BERT.

- The **token-type embeddings (segment embedding) and the pooler are removed** while the number of **layers is reduced by a factor of 2**.
  - Transformer architecture (linear layer and layer normalisation) are highly optimized
  - the last dimension of the tensor (hidden size dimension) have a smaller impact on computation efficiency (for a fixed parameters budget**)** than variations on other factors like the number of layers.

- Thus we **focus on reducing the number of layers**.

Student initialization / Distillation / Data and compute power

- **Student initialization**
  - we initialize the student from the teacher by taking one layer out of two.

- **Distillation**
  - very large batches leveraging gradient accumulation (up to 4K examples per batch)
  - using dynamic masking
  - without the next sentence prediction (NSP)

- **Data and compute power**
  - the same corpus as the original BERT model
  - 8 16GB V100 GPUs for approximately 90 hours.

Experiments

General Language Understanding

- We report scores on the development sets for each task by fine-tuning DistilBERT
  - Among the 9 tasks, DistilBERT is always on par or improving over the ELMo baseline.
  - DistilBERT also compares surprisingly well to BERT, retaining 97% of the performance with 40% fewer parameters.

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

https://arxiv.org/pdf/1910.01108.pdf

Downstream task / Size and inference speed

- **downstream tasks**
  - only 0.6% point behind BERT in test accuracy on the IMDb benchmark
  - On SQuAD, DistilBERT is within 3.9 points of the full BERT
  - two successive steps
    - one during the pre-training phase
    - one during the adaptation phase

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

- **Size and inference speed**
  - DistilBERT has 40% fewer parameters than BERT and is 60% faster than BERT.

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

- we investigate **the influence of various components of the triple loss** and **the student initialization on the performances** of the distilled model.

- removing the **Masked Language Modeling loss has little impact** while the two distillation losses account for a large portion of the performance

Table 4: **Ablation study.** Variations are relative to the model trained with triple loss and teacher weights initialization.

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

Conclusion

- general-purpose pre-trained version of BERT, 40% smaller, 60% faster, that retains 97% of the language understanding capabilities.

- We further demonstrated that DistilBERT is a compelling option for edge applications.