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

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

Abstract

As Transfer Learning from large-scale pre-trained models becomes more prevalent in Natural Language Processing (NLP), operating these large models in on-the-edge and/or under constrained computational training or inference budgets remains challenging. In this work, we propose a method to pre-train a smaller general-purpose language representation model, called DistilBERT, which can then be fine-tuned with good performances on a wide range of tasks like its larger counterparts. 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. To leverage the inductive biases learned by larger models during pre-training, we introduce a triple loss combining language modeling, distillation and cosine-distance losses. Our smaller, faster and lighter model is cheaper to pre-train and we demonstrate its capabilities for on-device computations in a proof-of-concept experiment and a comparative on-device study.

1 Introduction

The last two years have seen the rise of Transfer Learning approaches in Natural Language Processing (NLP) with large-scale pre-trained language models becoming a basic tool in many NLP tasks [Devlin et al., 2018, Radford et al., 2019, Liu et al., 2019]. While these models lead to significant improvement, they often have several hundred million parameters and current research[1] on pre-trained models indicates that training even larger models still leads to better performances on downstream tasks.

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

The trend toward bigger models raises several concerns. First is the environmental cost of exponentially scaling these models' computational requirements as mentioned in Schwartz et al. [2019], Strubell et al. [2019]. Second, 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.

[1] See for instance the recently released MegatronLM (https://nv-adlr.github.io/MegatronLM)

EMC^2: 5th Edition Co-located with NeurIPS'19

Background

BERT shows great performance in NLP tasks, but due to its large size and heavy computation, it's not ideal for real-time services or deployment on mobile/edge devices.

What this paper proposes

This paper introduces **DistilBERT**, a lighter and faster version of BERT that keeps most of its performance while being more efficient and practical for real-world use.

Key idea: Knowledge Distillation

The core idea is to use *knowledge distillation* — a process where a smaller model (the **student**, DistilBERT) learns from a larger, pre-trained model (the **teacher**, BERT). This allows the student to mimic the teacher's behavior and learn faster with fewer parameters.

Model Architecture

- Uses only **6 out of BERT-base's 12 layers** — so it's half the depth.
- Keeps the same hidden size and basic transformer structure.
- Removes some components like token-type embeddings and the pooler, which are not essential.

Training Loss

The training combines three types of loss:

- **Distillation loss** (to learn from the teacher's soft outputs)
- **Masked Language Modeling (MLM) loss** (same as BERT's original objective)
- **Cosine embedding loss** (to align internal representations between student and teacher)

Initialization

To kick-start training, every other layer from the teacher model is copied into the student — this gives the smaller model a strong starting point.

Results

- **Performance**: Achieves about **97% of BERT's performance** on GLUE, SQuAD, IMDb, etc.
- **Speed and efficiency**:
  - Inference is up to **60–71% faster** than BERT.
  - Model size is reduced by **40%**, down to around **207MB**.
- **On-device testing**: Runs in real-time on devices like the iPhone 7 Plus, proving it's usable for on-device AI.

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

Why it matters

This makes DistilBERT a strong fit for mobile or edge scenarios where computing power is limited. Before, large models like BERT could only be run on servers, but DistilBERT shows that **powerful NLP can now work on smaller devices too — a big step for bringing real-time language understanding to everyday applications.**

배경 :
BERT는 뛰어난 성능을 가지고 있지만 크키와 연산량이 커서 실제 서비스나 모바일 환경에서는 사용하기 어렵다는 리미테이션이 있음

이 논문에서는 BERT성능을 유지하지만 더 작고 빠르며 효율적인 DistillBert 를 제안함

Concept:

1. **Knowledge Distillation): 이미 학습된 (Bert, Teacher) 모델과 작은 모델 (**DistilBERT, Student**)로 전달하여 학습**

모델 구조 :

BERT-base의 12개 레이어 중 6개만 사용(레이어 수 절반).
- 히든 사이즈 등 기본 구조는 동일하게 유지.
- 토큰 타입 임베딩과 풀러(pooler) 등 일부 불필요한 요소 제거.

학습 손실:
- 증류 손실(distillation loss) + 마스킹 언어 모델링 손실(MLM loss) + 코사인 임베딩 손실을 조합.

초기화 방법

Teacher의 2개 중 1개 레이어를 Student의 각 레이어에 복사하여 초기화.

결과 :

성능:
- GLUE, SQuAD, IMDb 등 다양한 자연어 처리 벤치마크에서 BERT 대비 약 97%의 성능 유지.

속도 및 효율성:
- 추론 속도는 BERT 대비 최대 60~71% 빠름.
- 모델 크기는 40% 감소(207MB).

추론 속도가 BERT 대비 60~71% 더 빠르며, 모델 크기도 207MB로 줄어 모바일·엣지 디바이스에서도 실제로 동작 가능함을 실험으로 입증했습니다.

온디바이스 실험:
- iPhone 7 Plus 등 모바일 기기에서도 실시간 질문 응답이 가능함을 실험으로 입증.

실제 서비스와 온디바이스 AI에 매우 적합한 솔루션--> 대형 언어 모델이 서버 환경에서만 주로 사용될 수 있었지만, DistilBERT의 등장으로 모바일,엣지 디바이스 등 자원이 제한된 환경에서도 고성능 자연어처리 기술을 활용

DistilBERT 논문 요약
1. 배경 및 목적
- BERT는 뛰어난 성능을 자랑하지만, 크기와 연산량이 커서 실제 서비스나 모바일 환경에서 사용하기 어렵다는 한계가 있음.
- 본 논문에서는 BERT의 성능을 최대한 유지하면서, 더 작고 빠르며 효율적인 모델인 **DistilBERT**를 제안함.

2. 주요 기법 및 특징
- **지식 증류(Knowledge Distillation):**
이미 학습된 큰 모델(BERT, Teacher)의 예측 정보를 작은 모델(DistilBERT, Student)로 전달하여 학습.
- **모델 구조:**
- BERT-base의 12개 레이어 중 6개만 사용(레이어 수 절반).
- 히든 사이즈 등 기본 구조는 동일하게 유지.
- 토큰 타입 임베딩과 풀러(pooler) 등 일부 불필요한 요소 제거.
- **초기화 방법:**
- Teacher의 2개 중 1개 레이어를 Student의 각 레이어에 복사하여 초기화.
- **학습 손실:**
- 증류 손실(distillation loss) + 마스킹 언어 모델링 손실(MLM loss) + 코사인 임베딩 손실을 조합.

3. 실험 결과
- **성능:**
- GLUE, SQuAD, IMDb 등 다양한 자연어 처리 벤치마크에서 BERT 대비 약 97%의 성능 유지.
- **속도 및 효율성:**
- 추론 속도는 BERT 대비 최대 60~71% 빠름.
- 모델 크기는 40% 감소(207MB).
- **온디바이스 실험:**
- iPhone 7 Plus 등 모바일 기기에서도 실시간 질문 응답이 가능함을 실험으로 입증.

4. 의의 및 활용
- DistilBERT는 서버뿐 아니라 모바일, 엣지 디바이스 등 다양한 환경에서 빠르고 효율적으로 자연어 처리 모델을 사용할 수 있게 해줌.
- Hugging Face Transformers 라이브러리 등에서 쉽게 활용 가능.

한 줄 요약:
DistilBERT는 BERT의 성능을 거의 유지하면서도, 훨씬 작고 빠르게 동작하는 경량화 언어 모델로,
실제 서비스와 온디바이스 AI에 매우 적합한 솔루션입니다.

핵심 제안
- **DistilBERT**라는 새로운 경량화 언어 모델을 제안합니다.
- 기존 BERT 모델의 구조를 바탕으로 하되, 레이어(층) 수를 절반으로 줄이고, 불필요한 구성 요소를 제거해 모델 크기를 40% 이상 줄였습니다.
- **지식 증류(knowledge distillation)** 기법을 사용해, 큰 모델(teacher)의 예측 분포 등 풍부한 정보를 작은 모델(student)에 효과적으로 전달하여 학습합니다.
- 학생 모델의 초기값을 교사 모델의 일부 레이어 파라미터로 직접 복사해, 학습 효율과 수렴 속도를 높였습니다.

주요 성과
- **성능 유지:**
DistilBERT는 원본 BERT 대비 약 97%의 성능을 유지합니다.
(GLUE, SQuAD 등 주요 자연어 처리 벤치마크에서 검증)
- **속도 및 효율성:**
추론 속도가 BERT 대비 60~71% 더 빠르며, 모델 크기도 207MB로 줄어 모바일·엣지 디바이스에서도 실제로 동작 가능함을 실험으로 입증했습니다.
- **실제 적용 가능성:**
iPhone 7 Plus 등 스마트폰 환경에서 질문응답 모델로 구현해, 온디바이스 AI 서비스의 실현 가능성을 보여주었습니다.

즉, 이 논문은 **BERT의 성능을 거의 유지하면서도 훨씬 작고 빠른 모델을 만드는 방법**을 제안하고, 실제 다양한 환경에서의 실험을 통해 **경량 AI 모델의 실용성**을 입증한 것이 가장 큰 성과입니다.

주요 발견
- **지식 증류(knowledge distillation) 기법**을 활용해, 기존 BERT 모델의 성능을 거의 유지하면서도 크기와 속도를 크게 개선한 **DistilBERT**라는 경량화 모델을 개발했습니다.
- DistilBERT는 BERT 대비 **모델 크기가 40% 줄고, 추론 속도는 60~71% 빨라졌음에도 불구하고, 성능은 약 97%까지 유지**할 수 있음을 다양한 자연어처리 벤치마크(GLUE, SQuAD 등)에서 실험적으로 입증했습니다.
- **학생 모델의 초기화**를 교사 모델의 일부 레이어 파라미터로 직접 복사하는 전략을 적용해, 작은 모델이 더 빠르고 효과적으로 수렴할 수 있음을 보여주었습니다.
- DistilBERT는 실제 **모바일 기기(예: iPhone 7 Plus)**에서도 실시간 질문응답 등 온디바이스 AI 활용이 가능함을 실험으로 확인했습니다.

의미
- **대형 AI 모델의 실용화:**
기존에는 대형 언어 모델이 서버 환경에서만 주로 사용될 수 있었지만, DistilBERT의 등장으로 모바일·엣지 디바이스 등 자원이 제한된 환경에서도 고성능 자연어처리 기술을 활용할 수 있게 되었습니다.
- **효율적 AI 개발의 방향 제시:**
단순히 모델 크기를 줄이는 것이 아니라, 지식 증류와 같은 최적화 전략을 활용하면 성능 저하 없이 경량화가 가능하다는 점을 실증적으로 보여주었습니다.
- **AI 기술의 대중화와 확장:**
이 연구는 앞으로 더 많은 실생활 서비스(챗봇, 음성인식, 번역 등)에서 AI가 폭넓게 활용될 수 있는 기반을 마련했다는 점에서 큰 의의를 가집니다.

즉, 이 논문은 **"작고 빠르면서도 똑똑한 AI 모델"**의 실현 가능성을 입증했고,
이는 AI 기술의 접근성과 응용 가능성을 크게 넓힌 중요한 연구적 발견입니다.

학문적 측면
- **지식 증류(knowledge distillation)의 효과 실증**
기존 대형 언어모델의 지식을 효과적으로 작은 모델에 전달할 수 있음을 실험적으로 입증했습니다.
이는 모델 경량화 연구에서 이론적 근거와 실질적 방법론을 동시에 제시한 것으로,
향후 다양한 딥러닝 모델의 효율화 연구에 중요한 참고 사례가 됩니다.
- **모델 구조 최적화 전략 제시**
단순히 파라미터 수만 줄이는 것이 아니라, 레이어 수를 줄이는 것이 연산 효율에 더 효과적임을 밝혔습니다.
이는 트랜스포머 기반 모델 설계와 최적화에 있어 새로운 방향성을 제시한 것으로 평가받습니다.
- **학습 안정성 향상을 위한 초기화 방법**
교사 모델의 일부 레이어 파라미터를 학생 모델에 직접 복사하는 초기화 전략을 도입해,
작은 모델의 학습 수렴성과 성능을 동시에 높일 수 있음을 보여주었습니다.

실용적 측면
- **온디바이스 AI 실현 가능성 입증**
DistilBERT는 기존 BERT 대비 40% 더 작고, 60~71% 더 빠르면서도,
성능 저하를 최소화해 스마트폰 등 모바일·엣지 디바이스에서도 실시간 자연어처리 서비스가 가능함을 실험으로 증명했습니다.
- **AI 기술의 대중화와 접근성 확대**
서버가 아닌 다양한 환경(모바일, IoT, 임베디드 등)에서도 고성능 언어모델을 사용할 수 있게 되어,
챗봇, 음성인식, 번역, 요약 등 실생활 서비스의 확장성과 실용성을 크게 높였습니다.
- **산업적 활용 및 비용 절감 효과**
더 작은 모델로도 기존 대형 모델 수준의 품질을 제공할 수 있어,
서비스 운영 비용과 자원 소모를 줄이고, 다양한 기업과 개발자가 AI를 쉽게 도입할 수 있는 기반을 마련했습니다.

요약:
이 논문은 대형 언어모델의 경량화 및 실용화에 있어 학문적으로는 새로운 이론과 방법론을,
실용적으로는 실제 서비스와 산업 적용 가능성을 동시에 입증한 중요한 연구입니다.
이는 AI 기술의 발전 방향과 활용 범위를 크게 넓히는 데 기여합니다.