GLUE benchmark를 이용하여 자연어 모델 성능 조사 및 연구

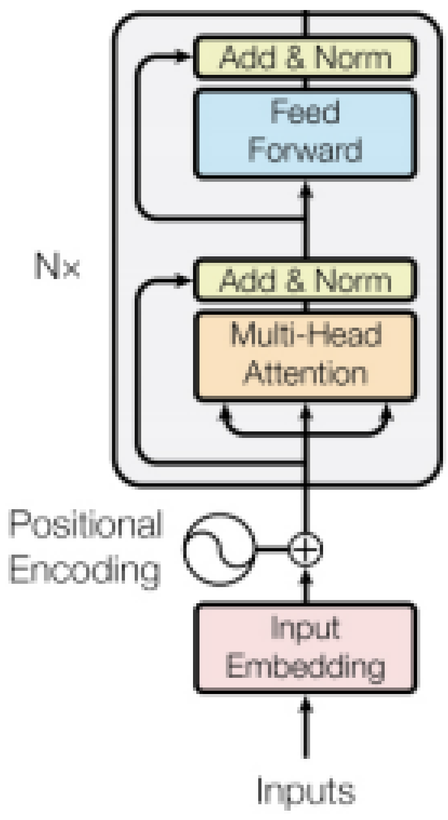
이 름 박재찬 지도교수 조현석

연구소개

딥러닝 기술과 딥러닝 모델의 성능이 나날이 발전해가면서 정보 추출, 맞춤법 교정, 언어 번역 등의 자연어 처리(Natural Language Processing) 분야도 탄력을 받고 있다. 자연어 모델의 성능이 개선되면서 우리는 실생활에서 더 알맞은 광고와 더 정확한 번역기 등을 이용할 수 있게 되었고, 이는 높은 가독성과 편의성으로 사용자들에게 만족감을 주었다.
그러나 자연어 모델의 성능이 개선될수록 더 많은 layer와 parameter가 필요했고, 모델의 크기는 점차 커지게 되어 일반 컴퓨터에서 학습시키기엔 너무 많은 시간과 비용, 메모리를 요구하게 되었다. 우리는 이런 부분에 주목하여 모델의 성능을 크게 떨어뜨리지 않으면서 모델의 크기를 획기적으로 줄일 수 있는 방법을 찾아보기로 했다.

연구배경

▶ GLUE
GLUE는 자연어 이해 시스템을 발전시키는 것을 목적으로 만들어진 9가지 NLU 작업 모음으로 되어있는 dataset을 모아둔 benchmark이다. 모델이 자연어에 대한 일반적인 이해능력을 가지고 있는지 테스트한다. score는 정확도를 의미한다.



▶ BERT
BERT는 구글에서 개발한 범용 language model로, 이 모델을 바탕으로 다양한 파생 모델이 연구되었다.

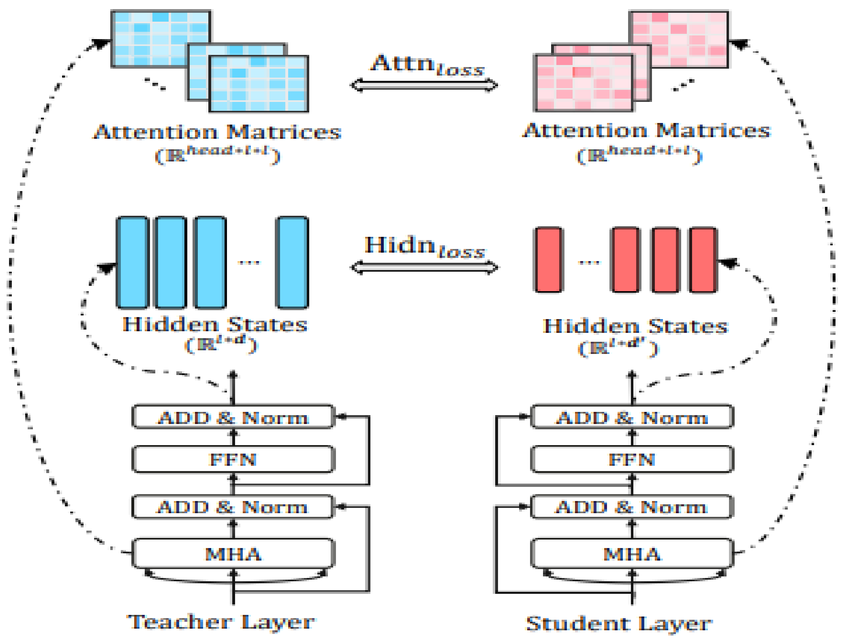
▶ 모델 경량화 기법 : 성능에 큰 차이가 없는 선에서 불필요한 head, layer나 parameter를 제거하여 모델의 크기를 줄임으로써, 메모리 크기와 프로세서 성능 등의 HW 문제를 해결하고 학습 속도를 높일 수 잇는 최적화 기법이다.
Ex) Factorization, sharing, Knowledge Distillation 등

모델 비교

▶ ALBERT : Factorization embedding parameterization을 이용해 input layer의 parameter 수를 줄이고, parameter sharing을 통해 각 layer 간에 같은 parameter를 공유하여 사용함으로써 모델 크기를 줄였다.

▶ DistilBERT : BERT 모델의 pre-train 과정에서 Knowledge Distillation 기법을 적용하여 layer 수를 줄이고 추론 속도를 향상시키면서 비슷한 성능을 보인다.

▶ TinyBERT : Attention과 hidden state 각각에 대해 Knowledge distillation을 하고 hidden state distillation과 비슷하게 embedding layer에 대해서도 distillation을 진행한다.



<TinyBERT 구조>

DistilBERT와 TinyBERT는 같은 Knowledge Distillation을 적용했지만 다음과 같은 차이가 있다.

KD Methods	KD at Pre-training Stage					KD at Fine-tuning Stage				
	INIT	Embd	Attn	Hidn	Pred	Embd	Attn	Hidn	Pred	DA
Distilled $BiLSTM_{SOFT}$									✓	✓
BERT-PKD	✓							✓[3]	✓	
DistilBERT	✓				✓[4]				✓	
TinyBERT (our method)		✓	✓	✓		✓	✓	✓	✓	✓

<DistilBERT와 TinyBERT의 KD 위치>

성능 비교 분석 및 결론

	parameter	speed	GLUE benchmark
BERT	Base : 110M Large : 340M	1x 0.2x	79.6 82.1
Pruning BERT	60% of BERT head	1.2x	79.4
ALBERT	1/18 of BERT-large	1.7x of BERT-large	89.4
Q-BERT(8bit)	1/4 of BERT-base	3.7x	99.3% of BERT
DistilBERT	6layer : 66M	1.6x	97% of BERT
TinyBERT	4layer : 14.5M	9.4x	96.8% of BERT

기존 BERT-base 모델과 비교했을 때 Knowledge Distillation을 적용한 TinyBERT가 다른 모델들보다 효율적이라는 결론을 내렸다. 그리고 pre-train 과정에서만 KD를 사용한 DistilBERT보다 pre-train 과정과 task-specific learning 과정 모두에 KD를 적용한 TinyBERT가 모델 크기가 훨씬 작고 훨씬 빠른 것을 통해 KD를 적용하는 layer의 수와 모델 성능이 비례한다는 결론을 내렸다.

후속 연구

KD를 사용한 모델의 경우 teacher로 어떤 모델을 쓰는지에 따라 추가적인 성능 변화가 일어나므로, 추후 teacher model size에 따른 성능 변화 및 개선에 대한 연구를 진행하려 한다.