MiniBridgeBERT: Efficient Compression of Transformer Models Through Stepwise Block Pruning and Progressive Knowledge Distillation

정성룡 (AI학과), 오규안 (AI학과)

2024 CUAI 중앙대학교 인공지능 학회 동계 컨퍼런스
Proceeding of 2024 Chung-Ang University Artificial Intelligence Winter Conference

Abstract

본 연구에서는 사전 훈련 없이, 미세 조정 중 지식 증류하는 MiniBridgeBERT(MB-BERT)를 제안한다. MB-BERT는 Block-wise Pruning과 Low Rank Bridge Layer를 통해 모델을 경량화하고, 점진적 지식 증류를 활용하여, 성능 저하를 최소화한다. CoLA 데이터셋에서 MB-BERT는 모델 크기를 37% 줄이면서, 높은 MCC 점수를 유지하는 데 성공했으며, 이는 자원 제한 환경에서도 효과적으로 활용가능한 모델 경량화 방법임을 입증한다.

Introduction

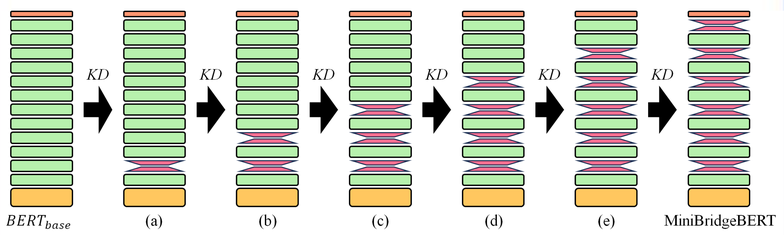
대형 언어 모델은 모델 크기가 증가할수록 성능이 향상되지만, Edge Device 실시간 애플리케이션에서 활용하기 위해서는 모델 경량화가 필수적이다. 기존의 모델 압축 기법으로는 양자화(Quantization), 가지치기(Pruning), 지식 증류(Knowledge Distillation, KD) 등이 있으며, 특히 KD는 성능을 유지하면서 모델을 경량화하는 데 효과적인 방법이다. 그러나 대부분의 KD 방식은 사전 학습된 교사 모델을 활용해야 하며, 이로 인해 높은 계산 비용이 요구된다. 본 연구에서는 사전 학습 없이도 파인튜닝 과정에서만 KD를 적용하는 MB-BERT를 제안하며, 이를 통해 모델 압축의 연산 비용을 줄이고 성능 저하를 최소화하고자 한다.

Aim

본 연구의 목표는 사전 학습 없이도 Transformer 모델을 효과적으로 경량화할 수 있는 방법을 제안하는 것이다. MB-BERT는 Block-wise Pruning)와 Low Rank Bridge Layer 삽입을 통해 모델 구조를 간소화하면서도, 점진적 지식 증류(Progressive Knowledge Distillation)를 적용하여 성능 저하를 최소화한다. 이를 통해 기존의 높은 연산 비용을 요구하는 KD 방식의 단점을 해결하고, Edge Device 환경에서도 효율적으로 대형 언어 모델을 배포할 수 있는 경량화 솔루션을 제공하는 것이 본 연구의 궁극적인 목표이다.

Methods

MB-BERT는 Low Rank Bridge Layer, Stepwise Block Pruning, Progressive Knowledge Distillation의 세 가지 핵심 기법을 기반으로 모델을 경량화한다. LoRB는 가지치기된 Transformer 블록에서 손실된 정보를 보완하는 역할을 하며, 기존 블록을 저차원 행렬로 대체함으로써 계산 비용을 줄이면서도 성능을 유지한다.



Stepwise Block Pruning은 Transformer의 짝수 인코더 블록을 단계적으로 가지치기하고 LoRB로 대체하는 방식으로, 기존의 일괄 가지치기 방식보다 성능 저하를 최소화하면서 안정적인 모델 압축을 가능하게 한다. 마지막으로, Progressive Knowledge Distillation은 각 가지치기 단계에서 직전 학생 모델을 새로운 교사 모델로 활용하여 지식 증류를 반복적으로 수행하는 방식으로, 모델이 점진적으로 최적화되도록 한다. 이러한 방법을 통해 MB-BERT는 기존의 사전 학습 기반 KD보다 낮은 연산 비용으로 모델을 경량화하면서도 성능을 유지할 수 있다.

Results

	$BERT_{base}$	(a)	(b)	(c)	(d)	(e)	**MiniBridgeBERT**
#Params	109.5M	102.8M	96.1M	89.4M	82.7M	76.0M	69.3M
#LoRB (Encoder)	0 (12)	1 (11)	2 (10)	3 (9)	4 (8)	5 (7)	6 (6)
CoLA	56.6	56.7	56.0	55.3	56.3	54.2	54.3

본 연구에서는 CoLA 데이터셋을 활용하여 MB-BERT의 성능을 평가하였다. 실험 결과, MB-BERT는 기존 BERT-base 대비 모델 크기를 37% 줄이면서, 높은 MCC 성능을 유지하는 데 성공했다.

	MiniBridgeBERT	w/o KD	w/o Progressive	DistillBERT
CoLA	**54.3**	38.3	52.1	54.2

단계별 가지치기와 LoRB를 함께 적용함으로써 가지치기된 블록에서도 정보 손실을 최소화할 수 있었으며, Progressive KD가 모델 성능 유지에 중요한 역할을 한다는 점을 실험적으로 확인하였다.

Conclusion

본 연구에서는 사전 학습 없이도 Transformer 모델을 효과적으로 경량화할 수 있는 MB-BERT를 제안하였다. MB-BERT는 파인튜닝 과정에서만 KD를 적용하는 방식으로 연산 비용을 줄이고, LoRB와 단계적 Block Pruning를 활용하여 성능 저하를 최소화하였다. 실험 결과, 기존 BERT-base 대비 모델 크기를 37% 줄이면서도 높은 MCC 성능을 유지할 수 있음을 확인하였다. 본 연구는 엣지 디바이스 등 자원 제한 환경에서도 LLM을 효율적으로 배포할 가능성을 제시하며, 향후 다양한 Transformer 모델에도 적용될 수 있을 것으로 기대된다.