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

CUAI 7 기 NLP 1 팀
정성룡(AI 학과), 오규안 (AI 학과)

[요약] 대형 언어 모델(LLM)의 성능은 모델 크기 증가로 향상되지만, 엣지 디바이스 배포를 위해 압축이 필수적이다. 본 연구는 사전 훈련 없이 미세 조정 중 지식 증류를 적용하는 MiniBridgeBERT (MB-BERT)를 제안한다. 블록 가지치기와 저차원 브릿지 레이어(LoRB)로 구조를 간소화하고, 점진적 증류로 성능 저하 없이 효율적 압축을 달성한다. CoLA 데이터셋 평가 결과, MB-BERT 는 높은 MCC 를 유지하며 경량화에 성공했다.

1. 서 론

대규모 언어 모델(LLM)의 성능은 모델 크기 증가에 따라 향상되지만, 엣지 디바이스에서의 배포를 위해 모델 경량화가 필수적이다. 이를 위해 양자화 (Quantization), 가지치기(Pruning), 지식 증류 (Knowledge Distillation) 등의 방법이 활용되며, 특히 지식 증류는 압축된 모델에서도 높은 성능을 유지하는 데 효과적이다. 그러나 기존 방법들은 사전 학습 단계에서 교사 모델을 활용해야 하며, 이는 높은 계산 비용을 초래한다.

본 연구에서는 사전 학습 없이 파인튜닝 과정에서만 지식 증류를 적용하는 MiniBridgeBERT (MB-BERT)를 제안한다. MB-BERT 는 블록 단위 가지치기(Block-wise Pruning)와 Low Rank Bridge Layer(LoRB) 삽입을 통해 모델 구조를 간소화하며, 성능 저하를 최소화하는 점진적 지식 증류(Progressive Knowledge Distillation) 전략을 활용한다. 특히, 짝수 인코더 블록을 LoRB 로 점진적으로 대체하면서, 각 단계에서 교사 모델로부터 지식을 전이하여 최적의 성능을 유지한다.

본 연구는 BERT-base 모델을 기반으로 CoLA 데이터셋에서 MB-BERT 의 성능을 평가하였으며, 높은 MCC 점수를 유지하면서도 모델 크기를 크게 줄이는 데 성공했다. 실험을 통해 LoRB 와 점진적 증류 전략이 모델 압축 과정에서 필수적인 역할을 함을 확인하였고, 제안된 방법이 자원이 제한된 환경에서도 효과적인 LLM 경량화 솔루션을 제공할 수 있음을 입증하였다.

2. 방 법 론

2.1 Low Rank Bridge Layer (LoRB)

LoRB 는 가지치기된 Transformer 블록에서 손실된 정보를 보완하는 핵심 모듈이다. Transformer 블록을 제거한 후, 해당 위치에 저차원 근사 행렬을 삽입하여 계산 복잡도를 줄이면서도 성능을 유지하도록 설계되었다. 블록 입력 x 와 출력 h 의 관계는 다음과 같이 표현된다.

$$h = x + \Delta W x = x + BAx$$

여기서 B와A 는 저차원 행렬로, 가지치기된 영역의 표현력을 복원하는 역할을 한다. LoRB 는 Skip Connection 형태로 설계되어 학습 안정성을 유지하며, 경량화된 모델에서도 성능을 보존할 수 있도록 한다.

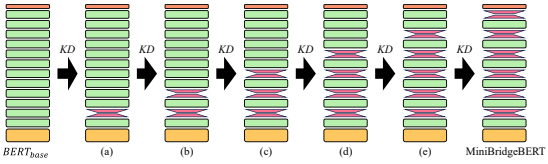

Figure 1. MiniBridgeBERT 에서 점진적으로 인코더 블록을 LoRB 로 대체하고, 대체된 모델을 다시 Teacher 모델로 설정하여 Knowledge Distillation 를 수행

2.2 Stepwise Block Pruning and Replacement

MB-BERT 는 Transformer 의 짝수 인코더 블록을 단계적으로 가지치기하고, 해당 블록을 LoRB 로 대체하는 방식을 따른다. (Figure 1) 먼저, 첫 번째 짝수 블록을 제거하고 LoRB 를 삽입한 후, 교사 모델로부터 지식 증류를 수행한다. 이후 동일한 과정을 반복하여 모든 짝수 블록을 LoRB 로 점진적으로 대체한다.

이러한 단계적 가지치기 전략은 모든 블록을 한 번에 제거하는 방법보다 성능 저하를 최소화하면서 안정적인 모델 압축을 가능하게 한다. 또한, Transformer 모델의 계층적 정보 분포를 고려하여, 가지치기가 성능에 미치는 영향을 줄이는 설계를 적용하였다.

Copyright © 2024 CUAI

2.3 Progressive Knowledge Distillation

MB-BERT 는 점진적 지식 증류 (Progressive Knowledge Distillation, KD) 기법을 활용하여 경량화된 모델에서도 성능을 유지한다. 기존 KD 와 달리, 각 가지치기 단계에서 교사 모델을 업데이트하는 방식으로 진행된다. 첫 번째 가지치기 단계에서는 BERT-base 가 교사 모델이 되며, 이후 단계적으로 가지치기가 진행됨에 따라 직전 단계의 학생 모델을 새로운 교사 모델로 활용한다. 이를 통해 지식 전이의 일관성을 유지하고, 가지치기 과정에서 발생하는 성능 저하를 완화할 수 있다.

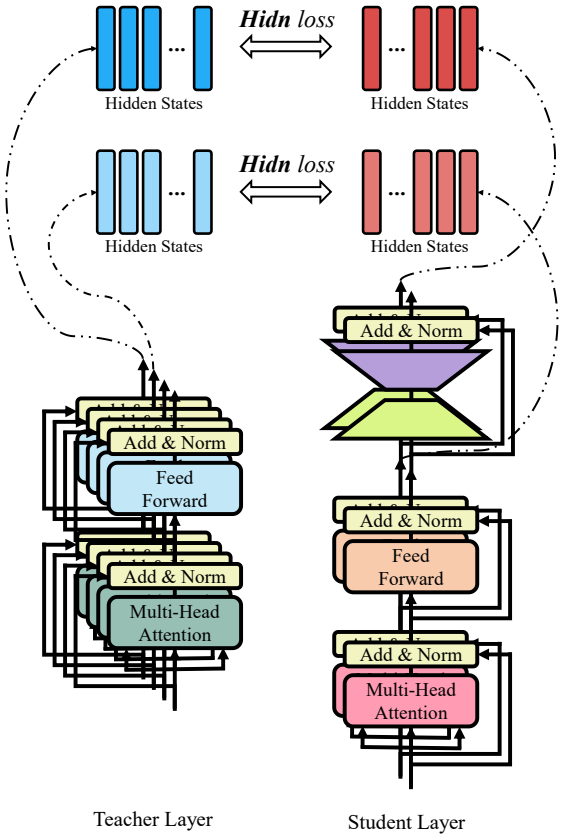

Figure 2: LoRB 과 Transformer-layer 에 대한 Hidden Loss 를 통한 distillation

지식 증류는 교사 모델과 학생 모델 간의 히든 스테이트(hidden states) 매칭을 통해 이루어지며, 평균제곱오차(MSE) 손실 함수를 활용하여 학생 모델이 교사 모델의 표현력을 모방하도록 학습한다. (Figure 2) 이를 통해 가지치기된 블록에서도 정보 손실을 최소화하며, 경량화된 모델에서도 안정적인 성능을 유지할 수 있다.

최종 손실 함수는 크로스 엔트로피 손실(L_{CE}) Hidden State 손실(L_{hidn})을 조합하여 정의되며, 이는 다음과 같다.

$$L_{total} = L_{CE} + \lambda L_{hidn}$$

여기서 λ 는 두 손실 항목 간의 균형을 조정하는 하이퍼파라미터이다. LoRB 와 점진적 KD 를 함께 적용한 MB-BERT 는 사전 학습 없이도 경량화된 모델에서 높은 성능을 유지할 수 있는 효과적인 접근법을 제공한다.

3. 실 험

3.1 데이터셋

본 연구에서는 GLUE 벤치마크의 CoLA(Corpus of Linguistic Acceptability) 데이터셋을 활용하여 MB-BERT 의 성능을 평가하였다. CoLA 는 문장의 문법적 수용 가능성을 판단하는 이진 분류 태스크를 다루며, 총 10,657 개 문장으로 구성된다. 주요 평가 지표로는 Matthews Correlation Coefficient (MCC)를 사용하며, 이는 불균형 데이터에서 모델의 예측 성능을 효과적으로 측정할 수 있는 척도이다.

3.2 MiniBridgeBERT 실험

MB-BERT 는 BERT-base 를 기반으로 설계된 경량화 모델로, 사전 학습 없이 파인튜닝 과정에서만 블록 단위 가지치기와 점진적 지식 증류(KD)를 적용하여 모델의 크기를 줄이는 동시에 성능 저하를 방지한다.

기존 BERT-base 모델(109M 파라미터, 12 개 Transformer 블록)을 교사 모델로 사용하였으며, 학생 모델은 짝수 블록을 LoRB 로 점진적으로 대체하면서 학습하였다. 가지치기된 블록의 학습을 촉진하기 위해, LoRB 에는 기존 인코더 블록보다 5 배 높은 학습률을 적용하여 빠르고 안정적인 수렴을 유도하였다.

	$BERT_{base}$	(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

Table 1: MiniBridgeBERT 의 CoLA 데이터셋 성능 비교

Table 1 에 나타난 실험 결과에 따르면, MiniBridgeBERT(MB-BERT)는 모델 크기를 줄이면서도 높은 MCC 성능을 유지하는 데 성공하였다. 즉, MB-BERT 는 모델 크기를 절반 가까이 줄이면서도

Copyright © 2024 CUAI

성능 저하를 최소화하였으며, 문법적 수용 가능성 판단(CoLA 태스크)에서도 효과적으로 동작함을 확인하였다.

3.3 Ablation Study

MB-BERT 의 성능 유지에 있어 Low Rank Bridge Layer(LoRB)와 점진적 지식 증류(KD)가 핵심적인 역할을 수행함을 확인하기 위해 Ablation Study 를 진행하였다.

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

Table 2: MiniBridgeBERT framework 의 다양한 절차에 대한 Ablation Study.

실험 결과, 점진적 과정을 적용하지 않고 모델을 학습한 경우 MCC 가 0.521 로 감소하였으며, 이는 점진적 KD 가 성능 유지에 중요한 역할을 한다는 것을 보여준다. 또한, KD 를 제외하고 학습한 경우 MCC 는 0.383 까지 감소하여, 교사 모델로부터의 지식 전이가 모델 성능에 미치는 영향을 명확히 확인할 수 있었다. 반면, 모든 구성 요소를 포함한 최종 모델은 MCC 0.543 을 기록하며 안정적인 성능을 달성하였다. Ablation study 는 Low Rank Bridge Layer 와 점진적 KD 가 모델 경량화 과정에서 필수적이며, 이들의 조합이 성능 유지와 경량화를 동시에 달성하는 데 핵심적인 요소임을 확인한다. (Table 2)

4. 결 론

본 연구에서는 사전 학습 없이도 효과적으로 경량화된 Transformer 모델을 설계할 수 있는 MiniBridgeBERT (MB-BERT)를 제안하였다. 기존 모델 압축 기법은 높은 연산 비용을 요구하는 사전 학습 기반 지식 증류를 필수적으로 활용하지만, 본 연구는 파인튜닝 과정에서만 지식 증류를 적용하는 새로운 접근법을 도입하여 경량화 과정의 효율성을 극대화하였다.

MB-BERT 는 블록 단위 가지치기(Block-wise Pruning)와 Low Rank Bridge Layer(LoRB) 삽입을 결합하여 모델 구조를 단순화하면서도 성능 저하를 최소화하였다. 또한, 점진적 지식 증류 (Progressive KD) 전략을 적용하여 가지치기된 영역에서도 안정적인 정보 전이를 가능하게 하였다.

CoLA 데이터셋을 활용한 실험 결과, MB-BERT 는 기존 BERT-base 대비 모델 크기를 37% 줄이면서도 높은 Matthews Correlation Coefficient (MCC) 성능을 유지하는 데 성공하였다. Ablation Study 를 통해 LoRB 와 점진적 KD 가 성능 유지에 필수적인 요소임을 확인하였으며, 기존 방법 대비 계산 비용을 크게 절감하면서도 모델 경량화 효과를 극대화할 수 있음을 입증하였다.

본 연구는 엣지 디바이스와 같은 자원 제한 환경에서도 대형 언어 모델(LLM)을 효율적으로 배포할 수 있는 경량화 방법을 제시하며, 향후 GPT, T5 와 같은 다양한 Transformer 기반 모델에도 일반화될 가능성이 있다. 또한, 추가적인 최적화 연구를 통해 다른 자연어 처리(NLP) 태스크에서도 MiniBridgeBERT(MB-BERT)의 성능을 확장하는 방향으로 연구를 발전시킬 수 있을 것이다.

참고 문헌

1) DistilBERT, a distilled version of BERT_ smaller, faster, cheaper and lighter, Victor Sanh et al., 2019
2) On the Effect of Dropping Layers of Pre-trained Transformer Models, Hassan Sajjad et al., 2020
3) Distilling the Knowledge in a Neural Network, Geoffrey Hinton et al., 2015
4) Fine-Tuned Transformers Show Clusters of Similar Representations Across Layers, Jason Phang et al., 2021
5) Neural Network Acceptability Judgments, Alex Warstadt et al., 2018
6) Born Again Neural Networks, Tommaso Furlanello et al., 2018

Copyright © 2024 CUAI