계층별 표현력 가중화 기반 BERT 지식 증류 기법

윤서희 [1], 문용혁 [2*]
[1] 성신여자대학교 컴퓨터공학과 학부생
[2] 성신여자대학교 컴퓨터공학과 교수

xyzysh@gmail.com, yhmoon@sungshin.ac.kr

BERT Distillation via Layer-wise Weighted Learning

Seo-Hee Yun[1], Yong-Hyuk Moon[1]
[1]Dept. of Computer Engineering, Sungshin Women's University

요 약

TinyBERT 의 중간층 증류는 교사 모델의 레이어를 학생 모델의 레이어와 1:1 로 대응시켜 수행된다. 본 연구에서는 이러한 중간층 증류 방식을 확장하여, 각 학생 레이어에 대해 교사 모델의 여러 레이어를 가중합 하여 증류하고, 학습 과정에서 해당 가중치를 동적으로 업데이트하는 전략을 제안한다. 이를 통해 모델의 성능을 향상시킬 수 있었으며, 태스크 특성에 적합한 교사 레이어 조합을 활용하는 적응적 지식 증류(Adaptive Knowledge Distillation)의 가능성을 검증하였다.

1. 서론

최근 대규모 사전학습 언어모델은 다양한 자연어 처리 과제에서 탁월한 성능을 보이고 있으나, 모델 규모가 크고 연산량이 방대하여 실제 응용에서 효율성이 떨어진다. 이를 해결하기 위해 지식 증류(Knowledge Distillation, KD)가 모델 효율화 기법으로 주로 활용되고 있으나, 기존 KD 기법은 주로 단순한 로짓(Logit) 증류나 균등 분할 기반의 레이어 매핑에 의존하여 정적인 학습을 수행하는 한계가 있다.

본 연구에서는 TinyBERT [1]의 중간층 증류 방식을 확장하여, 교사 레이어의 혼합 비율을 학습 과정에서 동적으로 조정하는 KD 학습 전략을 제안한다. 이를 통해 각 학생층이 교사 레이어를 태스크별로 다르게 참고하도록 하여 지식 전달의 효율성을 높인다.

2. 태스크 중심 지식 증류의 개선 가능성

DistilBERT [2]는 Teacher 모델의 최종 출력 로짓(Logit)만을 증류 대상으로 삼기 때문에 내부 표현에 담긴 세부적인 지식은 충분히 전달하지 못한다. 이에 비해 TinyBERT [1]는 중간 레이어의 표현까지 증류하여 보다 풍부하고 세밀한 지식을 Student 모델에 전달할 수 있다. 본 논문에서는 중간 레이어 증류를 확장하여 태스크 특화 표현력을 가중합 방식으로 반영하는 방법을 제안하고, 이를 3 종의 다운스트림 태스크에서 평가하여 소형 모델의 성능 개선을 확인한다.

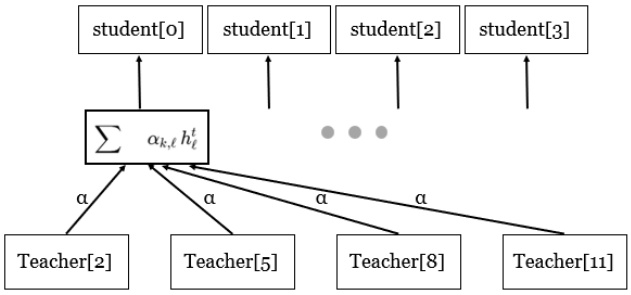

그림 1. 계층별 표현력 가중화 기반 제안 증류 방법론

3. 계층별 표현력 가중화 기반 지식 증류 기법

지식 증류를 위해 교사 모델(BERT Base, 12L)의 레이어 {2, 5, 8, 11}을 선택하였고, 각 레이어의 가중치에 따라 학생 모델(TinyBERT, 4L)의 각 레이어로 전달되는 증류 배합이 제안 수식 (1)과 같이 결정된다. 본 논문에서는 세 가지 방법론을 비교 실험한다.

- Fixed: 기준치가 되는 기존 TinyBERT 증류 방식
- Uniform: 모든 가중치 값을 균등 설정하는 방식
- Learned: 학습 과정에서 동적으로 업데이트되는 가중치를 기반으로 적응적 증류 배합 방식 (그림 1)

$$\tilde{h}_k^t = \sum_{l \in \{2,5,8,11\}} \alpha_{k,l} h_l^t \quad (1)$$

위 수식은 교사 레이어 $l = \{2, 5, 8, 11\}$에 대해 각 레이어 별 가중치 α를 정의하여, 교사의 l 번째 레이어의 은닉 상태(Hidden State) h_l^t의 배합 정도를 결정한

다. 학생의 k 번째 레이어는 왼쪽 항인 가중합 된 교사의 혼합 표현 $\tilde{h}_k^t$을 증류 받는다.

이때 가중치 α값을 직접 학습하는 것은 제한적인 문제가 있기에 아래 수식 (2)와 같이 SoftMax 함수를 통해 가중치 내부의 로짓 값 $s_{k,j}$을 학습에 사용한다.

$$\alpha_{k,l} = \frac{exp(s_{k,l})}{\sum_{j \in \{2,5,8,11\}} exp(s_{k,j})} \quad (2)$$

한편 학생 모델의 대응 레이어로 혼합 표현 $\tilde{h}_k^t$을 증류하기 위해 P 행렬로 선형 변환하여 차원을 일치시킨다. 이를 통해 학생 표현 이 교사 모델의 $\tilde{h}_k^t$ 혼합 표현을 MSE 기반으로 추종하도록 중간층 배합 증류의 표현력 손실함수 L_{hid}를 설계할 수 있다.

$$L_{hid} = \sum_{k=0}^{3} \left\| Ph_k^s - \widetilde{h_k^t} \right\|_2^2 \quad (3)$$

최종 손실 함수 [1]는 지도 학습 L_{sup}, 교사의 로짓 L_{logit}, 은닉 상태 L_{hid}, 레이어 별 매핑 가중치 α의 엔트로피 정규화 L_{ent}를 결합하여 구성된다. 중간 레이어 손실에는 별도의 가중치 λ 를 반영하여 배합 증류의 효과를 높이도록 설계하였다.

$$L_{total} = L_{sup} + \lambda_{logit} L_{logit} + \lambda_{hid} L_{hid} + L_{ent} \quad (4)$$

특히 초기 학습에서 L_{hid}의 큰 차이는 더 강한 페널티(Penalty)로 작용하여 학습 효과를 높이는데 기여한다. 또한 L_{att}는 성능 향상에 미치는 영향이 미비함을 사전 실험을 통해 확인하여 최종 실험 단계에서 제외하였다.

4. 가중합 증류 실험

4.1. 실험 설정

본 실험은 GLUE 의 세 가지 태스크를 대상으로 수행하였다. 데이터셋 규모가 결과에 영향을 줄 수 있다고 판단하여, 상대적으로 크기가 작아 과적합이 발생하기 쉬운 MRPC, RTE 2 종과 규모가 큰 QNLI 1 종을를 선정하였다.

Parameters	*Values*
Batch size for train and evaluation	16, 32
Learning rate for student and α	3e-5, 1e-3
$\lambda_{logit}, \lambda_{hid}, \lambda_{att}$	1.0, 1.0, 0.0

표 1. 하이퍼파라미터 설정

하이퍼파라미터는 다수의 요소 소거 분석(Ablation Study)와 값 조정을 통해 최적화하였다. 특히 Warmup 값을 지나치게 크거나 작게 설정할 경우 성능이 급격히 저하되는 현상을 관찰하여, 이를 300 으로 고정하였다. 또한 학생 모델의 학습률과 가중치 α에 적용되는 학습률을 별도로 설정하여 실험을 진행하였다.

Task-Metrics	*Fixed*	*Uniform*	*Learned*
MRPC-accuracy	0.7843	0.7770	**0.8039**
MRPC-F1	0.8576	0.8520	**0.8662**
RTE-accuracy	0.5668	0.5957	**0.5993**
QNLI - accuracy	0.8259	0.8266	**0.8272**

표 2. 가중합 증류 실험 결과

4.2 실험 결과

TinyBERT 증류 과정에서 레이어 매핑을 세 가지 방식(Fixed, Uniform, Learned)으로 설정하여 성능을 비교하였다. 실험 결과는 10 epochs 중 최적 성능을 기준으로 표에 정리하였으며, 약 3–4 epochs 이후부터는 값이 크게 변하지 않음을 확인하였다. 전반적으로 모든 태스크에서 Learned 방식이 근소하게 우세하였으나, 대규모 데이터셋에서는 세 방식 간 성능 차이가 거의 나타나지 않았다. 반면, 소규모 데이터셋에서는 Learned 방식이 상대적으로 더 큰 이점을 보였다. 이는 제한된 데이터 환경에서 학습 가능한 가중치 분포를 활용해 교사와 학생 레이어 간 동적 매핑을 수행하는 것이, 단순한 고정 매핑보다 일반화 성능 향상에 유리함을 시사한다.

5. 결과 및 재언

본 연구에서는 계층별 표현력 가중화 기반의 지식 증류 기법을 제안하였다. 실험 결과, Fixed 와 Uniform 방식은 안정적인 성능을 보였으나, 데이터셋 규모와 태스크 특성에 따라 Learned 방식이 추가적인 성능 향상을 제공함을 확인하였다. 이는 교사–학생 레이어 간 단순한 일대일 매핑을 넘어, 가중치 기반의 적응적 증류 전략이 실제 과제에서 더 우수한 일반화 능력을 발휘할 수 있음을 시사한다. 특히 본 제안 방식은 교사 모델의 중간층 표현을 배합하는 가중치를 학습 대상으로 정의함으로써 단일 단계(Single-Stage) 지식 증류 효과를 달성한 차별성이 있다.

Acknowledgement. 이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (RS-2025-24292968).

참고문헌

[1] Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., Liu, Q., TinyBERT: Distilling BERT for Natural Language Understanding, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 2020, pp. 4163–4174.

[2] Sanh, V., Debut, L., Chaumond, J., Wolf, T., DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter, Proceedings of the 5th Workshop on Energy-Efficient Machine Learning and Cognitive Computing, Vancouver, Canada, 2019, pp. 1–6.