# Chapter 5. BERT 파생모델II: 지식 증류 기반

발표자: 박채원

22-07-29

# 목차

- 지식 증류 소개

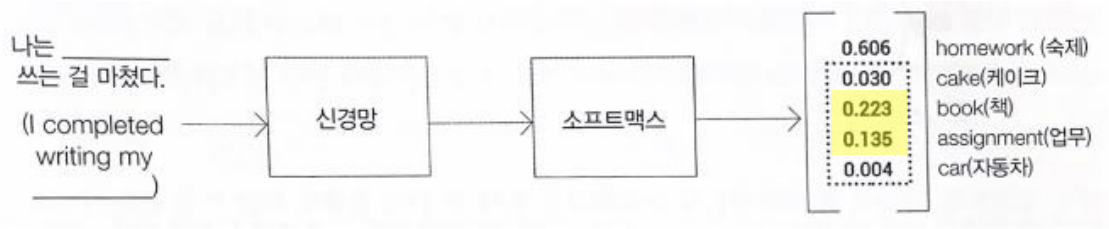
- DistilBERT
  - BERT의 증류 버전
  - 대형 BERT에서 소형 BERT로 지식 전달 방법

- TinyBERT
  - 정의
  - 지식 증류를 사용해 사전학습된 대형 BERT에서 지식을 얻는 방법
  - 데이터 증식 방법

- 대규모 BERT에서 **신경망**으로의 지식 전달 방법

# 5.1 지식 증류 소개

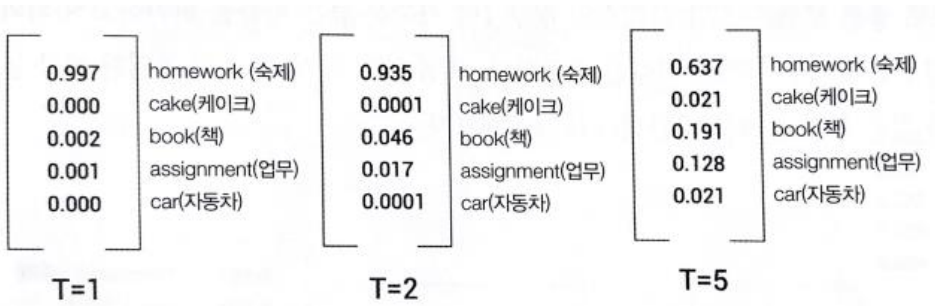
- 지식 증류: 사전 학습된 대형 모델의 동작을 재현하기 위해 소형 모델을 학습시키는 모델 압축 기술
  - 다음 단어 예측하는 사전 학습된 네트워크 (교사 네트워크)




- 암흑 지식이 적은 경우
  - 출력 레이어에 소프트맥스 템퍼러처 사용해, 확률 분포를 평활화

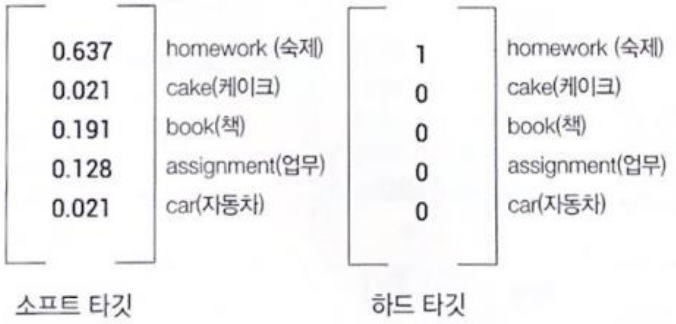
$$P_i = \frac{exp\left(\frac{z_i}{T}\right)}{\sum_j exp\left(\frac{z_j}{T}\right)}$$

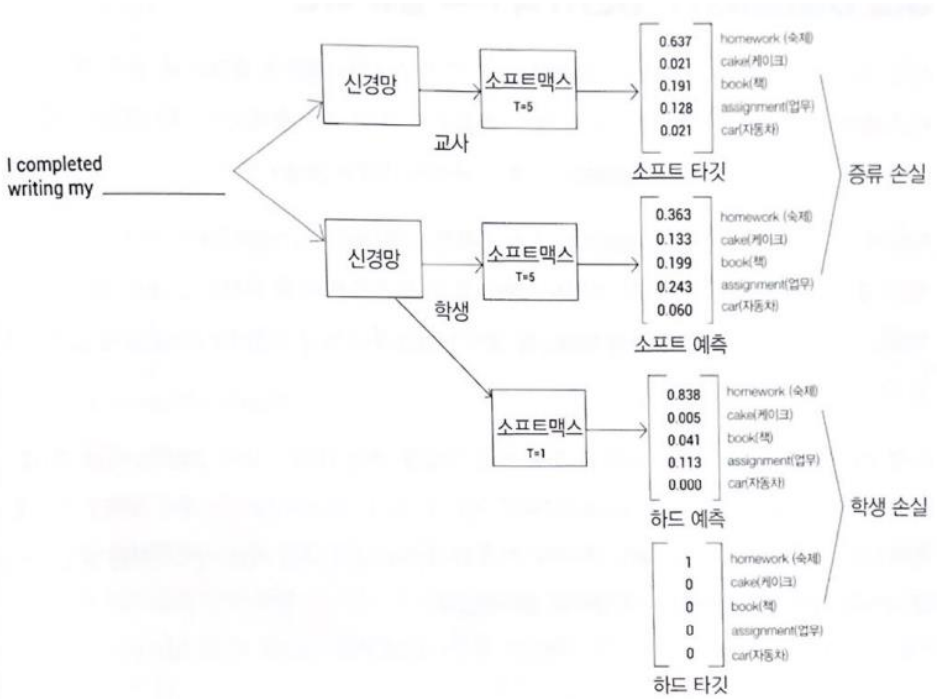
소프트맥스 템퍼러처 (T=템퍼러처)



# 5.1 지식 증류 소개

- 5.1.1 학생 네트워크 학습
  - 교사 네트워크가 반환하는 확률 분포(=소프트 타깃)가 학생 네트워크 반환 확률분포(=소프트 예측)의 목표가 됨
  - 소프트 타깃과 소프트 예측 간 교차 엔트로피 손실 계산 -> 증류 손실
  - 학생 네트워크에서 두개의 손실 사용
    - 증류 손실 (소프트 간 손실)
    - 학생 손실 (하드 간 손실)
  - 최종 손실 = $\alpha$ * 학생 손실 + $\beta$ * 증류 손실



# 5.2 DistilBERT: BERT의 지식 증류 버전

- DistilBERT
  - 허깅페이스 연구원들이 제안한 작고 빠르며 가벼운 버전의 BERT
  - 소형(학생) BERT가 DistilBERT이고, 대형 모델에 비해 60% 더 빠르며 크기는 40% 더 작음
  - 교사 레이어의 출력 레이어에서 학생 BERT로 지식을 전달하는 방법

- 5.2.1 교사-학습 아키텍처
  - BERT-base 모델을 교사 모델로 사용
  - 마스크 된 문장이 입력되면 사전학습된 BERT가 어휘 사전의 모든 단어에 대해 마스크된 단어 확률 분포를 제공. 여기엔 암흑 지식이 포함되어 있으며 이를 학생 BERT에 전달 해야한다.

- 학생 BERT
  - 레이어가 교사 BERT에 비해 적고, 매개변수 또한 5500만개 적다.
  - 은닉 상태 차원은 동일하게 768

# 5.2 DistilBERT: BERT의 지식 증류 버전

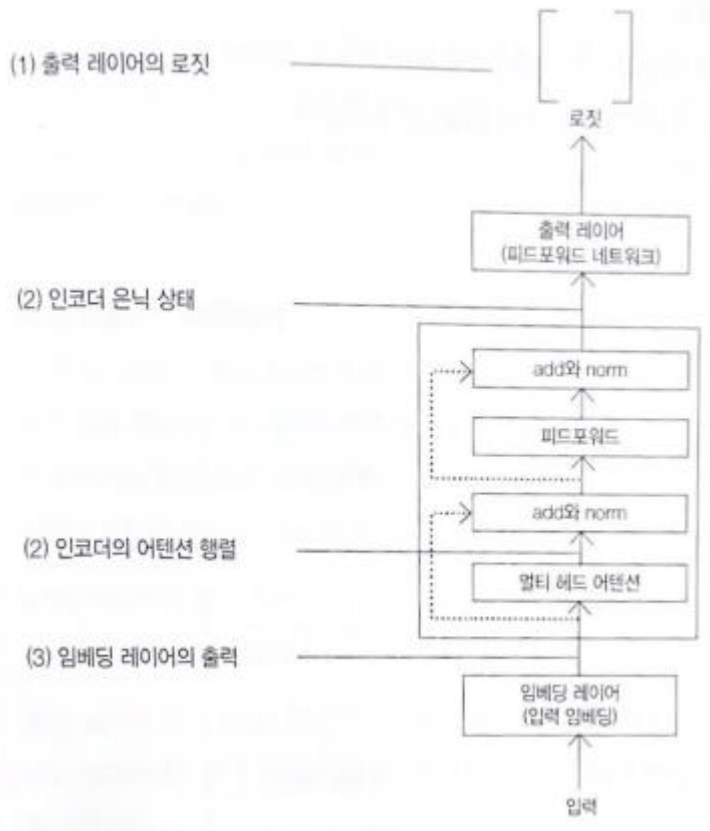
- 5.2.2 학생 BERT(DistilBERT 학습)
  - 교사 BERT 사전학습에 사용한 것과 동일한 데이터셋으로 학습 진행
  - MLM 태스크만 사용해 학습 진행
  - 즉, 여기선 학생 손실 == MLM 손실
  - 코사인 임베딩 손실 계산 -> 교사와 학생 BERT가 출력하는 벡터 사이의 거리 측정

  - BERT-base 모델의 97% 정도의 성능 제공
  - 추론 시 60% 더 빠름
  - 더 가볍기 때문에 다양한 엣지 디바이스에 쉽게 배포 가능

# 5.3 TinyBERT 소개

- 교사의 출력 레이어에서 학생에게 지식을 전달하는 것 외에 임베딩 및 여러 인코더 레이어에서 지식을 전달
  - 출력 레이어의 로짓
  - 은닉 상태
  - 어텐션 행렬: 어텐션 행렬은 언어 정보를 캡슐화
  - 임베딩 레이어의 출력

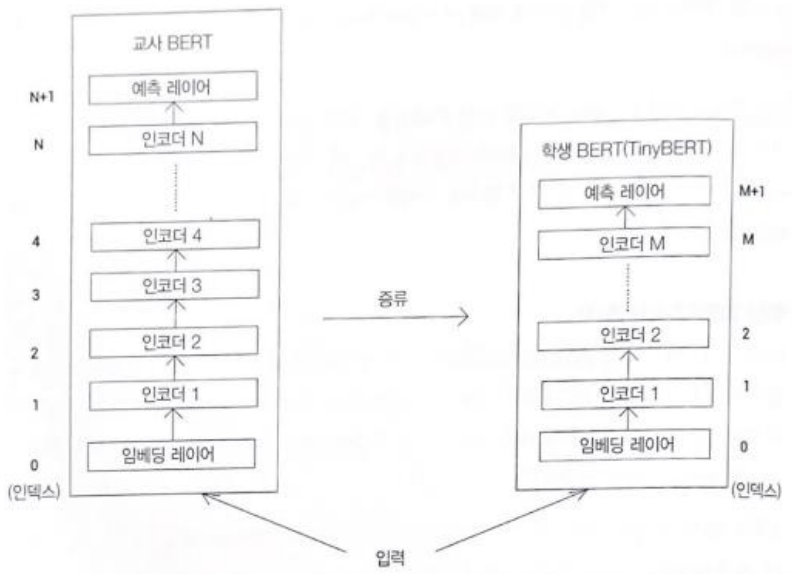
- 사전 학습 및 파인 튜닝 단계에서 증류를 적용하는 2단계 학습 프레임워크 사용



# 5.3 TinyBERT 소개

- 5.3.1 교사-학생 아키텍처
  - 학생 모델
    - 교사 BERT보다 작은 수의 인코더 layer를 가짐
    - 은닉 상태 차원은 312
    - 1450만 개의 매개변수 가짐

- 5.3.2 TinyBERT 지식 증류
  - n=g(m)
  - 위의 수식 처럼 교사 BERT의 n번째 레이어에서 학생의 m번째 레이어로 지식 전달



# 5.3 TinyBERT 소개

- 트랜스포머 레이어의 증류
  - 어텐션 기반 증류
    - 어텐션 행렬 지식을 교사 BERT에서 학생 BERT로 이전
    - 언어를 이해하는 데 매우 유용
    - 학생의 어텐션 행렬과 교사 BERT의 어텐션 행렬 간의 평균 제곱 오차를 최소화해 학생 네트워크 학습
    - 비정규화된 어텐션 행렬 사용이 이 설정에서 더 나은 성능을 발휘하고 더 빠른 수렴을 달성함

$$\mathcal{L}_{\text{attn}} = \frac{1}{h}\sum_{i=1}^{h} \text{MSE}(\boldsymbol{A}_i^S, \boldsymbol{A}_i^T)$$

  - 은닉 상태 기반 증류
    - 은닉 상태는 인코더의 출력 (=표현 벡터)
    - 교사 BERT와 은닉 상태 차원이 다름 -> 학생의 은닉 상태에 행렬을 곱해 선형 변환을 수행

$$\mathcal{L}_{\text{hidn}} = \text{MSE}(\boldsymbol{H}^S\boldsymbol{W}_h, \boldsymbol{H}^T)$$

# 5.3 TinyBERT 소개

- 임베딩 레이어의 증류

$$\mathcal{L}_{\text{embd}} = \text{MSE}(\boldsymbol{E}^S \boldsymbol{W}_e, \boldsymbol{E}^T),$$

- 예측 레이어 증류
  - 소프트 타깃과 소프트 예측 간의 교차 엔트로피 손실을 최소화해 예측 레이어 증류 수행
  - soft cross-entropy loss 사용
  - $L_{pred} = -softmax(Z^T) * \log\_softmax(Z^S)$ $\mathcal{L}_{\text{pred}} = \text{CE}(\boldsymbol{z}^T/t, \boldsymbol{z}^S/t),$

- 최종 손실 함수
  - m=0일 땐, 임베딩 레이어
  - m이 0보다 크고, M보다 작거나 같을 때는 트랜스포머 레이어
  - m이 M+1이면 예측 레이어
  - $L = \sum_{m=0}^{M+1} \lambda_m L_{layer}(S_m, T_{g(m)})$
  - $\lambda_m$는 계층의 중요도를 제어하는 하이퍼 파라미터 역할

$$\mathcal{L}_{\text{layer}} = \begin{cases} \mathcal{L}_{\text{embd}}, & m = 0 \\ \mathcal{L}_{\text{hidn}} + \mathcal{L}_{\text{attn}}, & M \geq m > 0 \\ \mathcal{L}_{\text{pred}}, & m = M+1 \end{cases}$$

# 5.3 TinyBERT 소개

- 5.3.4 TinyBERT 학습
  - 사전 학습 및 파인 튜닝 단계 모두에서 증류 가능

  - 일반 증류(=사전학습)
    - 사전 학습 단계에서 교사 BERT가 사용하는 데이터셋을 동일하게 사용해서 지식 전달
    - 이렇게 사전 학습 된 학생 BERT를 일반 TinyBERT라고 부름
  - 태스크 특화 증류 (=파인튜닝)
    - DistilBERT와 달리 사전 학습 단계에서 증류를 적용하는 것 외에 파인 튜닝 단계에서도 증류 적용 가능
    - 파인튜닝된 BERT-base를 교사로 사용해 지식 증류할 경우 이 학생 모델을 파인튜닝 된 TinyBERT라고 부른다.

| | 일반 증류(사전 학습) | 태스크 특화 증류(파인 튜닝) |
|---|---|---|
| 교사 | 사전 학습된 BERT-base | 파인 튜닝된 BERT-base |
| 학습 | 작은 BERT | 일반 TinyBERT<br>(사전학습된 TinyBERT) |
| 결과 | 일반 TinyBERT<br>(사전학습된 학생 BERT) | 파인튜닝 된 TinyBERT |

# 5.3 TinyBERT 소개

- 데이터 증식 방법
  - 문장에서 각 단어에 대해 다음 과정 진행

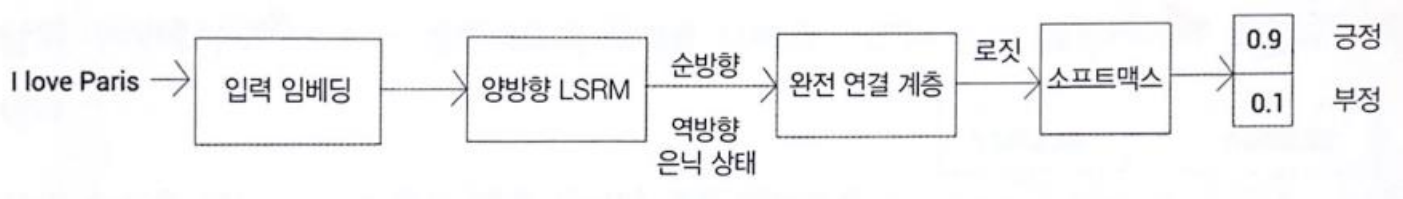
  - 단일 단어의 경우: 해당 단어 마스킹 후 BERT-base 모델을 사용해 마스킹 된 단어 예측해서 k개 후보 리스트
  - 단일 단어가 아닌 경우: 마스킹 하지 않고 글로브 임베딩 사용해 유사한 단어 K개 후보 리스트
  - P가 Pt(threshold)보다 작거나 같으면 후보 목록의 임의의 단어로 교체
  - P가 Pt보다크면 실제 단어를 그대로 둔다

  - 모든 문장에 대해 N번 반복 -> N개의 새로운 문장을 얻게 됨
  - 이렇게 증식된 데이터셋을 사용해 일반 TinyBERT를 파인튜닝

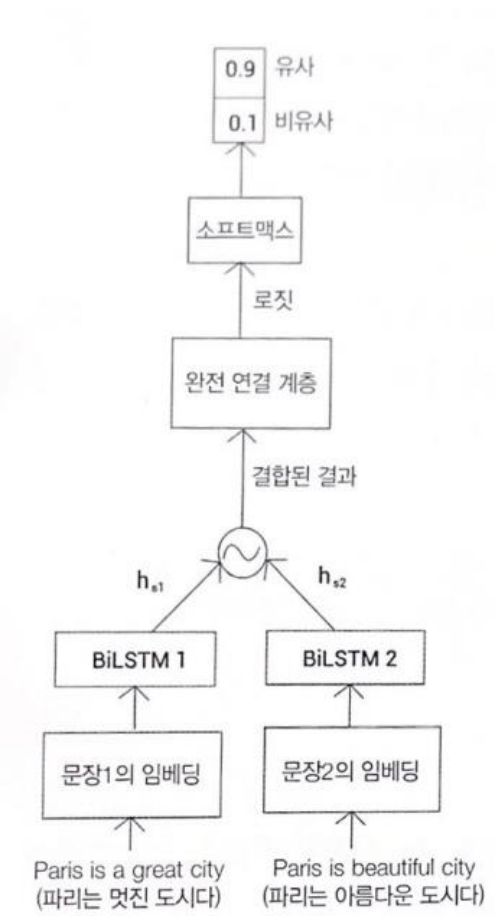
- TinyBERT는 BERT-base 모델보다 추론 효율 면에서 96% 좋고, 7.5배 더 작으며, 9.4배 더 빠름

# 5.4 BERT에서 신경망으로 지식 전달

- 5.4.1 교사-학생 아키텍처
  - 교사 BERT : 사전학습된 BERT-large를 파인튜닝 후 사용
- 학생 네트워크 : 단순 양방향 LSTM (BiLSTM) – 순방향 및 역방향 은닉 상태를 얻을 수 있음
  - Ex) 감정 분석




  - Ex) 문장 매칭 태스크
    - 학생 네트워크 = 샴 BiLSTM
    - 각 문장의 전방 및 후방 은닉 상태를 연결 비교 연산으로 결합하고
    - ReLU를 사용해 활성화 결과를 완전 연결 계층에 입력해 로짓을 얻는다.
    - 소프트맥스 함수에 입력



# 5.4 BERT에서 신경망으로 지식 전달

- 5.4.2 학생 네트워크 학습
  - 교사: 사전학습 되고 파인 튜닝된 BERT, 학생: BiLSTM
  - 학생 손실과 증류 손실의 가중 합계 손실을 최소화하는 방향으로 학생 네트워크를 학습 시킴
  - $L = \alpha * CE(Z^T, Z^S) + (1 - \alpha) * MSE(Z^T, Z^S)$
  - 교사 BERT에서 학생 네트워크로 지식 추출을 위해선 대규모 데이터셋 필요 -> 태스크 독립적인 데이터 증식

- 5.4.3 데이터 증식 방법
  - 마스킹
    - Pmask 확률로 문장의 단어를 [MASK]토큰으로 무작위 마스킹 해 새로운 문장을 만든다.
    - 모델이 클래스 레이블에 대해 각 단어의 기여도가 어느정도인지 이해하는 방법
  - 형태소 기반 단어 대체 방법
    - Ppos 확률로 문장의 한 단어를 같은 품사의 다른 단어로 대체
  - 엔그램 샘플링 방법
    - Png 확률로 문장에서 앤그램을 무작위로 샘플링하는 방법

# 5.4 BERT에서 신경망으로 지식 전달

- 데이터 증식 프로세스
  - 문장의 각 단어 Wi에 대해 Xi(0~1 사이의 무작위 값)라는 변수 생성
  - Xi<Pmask면 Wi를 마스킹
  - Pmask <= Xi*Pmask+Ppos면 형태소 기반 단어 대체 진행
  - 위의 두 로직은 겹치지 않는다.
  - 이후 수정된 문장에 Png의 확률로 엔그램 샘플링 적용해 최종 합성 문장을 얻는다.

  - 모든 문장에 대해 앞의 단계를 N번 수행하고 N개의 새로운 합성 문장을 얻는다.

# 5.5 마치며

- 지식 증류가 무엇인지, 어떻게 작동하는지 이해
  - 사전 학습된 대형 모델의 성능을 재현하기 위해 소형 모델을 학습시키는 모델 압축 기법
  - 교사(대형)-학생(소형) 학습
- DistilBERT
- TinyBERT의 동작 방식
- BERT에서 간단한 신경망으로 태스크 특화 지식 전달 방법 확인