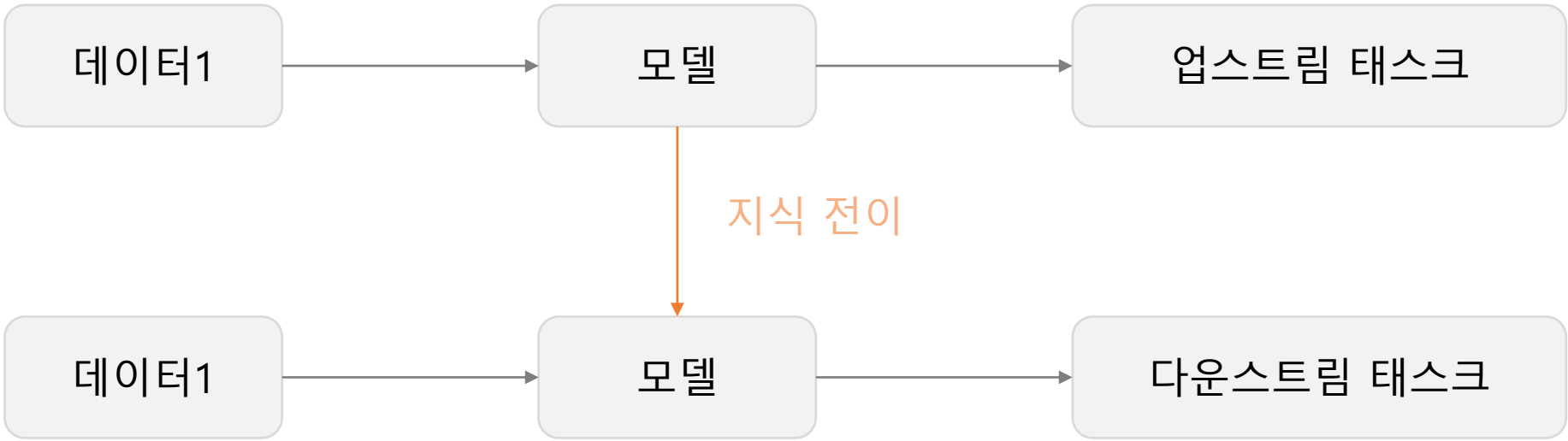
트랜스포머, GPT, BERT, ViT 알아보기

트랜스퍼 러닝transfer learning

- 특정 태스크를 학습한 모델을 다른 태스크 수행에 재사용



업스트림upstream 태스크

- 지도학습과 대비해서 뉴스, 웹문서, 백과사전과 같이 쉽게 구할 수 있는 글로 수작업 없이 다량의 학습 데이터를 만들 수 있음

- 다음 단어 맞히기(GPT)
  - 티끌 모아 [ ]

- 빈칸 채우기(BERT)
  - 티끌 [ ] 태산

다운스트림downstream 태스크

- 프리트레인pretrain을 마친 모델을 구조 변경 없이 그대로 사용하거나 태스크 모듈을 덧붙여서 수행

- 다운스트림 태스크 종류
  - 문서 분류 : 긍정/부정
  - 자연어 추론 : 참/거짓/중립
  - 개체명 인식 : 기관명, 인명, 지명 등
  - 질의응답
  - 문장 생성

다운스트림 태스크 학습 방식

- 파인튜닝fine-tuning
  - 다운스트림 태스크 데이터 전체를 사용하여 모델 전체를 업데이트

- 프롬프트 튜닝prompt tuning
  - 다운스트림 데이터 전체를 사용하고 모델 일부만 업데이트

- 인컨텍스트 러닝in-context learning
  - 다운스트림 데이터 일부만 사용하고 모델을 업데이트하지 않음

인컨텍스트 러닝 방식

- 제로샷 러닝zero-shot learning
  - 다운스트림 태스크 데이터를 전혀 사용하지 않고 모델이 바로 다운스트림 태스크를 수행
- 원샷 러닝one-shot learning
  - 다운스트림 태스크 데이터를 1건만 사용하여 다운스트림 태스크 수행
- 퓨샷 러닝few-shot learning
  - 다운스트림 태스크 데이터를 몇건만 사용하여 다운스트림 태스크 수행

토큰화tokenization

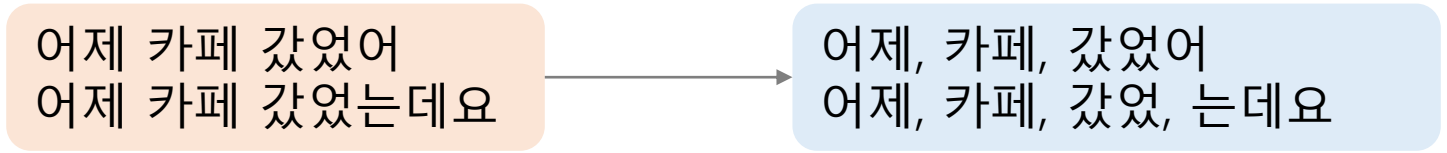
- 문장을 토큰 시퀀스로 나누는 과정

- 토큰화의 목적
  - 데이터의 기본단위는 텍스트 형태의 문장
  - 트랜스포머 모델은 토큰 시퀀스를 입력 받음
  - 문장에 토근화를 수행 해야함

단어단위 토큰화

- 공백으로 분리
  - 어휘 집합의 크기가 커지는 문제가 있음

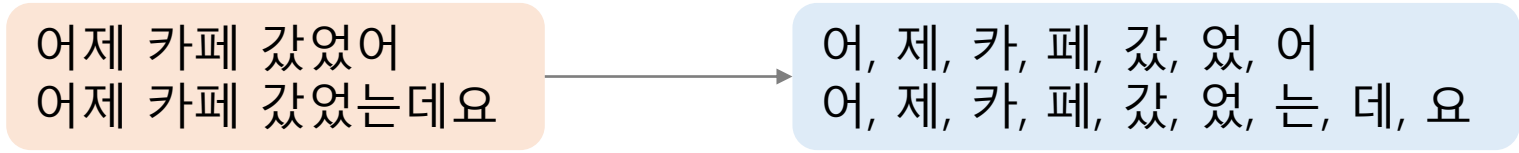
- 학습된 토크나이저를 사용
  - 의미있는 단위로 토큰화 해서 어휘 집합이 커지는 것을 다소 막을 수 있음



문자단위 토큰화

- 한글, 알파벳, 숫자, 기호를 고려해도 어휘 집합은 1만 5000개를 넘지 않고 미등록 토큰 문제로부터 자유로움

- 각 문자 토큰이 의미있는 단위가 되기 어렵고 토큰 시퀀스의 길이가 상대적으로 길어져서 학습이 어렵고 성능이 떨어짐



서브워드 단위 토큰화

- 단어와 문자 단위 토큰화의 중간에 있는 형태로 장점만 취함

- 대표적인 기법으로 바이트 페어 인코딩
  - BPE (GPT)
  - 워드피스wordpiece (BERT)

BPE

- 1994년 제안된 정보압축 알고리즘으로 데이터에서 가장 많이 등장한 문자열을 병합해서 데이터를 압축하는 기법
  - aaabdaaabac → ZabdZabac → ZYdZYac → XdXac
  - 사전의 크기는 4개에서 7개로 늘고 데이터 길이는 11에서 5로 축소
  - 분석 대상 언어에 대한 지식이 필요 없음

워드피스

- BPE와 본질적으로 유사하나 단순히 빈도로 병합하는 것이 아니라 우도likehood를 가장 높이는 쌍을 병합

  https://colab.research.google.com/drive/1GRSKcPKjbwr1NFM_6pBY3S2j2Me47hcw?usp=sharing

언어 모델language model

- 단어 시퀀스에 확률을 부여하는 모델

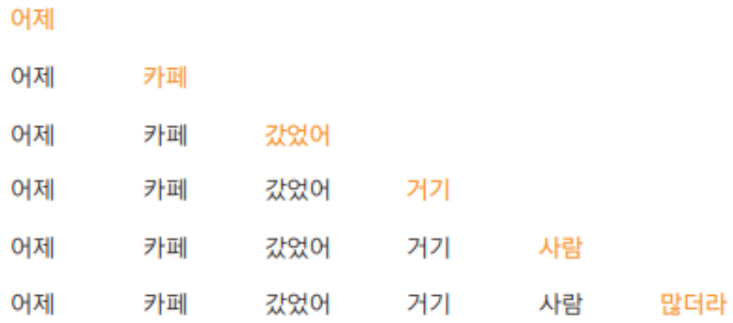
- 언어 모델의 출력은 문장에서 i번째로 등장할 단어를 w_i로 표시 한다면 n개 단어로 구성된 문장이 해당 언어에서 등장할 확률
  - $P(w_1, w_2, w_3, \ldots, w_n)$

- P(무모, 운전)보다는 P(난폭, 운전)이 큰 확률값

단방향 언어 모델

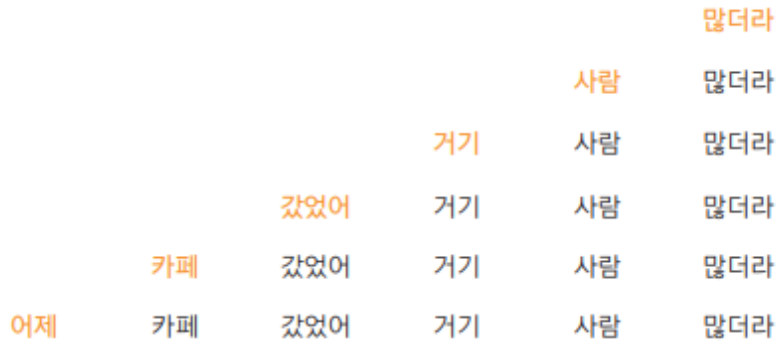
순방향 언어 모델forward language model

- 문장 앞부터 뒤로, 사람이 이해하는 순서대로 계산하는 모델
- GPT, ELMo 같은 모델이 이런 방식으로 프리트레인을 수행



역방향 언어 모델backward language model

- 문장 뒤부터 앞으로 계산
- ELMo(순방향과 역방향 모두 활용) 같은 모델이 이 방식을 활용



마스크 언어모델masked language model

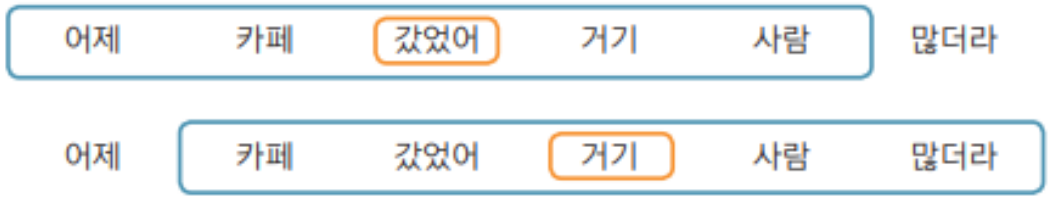
- 학습 대상 문장에 빈칸을 만들어 놓고 해당 빈칸에 올 단어로 적절한 단어가 무엇일지 분류하는 과정으로 학습하며 BERT가 주로 활용
- 단어를 계산할 때 문장 전체 맥락을 참고할 수 있다는 장점

어제	카페	갔었어	거기	사람	많더라
어제	카페	갔었어	거기	사람	많더라
어제	카페	갔었어	거기	사람	많더라
어제	카페	갔었어	거기	사람	많더라
어제	카페	갔었어	거기	사람	많더라
어제	카페	갔었어	거기	사람	많더라

스킵-그램 모델skip-gram model

- 어떤 단어 앞 뒤에 특정 범위를 정해 두고 이 범위 내에 어떤 단어들이 올지 분류하는 방식으로 학습

- '갔었어' 주변에 어제, 카페, 거기, 사람이 나타날 확률을 각각 높이는 방식으로 학습



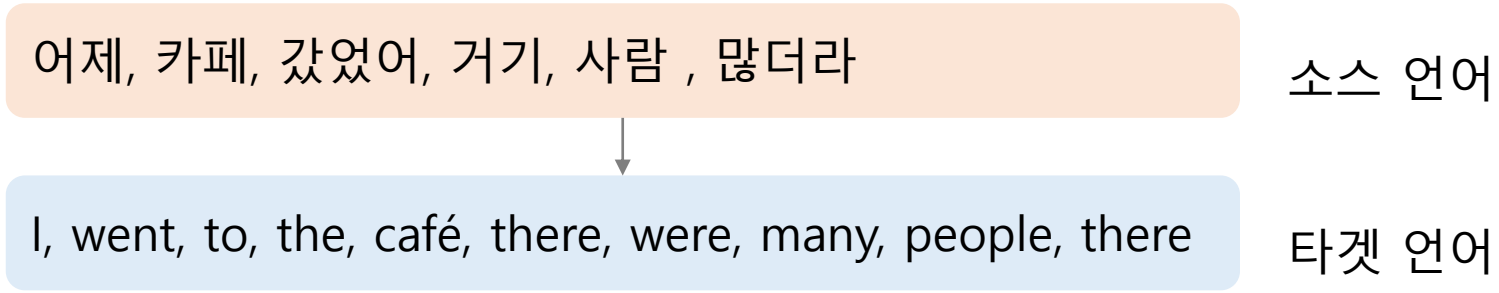
- Word2Vec이 스킵-그램 모델방식으로 학습

트랜스포머

2017년 구글이 제안한 시퀀스-투-시퀀스 모델

시퀀스-투-시퀀스sequence-to-sequence

- 특정 속성을 지닌 시퀀스를 다른 속성의 시퀀스로 변환
- 기계 번역의 경우 소스 언어의 토큰 시퀀스를 타겟 언어의 토큰 시퀀스로 변환하는 과제
- 소스 시퀀스의 길이(6개)와 타겟 시퀀스의 길이(10개)가 다르더라도 수행 가능

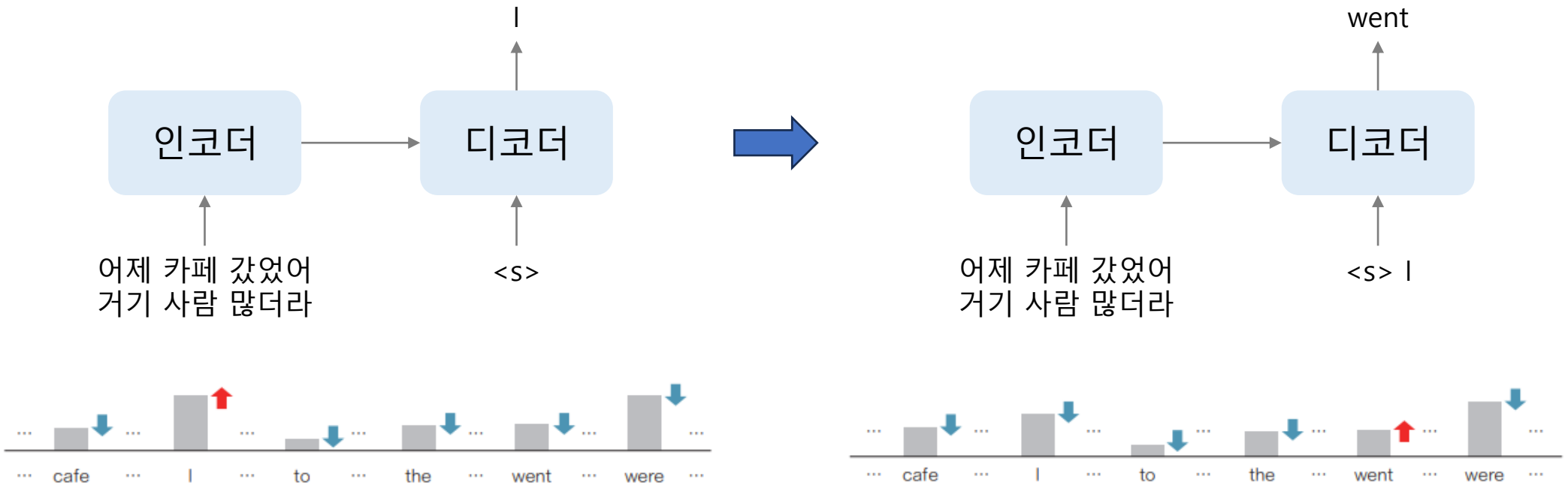


인코더와 디코더

- 시퀀스-투-시퀀스 과제를 수행하는 모델은 대개 인코더와 디코더 2개 파트로 구성됨
- 인코더는 소스 시퀀스의 정보를 압축해 디코더로 보내고 디코더는 인코더가 보내준 소스 시퀀스 정보를 받아서 타겟 시퀀스를 생성
- 기계 번역에서는 인코더가 한국어 문장을 압축해 디코더에 보내고 디코더는 이를 영어로 번역

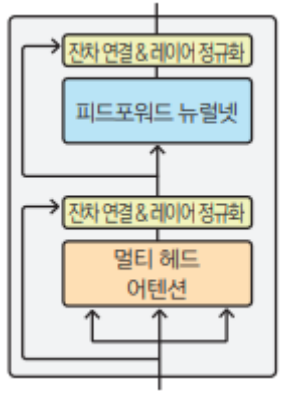
모델 학습과 인퍼런스

- 인코더에 소스 시퀀스 전체를 입력하고 디코더의 입력은 타겟 시퀀스를 순차적으로 전달하며 다음 토큰 값의 확률이 높아지도록 모델을 갱신함

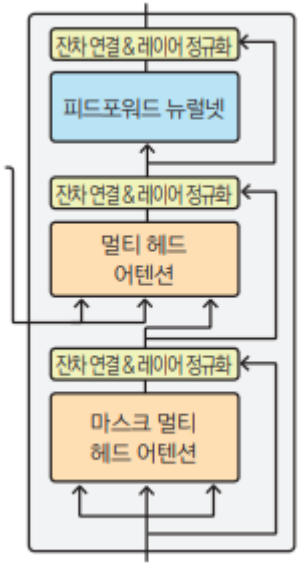


트랜스모머 블록

- 인코더 블록
  - 멀티 헤드 어텐션
  - 피드포워드 뉴럴 네트워크
  - 잔차 연결 및 레이어 정규화



- 디코더 블록
  - 마스크 멀티 헤드 어텐션 추가
  - 멀티 헤드 어텐션에서 인코더 입력을 함께 사용

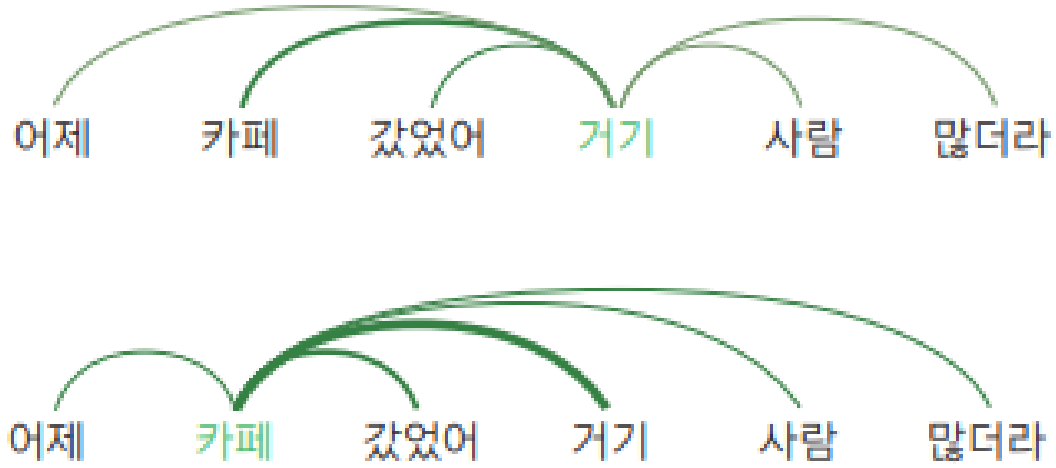


셀프 어텐션

- 시퀀스 입력에 수행하는 기계학습 방법의 일종으로 요소 가운데 중요한 요소에 집중하고 그렇지 않은 요소는 무시해 성능을 끌어올리는 기법
- CNN은 필터 크기를 넘어서는 문맥은 읽어내기 어려움
- RNN은 오래 전에 입력된 단어를 잊어버리거나 특정 단어 정보를 과도하게 반영하여 전체 정보를 왜곡하는 문제
- 어텐션은 소스 시퀀스 전체 단어와 타깃 시퀀스 하나 사이를 연결하지만 셀프 어텐션은 전체 입력시퀀스 모두 어텐션 계산을 수행

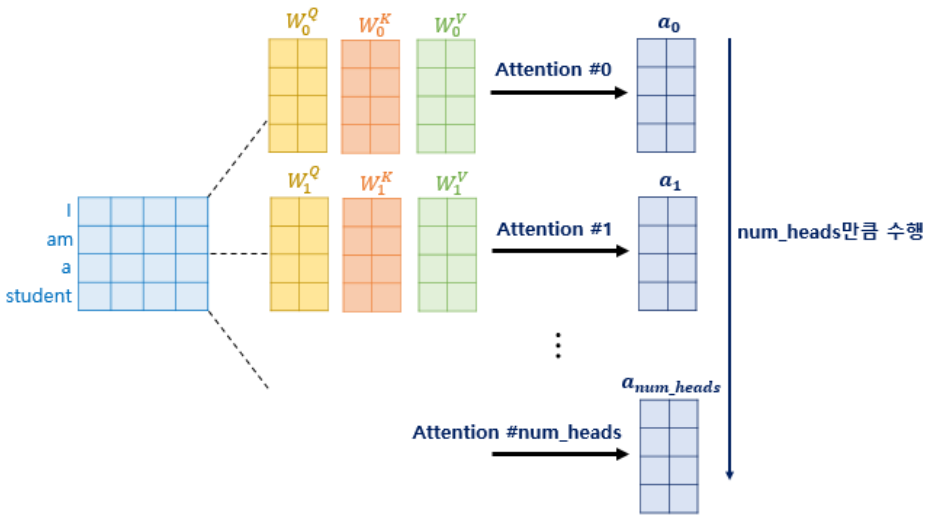
특징 및 장점

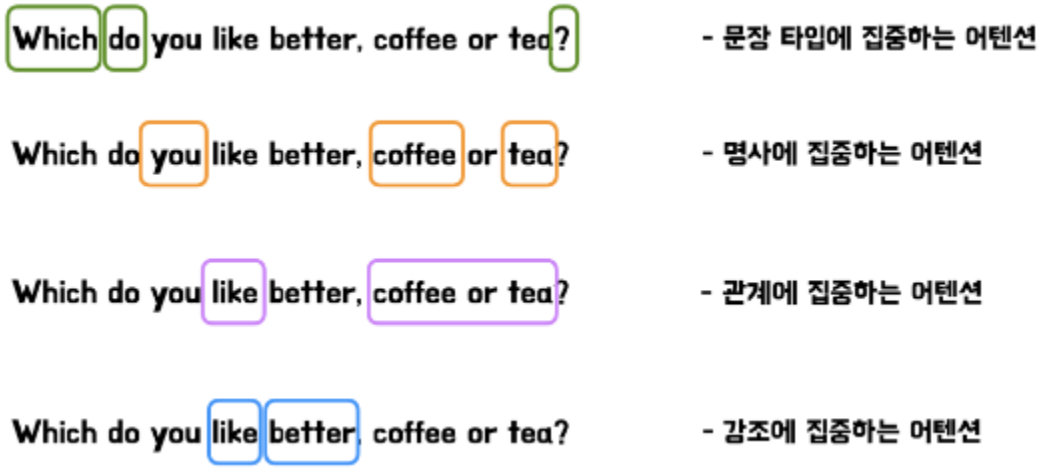
- 개별 단어와 전체 입력 시퀀스를 대상으로 어텐션 계산을 수행해 문맥 전체를 고려하며 모든 경우의 수를 고려하기 때문에 시퀀스 길이가 길어지더라도 정보를 잊거나 왜곡하지 않음



멀티 헤드 셀프 어텐션

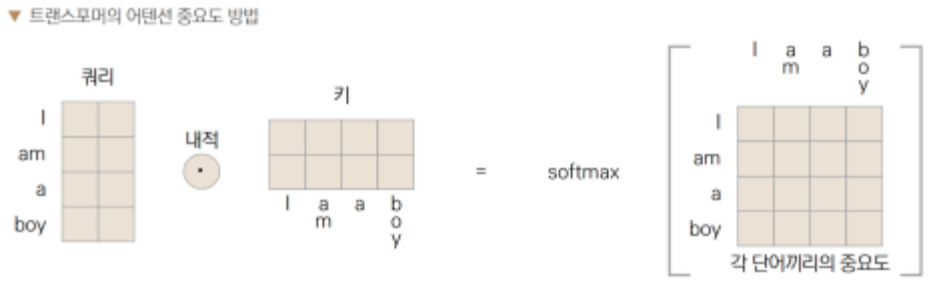
- 어텐션을 병렬로 수행하여 다른 시각으로 정보를 수집
- 모델이 입력 토큰 간의 다양한 유형의 종속성을 포착하고 다양한 소스의 정보를 결합가능

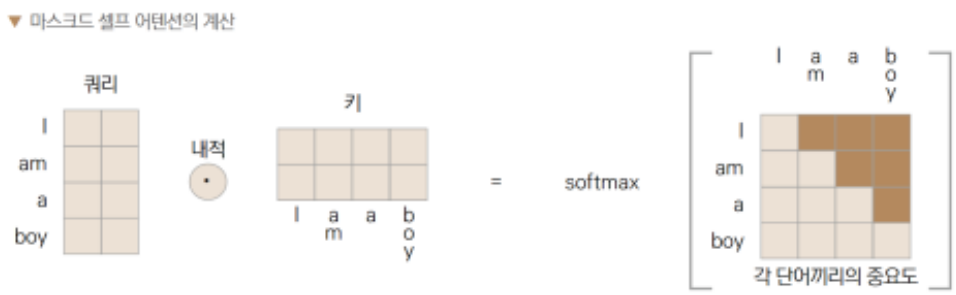




어텐션 중요도 계산

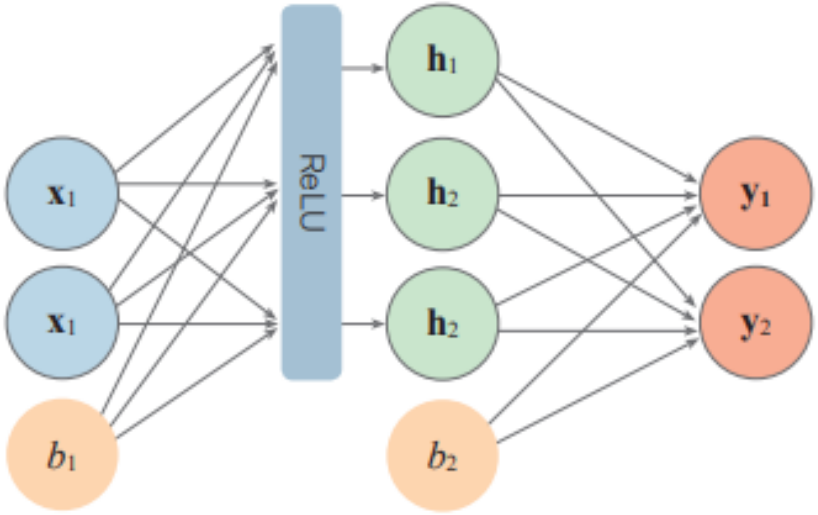
- 쿼리와 키를 곱한 결과를 소프트맥스 함수의 입력으로 사용해 각 단어끼리 중요도를 구하고 밸류에 곱함
- 디코더 입력의 셀프 어텐션에서는 앞으로 오게 될 단어를 마스킹 하기 위해 아주 큰 음수를 넣어서 소프트맥스 계산 시 0에 가깝게 나오도록 함





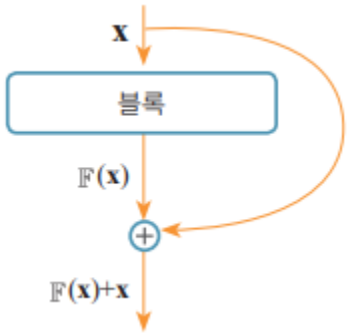
피드포워드 뉴럴 네트워크

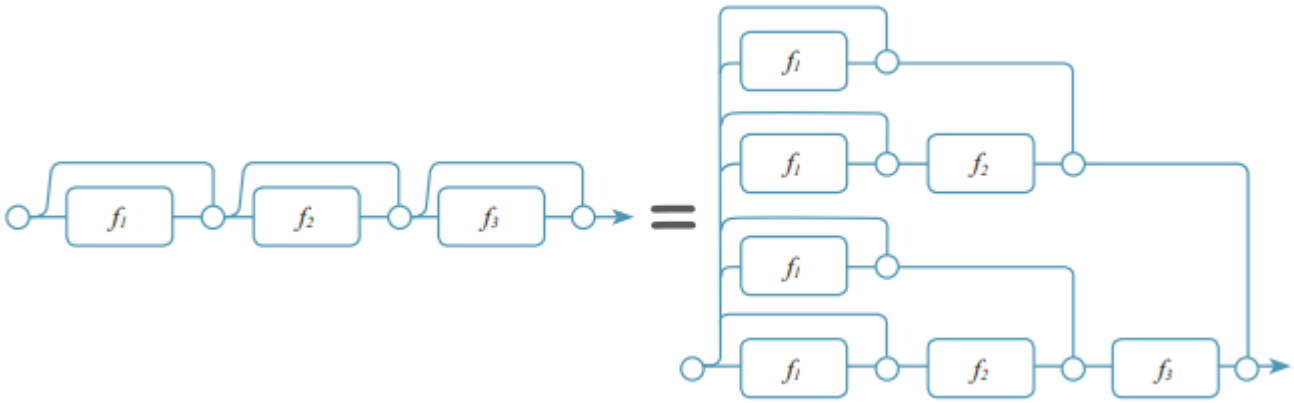
- 신경망의 한 종류로 입력층, 은닉층, 출력층 3개 계층으로 구성
- 오직 입력층에서 출력층 방향으로만 연산이 전개



잔차 연결

- 블록이나 레이어 계산을 건너뛰는 경로를 하나 두는 것
- 잔차 연결을 하지 않으면 하나의 경로만 존재했으나 잔차 연결을 추가하여 새로운 경로가 생김
- 레이어가 많아지면 학습이 어려운 경향이 있는데 잔차 연결을 통해 블록을 건너뛰는 경로를 설정하여 학습을 쉽게함

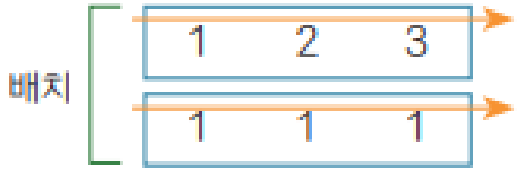




레이어 정규화

- 미니 배치의 인스턴스(x) 별로 평균(E[x])을 빼주고 표준편차로 나눠 정규화를 수행하는 기법.
- β와 γ는 학습 과정에서 업데이트되는 가중치이고 ε는 분모가 0이 되는 것을 방지하기 위해 더해주는 고정값(보통 1e-5)
- 학습이 안정되고 속도가 빨라지는 효과
  (Batch size에 영향이 적음)

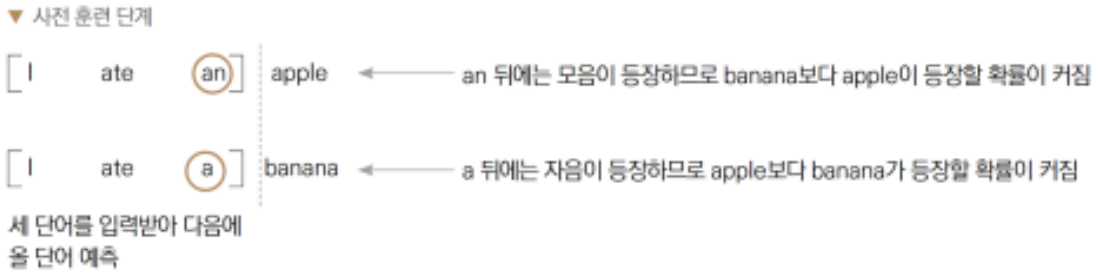
$$\mathbf{y} = \frac{\mathbf{x} - \mathbb{E}[\mathbf{x}]}{\sqrt{\mathbb{V}[\mathbf{x}] + \epsilon}} \gamma + \beta$$

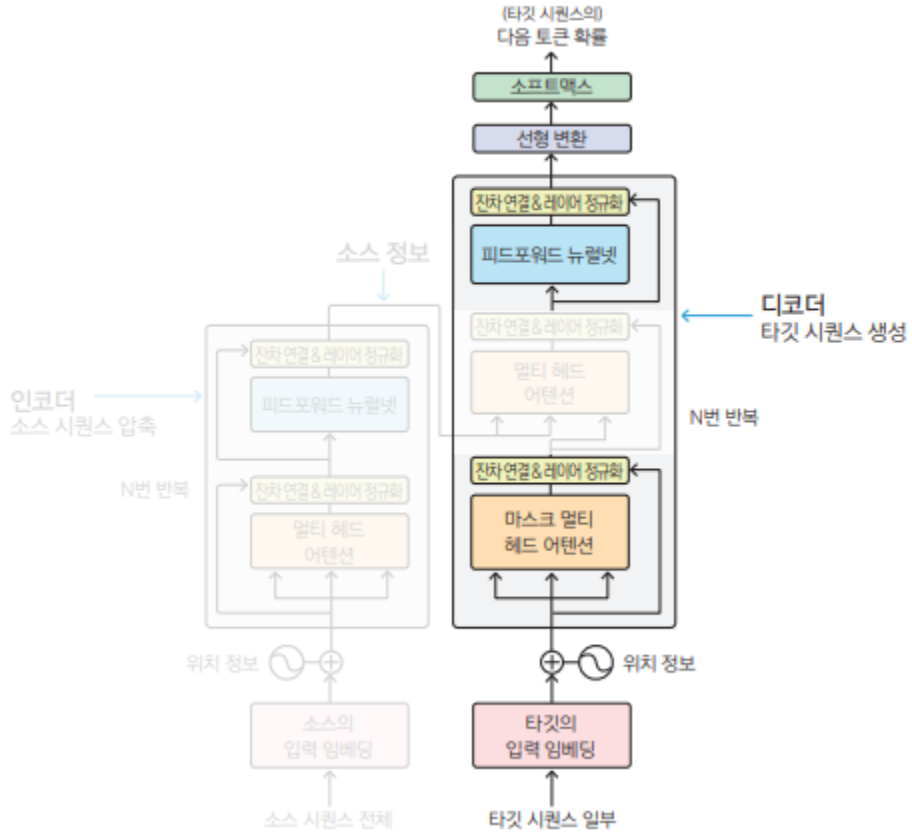


평균	표준편차
2	0.8164
1	0

GPT의 특징

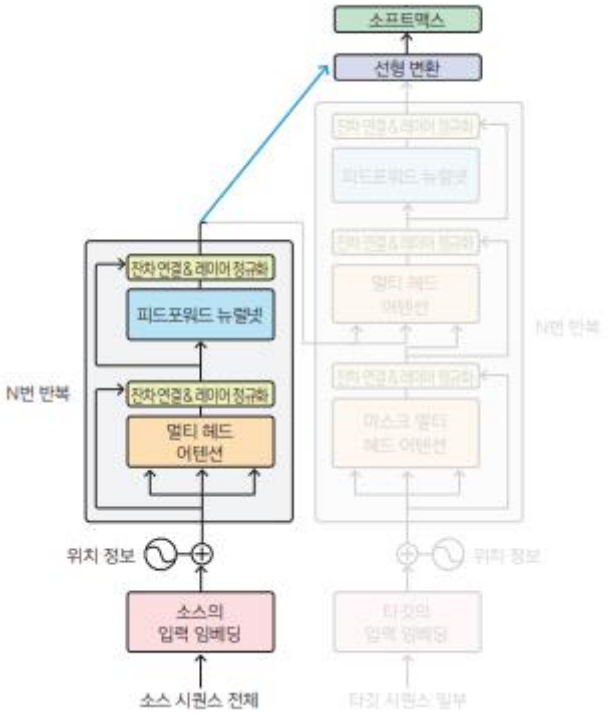
- 문장 왼쪽부터 오른쪽으로 일방향 성격을 가짐.
- 트랜스포머에서 디코더만 사용하고 인코더에서 보내는 정보를 받는 멀티 헤드 어텐션을 사용하지 않음





BERT의 특징

- 마스크 언어 모델으로 빈칸 앞뒤 문맥을 모두 파악 할 수 있는 양방향 성격을 가짐
- 디코더를 제외하고 인코더만 사용



ViT

- Vision Transformer의 약자로 최초로 트랜스포머를 이미지에 활용한 모델
- 이미지의 가로 세로를 일정 비율로 나눈 다음 2차원 이미지를 1차원 벡터로 변환, MLP 모델을 이용해 픽셀에 임베딩 정보를 부여하고 트랜스포머 인코더로 모델 예측값 출력