머신러닝 기반 감성 분석

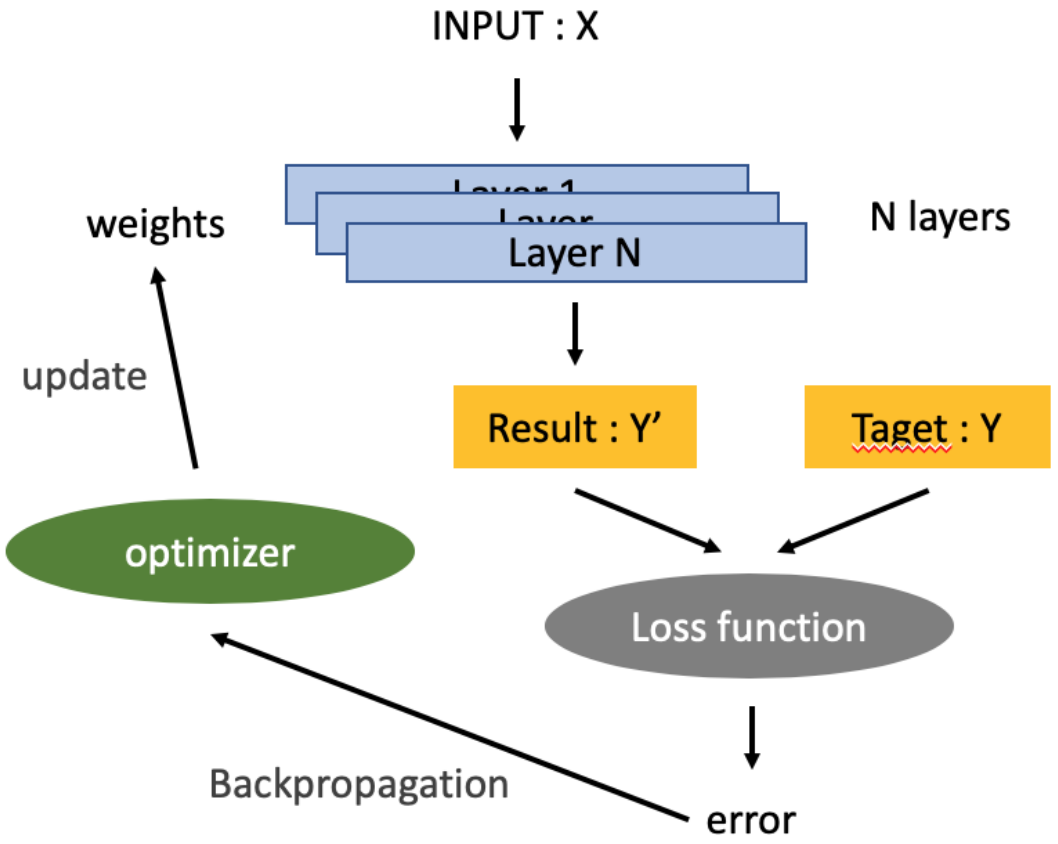
딥러닝 개요
+ 현장에서의 머신러닝

한국폴리텍대학 성남캠퍼스 인공지능소프트웨어과
이혜정 교수

딥러닝

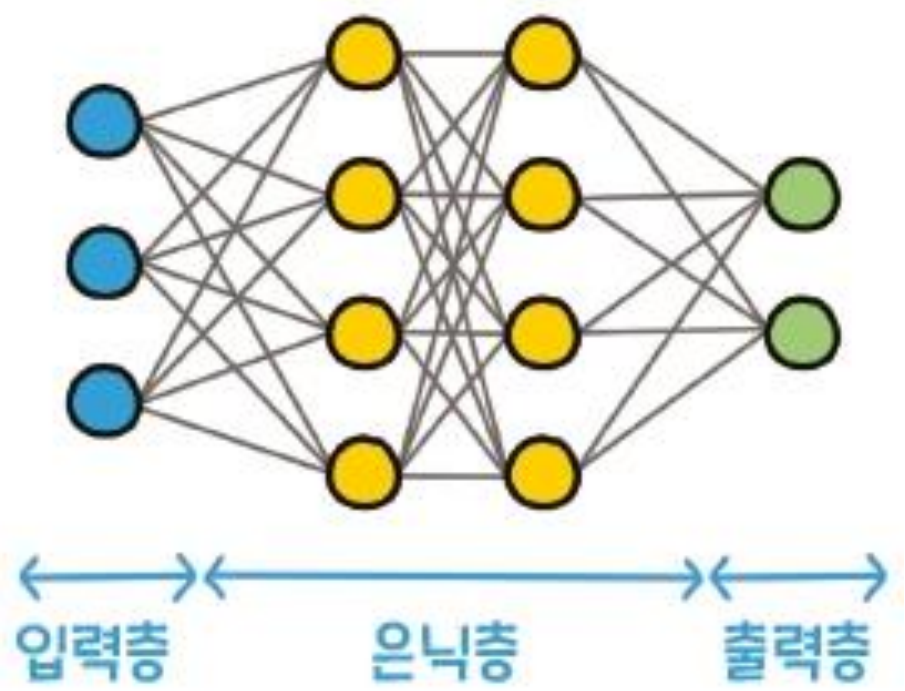
- 다양한 유형의 레이어로 구성된 인공 신경망을 사용하여 복잡한 패턴과 특징을 학습하고 이해하는 모델을 만드는 기술

- 구성요소
  - Layer
  - Activation Function
  - Loss Function
  - Backpropagation
  - Optimizer



DFN, Deep Feedforward Network

- **딥러닝에서 가장 기본으로 사용하는 인공신경망 (심층신경망)**
  - 입력층, 은닉층, 출력층으로 구성, 은닉층이 2개 이상



  - 입력 데이터가 시간 순서에 따른 종속성을 가질 경우 시계 열데이터 처리에 한계가 있음

RNN, Recurrent Neural Network

- **시계열 데이터와 같이 시간적으로 연속성이 있는 데이터를 처리하기 위해 고안된 인공신경망**
  - 은닉층의 각 뉴런에 순환 구조를 추가하여 이전에 입력된 데이터가 현재 데이터를 예측할 때 다시 사용될 수 있도록 함
  - 따라서 현재 데이터를 분석할 때 과거 데이터를 고려한 정확한 데이터 예측

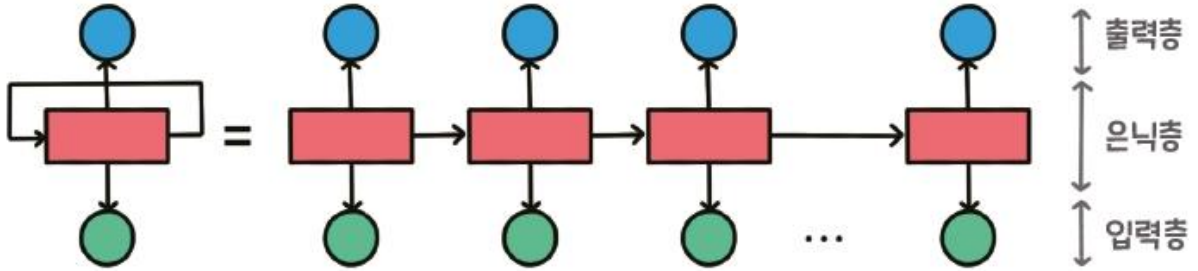


그림 9-23 RNN

  - 신경망 층이 깊어질수록(은닉층 수가 많을수록) 먼 과거의 데이터가 현재에 영향을 미치지 못하는 문제가 발생함 –〉 '장기 의존성(Long-Term Dependency)' 문

  - 이를 해결하기 위해 제안된 것이 LSTM

LSTM(Long Short-Term Memory)

- **신경망 내에 메모리를 두어 먼 과거의 데이터도 저장할 수 있도록 함**
  - 입출력을 제어하기 위한 소자를 두었는데, 이것을 게이트(Gate)라고 함
    - 게이트는 입력 게이트, 출력 게이트, 망각 게이트가 있음
      - 입력 게이트 : 현재 정보를 기억
      - 망각 게이트 : 과거 정보를 어느 정도 기억할지 결정
      - 출력 게이트 : 출력층으로 출력할 정보의 양을 결정
  - 시퀀스에서 작동하고 후속 단계의 입력으로 자신의 출력을 사용

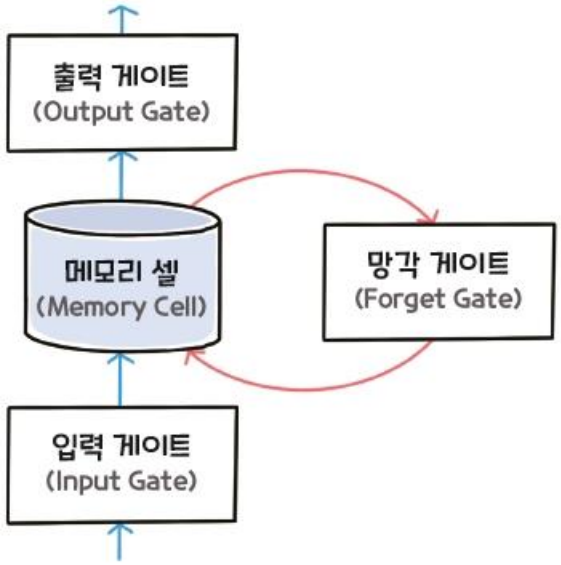
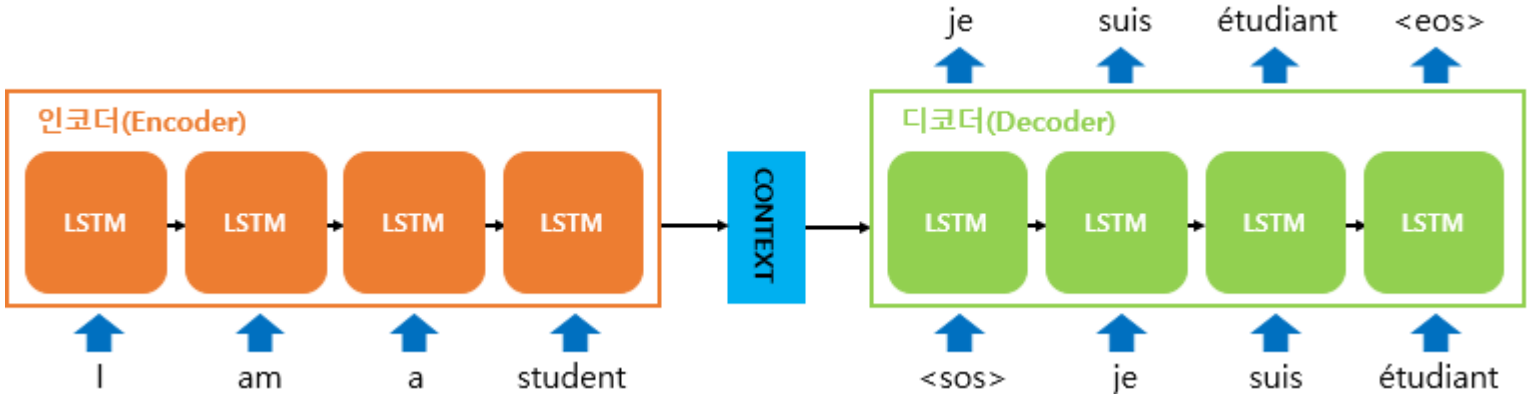


그림 9-24 LSTM

- **RNN, LSTM 문제**
  - Vanishing Gradient 문제

Seq2Seq

- **Seq2Seq (Encoder Decoder Network)**
  - 2개의 RNN으로 구성
  - Encoder : 입력 시퀀스를 읽고 단일 벡터 출력 (Context Vector)
  - Decoder : Context Vector를 읽어서 출력 시퀀스 생성

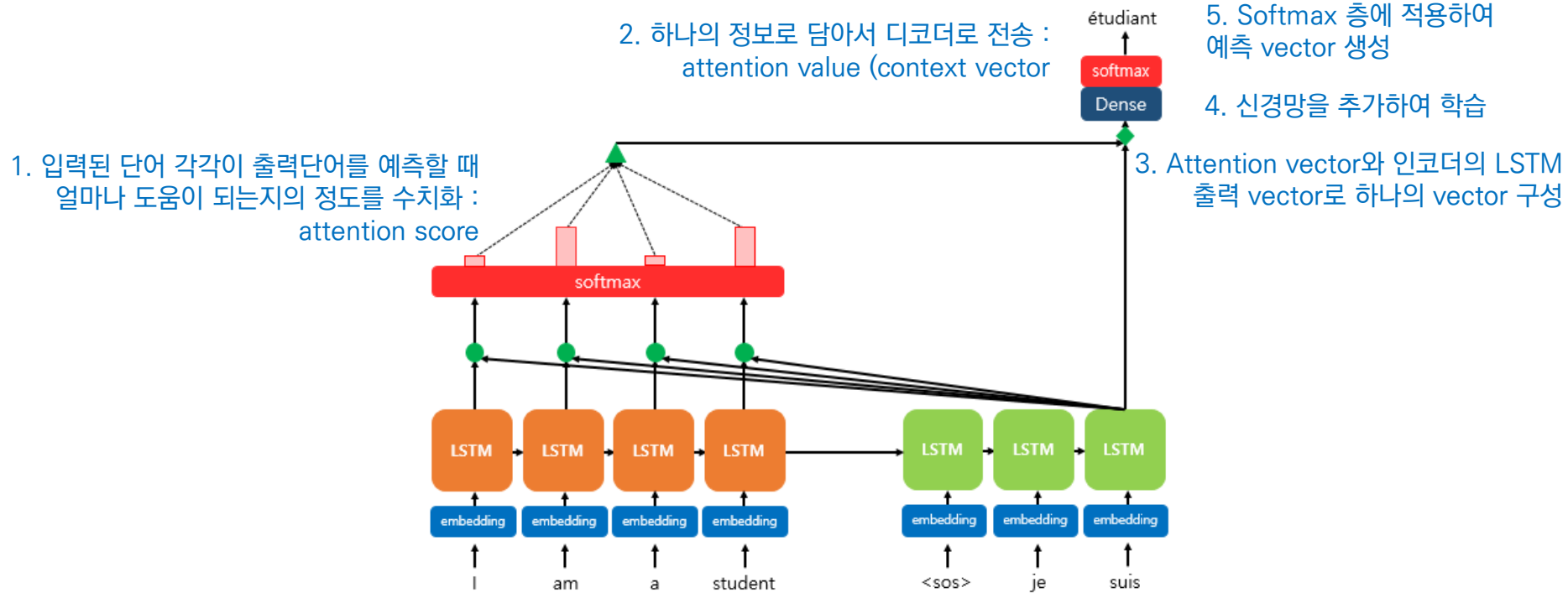


- **문제점**
  - 하나의 고정된 크기의 벡터에 모든 정보를 압축하려다 보니 정보 손실 많음
  - RNN(LSTM)의 고질적인 문제인 Vanishing Gradient 문제 존재

  –〉 중요한 단어에 집중하여 Decoder에 바로 전달하는 Attention 기법 등장

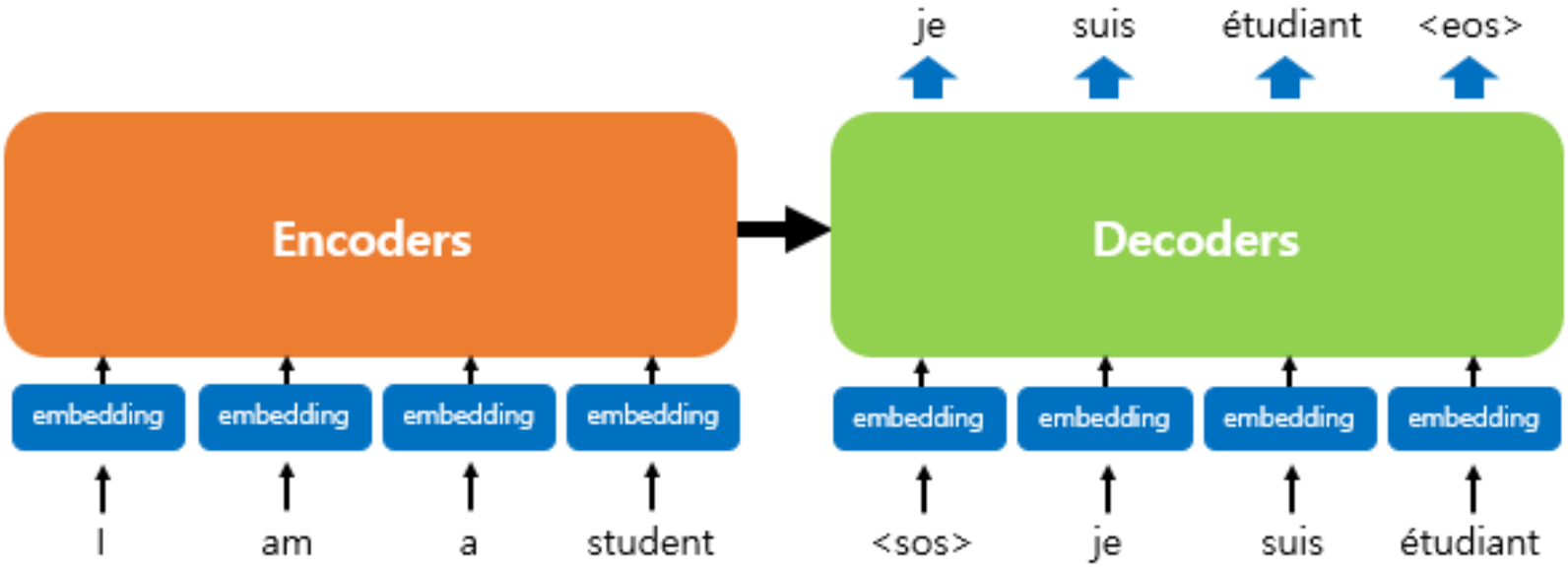
Attention

- 기본 아이디어
  - 디코더에서 출력 단어를 예측하는 매 시점(time step)마다, 인코더에서의 전체 입력 문장 중 해당 시점에서 예측해야할 단어와 연관이 있는 입력 단어 부분을 좀 더 집중(attention)해서 보자



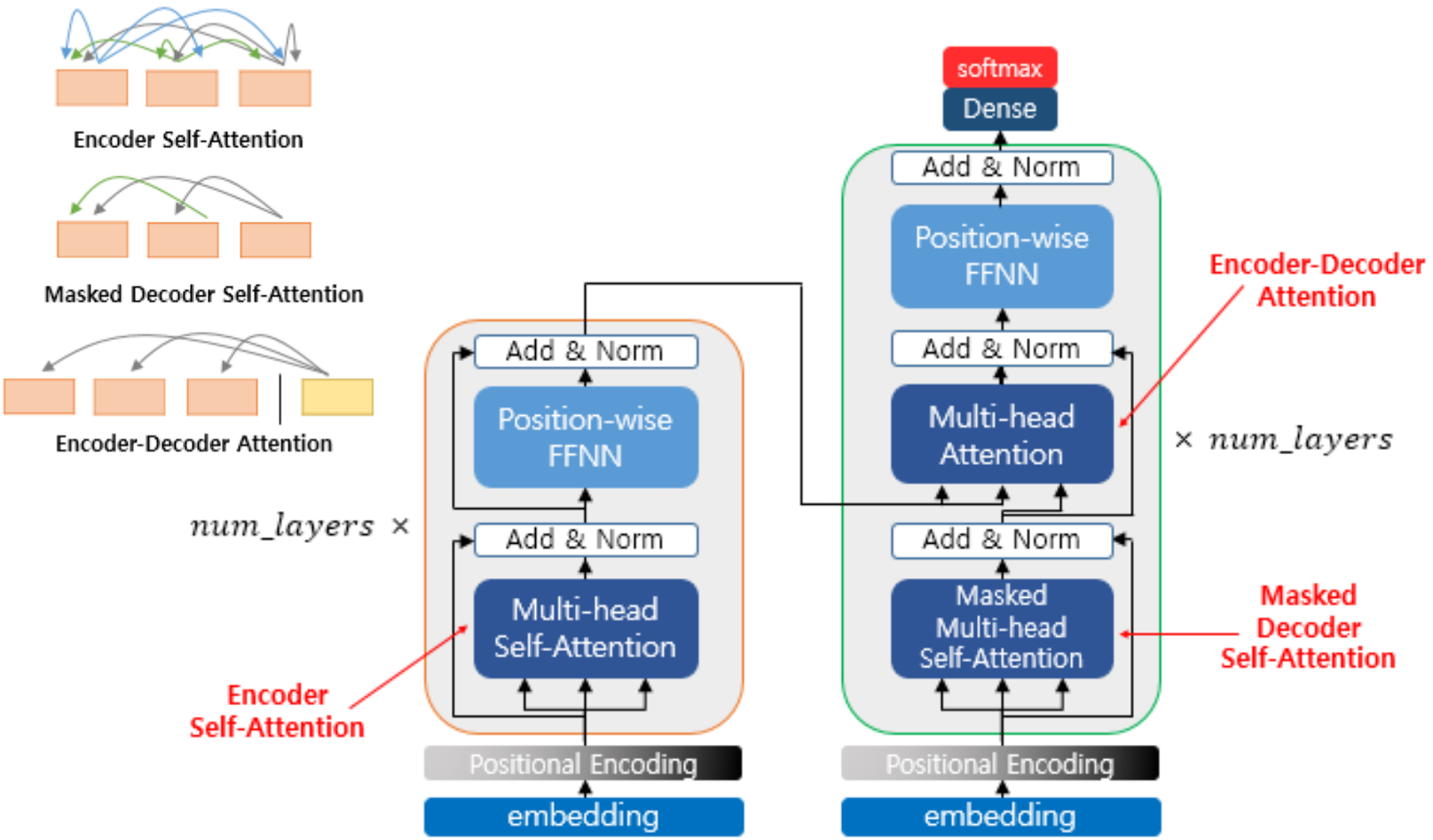
Transformer

- 기본 아이디어
  - attention을 RNN을 보정하기 위한 용도가 아닌, attention만으로 encode와 decoder를 만들어 보자



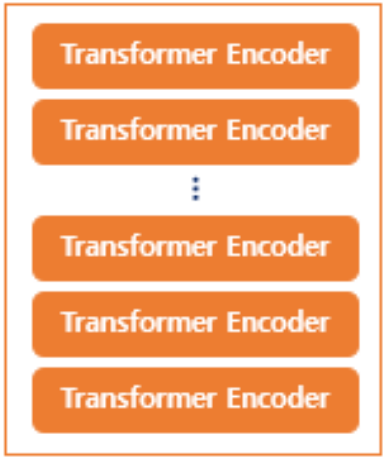
Transformer

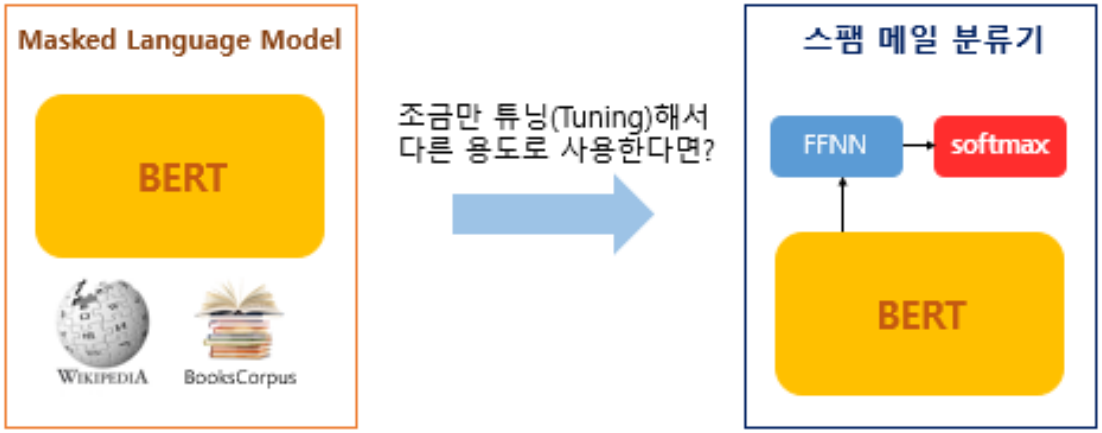
- Transformer = encoders + decoders
  - Encorder or Decorder = attention + positional encoding + feedforward net



BERT

- Bidirectional Encoder Representations from Transformers
- 트랜스포머의 인코더 12 layer (BERT-Base), 24 layer (BERT-Large)로 구현
  - 사전 훈련된 언어 모델
    - 위키피디아(25억 단어)와 BooksCorpus(8억 단어)와 같은 레이블이 없는 텍스트 데이터로 학습
  - 파인 튜닝(Fine-tuning)
    - 레이블이 있는 다른 작업(Task)에서 추가 훈련과 함께 하이퍼파라미터를 재조정하여 이 모델을 사용하면 성능 확보



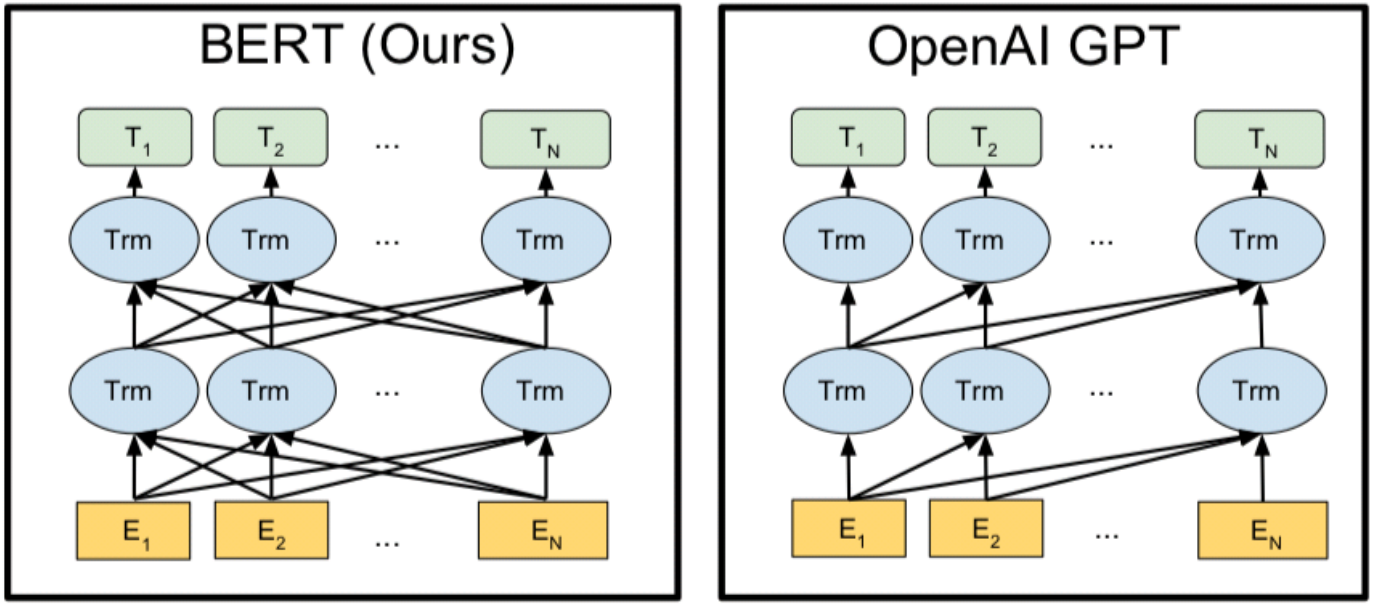


33억 단어에 대해서 4일간 학습시킨 언어 모델

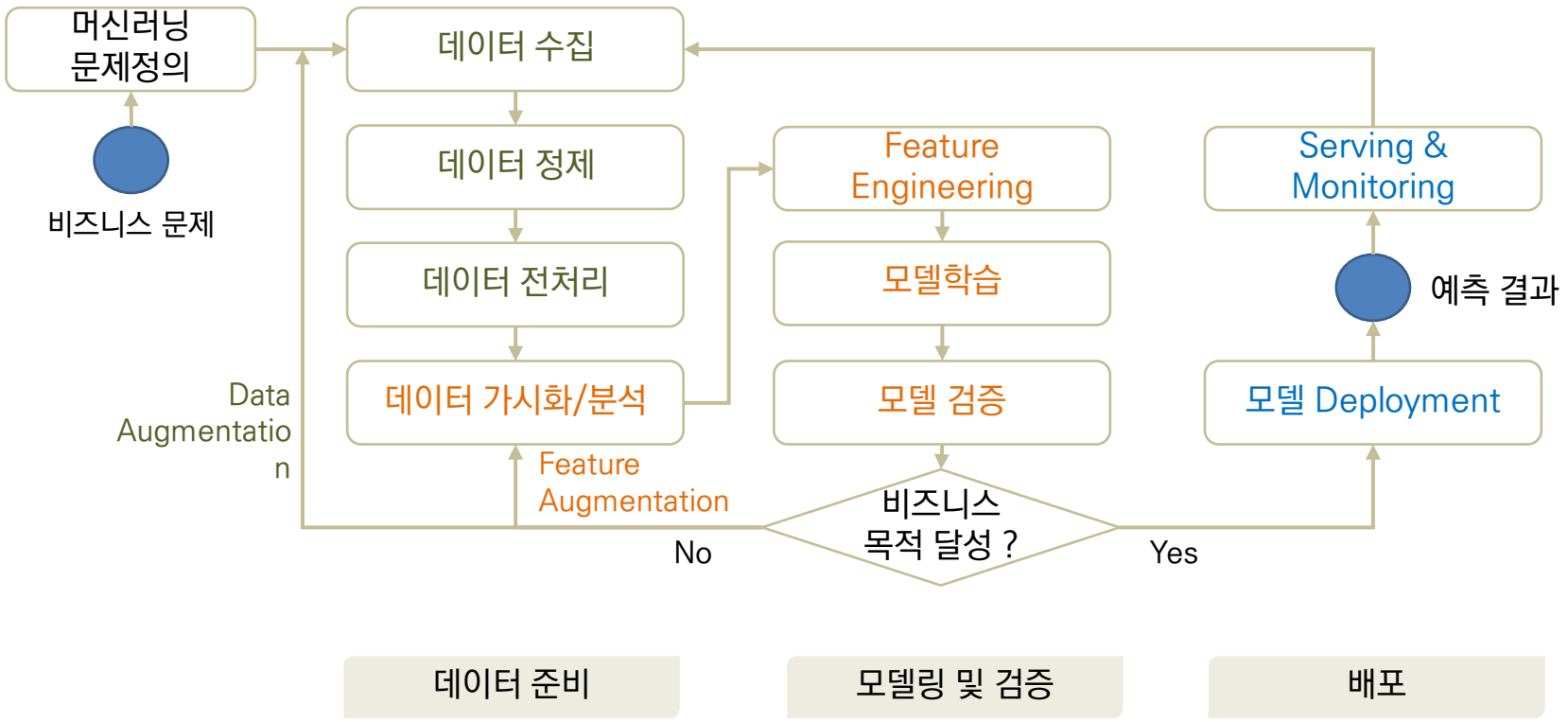
BERT의 지식을 이용한 스팸 메일 분류기

BERT

- **BERT의 사전 훈련 (Pre-training)**
  - Masked 언어 모델 (MLM)
    - 인공 신경망의 입력으로 들어가는 입력 텍스트의 15%의 단어를 랜덤으로 마스킹 (Masking) -> masking된 단어를 예측하도록 학습
  - 다음 문장 예측(Next Sentence Prediction, NSP)
    - 두 개의 문장을 준 후에 이 문장이 이어지는 문장인지 아닌지를 맞추는 방식으로 훈련



MLOps, 머신러닝 파이프라인



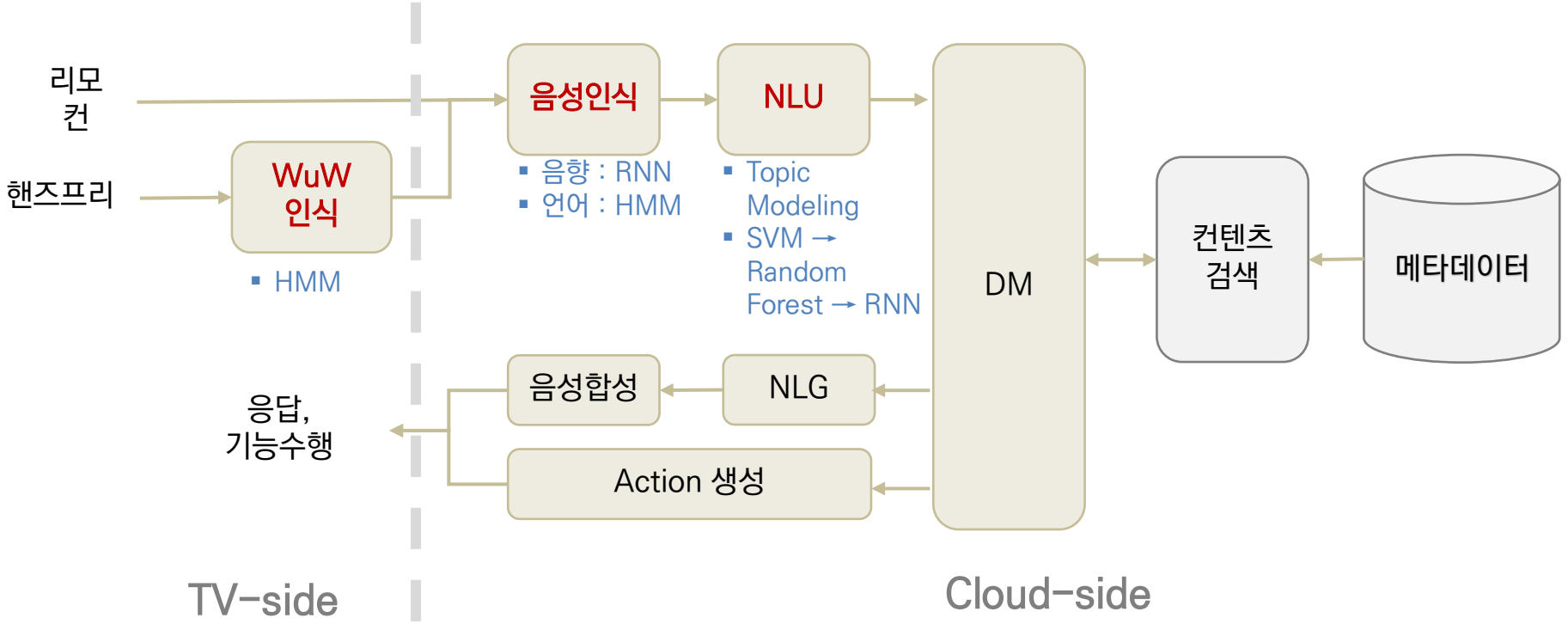
머신러닝 기반 감성 분석

현장에서의 머신러닝

TV용 음성대화 솔루션

머신러닝 모델 선택

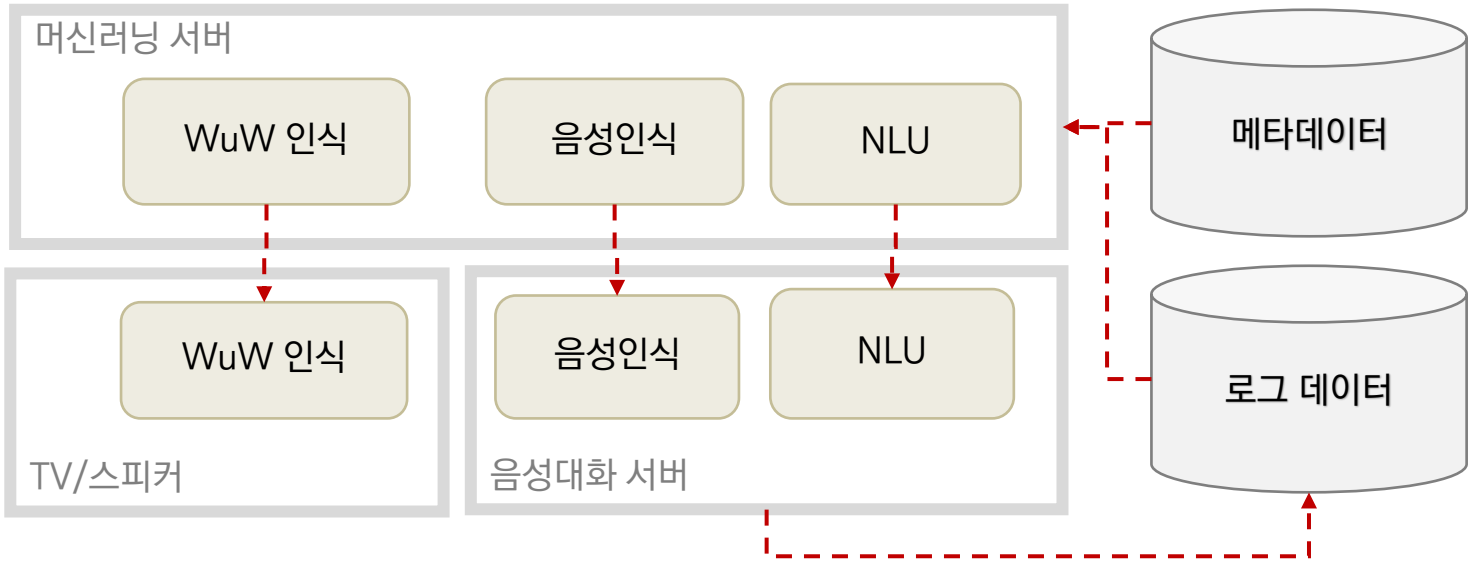
- WuW인식, 음성인식, NLU에 머신러닝 적용



TV용 음성대화 솔루션

머신러닝 파이프라인

- 메타데이터와 로그데이터를 학습 데이터에 반영하여 지속적으로 성능 개선 (주기 : 2주)



보안

- 사용자가 의도한 음성만 서버로 전송하기 위해, WuW 인식을 Device에서 수행

편향

- 정치, 사회적인 이슈에 대해서 NLU 이전에 Topic Modeling 적용하여 Out-of-Domain 처리