딥러닝팀

1팀

안세현
이수정
이승우
전효림
홍지우

INDEX

1 자연어의 특징

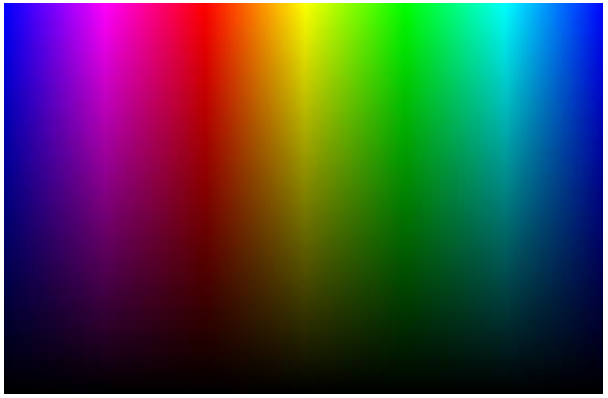
자연어 특징

불연속적 데이터

- 단어의 모습만으로 의미가 유사하다고 말하기에는 어려움이 있음.
- 예) 배

✓ 연속적 데이터

예) 컬러 이미지



2 자연어의 전처리

정제(Cleaning)

정제

자연어 데이터를 다룰 때 사용할 여러 문장 데이터 집합

코퍼스에서 용도에 맞게 노이즈 데이터를 제거하는 일

특수 문자 제거, 영어의 대/소문자 통일

예1.

There is an apple

there is an apple

컴퓨터는 다르게 인식하나, 의미적인 변화가 없는 두 문장이므로 통일

예2.

US(미국)

us(우리)

서로 다른 의미를 갖는 경우가 있으므로 기계적으로 변환해서는 안 됨.

2 자연어의 전처리

토큰화(Tokenization)

토큰화

일반적으로 의미 있는 단위로 정의

- 코퍼스를 토큰이라 불리는 단위로 나누는 작업.

예1.

입력: Time is an illusion. Lunchtime double so!

출력: “Time”, “is“, “an”, “illusion”, “Lunchtime”, “double”, “so”

✓ 특수문자 제거

띄어쓰기 단위로 끊음

Subword Segmentation

Subword Segmentation

- '단어는 의미를 가진 더 작은 서브워드들의 조합으로 이루어진다'는 가정 하에 단어를 쪼개는 것
- OOV 문제를 해소하기 위한 하나의 방법

OOV 문제

모델이 학습하지 못한 단어(OOV, UNK)가 나타나는 문제

OOV: Out-Of-Vocabulary

UNK: Unknown Token

예) rebirth

re + birth

학습하지 못한 단어, 신조어 등에 더 효율적인 대처 가능

One Hot Encoding

One Hot Encoding

- N개의 단어를 각각 N차원의 벡터로 표현하는 방법
- 표현하고 싶은 단어 인덱스에 1, 나머지는 0 부여

- 예)

Pet
Cat
Dog
Turtle
Fish

→

Cat	Dog	Turtle	Fish
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

3 Embedding

Word Embedding

Word Embedding

- 단어를 밀집 벡터로 표현하는 방법

I	0.2	0.58	0.72
am	0.4	0.3	0.39
a	0.5	0.82	0.22
blogger	0.1	0.03	0.81
studying	0.05	0.76	0.64
data	0.81	0.52	0.33
analysis	0.65	0.72	0.4

임베딩 벡터

임베딩을 거친 결과인 밀집벡터

벡터들로 단어 간 유사도 계산 가능

분산표현

- 단어의 의미를 유지하면서 밀집 벡터를 만드는 것
  - 분포 가설을 기반

분포가설

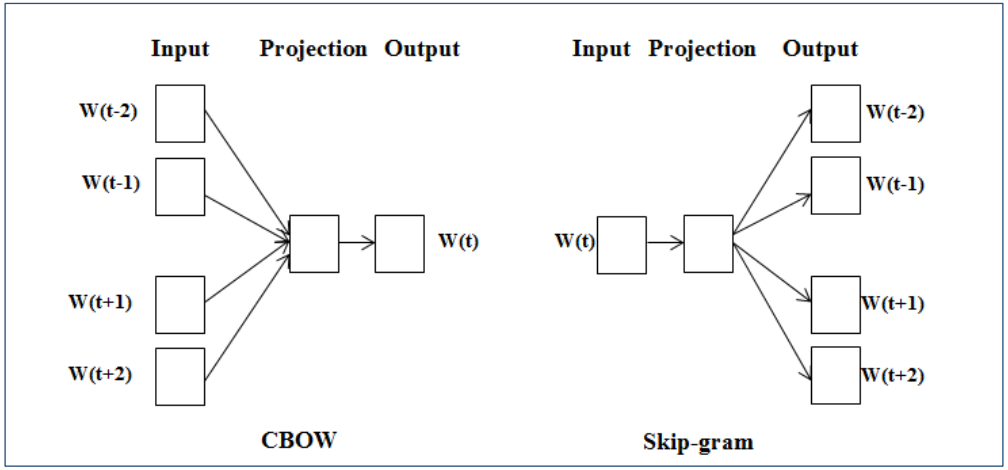
- '비슷한 위치에 등장하는 단어는 비슷한 의미를 가진다 ' 는 가정

3 Embedding

Word2vec

Word2vec

- 2013년 구글에서 개발
- 분산 표현 방법을 이용해 단어를 벡터로 바꾸는 방법



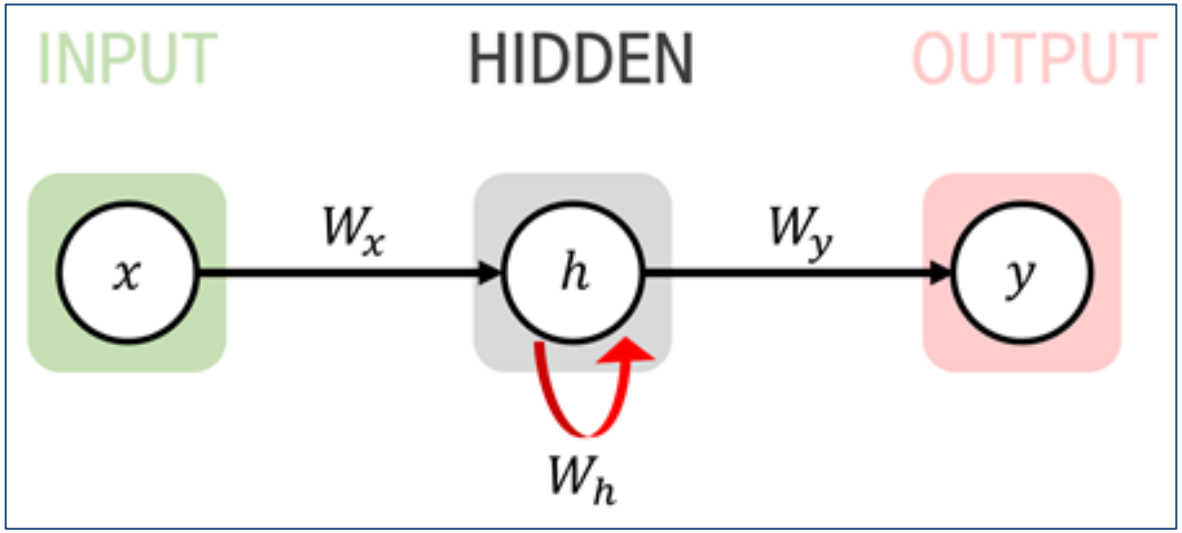
두 가지 방식

1. CBOW: 여러 개의 주변 단어를 바탕으로 하나의 중심 단어를 예측하는 모델
2. Skip-gram: 하나의 입력으로 여러 개의 주변 단어를 예측하는 모델

4 RNN

- RNN이란?

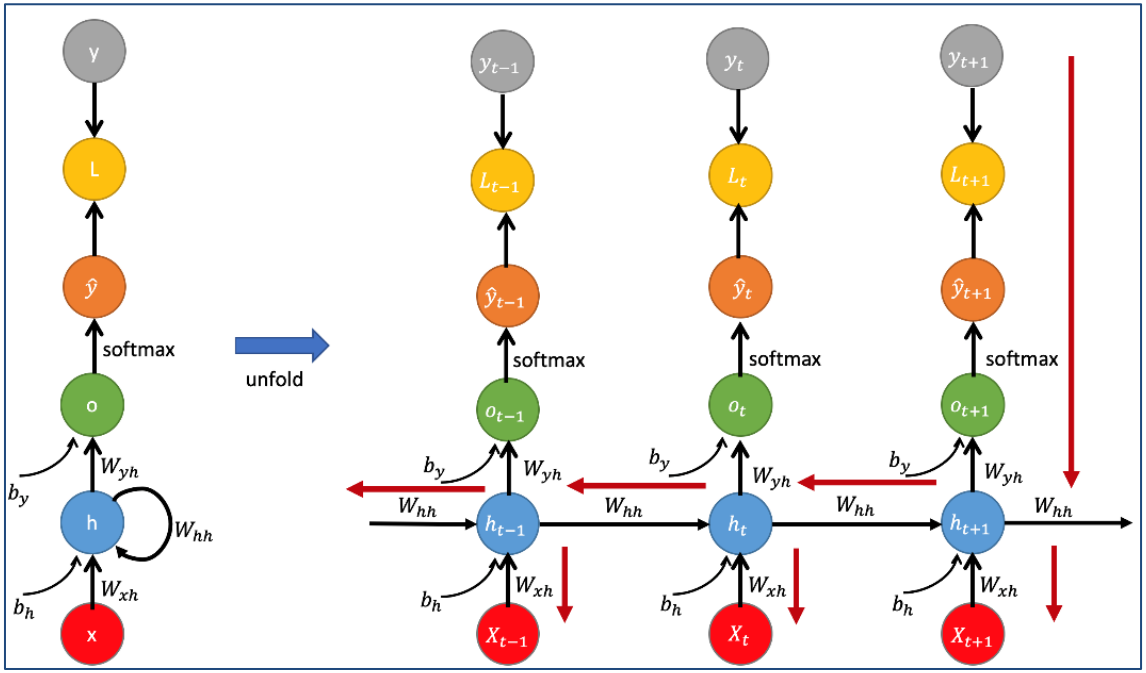
RNN(Recurrent Neural Network)



Sequential Data를 다루기 위한 신경망으로, 순환적인 구조를 갖는 모델

- RNN의 역전파

BPTT(Back Propagation Through Time)

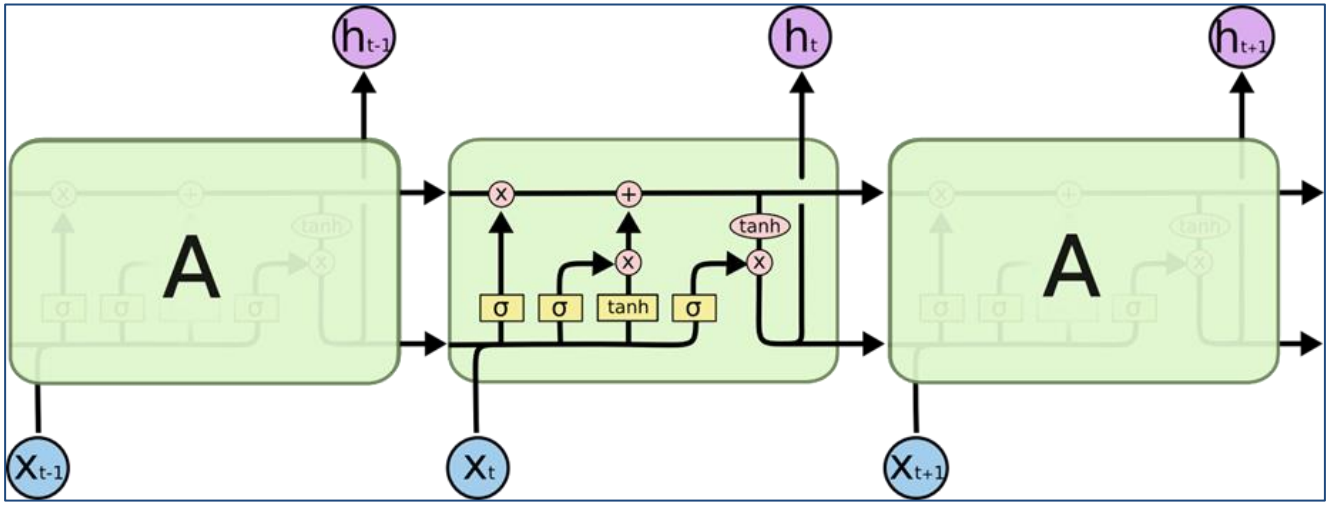


RNN: 매 time step마다 예측값이 출력 → 매 time step마다 loss, gradient 계산이 가능

RNN의 역전파: gradient들의 평균으로 가중치를 업데이트

- LSTM

LSTM(Long Short-Term Memory)이란?



LSTM은 RNN에 **단기 기억과 장기 기억**의 원리를 추가한 모델

RNN의 출력은 1개이지만, LSTM의 출력은 2개

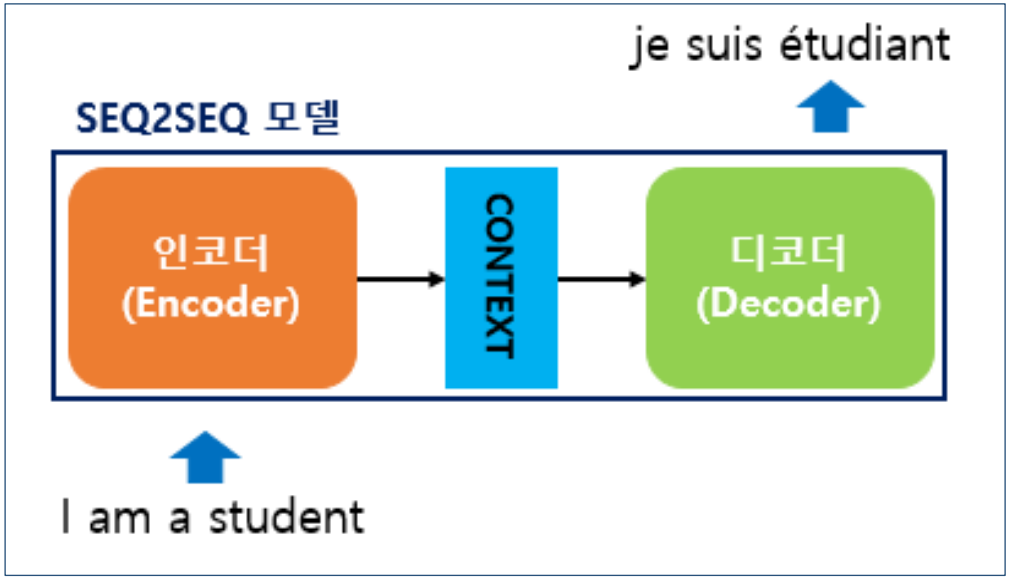
5 Seq2Seq

- Seq2Seq란

Sequence-to-Sequence

입력 시퀀스를 받아 새로운 출력 시퀀스를 내는 모델

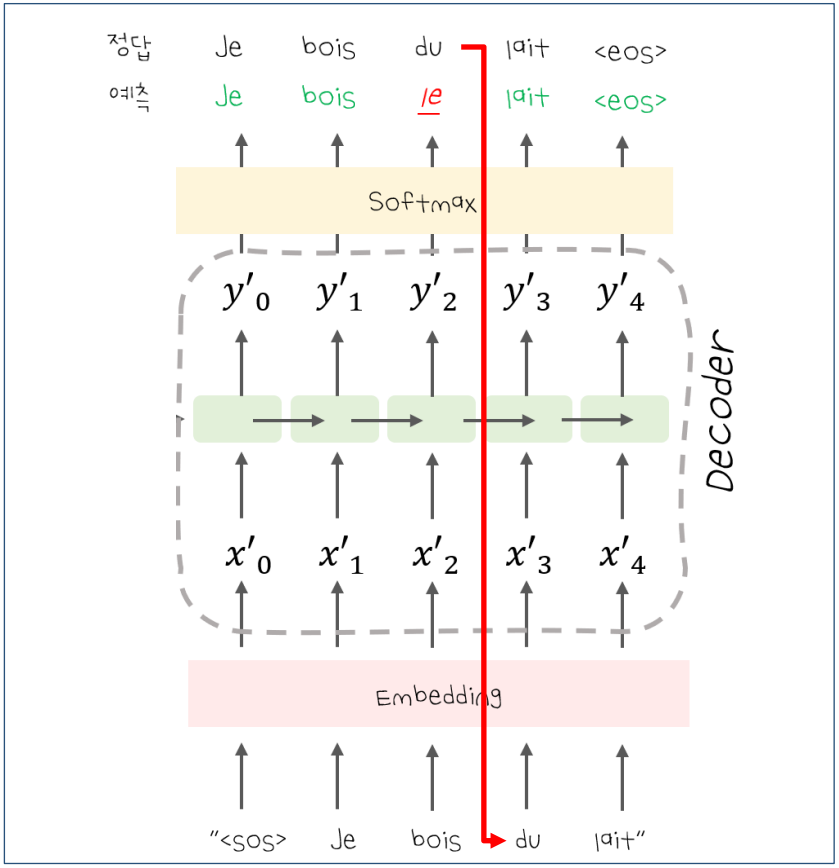
번역, 문서 요약 등에 쓰임



5 Seq2Seq

- Seq2Seq의 구조

디코더 (Decoder)



교사 강요 (Teaching Forcing)

학습 중 디코더가 예측을 잘못한 경우

원활한 학습을 위해

디코더의 출력 대신

실제 정답을 디코더의 입력으로 넣는 것

5 Seq2Seq

- 출력 단어 선택 방식

Greedy search decoding

Softmax 함수를 거친 단어 후보군들 중 가장 높은 값을 가진 단어 선택

틀린 예측을 했을 경우 돌이킬 수 없어 연쇄적인 오류 발생

Exhaustive search decoding

각 시점에서 가능한 모든 단어를 이용해 문장을 생성한 후 확률이 가장 높은 문장 선택

모든 경우의 수를 계산해 시간과 비용 측면에서 비효율적

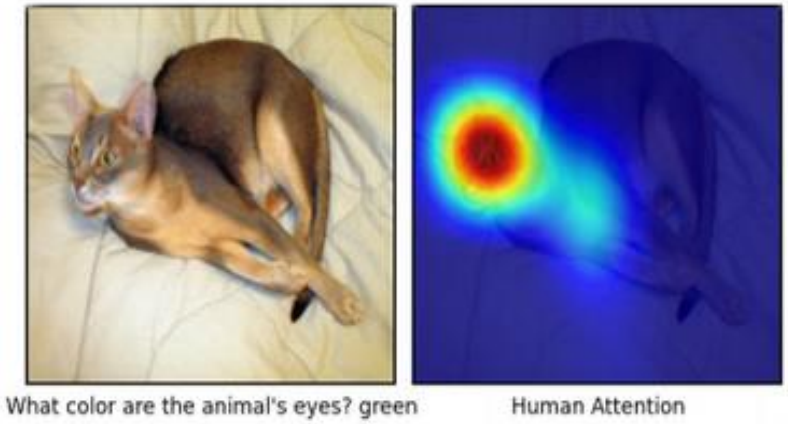
Beam search Decoding

위의 두 방법의 중간지점에 있는 방식

임의로 정한 k개의 후보군을 유지하며 출력할 단어를 선택하는 방법

6 Attention

- Attention이란

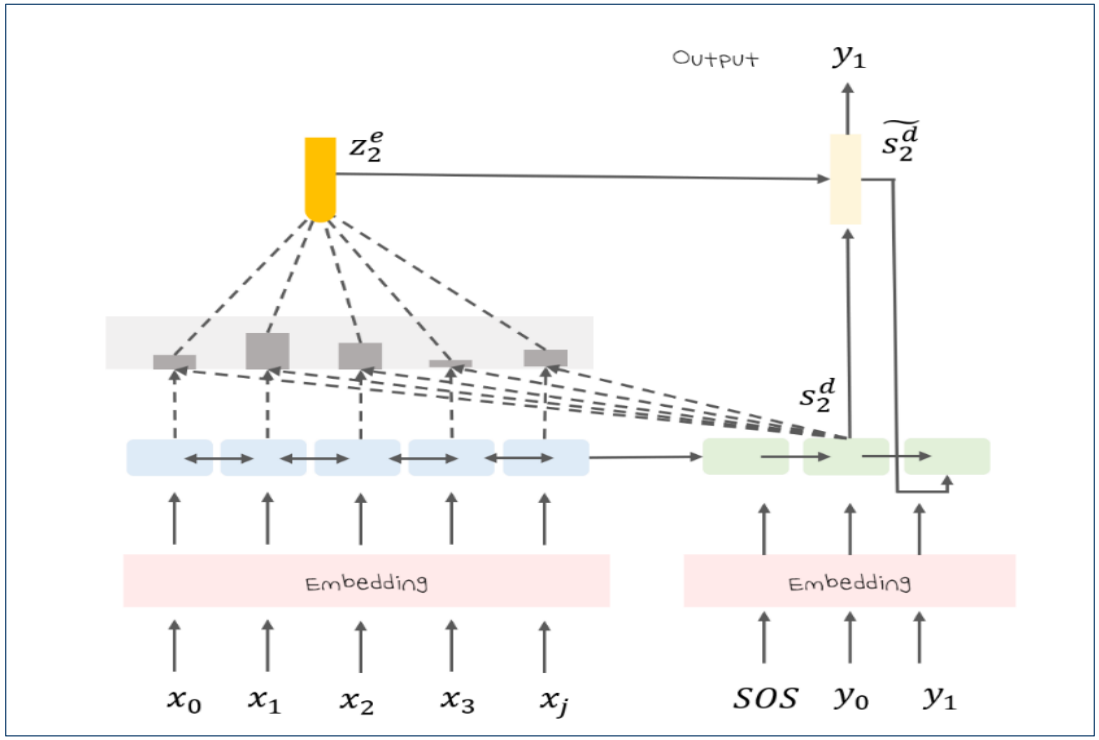


이와 같이 필요한 정보에 집중하는 방식에서 착안하여

디코더가 예측할 때 입력 시퀀스를 참조하여 필요한 부분에 집중하는 것

6 Attention

- Attention의 진행과정



인코더 부분은 기존의 seq2seq와 동일