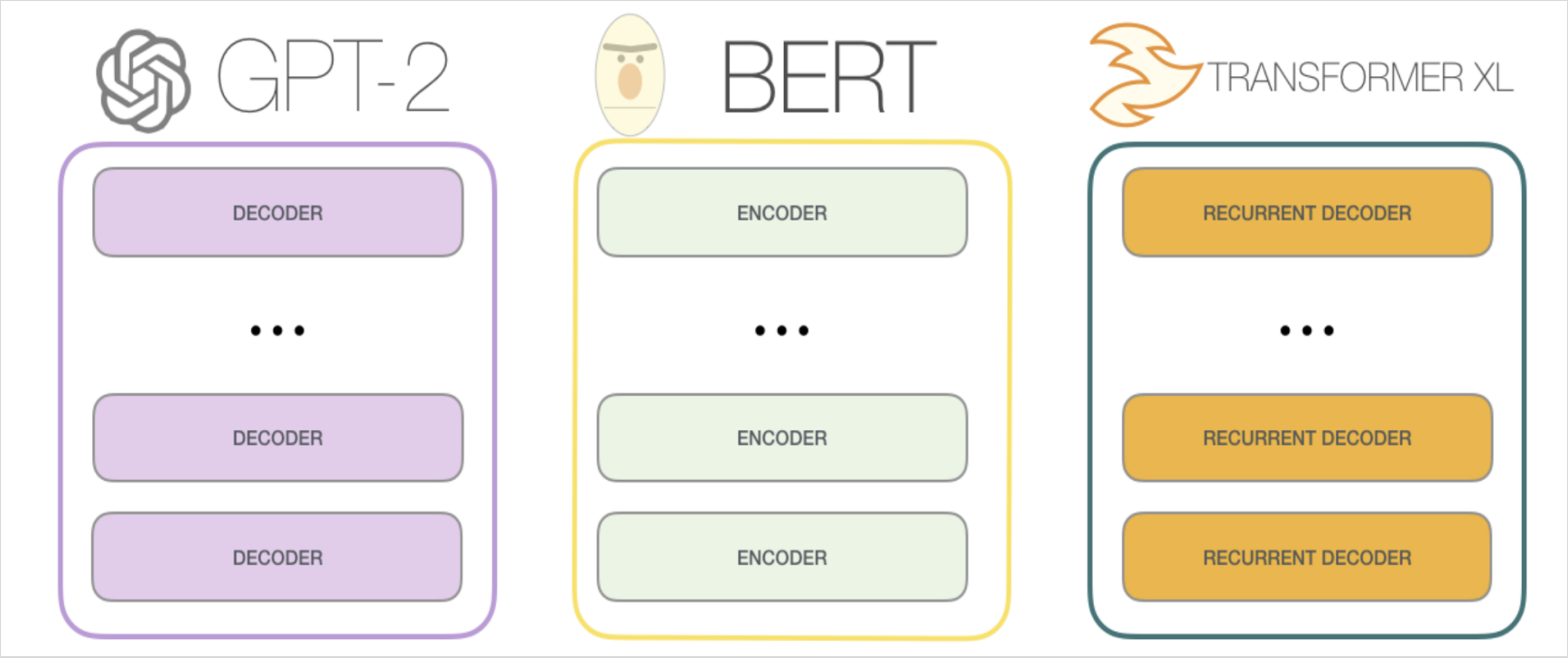
한국어 임베딩

서론

1. 임베딩이란?
2. 임베딩의 역할
3. 임베딩 기법의 역사와 종류
4. 이 책에서 다루는 데이터와 주요 용어
5. 5. 요약

1. 1 임베딩이란?

"만약 컴퓨터가 인간을 속여 자신을 마치 인간인 것처럼 믿게 할 수 있다면 컴퓨터를 인텔리전트 하다고 부를 만한 가치가 충분히 있다."



https://talktotransformer.com

1. 1 임베딩이란?

"컴퓨터는 어디까지나 빠르고 효율적인 계산기이다."

- 컴퓨터는 인간의 언어(자연어)를 그대로 이해하지 못한다.
- 컴퓨터는 숫자로 (변형된 말이나 글을) 계산한다.
- 기계의 자연어 이해와 생성은 연산이나 처리의 영역이다.

1. 1 임베딩이란?

Embedding = (Vector space)벡터공간 + (Embed)끼워 넣는다

구분	메밀꽃 필 무렵	운수 좋은 날	사랑 손님과 어머니	삼포 가는 길
기차	0	2	10	7
막걸리	0	1	0	0
선술집	0	1	0	0

- (운수좋은날) 문서의 임베딩은 [2,1,1]
  (막걸리) 단어의 임베딩은 [0,1,0,0]
- (사랑 손님과 어머니)와 (삼포 가는길)이 (기차)라는 소재를 공유한다는 점에서 비슷한 작품이라고 추정 가능하다.
- (막걸리)와 (선술집)이라는 단어가 (운수 좋은 날)이라는 작품에만 등장한 것을 알 수 있다.
- (막걸리 - 선술집) 간 의미 차이가 (막걸리 - 기차) 보다 작을 것이라고 추정 할 수 있다.

1. 2 임베딩의 역할

I. 단어/문장 간 관련도 계산
II. 의미적/문법적 정보 함축

https://github.com/hansw90/ModuLABS/blob/master/MeaniNLP/Word2VecKor_code.ipynb

III. 전이 학습

1. 2 임베딩의 역할

1.2.3 전이 학습 ??

품질이 좋은 임베딩을 쓰면 문서 분류 정확도와 학습 속도가 올라간다. 이렇게 임베딩을 다른 딥러닝 모델의 입력값으로 쓰는 기법을 (전이 학습)이라고 한다.

1. ㅋ 임베딩의 기법의 역사와 종류

통계기반 임베딩에서 뉴럴 네트워크 기반 임베딩

- 통계 기반
  - 잠재 의미 분석 (Latent Semantic Analysis) : 단어 사용 빈도 등 말뭉치의 통계량 정보가 들어 있는 커다란 행렬에 수학적 기법을 적용해 행렬에 속한 벡터들의 차원을 축소하는 방법.
    Ex) TF-IDF 행렬, 단어-문맥 행렬, 단어 – 문서행렬 등

- 뉴럴 네트워크 기반
  - 이전 단어들이 주어졌을 때 다음 단어가 뭐가 될지 예측하거나, 문장 내 일부분에 구멍을 뚫어 놓고(masking) 해당 단어가 무엇일지 맞추는 과정에서 학습된다.

1. ㅋ 임베딩의 기법의 역사와 종류

단어 수준에서 문장수준의 임베딩

- 단어 수준의 임베딩 : NPLM, W2V, Glove, FastText, Swivel 등
  단어 임베딩 기법들은 각각의 벡터에 해당 단어의 문맥적 의미를 함축한다. But 동음이의어(homonym)를 분간하기 힘들다.

- 문장 수준의 임베딩 : ELMo, BERT, GPT 등
  문장 수준의 임베딩 기법은 개별 단어가 아닌 단어 시퀀스 전체의 문맥적 의미를 함축하기 때문에 단어 임베딩 기법보다 전이 학습 효과가 좋은 것으로 알려져 있다.

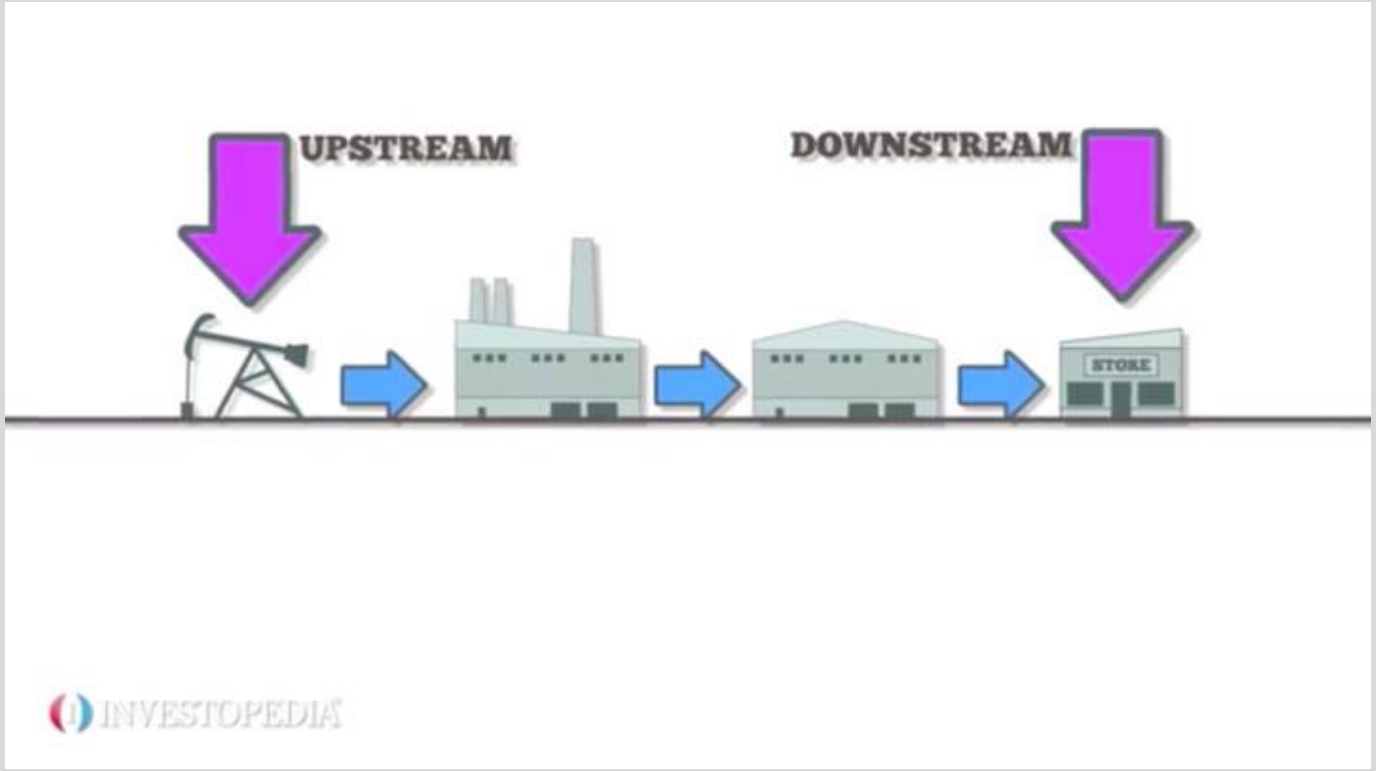
1. ㅋ 임베딩의 기법의 역사와 종류

룰 -> 엔드투엔드 -> 프리트레인/파인 튜닝

- 룰 : 언어학적인 지식을 이용하여 사람이 feature를 직접 뽑는다.
  * feature : 모델의 입력값
- 엔드투엔드 : 데이터를 통째로 모델에 넣고 입출력 사이의 관계를 사람의 개입 없이 모델 스스로 처음부터 끝까지 이해하도록 유도한다.
  *seq2seq 모델이 엔드투 엔드의 대표 사례
- 프리트레인, 파인튜닝 : 대규모 말뭉치로 임베딩을 만들고(이 임베딩엔 의미적 문법적 맥락이 포함됨)우리가 풀고 싶은 구체적 문제에 맞는 수규모 데이터에 맞게 임베딩을 포함한 모델 전체를 업데이트 한다.
  * ELMo, GPT, BERT 등이 이방식에 해당한다.

1. ㅋ 임베딩의 기법의 역사와 종류

DownStream VS Upstream Task (p40)



1. ㅋ 임베딩의 기법의 역사와 종류

임베딩의 종류

1. 행렬 분해 기반 방법
2. 예측 기반 방법
3. 토픽 기반 방법

임베딩 종류별 성능

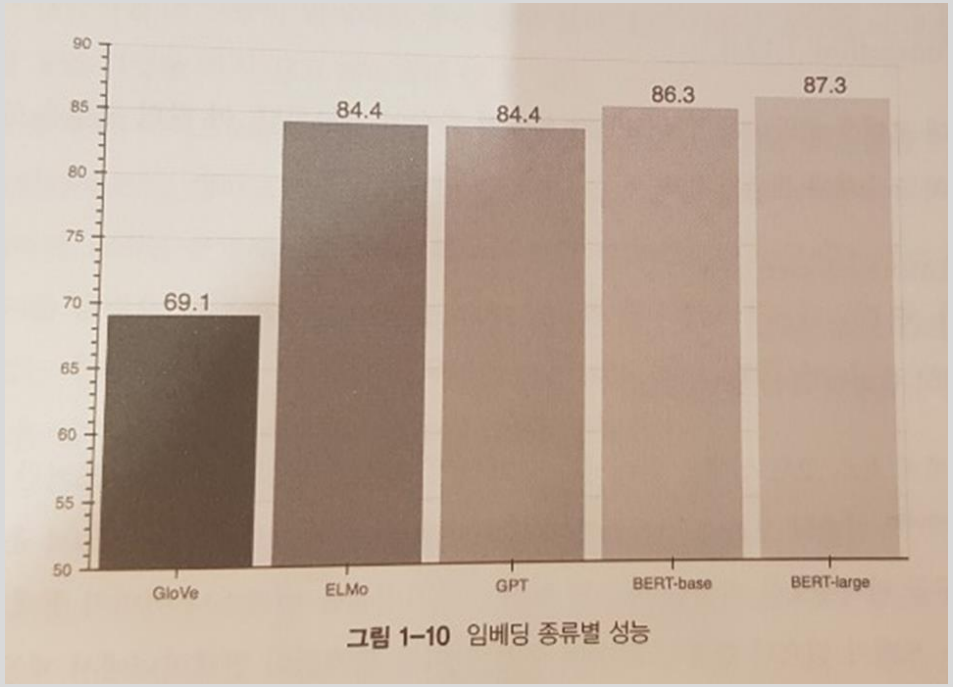


그림 1-10 임베딩 종류별 성능

1. 5 이 책이 다루는 데이터와 주요 용어

Corpus : 임베딩 학습이라는 특정한 목적을 가지고 수집한 표본(sample)이다.
Ex) 한국어 위키백과와 네이버 영화리뷰 데이터를 모두 합친 말뭉치

Collection : Corpus에 속한 각각의 집합
Ex) 한국어 위키백괴와 네이버 영화리뷰 각각

Sentence : 이 책이 다루는 데이터의 기본 단위,

Document : 생각이나 감정 정보를 공유하는 문장(sentence) 집합이 문서(Document)이다.

Token : 문장은 여러 개의 토큰으로 구성된다, 문맥에 따라 토큰을 단어, 형태소, 서브워드라고 부르기도 한다.

Tokenize : 문장을 아래처럼 토큰 시퀀스로 분석하는 과정

Ex) mecab : 실수, 인, 초월수, 는, 모두, 무리수, 이, 다, .

Vocabulary :말뭉치에 있는 모든 문서를 문장으로 나누고 여기에 토크나이즈를 실시한 후 중복을 제거한 토큰들의 집합이다.