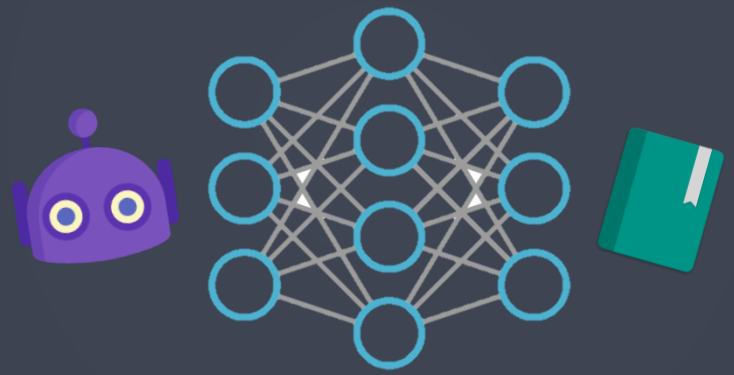
Deep Learning

Chapter 6 언어지능 part1
(Pretrained Language Model)



START

Smart Media
스마트미디어인재개발원

학습목표

- 자연어처리를 위한 PLM(Pretrained Language Model) 종류를 알 수 있다.
- PyTorch를 이용해 모델 구성을 할 수 있다.
- 언어모델과 관련된 오픈소스를 사용하는 환경을 구축할 수 있다.

학습 로드맵

Part 1.

- 퍼셉트론 (Perceptron)
- 다층 퍼셉트론 (Multi Layer Perceptron)
- 오차 역전파 (Backpropagation)

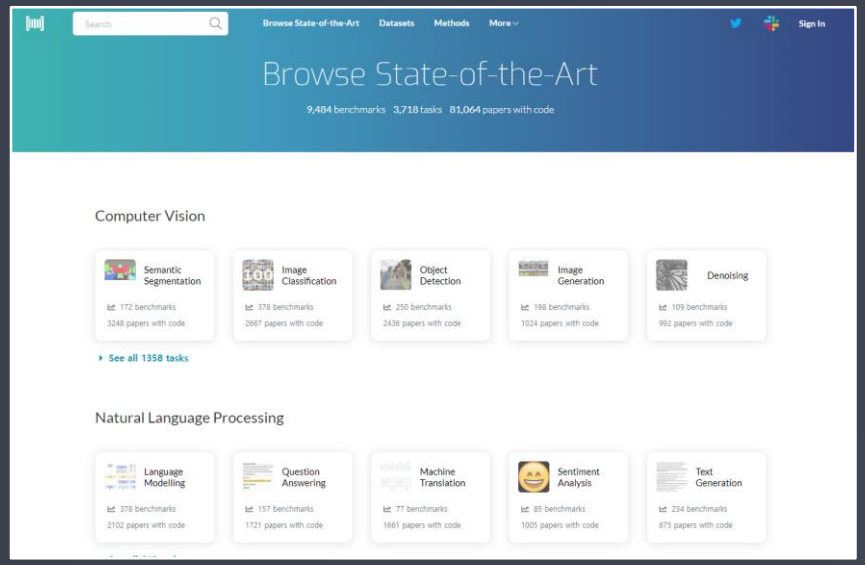
Part 2.

- 합성곱 신경망 (Convolutional Neural Network)
- 순환 신경망 (Recurrent Neural Network)

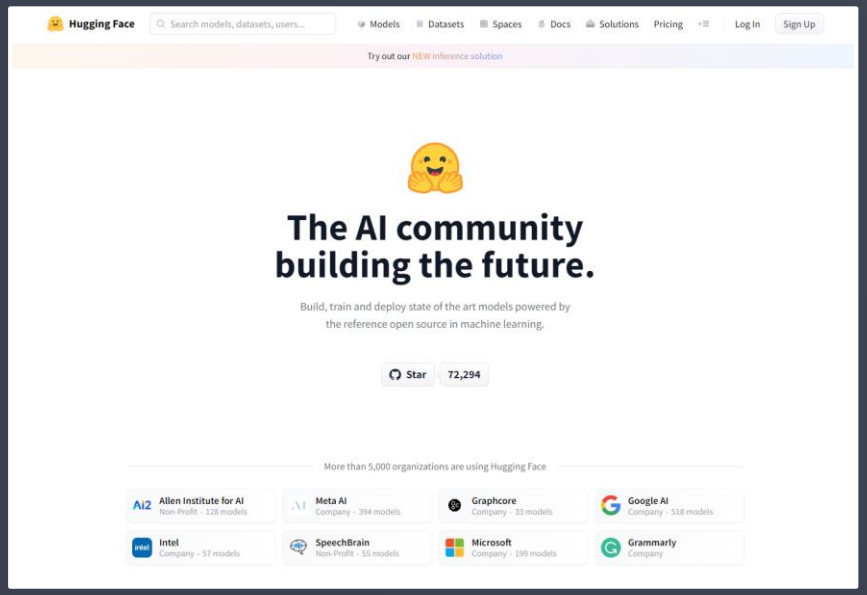
Part 3.

- 이미지 데이터 관련 알고리즘
- 음성 데이터 관련 알고리즘
- **텍스트 데이터 관련 알고리즘**
- 생산적 적대 신경망 (Generative Adversarial Networks)
- 심층 강화 학습 (Deep Reinforcement Learning)

다양한 AI모델의 정보를 볼 수 있는 유용한 사이트



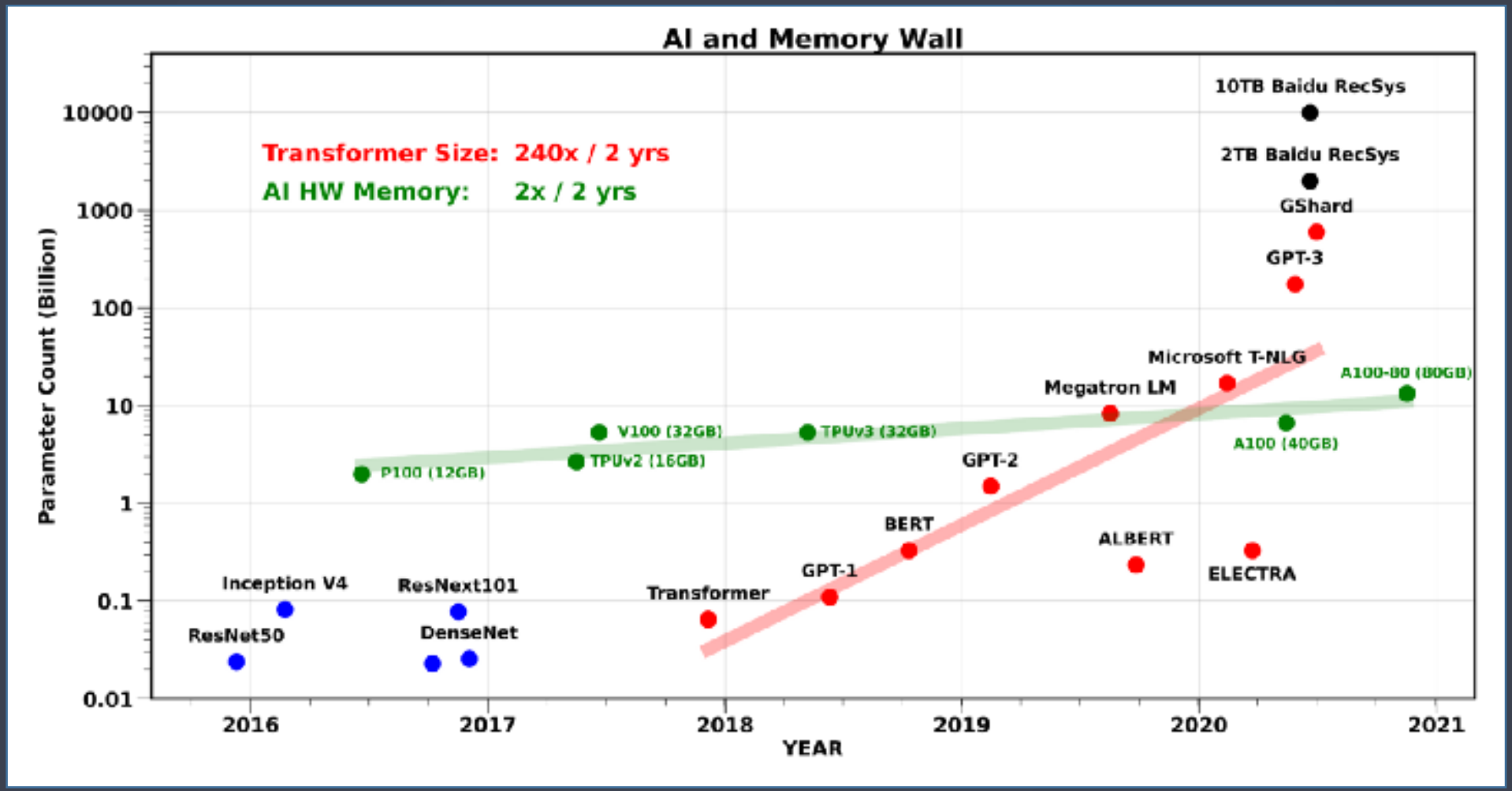
https://paperswithcode.com/



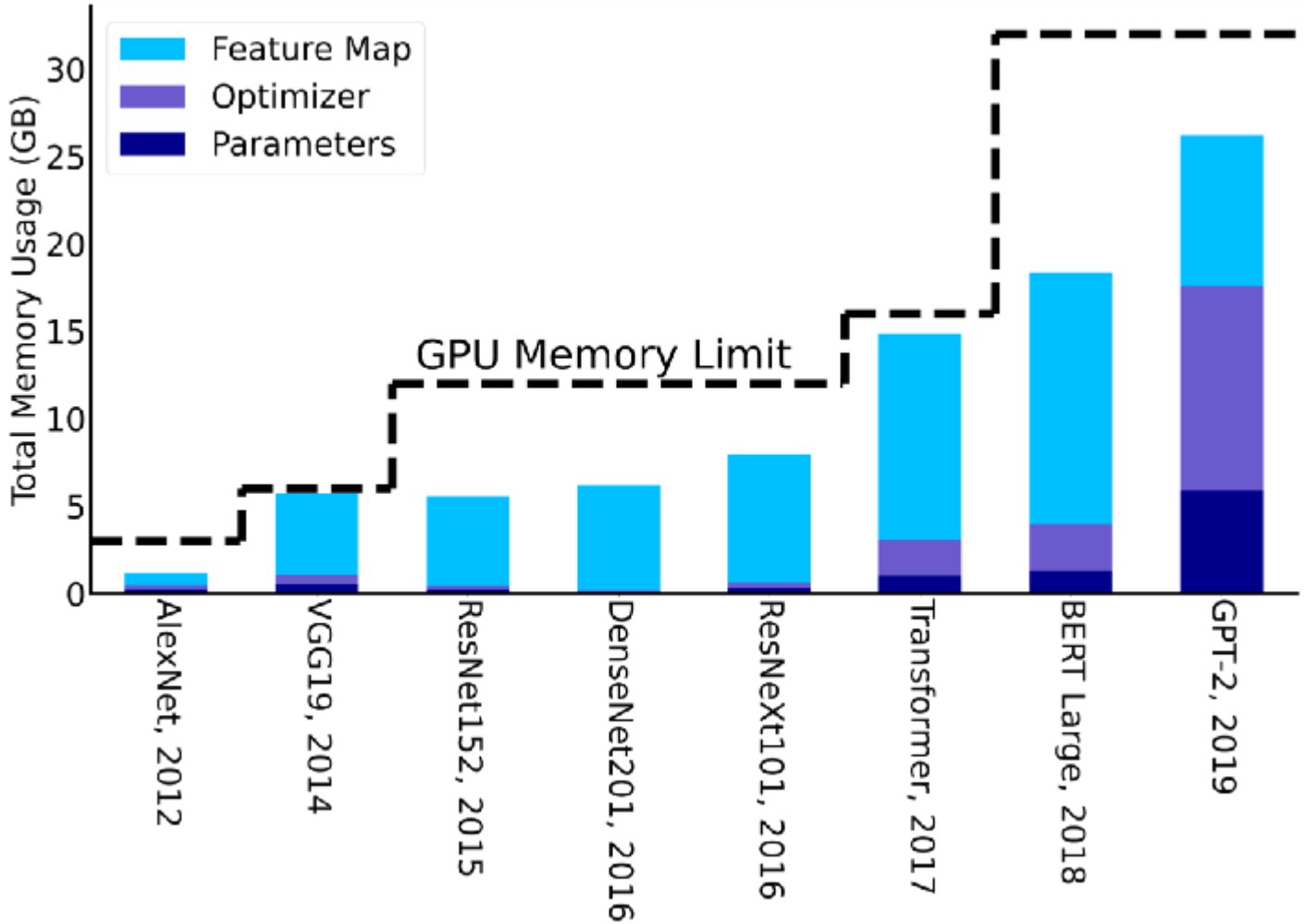
https://huggingface.co/

Smart Media
스마트미디어인재개발원

AI모델과 파라미터의 크기 변화

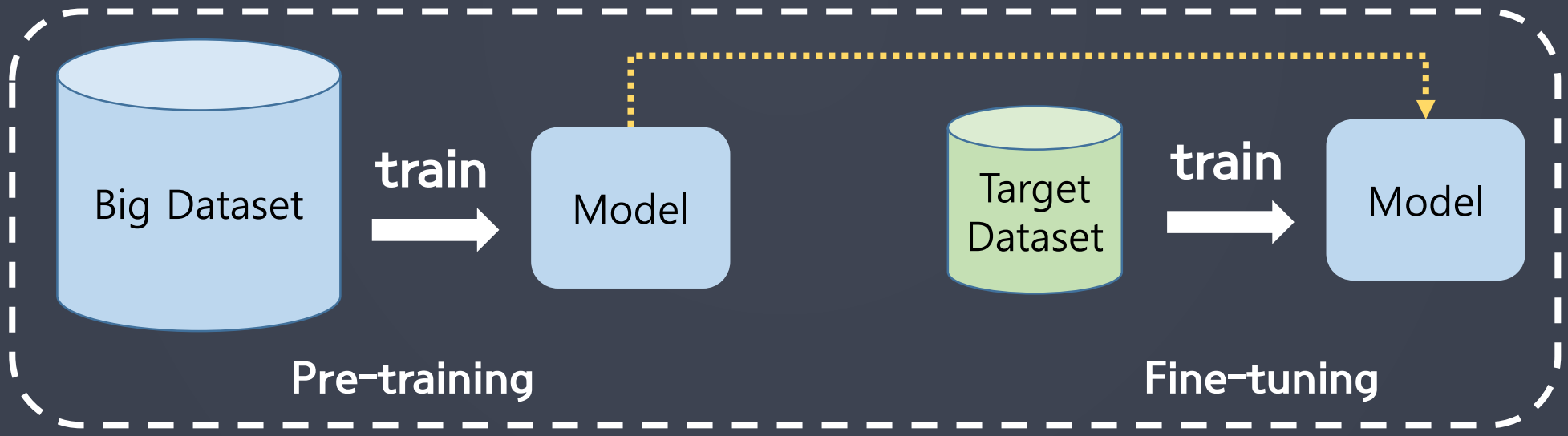


AI모델과 파라미터의 크기 변화



Transfer Learning

- 언어모델의 경우 단어사전의 크기가 크고 품질이 좋을 수 록 성능이 좋아지기 때문에 많은 데이터가 필요하다. 하지만 직접 데이터를 수집하는 방법은 한계가 있다.
- 언어모델도 사전학습된 모델을 fine-tuning하여 사용하는 전이학습 방법이 좋은 성능을 기대해 볼 수 있어 많은 관심을 받고 있다.



PLM을 이용한 성능 향상 방법

- Feature-based Approach
  - 더 좋은 입력 representation을 갖게 하여 성능을 개선하는 방법
  - Word Embedding

Frequency Based	Prediction Based
BOW, TF-IDF	Word2Vec, GloVe, FastText

- Fine-tuning Approach
  - 더 좋은 weight parameter를 갖게 하여 성능을 개선하는 방법
  - ELMo, GPT, BERT

word embedding과 word2vec

- Word embedding
  - 사람이 사용하는 자연어를 컴퓨터가 이해 할 수 있는 숫자의 나열인 형태인 밀집벡터(dense vector)로 변환하는 방법
  - 워드 임베딩을 통해 나온 밀집벡터가 임베딩 벡터(Embedding vector)
  - 임베딩 벡터를 구성하기 위한 모델 -> word2vec, GloVe, fasttext

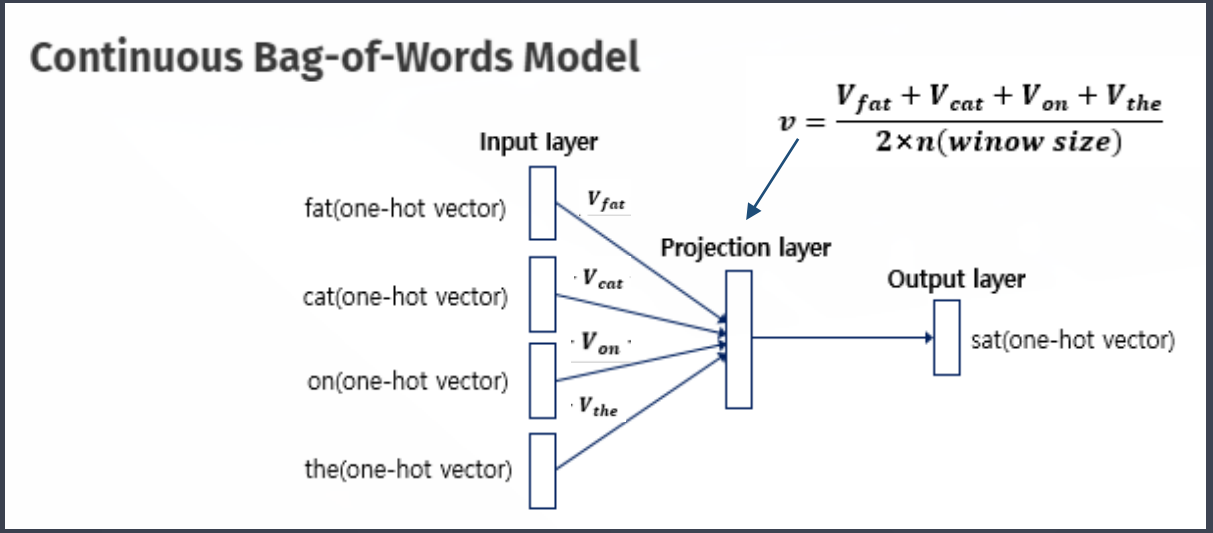
- Word2vec
  - 2013년에 google의 researcher 팀이 논문으로 발표
  - Continuous Bag-of-Words Model과 Continuous Skip-gram Model 2가지 형태를 제시
  - 입력벡터의 가중합을 평균내는 CBOW에 비해 skip-gram 성능이 더 좋다
  - 대표적인 Self supervised learning 방법 중 하나

word2vec 모델 형태 비교

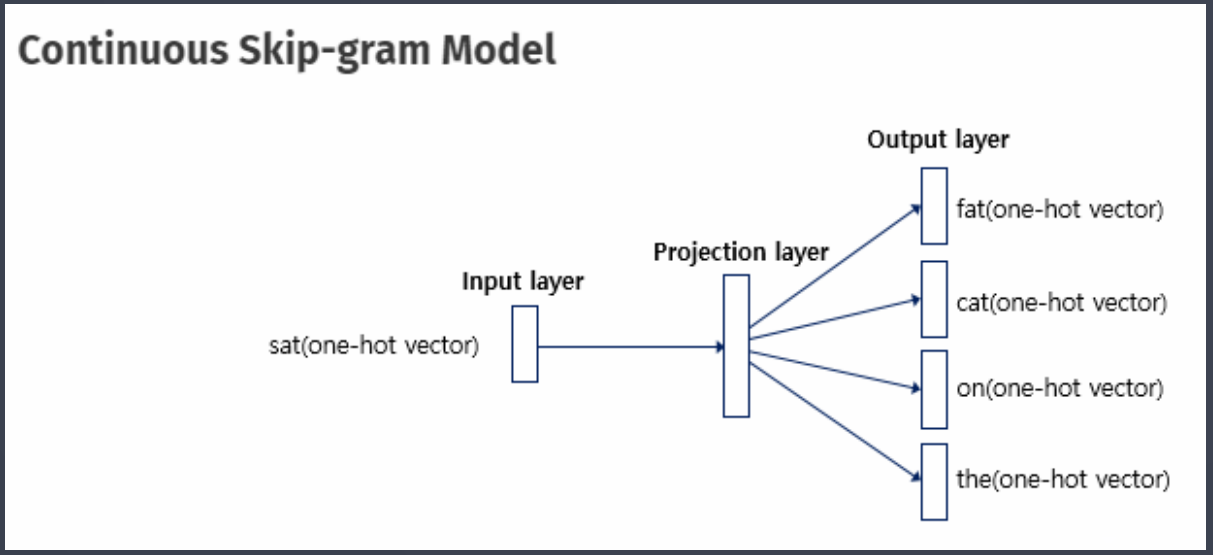
Skip - Gram	CBOW (continuous Bag Of Word)
중간에있는 단어 (타깃단어) 를 가지고 주변에있는 단어 (문맥단어) 를 예측하는 방법	주변에있는 단어 (문맥단어) 를 가지고 중간에있는 단어 (타깃단어) 를 예측하는 방법
개울가 에서 ____ 빨래 ____ 하는	개울가 에서 속옷 _____를 하는

word2vec 모델 형태 비교

fat cat sat on the mat



CBOW
(continuous Bag-of-Words)



Skip-gram

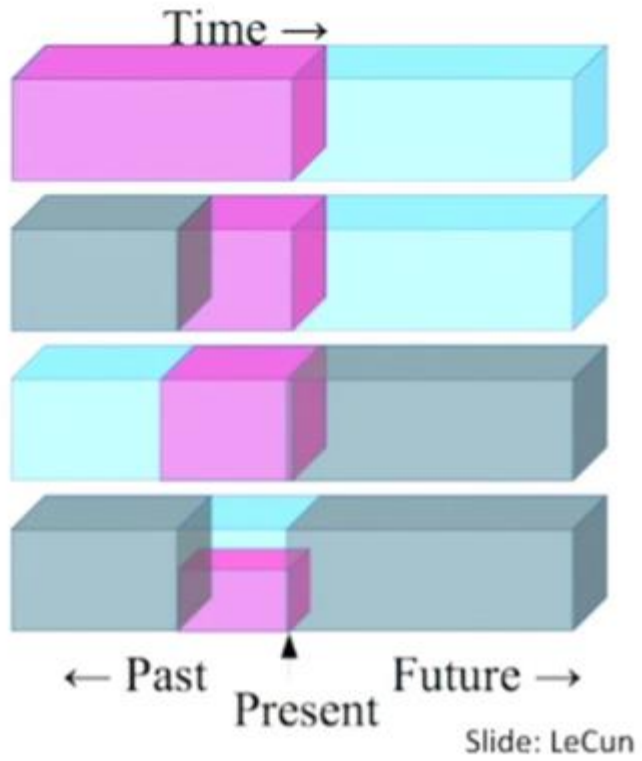
self-supervised learning

- 자기지도학습
  - 지도학습을 위해 data를 labeling하는 작업은 비용이 많이 들어간다
  - unlabeled dataset으로부터 좋은 representation을 얻고자하는 방법
  - input 데이터에서 target으로 쓰일만 한 것을 스스로 정해서 학습
  - 비지도학습과 달리 정답을 스스로 정하기때문에 일반적인 평가가 가능
  - Self-prediction과 Contrastive learning으로 구분 할 수 있다

Self-prediction	Contrastive learning
개별 샘플 데이터의 일부를 이용해 나머지를 예측	개별 데이터 샘플들 사이의 관계를 이용한 예측

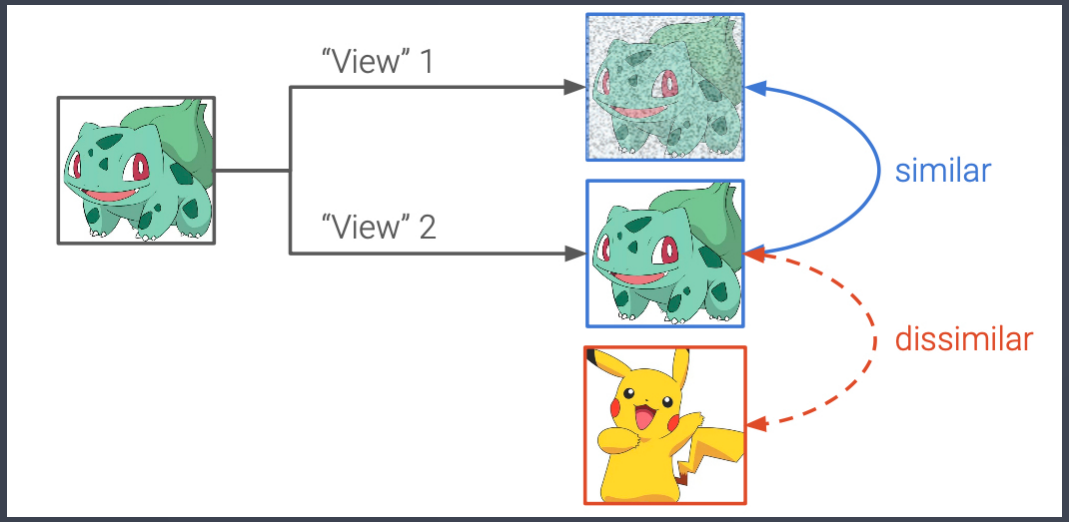
self-supervised learning

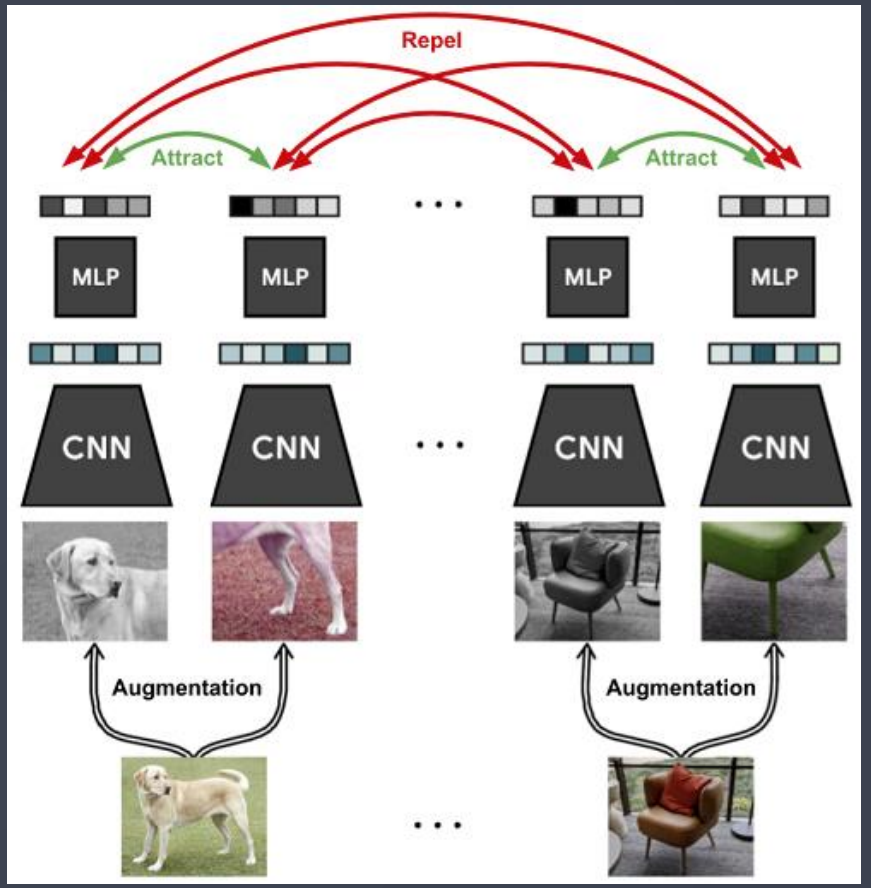
- Predict any part of the input from any other part.
- Predict the future from the past.
- Predict the future from the recent past.
- Predict the past from the present.
- Predict the top from the bottom.
- Predict the occluded from the visible
- Pretend there is a part of the input you don't know and predict that.



Self-prediction

self-supervised learning





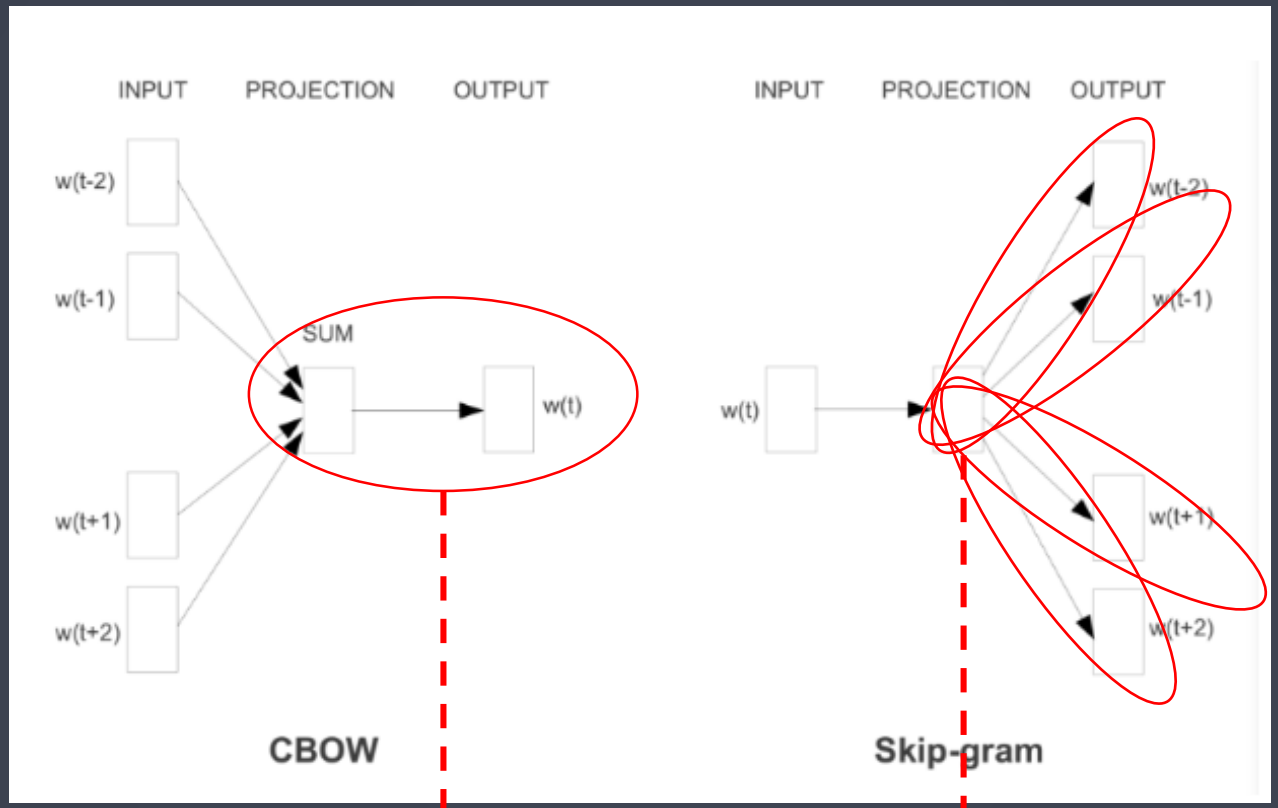
Contrastive learning

self-supervised learning - word2vec 데이터 구축 task

Window Size	Text	Skip-grams
2	[The **wide** road shimmered] in the hot sun.	wide, the wide, road wide, shimmered
	The [wide road **shimmered** in the] hot sun.	shimmered, wide shimmered, road shimmered, in shimmered, the
	The wide road shimmered in [the hot **sun**].	sun, the sun, hot
3	[The **wide** road shimmered in] the hot sun.	wide, the wide, road wide, shimmered wide, in
	[The wide road **shimmered** in the hot] sun.	shimmered, the shimmered, wide shimmered, road shimmered, in shimmered, the shimmered, hot
	The wide road shimmered [in the hot **sun**].	sun, in sun, the sun, hot

Skip-gram 방식을 이용한 데이터 샘플 구축 task

self-supervised learning - word2vec 데이터 구축 task



한쌍

문맥단어4개, 타깃단어

4쌍

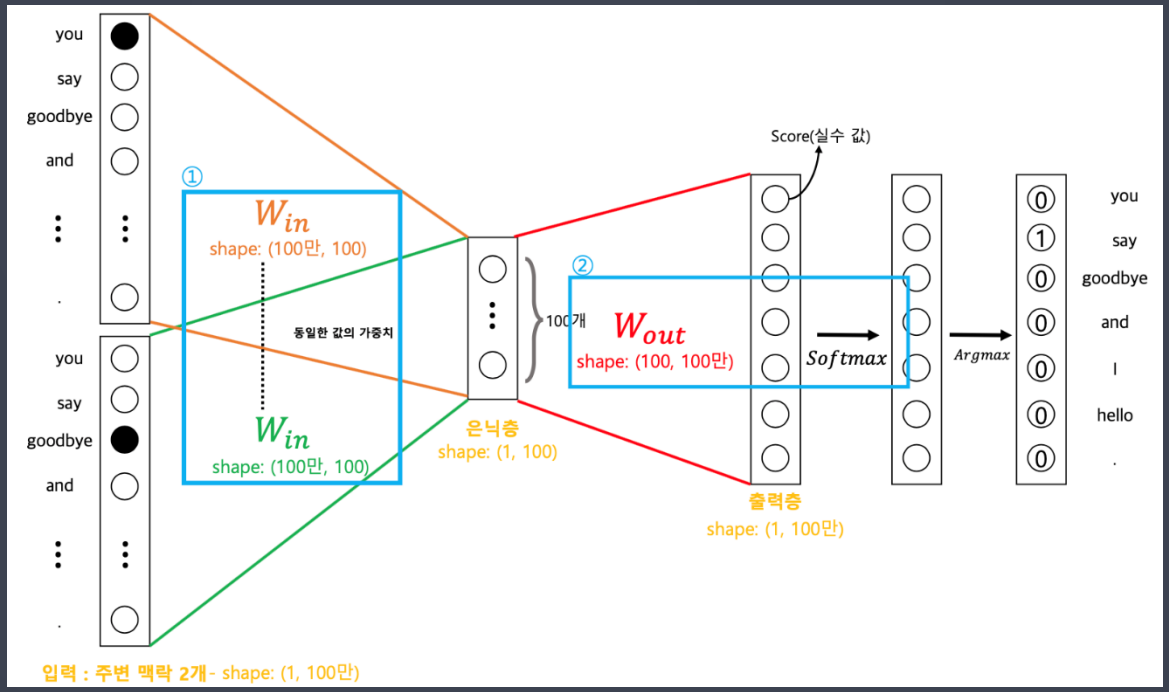
타깃단어, 타깃직전두번째단어
타깃단어, 타깃직전단어
타깃단어, 타깃다음단어
타깃단어, 타깃다음 두번째 단어

- CBOW 경우 입출력 학습데이터 쌍이 하나인 반면 Skip-gram 4개 쌍을 가지고 있다
- Skip-gram 이 같은 말뭉치로도 더 많은 학습데이터를 확보할 수 있어 임베딩 품질이 좋은 경향이 있다.

임베딩 레이어를 이용한 임베딩 벡터 생성(skip-gram)

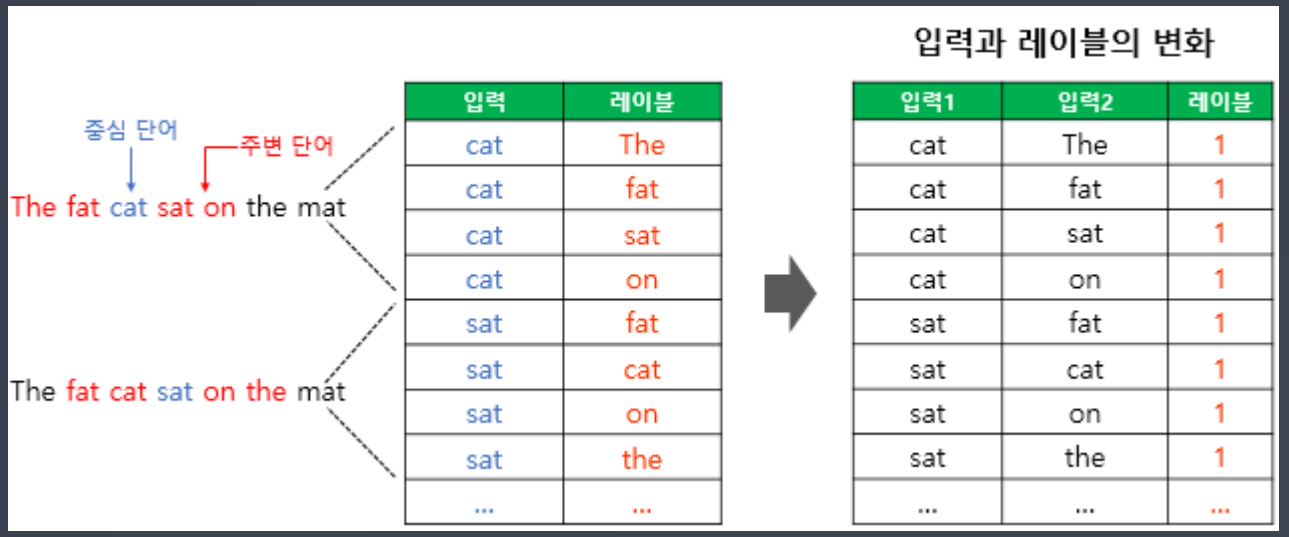
네거티브 샘플링(Negative Sampling)

- 단어가 적을 때는 상관없지만 단어가 100만 개 정도로 많아진다면 신경망 학습하는데 시간이 너무 많이 걸린다
- 특히 출력층의 softmax를 통해 나온 모든 단어의 확률을 계산하여 업데이트하는 비효율적인 연산이 진행된다
- 모든 단어 대상으로 확률을 구하지 않고 일부 단어만 사용하여 계산을 하도록 다중분류 문제에서 이진분류 문제로 변경



Smart Media
스마트미디어인재개발원

네거티브 샘플링(Negative Sampling)



입력	레이블
cat	The
cat	fat
cat	sat
cat	on
sat	fat
sat	cat
sat	on
sat	the
...	...

입력과 레이블의 변화

입력1	입력2	레이블
cat	The	1
cat	fat	1
cat	sat	1
cat	on	1
sat	fat	1
sat	cat	1
sat	on	1
sat	the	1
...	...	...

- 중심단어가 들어가면 말뭉치 단어 중에 어떤 단어가 등장할지 확률을 구하는 문제에서
- 중심단어와 비교단어를 넣고 주변단어인지 아닌지 확률을 구하는 문제로 변경

입력과 레이블의 변화

입력1	입력2	레이블
cat	The	1
cat	fat	1
cat	sat	1
cat	on	1

Negative Sampling

입력1	입력2	레이블
cat	The	1
cat	fat	1
cat	**pizza**	**0**
cat	**computer**	**0**
cat	sat	1
cat	on	1

단어 집합에서 랜덤으로 선택된 단어들을 레이블 0의 샘플로 추가.

- 주변단어만 레이블을 1로 잡으면 학습이 잘 안되니 주변단어가 아닌 것도 샘플링해서 0으로 레이블링 한다
- 작은 말뭉치는 5~20 큰 말뭉치는 2~5정도 샘플링

네거티브 샘플링(Negative Sampling)

포지티브샘플	네거티브샘플
타깃단어(t)와 실제 등장한 문맥단어(c) 쌍	타깃단어(t)와 주변에 등장하지 않은 단어 쌍 (말뭉치전체에서 랜덤추출)

t	C
빨래	에서
빨래	속옷
빨래	를
빨래	하는

t	C
빨래	책상
빨래	안녕
빨래	자동차
빨래	숫자

t	C
빨래	커피
빨래	떡
빨래	사과
빨래	노트북

개울가 (에서 속옷 빨래 를 하는) 남녀

실습

gensim라이브러리 활용 word2vec 활용

GloVe(Global Vectors for Word Representation)

- Word embedding
  - 카운트 기반과 예측 기반을 모두 사용하는 방법론으로 2014년에 미국 스탠포드대학에서 개발한 단어 임베딩 방법론
  - 카운트 기반의 LSA(Latent Semantic Analysis)는 전체적인 통계정보는 고려하지만 비슷한 의미의 단어 관계는 유추하지 못한다
  - word2vec는 예측기반으로 비슷한 의미의 단어 관계 유추는 뛰어나지만 윈도우 사이즈의 단어끼리만 학습하기 때문에 전체적인 통계정보는 고려하지 못한다
  - 말뭉치의 전체 통계정보와 단어 관계 유추의 장점을 모두 고려하도록 구성

GloVe(Global Vectors for Word Representation)

- 윈도우 기반 동시 등장 행렬(Window based Co-occurrence Matrix)
  - 전체 단어 집합의 단어들로 구성하고, i 단어의 윈도우 크기(Window Size) 내에서 k 단어가 등장한 횟수를 i행 k열에 기재한 행렬

k

i

카운트	I	like	enjoy	deep	learning	NLP	flying
I	0	2	1	0	0	0	0
like	2	0	0	1	0	1	0
enjoy	1	0	0	0	0	0	1
deep	0	1	0	0	1	0	0
learning	0	0	0	1	0	0	0
NLP	0	1	0	0	0	0	0
flying	0	0	1	0	0	0	0

- I like deep learning
- I like NLP
- I enjoy flying

말뭉치(corpus)

window size = 1

GloVe(Global Vectors for Word Representation)

- 동시 등장 확률(Co-occurrence Probability)
  - 동시 등장 확률 P(k | i)는 동시 등장 행렬로부터 특정 단어 i가 등장했을 때 등장한 나머지 단어의 전체 횟수를 카운트하고, 특정 단어 i가 등장했을 때 어떤 단어 k가 등장한 횟수를 카운트하여 계산한 조건부 확률

동시 등장 확률과 크기 관계 비(ratio)	k=solid	k=gas	k=water	k=fasion
P(k I ice)	0.00019	0.000066	0.003	0.000017
P(k I steam)	0.000022	0.00078	0.0022	0.000018
P(k I ice) / P(k I steam)	8.9	0.085	1.36	0.96

  - ice가 등장했을 때 solid가 등장할 확률이 steam이 등장했을 때 solid가 등장할 확률보다 높다
  - solid라는 단어는 steam과 등장할 확률 대비 ice와 등장할 확률이 8.9배 높다
  - 임베딩 된 중심 단어와 주변 단어 벡터의 내적이 전체 코퍼스에서의 동시 등장 확률이 되도록 만드는 것이 GloVe의 목표

FastText

- FastText
  - 2015년에 페이스북의 AI research 팀에서 개발한 word2vec의 확장판 방법론
  - Word2Vec는 단어를 쪼개질 수 없는 단위로 생각한다면, FastText는 하나의 단어 안에도 여러 단어들이 존재하는 것으로 간주
  - 내부 단어. 즉, 서브워드(subword)를 고려하여 학습하기 때문에 모르는 단어 (Out Of Vocabulary, OOV)에 대해서도 다른 단어와의 유사도를 계산할 수 있고, 빈도가 적은 단어도 서브워드가 존재한다면 임베딩의 정확성을 높일 수 있다

<시나 / 시나브 / 나브로 / 브로> / <시나브로>

n = 3인 경우 시나브로 단어의 서브워드

FastText

- FastText 강점
  - 조사나 어미가 발달한 한국어에 좋은 성능을 낼 수 있다
  - 용언(동사, 형용사) 활용이나 그와 관계된 어미들이 벡터공간상 가깝게 임베딩 된다
  - 임베딩을 문자단위 ngram 벡터의 합으로 표현하기때문에 오타나 미등록단어에도 강건하다
  - '서울특벌시'라는 오타가 나도 서울 어휘가 포함이 되면서 나머지 '울특' '특벌'이 단어라도 서울특별시에대한 임베딩을 추정할 수가 있게 된다

실습

gensim라이브러리 활용 FastText 활용

InveptionV3와 GloVe를 이용한 이미지 캡셔닝