

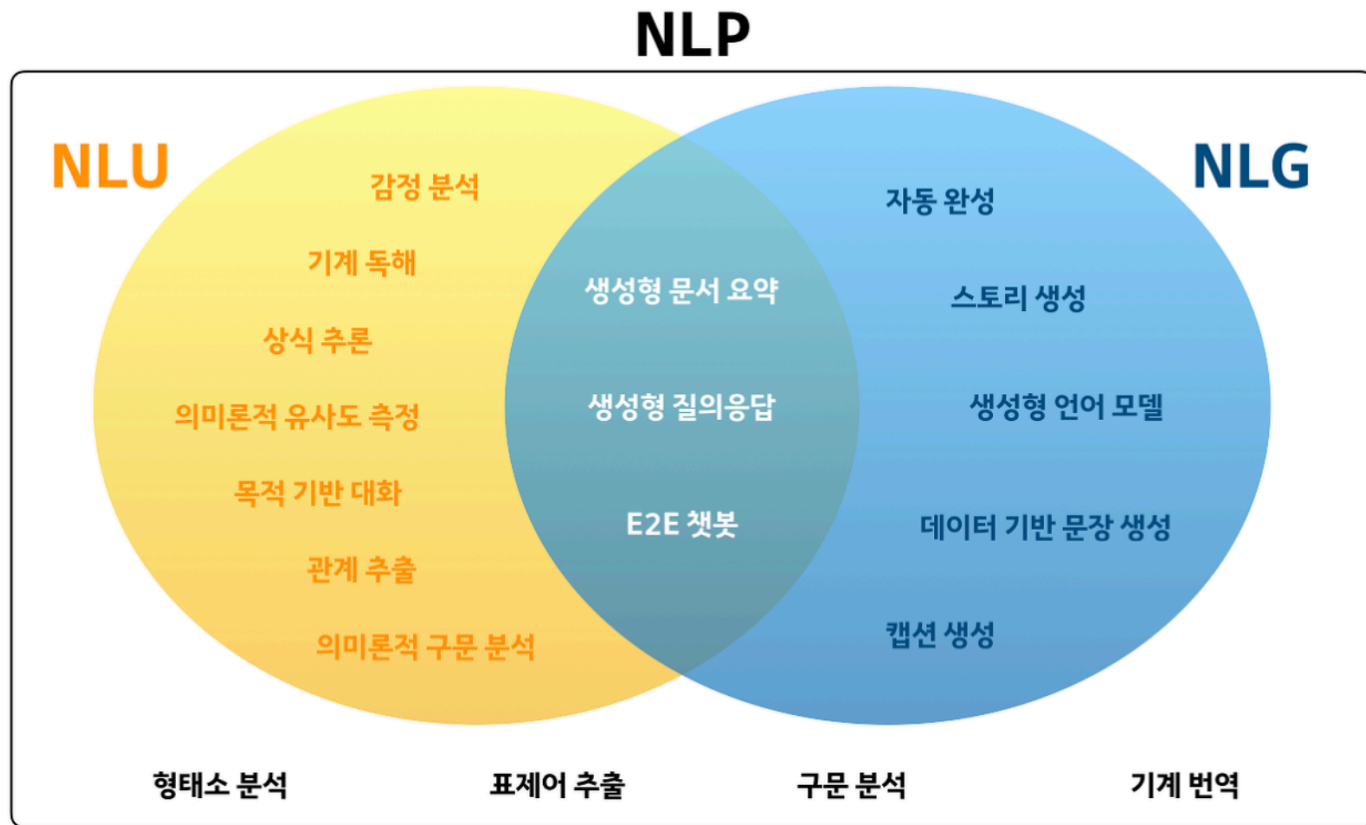
6장. 순환신경망

자연어 처리란?

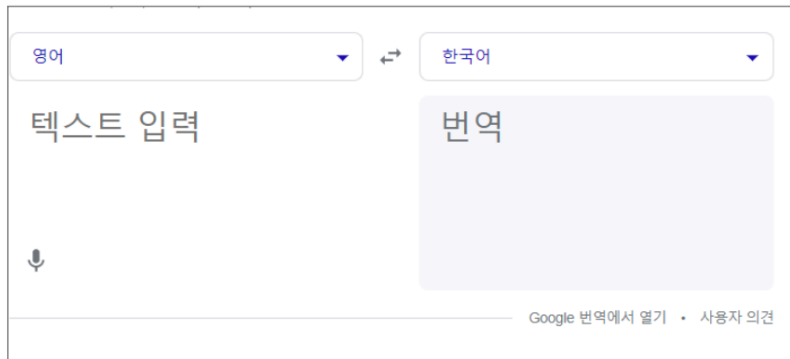
NLP(Natural Language Processing) : 인간의 언어를 컴퓨터에게 가르치는 과정

NLU(Natural Language Understanding) : 자연어의 의미를 모델이 이해하도록 하는 것

NLG(Natural Language Generation) : 자연어를 모델이 생성하도록 하는 것



실제활용



NLP(Natural Language Processing)

컴퓨터가 사람이 일상 생활에 사용하는 언어(자연어)의 의미를 분석하고 처리할 수 있도록 하는 일. (음성 인식, 내용 요약, 번역, 텍스트 분류, 감성 분석 등)

Vector : 인간의 언어를 컴퓨터가 인식할 수 있도록 수치적으로 변환.

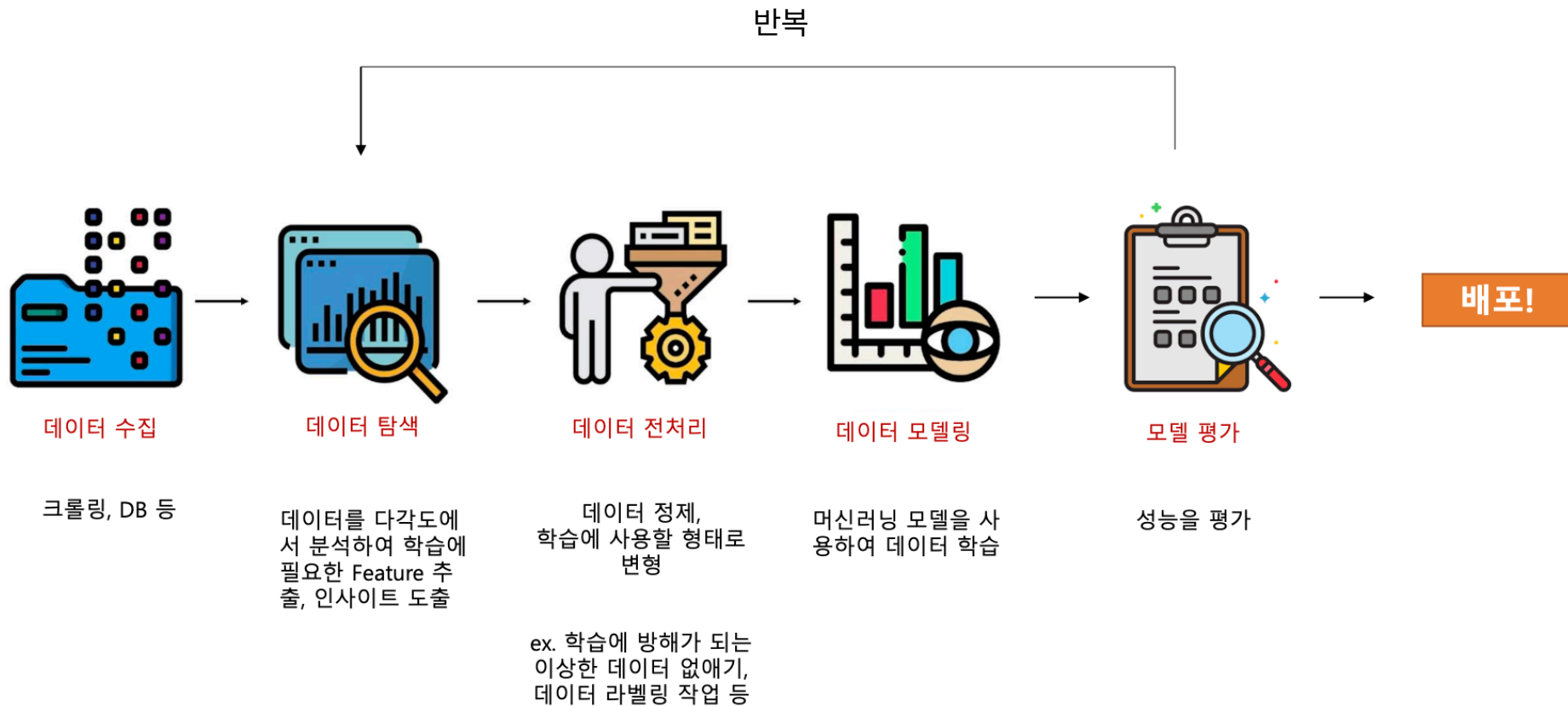
크기, **방향**을 포함하는 개념(N차원 공간 표현)
=> 거리감(얼마나 비슷한지) 수치로 나타내는 것이 가능함.

Corpus(data)

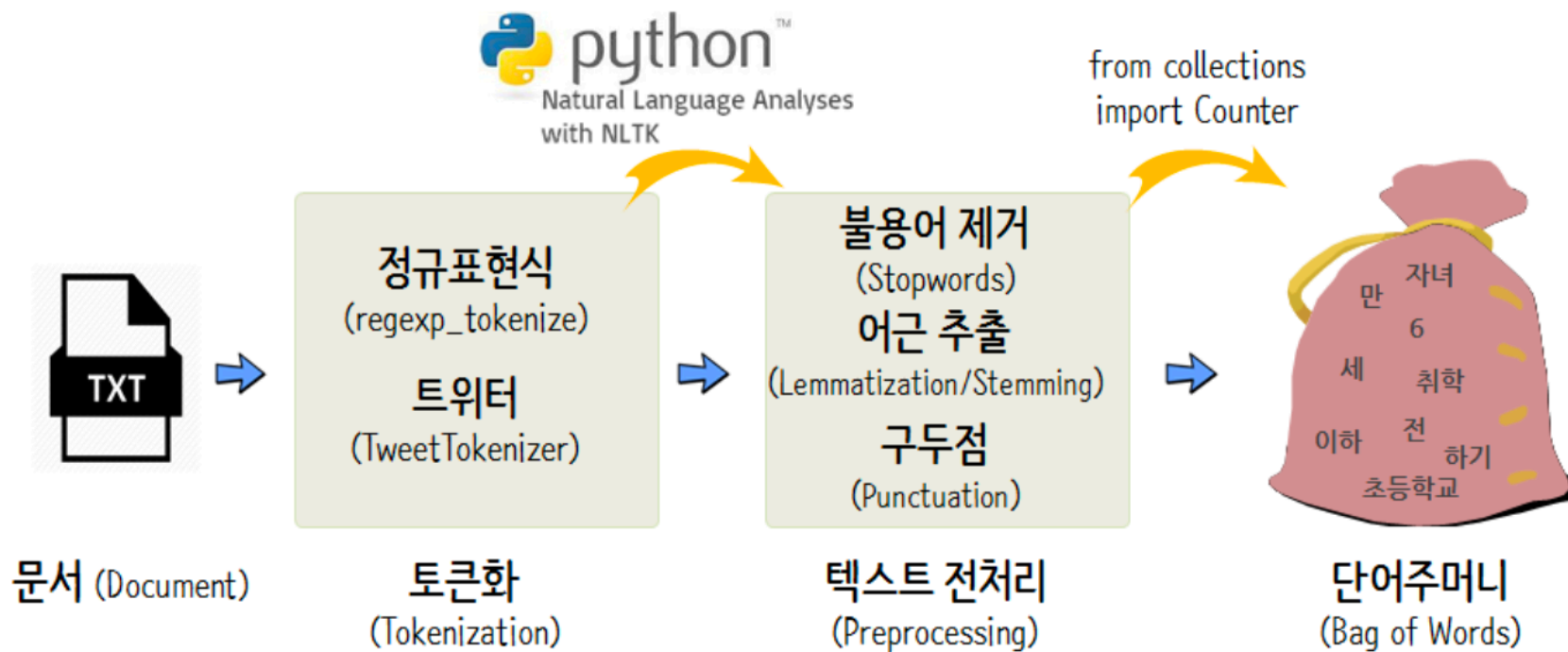
["hello", "world", "natural", "language", "processing"]				
1	2	3	4	5

index

머신러닝 워크플로우



자연어 처리를 위한 전처리



Cleansing 예시:

반복 문자열 대체 (ex. ㅋㅋㅋㅋㅋㅋㅋㅋ)

특수문자 제거 (ex. ^^)

토큰화 - 영어 vs 한국어

영어

"This book is for deep learning learners"

토큰화



띄어쓰기만 분리해도 가능?

한국어

가방에 들어가신다.

가방/NNG

에/JKM

들어가/VV

시/EPH

니다/EFN

어렵다!

- 1) 다양한 조사 존재
(ex. '그가', '그에게', '그를', '그와', '그는' ...)
- 2) 띄어쓰기가 잘 지켜지지 않는다

One-Hot Encoding

	hello	world	processing
"unk": 0	$\begin{bmatrix} 0 \end{bmatrix}$	$\begin{bmatrix} 0 \end{bmatrix}$	$\begin{bmatrix} 0 \end{bmatrix}$
"hello": 1	$\begin{bmatrix} 1 \end{bmatrix}$	$\begin{bmatrix} 0 \end{bmatrix}$	$\begin{bmatrix} 0 \end{bmatrix}$
"world": 2	$\begin{bmatrix} 0 \end{bmatrix}$	$\begin{bmatrix} 1 \end{bmatrix}$	$\begin{bmatrix} 0 \end{bmatrix}$
"natural": 3	$\begin{bmatrix} 0 \end{bmatrix}$	$\begin{bmatrix} 0 \end{bmatrix}$	$\begin{bmatrix} 0 \end{bmatrix}$
"language": 4	$\begin{bmatrix} 0 \end{bmatrix}$	$\begin{bmatrix} 0 \end{bmatrix}$	$\begin{bmatrix} 0 \end{bmatrix}$
"processing": 5	$\begin{bmatrix} 0 \end{bmatrix}$	$\begin{bmatrix} 0 \end{bmatrix}$	$\begin{bmatrix} 1 \end{bmatrix}$

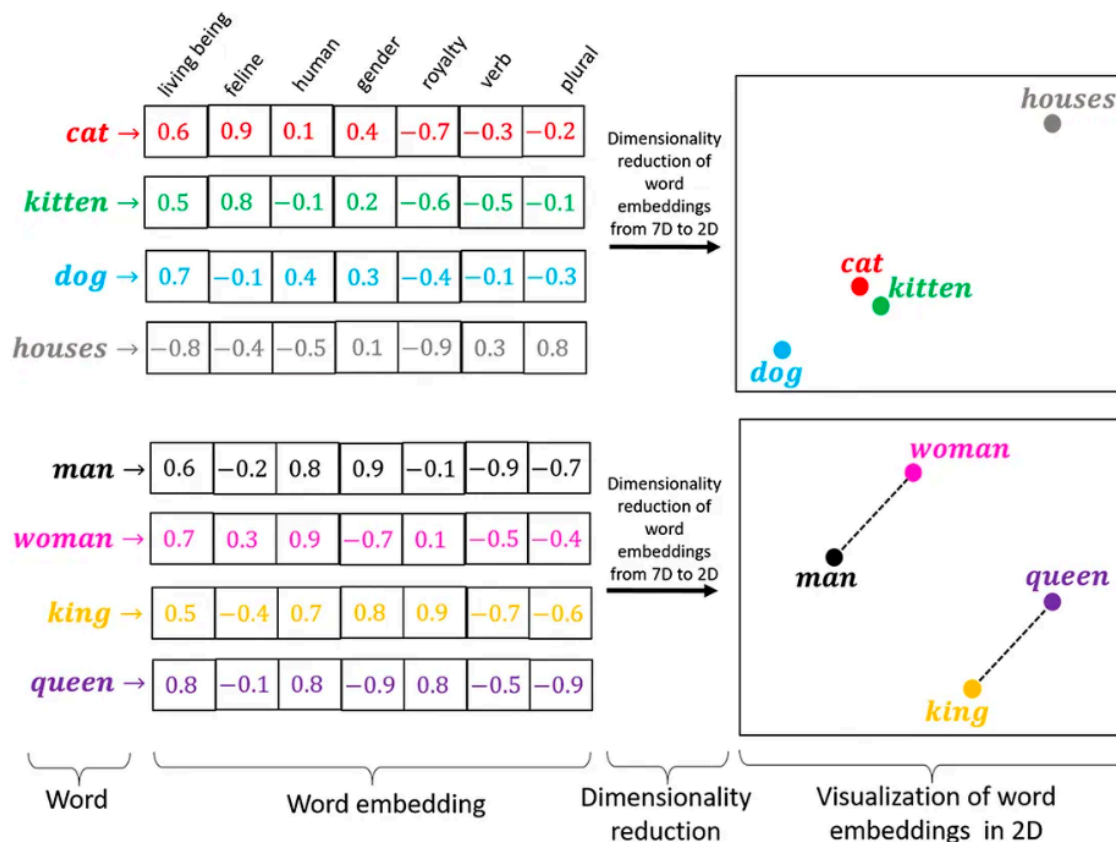
한계점 :

1. 데이터가 커질수록 차원이 늘어남
2. 단어간의 관계(유사도)를 표현할 수 없음.

Word Embedding (Word Vector)

단어를 N차원의 vector로 만든 것.

- Frequency based Embedding: Count Vectors, TF-IDF, Co-Occurrence Matrix
- Prediction based Embedding: Word2Vec, GloVe, FastText




Transformer 계열 모델 (BERT, GPT)

전체 문맥을 입력해서, 빈칸을 예측.


GPT


어제 카페 갔었어 거기 사람 많더라



문장 생성에 강점

BERT

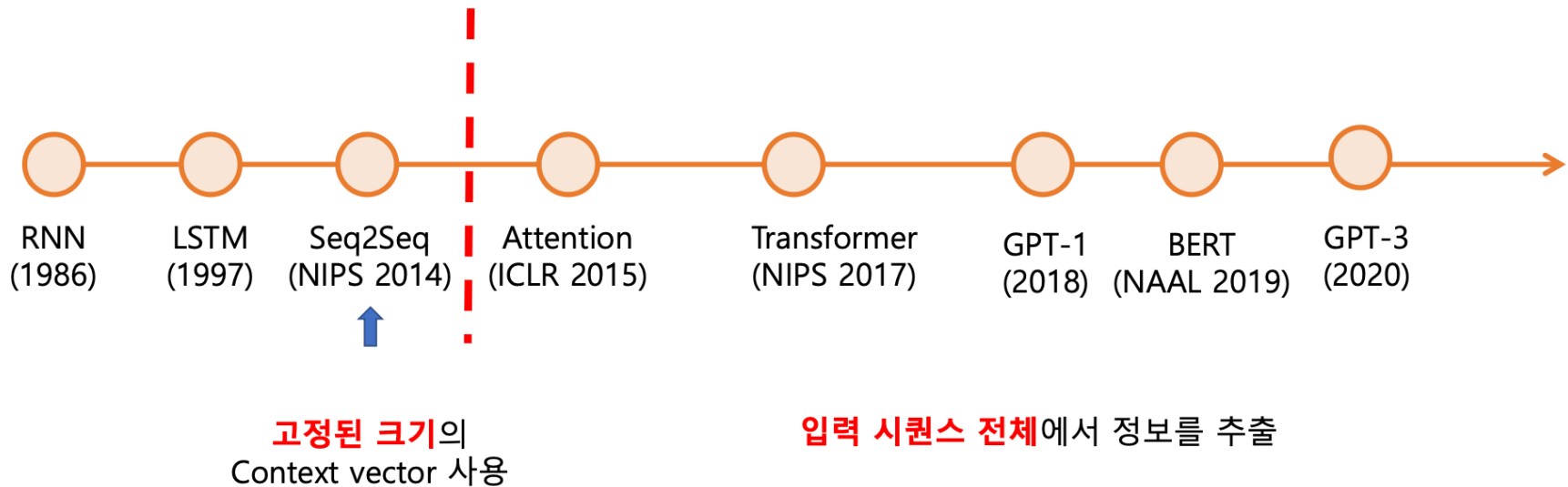
어제 카페 갔었어  사람 많더라



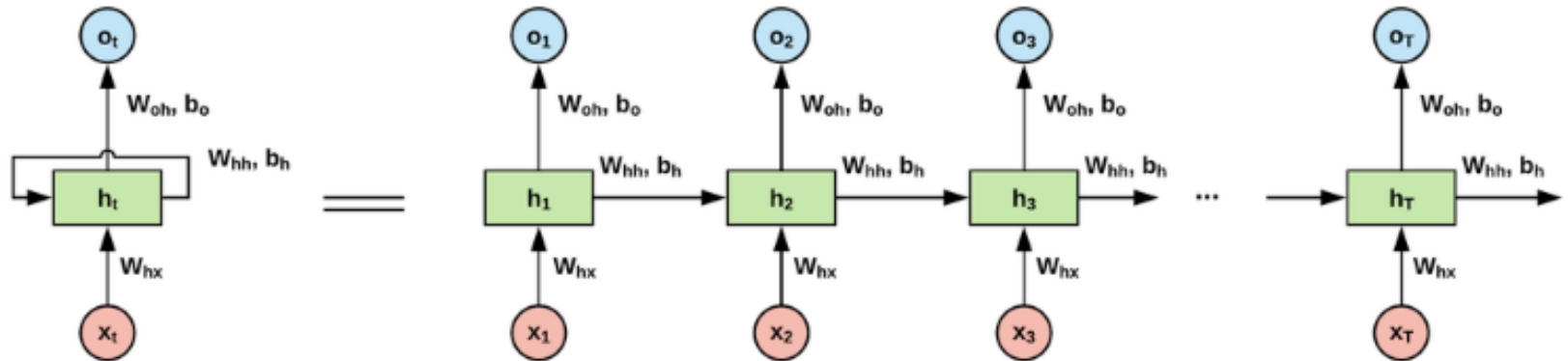
문장(문맥) 이해에 강점

딥러닝 기반의 기계 번역 발전

최신 고성능 모델들은 **Transformer** 아키텍처를 기반으로 함



RNN (Recurrent Neural Network, 순환 신경망)

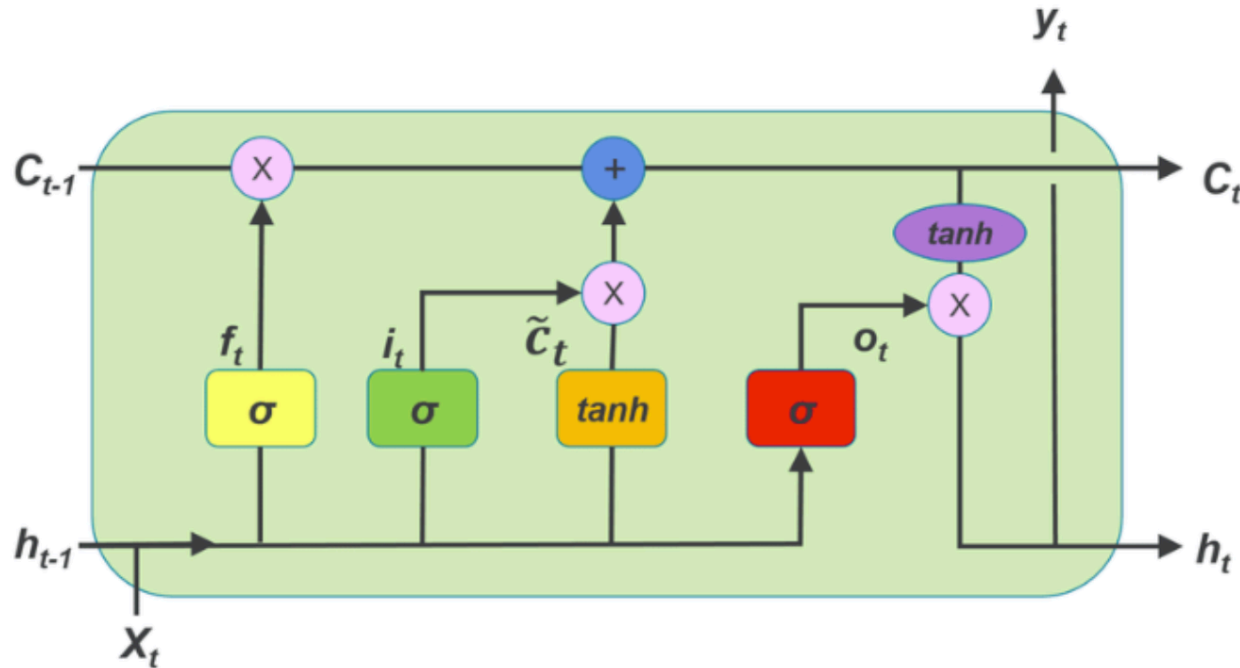


RNN은 시퀀스 데이터를 처리하는 데 특화된 인공 신경망의 한 유형입니다.

이 신경망은 시간적으로 연속된 데이터를 다루기 위해 과거의 정보를 기억하면서 현재의 입력과 함께 처리하는 능력을 가지고 있습니다.

기본적으로 RNN은 내부에 루프를 가진 네트워크 구조로서, 이 루프를 통해 정보가 순환 되면서 이전의 출력이 다시 입력으로 사용됩니다.

LSTM(Long Short-Term Memory, 장단기 메모리)



LSTM은 순환 신경망(RNN)의 한 유형으로, 특히 장기 의존성 문제를 해결하기 위해 고안된 구조입니다. 이 구조는 기본 RNN에 비해 더욱 복잡한 작업을 효과적으로 수행할 수 있으며, 긴 시간 간격에 걸친 데이터의 중요한 특징을 학습할 수 있는 능력이 탁월합니다.

주요 게이트는 Input Gate, Forget Gate, Output Gate로 이루어져 있습니다.