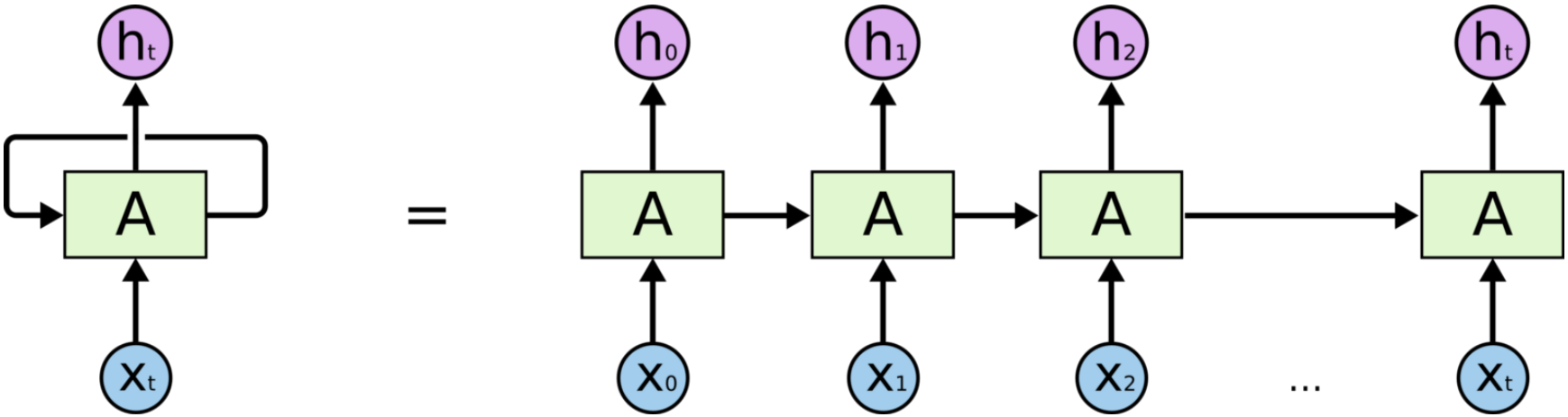
순환 신경망(RNN)

2022.05

1. 개요

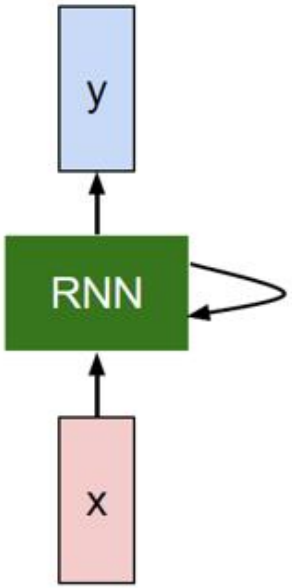
❖ 순환 신경망(RNN: Recurrent Neural Network)

- Sequence Data
  - 문장(Text), 음성 신호, 주가 데이터
- DNN, CNN: 입력층 → 출력층 한 방향으로만 흐르는 Feed forward 신경망
- RNN
  - 시퀀스 데이터의 모델링에 사용되는 신경망

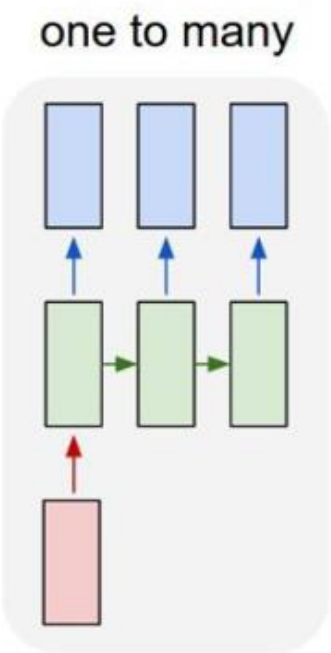


1. 개요

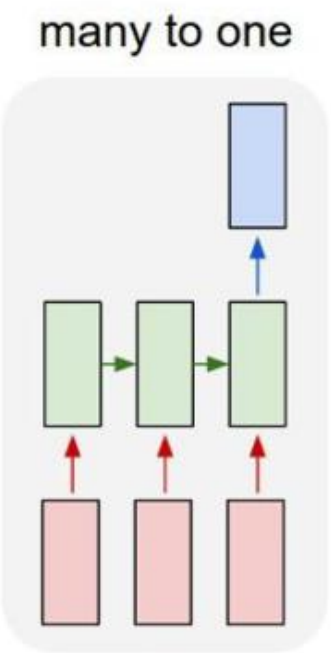
❖ 입력/출력 시퀀스



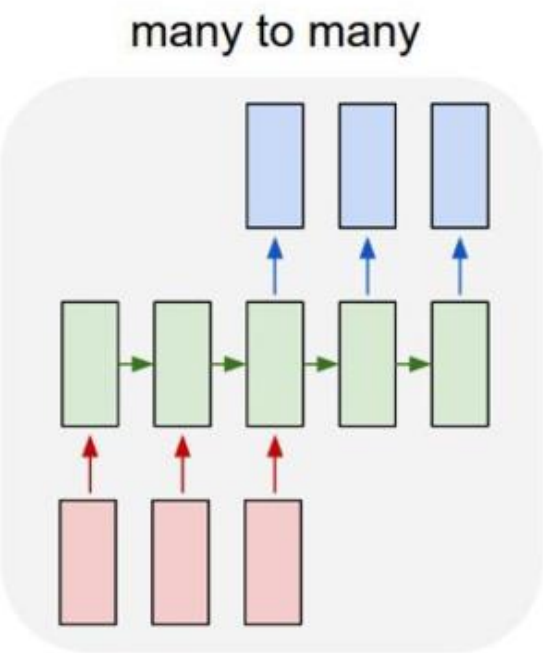
<기본구조>



<사진설명 붙이기>
사진 → 단어들



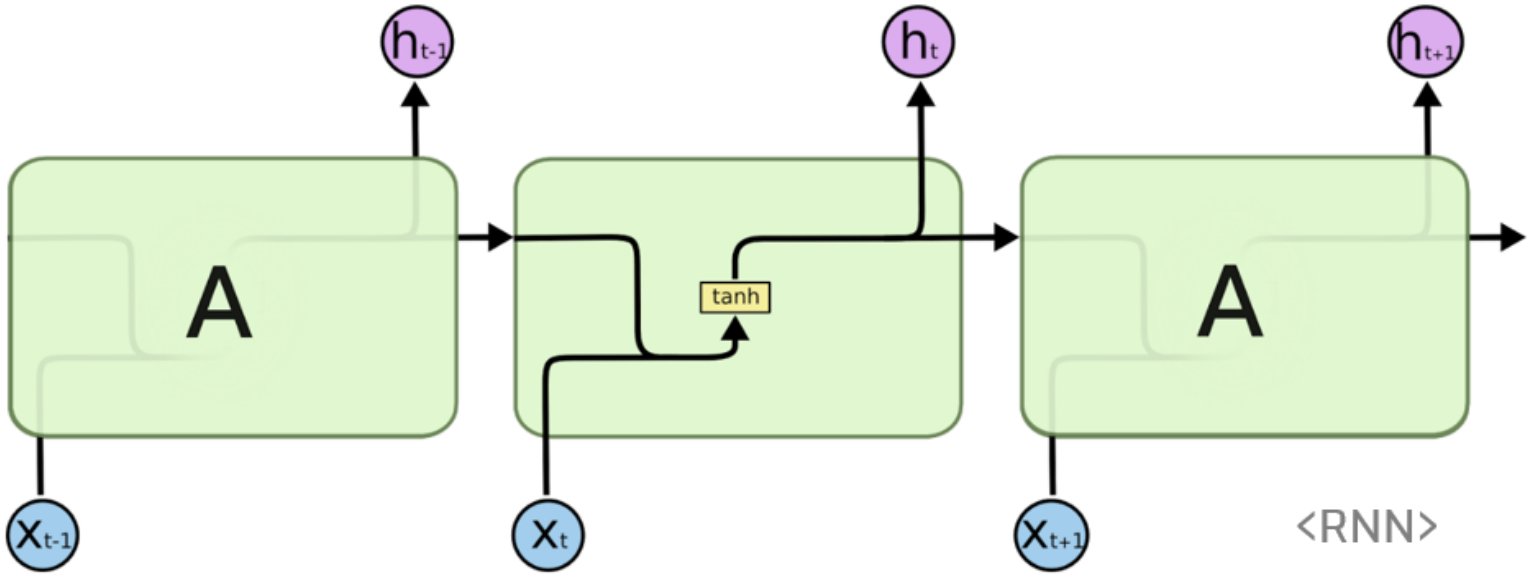
<감성분석>
단어들 → 감성점수



<번역>
단어들 → 단어들

1. 개요

❖ 기본적인 구조(Simple RNN, Vanilla RNN)

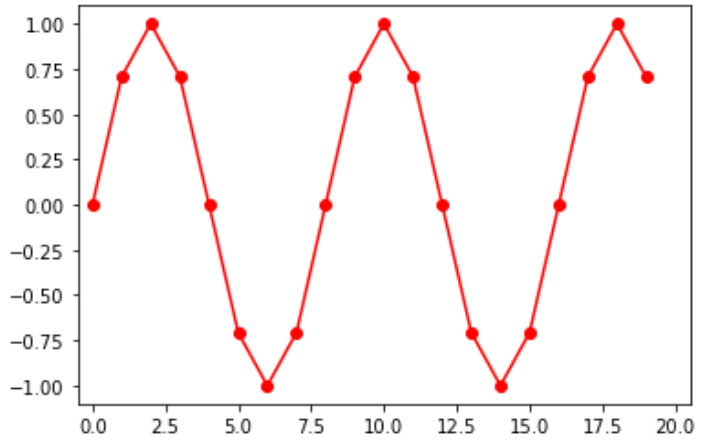


$$\boldsymbol{h_t = \tanh(W_x x_t + W_h h_{t-1} + b)}$$

활성화 함수 가중치 바이어스

1. 개요

❖ 시계열 예측 문제 (Many-to-one 문제)



- 3개의 순서열 입력

  → 출력값이 타겟과 일치하게 학습

- 모델 설계

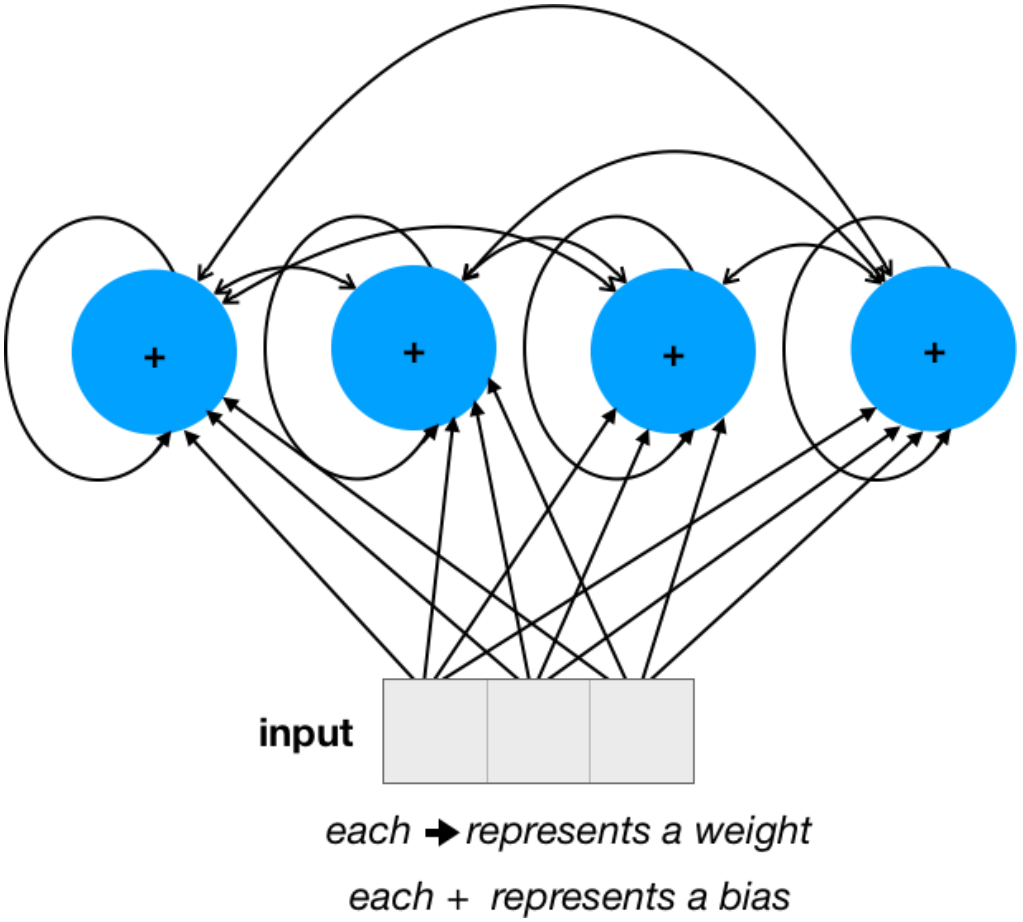
```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
simple_rnn_1 (SimpleRNN)     (None, 10)                120
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 11
=================================================================
Total params: 131
Trainable params: 131
Non-trainable params: 0
_________________________________________________________________
```

1. 개요

❖ 시계열 예측 문제 (Many-to-one 문제)

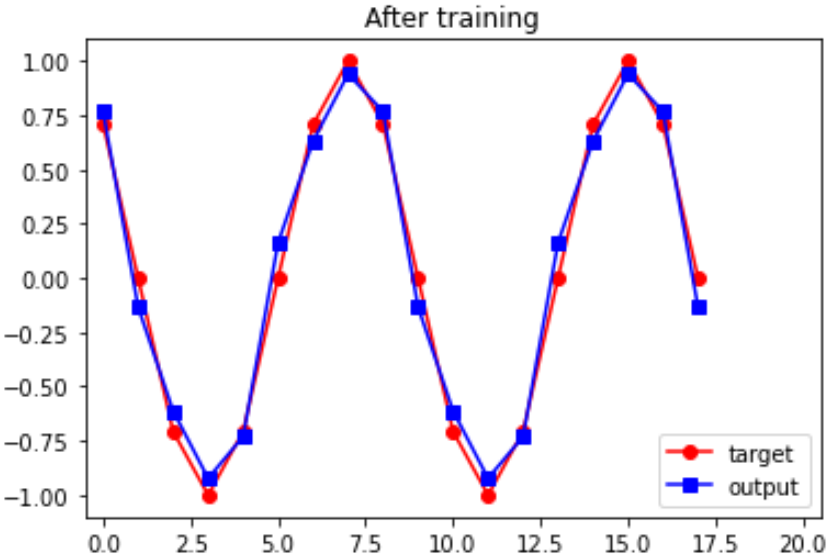
- 파라메터 갯수= 순환 weights + 입력 weights + bias
- 즉, (피쳐 갯수 + 유닛 갯수) * 유닛 갯수 + 유닛 개수

- 좌측 그림에서
  피쳐 갯수(input dim)는 3,
  유닛 개수는 4 이므로
  파라메터 갯수는
  (3 + 4) * 4 + 4 = 32

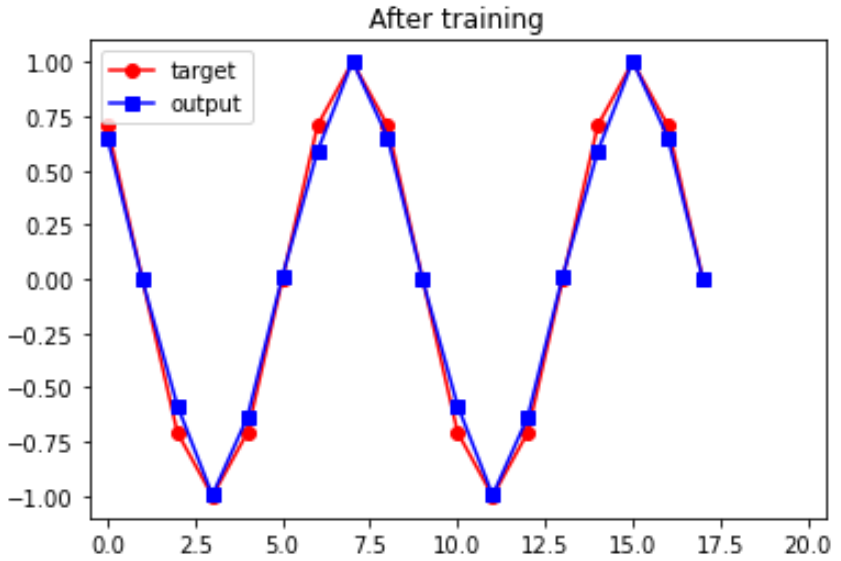


1. 개요

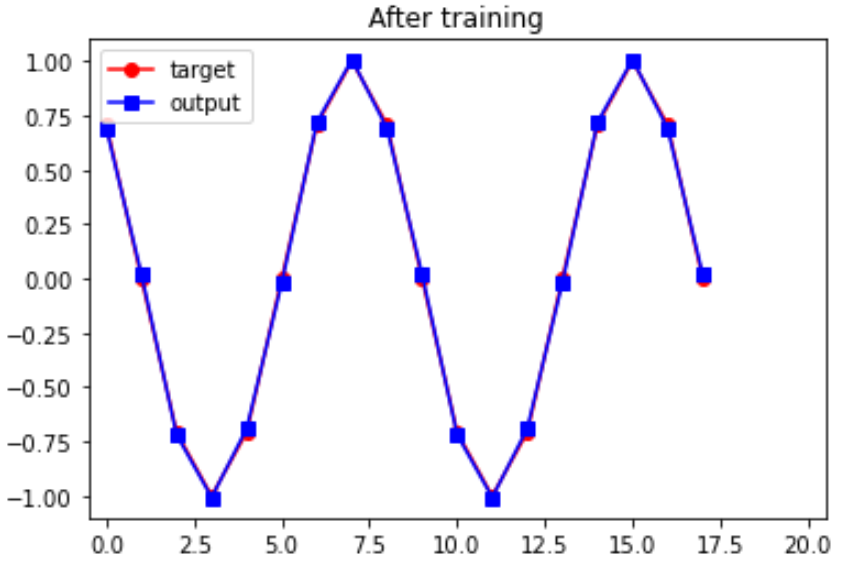
❖ 시계열 예측 문제 (Many-to-one 문제)



RNN 노드 개수 = 10



5



20

1. 개요

❖ RNN을 이용한 텍스트 생성(One-to-many 구조)

- 문맥을 반영해서 텍스트 생성
- 입력 문장

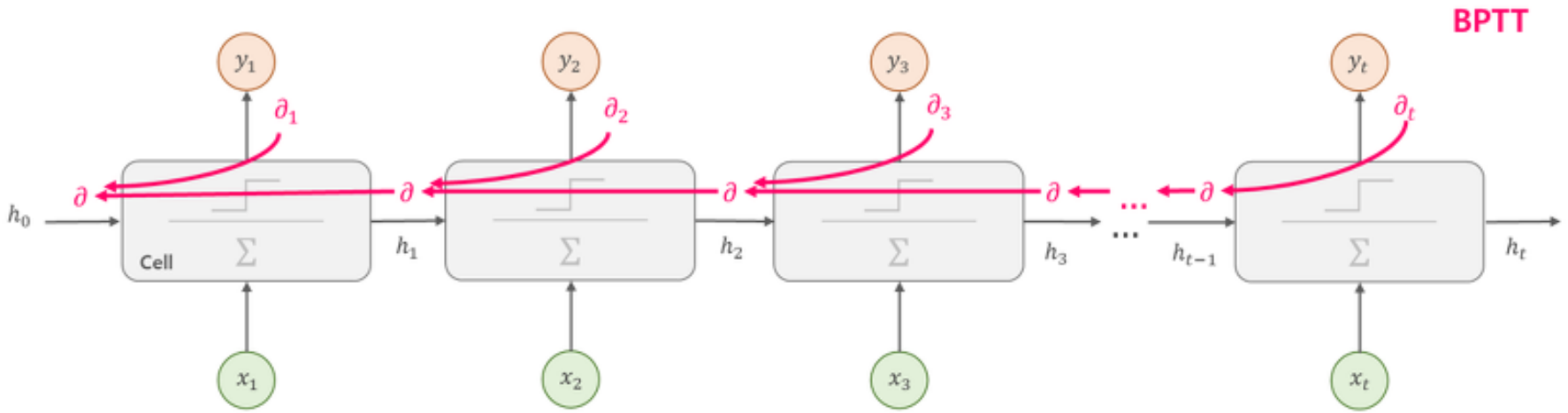
  '경마장에 있는 말이 뛰고 있다'

  '그의 말이 법이다'

  '가는 말이 고와야 오는 말이 곱다'

❖ BPTT(Back Propagation Through Time) 문제점

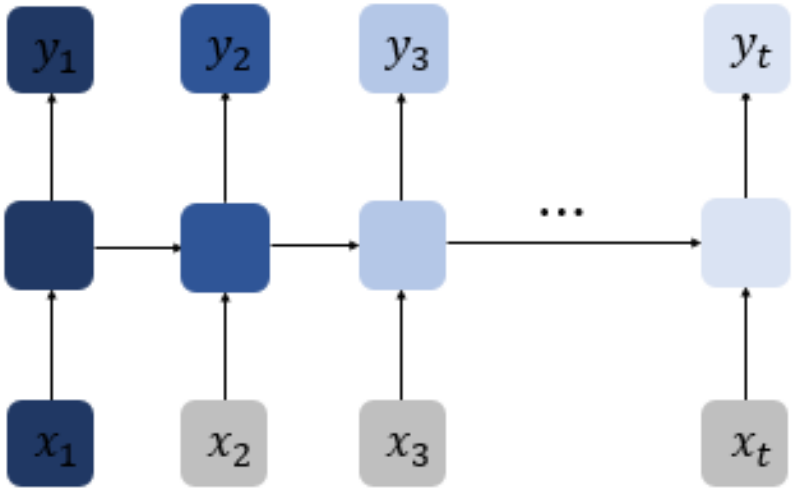
- 그래디언트 소실 및 폭주(vanishing & exploding gradient) 문제가 발생할 가능성



❖ 장단기 메모리(Long Short-Term Memory, LSTM)

- 바닐라 RNN의 한계
  - 비교적 짧은 시퀀스에 대해서만 효과를 보임
  - 시점(time step)이 길어질수록 앞의 정보가 뒤로 충분히 전달되지 못함

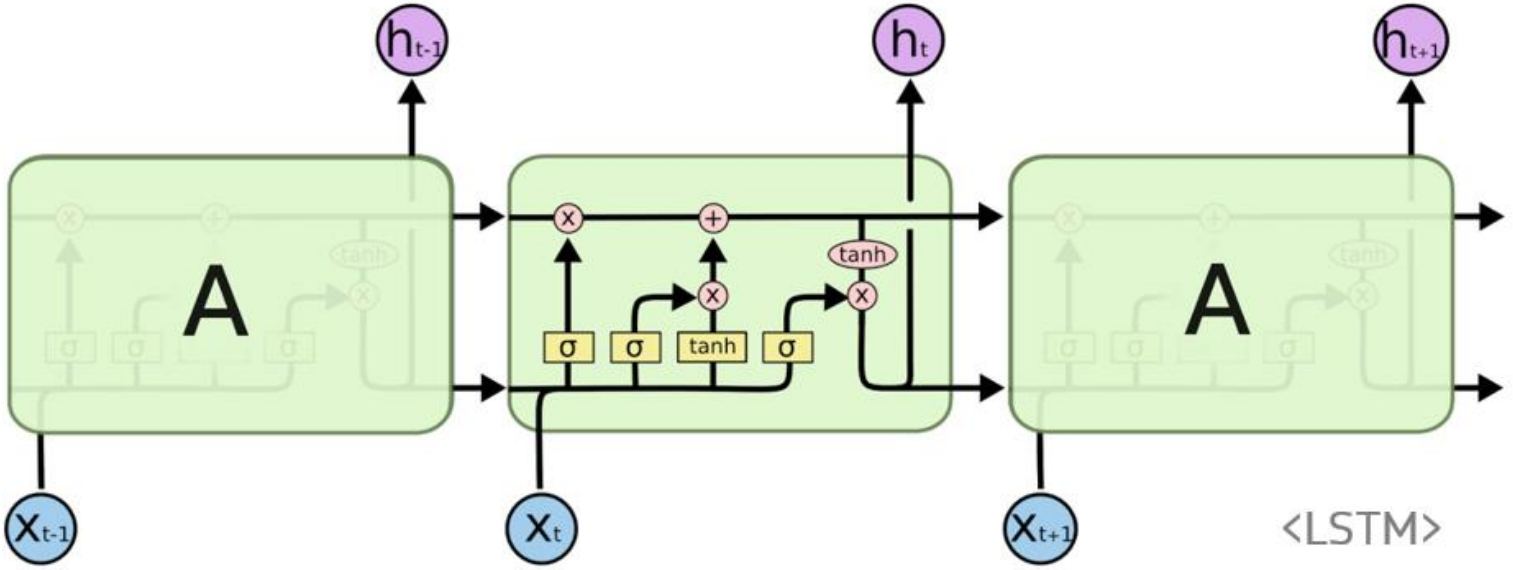
    ➔ 장기 의존성 문제(Problem of Long-Term Dependencies)



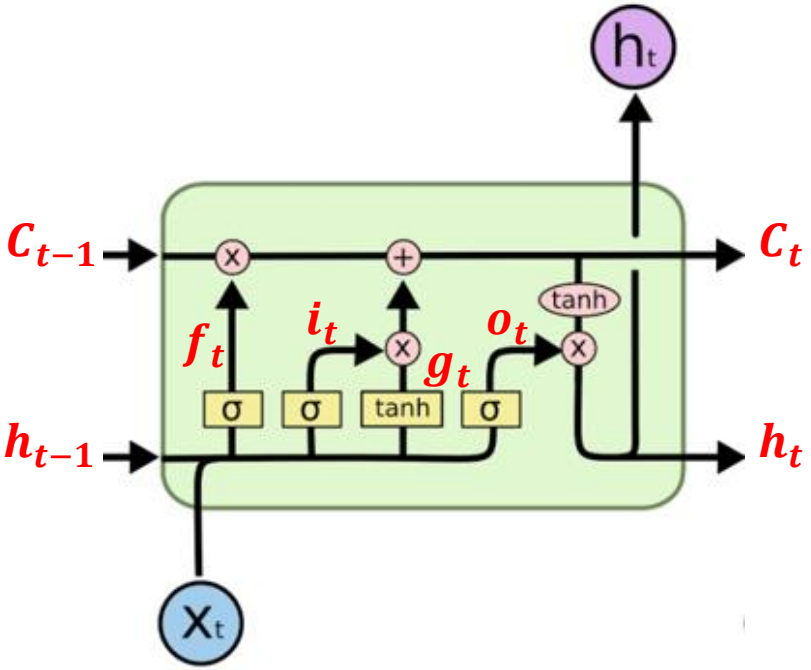
"모스크바에 여행을 왔는데 건물도 예쁘고 먹을 것도 맛있었어. 그런데 글쎄 직장 상사한테 전화가 왔어. 어디냐고 묻더라구 그래서 나는 말했지. 저 여행왔는데요. 여기 ____"

❖ 장단기 메모리(Long Short-Term Memory, LSTM)

- 은닉층의 메모리 셀에 입력 게이트, 망각(삭제) 게이트, 출력 게이트를 추가
- 불필요한 기억은 지우고, 기억해야 할 것들은 기억함
- LSTM은 한 층 안에서 반복을 많이 해야 하는 RNN의 특성상 일반 신경망보다 기울기 소실 문제가 더 많이 발생하고 이를 해결하기 어렵다는 단점을 보완한 방법
- 즉, 반복되기 직전에 다음 층으로 기억된 값을 넘길지 안 넘길지를 관리하는 단계를 하나 더 추가하는 것



❖ 장단기 메모리(Long Short-Term Memory, LSTM)



- 입력 게이트 :

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i)$$

$$g_t = \tanh(W_{xg}x_t + W_{hg}h_{t-1} + b_g)$$

- 삭제 게이트 :

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f)$$

- 셀 상태(장기 상태) :

$$C_t = f_t \circ C_{t-1} + i_t \circ g_t$$

- 출력 게이트와 은닉 상태(단기 상태) :

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o)$$

$$h_t = o_t \circ \tanh(C_t)$$

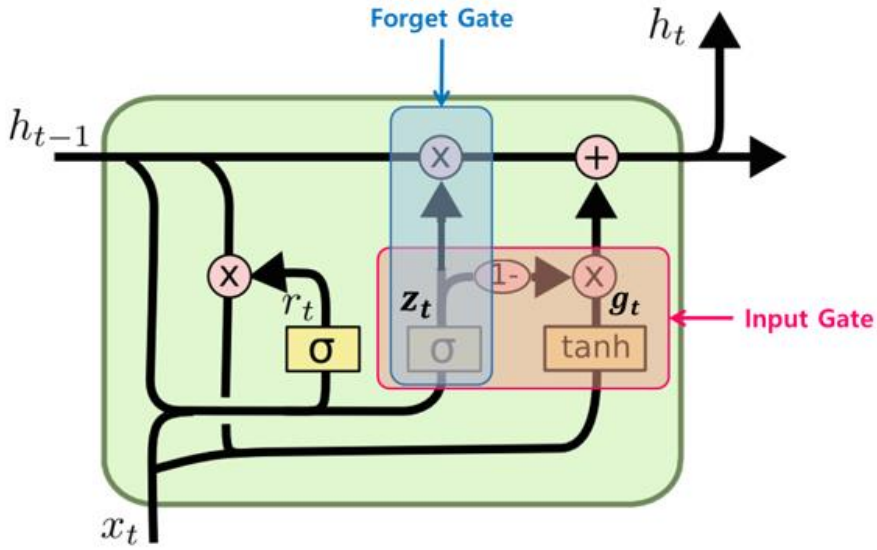
1. 개요

❖ LSTM을 이용한 텍스트 생성(One-to-many 구조)

Kaggle – New York Times Comments

❖ 게이트 순환 유닛(Gated Recurrent Unit, GRU)

- LSTM의 장기 의존성 문제에 대한 해결책을 유지하면서,
  은닉 상태를 업데이트하는 계산을 줄인 RNN
- 업데이트 게이트(z)와 리셋 게이트(r)
  두 가지 게이트만이 존재
- LSTM보다 학습 속도가 빠르다고
  알려져 있지만 여러 평가에서
  GRU는 LSTM과 비슷한 성능을
  보인다고 알려져 있음
- 사용 방법은 SimpleRNN이나
  LSTM과 동일




```
GRU(hidden_size, input_shape=(timesteps, input_dim))
```

❖ 워드 임베딩

▪ 워드 임베딩

	One-hot vector	Embedding vector
차원	고차원(단어 집합의 크기)	저차원
다른 표현	희소 벡터의 일종	밀집 벡터의 일종
값의 지정	수동	훈련 데이터로부터 학습함
값의 타입	1과 0	실수(float)

▪ 케라스의 Embedding()

- 단어를 밀집 벡터로 변환한 수
- 인공신경망 학습과 같이 단어 벡터를 학습하는 방법 사용
- `Embedding(vocab_size, output_dim, input_length)`

  `입력: (number of samples, input_length)`

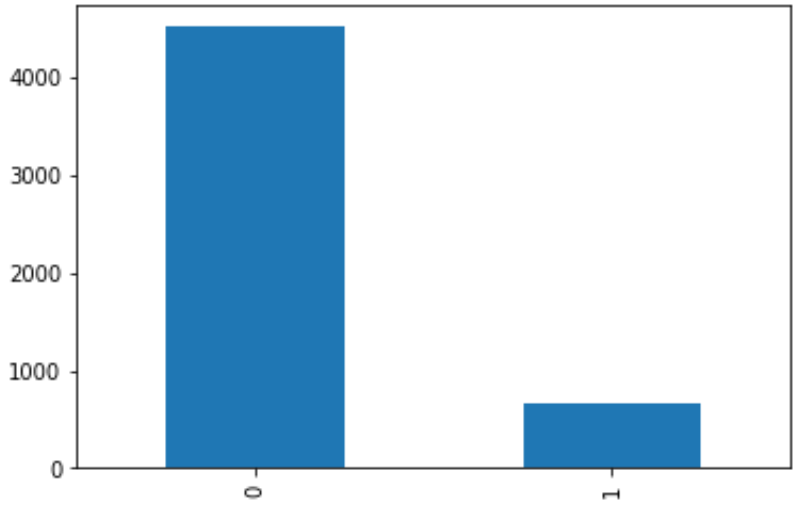
2. 분류

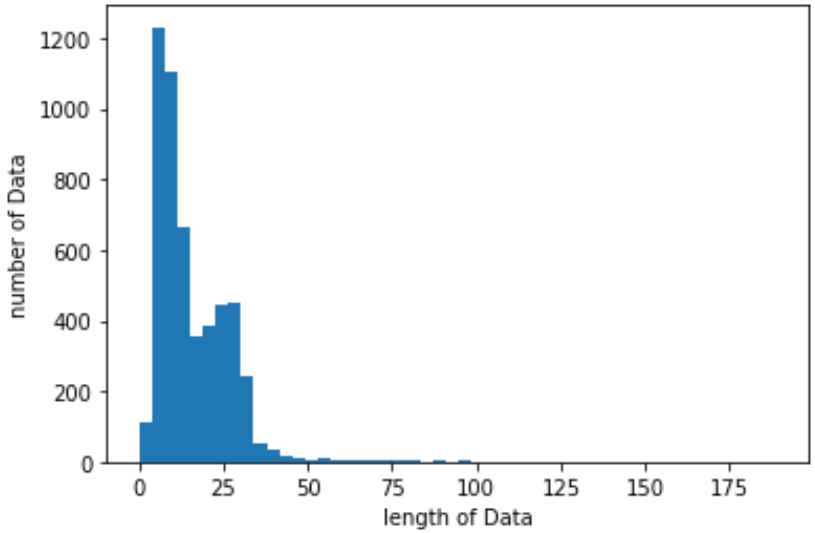
❖ 텍스트 분류 개요

- 지도 학습
- RNN의 다-대-일(Many-to-One) 문제
- 모든 시점(time step)에 대해서 입력을 받지만 최종 시점의 RNN 셀만이 은닉 상태를 출력하고, 이것이 출력층으로 가서 활성화 함수를 통해 정답을 고르는 문제
- 이진 분류
  - sigmoid 함수, binary_crossentropy, 출력층 크기 = 1
  - 스팸 메일 분류하기, IMDB 리뷰 감성 분류하기
- 다중 클래스 분류
  - softmax 함수, categorical_crossentropy, 출력층 크기 = N
  - 로이터 뉴스 분류하기

❖ 스팸 메일 분류

- 캐글에서 제공하는 스팸메일 데이터 활용
  - 총 5,572개의 데이터, 중복 데이터 403개
  - 중복 제거한 5,169개의 데이터 중 햄 4,516개, 스팸 653개
  - Unique 단어의 개수 : 8,920 개
  - 메일의 최대 단어 수 : 189 (평균 15.61)
  - 훈련 데이터 80%, 테스트 데이터 20%로 분리



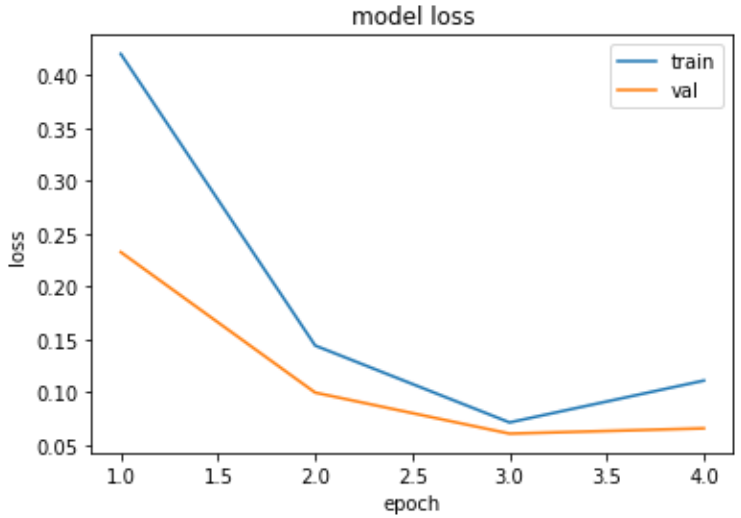


❖ 스팸 메일 분류

- 모델 설계

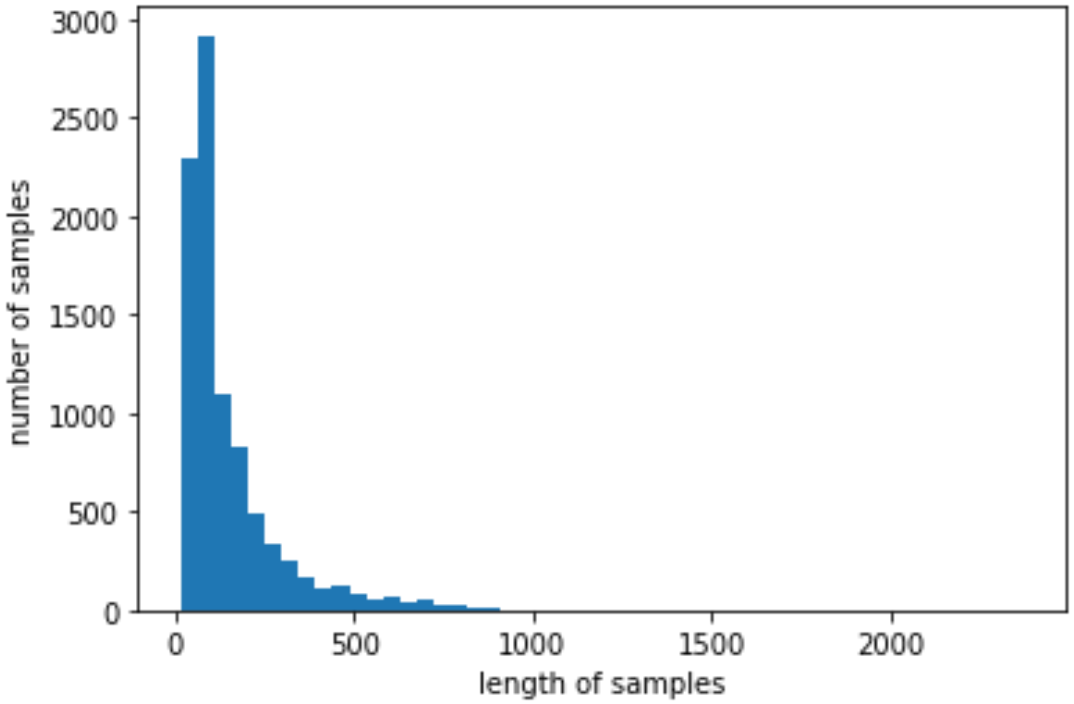
```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, None, 32)          285472
_________________________________________________________________
simple_rnn (SimpleRNN)       (None, 32)                2080
_________________________________________________________________
dense (Dense)                (None, 1)                 33
=================================================================
```

- optimizer='rmsprop', loss='binary_crossentropy'
- 훈련용 데이터 중 20%는 검증용으로 사용
- 테스트 정확도는 98.16%



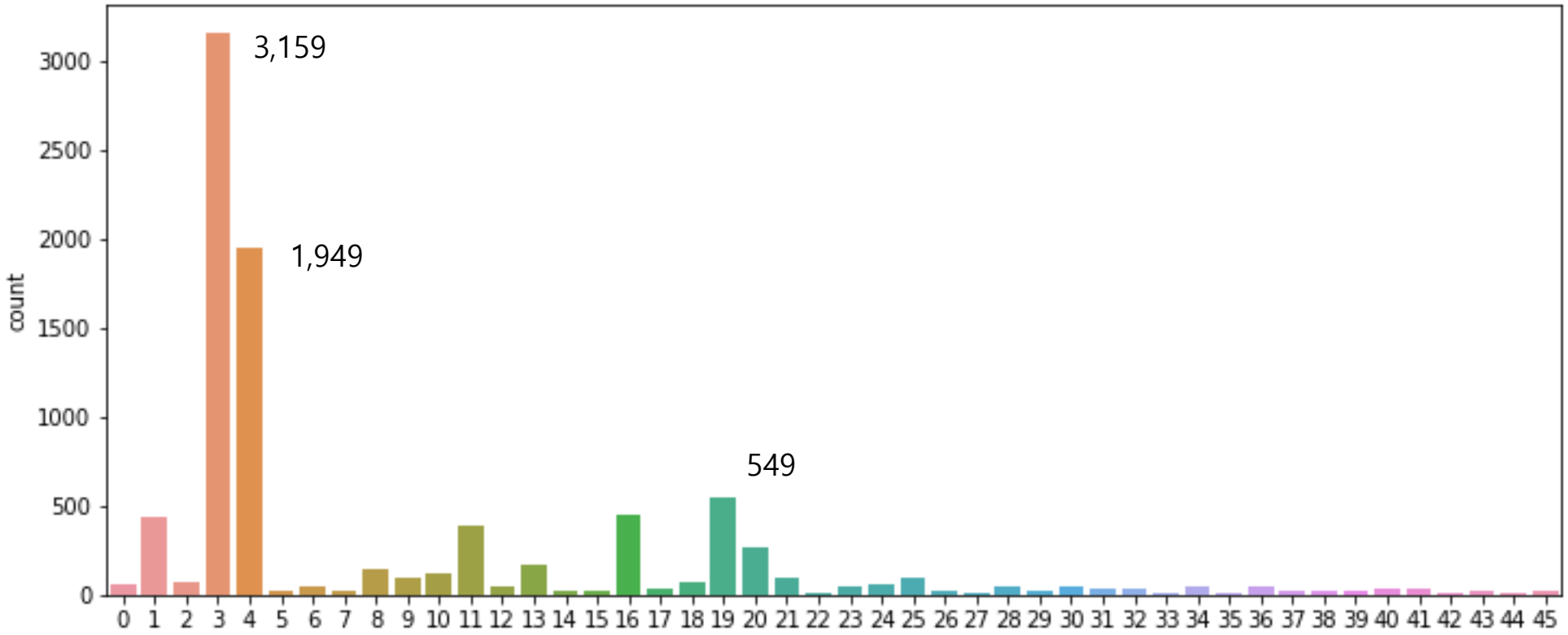
❖ 로이터 뉴스 분류

- Keras에서 제공하는 데이터 활용 (tensorflow.keras.datasets.reuters)
- 46개 카테고리의 총 11,228개의 뉴스 기사 데이터
- 전처리가 된 상태로 제공
- 뉴스 기사의 길이
  - 평균: 145.5 개의 단어
  - 최대: 2,376 단어



❖ 로이터 뉴스 분류

- 뉴스의 카테고리별 분포



❖ 로이터 뉴스 분류

- 단어 분포
  - 고유 단어수: 30,979개
  - 빈도 순으로 index 부여
  - get_word_index() 메소드 제공
  - 빈도수 1위: 'the'
- 첫번째 뉴스

X_train[0]	y_train[0]
[1, 27595, 28842, 8, 43, 10, 447, 5, 25, 207, 270, 5, 3095, 111, 16, 369, 186, 90, 67, 7, 89, 5, 19, 102, 6, 19, 124, 15, 90, 67, 84, 22, 482, 26, 7, 48, 4, 49, 8, 864, 39, 209, 154, 6, 151, 6, 83, 11, 15, 22, 155, 11, 15, 7, 48, 9, 4579, 1005, 504, 6, 258, 6, 272, 11, 15, 22, 134, 44, 11, 15, 16, 8, 197, 1245, 90, 67, 52, 29, 209, 30, 32, 132, 6, 109, 15, 17, 12]	3
the wattie nondiscriminatory mln loss for plc said at only ended said commonwealth could 1 traders now april 0 a after said from 1985 and from foreign 000 april 0 prices its account year a but in this mln home an states earlier and rise and revs vs 000 its 16 vs 000 a but 3 psbr oils several and shareholders and dividend vs 000 its all 4 vs 000 1 mln agreed largely april 0 are 2 states will billion total and against 000 pct dlrs	

❖ 로이터 뉴스 분류

- LSTM 모델
  - 빈도수 1000 까지의 단어만 사용
  - 모든 문장이 아니라 100 단어 까지만 사용
- 120 차원의 Embedding vector

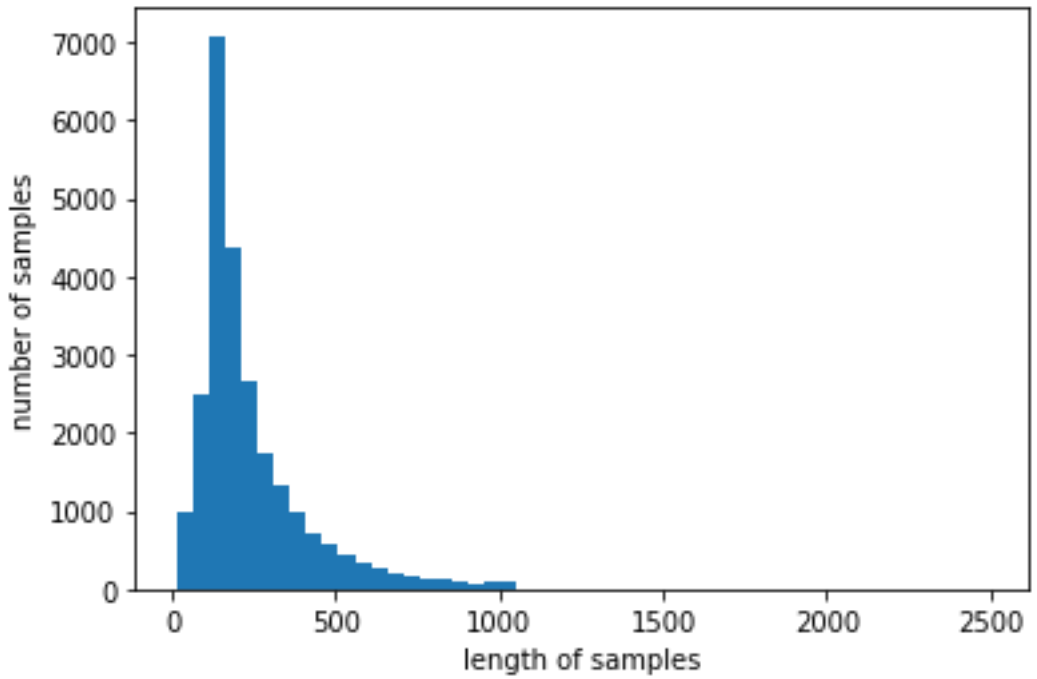
```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, None, 120)         120000
_________________________________________________________________
lstm (LSTM)                  (None, 120)               115680
_________________________________________________________________
dense (Dense)                (None, 46)                5566
=================================================================
```

- optimizer='adam', loss='categorical_crossentropy'
- 훈련용 데이터 중 20%는 검증용으로 사용

❖ IMDB 영화 리뷰 감성 분류

- Keras에서 제공하는 데이터 활용 (tensorflow.keras.datasets.imdb)
- 2개 카테고리(긍정: 1, 부정: 0)의 총 50,000개의 영화 리뷰 데이터
- 스탠포드 대학교에서 2011년에 낸 논문에서 이 데이터를 소개

  훈련 데이터와 테스트 데이터를 50:50대 비율로 분할하여 88.89%의 정확도
- 길이
  - 평균: 238.7 개의 단어
  - 최대: 2,494 단어



❖ IMDB 영화 리뷰 감성 분류

- 단어 분포
  - 고유 단어수: 88,584개
- 일곱번째 리뷰

X_train[6]	y_train[6]
[1, 6740, 365, 1234, 5, 1156, 354, 11, 14, 5327, 6638, 7, 1016, 10626, 5940, 356, 44, 4, 1349, 500, 746, 5, 200, 4, 4132, 11, 16393, 9363, 1117, 1831, 7485, 5, 4831, 26, 6, 71690, 4183, 17, 369, 37, 215, 1345, 143, 32677, 5, 1838, 8, 1974, 15, 36, 119, 257, 85, 52, 486, 9, 6, 26441, 8564, 63, 271, 6, 196, 96, 949, 4121, 4, 74170, 7, 4, 2212, 2436, 819, 63, 47, 77, 7175, 180, 6, 227, 11, 94, 2494, 33740, 13, 423, 4, 168, 7, 4, 22, 5, 89, 665, 71, 270, 56, 5, 13, 197, 12, 161, 5390, 99, 76, 23, 77842, 7, 419, 665, 40, 91, 85, 108, 7, 4, 2084, 5, 4773, 81, 55, 52, 1901]	1
the boiled full involving to impressive boring this as murdering naschy br villain council suggestion need has of costumes b message to may of props this echoed concentrates concept issue skeptical to god's he is dedications unfolds movie women like isn't surely i'm rocketed to toward in here's for from did having because very quality it is captain's starship really book is both too worked carl of mayfair br of reviewer closer figure really there will originals things is far this make mistakes kevin's was couldn't of few br of you to don't female than place she to was between that nothing dose movies get are 498 br yes female just its because many br of overly to descent people time very bland	positive

❖ IMDB 영화 리뷰 감성 분류

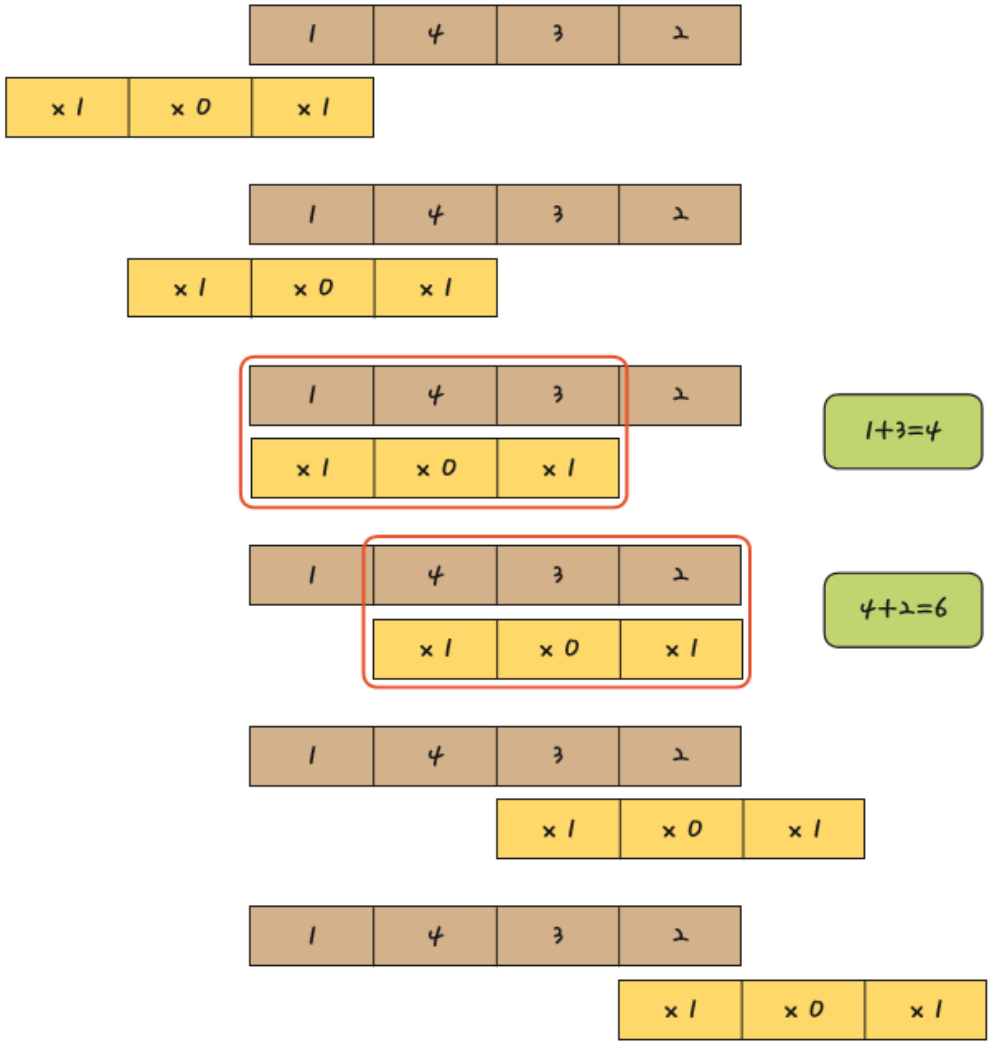
- LSTM 모델
  - 빈도수 5000 까지의 단어만 사용
  - 모든 문장이 아니라 500 단어 까지만 사용
- 120 차원의 Embedding vector

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, None, 120)         600000
_________________________________________________________________
lstm (LSTM)                  (None, 120)               115680
_________________________________________________________________
dense (Dense)                (None, 1)                 121
=================================================================
```

- optimizer='adam', loss='binary_crossentropy'

❖ IMDB 영화 리뷰 감성 분류 – LSTM + CNN

- Conv1D



❖ IMDB 영화 리뷰 감성 분류 – LSTM + CNN

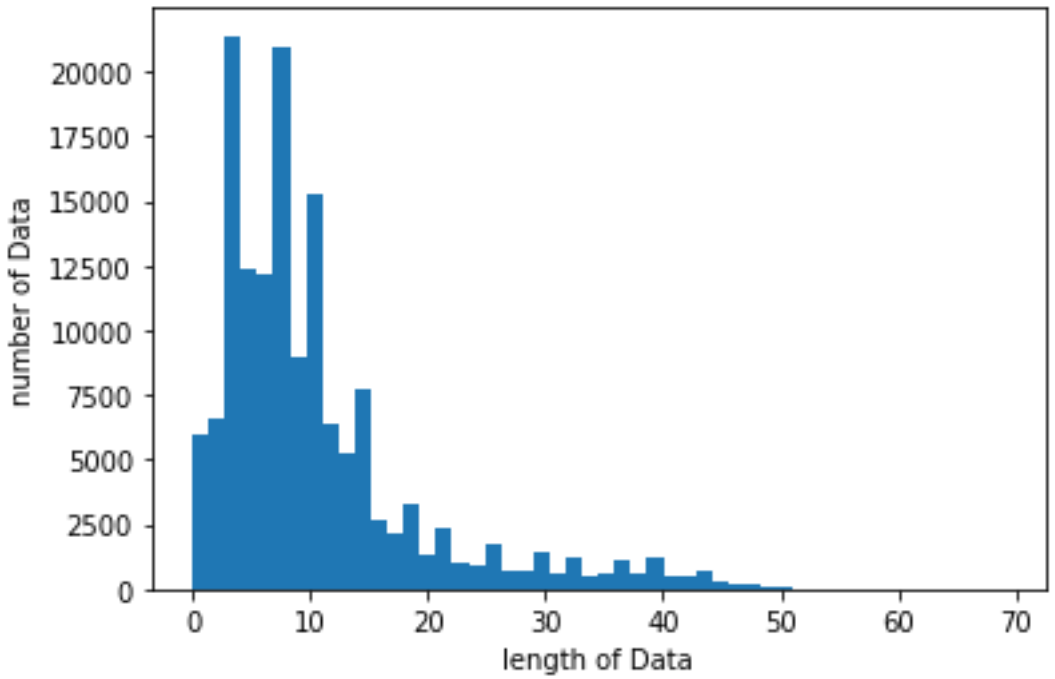
- MaxPooling1D

- Model

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, None, 120)         600000
_________________________________________________________________
dropout (Dropout)            (None, None, 120)         0
_________________________________________________________________
conv1d (Conv1D)              (None, None, 64)          38464
_________________________________________________________________
max_pooling1d (MaxPooling1D) (None, None, 64)          0
_________________________________________________________________
lstm (LSTM)                  (None, 55)                26400
_________________________________________________________________
dense (Dense)                (None, 1)                 56
=================================================================
```

❖ 네이버 영화 리뷰 감성 분류

- Github에 올라가 있는 데이터 활용 (https://github.com/e9t/nsmc/)
- 2개 카테고리(긍정: 1, 부정: 0)의 총 200,000개의 영화 리뷰 데이터
  리뷰 점수: 10~9 → 긍정, 4~1 → 부정, 8~5 → 미사용
- 학습용: 150,000개, 테스트용: 50,000개
- 길이
  - 평균: 10.65 개의 단어
  - 최대: 69 단어



❖ 네이버 영화 리뷰 감성 분류

- 한글 데이터 전처리
  - 특수 문자, 영어 등을 모두 제거한 후 한글만 남김 ← 정규 표현식 활용

```
train_data['document'] = train_data['document'].str.replace("[^ㄱ-ㅎㅏ-ㅣ가-힣 ]","")
```

  - 토큰화
  - 불용어 제거

- 모델

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, None, 100)         3500000
_________________________________________________________________
lstm (LSTM)                  (None, 128)               117248
_________________________________________________________________
dense (Dense)                (None, 1)                 129
=================================================================
```