긴 시퀀스 다루기

긴 시퀀스(타임스텝 개수 多)

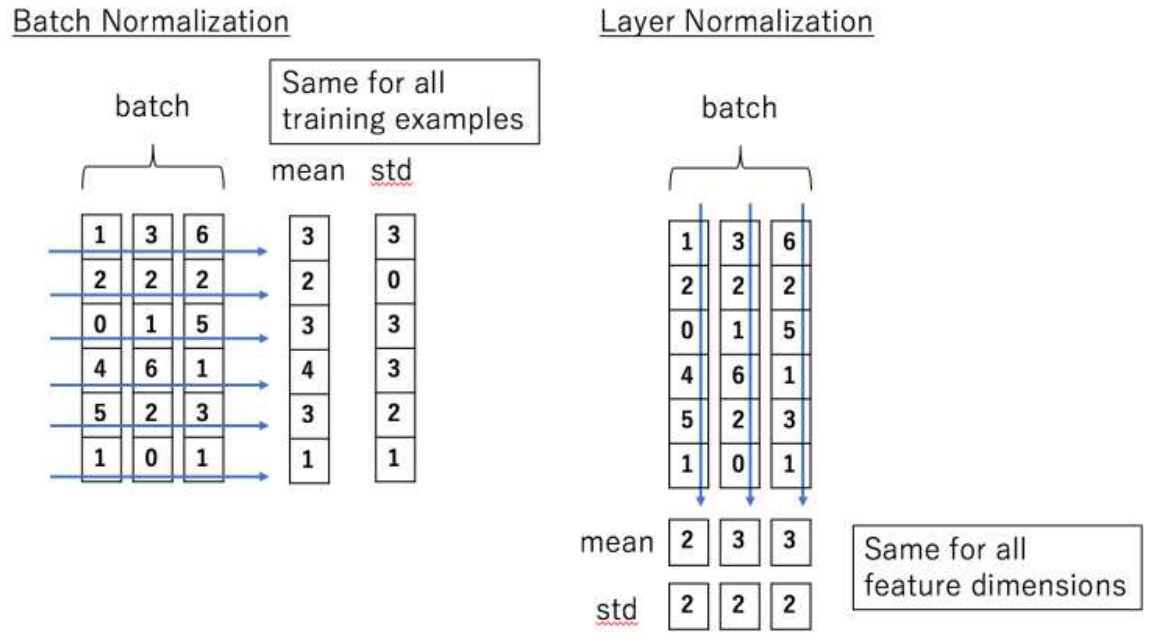
- 그레디언트 소실 · 폭주
- 오랜 훈련시간

심층 신경망의 기법들 차용

- 가중치 초기화 기법, 옵티마이저, 드롭아웃

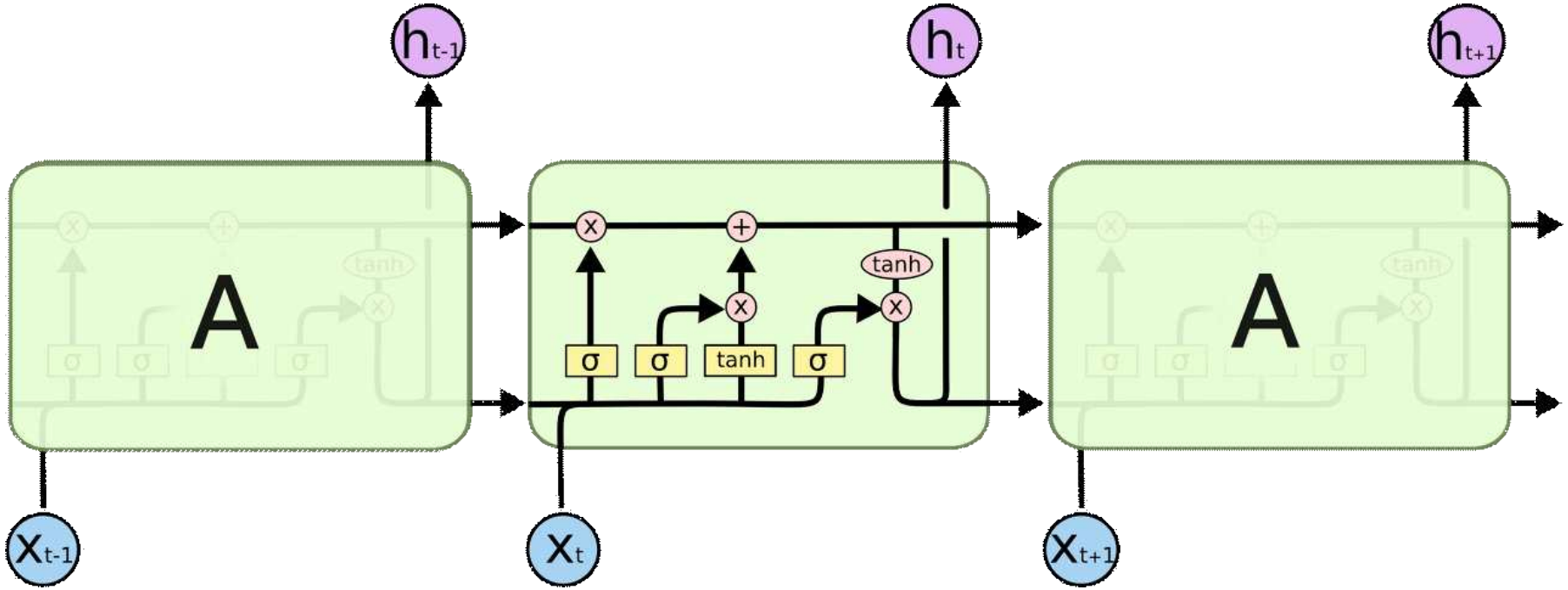
층 정규화

- 배치 차원이 아닌 특성 차원에서 정규화
- 샘플에 독립적으로 타임스텝마다 필요한 통계 계산



이미지 출처: https://yonghyuc.wordpress.com/2020/03/04/batch-norm-vs-layer-norm/

LSTM

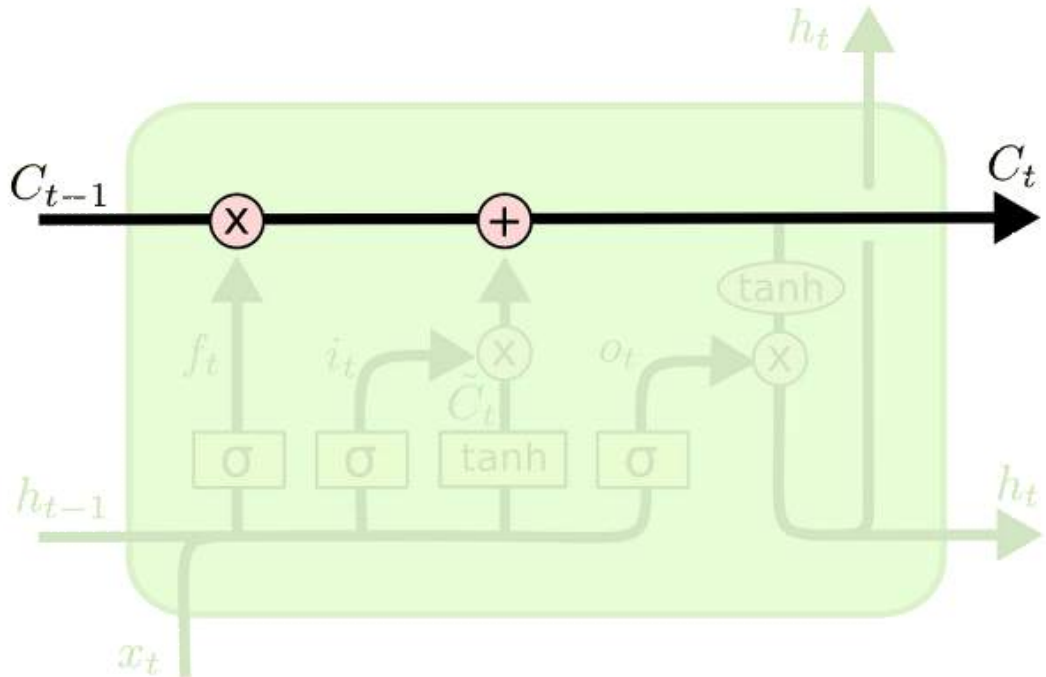


이미지 출처: https://dgkim5360.tistory.com/entry/understanding-long-short-term-memory-lstm-kr

- 장기 기억, 단기 기억 따로 처리
- σ: 로지스틱 (0 ~ 1), tanh: 하이퍼볼릭 탄젠트 (-1 ~ 1)

LSTM 구조

장기 기억



이미지 출처: https://dgkim5360.tistory.com/entry/understanding-long-short-term-memory-lstm-kr

C_{t-1}: t-1 타임스텝의 장기 기억

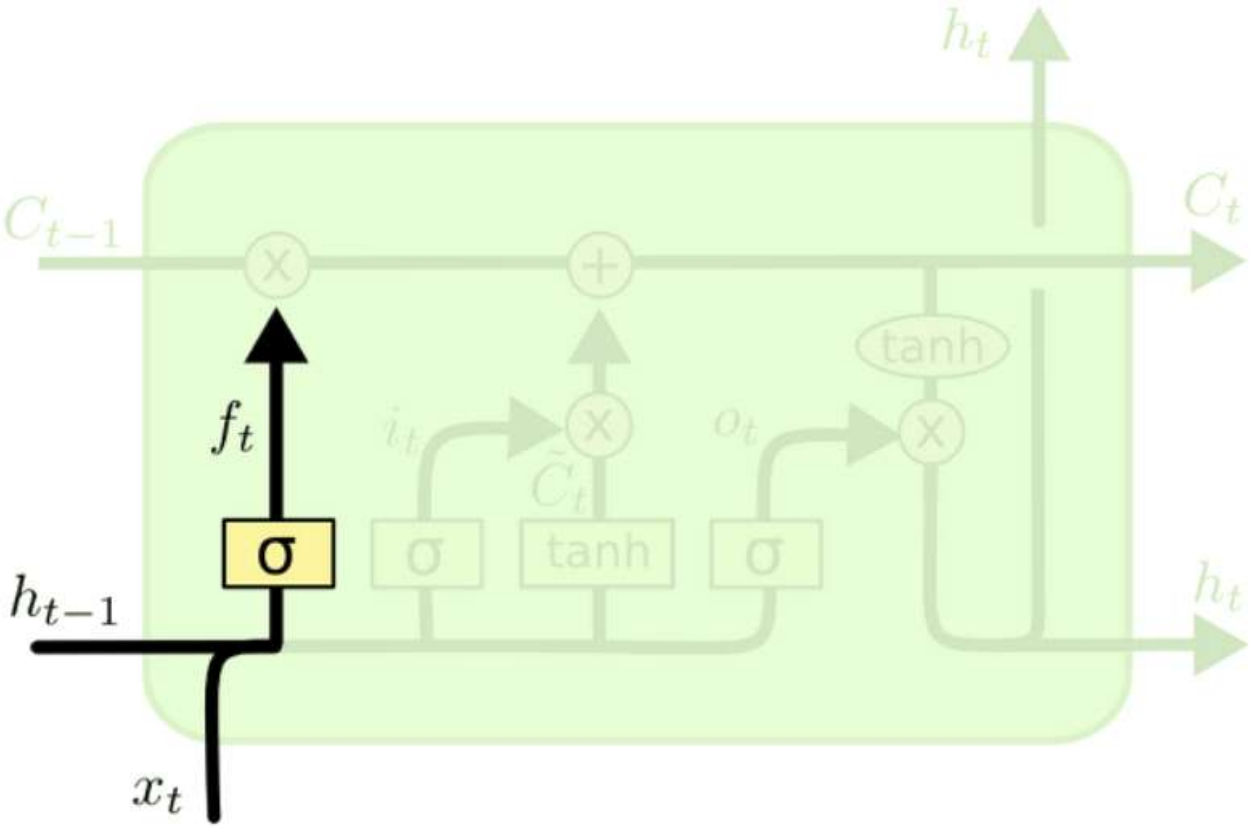
C_t: t 타임스텝의 장기 기억

LSTM 구조

삭제 게이트

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

시그모이드 활성화함수: 0 ~ 1 값 출력



이미지 출처: https://dgkim5360.tistory.com/entry/understanding-long-short-term-memory-lstm-kr

장기 상태의 어느 부분이 삭제될지 제어

LSTM 구조

입력 게이트

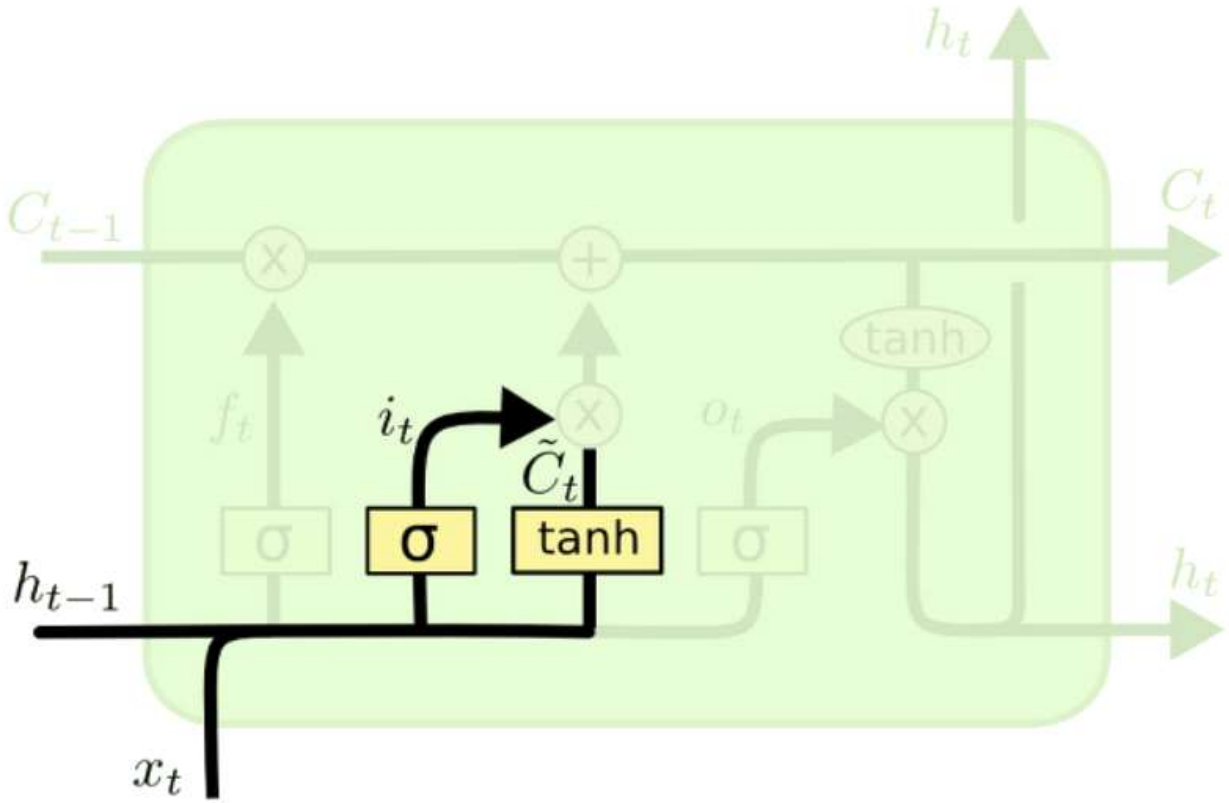
$$i_t = \sigma(W_{xi}^T x_t + W_{hi}^T h_{t-1} + b_i)$$

0 ~ 1 사이 값

주 층

$$\widetilde{C}_t = \tanh(W_{xg}^T x_t + W_{hg}^T h_{t-1} + b_g)$$

-1 ~ 1 사이 값



이미지 출처: https://dgkim5360.tistory.com/entry/understanding-long-short-term-memory-lstm-kr

입력 게이트가 주 층의 어느 부분을 장기 상태에 더할지 제어

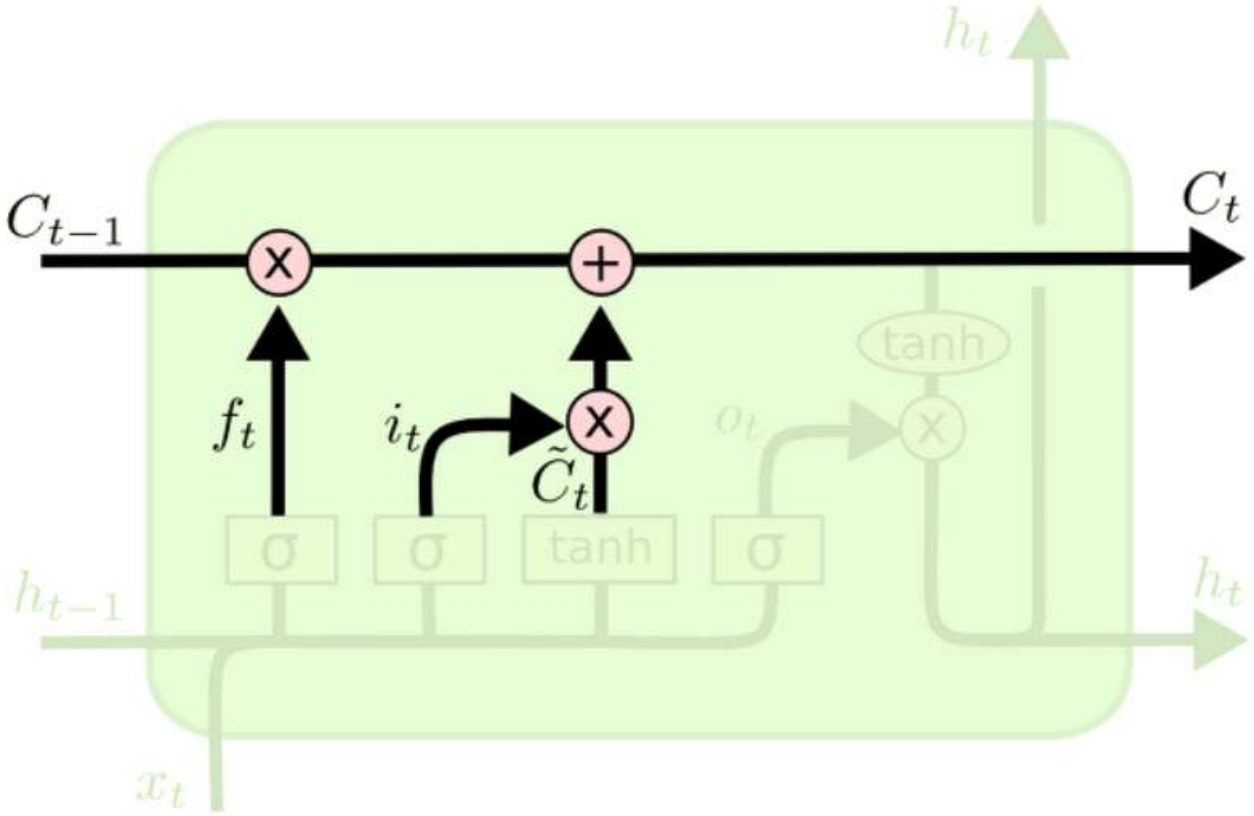
LSTM 구조

t 타임스텝의 장기 기억

$$C_t = f_t \otimes c_{t-1} + i_t \otimes \widetilde{C}_t$$

삭제 게이트로 손실된 t-1 타임스텝의 장기 기억

\+ 입력 게이트로 제어된 주 층



이미지 출처: https://dgkim5360.tistory.com/entry/understanding-long-short-term-memory-lstm-kr

LSTM 구조

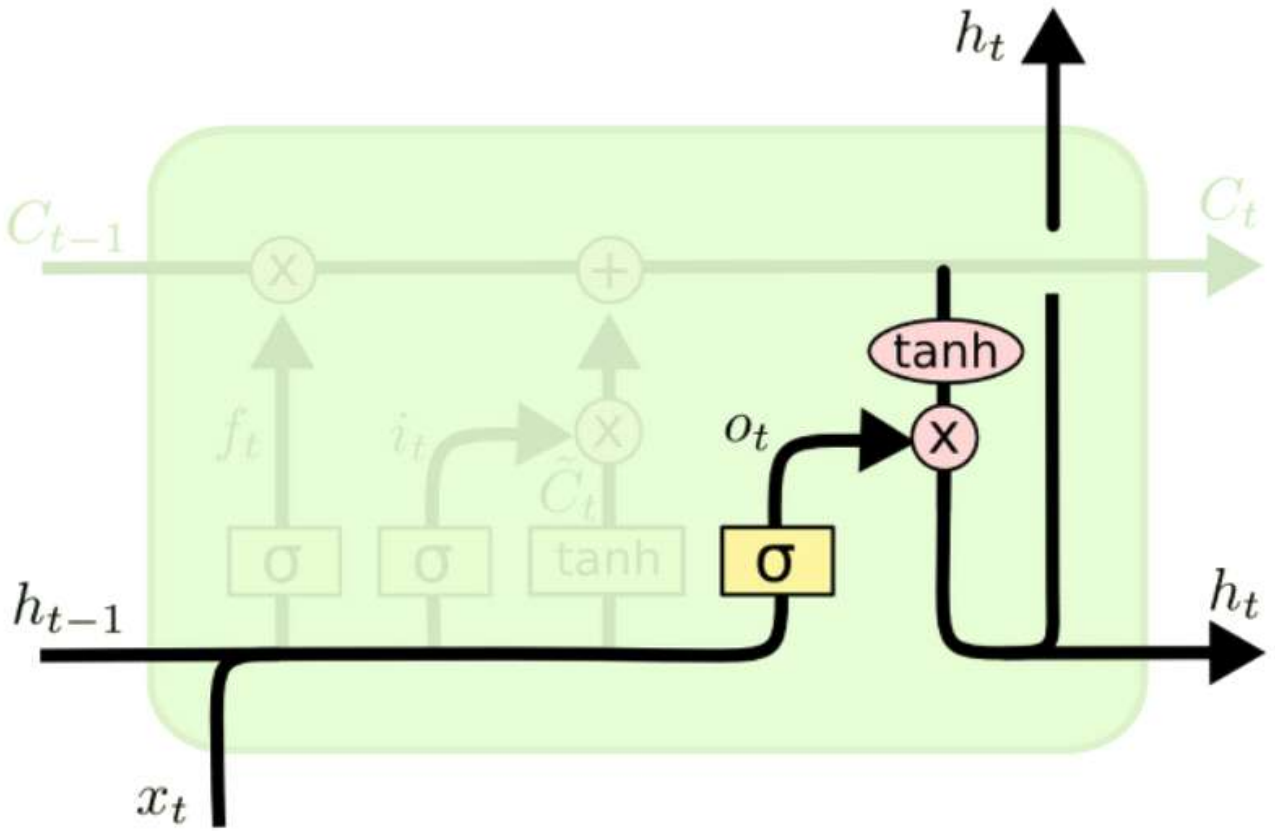
출력 게이트

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

0 ~ 1 사이 값

출력 값, 은닉 값

$$y_t = h_t = o_t \otimes \tanh(c_t)$$



이미지 출처: https://dgkim5360.tistory.com/entry/understanding-long-short-term-memory-lstm-kr

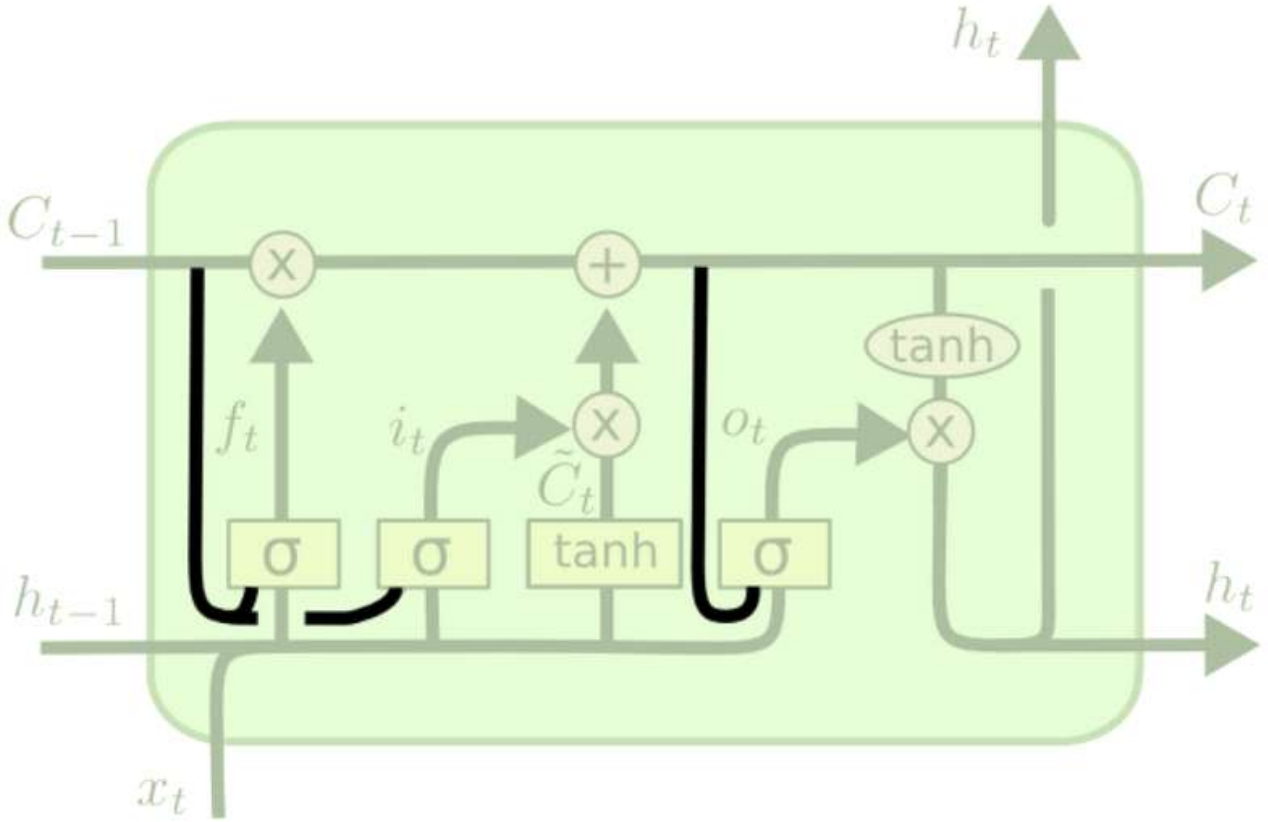
출력 게이트에서 제어된 장기 기억이 출력 또는 다음 타임스텝으로 넘겨짐

LSTM의 변형: 핍홀

삭제 게이트를 결정하는 요소에 이전 장기기억 추가

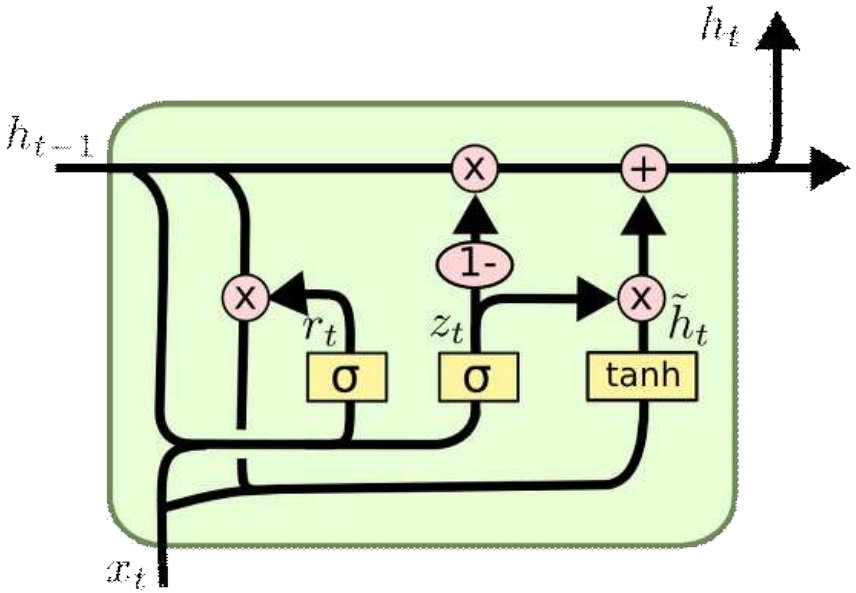
입력 게이트를 결정하는 요소에 이전 장기 기억 추가

출력 게이트를 결정하는 요소에 현재 장기 기억 추가



이미지 출처: https://dgkim5360.tistory.com/entry/understanding-long-short-term-memory-lstm-kr

GRU 구조



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

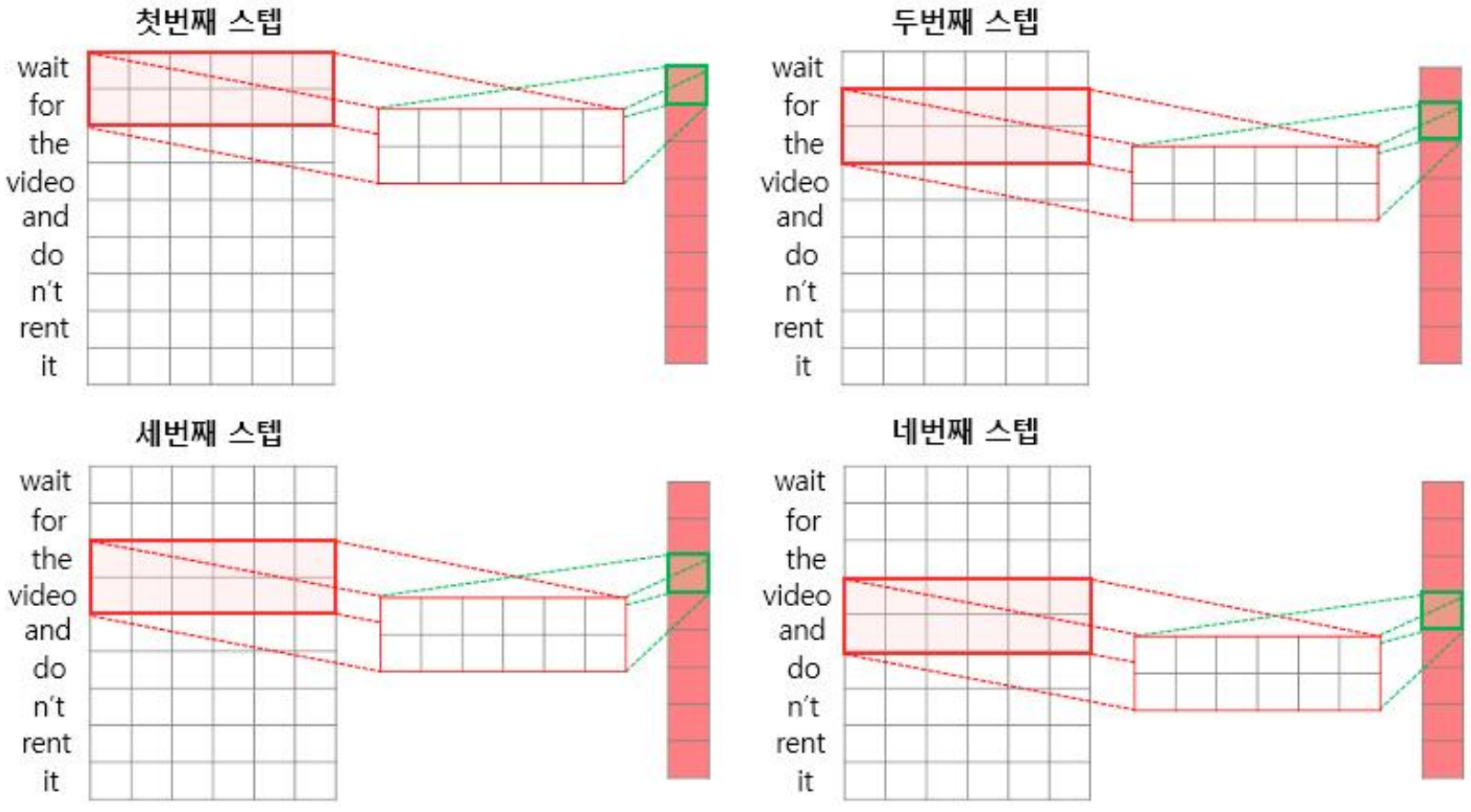
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

이미지 출처: https://dgkim5360.tistory.com/entry/understanding-long-short-term-memory-lstm-kr

- 장기 상태와 단기 상태가 h로 합쳐짐
- r_t: 이전 타임스텝의 출력 h_{t-1} 중 어느 부분이 주 층에 노출될지 결정
- 하나의 게이트 z_t가 삭제 게이트와 입력 게이트를 겸함

z_t: 0 ~ 1 사이 값 입력 게이트 / $1-z_t$: 0 ~ 1 사이 값 삭제 게이트

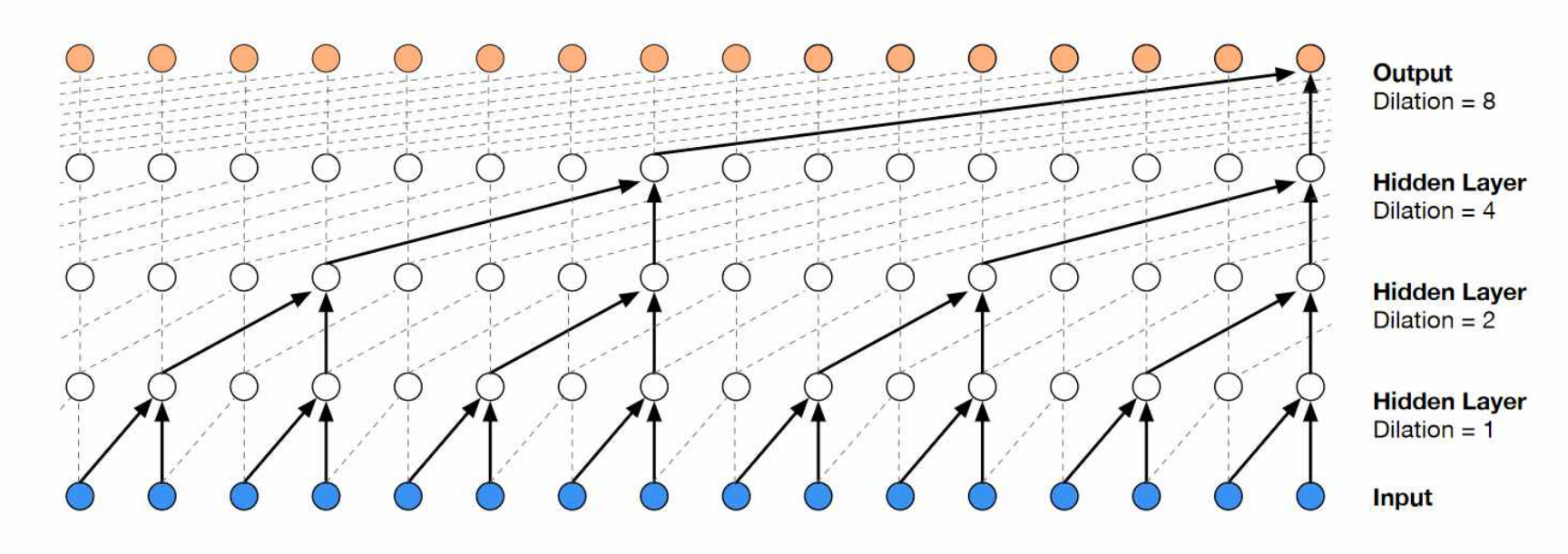
1D 합성곱 층으로 시퀀스 제어



이미지 출처: https://wikidocs.net/80437

1D: 필터의 슬라이싱 축이 한 방향만 존재

WAVENET



이미지 출처: https://arxiv.org/pdf/1609.03499.pdf

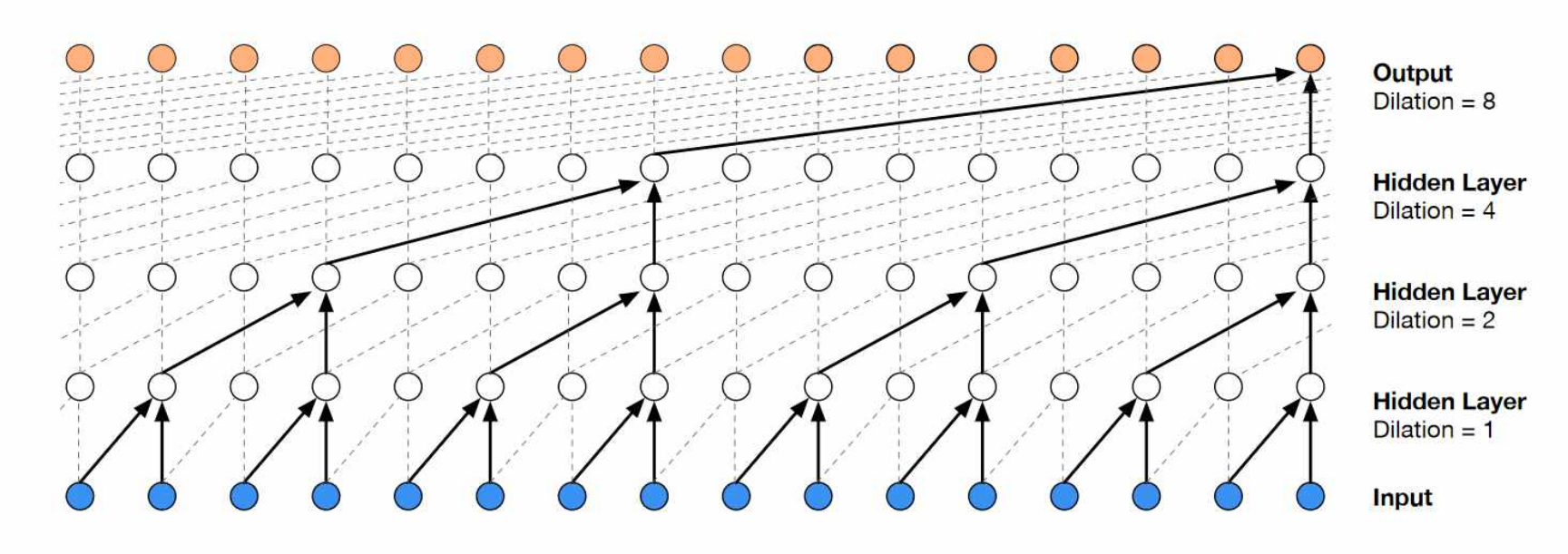
팽창비율을 1, 2, 4, 8, … , 256, 512로 하는 합성곱 층을 쌓음.

- 예) 팽창비율 2: [1, 2] → [1, 0, 2]

  팽창비율 4: [1, 2] → [1, 0, 0, 0, 2]

쌓은 합성곱 층 10개(팽창비율 1 ~ 512)를 하나의 층으로 두고 동일한 10개의 층을 쌓음

WAVENET



이미지 출처: https://arxiv.org/pdf/1609.03499.pdf

- 빠르고 강력, 더 적은 파라미터 사용

- 시퀀스가 긴 오디오 데이터에서 강력한 성능

참고 자료

- 오헬리앙 제롱. *핸즈온 머신러닝 2판*(한빛미디어, 2020). 614-625

- Batch Norm vs Layer Norm, *Lifetime behind every seconds*, 2020년 3월 4일 수정, 2021년 9월 4일 접속. https://yonghyuc.wordpress.com/2020/03/04/batch-norm-vs-layer-norm/.

- Long Short-Term Memory (LSTM) 이해하기, *개발새발로그*, 2018년 4월 10일 수정, 2021년 9월 4일 접속. https://dgkim5360.tistory.com/entry/understanding-long-short-term-memory-lstm-kr.

- 자연어 처리를 위한 1D CNN, *딥 러닝을 통한 자연어 처리 입문*, 2021년 9월 4일 수정, 2021년 9월 4일 접속. https://wikidocs.net/80437.

- Aäron van den Oor 외 8명, “WAVENET: A GENERATIVE MODEL FOR RAW AUDIO”, arXiv preprint arXiv:1609.03499 (2016). https://arxiv.org/pdf/1609.03499.pdf.