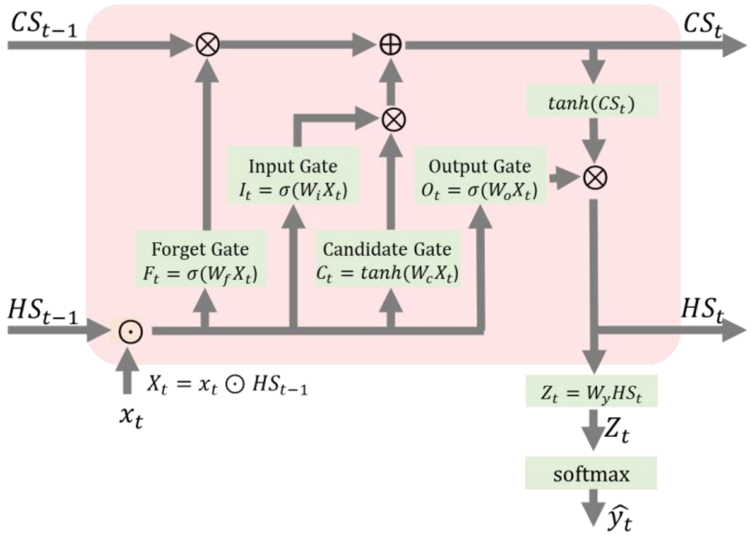
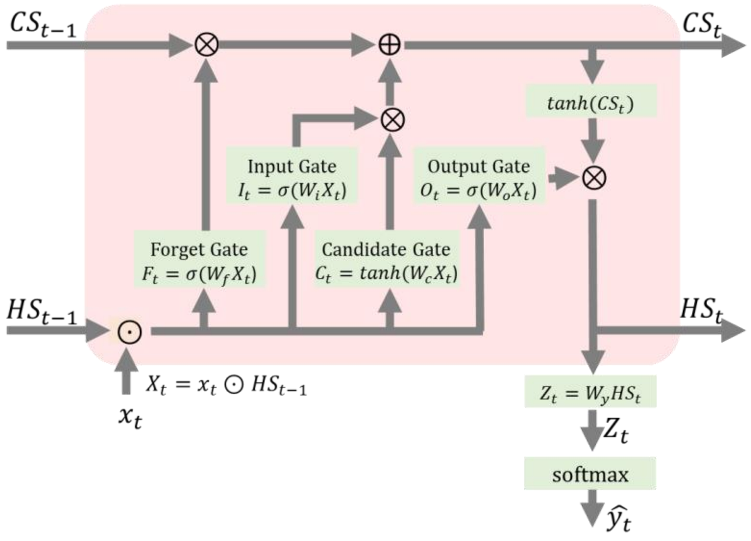
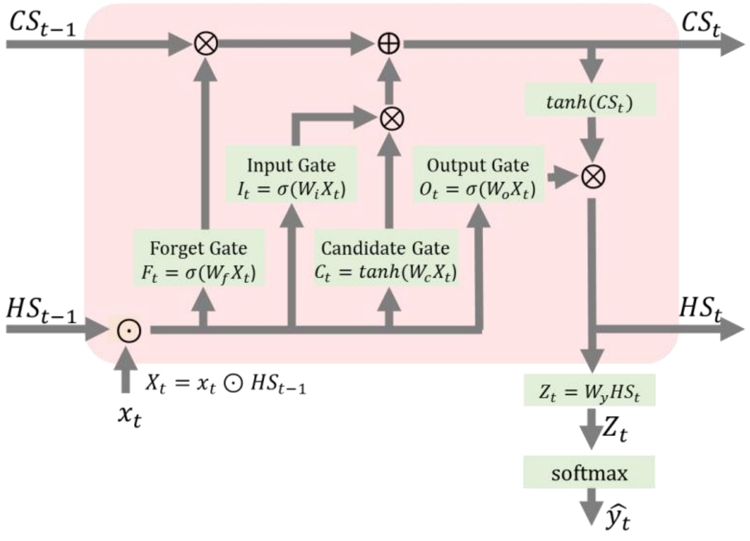
Deep Learning101

LSTM 신경망

초보자를 위한 안내서







신박AI

안녕하세요 신박AI입니다

오늘은 LSTM에 대해 같이 알아보는 시간을
가져보도록 하겠습니다

신박AI

LSTM은 Long Short-Term Memory
의 약자로,

RNN처럼 시계열 데이터를 처리할 때
사용되는 신경망입니다

신박AI

RNN은 시계열 데이터를 처리함에 있어서
한가지 중요한 약점이 있었는데요

LSTM 은 그러한 RNN의 약점을 극복하기
위해 탄생한 신경망입니다

그래서 오늘은 이 LSTM이
탄생하게 된 배경과

신박AI

LSTM의 구조와 개념에 대해
간략하게 소개해드리고

LSTM이 시계열 정보를 학습하는 알고리즘에 대해

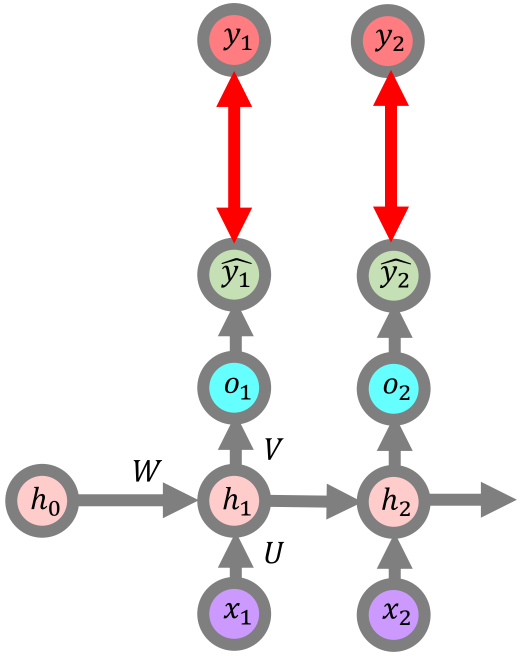
간단한 산수(?)와 함께
소개해드리고자 합니다

신박AI

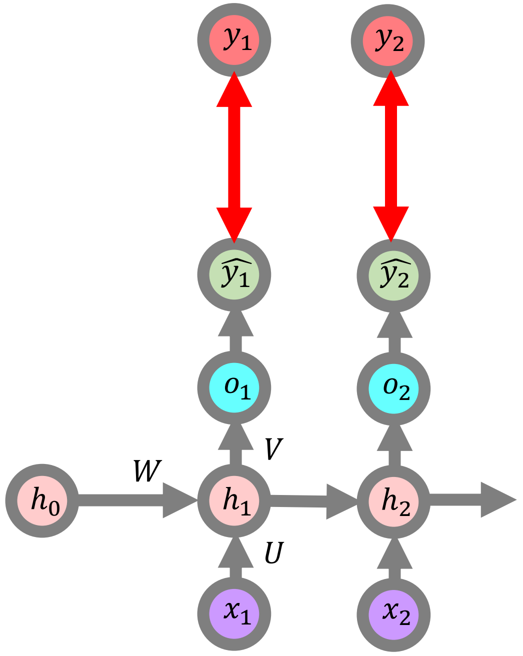
먼저 LSTM이 탄생하게 된 배경에 대해 말씀드리겠습니다

신박AI

RNN은 시계열 데이터를 처리하는데 있어서 크리티컬한 약점이 있었습니다



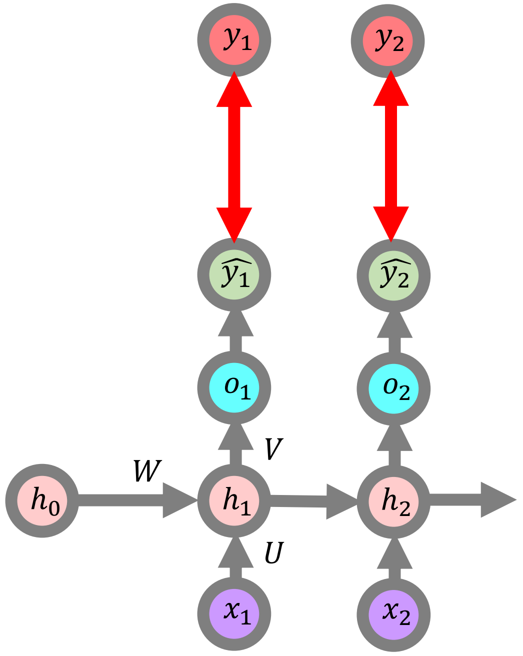
바로 장기 의존성 (long-term dependency)라는 약점이 그것입니다



신박AI

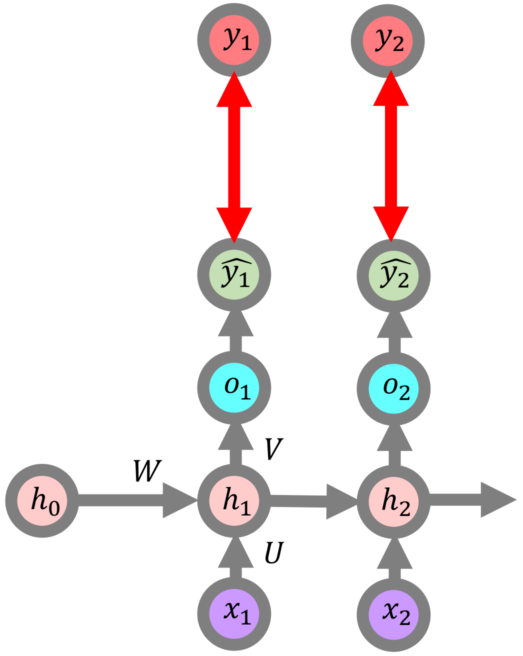
장기 의존성을 설명하기 위해 지난 영상의 식을 잠간 빌려 오겠습니다

$$\frac{\partial L_2}{\partial W}=\frac{\partial L_2}{\partial \widehat{y_2}}\frac{\partial \widehat{y_2}}{\partial o_2}\frac{\partial o_2}{\partial h_2}\frac{\partial h_2}{\partial W}+\frac{\partial L_2}{\partial \widehat{y_2}}\frac{\partial \widehat{y_2}}{\partial o_2}\frac{\partial o_2}{\partial h_2}\frac{\partial h_2}{\partial h_1}\frac{\partial h_1}{\partial W}$$



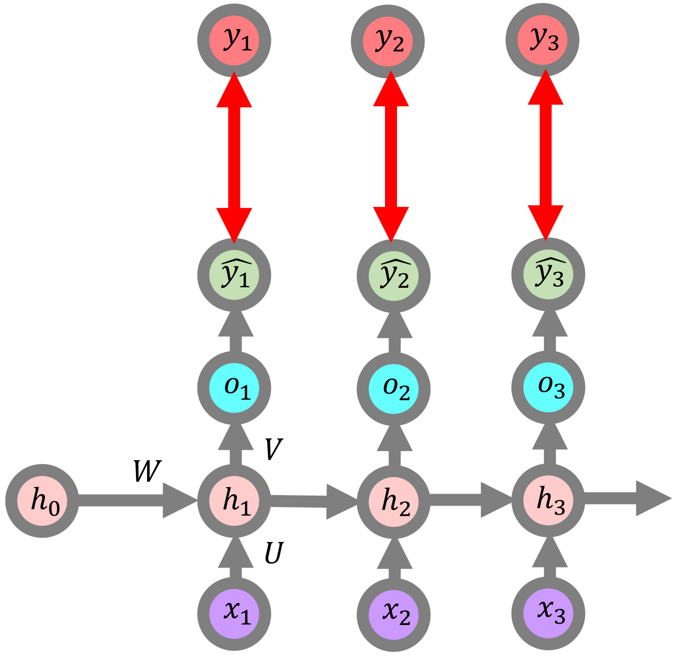
여기보시면 시간이 늘어날수록,

$$\frac{\partial L_2}{\partial W} = \frac{\partial L_2}{\partial \widehat{y_2}}\frac{\partial \widehat{y_2}}{\partial o_2}\frac{\partial o_2}{\partial h_2}\frac{\partial h_2}{\partial W} + \frac{\partial L_2}{\partial \widehat{y_2}}\frac{\partial \widehat{y_2}}{\partial o_2}\frac{\partial o_2}{\partial h_2}\boxed{\frac{\partial h_2}{\partial h_1}}\frac{\partial h_1}{\partial W}$$



여기보시면 시간이 늘어날수록, 체인룰로 계산해야할 부분이

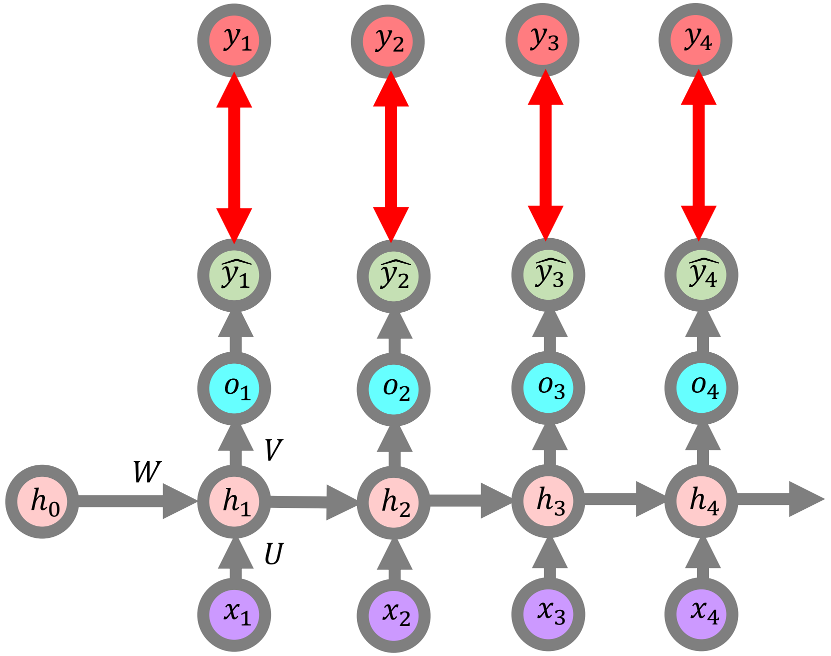
$$\frac{\partial L_3}{\partial W}=\frac{\partial L_3}{\partial \widehat{y_3}}\frac{\partial \widehat{y_3}}{\partial o_3}\frac{\partial o_3}{\partial h_3}\frac{\partial h_3}{\partial W}+\frac{\partial L_3}{\partial \widehat{y_3}}\frac{\partial \widehat{y_3}}{\partial o_3}\frac{\partial o_3}{\partial h_3}\frac{\partial h_3}{\partial h_2}\frac{\partial h_2}{\partial W}+\frac{\partial L_3}{\partial \widehat{y_3}}\frac{\partial \widehat{y_3}}{\partial o_3}\frac{\partial o_3}{\partial h_3}\frac{\partial h_3}{\partial h_2}\frac{\partial h_2}{\partial h_1}\frac{\partial h_1}{\partial W}$$



여기보시면 시간이 늘어날수록, 체인룰로 계산해야할 부분이 계속

$$\frac{\partial L_4}{\partial W}=\frac{\partial L_4}{\partial \widehat{y_4}}\frac{\partial \widehat{y_4}}{\partial o_4}\frac{\partial o_4}{\partial h_4}\frac{\partial h_4}{\partial W}+\frac{\partial L_4}{\partial \widehat{y_4}}\frac{\partial \widehat{y_4}}{\partial o_4}\frac{\partial o_4}{\partial h_4}\frac{\partial h_4}{\partial h_3}\frac{\partial h_3}{\partial W}+\frac{\partial L_4}{\partial \widehat{y_4}}\frac{\partial \widehat{y_4}}{\partial o_4}\frac{\partial o_4}{\partial h_4}\frac{\partial h_4}{\partial h_3}\frac{\partial h_3}{\partial h_2}\frac{\partial h_2}{\partial W}$$

$$+\frac{\partial L_4}{\partial \widehat{y_4}}\frac{\partial \widehat{y_4}}{\partial o_4}\frac{\partial o_4}{\partial h_4}\frac{\partial h_4}{\partial h_3}\frac{\partial h_3}{\partial h_2}\frac{\partial h_2}{\partial h_1}\frac{\partial h_1}{\partial W}$$

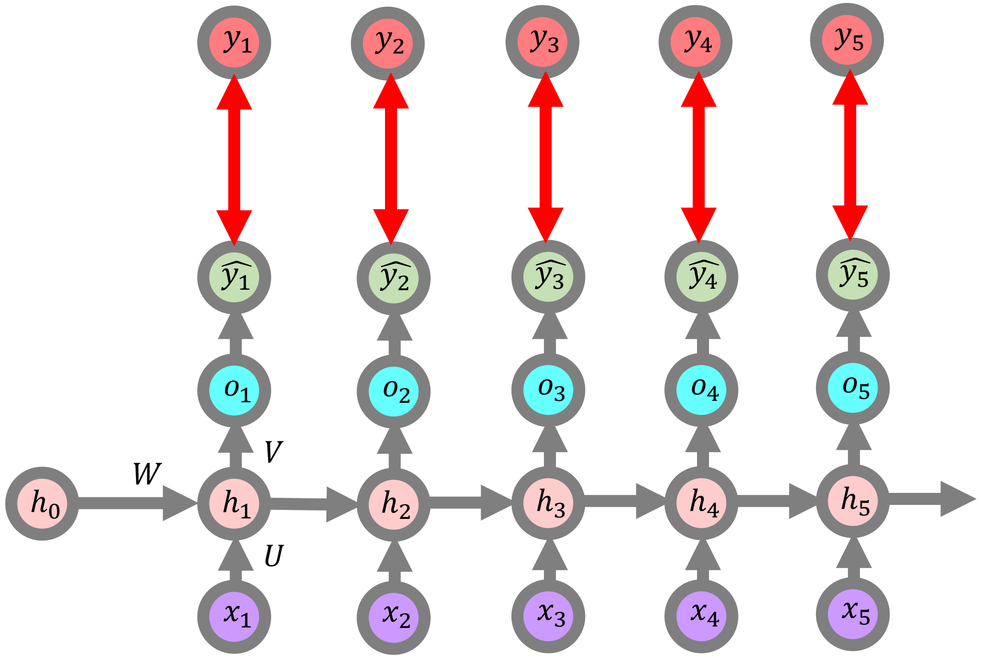


신박AI

여기보시면 시간이 늘어늘수록, 체인룰로 계산해야할 부분이 계속 늘어나는 것을 알 수 있습니다

$$\frac{\partial L_5}{\partial W}=\frac{\partial L_5}{\partial \widehat{y_5}}\frac{\partial \widehat{y_5}}{\partial o_5}\frac{\partial o_5}{\partial h_5}\frac{\partial h_5}{\partial W}+\frac{\partial L_5}{\partial \widehat{y_5}}\frac{\partial \widehat{y_5}}{\partial o_5}\frac{\partial o_5}{\partial h_5}\frac{\partial h_5}{\partial h_4}\frac{\partial h_4}{\partial W}+\frac{\partial L_5}{\partial \widehat{y_5}}\frac{\partial \widehat{y_5}}{\partial o_5}\frac{\partial o_5}{\partial h_5}\frac{\partial h_5}{\partial h_4}\frac{\partial h_4}{\partial h_3}\frac{\partial h_3}{\partial W}$$

$$+\frac{\partial L_5}{\partial \widehat{y_5}}\frac{\partial \widehat{y_5}}{\partial o_5}\frac{\partial o_5}{\partial h_5}\frac{\partial h_5}{\partial h_4}\frac{\partial h_4}{\partial h_3}\frac{\partial h_3}{\partial h_2}\frac{\partial h_2}{\partial W}+\frac{\partial L_5}{\partial \widehat{y_5}}\frac{\partial \widehat{y_5}}{\partial o_5}\frac{\partial o_5}{\partial h_5}\frac{\partial h_5}{\partial h_4}\frac{\partial h_4}{\partial h_3}\frac{\partial h_3}{\partial h_2}\frac{\partial h_2}{\partial h_1}\frac{\partial h_1}{\partial W}$$



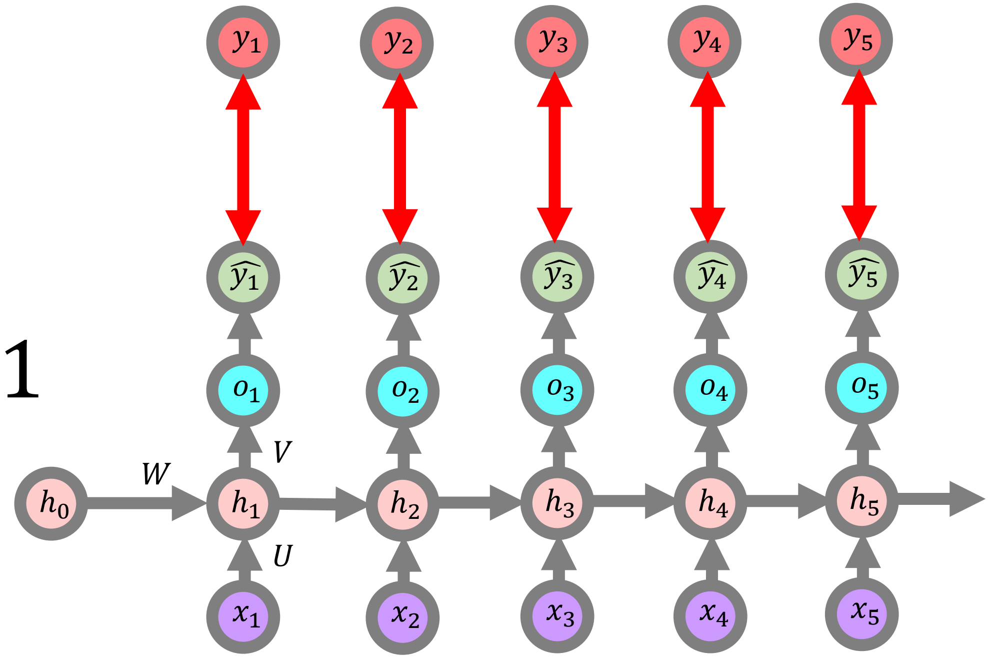
신박AI

만약에 이런 부분들이 1보다 작다면,

$$\frac{\partial L_5}{\partial W}=\frac{\partial L_5}{\partial \widehat{y_5}}\frac{\partial \widehat{y_5}}{\partial o_5}\frac{\partial o_5}{\partial h_5}\frac{\partial h_5}{\partial W}+\frac{\partial L_5}{\partial \widehat{y_5}}\frac{\partial \widehat{y_5}}{\partial o_5}\frac{\partial o_5}{\partial h_5}\frac{\partial h_5}{\partial h_4}\frac{\partial h_4}{\partial W}+\frac{\partial L_5}{\partial \widehat{y_5}}\frac{\partial \widehat{y_5}}{\partial o_5}\frac{\partial o_5}{\partial h_5}\frac{\partial h_5}{\partial h_4}\frac{\partial h_4}{\partial h_3}\frac{\partial h_3}{\partial W}$$

$$+\frac{\partial L_5}{\partial \widehat{y_5}}\frac{\partial \widehat{y_5}}{\partial o_5}\frac{\partial o_5}{\partial h_5}\frac{\partial h_5}{\partial h_4}\frac{\partial h_4}{\partial h_3}\frac{\partial h_3}{\partial h_2}\frac{\partial h_2}{\partial W}+\frac{\partial L_5}{\partial \widehat{y_5}}\frac{\partial \widehat{y_5}}{\partial o_5}\frac{\partial o_5}{\partial h_5}\frac{\partial h_5}{\partial h_4}\frac{\partial h_4}{\partial h_3}\frac{\partial h_3}{\partial h_2}\frac{\partial h_2}{\partial h_1}\frac{\partial h_1}{\partial W}$$

< 1



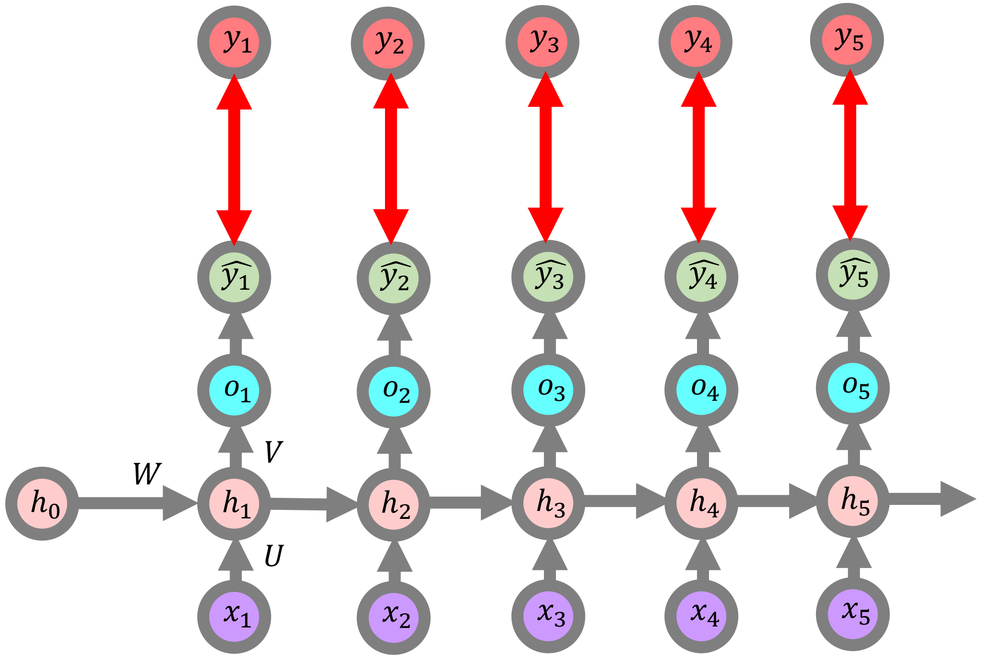
신박AI

이렇게 계속 체인룰로 곱해 나가면 멀리있는 부분의 기울기 값이 작아지게 됩니다

$$\frac{\partial L_5}{\partial W}=\frac{\partial L_5}{\partial \widehat{y_5}}\frac{\partial \widehat{y_5}}{\partial o_5}\frac{\partial o_5}{\partial h_5}\frac{\partial h_5}{\partial W}+\frac{\partial L_5}{\partial \widehat{y_5}}\frac{\partial \widehat{y_5}}{\partial o_5}\frac{\partial o_5}{\partial h_5}\frac{\partial h_5}{\partial h_4}\frac{\partial h_4}{\partial W}+\frac{\partial L_5}{\partial \widehat{y_5}}\frac{\partial \widehat{y_5}}{\partial o_5}\frac{\partial o_5}{\partial h_5}\frac{\partial h_5}{\partial h_4}\frac{\partial h_4}{\partial h_3}\frac{\partial h_3}{\partial W}$$

$$+\frac{\partial L_5}{\partial \widehat{y_5}}\frac{\partial \widehat{y_5}}{\partial o_5}\frac{\partial o_5}{\partial h_5}\frac{\partial h_5}{\partial h_4}\frac{\partial h_4}{\partial h_3}\frac{\partial h_3}{\partial h_2}\frac{\partial h_2}{\partial W}+\boxed{\frac{\partial L_5}{\partial \widehat{y_5}}\frac{\partial \widehat{y_5}}{\partial o_5}\frac{\partial o_5}{\partial h_5}\frac{\partial h_5}{\partial h_4}\frac{\partial h_4}{\partial h_3}\frac{\partial h_3}{\partial h_2}\frac{\partial h_2}{\partial h_1}\frac{\partial h_1}{\partial W}}$$

예를 들자면, 0.1x0.3x0.2x0.1 = 0.0006

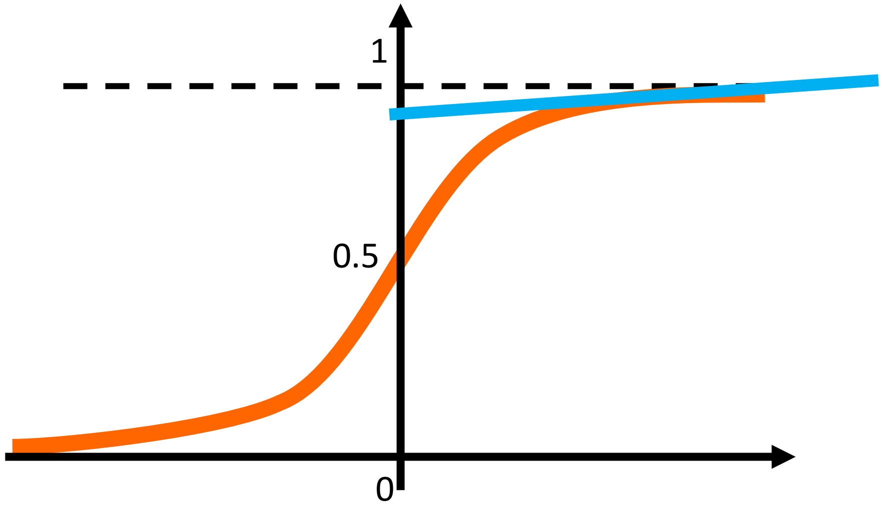


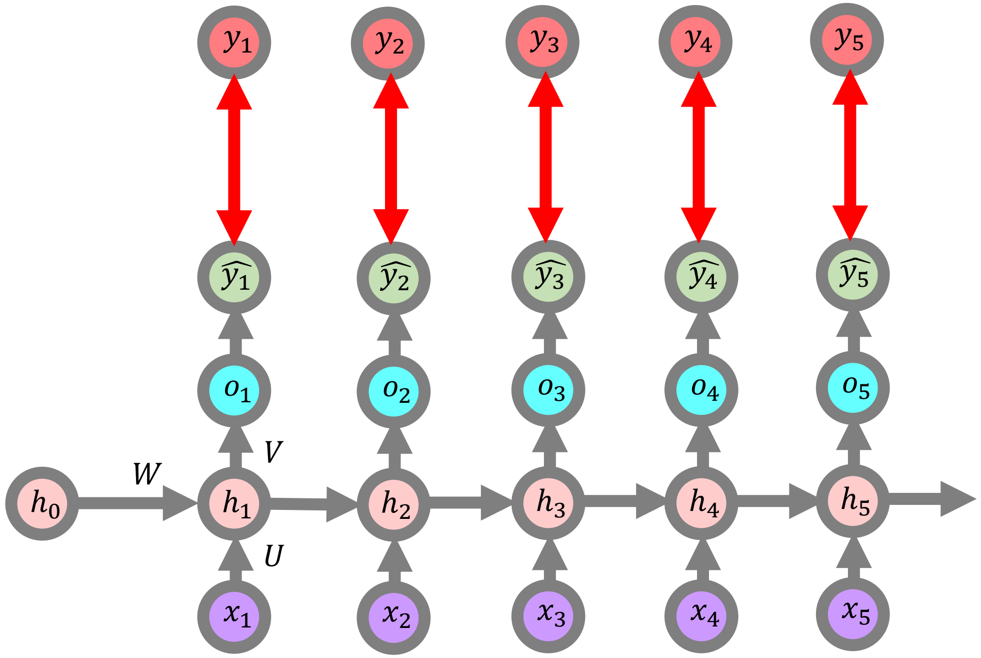
기울기가 작다는 것은 학습에 미치는 영향이 미미하다는 것을 뜻하고,

$$\frac{\partial L_5}{\partial W}=\frac{\partial L_5}{\partial \widehat{y_5}}\frac{\partial \widehat{y_5}}{\partial o_5}\frac{\partial o_5}{\partial h_5}\frac{\partial h_5}{\partial W}+\frac{\partial L_5}{\partial \widehat{y_5}}\frac{\partial \widehat{y_5}}{\partial o_5}\frac{\partial o_5}{\partial h_5}\frac{\partial h_5}{\partial h_4}\frac{\partial h_4}{\partial W}+\frac{\partial L_5}{\partial \widehat{y_5}}\frac{\partial \widehat{y_5}}{\partial o_5}\frac{\partial o_5}{\partial h_5}\frac{\partial h_5}{\partial h_4}\frac{\partial h_4}{\partial h_3}\frac{\partial h_3}{\partial W}$$

$$+\frac{\partial L_5}{\partial \widehat{y_5}}\frac{\partial \widehat{y_5}}{\partial o_5}\frac{\partial o_5}{\partial h_5}\frac{\partial h_5}{\partial h_4}\frac{\partial h_4}{\partial h_3}\frac{\partial h_3}{\partial h_2}\frac{\partial h_2}{\partial W}+\boxed{\frac{\partial L_5}{\partial \widehat{y_5}}\frac{\partial \widehat{y_5}}{\partial o_5}\frac{\partial o_5}{\partial h_5}\frac{\partial h_5}{\partial h_4}\frac{\partial h_4}{\partial h_3}\frac{\partial h_3}{\partial h_2}\frac{\partial h_2}{\partial h_1}\frac{\partial h_1}{\partial W}}$$

예를 들자면, 0.1x0.3x0.2x0.1 = 0.0006





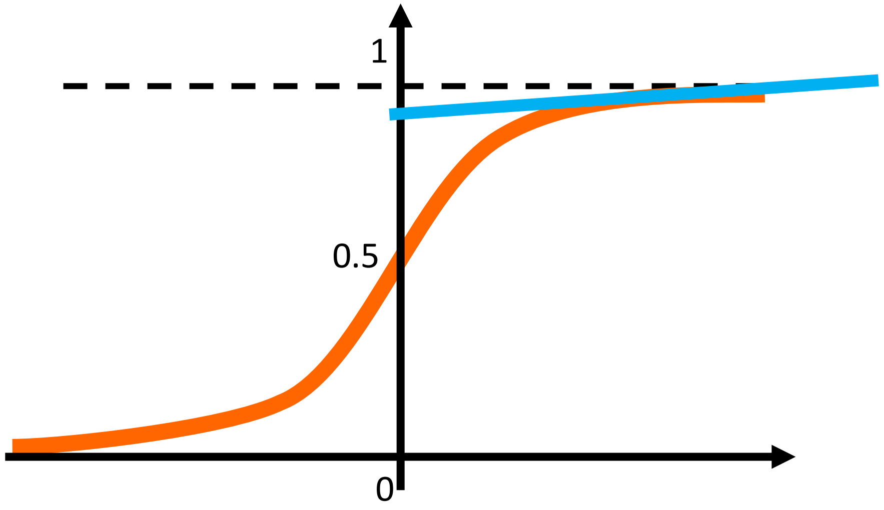
신박AI

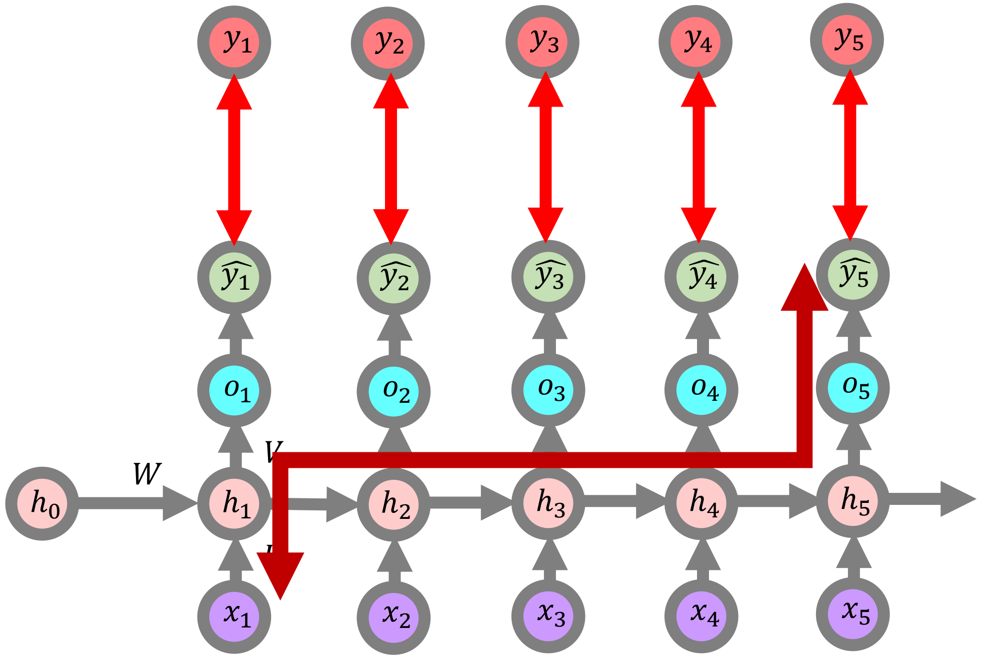
결과적으로는, 시간적으로 먼 입력값 일수록 학습에 미치는 영향도 미미하다는 것을 뜻합니다

$$\frac{\partial L_5}{\partial W}=\frac{\partial L_5}{\partial \widehat{y_5}}\frac{\partial \widehat{y_5}}{\partial o_5}\frac{\partial o_5}{\partial h_5}\frac{\partial h_5}{\partial W}+\frac{\partial L_5}{\partial \widehat{y_5}}\frac{\partial \widehat{y_5}}{\partial o_5}\frac{\partial o_5}{\partial h_5}\frac{\partial h_5}{\partial h_4}\frac{\partial h_4}{\partial W}+\frac{\partial L_5}{\partial \widehat{y_5}}\frac{\partial \widehat{y_5}}{\partial o_5}\frac{\partial o_5}{\partial h_5}\frac{\partial h_5}{\partial h_4}\frac{\partial h_4}{\partial h_3}\frac{\partial h_3}{\partial W}$$

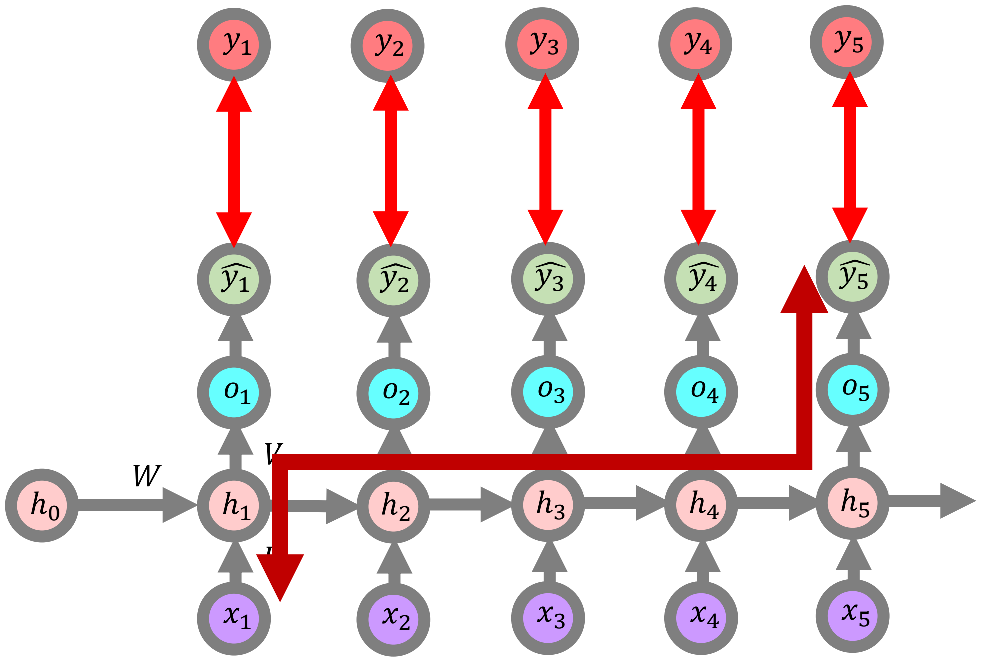
$$+\frac{\partial L_5}{\partial \widehat{y_5}}\frac{\partial \widehat{y_5}}{\partial o_5}\frac{\partial o_5}{\partial h_5}\frac{\partial h_5}{\partial h_4}\frac{\partial h_4}{\partial h_3}\frac{\partial h_3}{\partial h_2}\frac{\partial h_2}{\partial W}+\boxed{\frac{\partial L_5}{\partial \widehat{y_5}}\frac{\partial \widehat{y_5}}{\partial o_5}\frac{\partial o_5}{\partial h_5}\frac{\partial h_5}{\partial h_4}\frac{\partial h_4}{\partial h_3}\frac{\partial h_3}{\partial h_2}\frac{\partial h_2}{\partial h_1}\frac{\partial h_1}{\partial W}}$$

예를 들자면, 0.1x0.3x0.2x0.1 = 0.0006



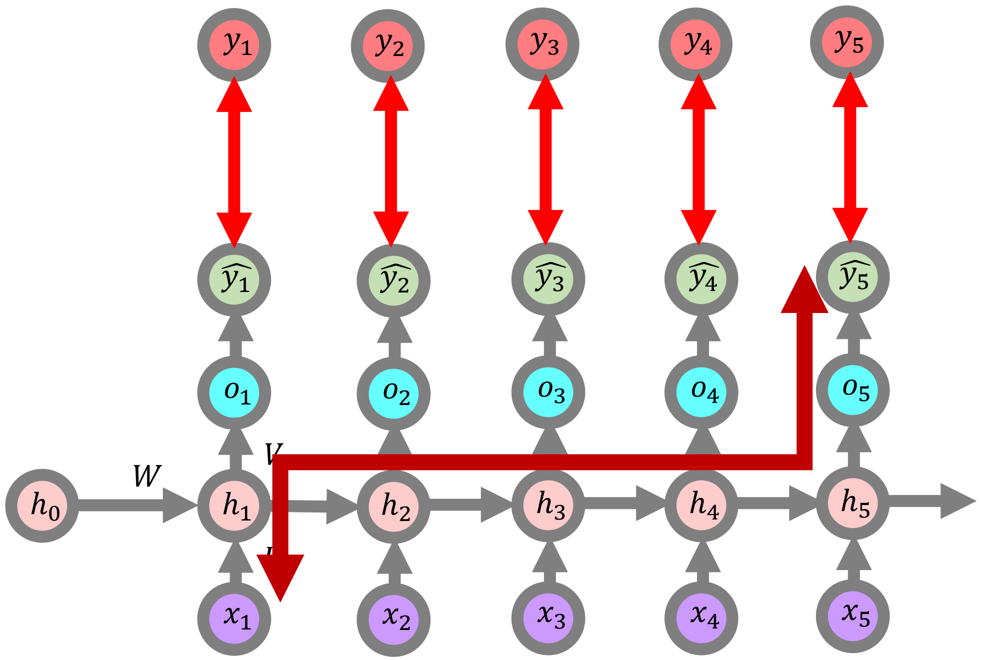


그럴경우, 다음과 같은 기계번역을 학습하는데 문제가 발생할 수 있습니다



예를들어, 다음 영어 문장을 한국어로 번역한다고 가정해 봅시다

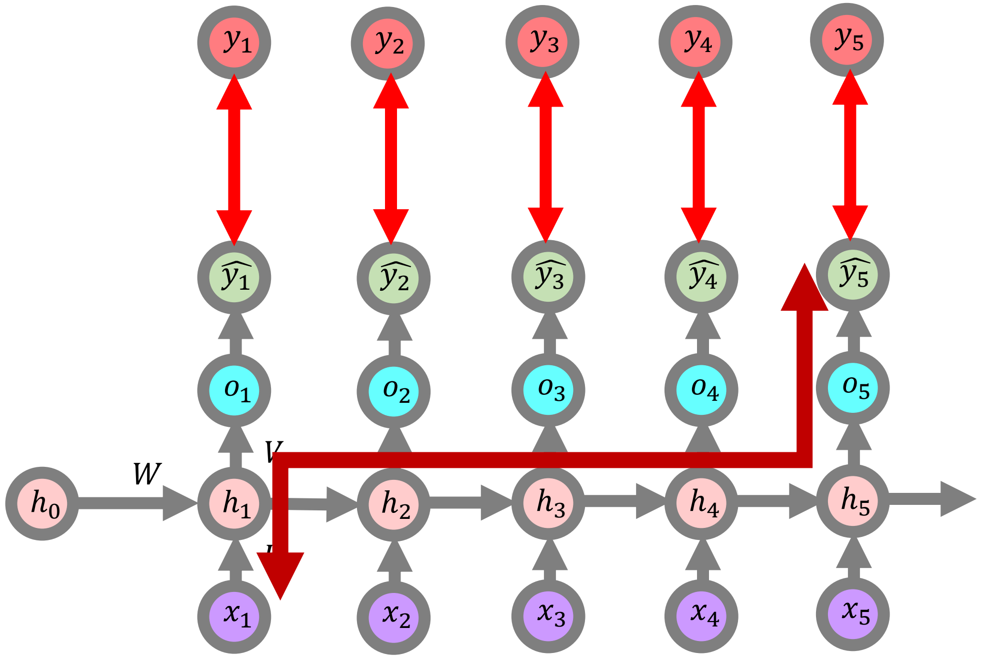
Don't underestimate your inner strength.



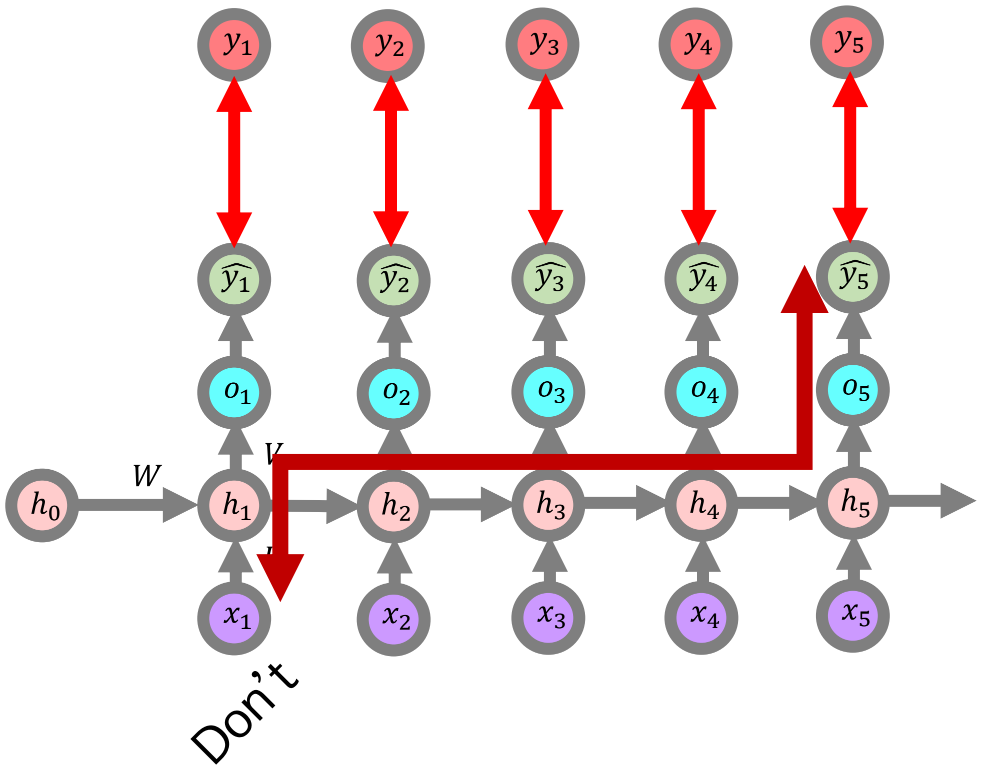
예를들어, 다음 영어 문장을 한국어로 번역한다고 가정해 봅시다

Don’t underestimate your inner strength

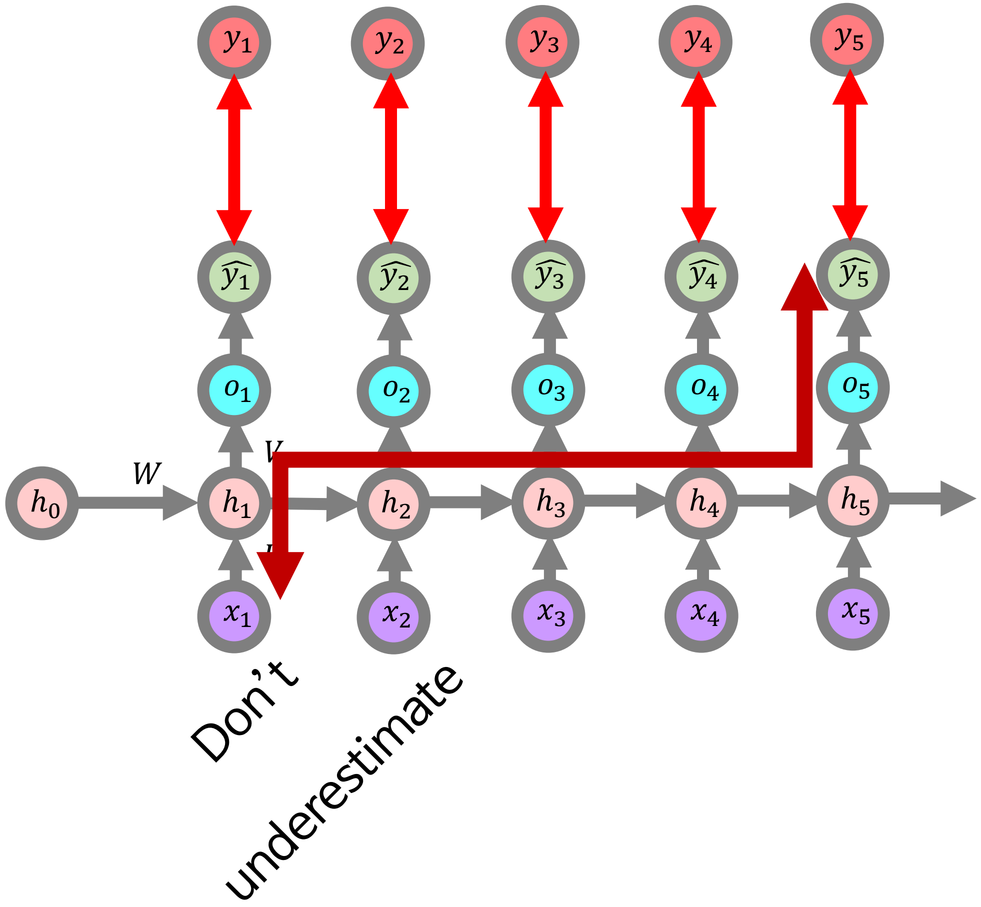
당신의 내면의 힘을 과소평가하지 마세요



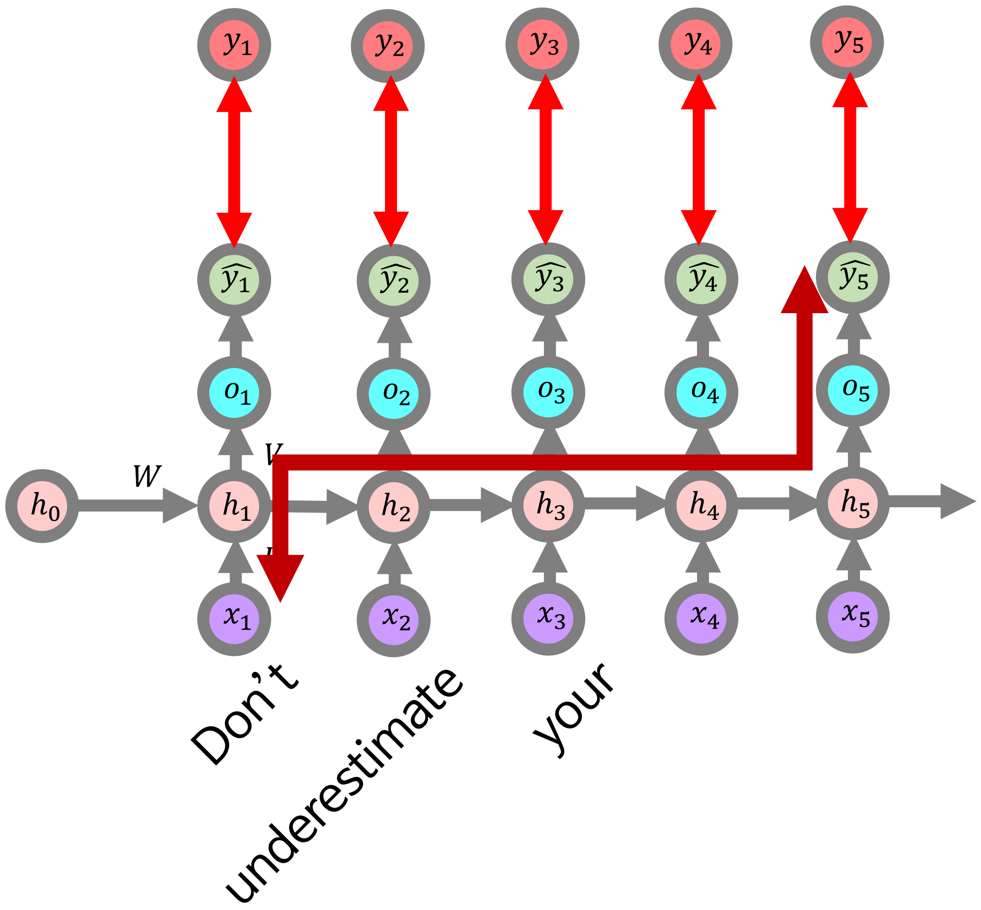
그러면 다음과 같이 RNN에 영어 단어를 입력할 수 있습니다



그러면 다음과 같이 RNN에 영어 단어를 입력할 수 있습니다

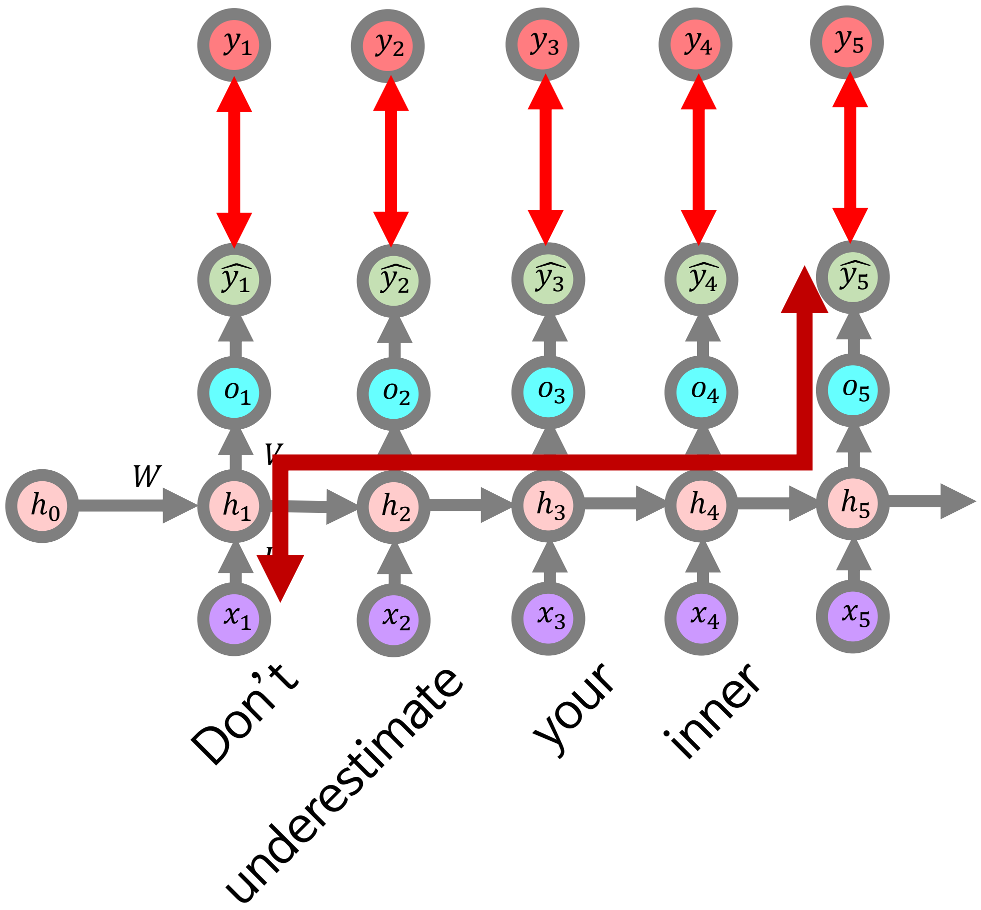


그러면 다음과 같이 RNN에 영어 단어를 입력할 수 있습니다

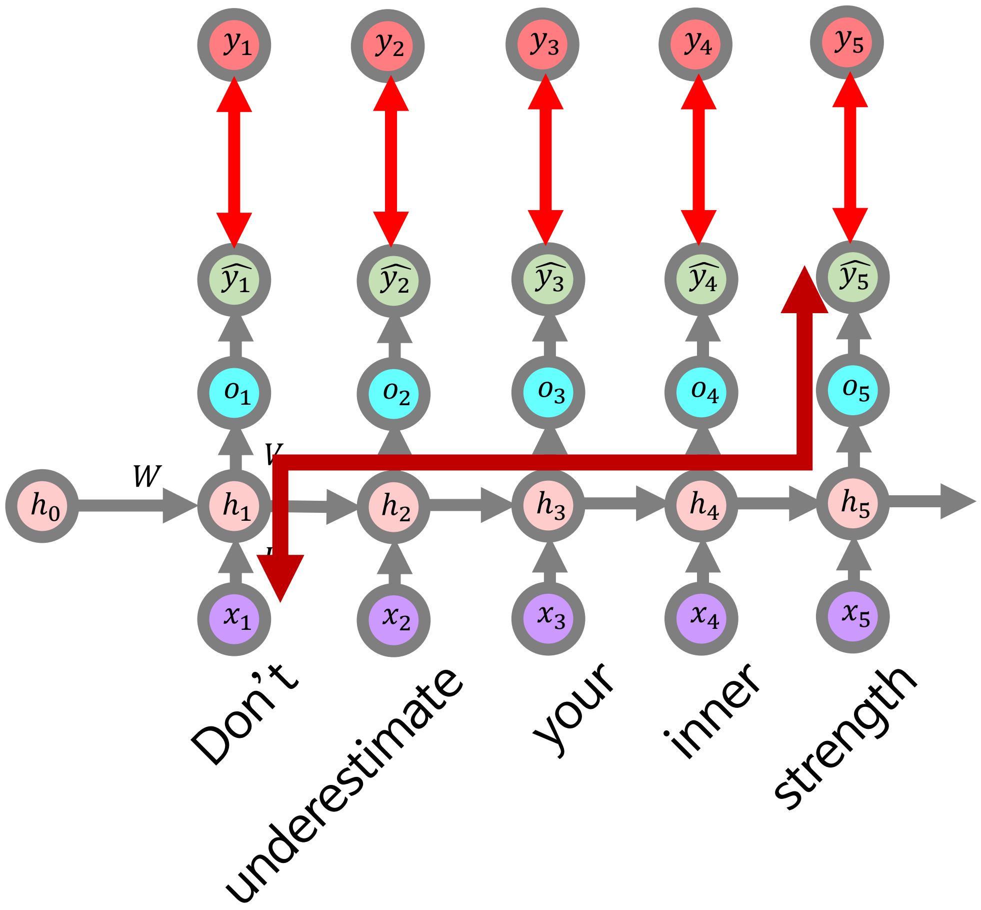


신박AI

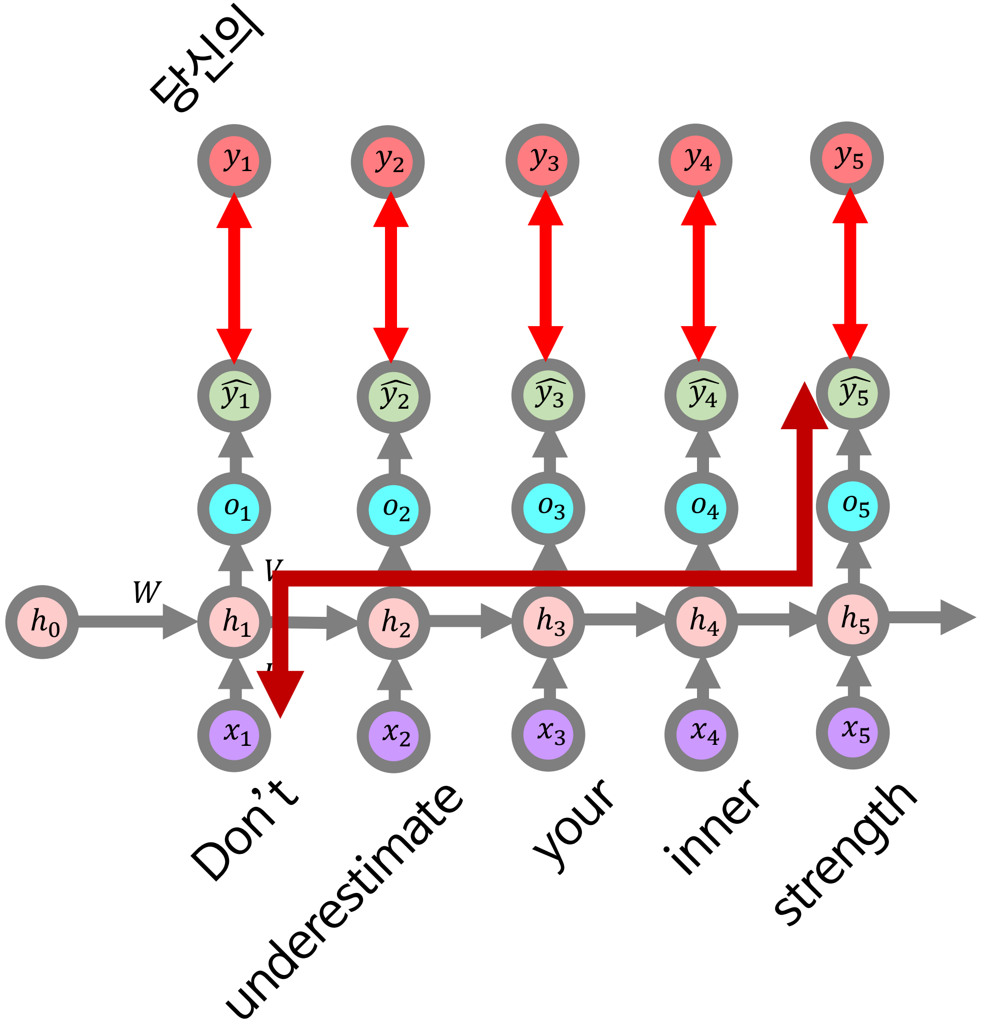
그러면 다음과 같이 RNN에 영어 단어를 입력할 수 있습니다



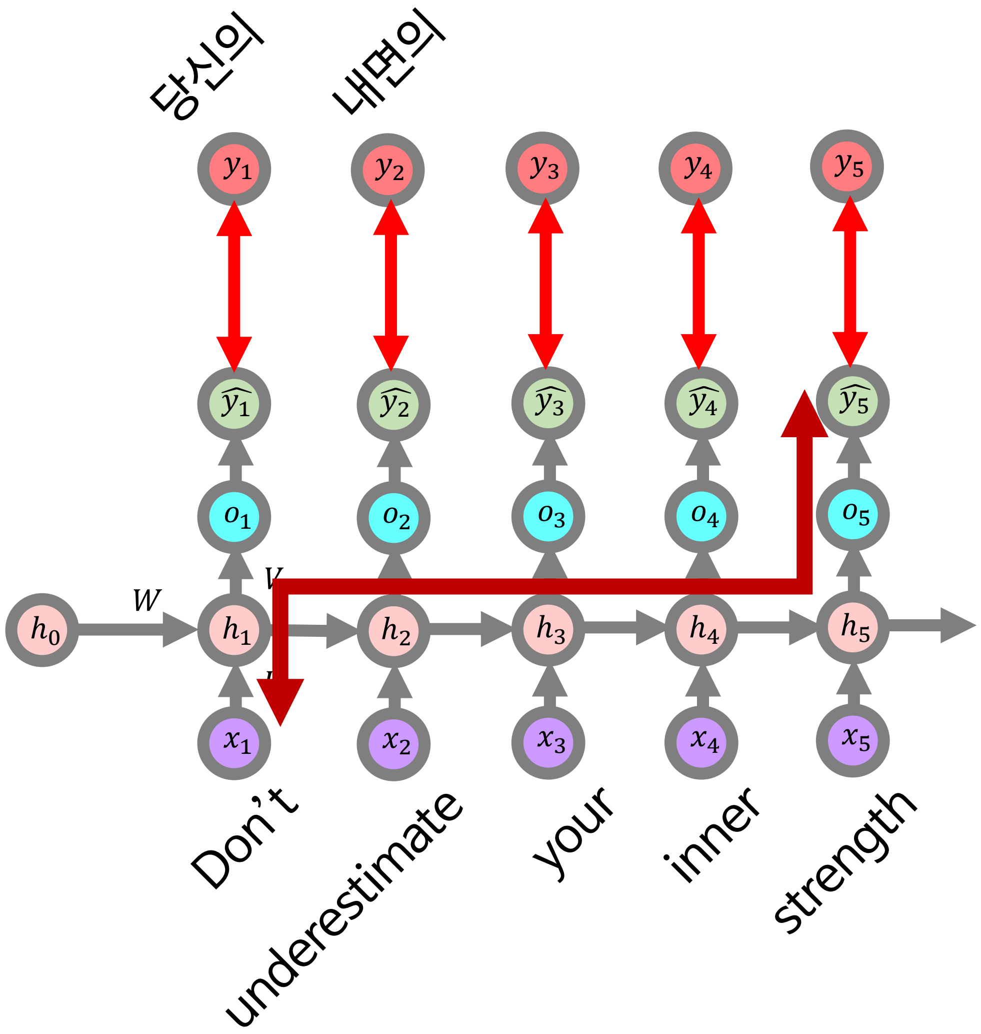
그러면 다음과 같이 RNN에 영어 단어를 입력할 수 있습니다



그러면 다음과 같이 한국어 단어로 번역이 됩니다

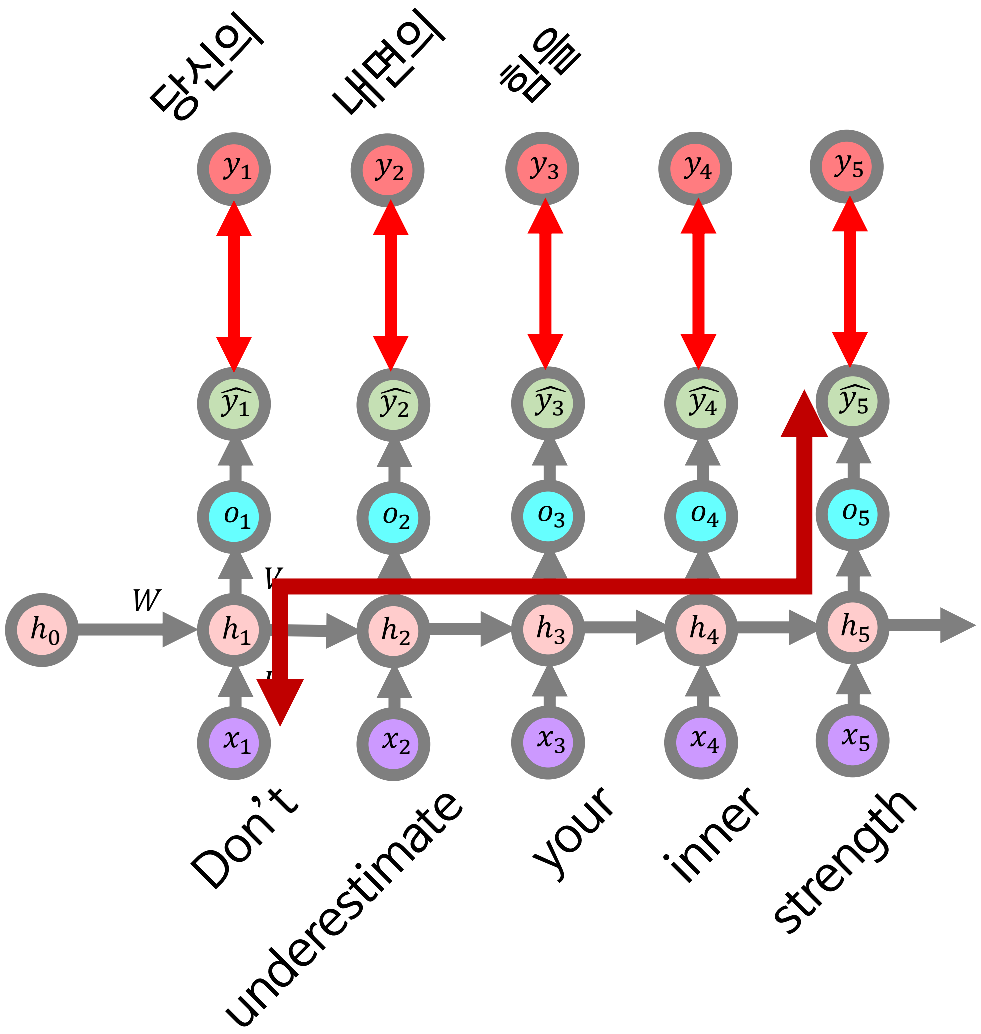


그러면 다음과 같이 한국어 단어로 번역이 됩니다

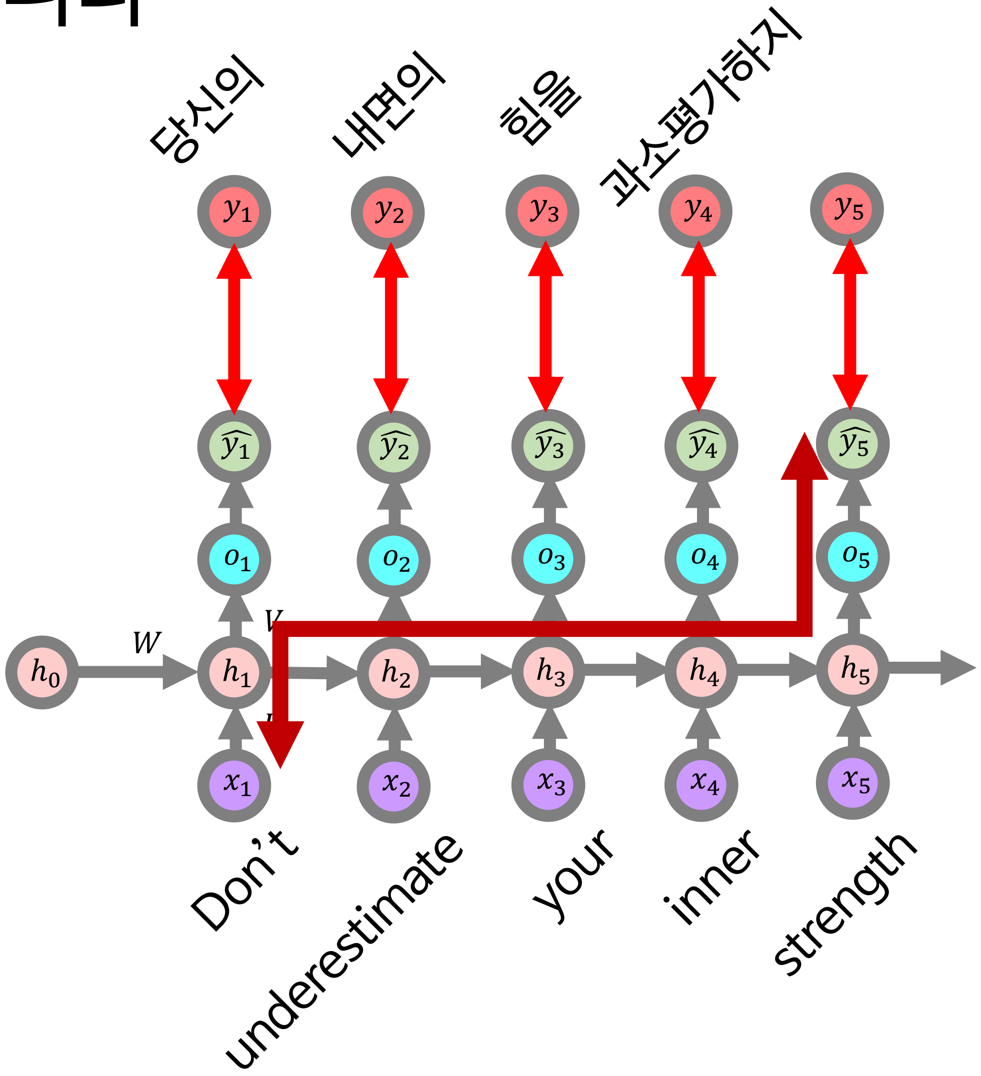


신박AI

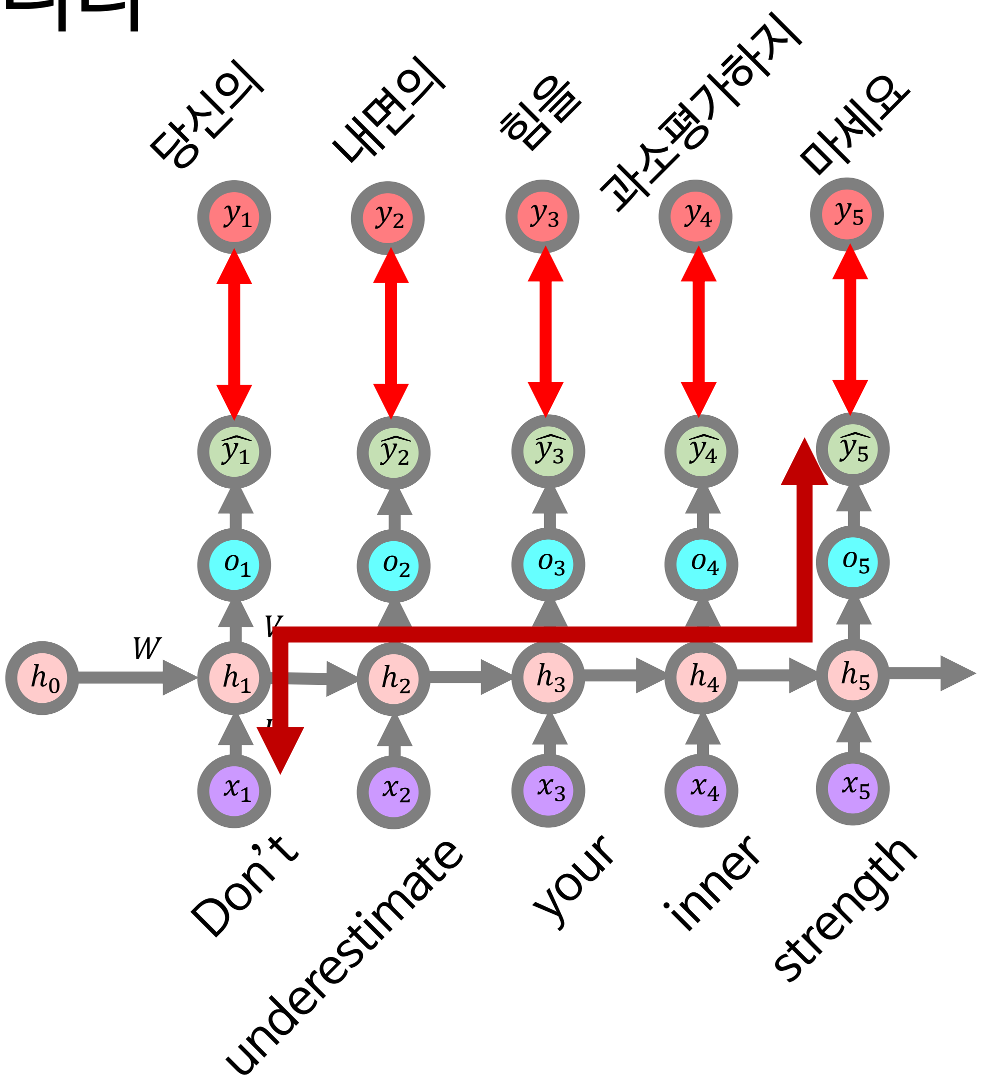
그러면 다음과 같이 한국어 단어로 번역이 됩니다



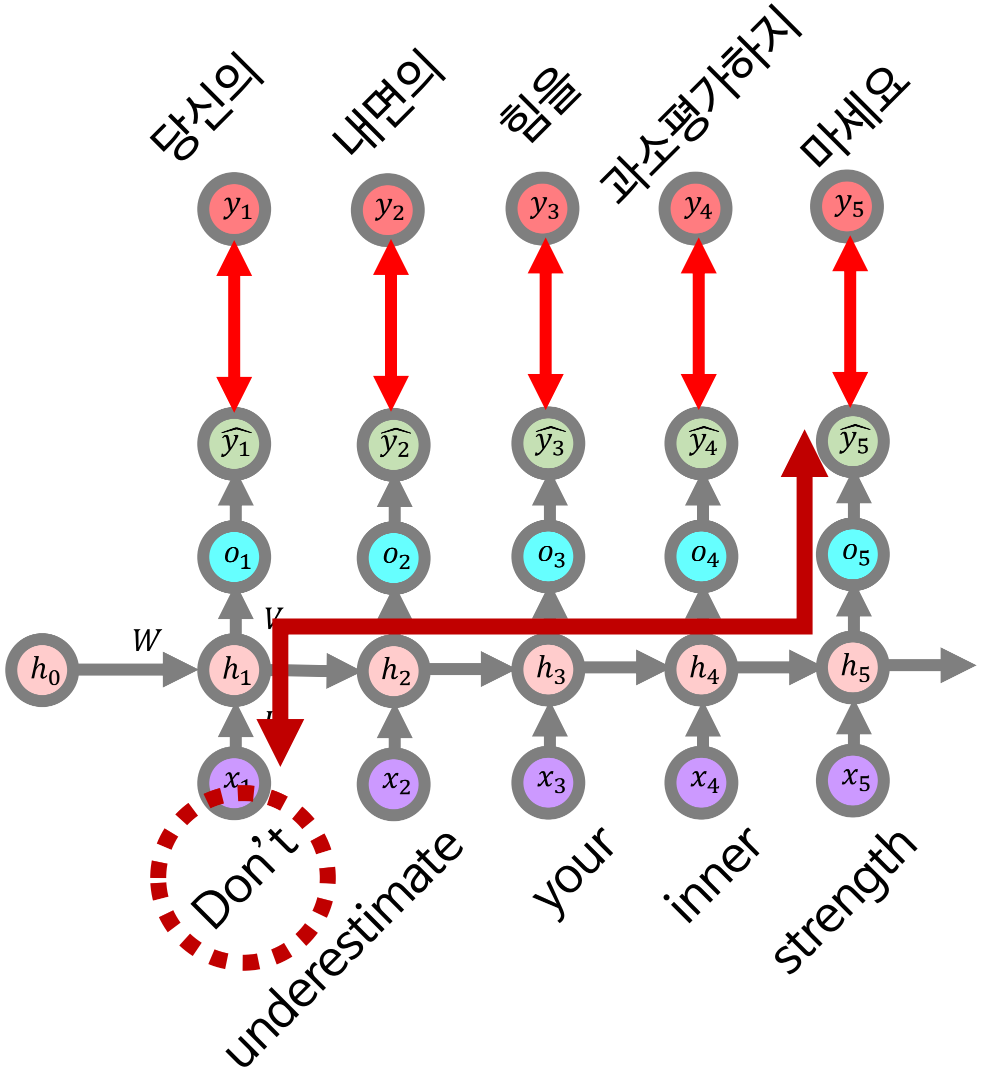
그러면 다음과 같이 한국어 단어로 번역이 됩니다



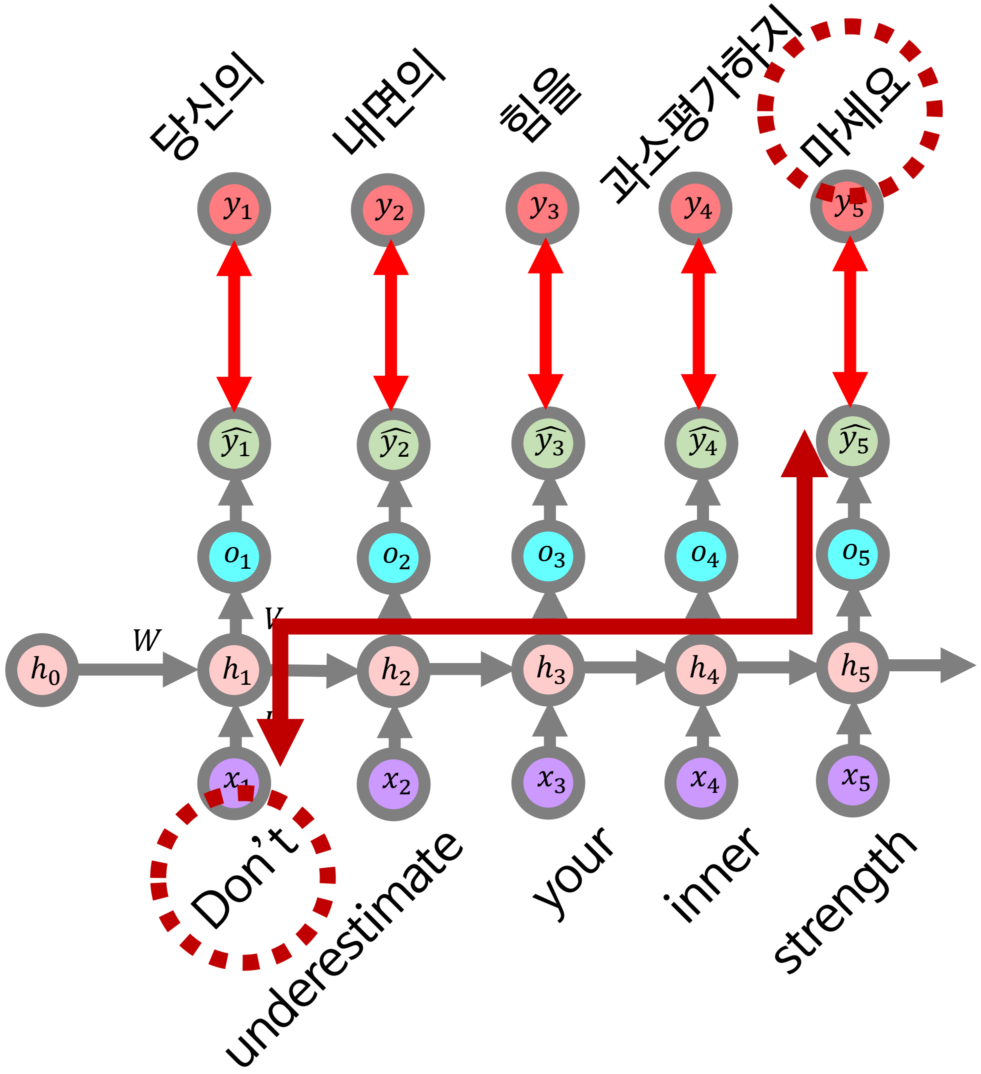
그러면 다음과 같이 한국어 단어로 번역이 됩니다



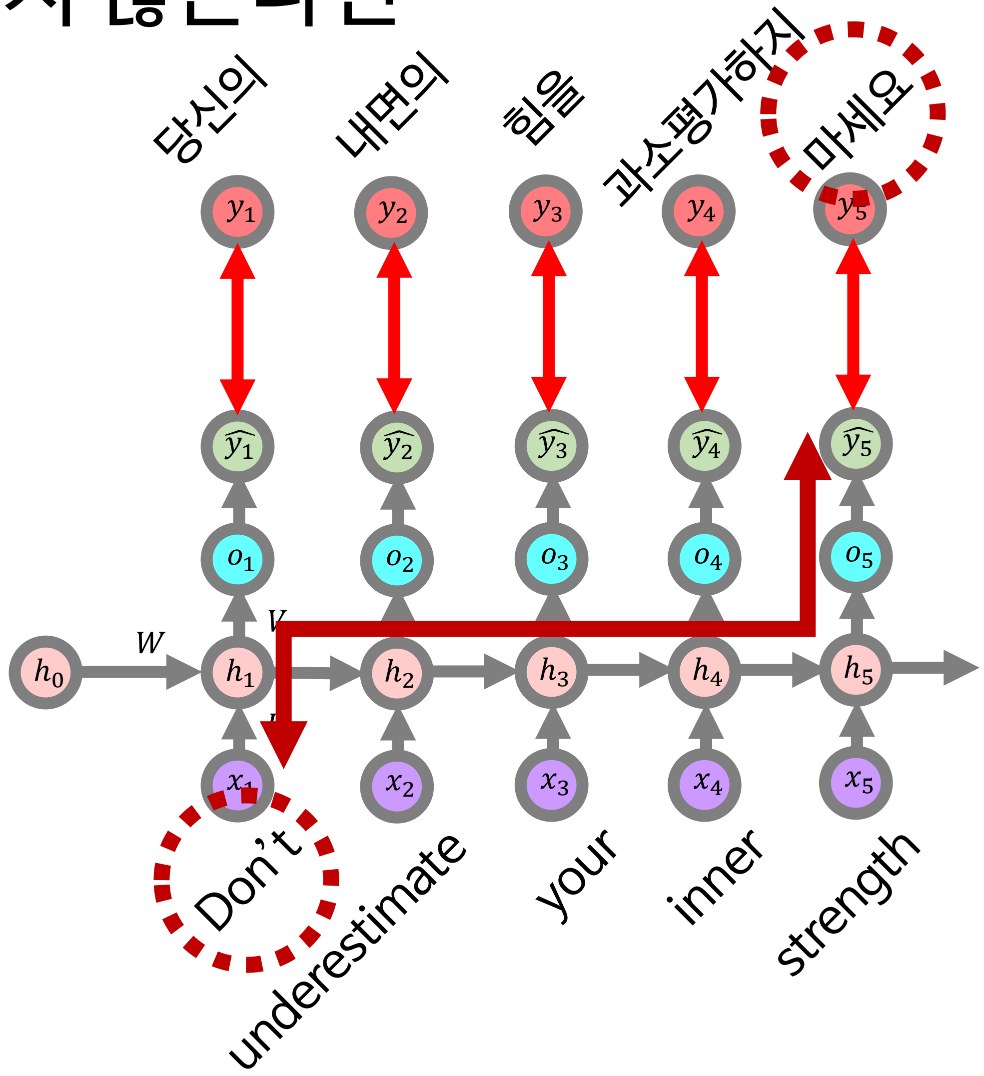
자 이와 같은 경우, “Don’t”와 “마세요”는



의미적으로 상당히 가까운 단어들입니다

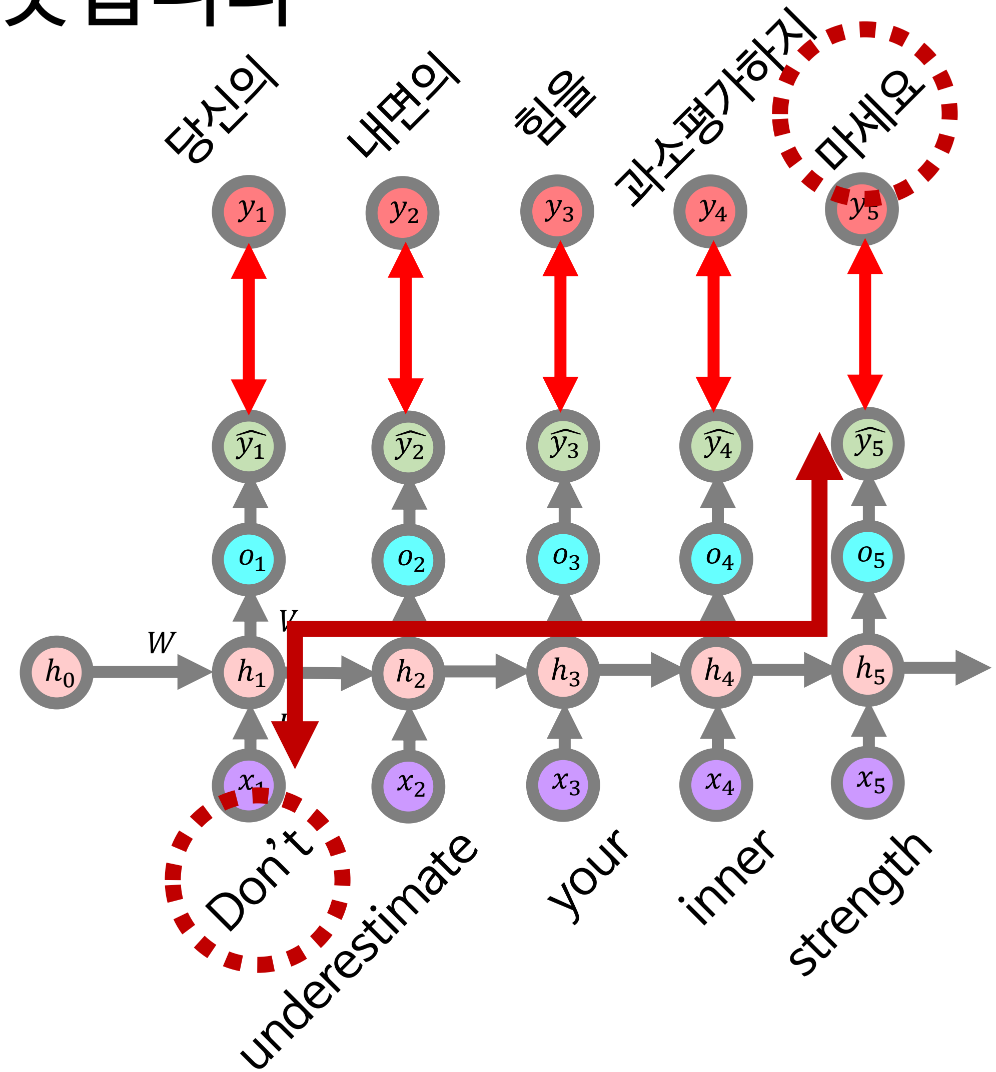


그런데 이 두 단어의 관련성이 학습에 반영되지 않는다면



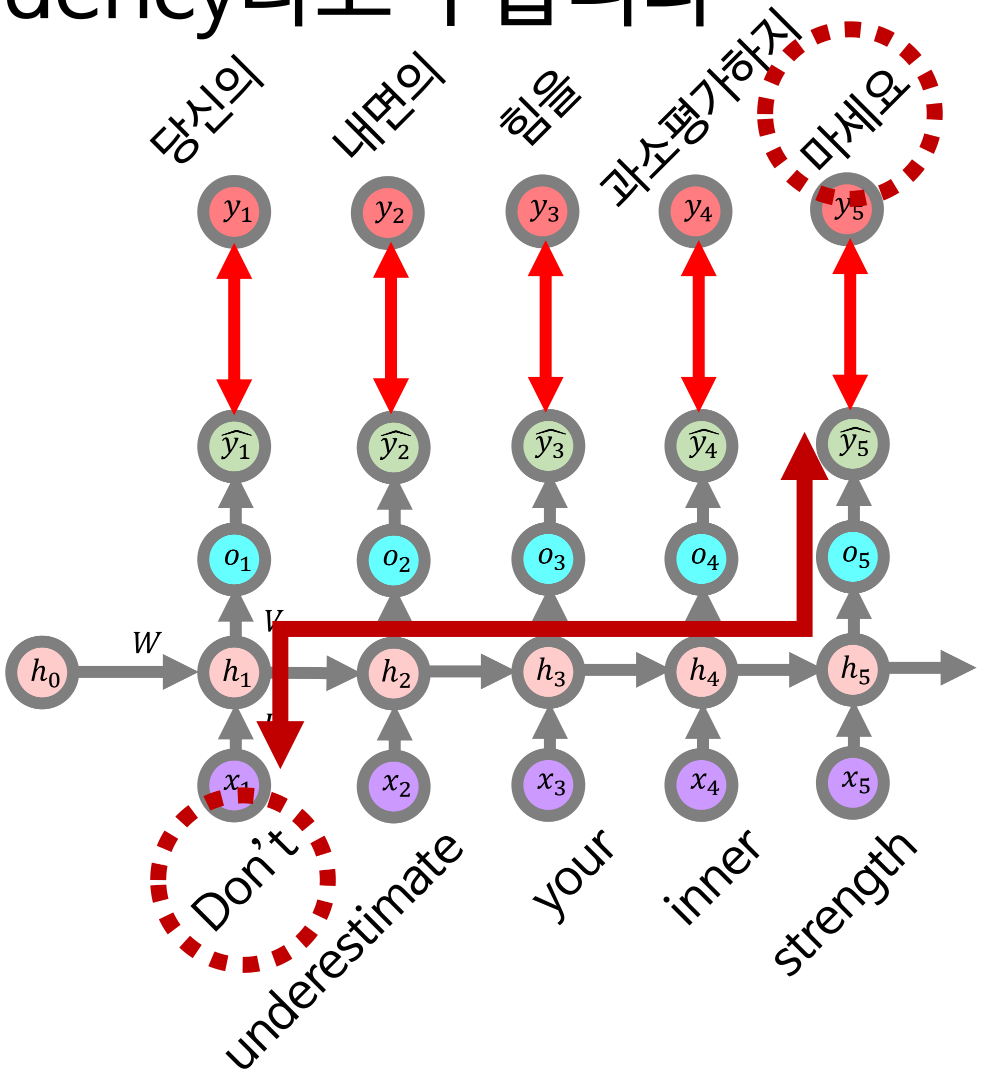
신박AI

결과적으로 기계번역의 정확성은 높지 않을 것입니다



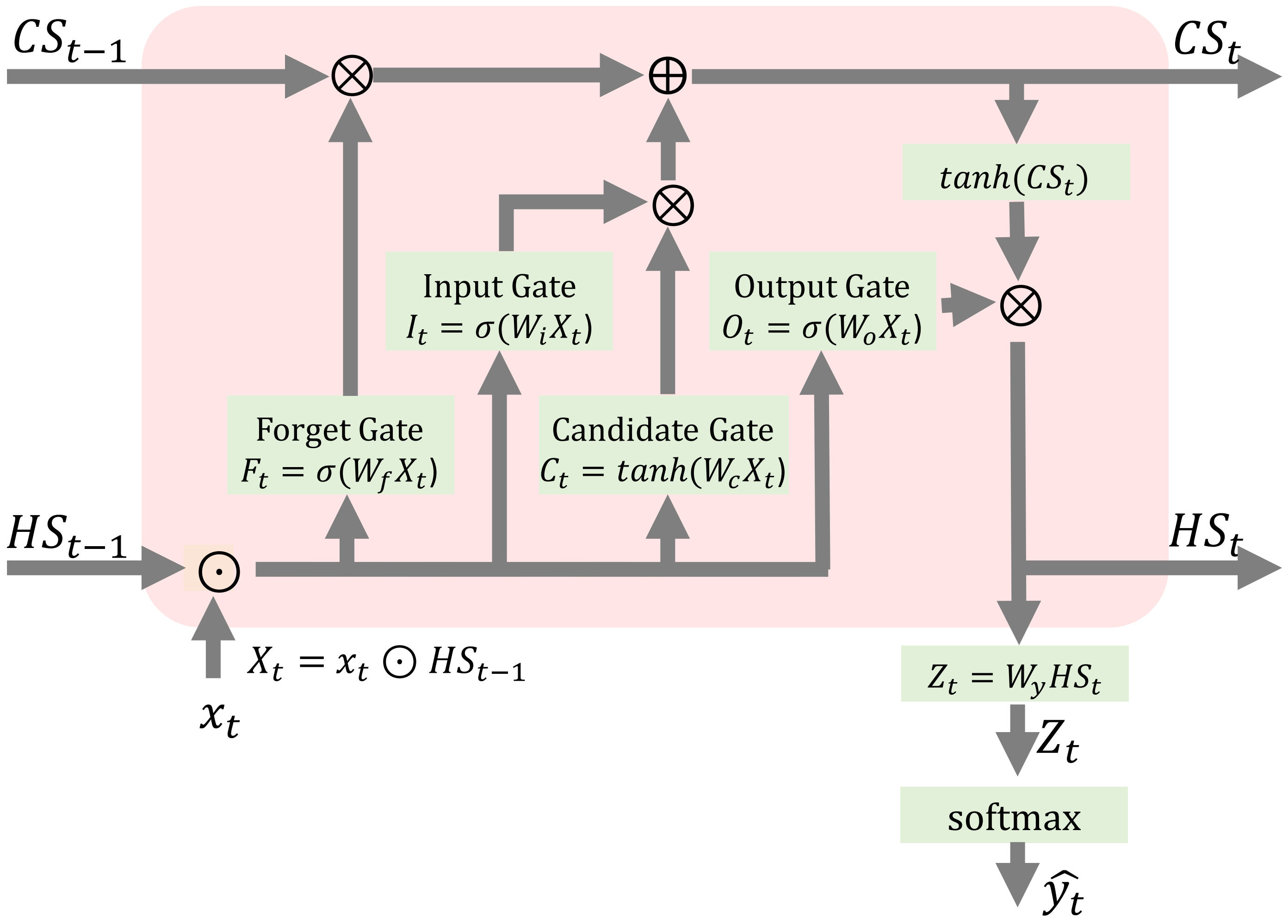
신박AI

이런 현상을 장기의존성 long-term dependency라고 부릅니다

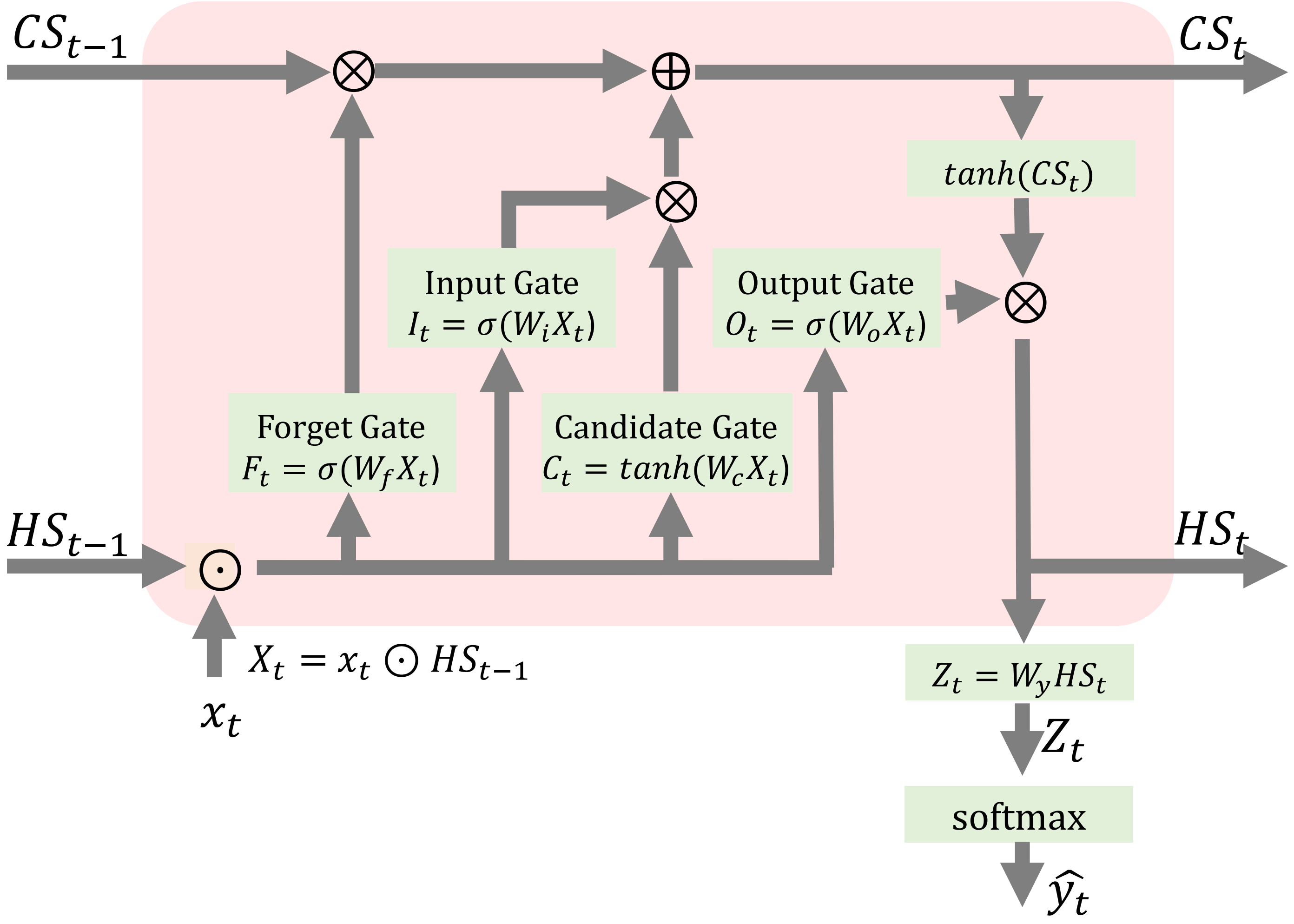


그래서 이러한 RNN의 약점을 극복하기 위해

LSTM이라는 신경망이 개발된 것입니다

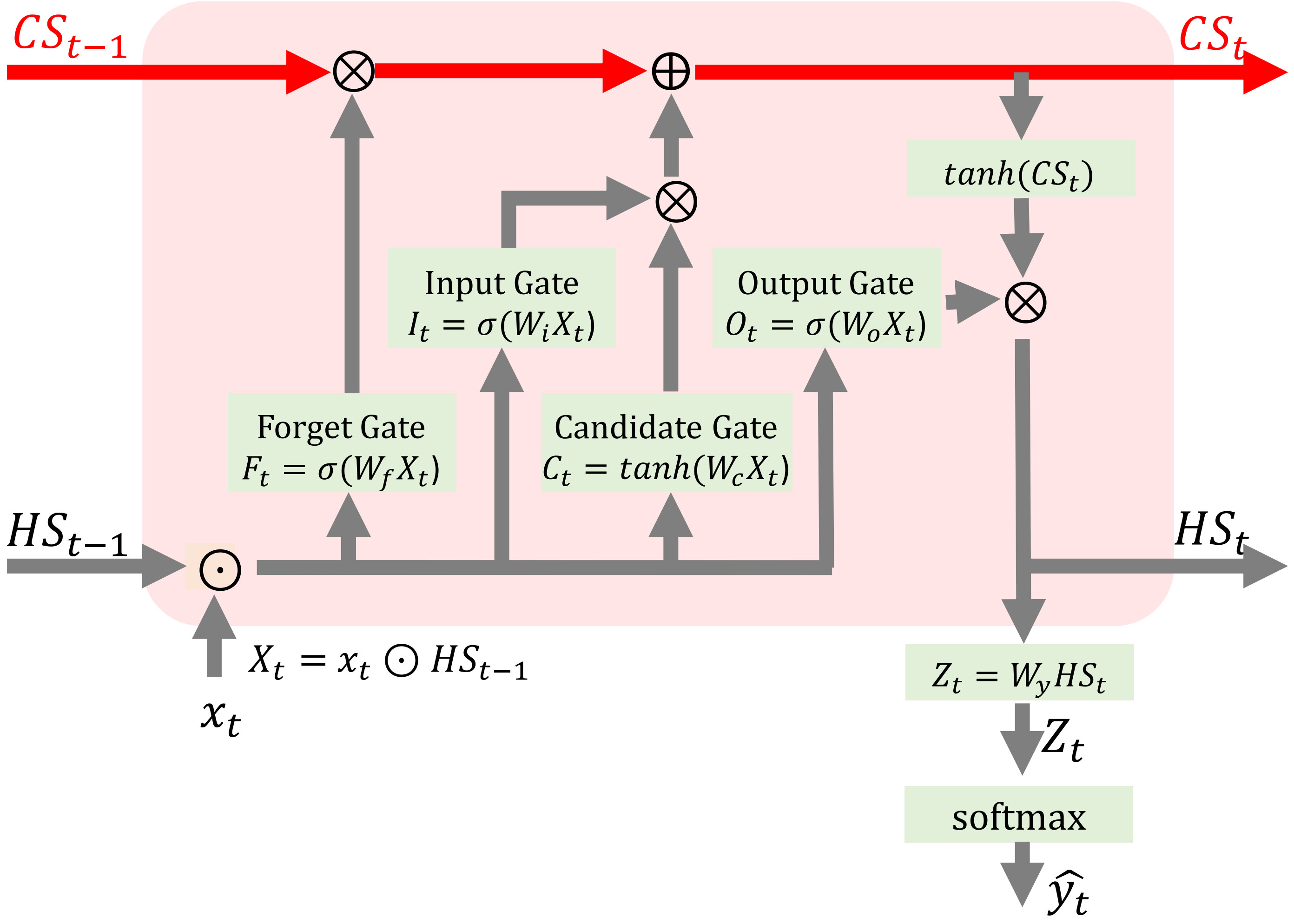


그러면 LSTM은 어떻게 장기의존성 문제를 극복하는 것일까요?



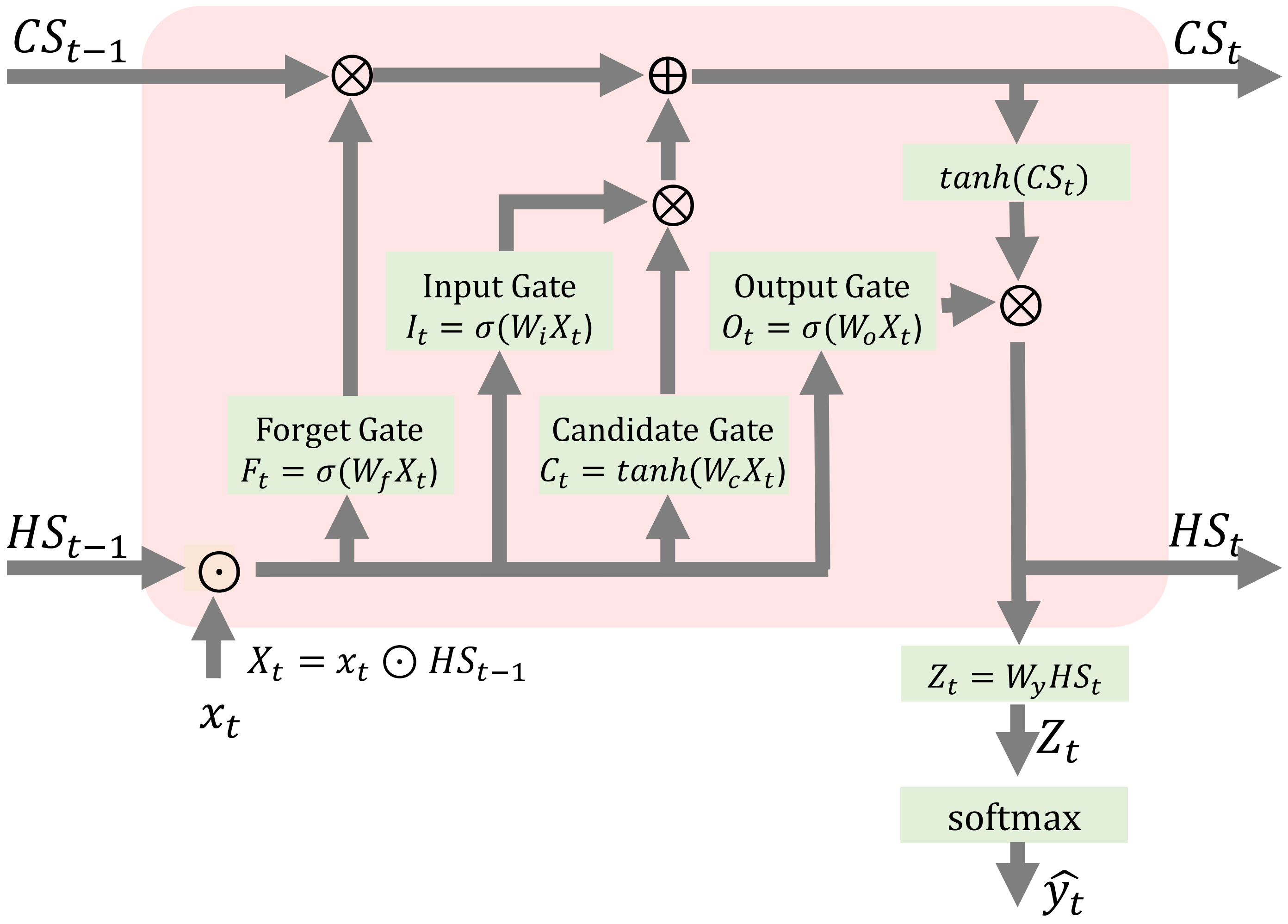
신박AI

그 비밀은 바로 셀 상태 (Cell State, CS)라 불리는 정보에 있습니다

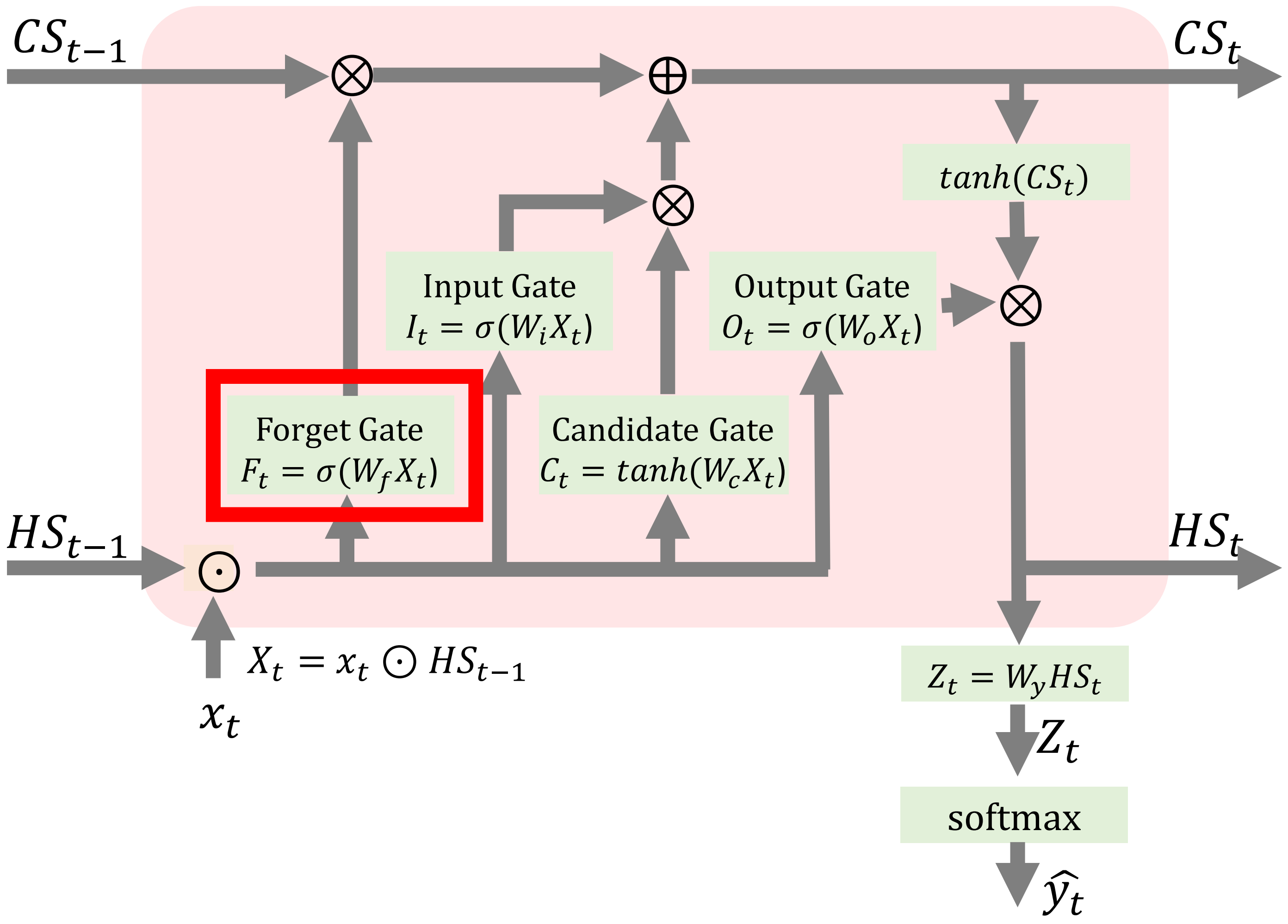


신박AI

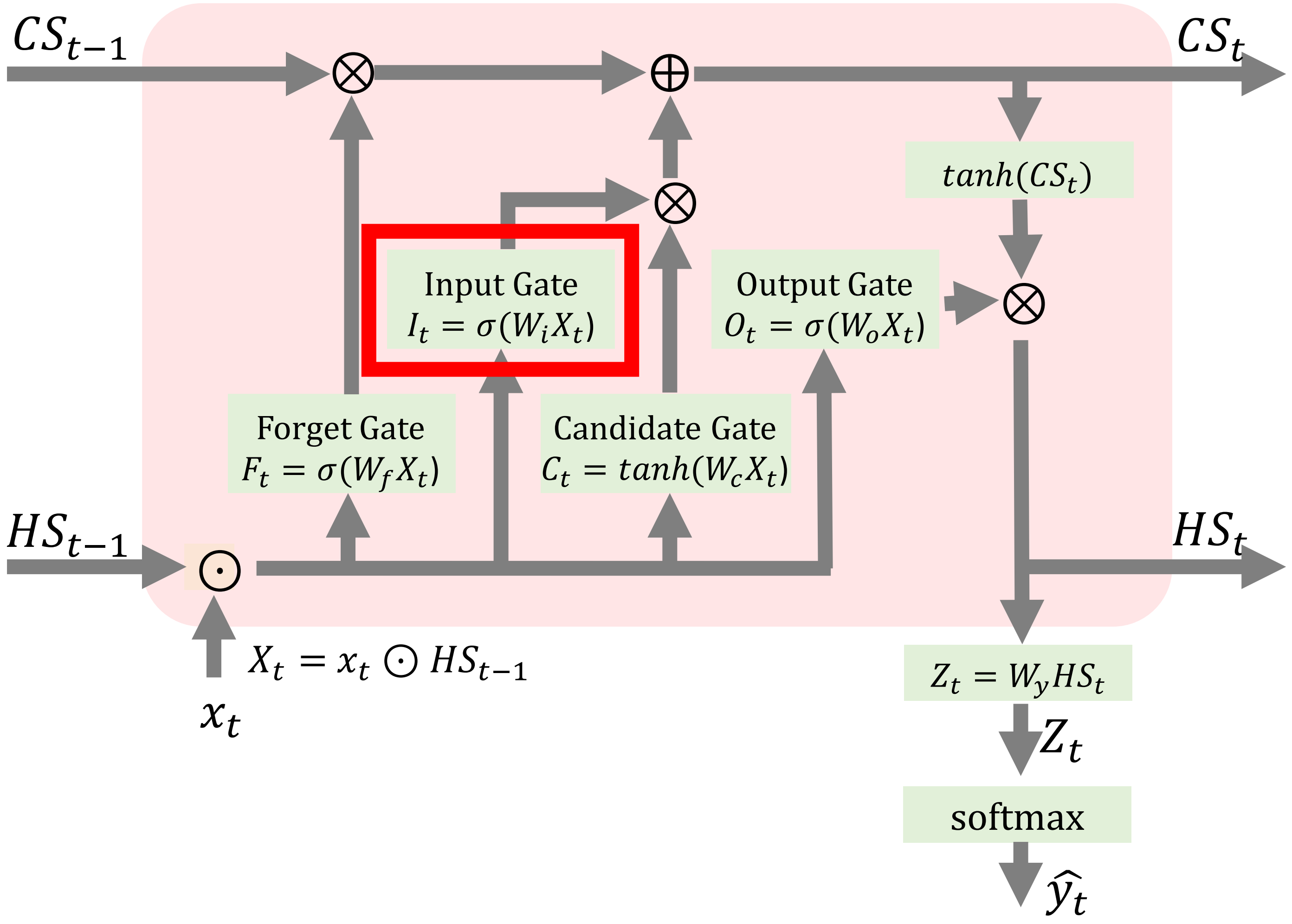
그리고 LSTM에는 RNN과 다른 4개의 게이트가 있습니다



그리고 LSTM에는 RNN과 다른 4개의 게이트가 있습니다

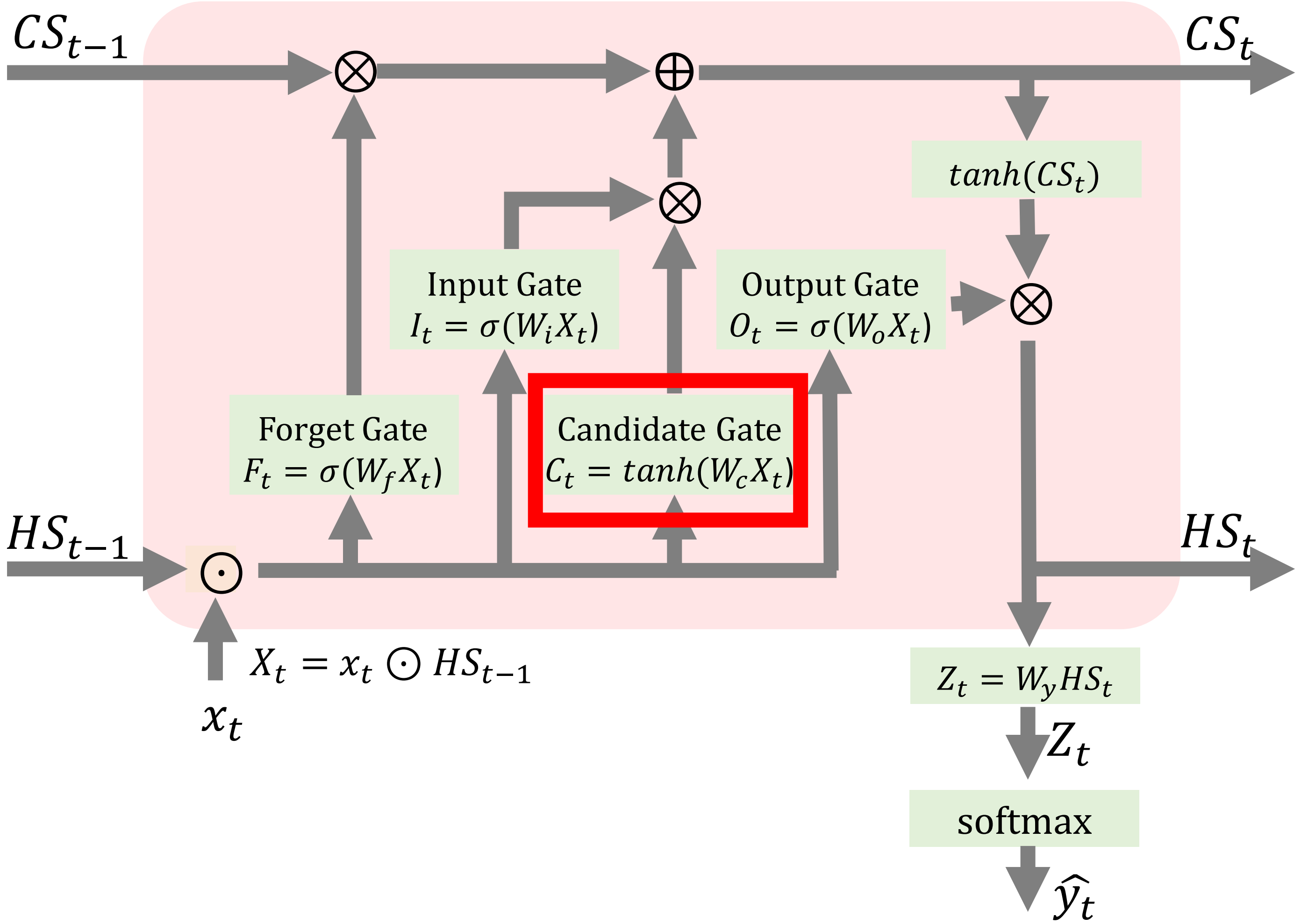


그리고 LSTM에는 RNN과 다른 4개의 게이트가 있습니다

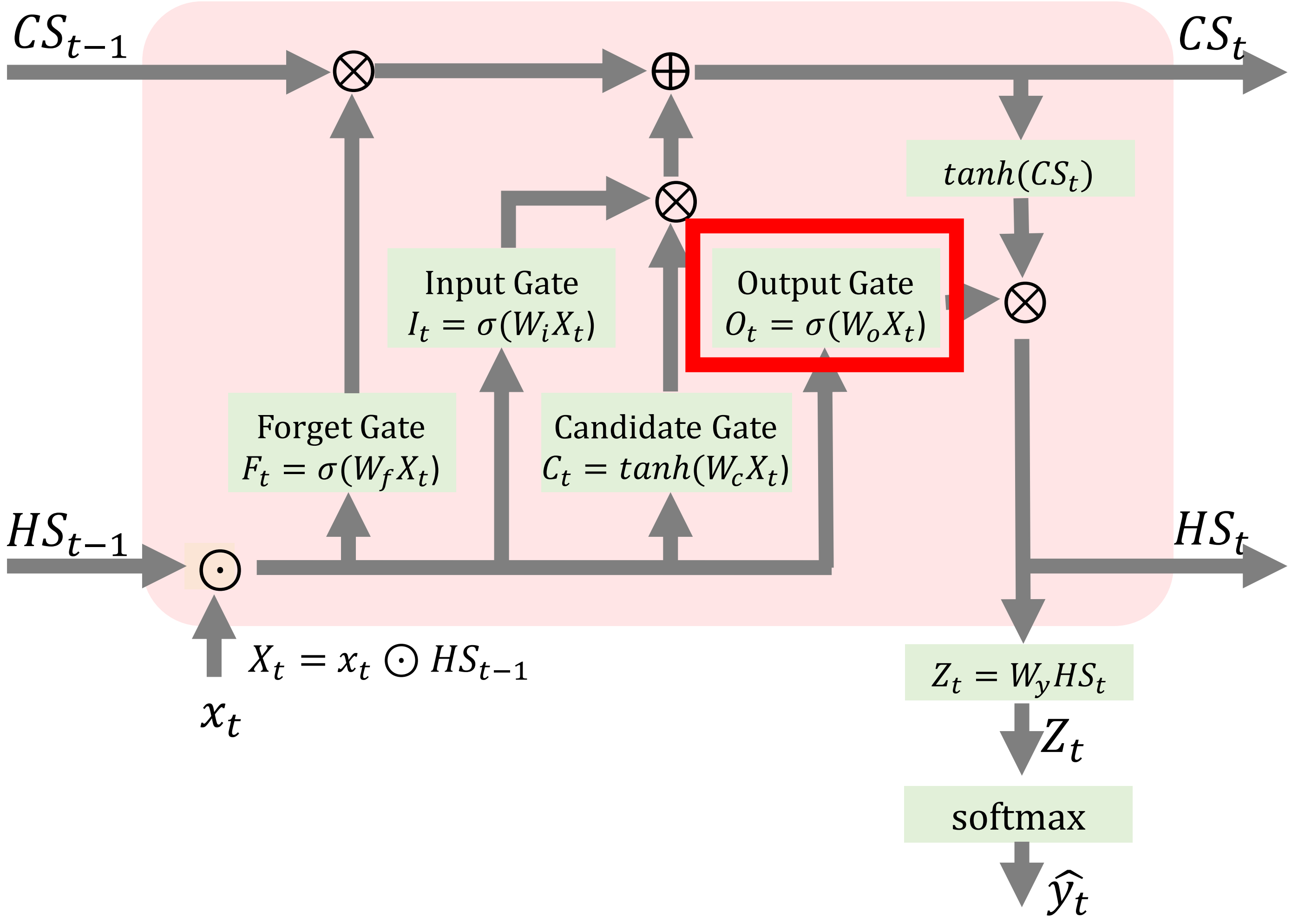


신박AI

그리고 LSTM에는 RNN과 다른 4개의 게이트가 있습니다

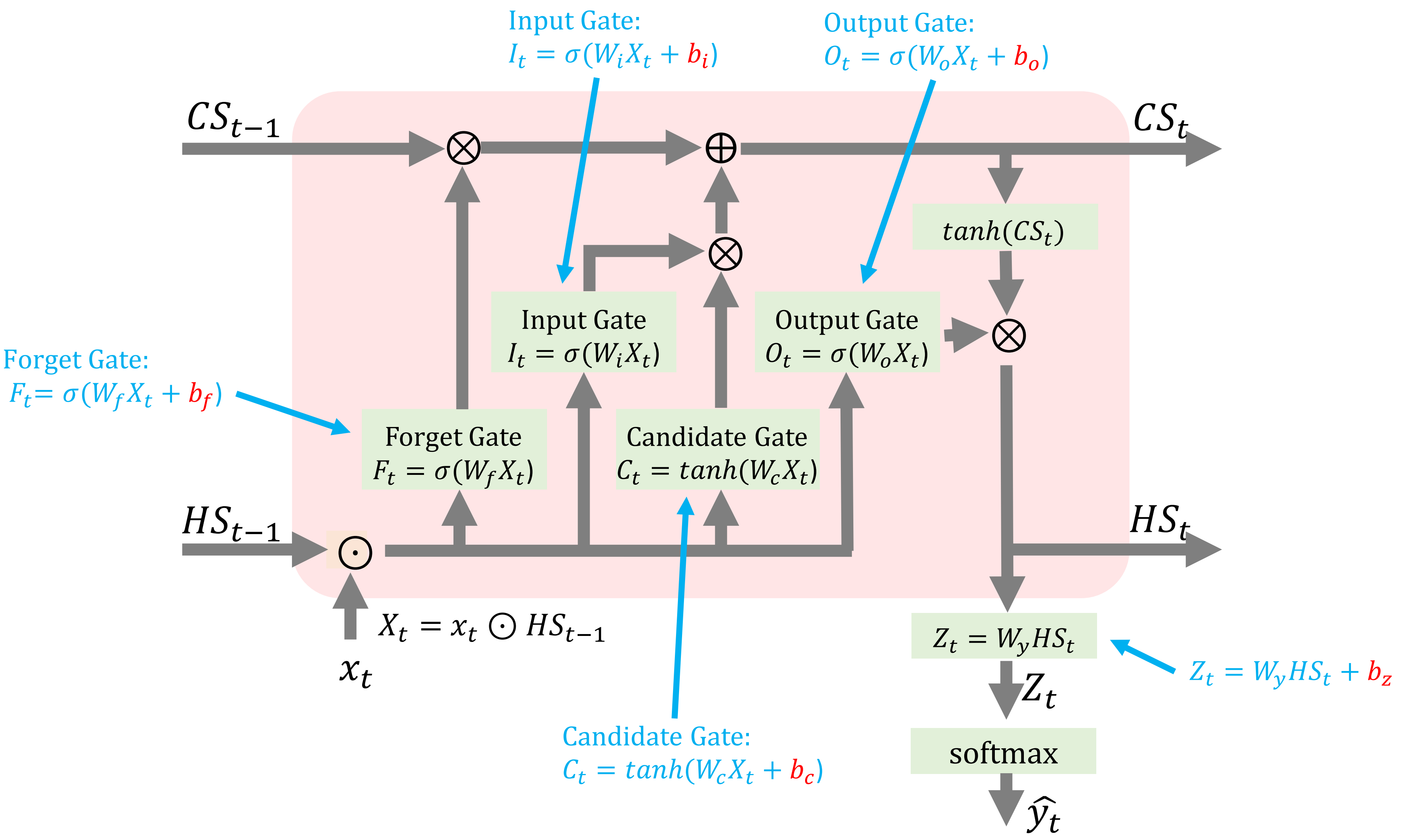


그리고 LSTM에는 RNN과 다른 4개의 게이트가 있습니다



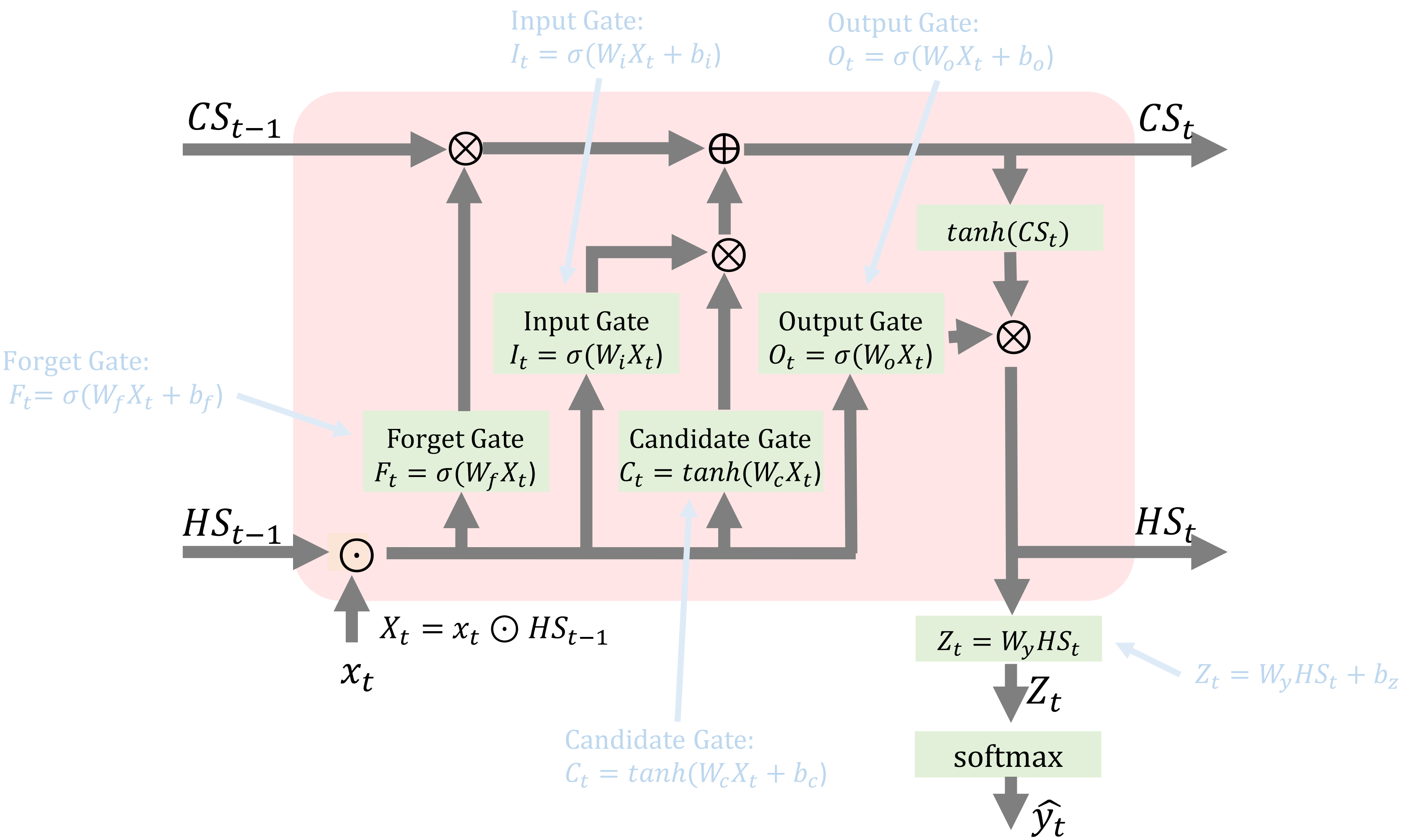
신박AI

원래 각각의 게이트와 레이어는 편향을 포함시켜야 합니다

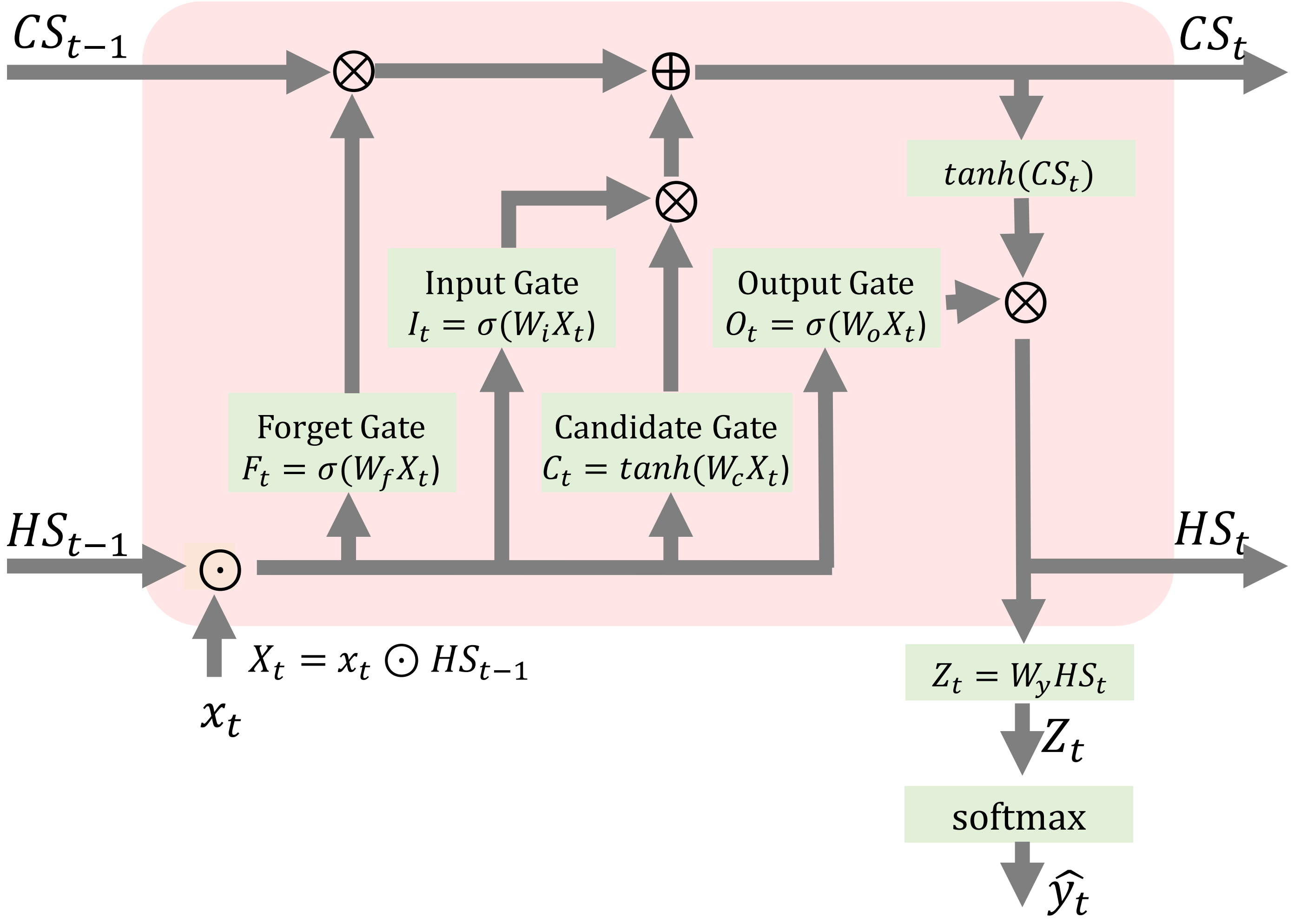


신박AI

그러나 오늘은 계산의 편의상 편향은 생략..하도록 하겠습니다

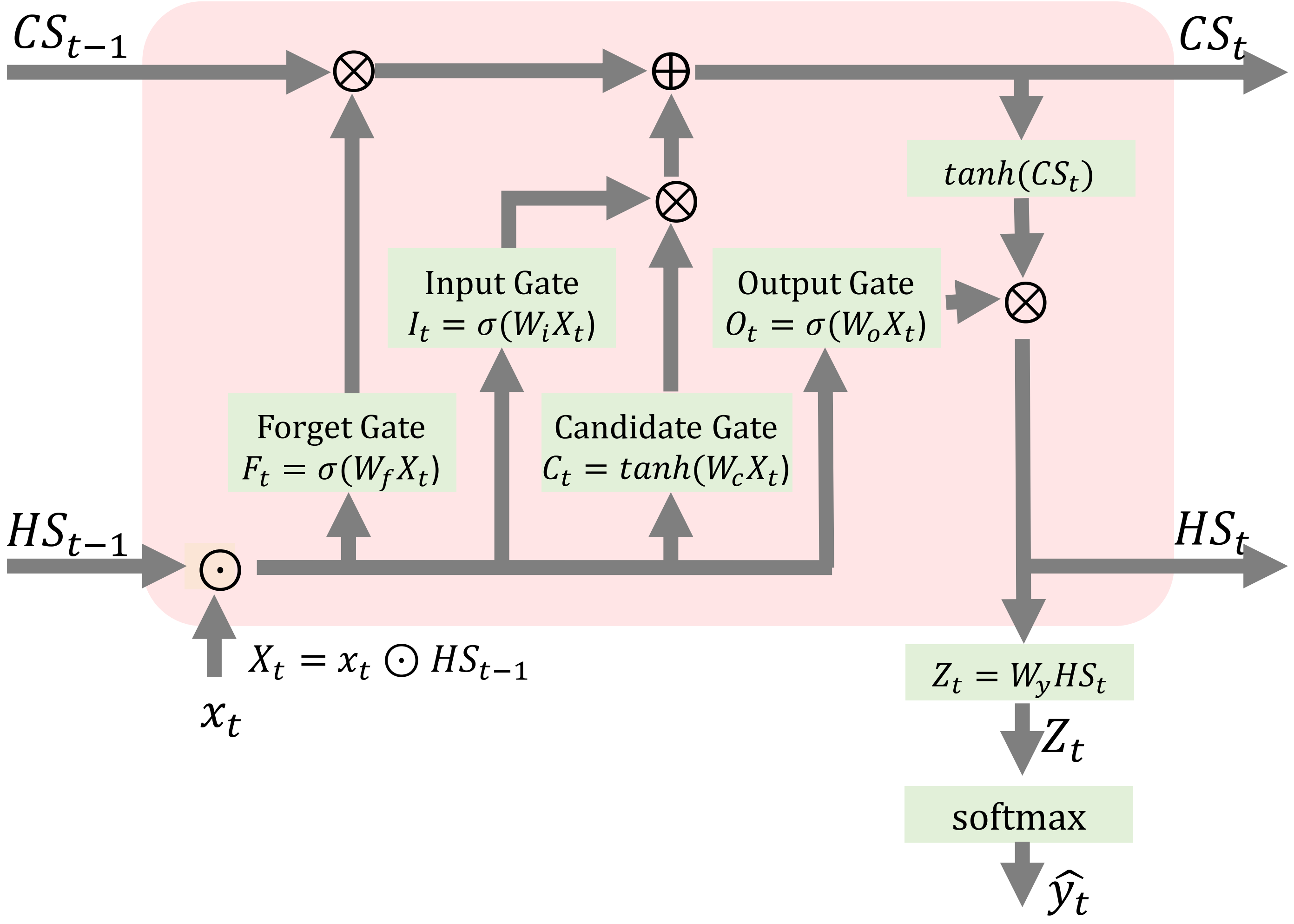


그러면 각각의 게이트가 정보들을 어떻게 처리하는지 알아보도록 하겠습니다



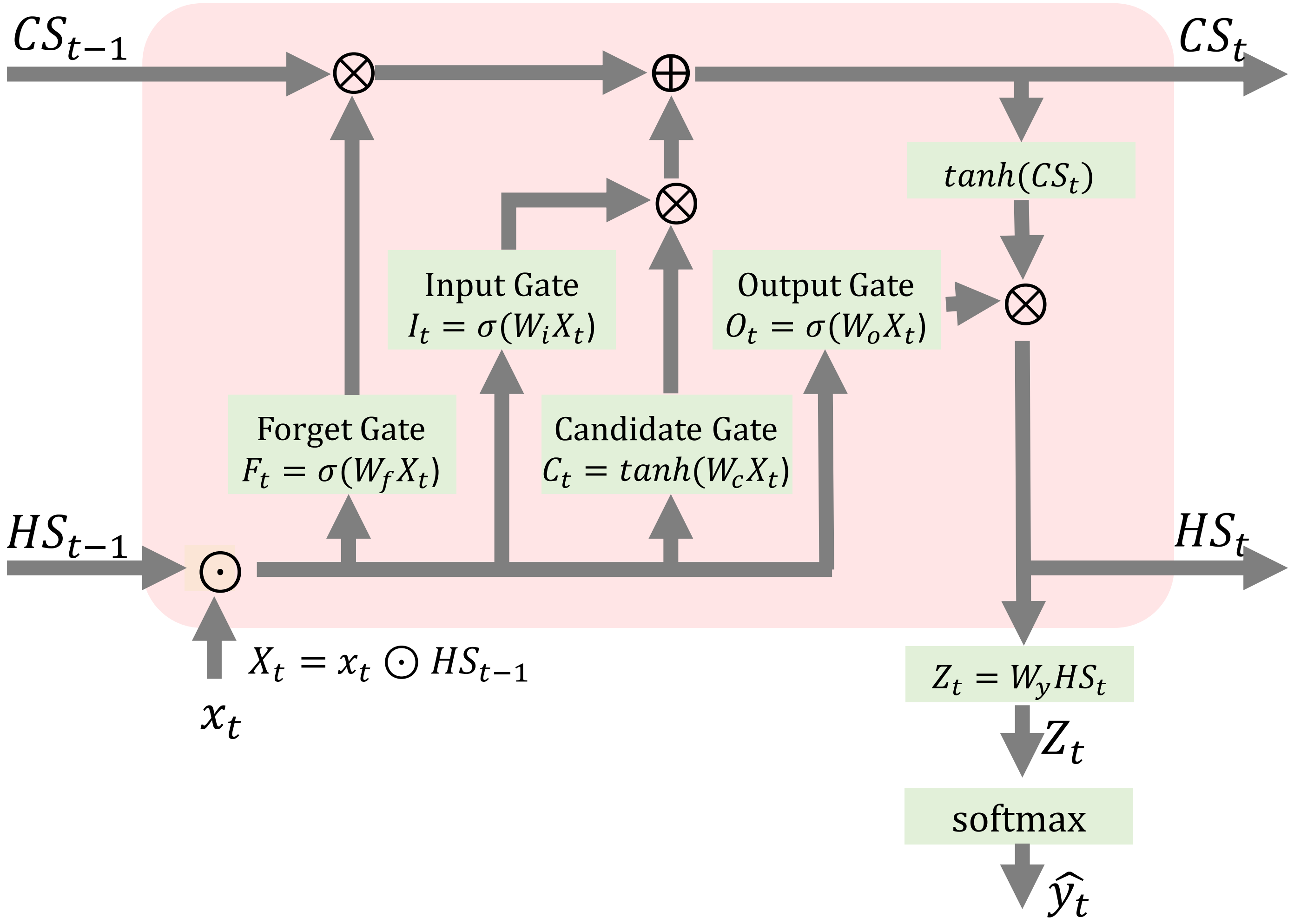
신박AI

숫자를 이용하여 계산과정을 알아보기 전에,

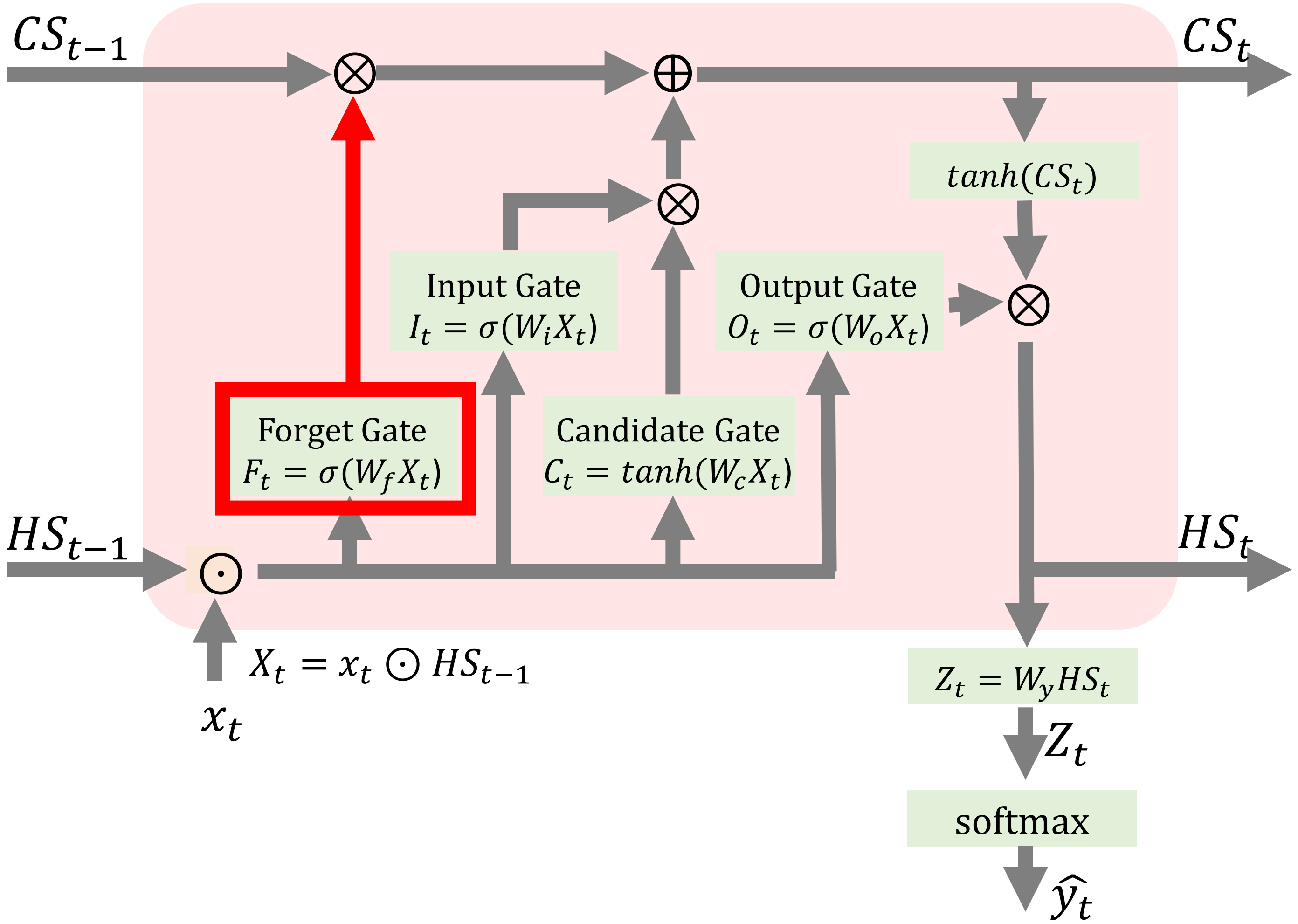


신박AI

각각 게이트들의 개념, 빅픽처를 살펴보도록 하겠습니다

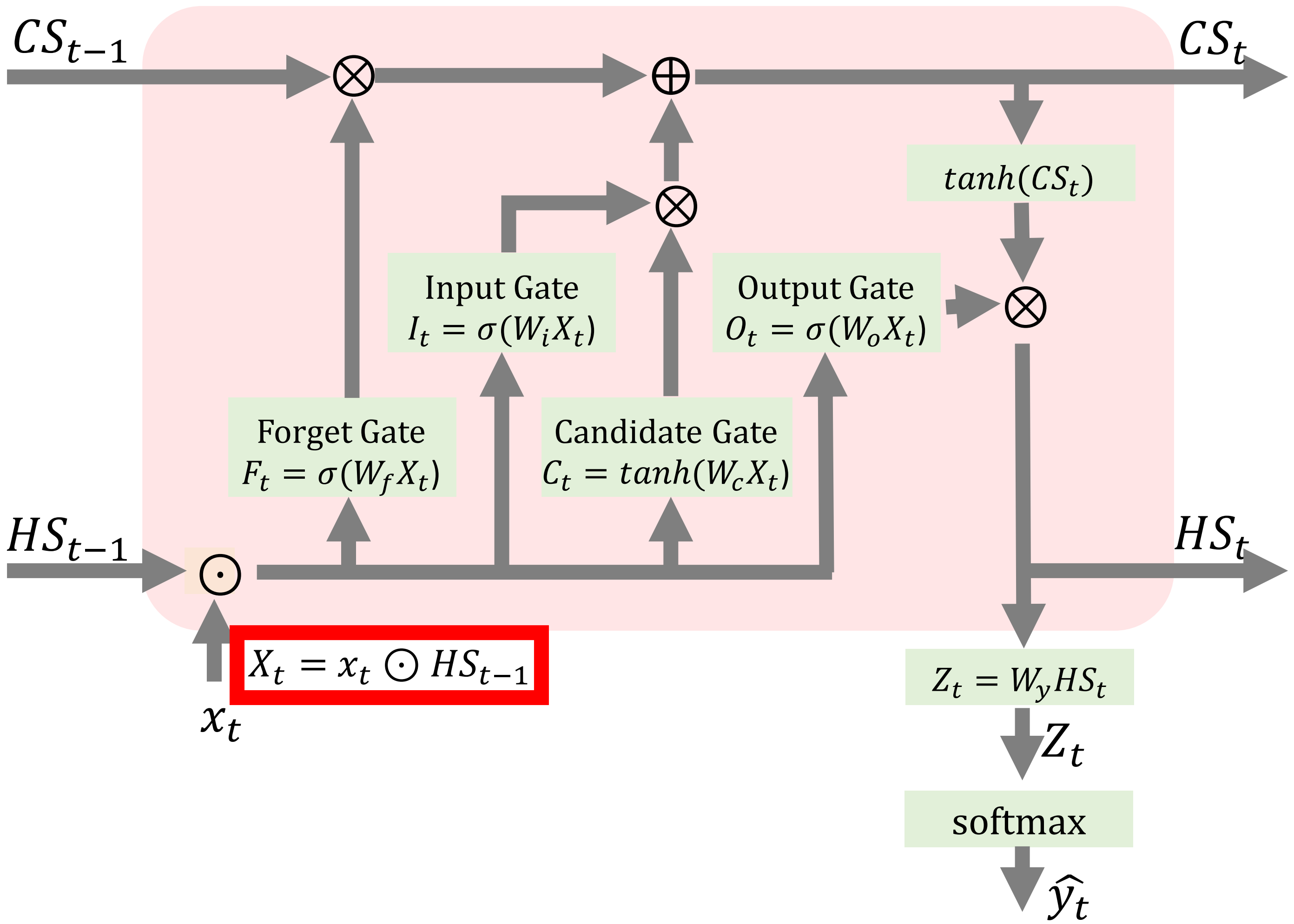


먼저 Forget Gate가 하는 일은 이름에서 알 수 있듯이 어떤 정보를 지울 것 (망각, forget)인가를 결정합니다



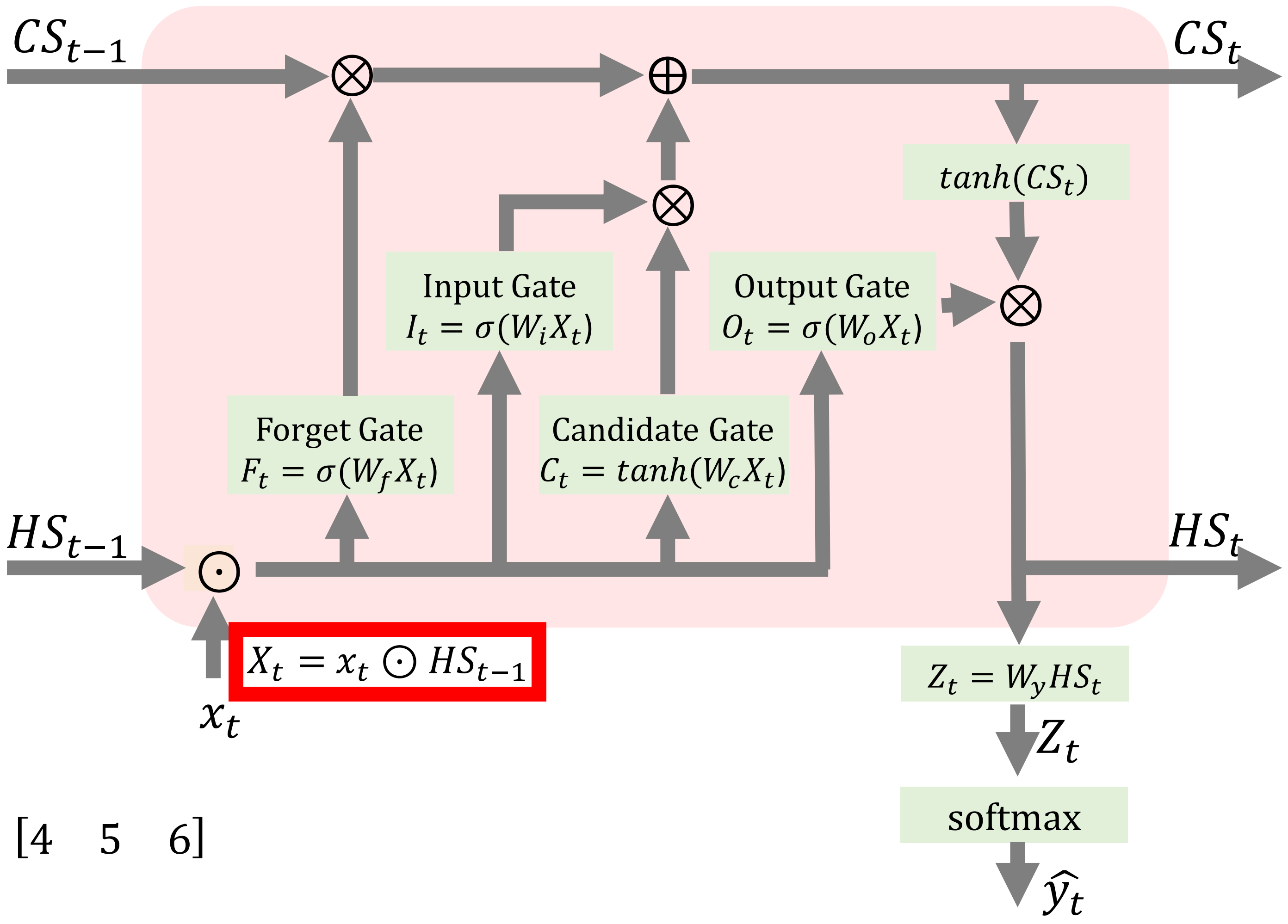
신박AI

우선 Forget Gate로 들어오는 정보는 지난 은닉상태 (HS_{t-1})와 현재입력값 (x_t)를 concatenate한 값입니다



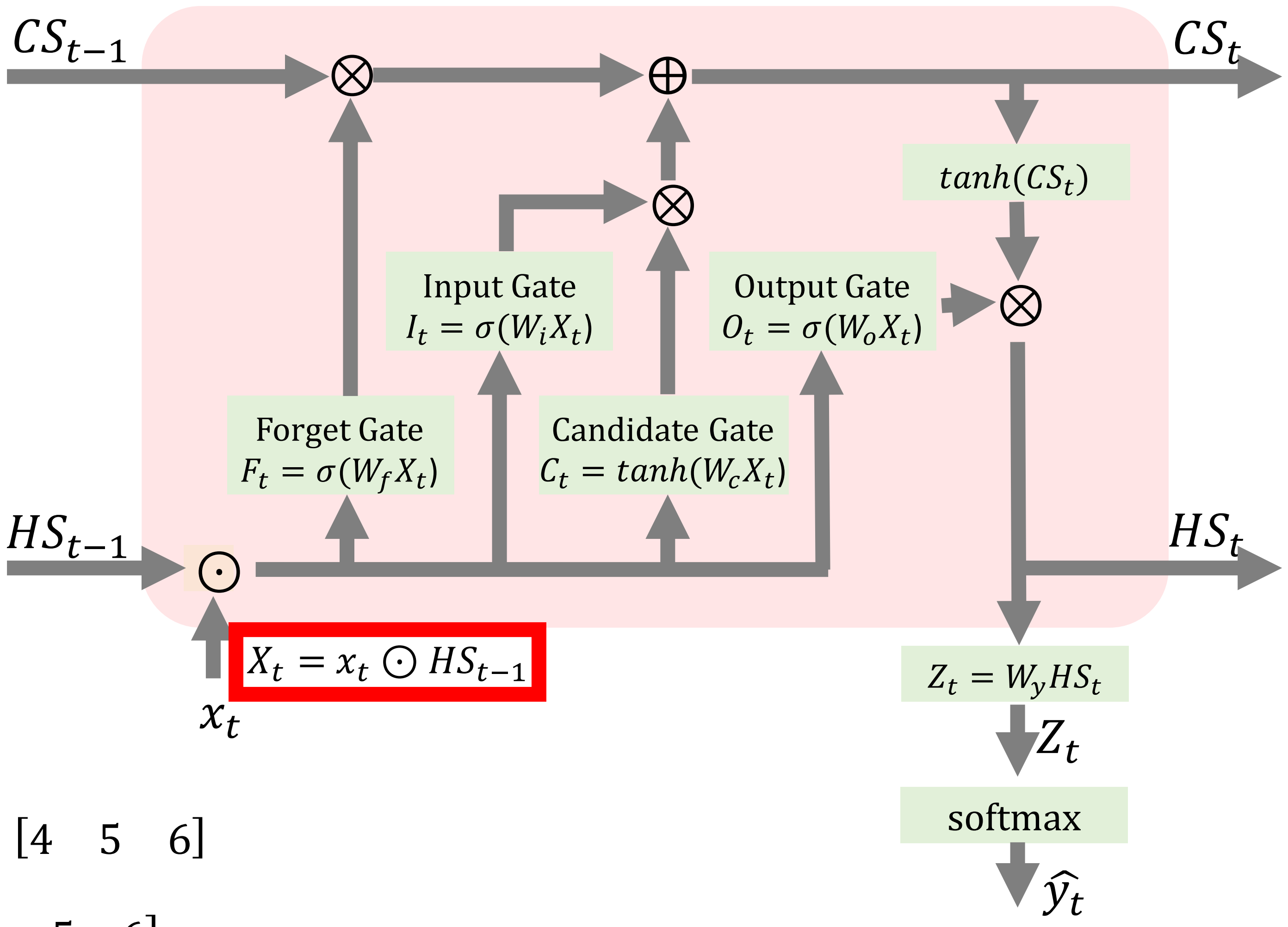
신박AI

Concatenate 한다는 것은 예를들어 이 두 행렬을



$[1 \quad 2 \quad 3] \odot [4 \quad 5 \quad 6]$

나란히 붙이는 것을 말합니다

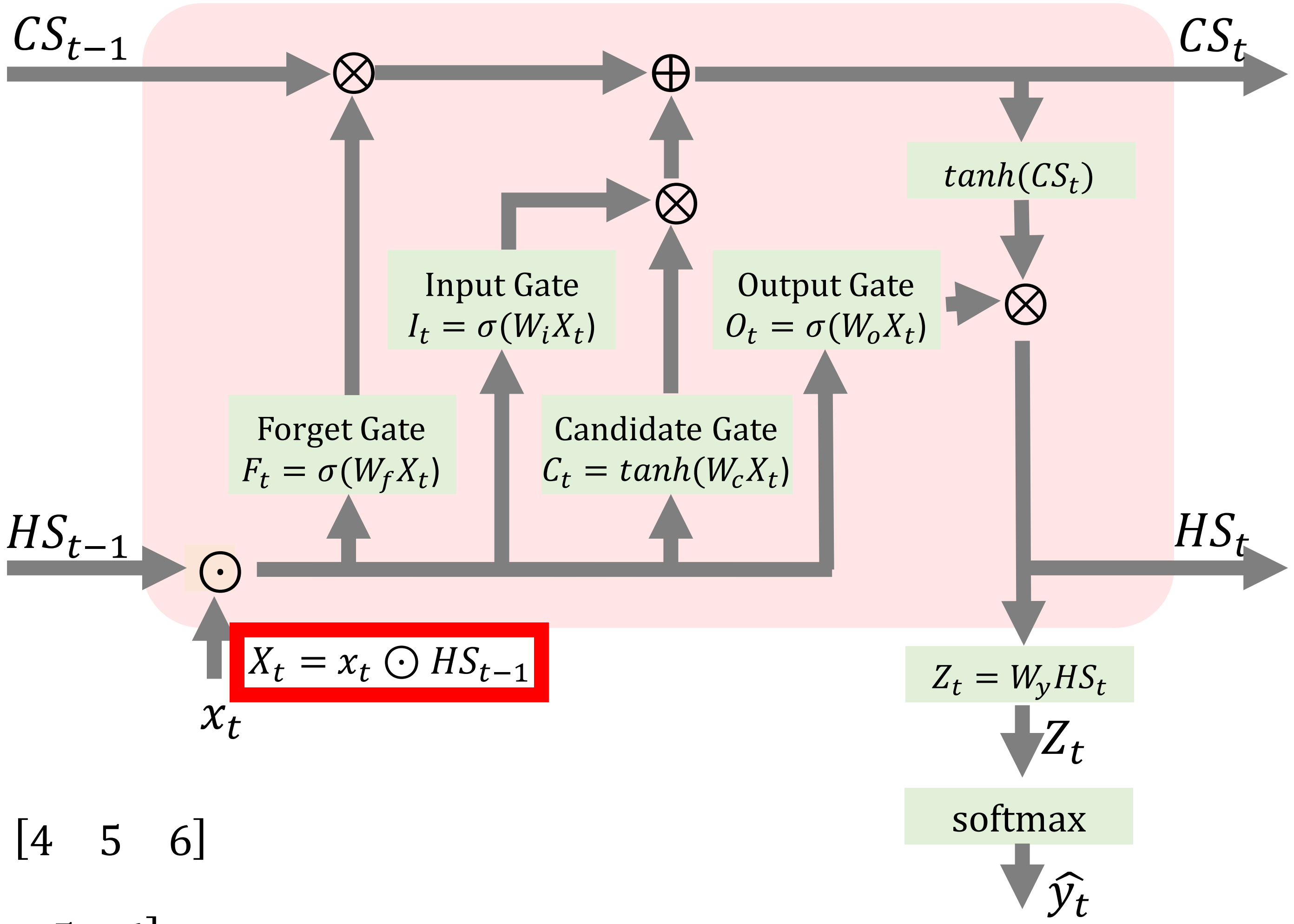


$$[1 \quad 2 \quad 3] \odot [4 \quad 5 \quad 6]$$

$$= [1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6]$$

신박AI

이렇게 함으로써, concatenate된 X_t는 이전 은닉상태와 현재 입력값이 한데 묶여진 일종의 단기기억 (short-term memory)처럼 됩니다

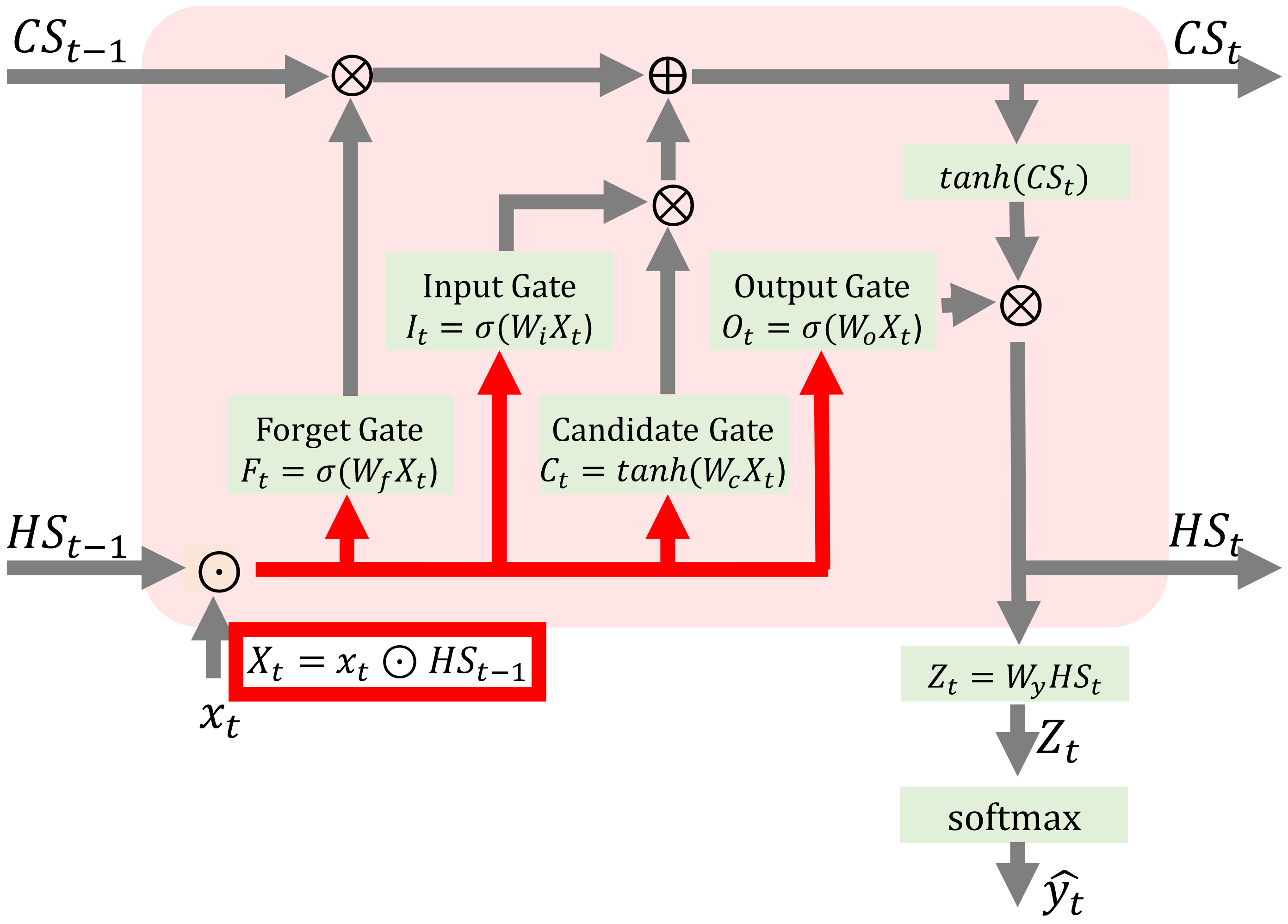


$$[1 \quad 2 \quad 3] \odot [4 \quad 5 \quad 6]$$

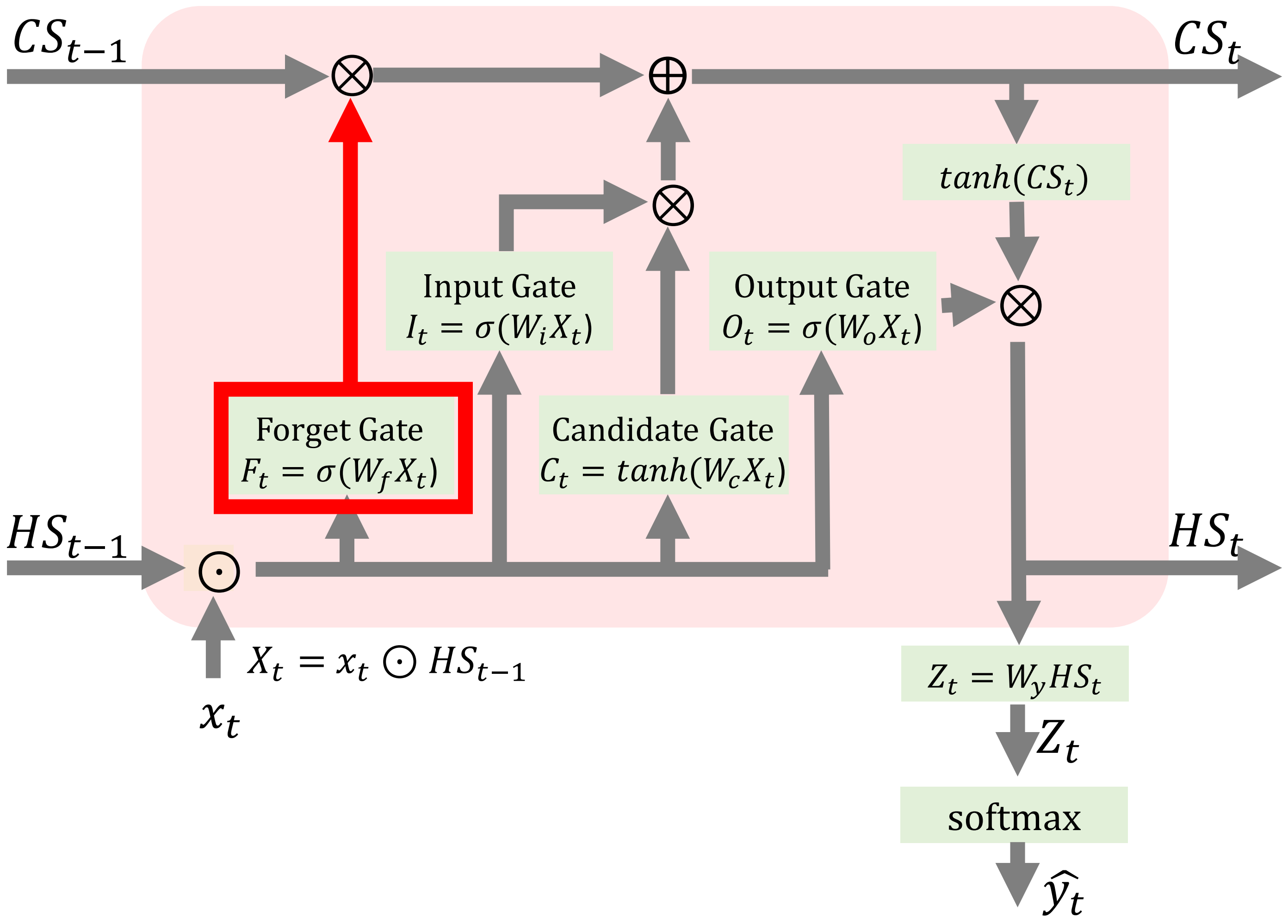
$$= [1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6]$$

신박AI

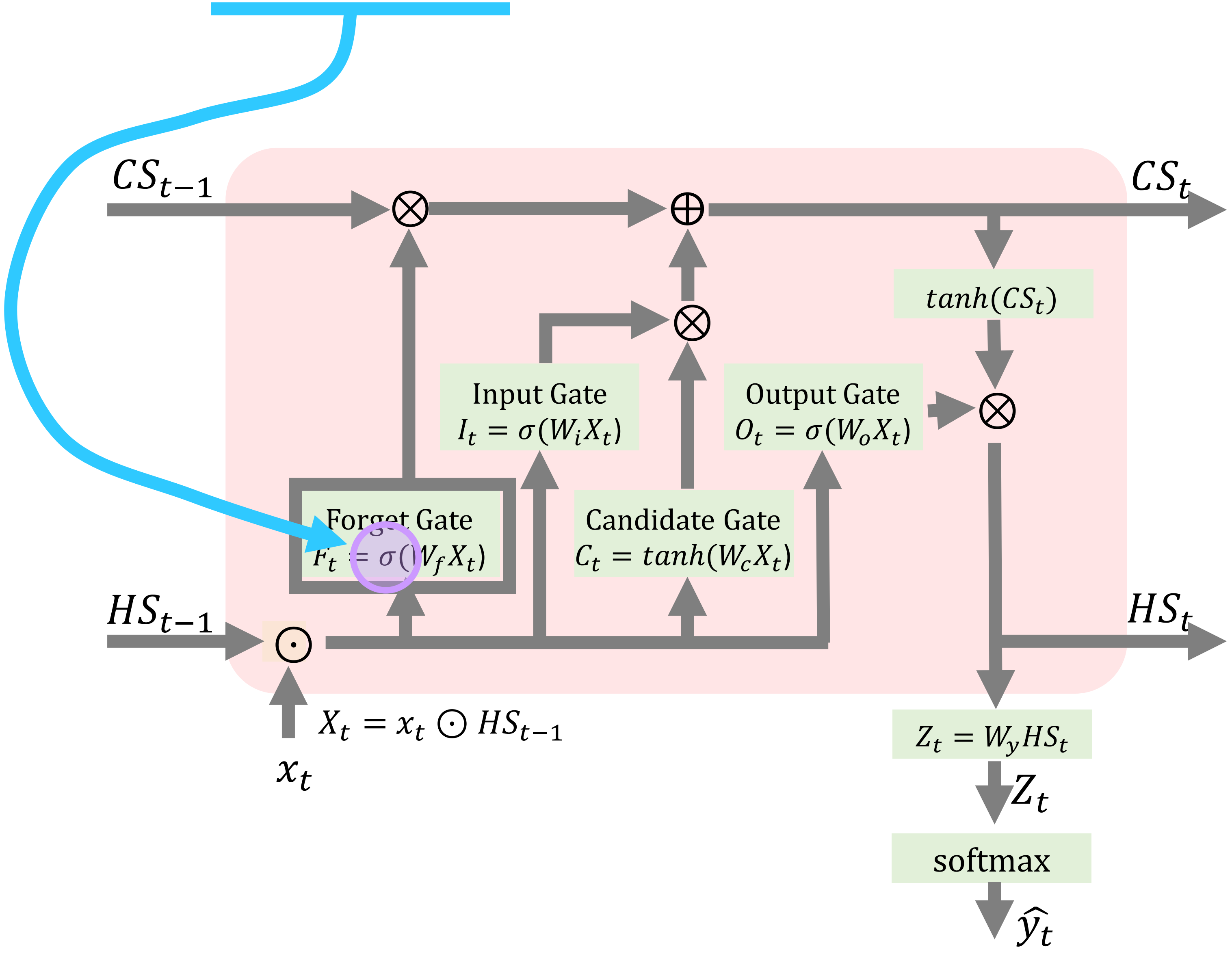
이 X_t는 LSTM내의 모든 게이트들의 입력값이 되는 것을 기억해주세요



그러면 첫번째 Forget Gate 에서 주목해야 할 것은

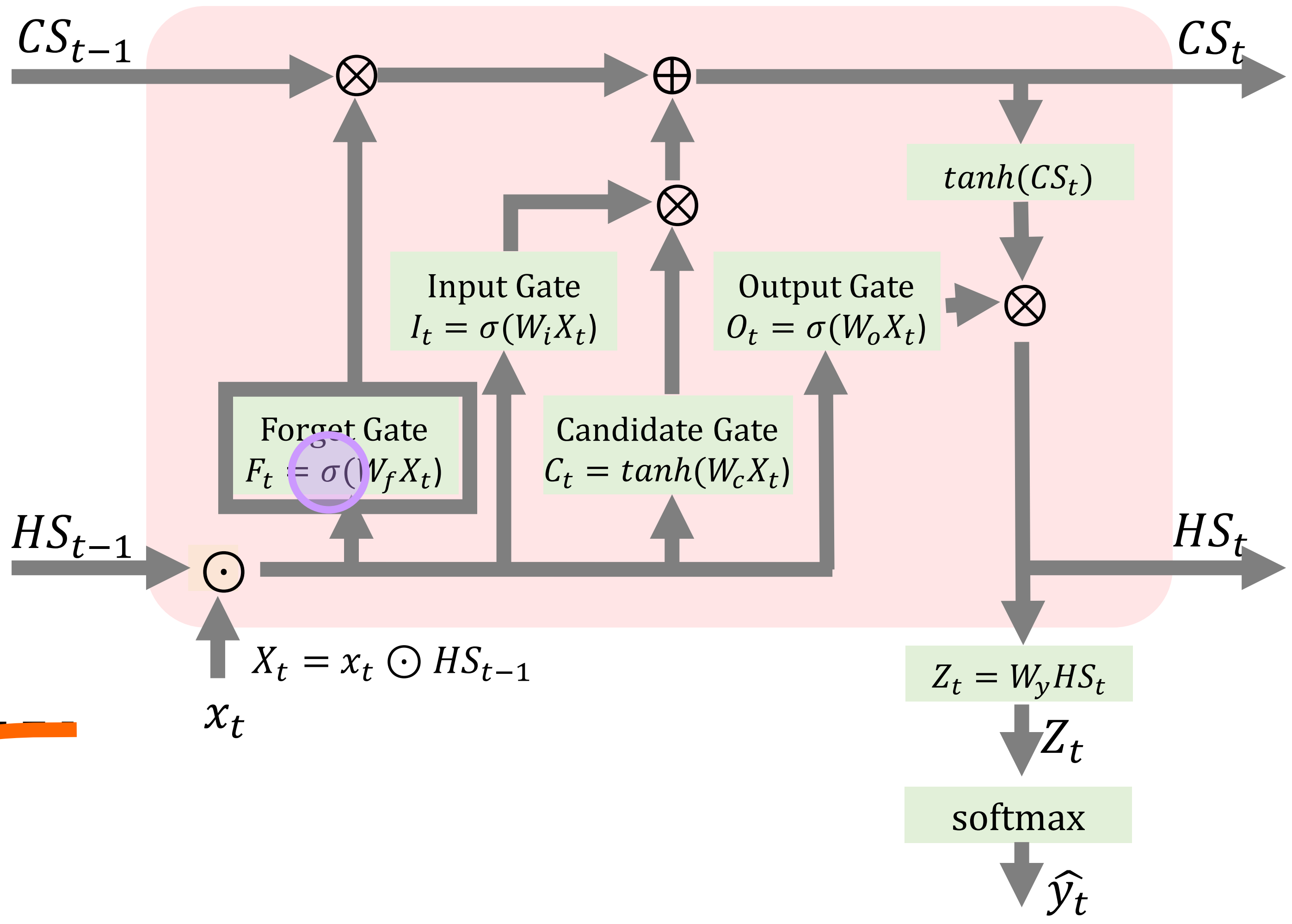


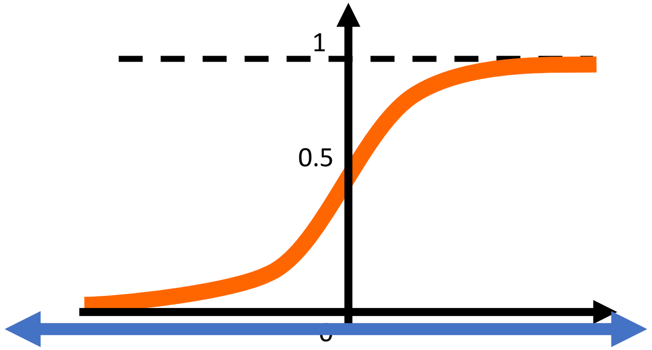
Forget Gate안에 시그모이드 함수가 있다는 점입니다



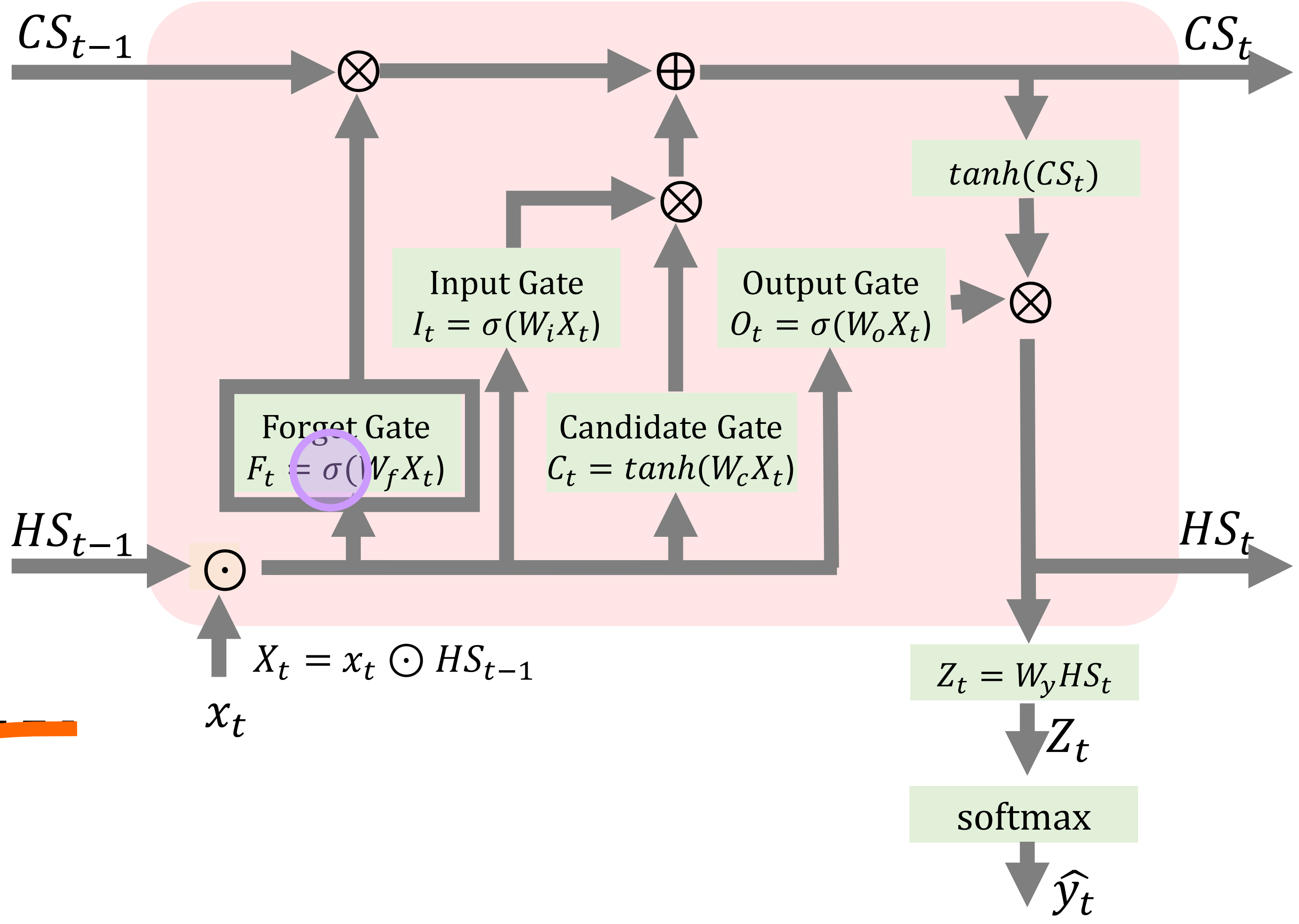
신박AI

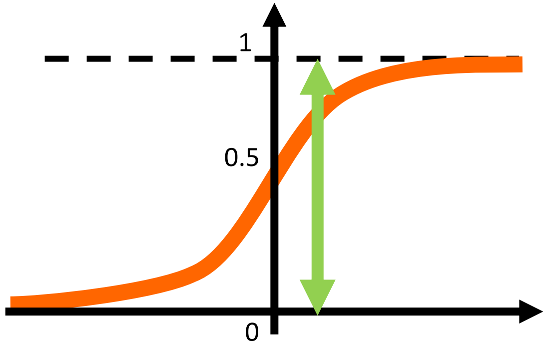
우리가 시그모이드를 배워 본바와 같이 시그모이드 함수는 어떤 입력이 들어와도



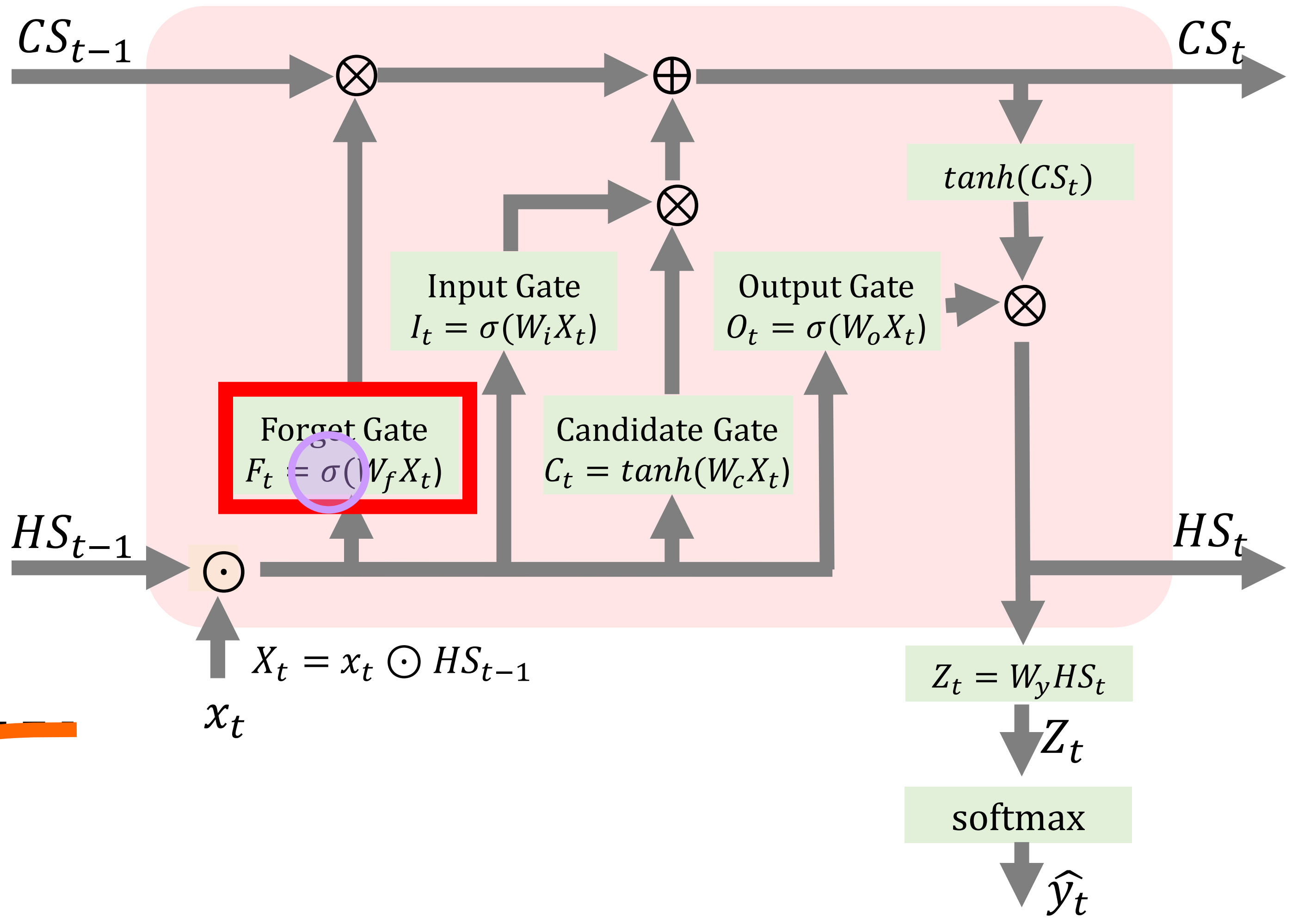


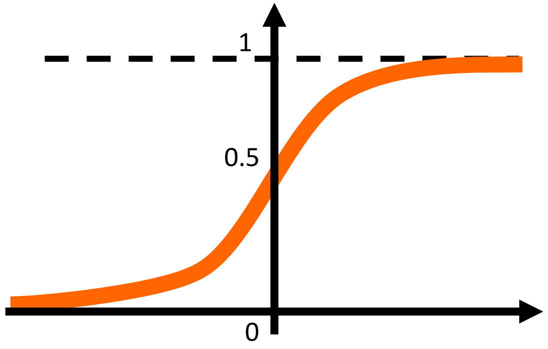
0과 1 사이의 값을 리턴하는 함수입니다





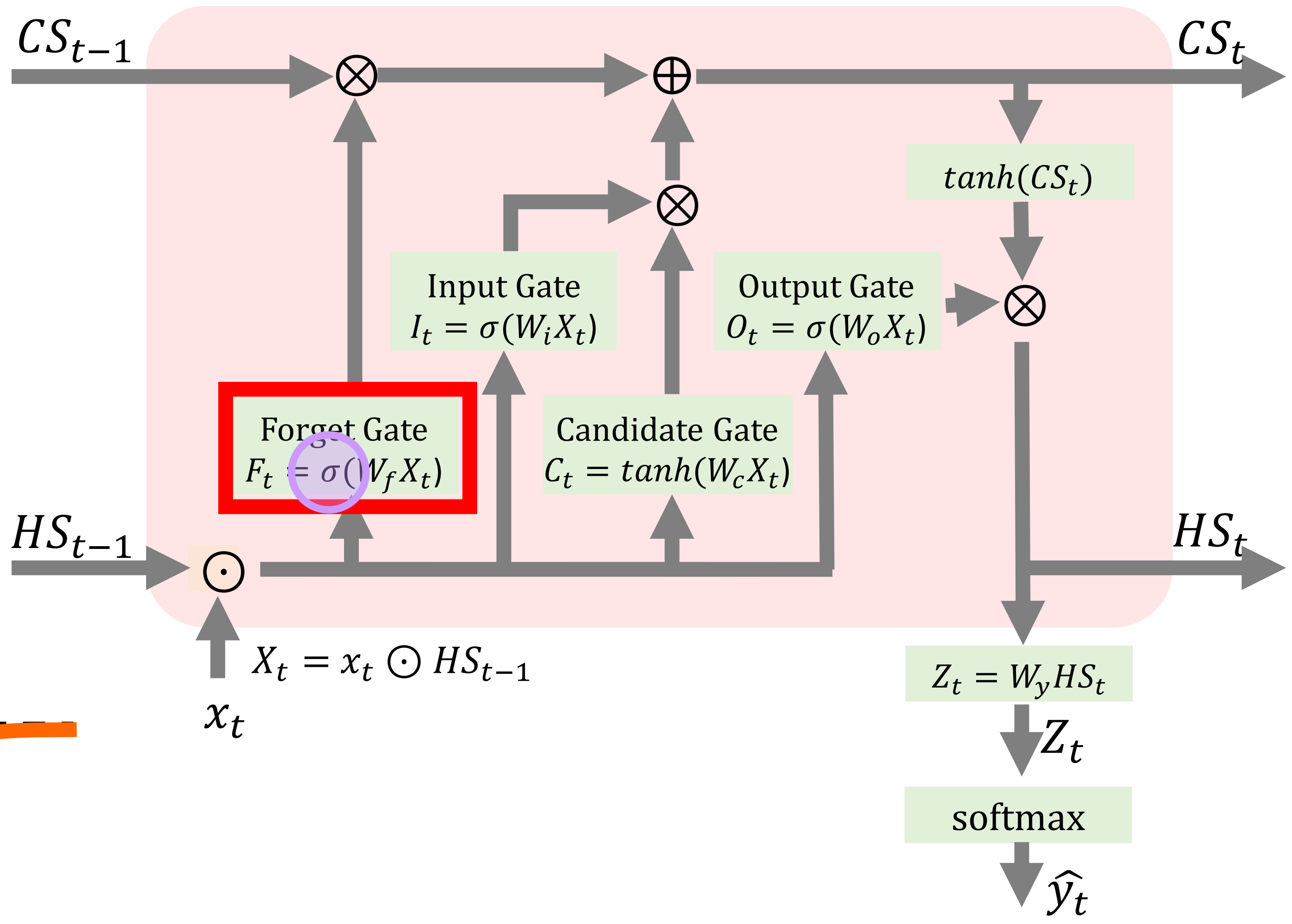
즉 Forget Gate가 하는 일은 들어오는 (바로 앞의 과거+현재) 입력값 받아서 가중치를 곱한 뒤,

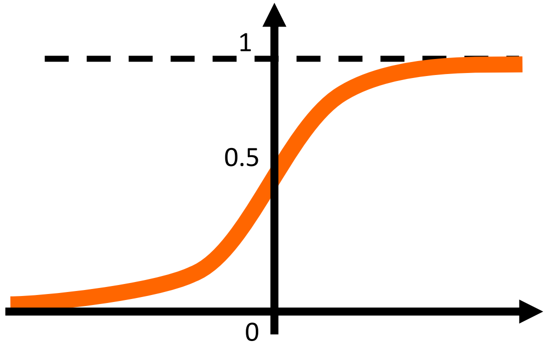




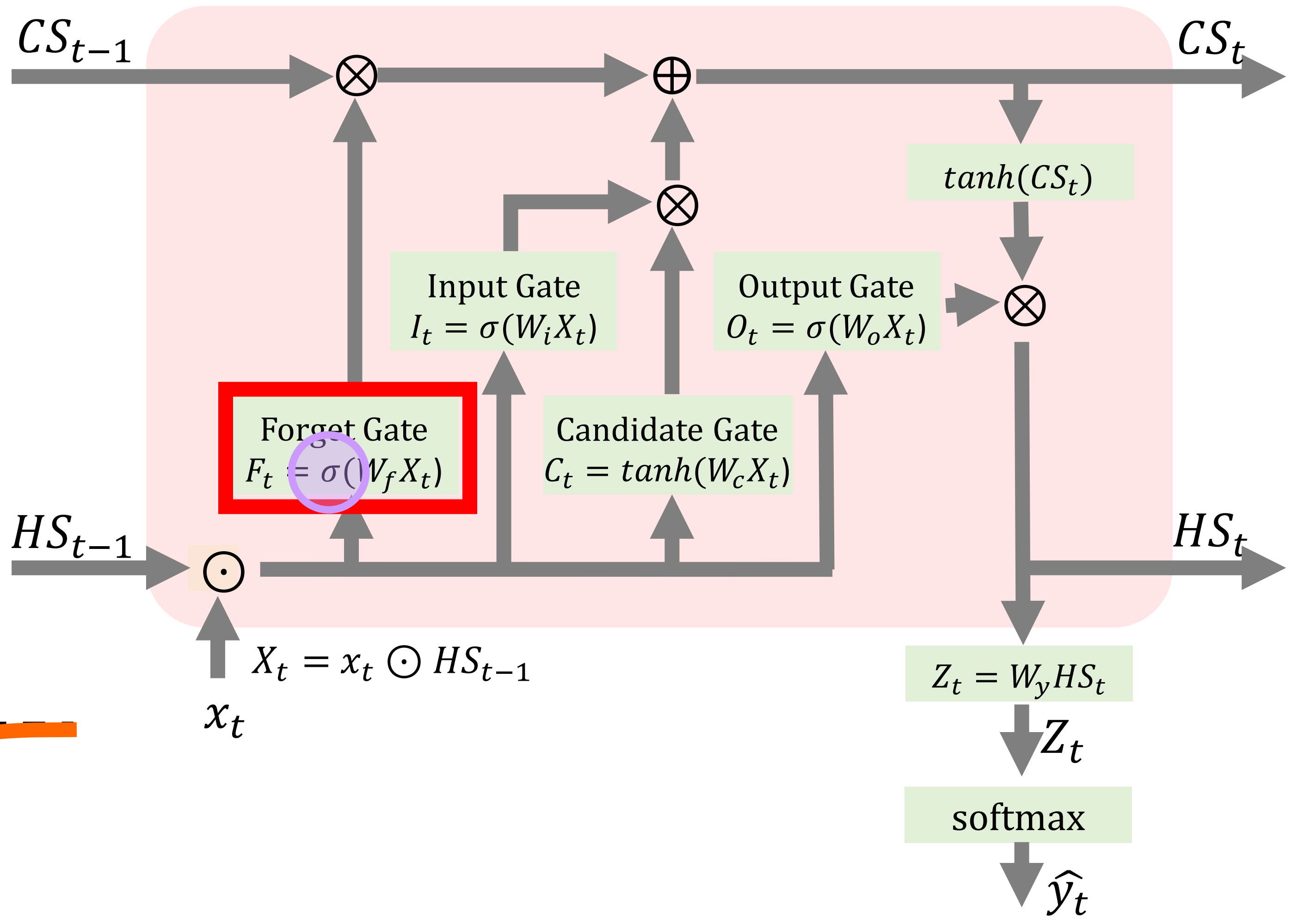
신박AI

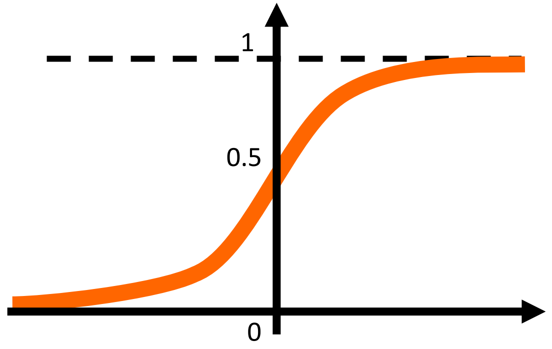
0과 1 사이의 값으로 바꾸어주는 역할을 합니다



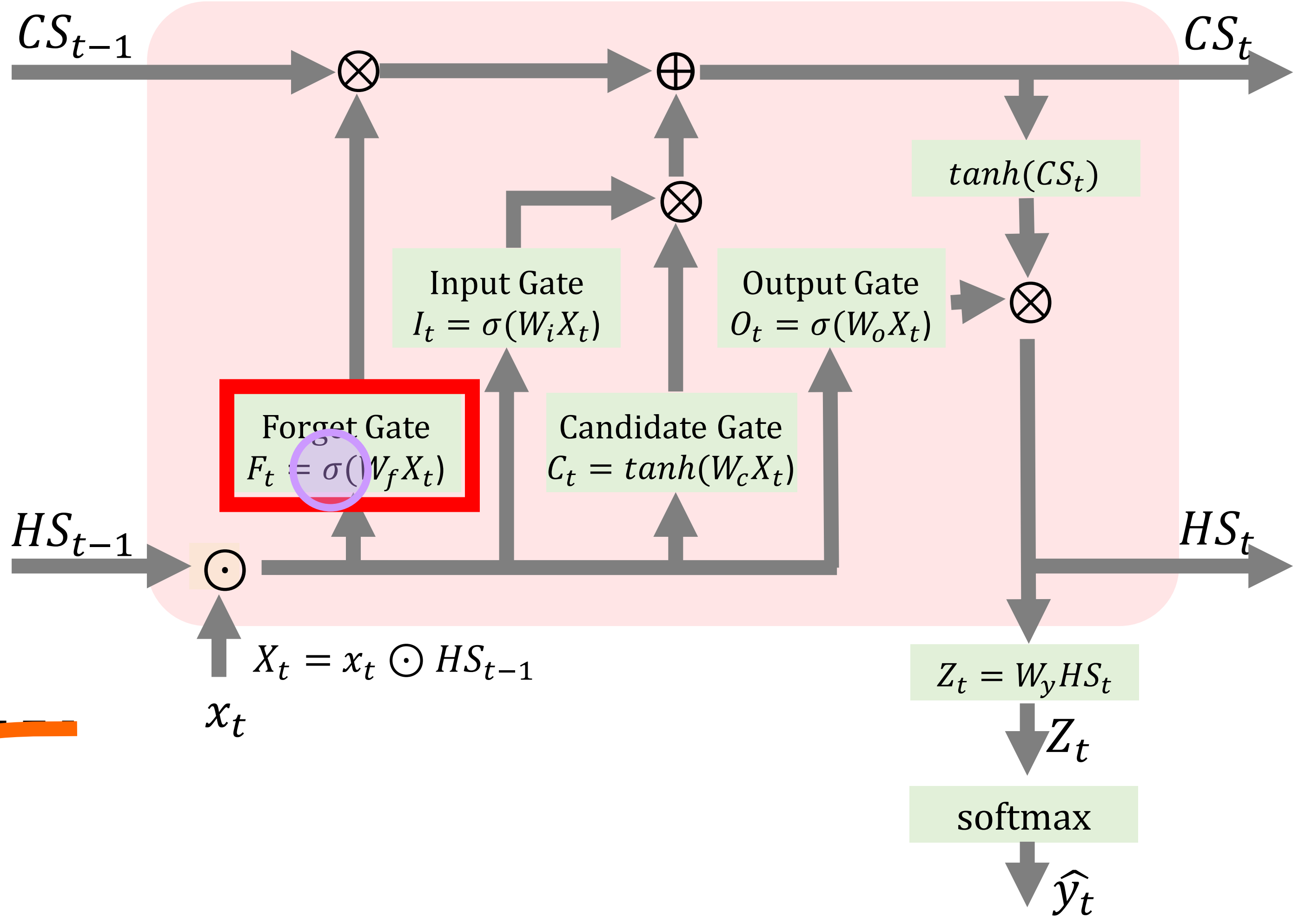


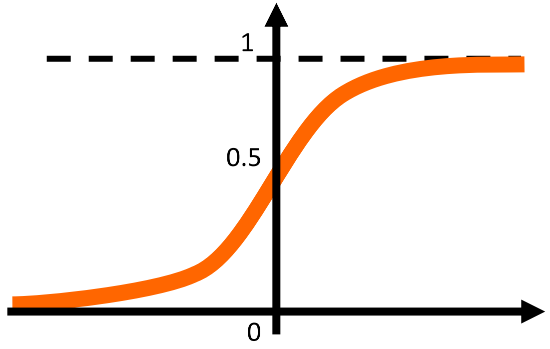
이때, $W_f X_t$ 값들중 마이너스 값들은 0에 가까워 질 것이고





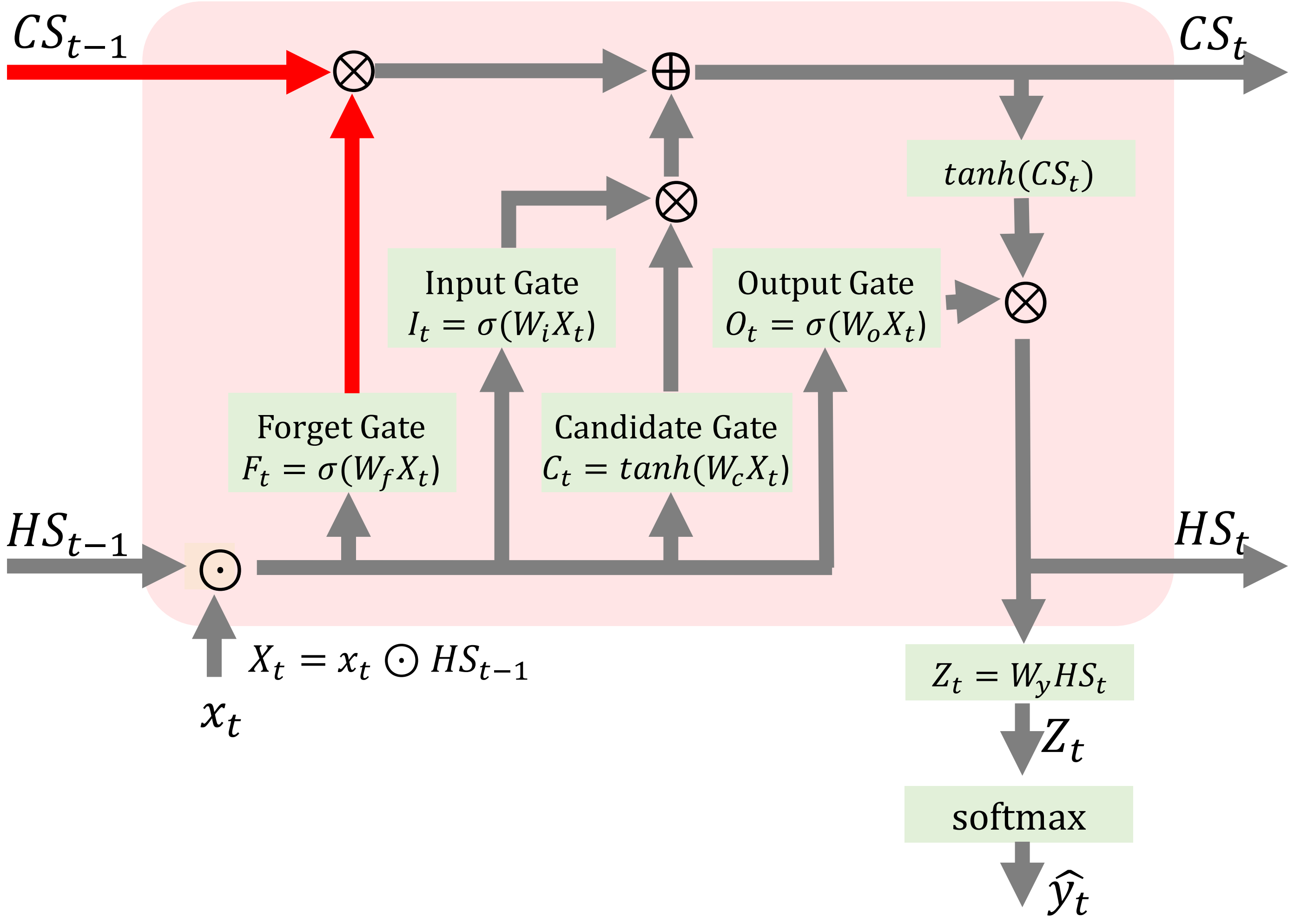
양수는 1에 가까워 질 것입니다





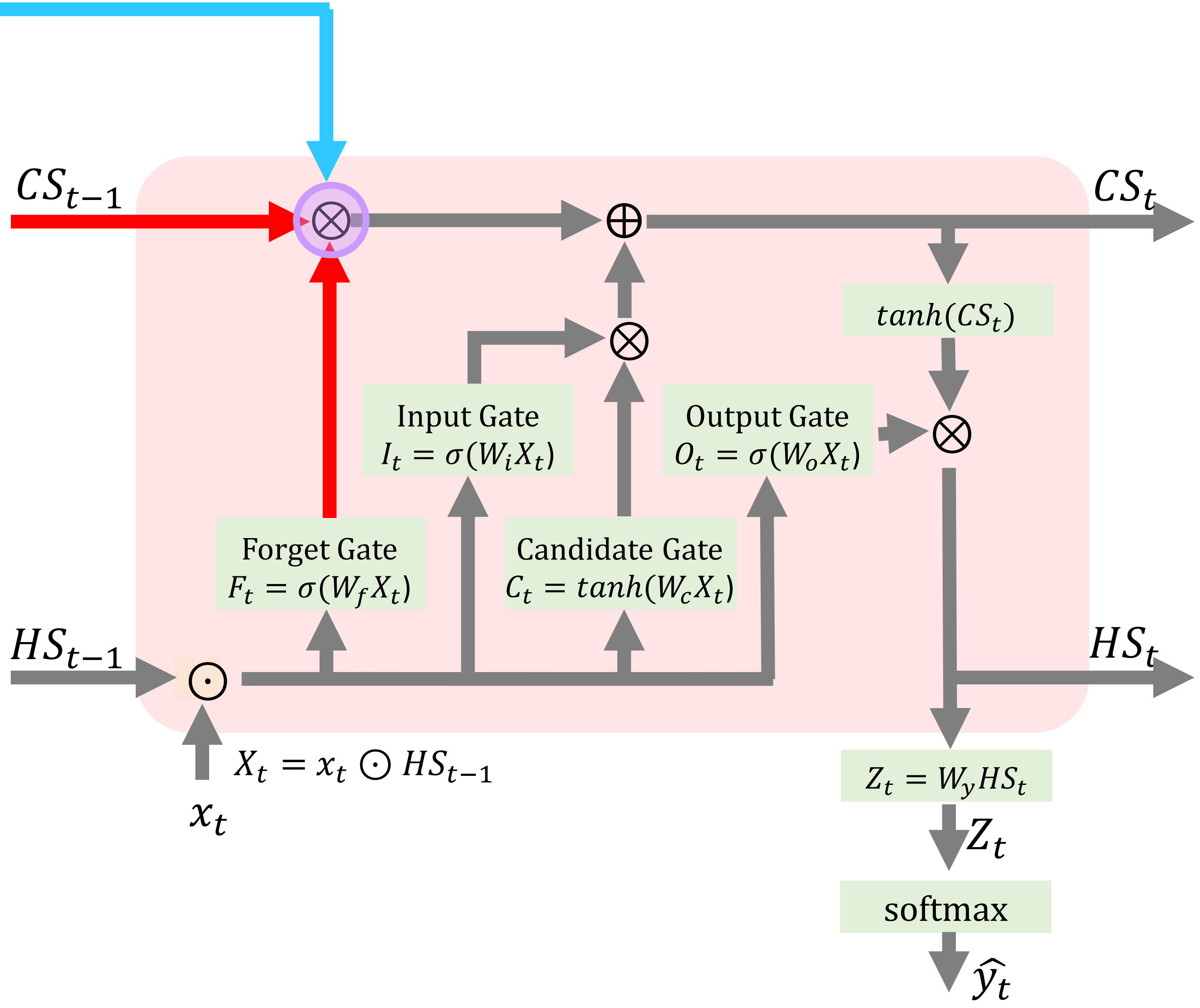
신박AI

그러면 이렇게 0과 1사이로 바뀐 값들은 셀상태의 값들을 만나



신박AI

Element-wise 곱셈연산을 하게 됩니다

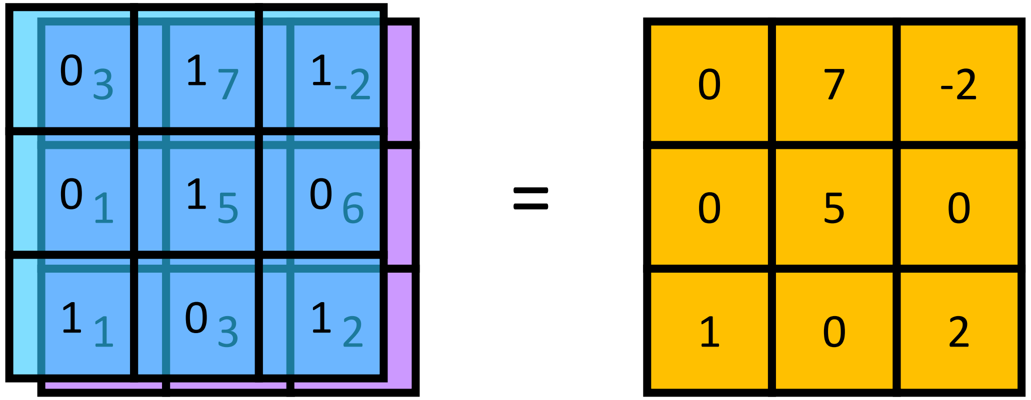


신박AI

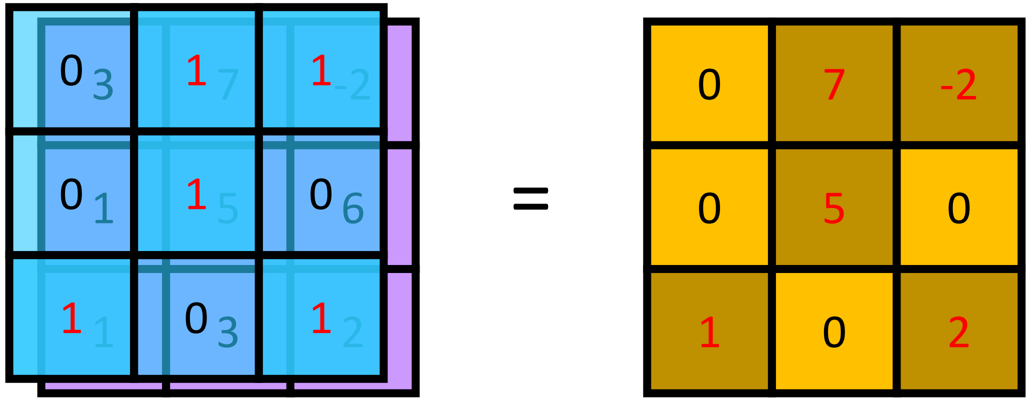
이 element-wise연산은 두 행렬을 곱하는데 각각의 원소(element)별로 곱하는 것을 말합니다

$$\begin{bmatrix} 0 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 3 & 7 & -2 \\ 1 & 5 & 6 \\ 1 & 3 & 2 \end{bmatrix}$$

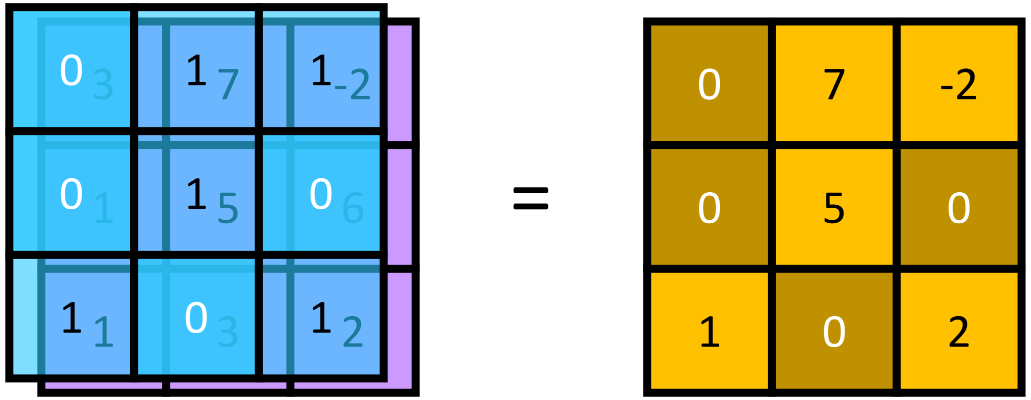
이 element-wise연산은 두 행렬을 곱하는데 각각의 원소(element)별로 곱하는 것을 말합니다



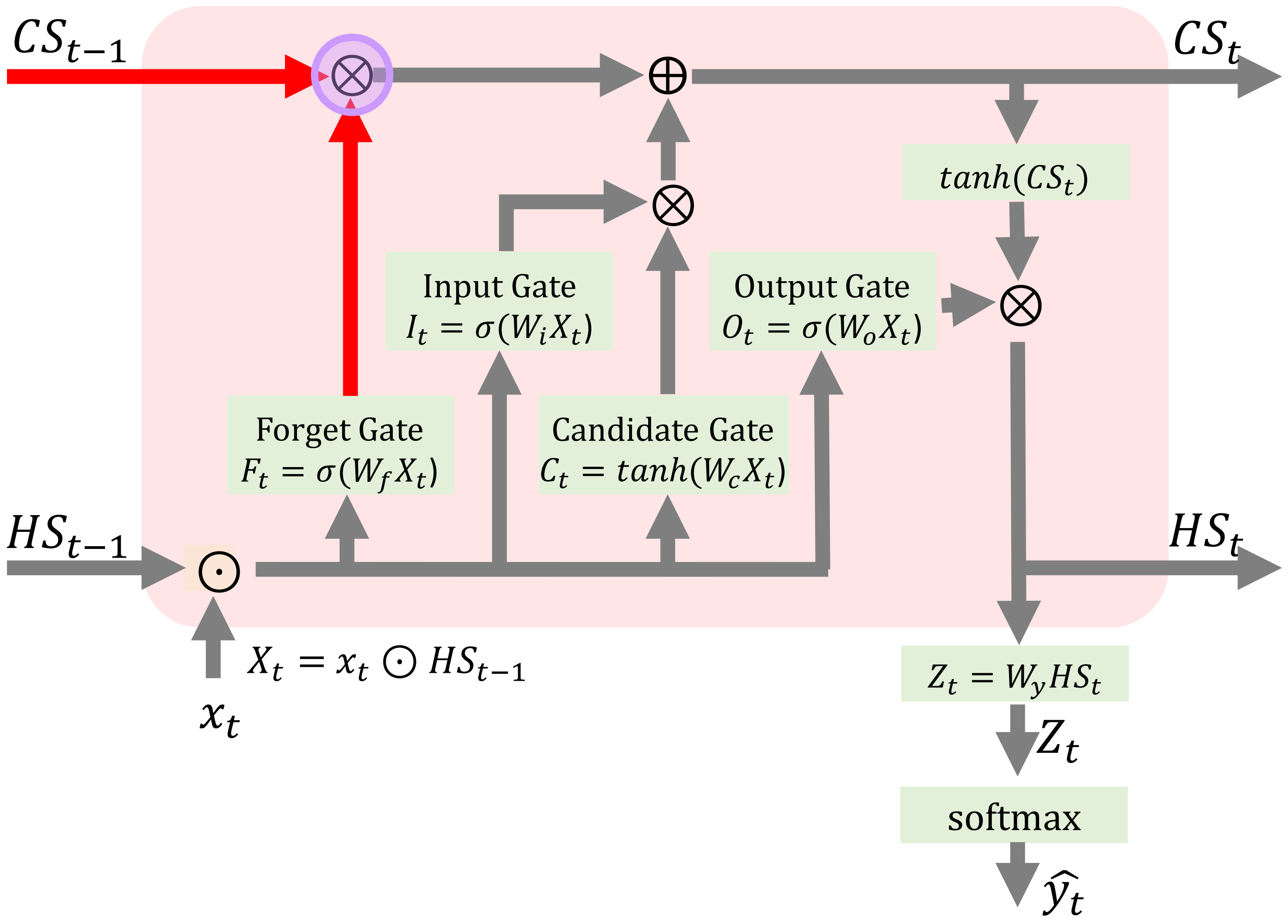
이렇게 연산하는 이유는 원소가 1인 행렬의 정보는 남기고



원소가 0인 행렬의 정보는 지워버리기 위함 (망각) 입니다

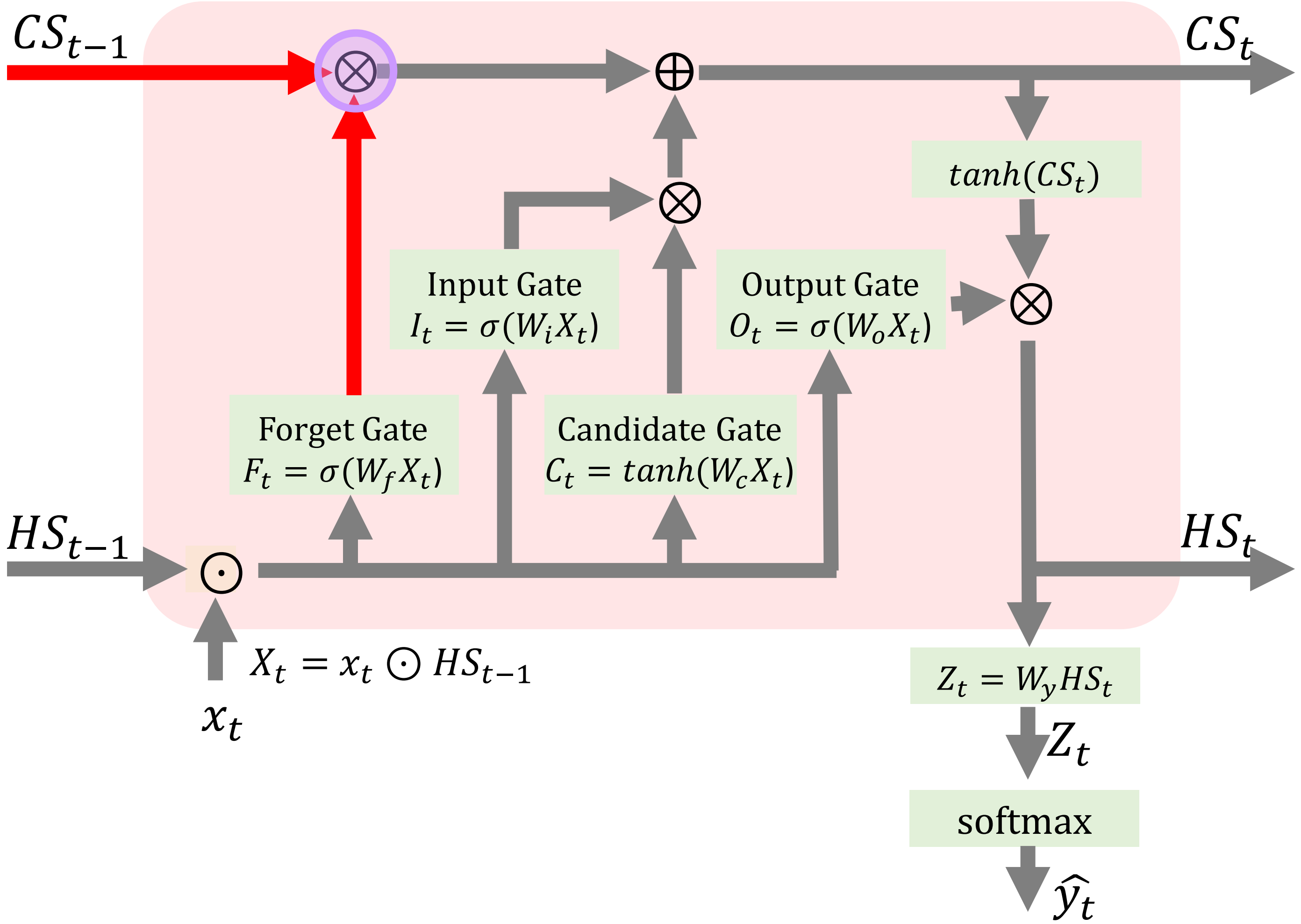


만약에 예를들어 Forget Gate의 출력값이 0과 1로만 이루어져 있다고 가정해보면



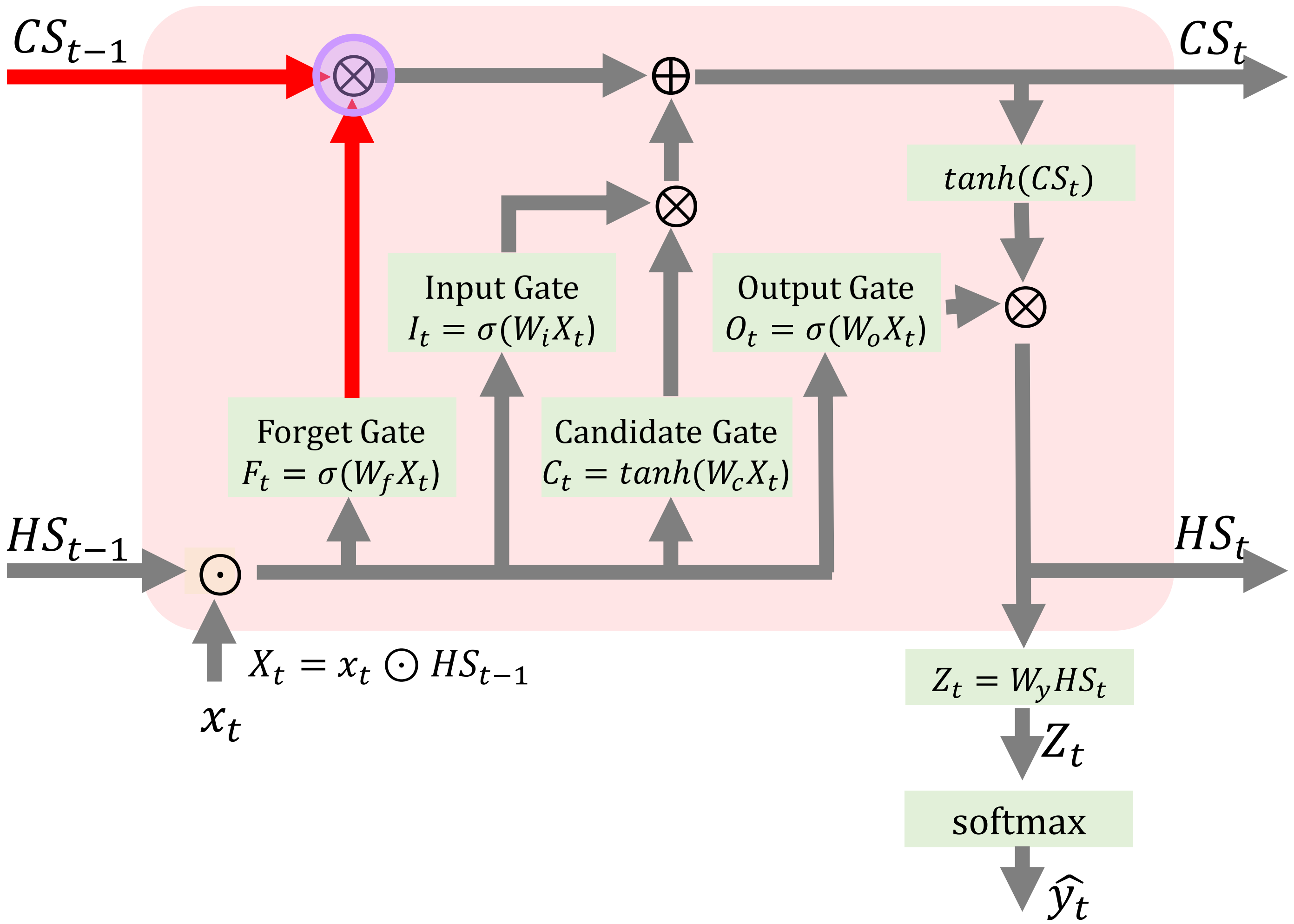
신박AI

Forget Gate의 출력값이 0인 곳은 element-wise곱에 의해서



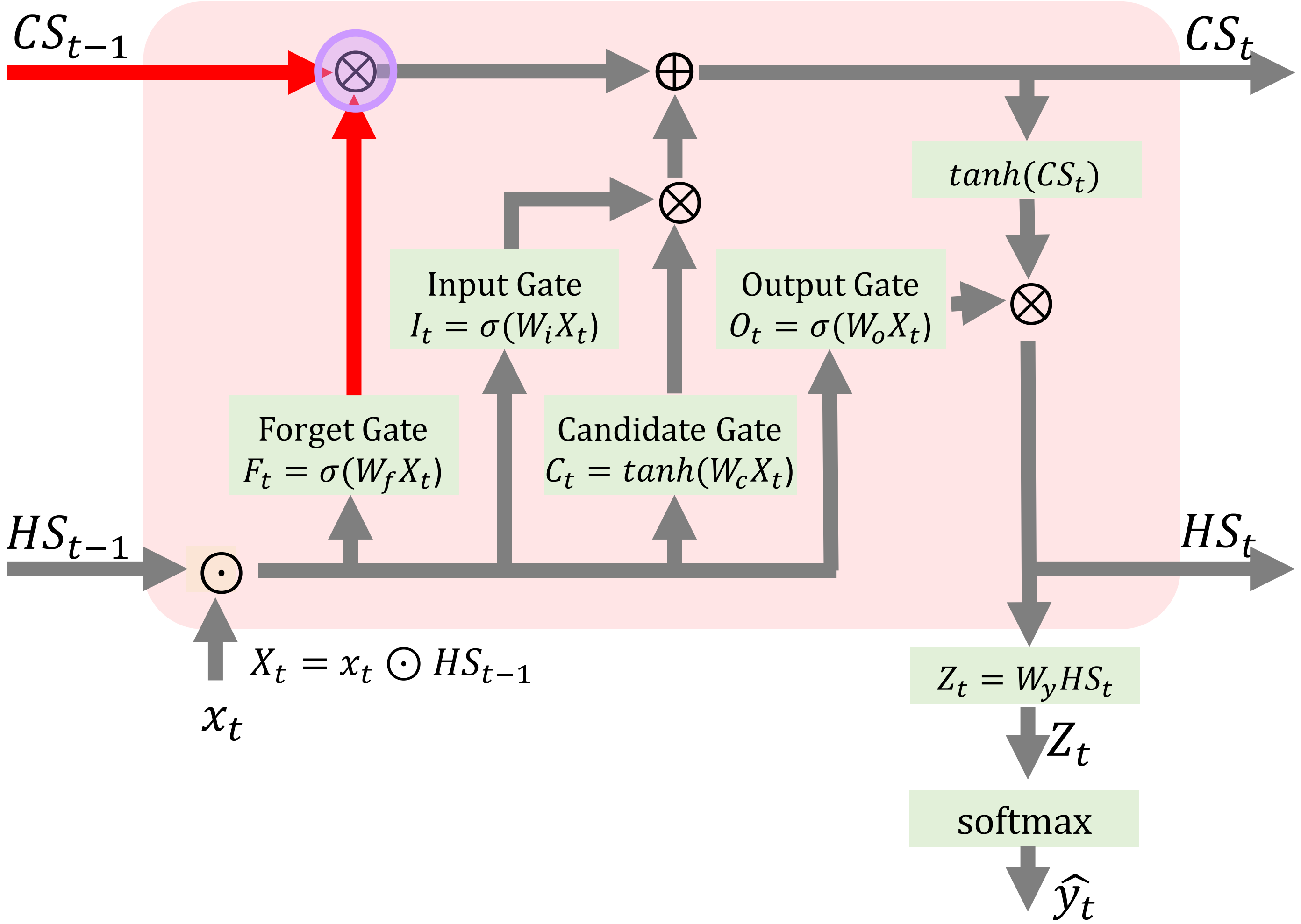
신박AI

셀상태의 원소값은 0으로 바뀌게 됩니다

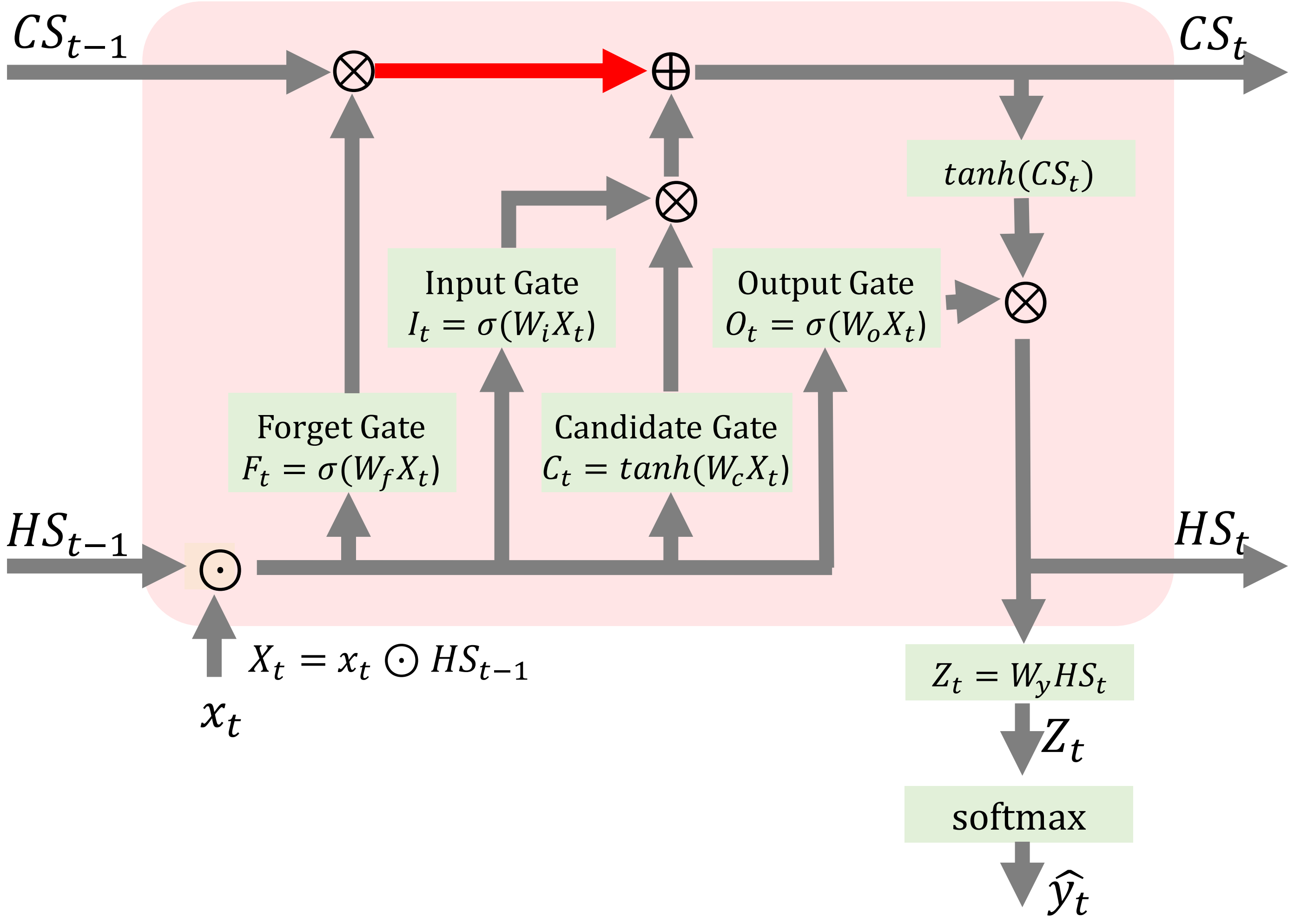


신박AI

셀상태의 값이 0이 되는 (혹은 그에 준하게 작아지는) 것을 망각(Forget)이라 정의합니다

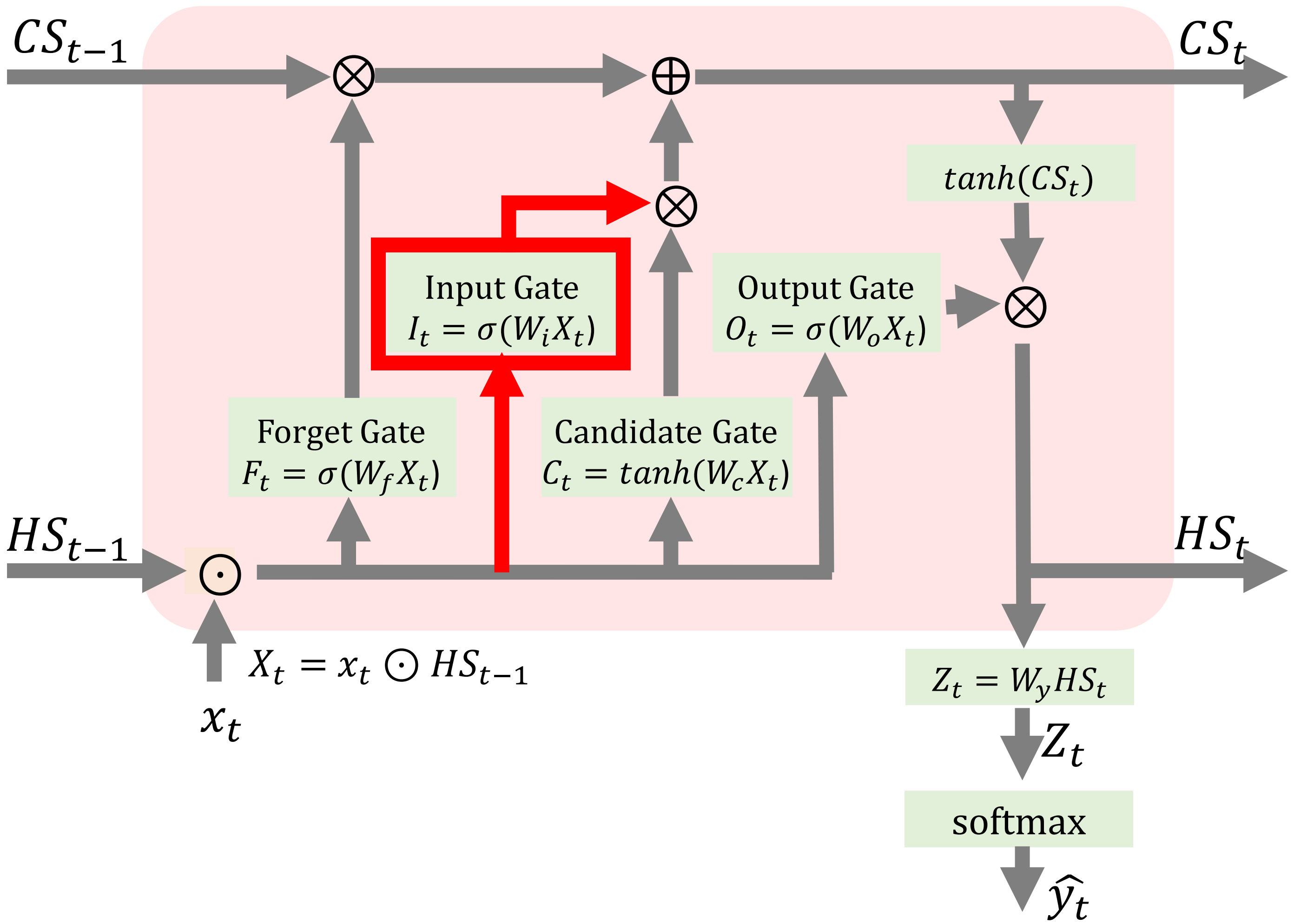


즉 셀상태는 망각게이트를 지나면서 잊어버려야할 것들을 잊어버리게 됩니다

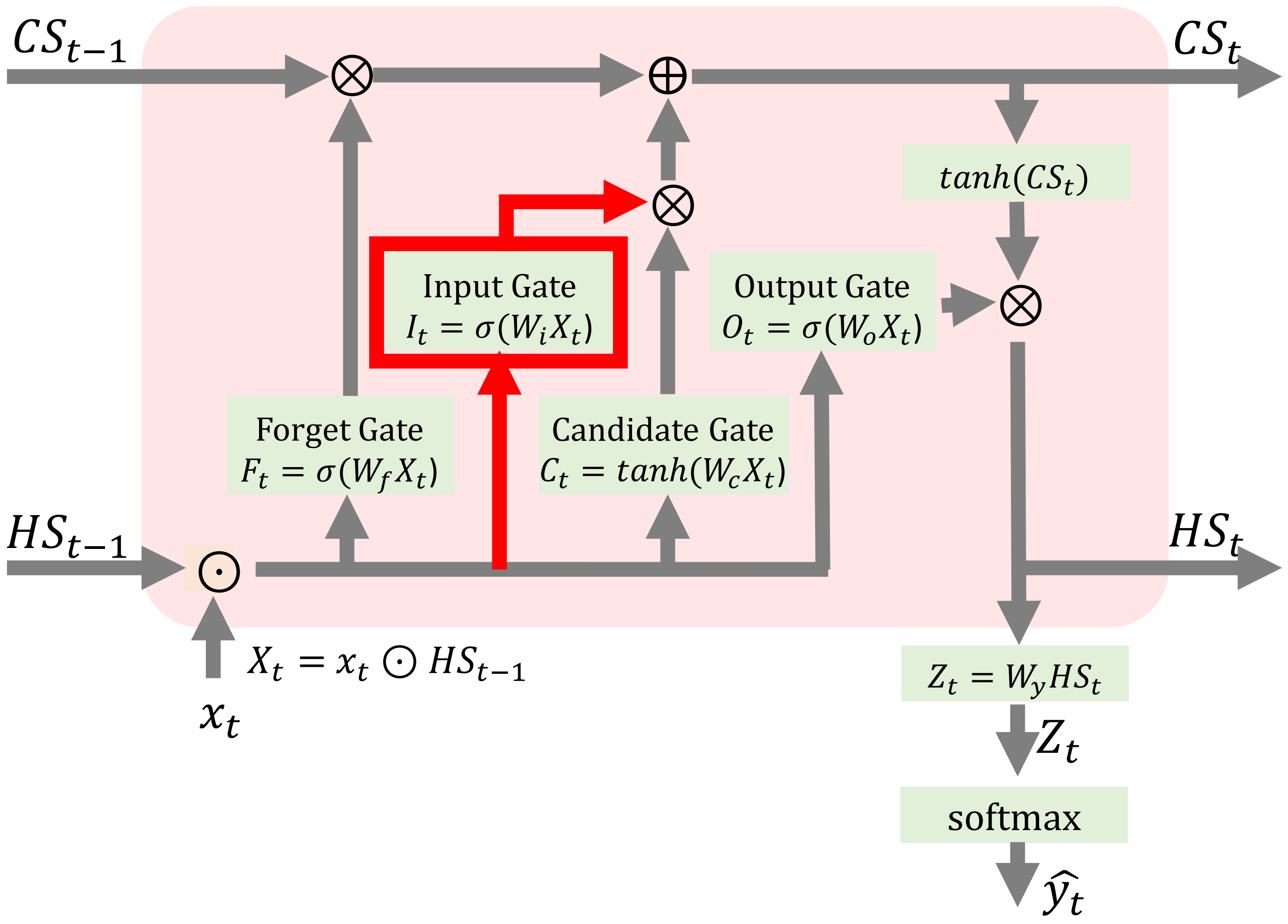


신박AI

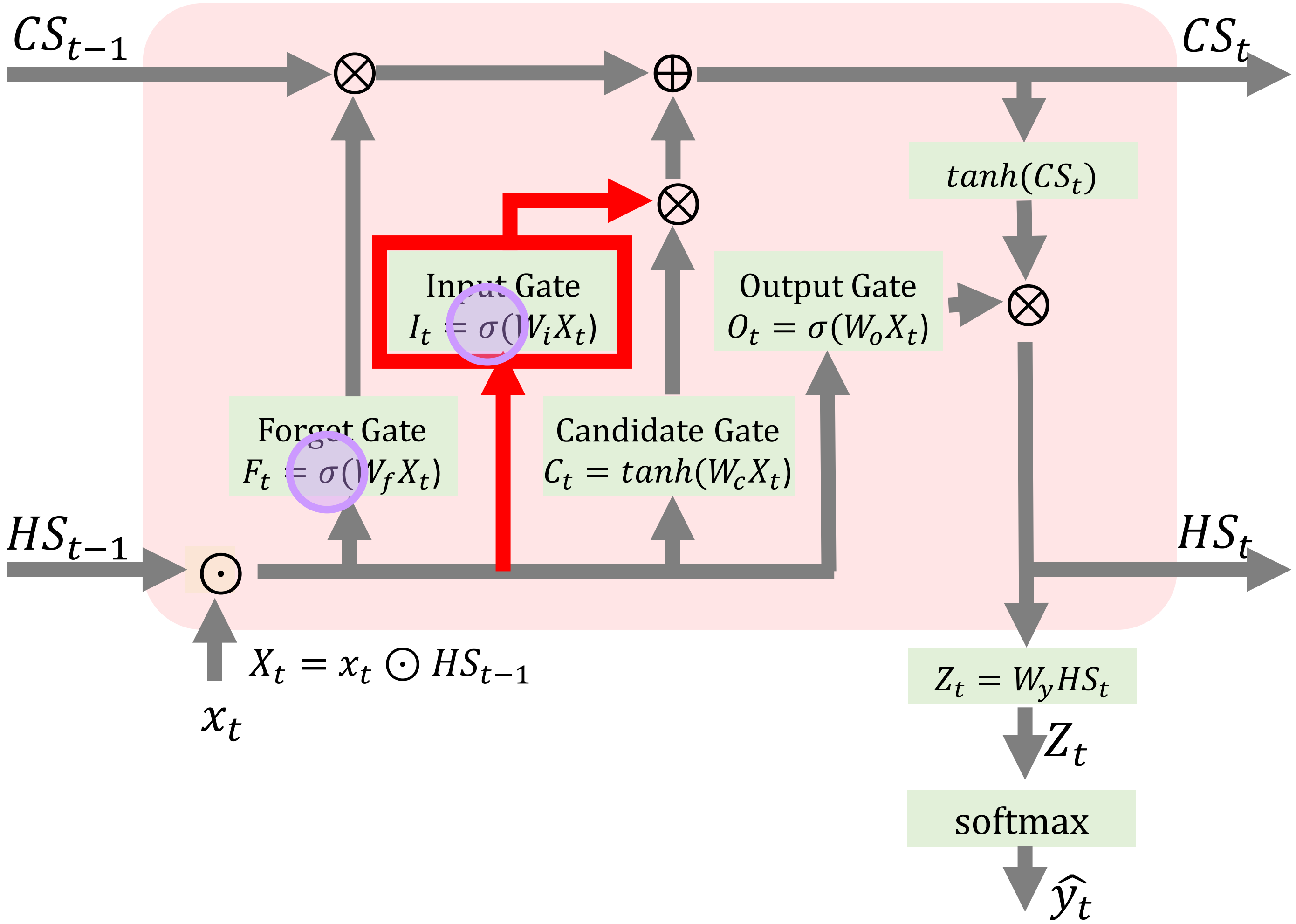
그 다음 Input Gate를 알아보겠습니다



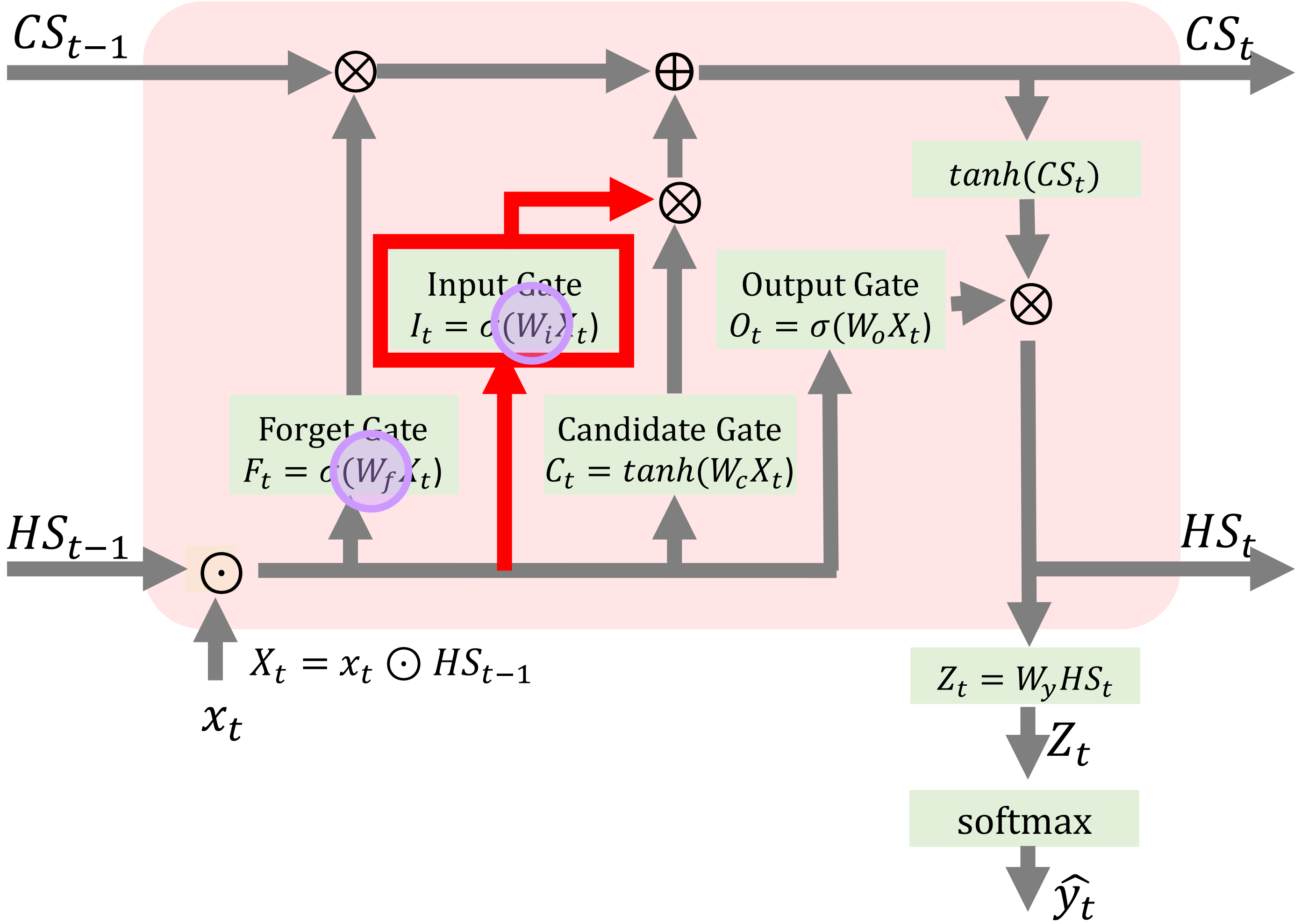
Input Gate는 실질적으로 Forget Gate와 연산과정은 동일합니다



왜냐하면 둘다 같은 시그모이드 함수를 사용하기 때문입니다

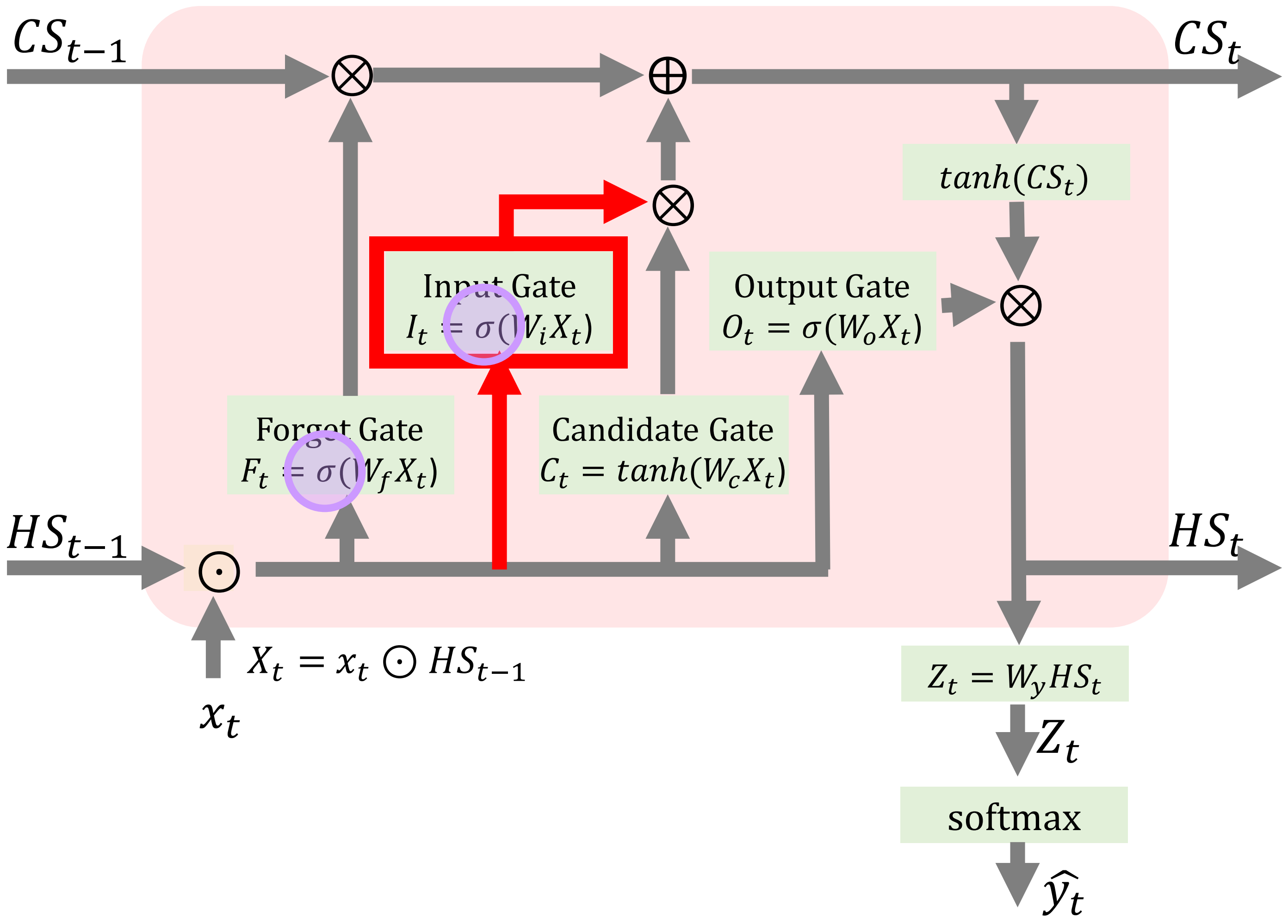


다만 가중치값만 다를 뿐입니다

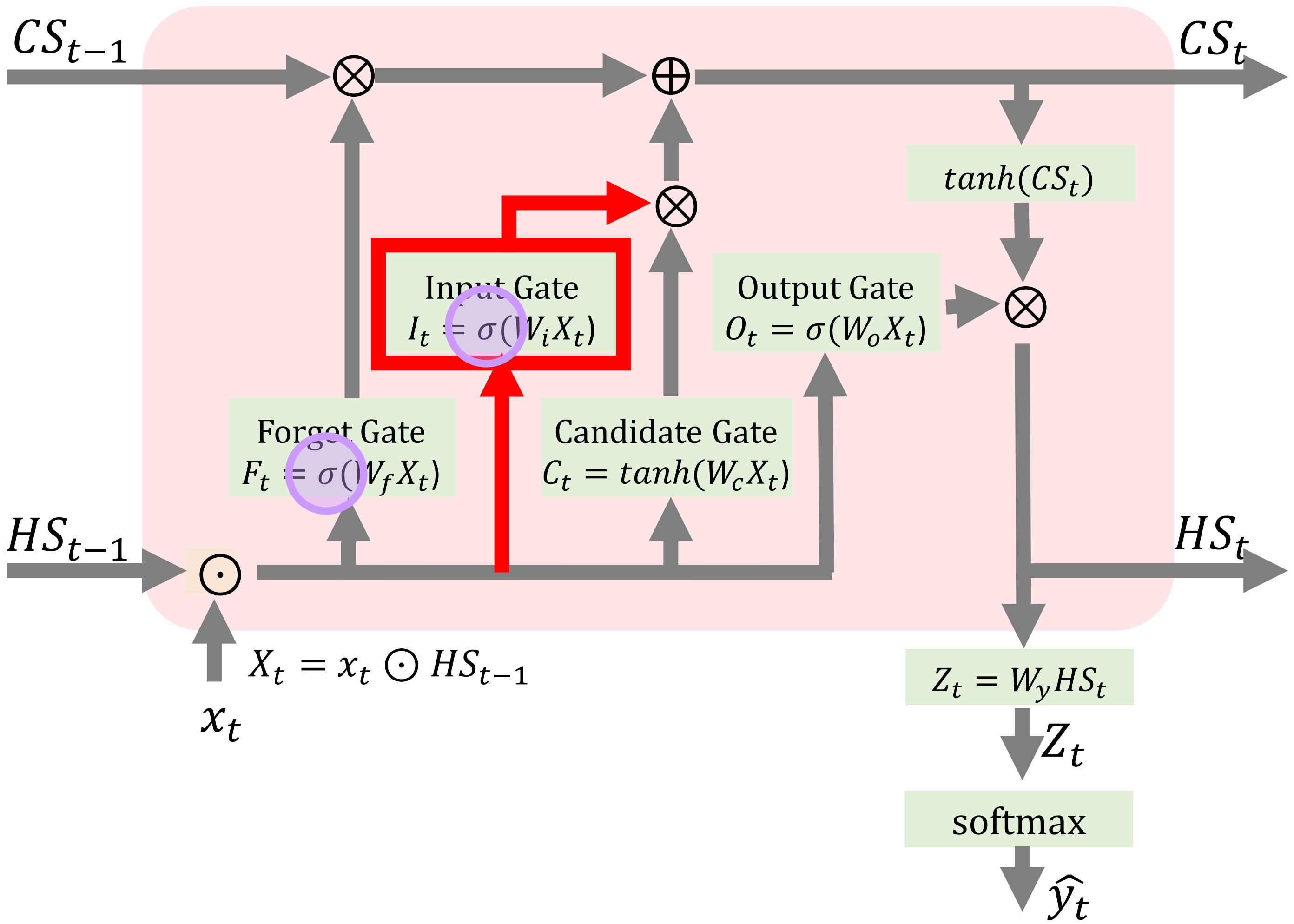


신박AI

생각해보면, 무엇을 망각할 것이냐와 무엇을 기억할 것이냐는

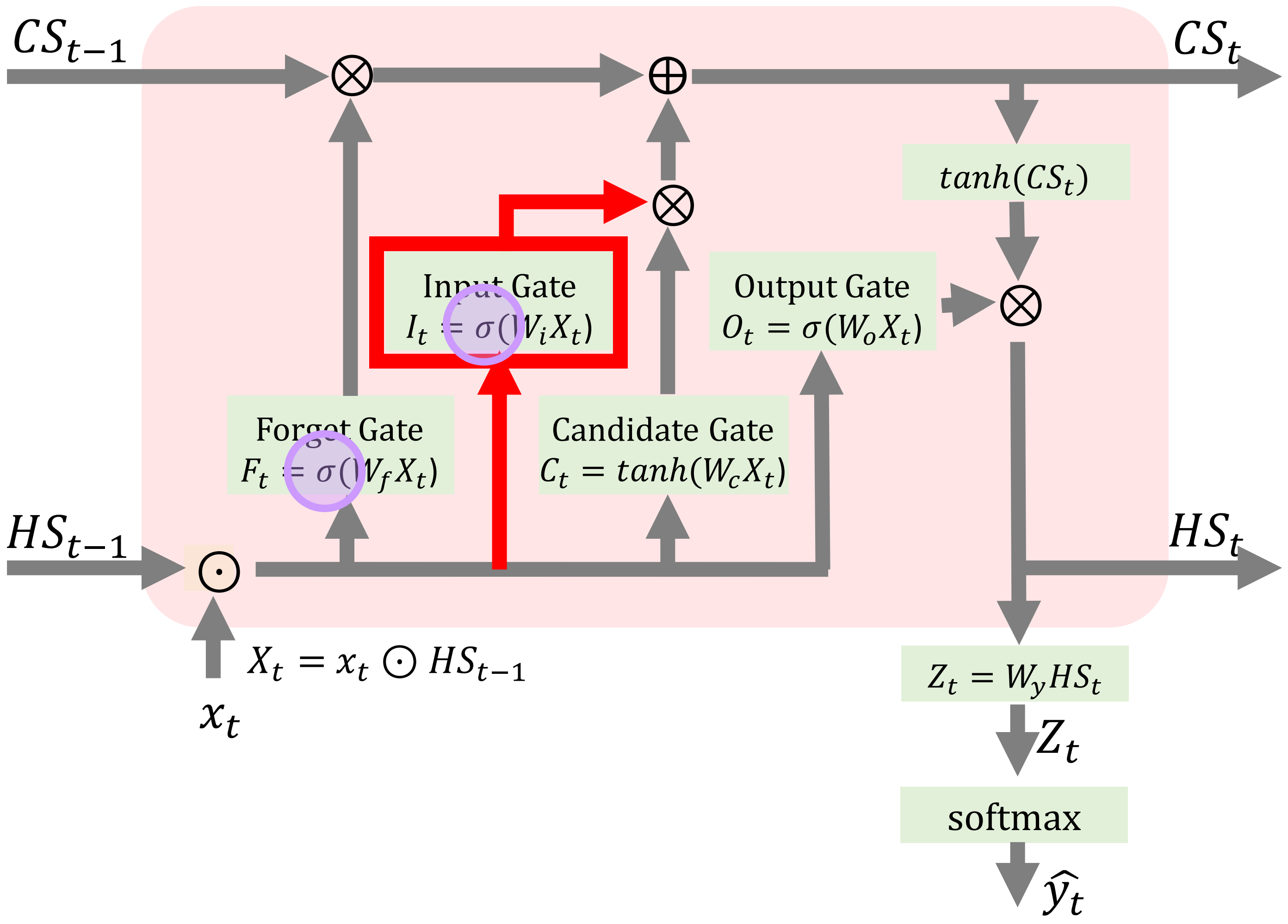


밤이 아니면 낮인 것처럼, 여자가 아니면 남자인 것처럼

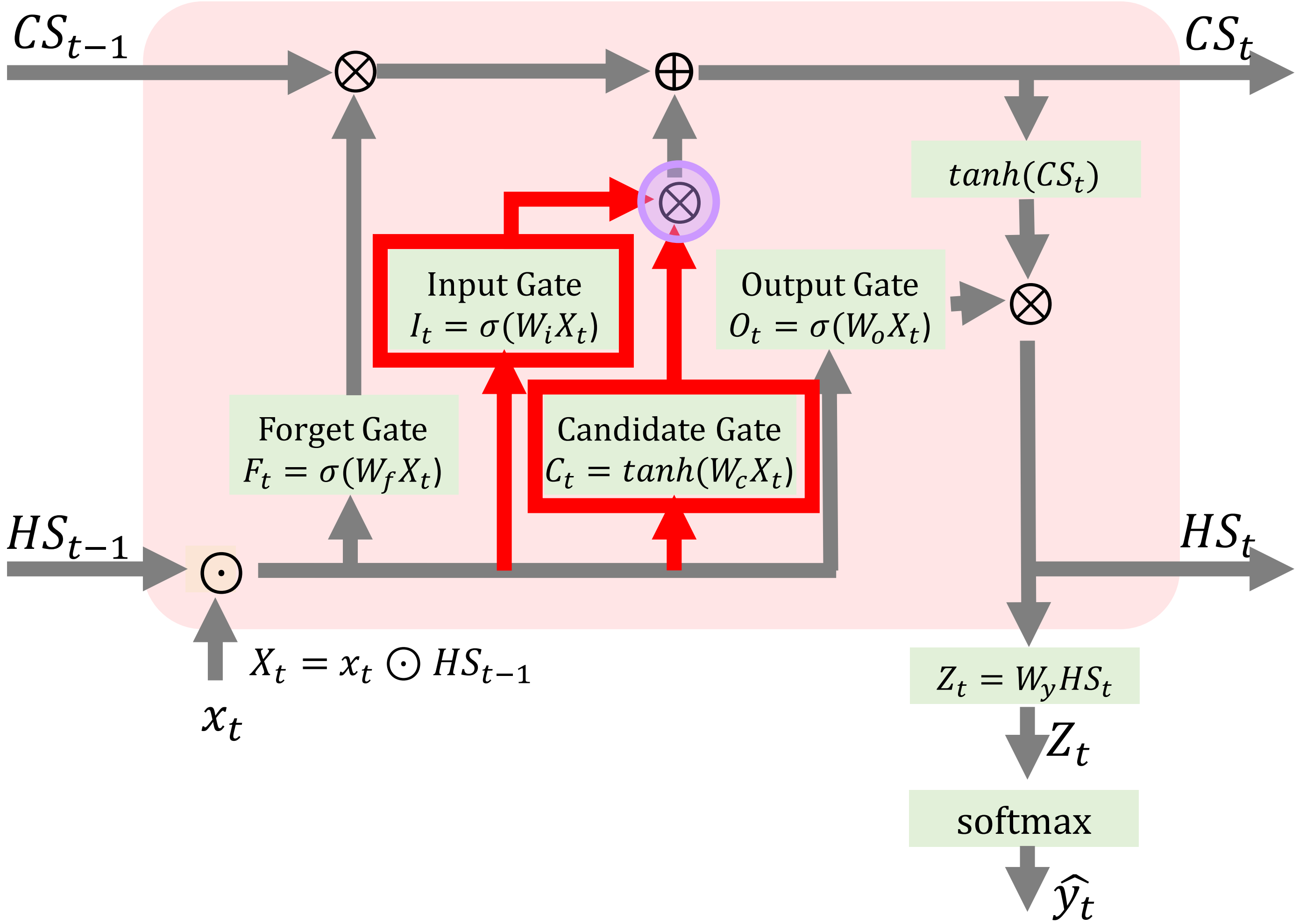


신박AI

사실상 의미적으로 같은 것입니다

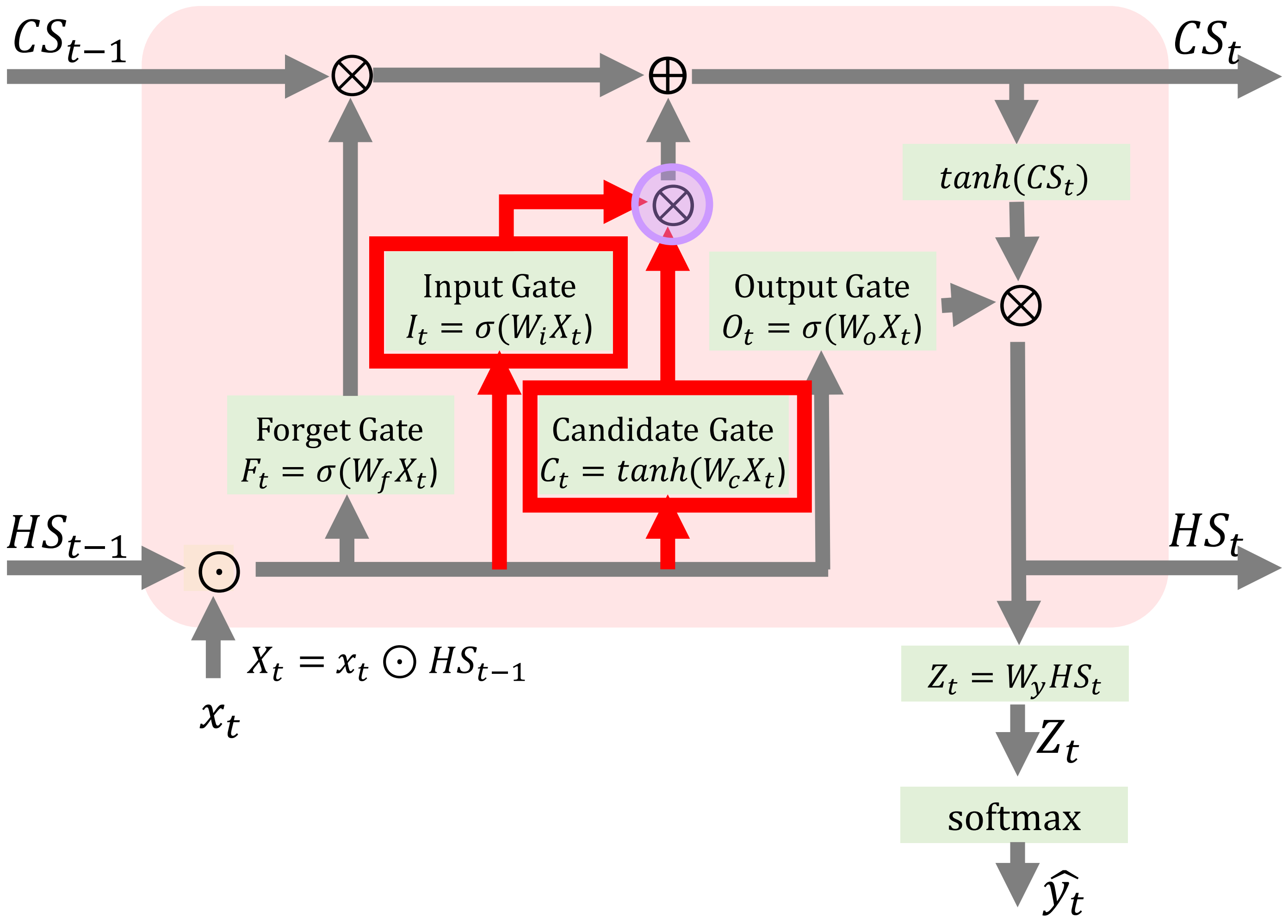


다만 이 Input Gate는 Candidate Gate와 같이 연산하여,

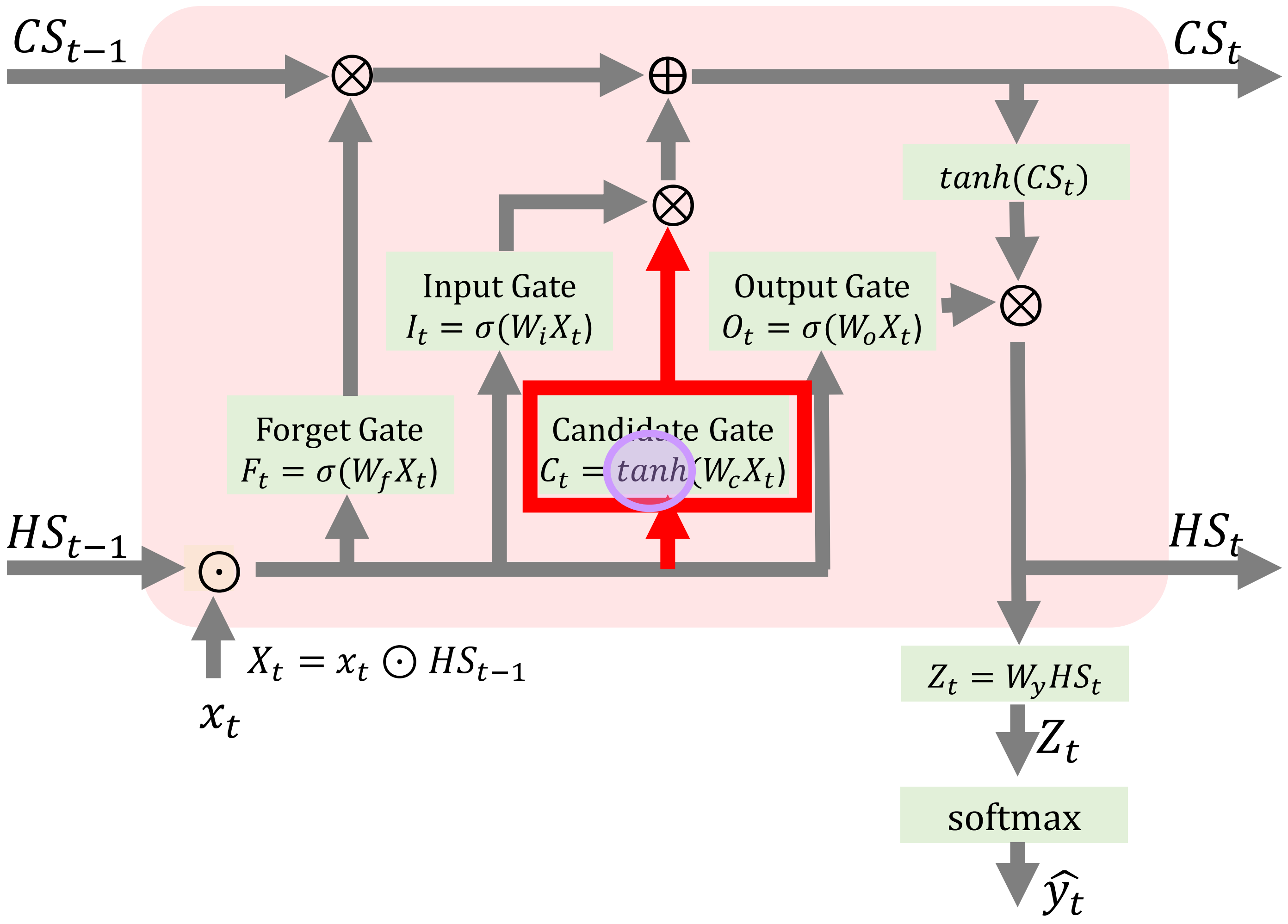


신박AI

셀상태를 '기억'해야할 것들로 업데이트 합니다

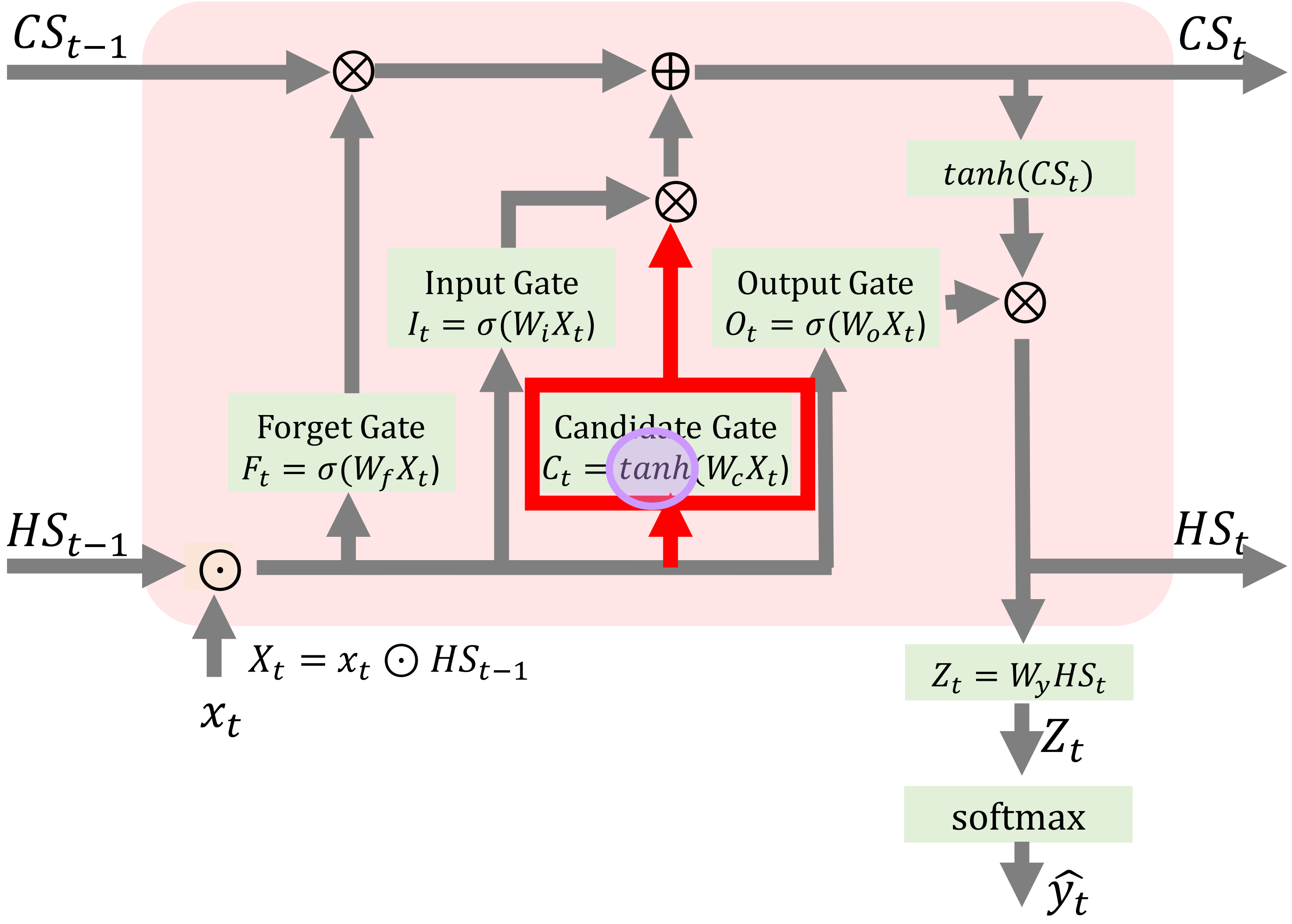


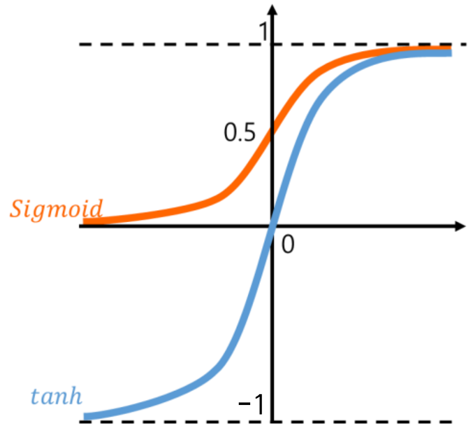
Candidate Gate는 내부 연산이 시그모이드가 아닌 tanh 함수입니다



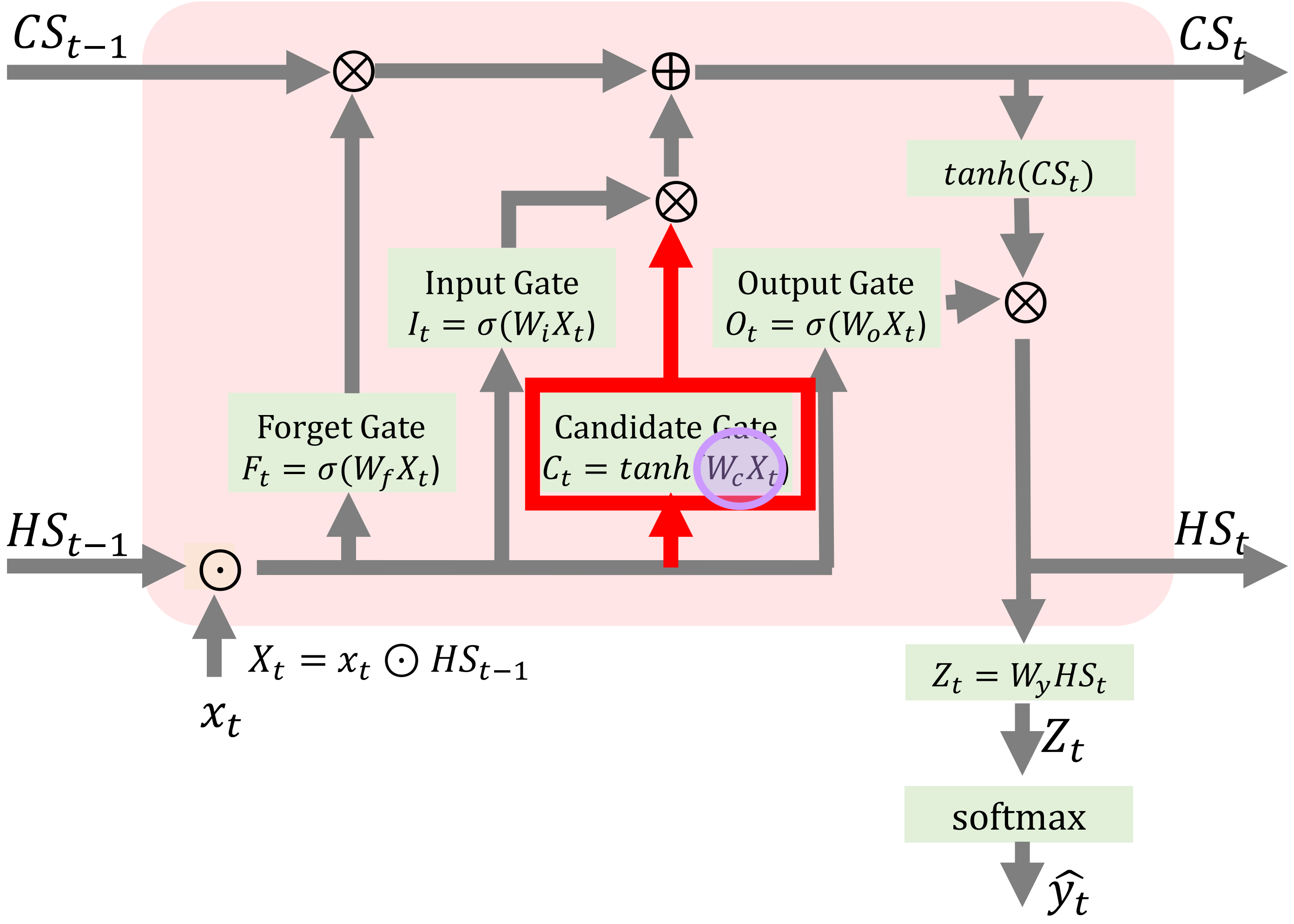
신박AI

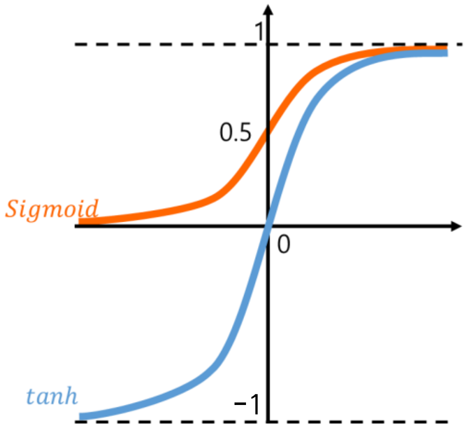
지난 영상에서 보셨듯, tanh함수는 들어오는 값을 −1과 1 사이 값으로 바꾸어줍니다



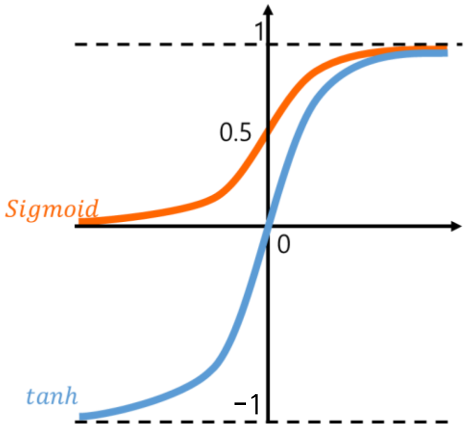


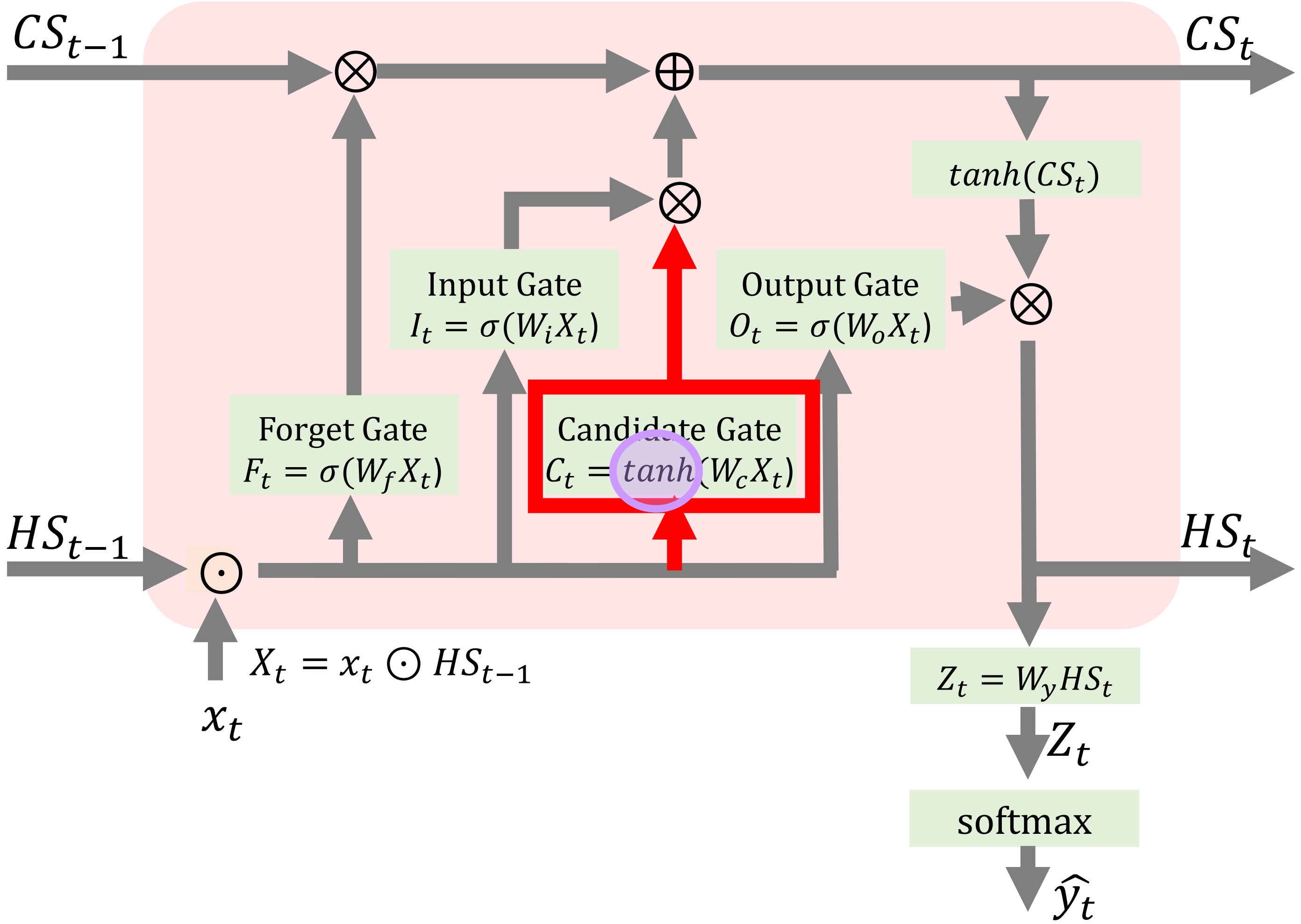
즉 Candidate Gate가 하는 일은, 입력값에 가중치를 곱한 뒤,





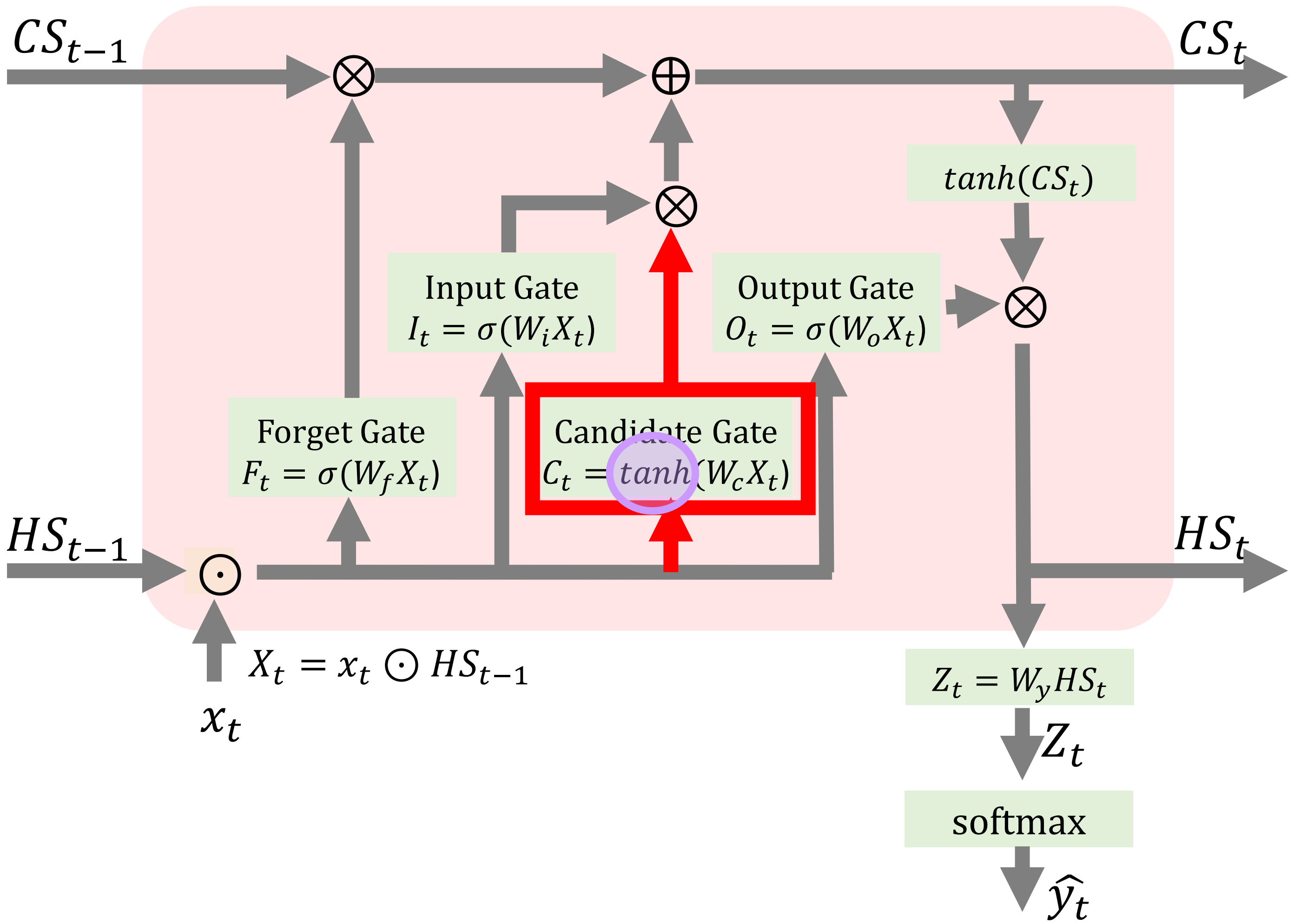
그 계산값이 마이너스 인 것은 그대로 마이너스로,

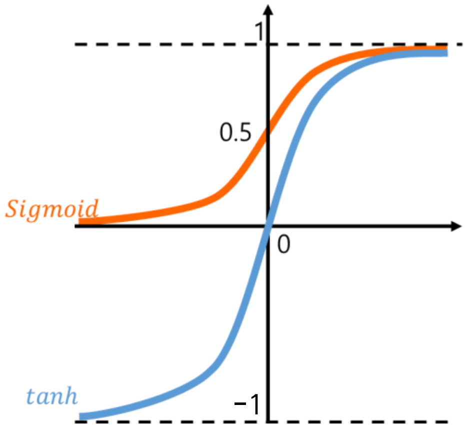




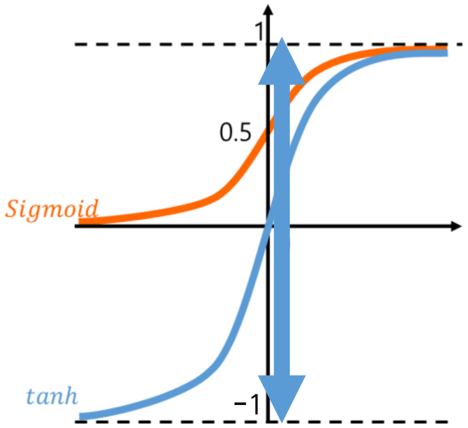
신박AI

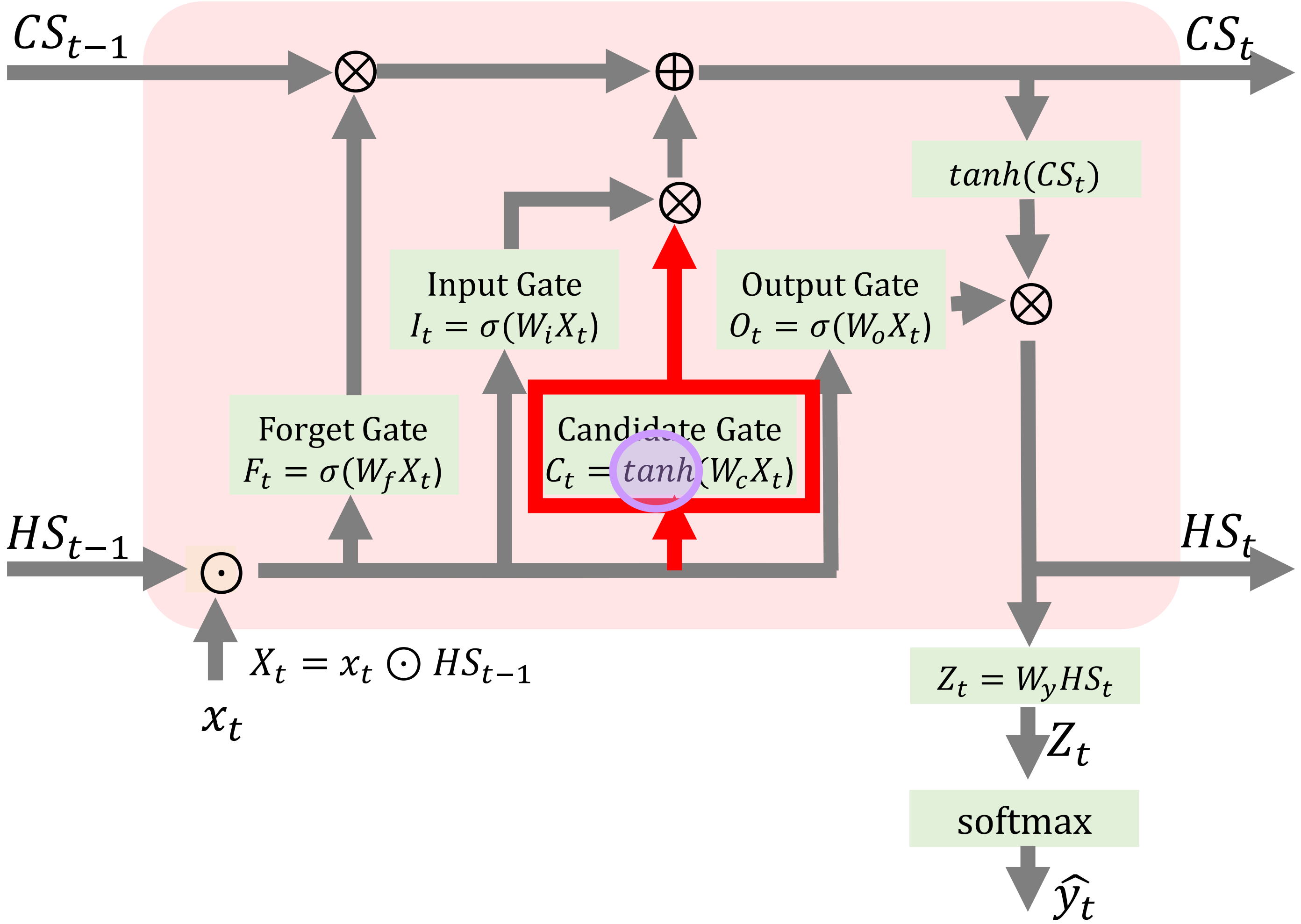
또 플러스 인 것은 그대로 플러스로 이렇게 극성은 보존하되,



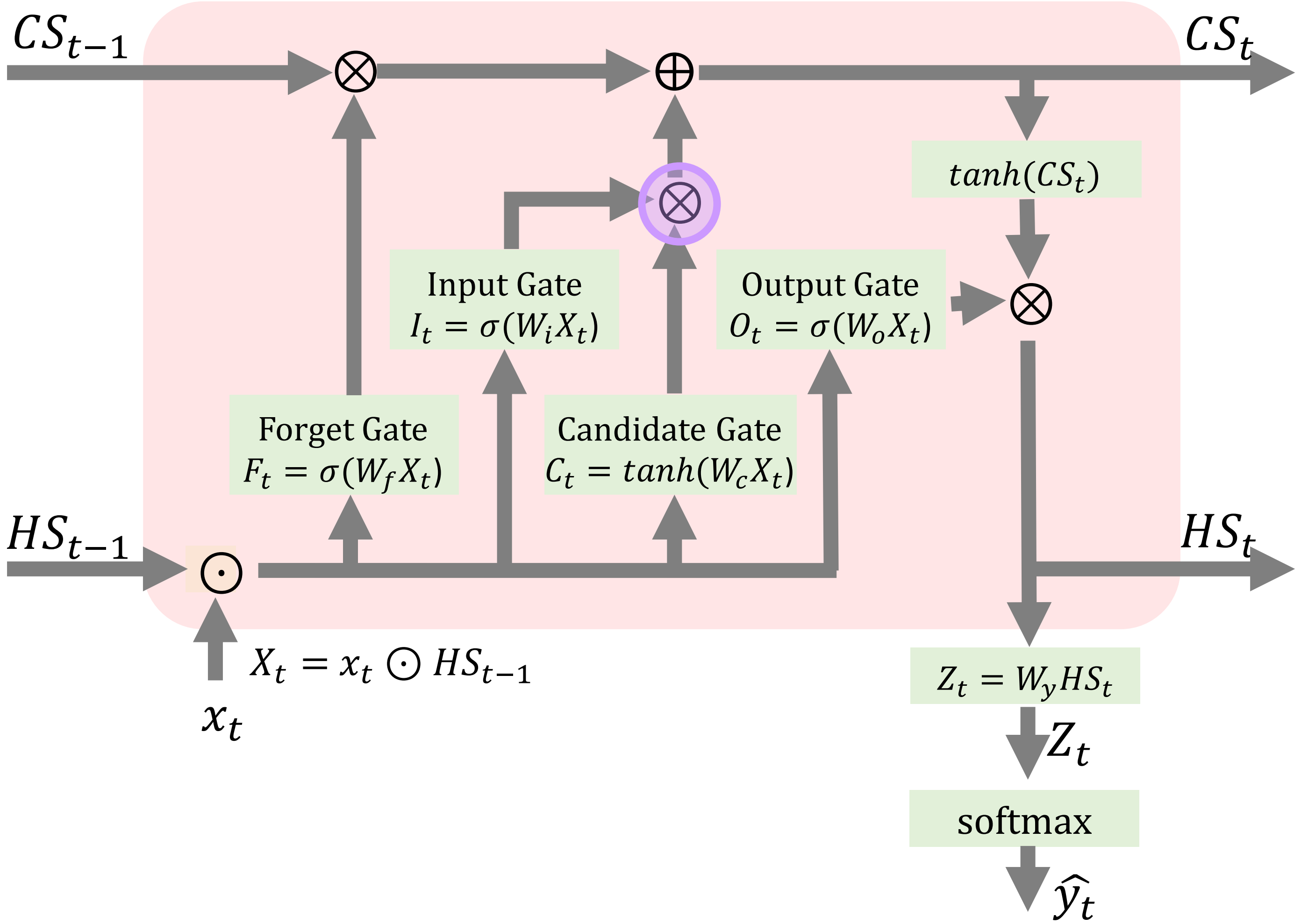


범위를 -1에서 1사이가 되도록 정규화하는 역할이라고 보시면 되겠습니다



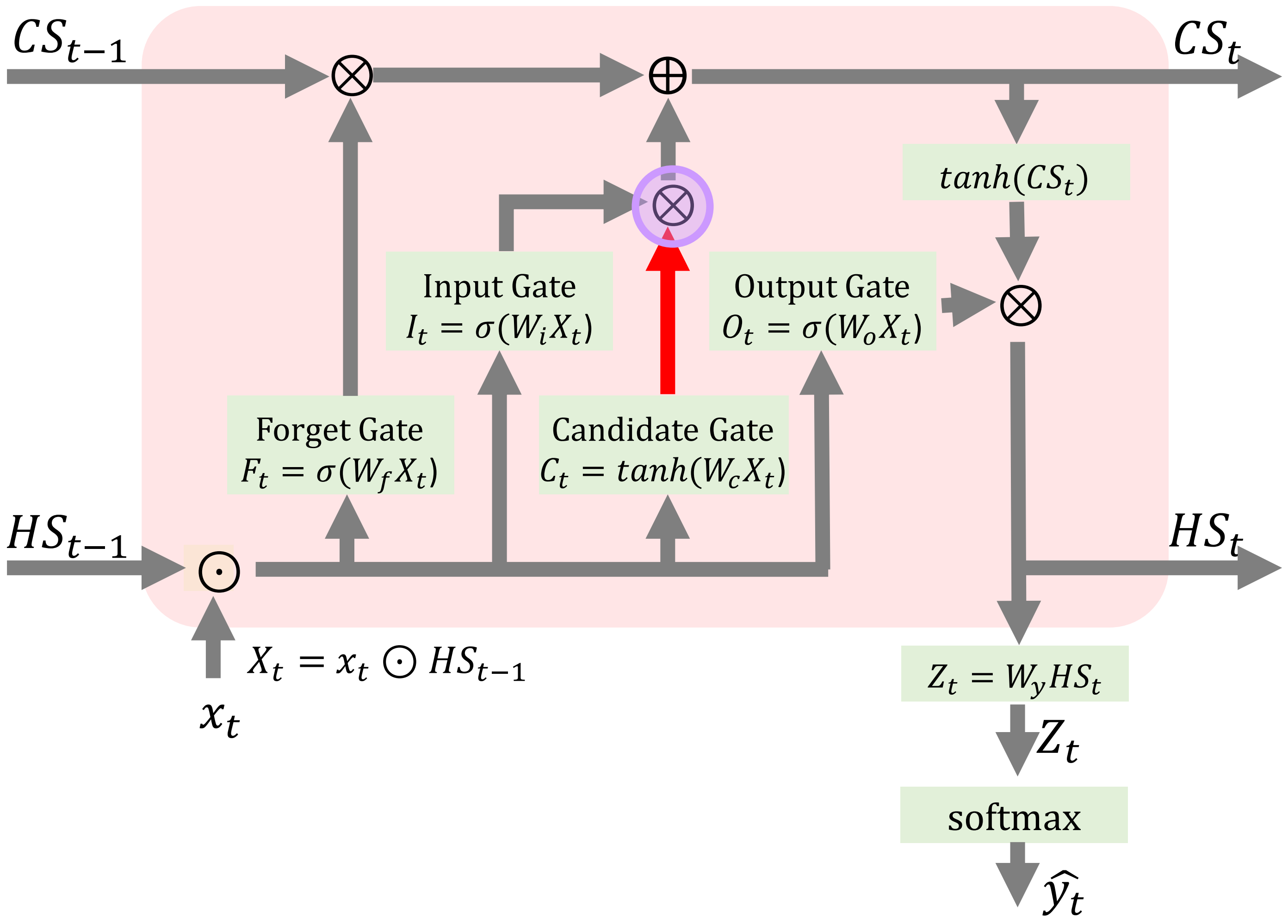


그런 다음 여기 Input Gate에서 나온 0과 1 사이 값들과 element-wise연산을 통해서



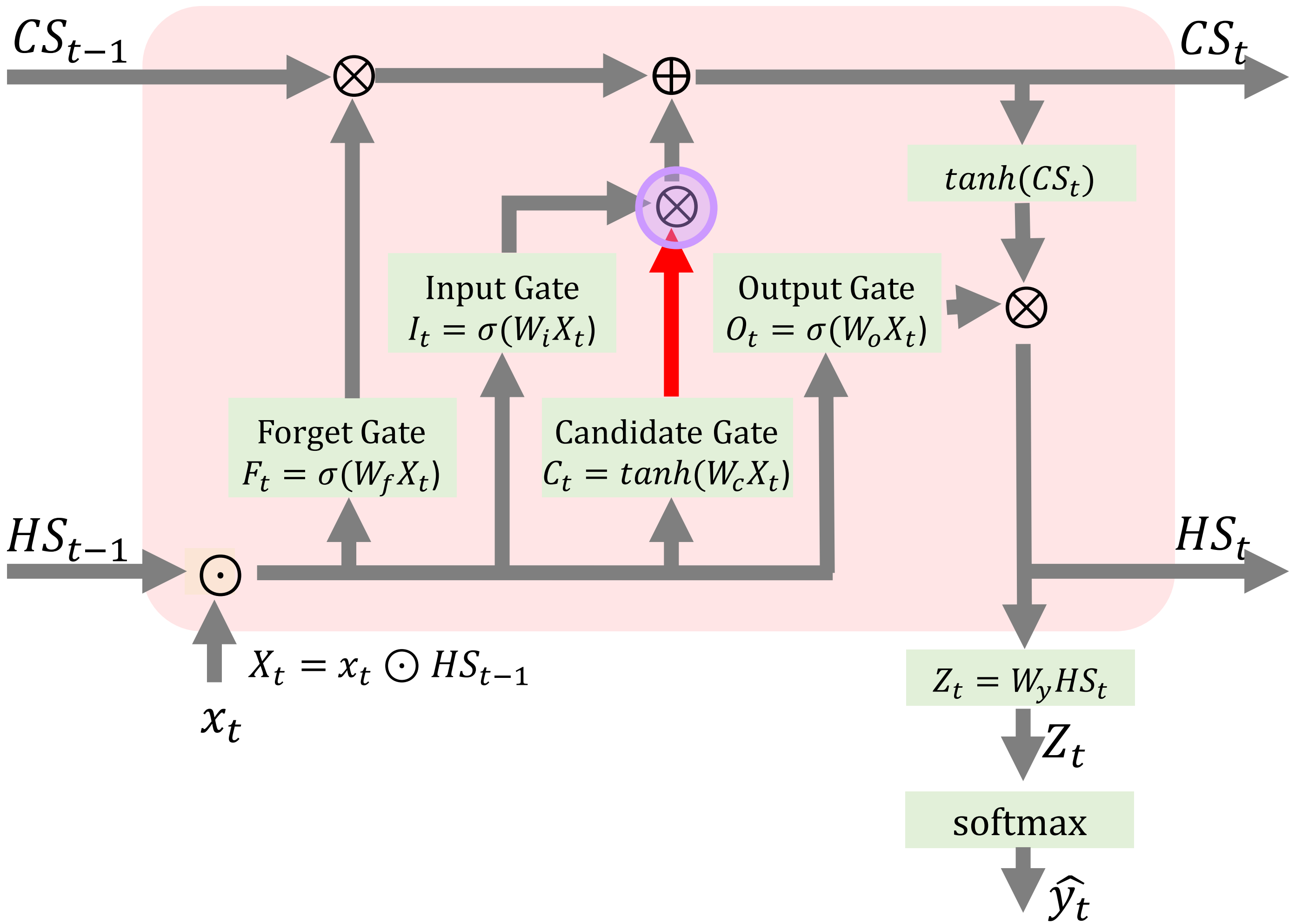
신박AI

Candidate Gate에서 나온 값들 중 어떤 값들은 0에 가깝게 만들고



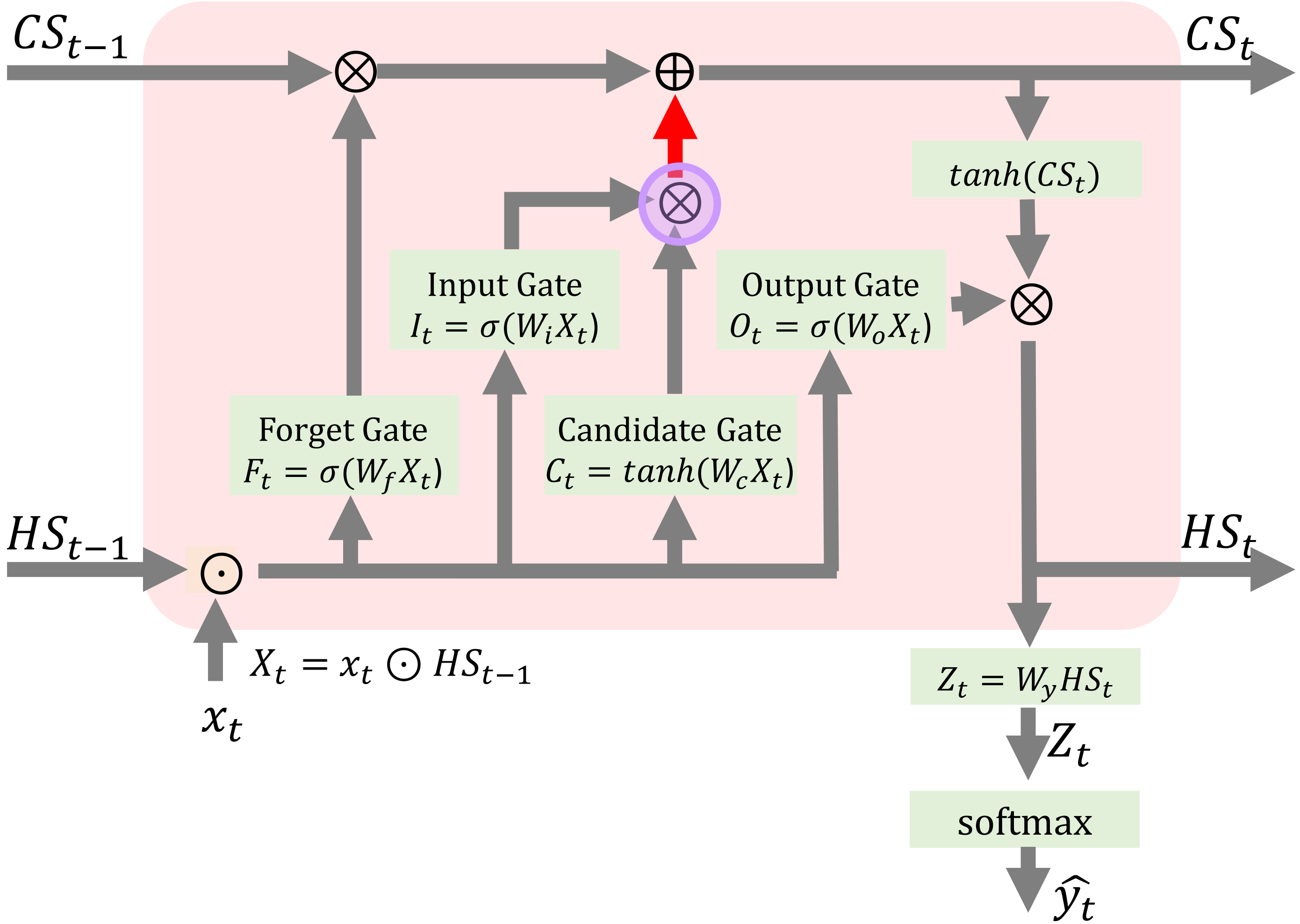
신박AI

어떤 값들은 그대로 놔두는 역할을 합니다

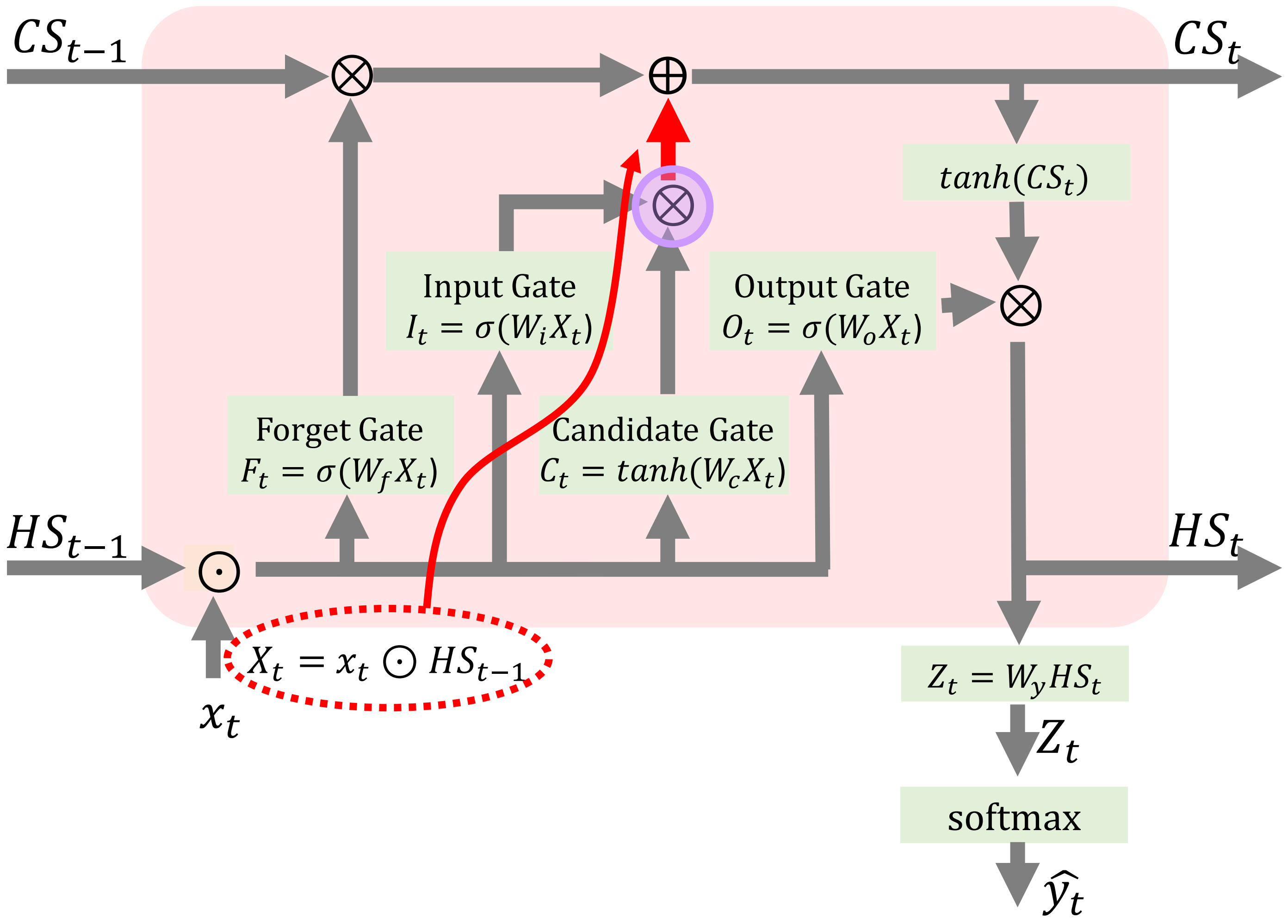


신박AI

이렇게 그대로 놔두게 되는 값들의 의미가 말하자면,

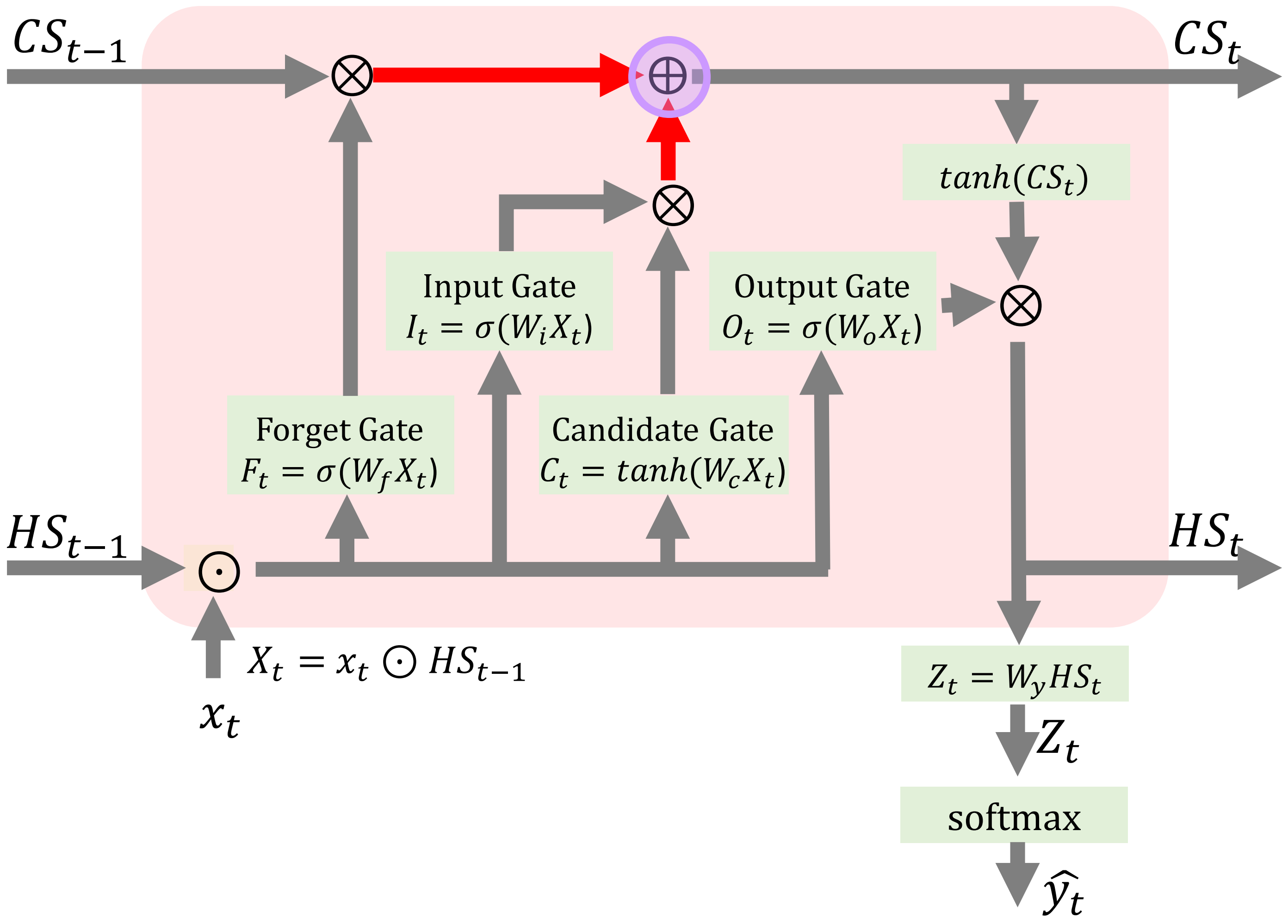


현재 입력 (short-term)중 기억할 부분이 되겠습니다

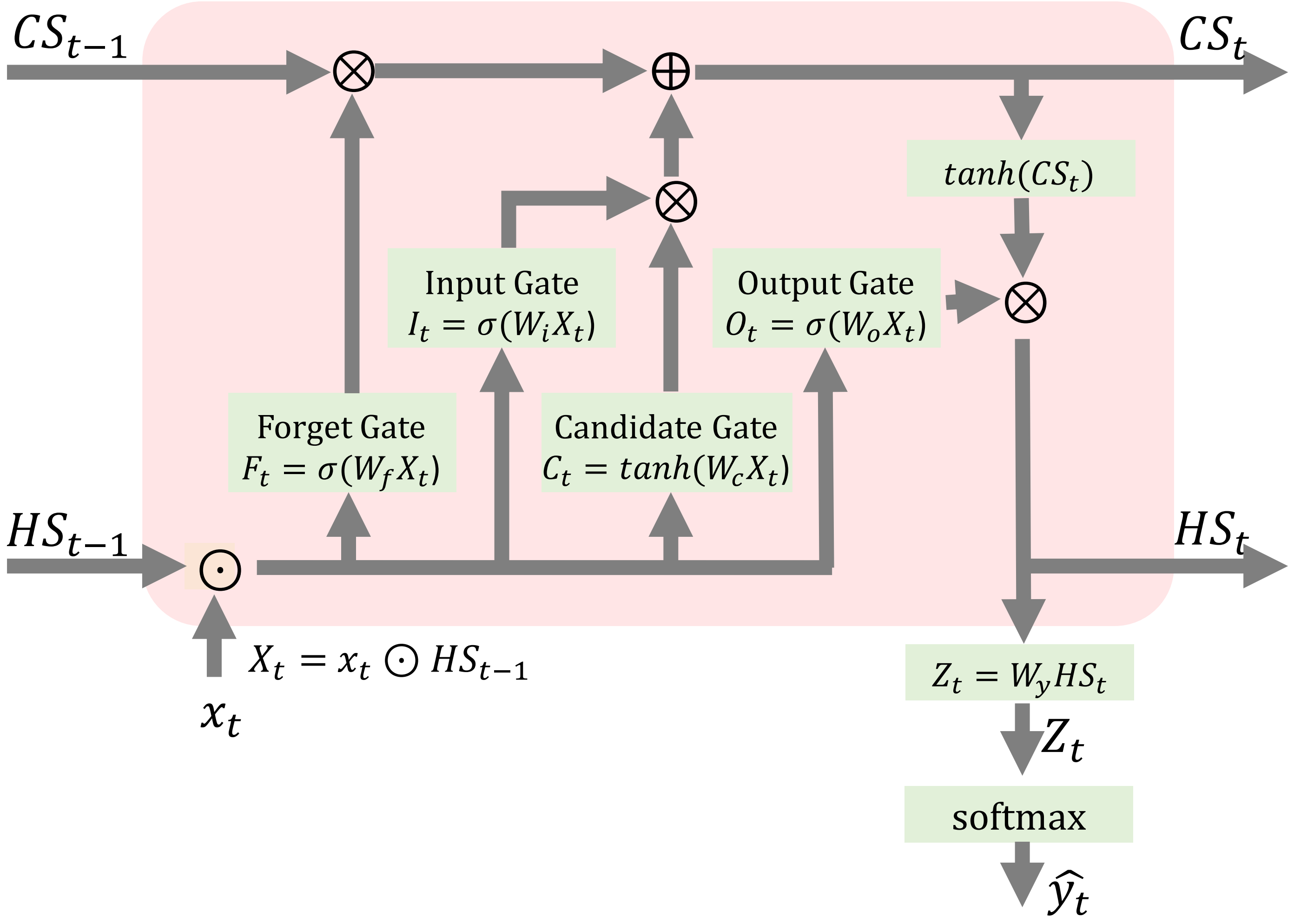


신박AI

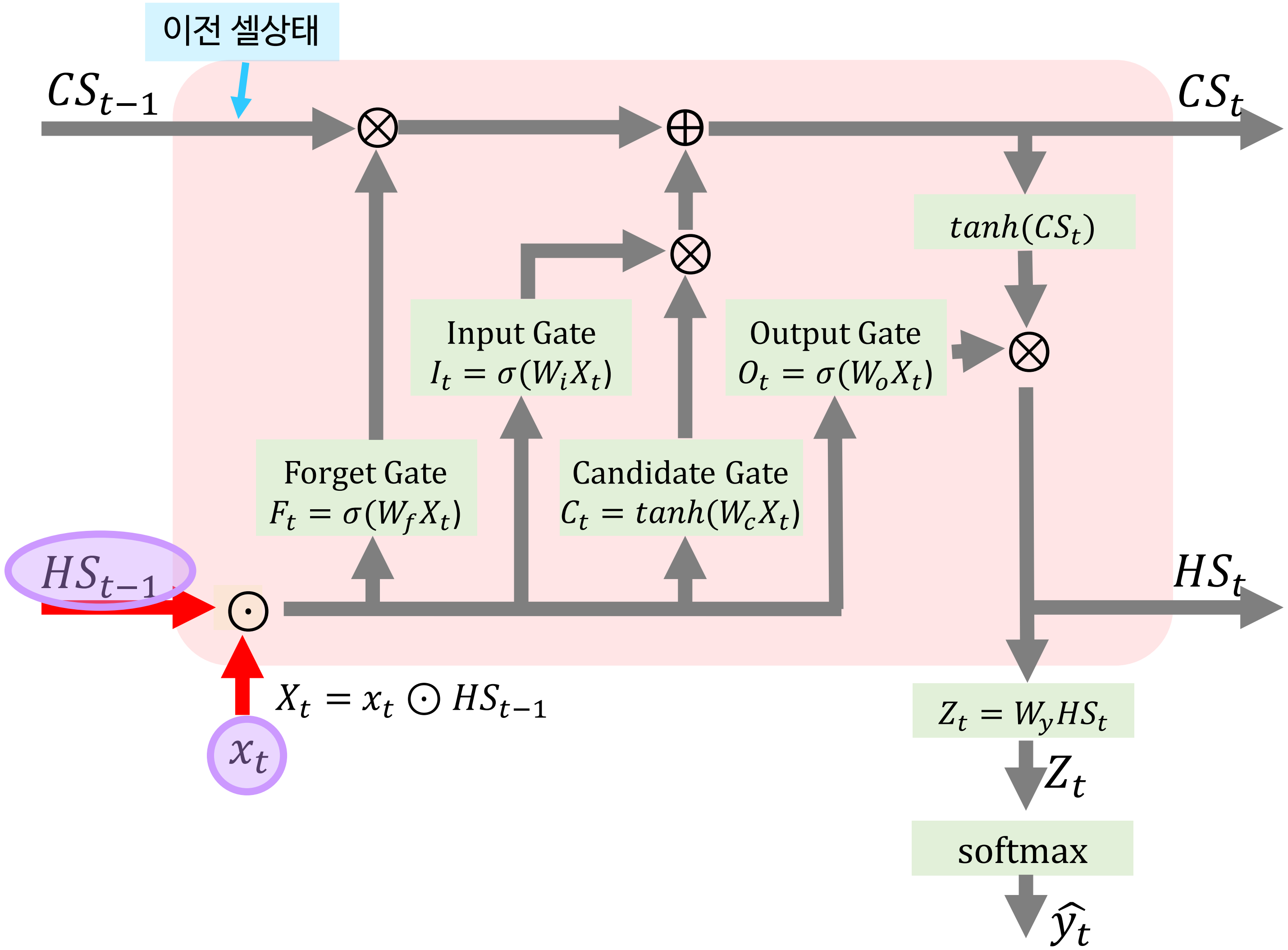
그런 다음 그 남은 값들을 셀상태에 더하여 업데이트 하게 됩니다



자 여기까지 뭔가 장황하게 설명했지만, 결국 자세히 보면 LSTM이 하는 일은,

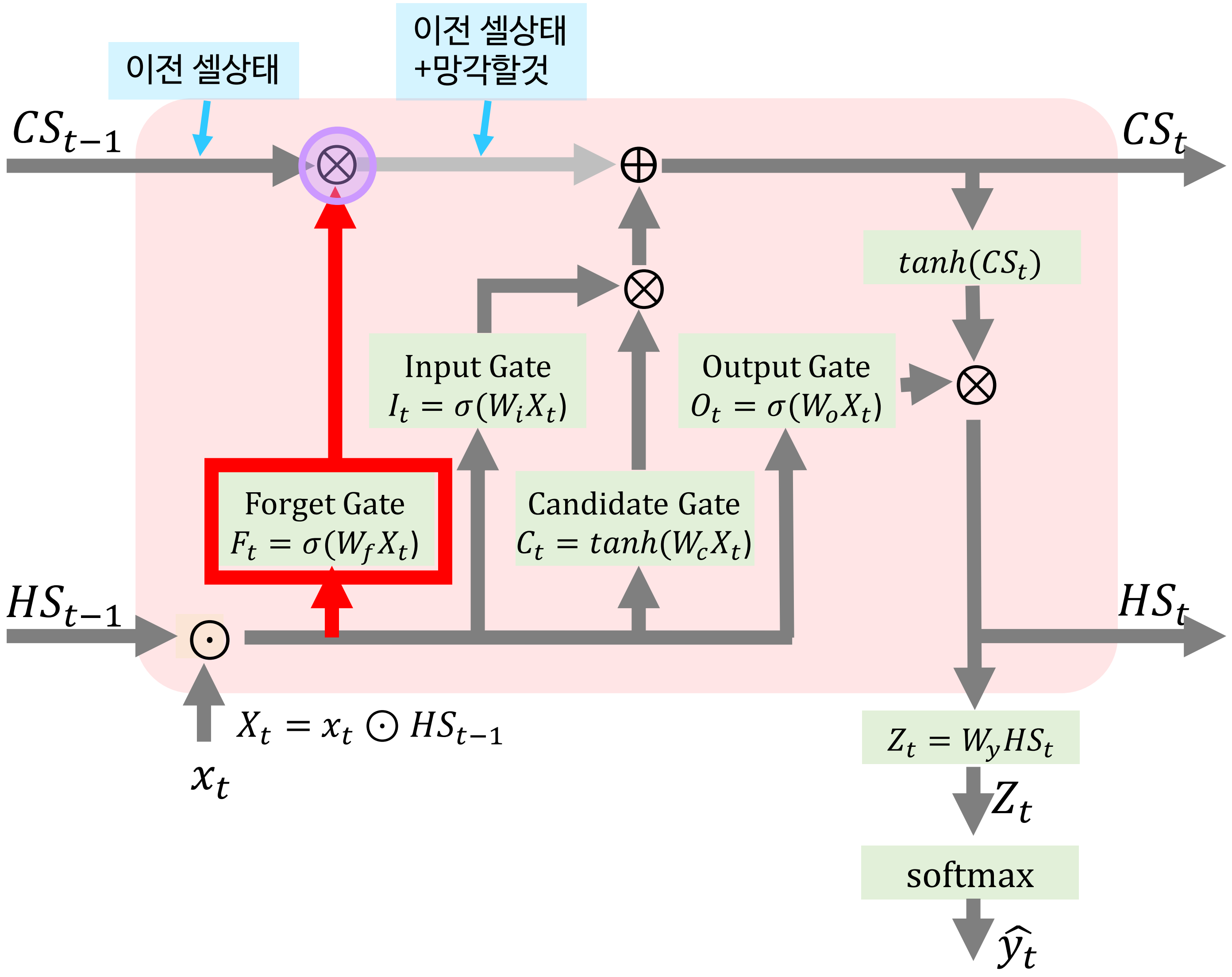


바로 이전 히든상태와 현재 입력값을 받아서,



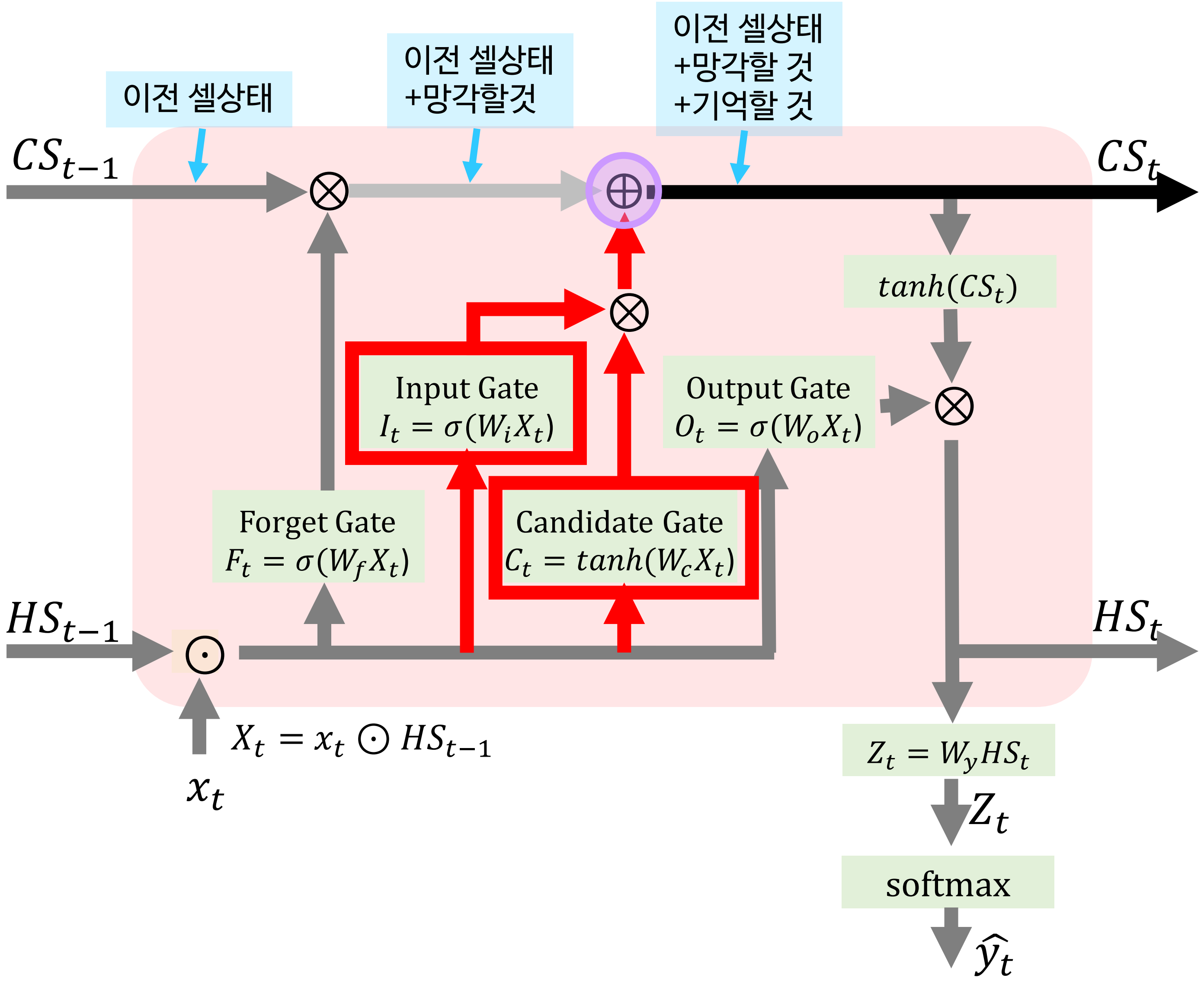
신박AI

이전 셀상태에서 망각할 것은 망각하고,



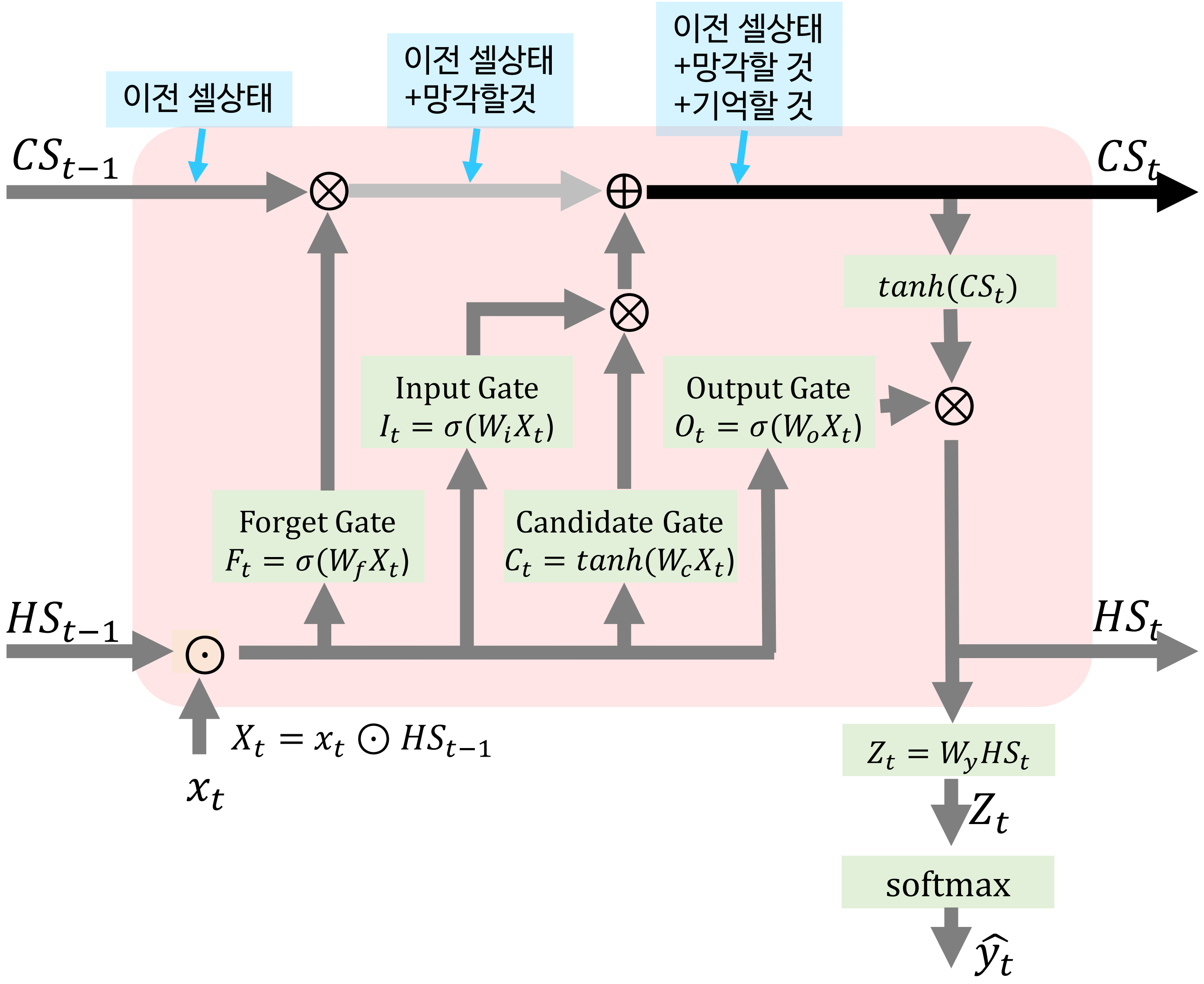
신박AI

기억할 것은 기억해서

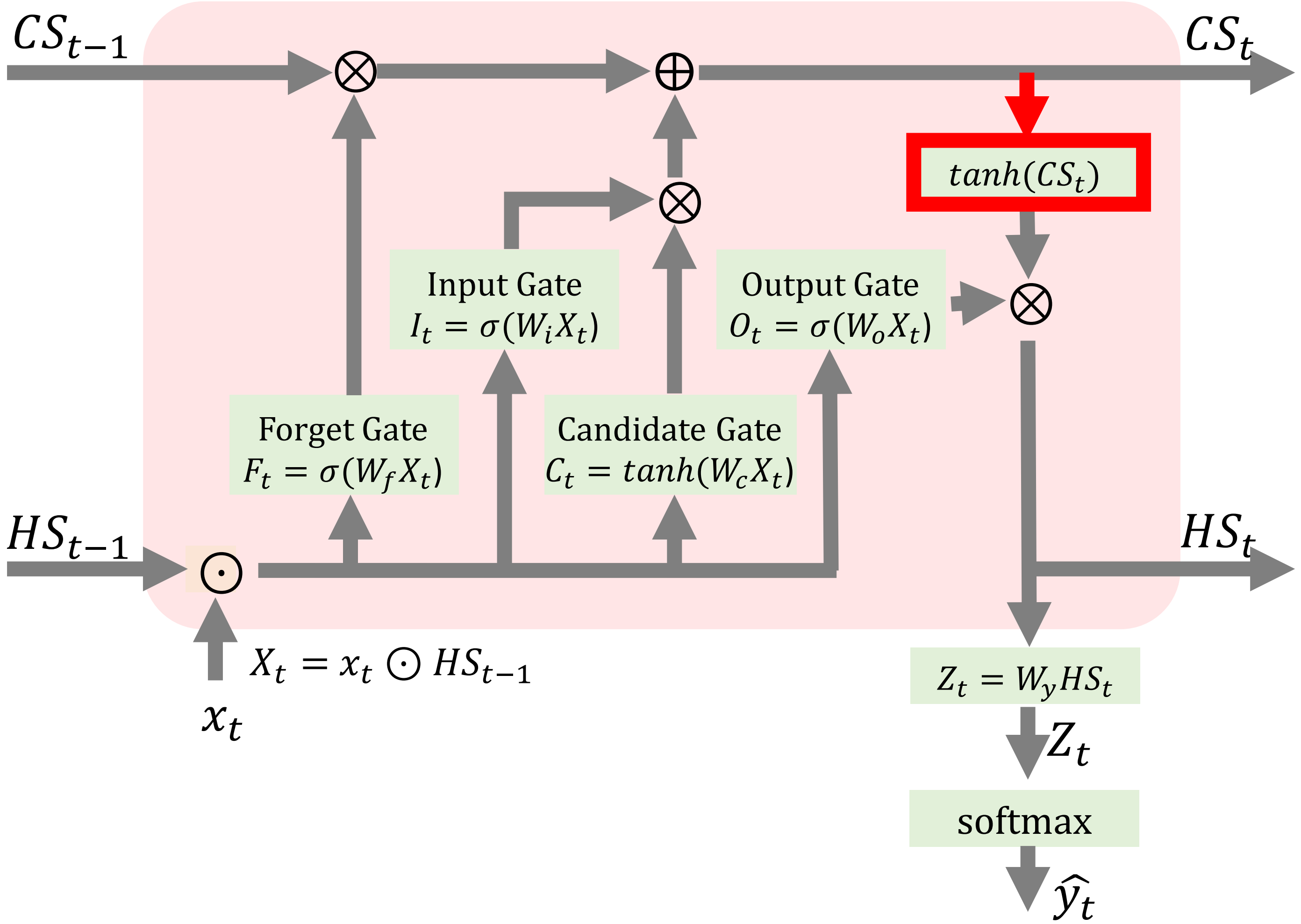


신박AI

LSTM의 장기기억 상태를 업데이트하는 것입니다

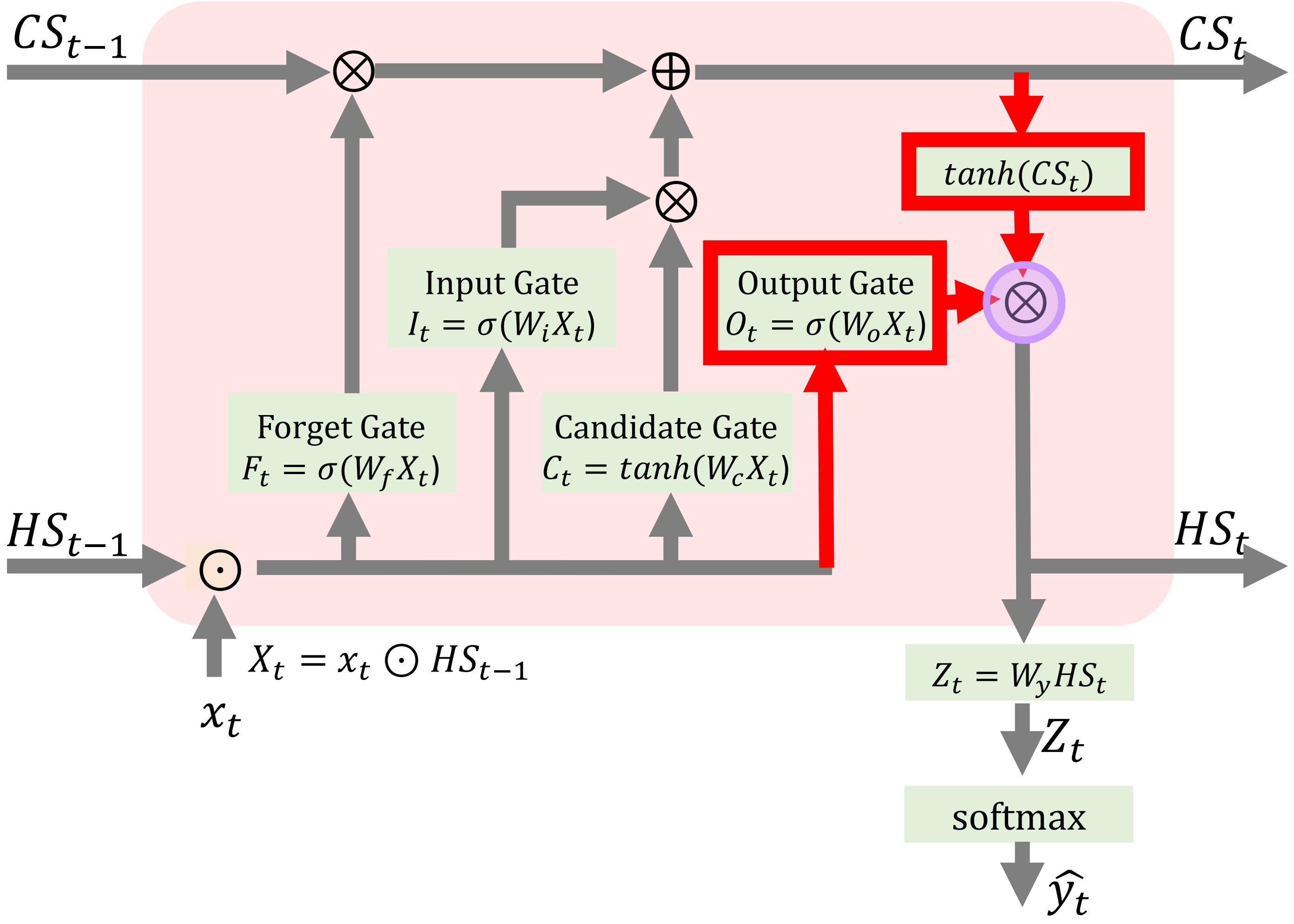


그 다음 이러한 장기기억 상태를 Tanh를 통해서 정규화 (-1~1) 한 다음,

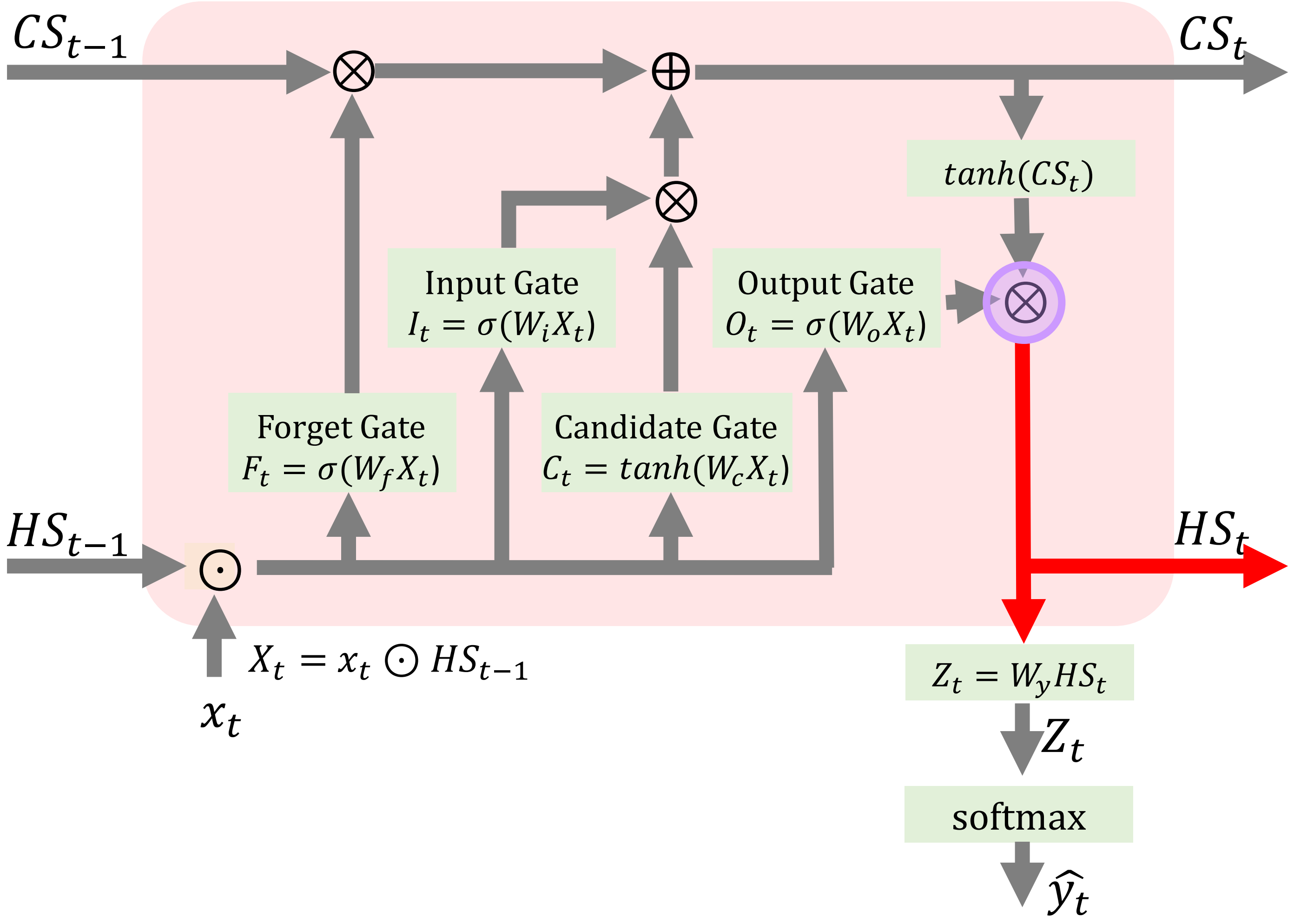


신박AI

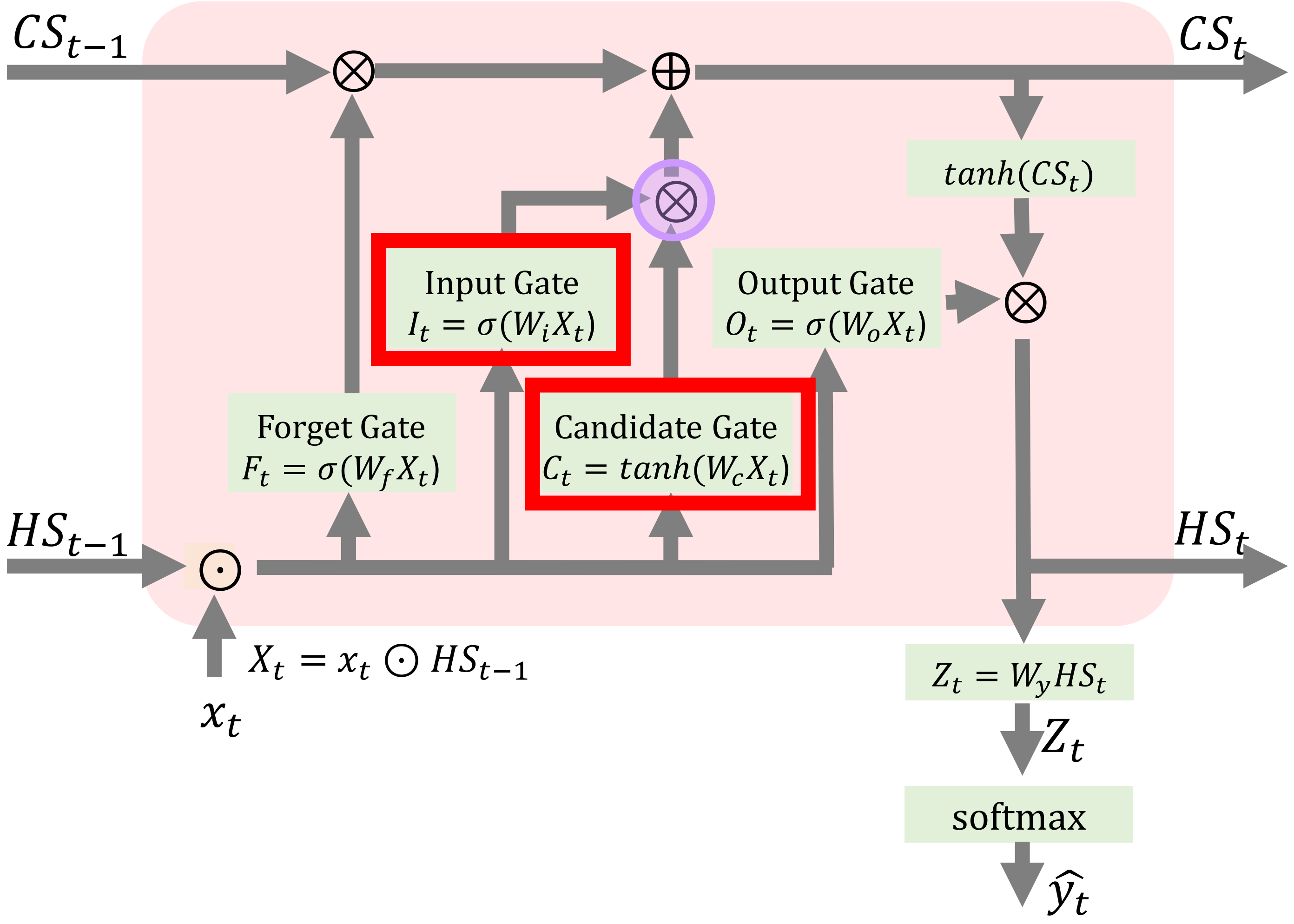
Output Gate에서 나온 값과 element-wise곱을 하여,



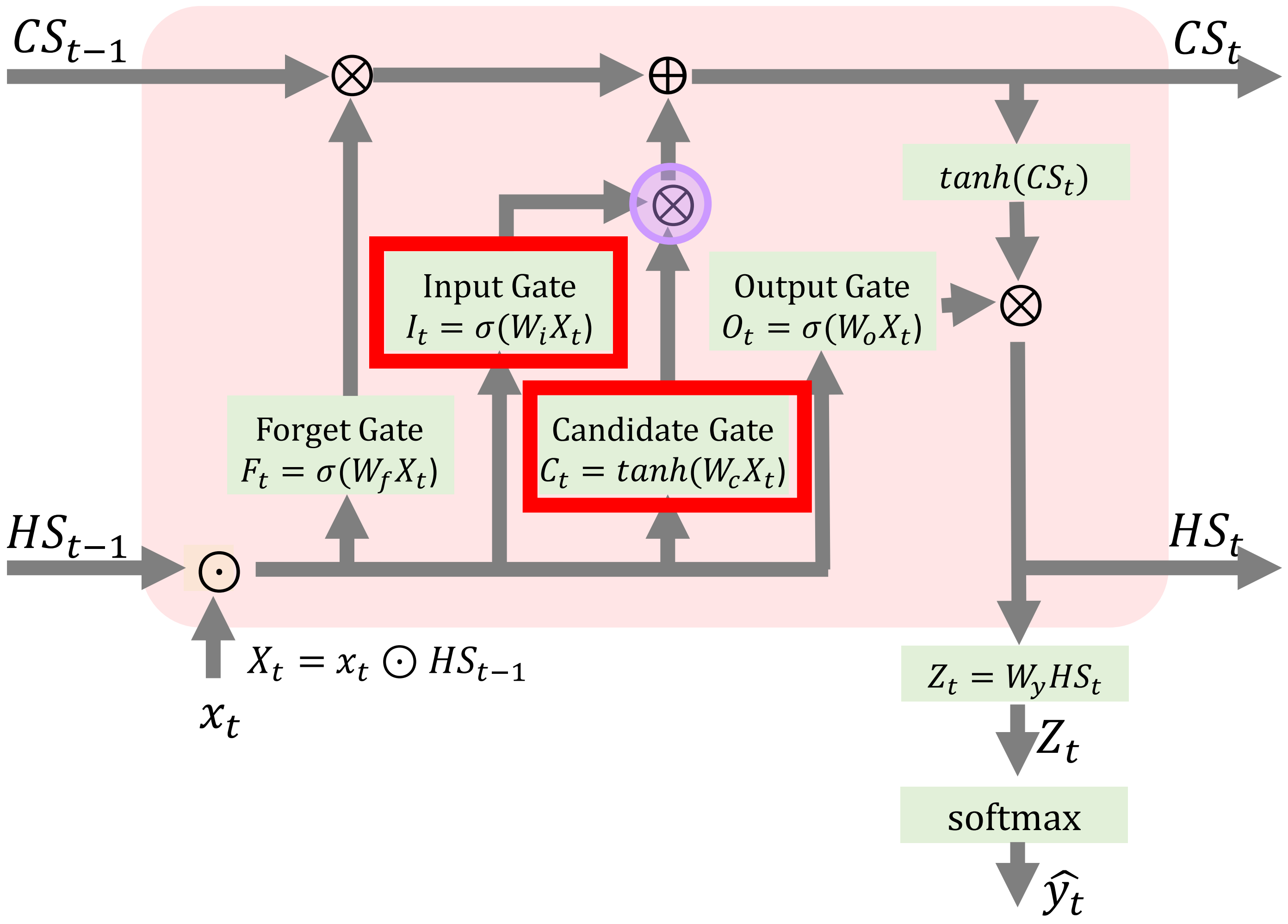
새로운 히든상태 HS_t 를 만어 냅니다



마치 Input Gate와 Candidate Gate의 콜라보가

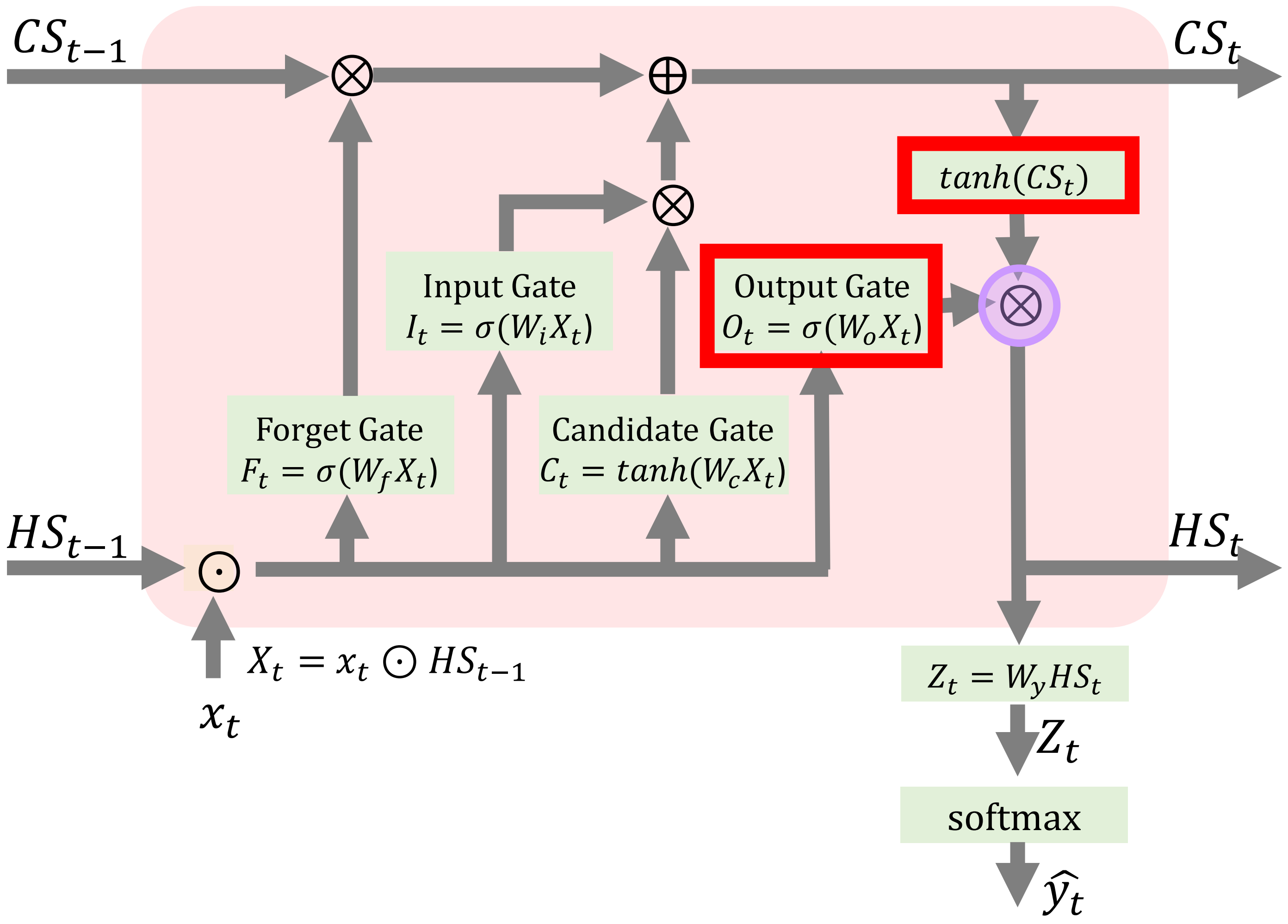


현재 입력 (short-term)중 기억할 부분을 남기는 것 처럼

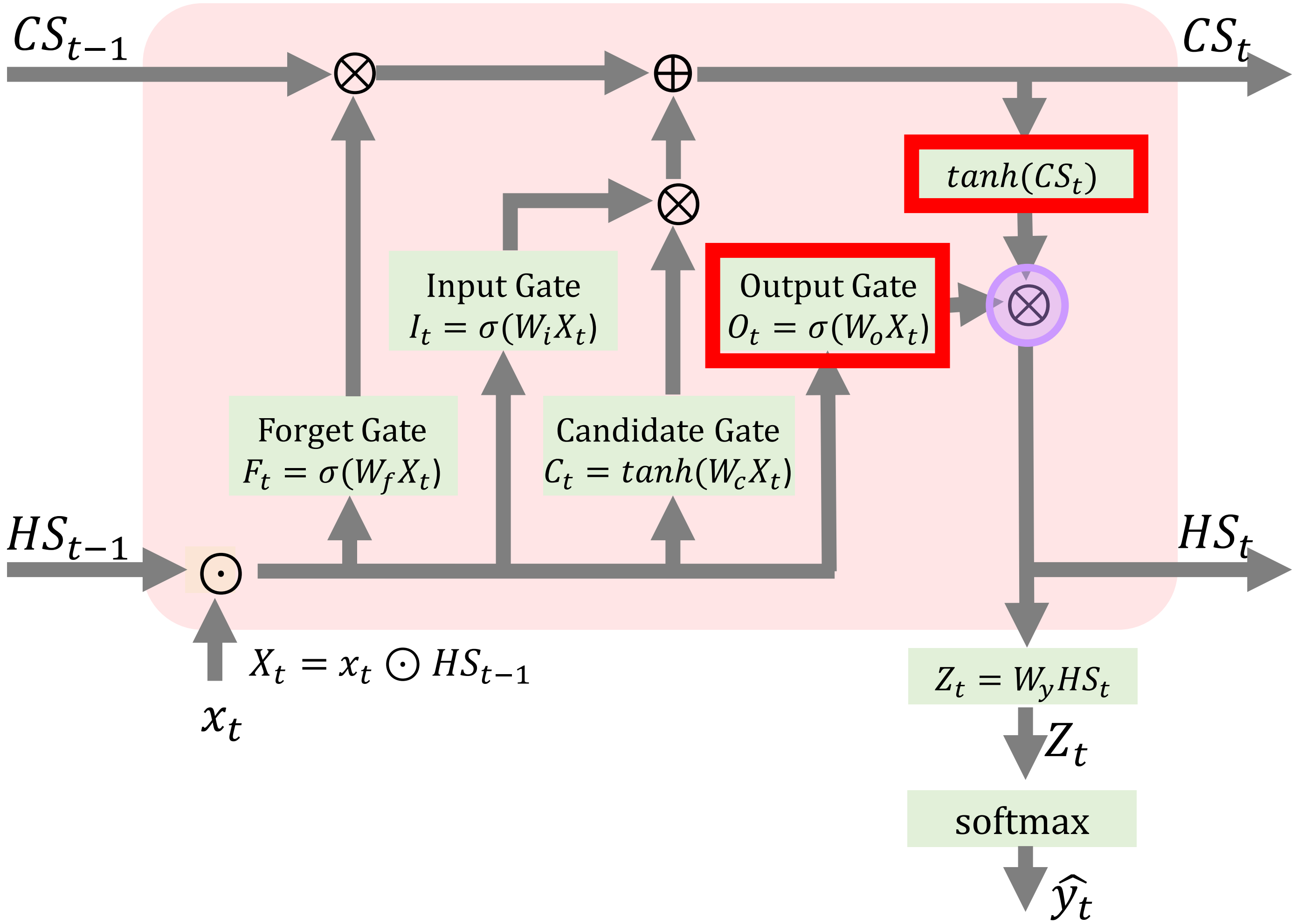


신박AI

Output Gate와 $tanh(CS_t)$ 의 콜라보는

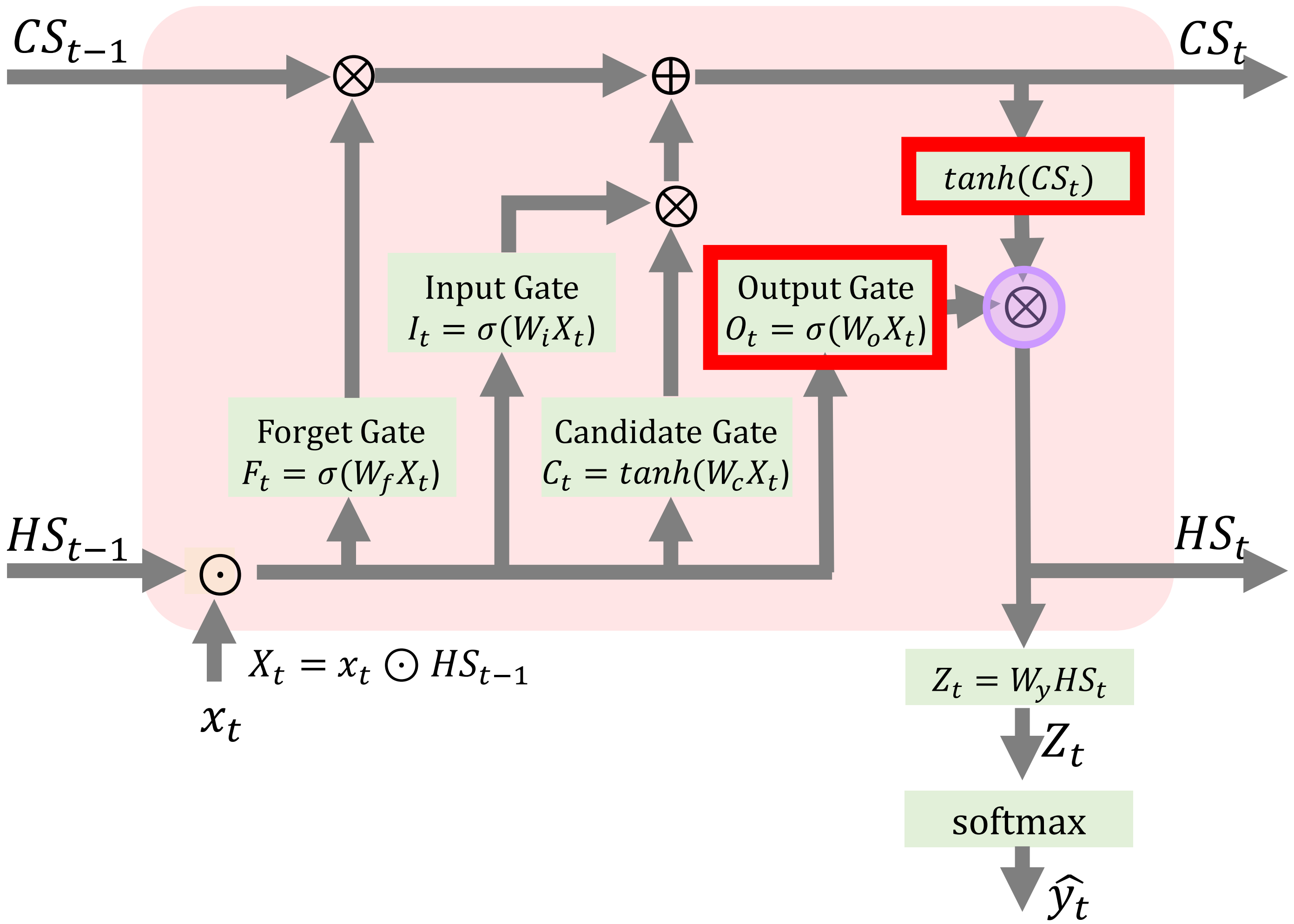


업데이트된 셀 상태(CS_t)에서 현재 입력값 (X_t)의 특성을 더 반영하는



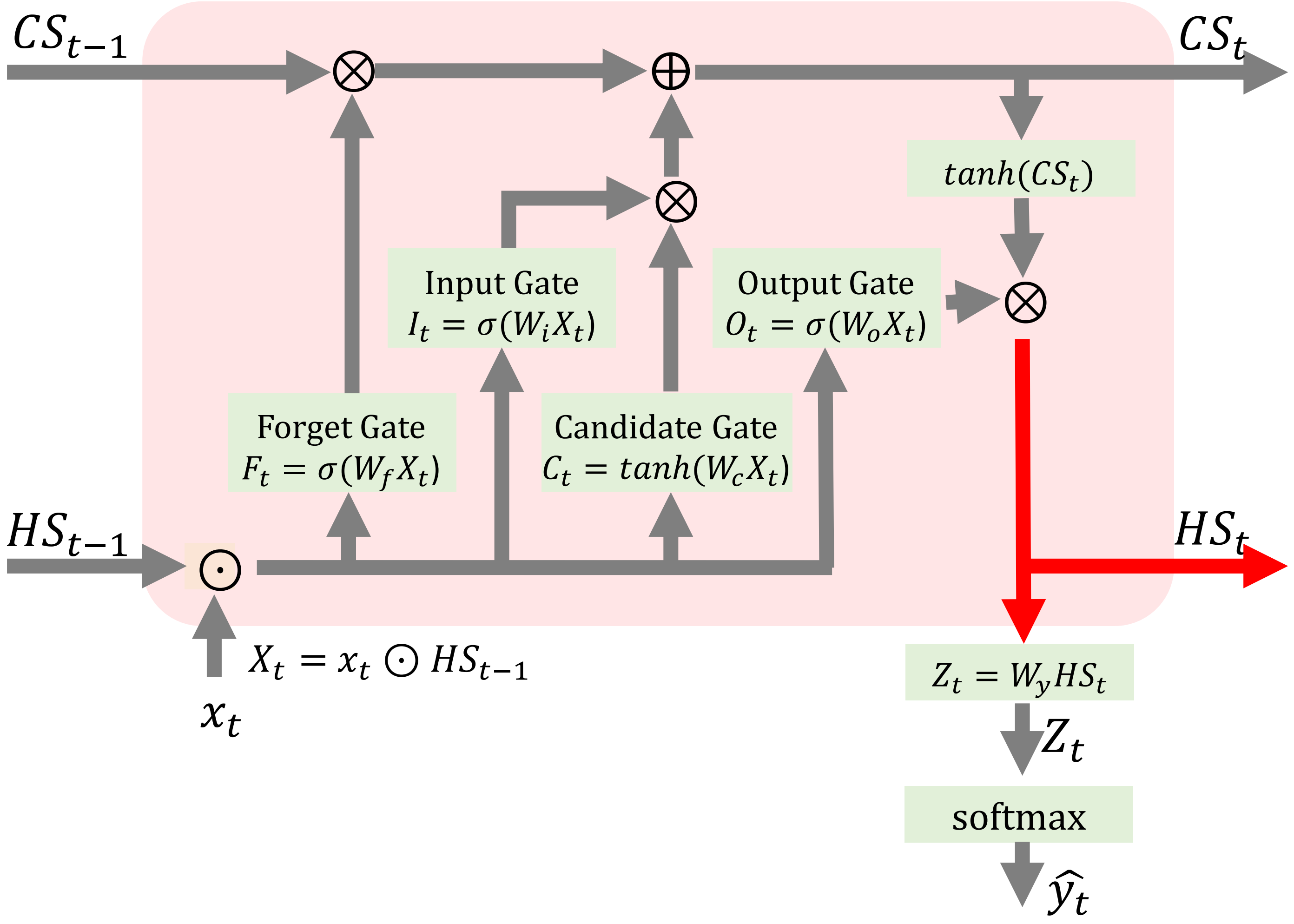
신박AI

새로운 히든상태 HS_t 를 만들어내는 것으로 보시면 됩니다

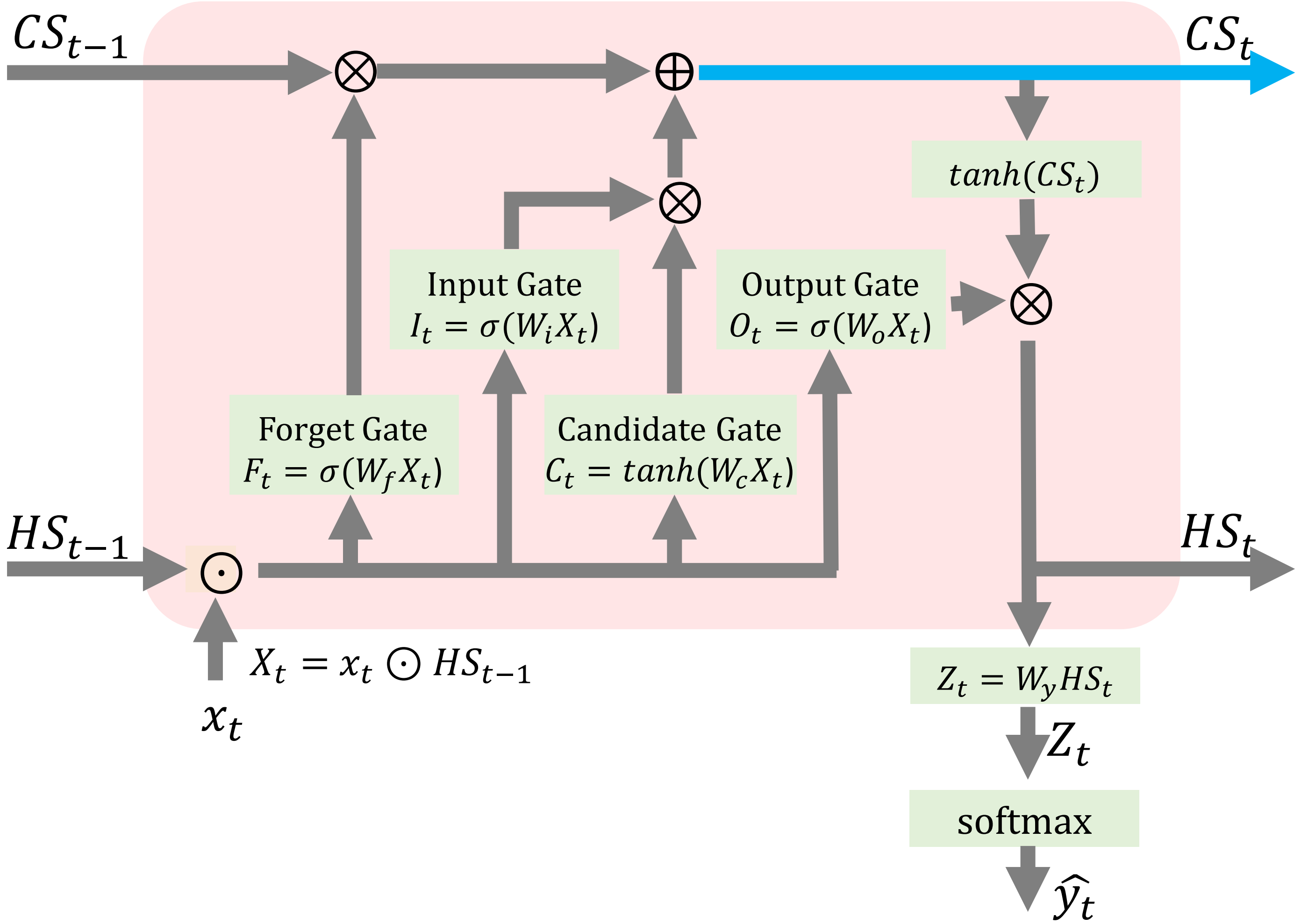


신박AI

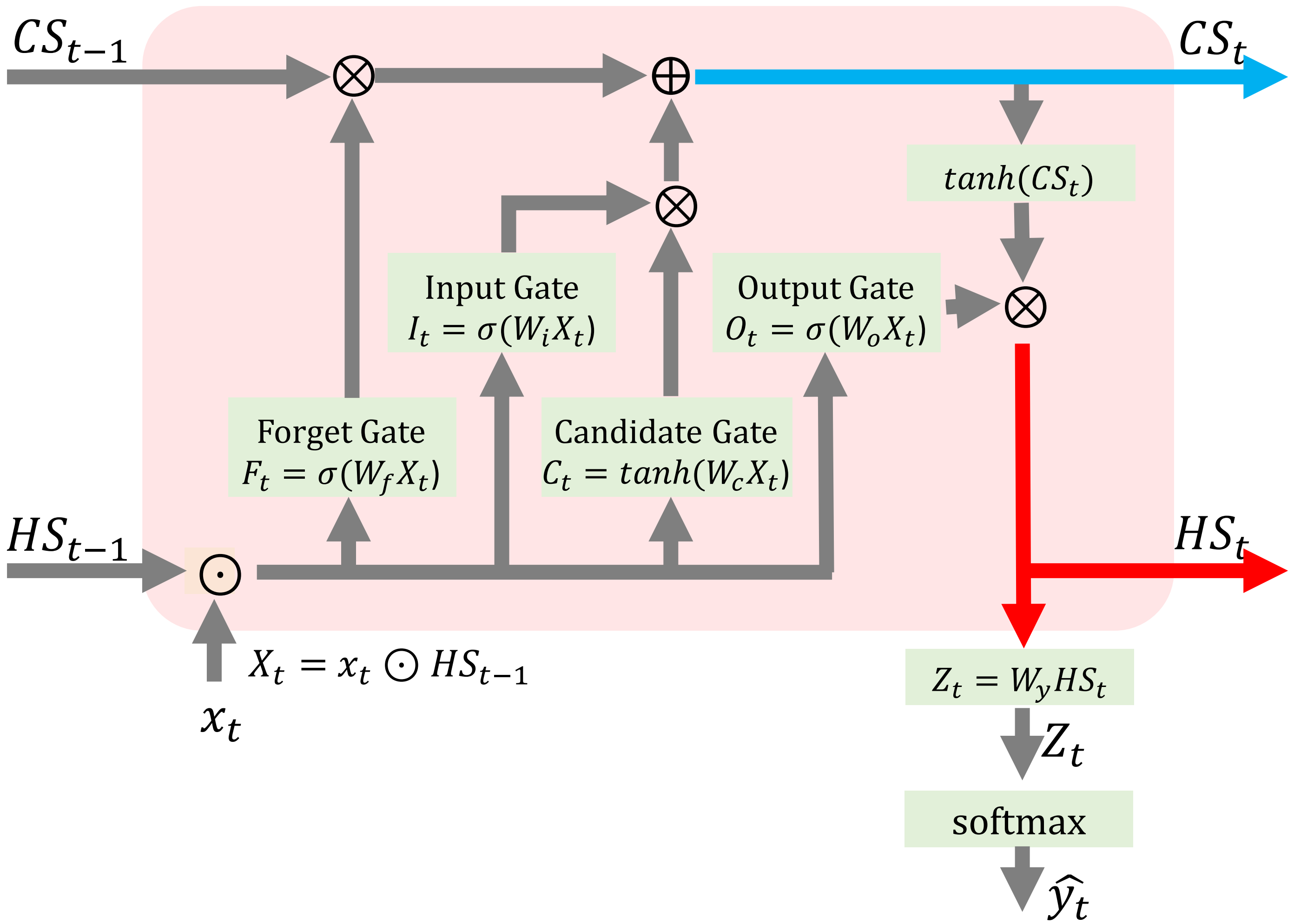
그러면 이 히든상태 HS_t 는 CS_t 에 비해서 좀 더 short-term 특성을 보이게 될 것입니다



그래서 CS_t 가 long-term 정보를 더 많이 담는다면,

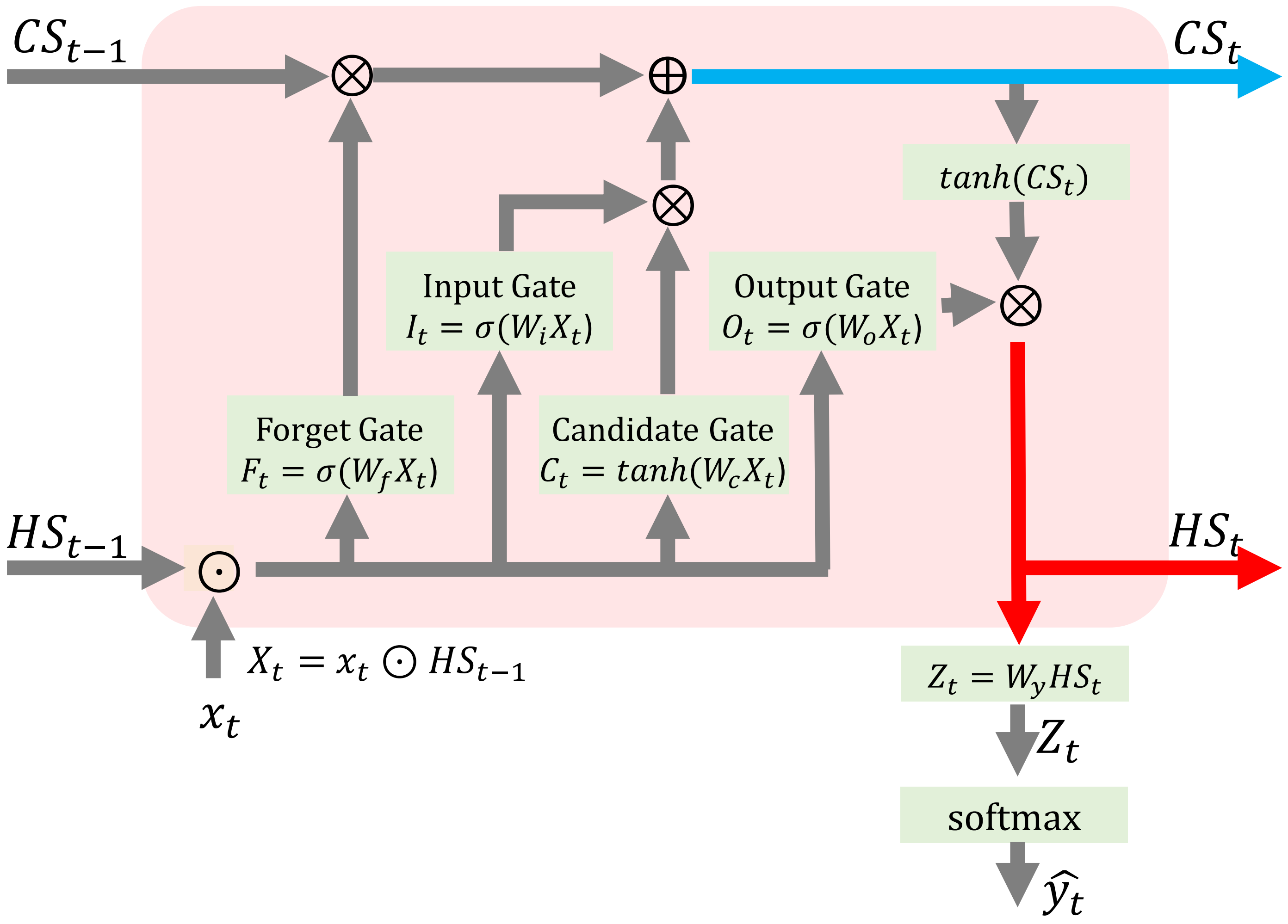


HS_t 는 같은 입력으로 short-term에 좀 더 가까운 정보를 담게 되므로

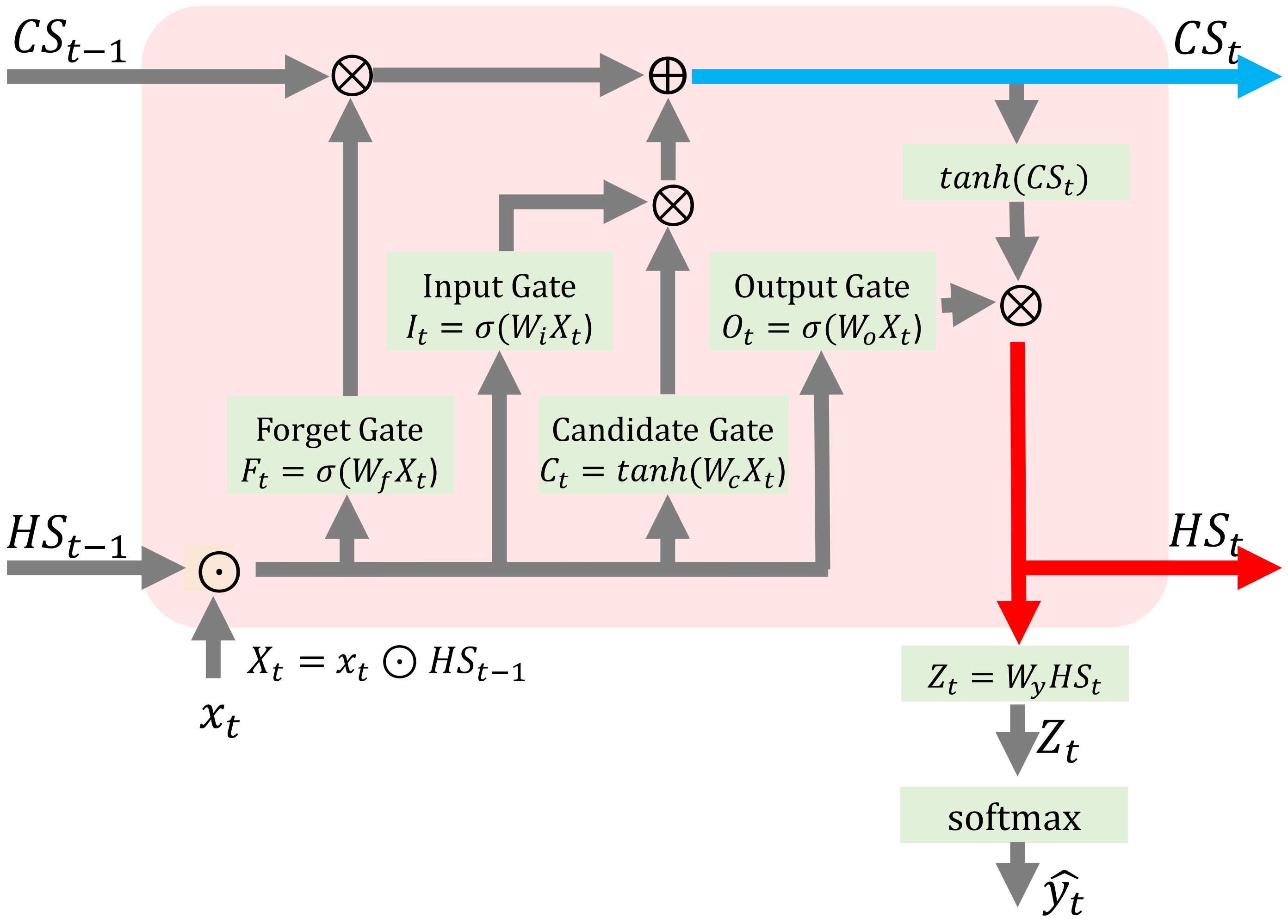


신박AI

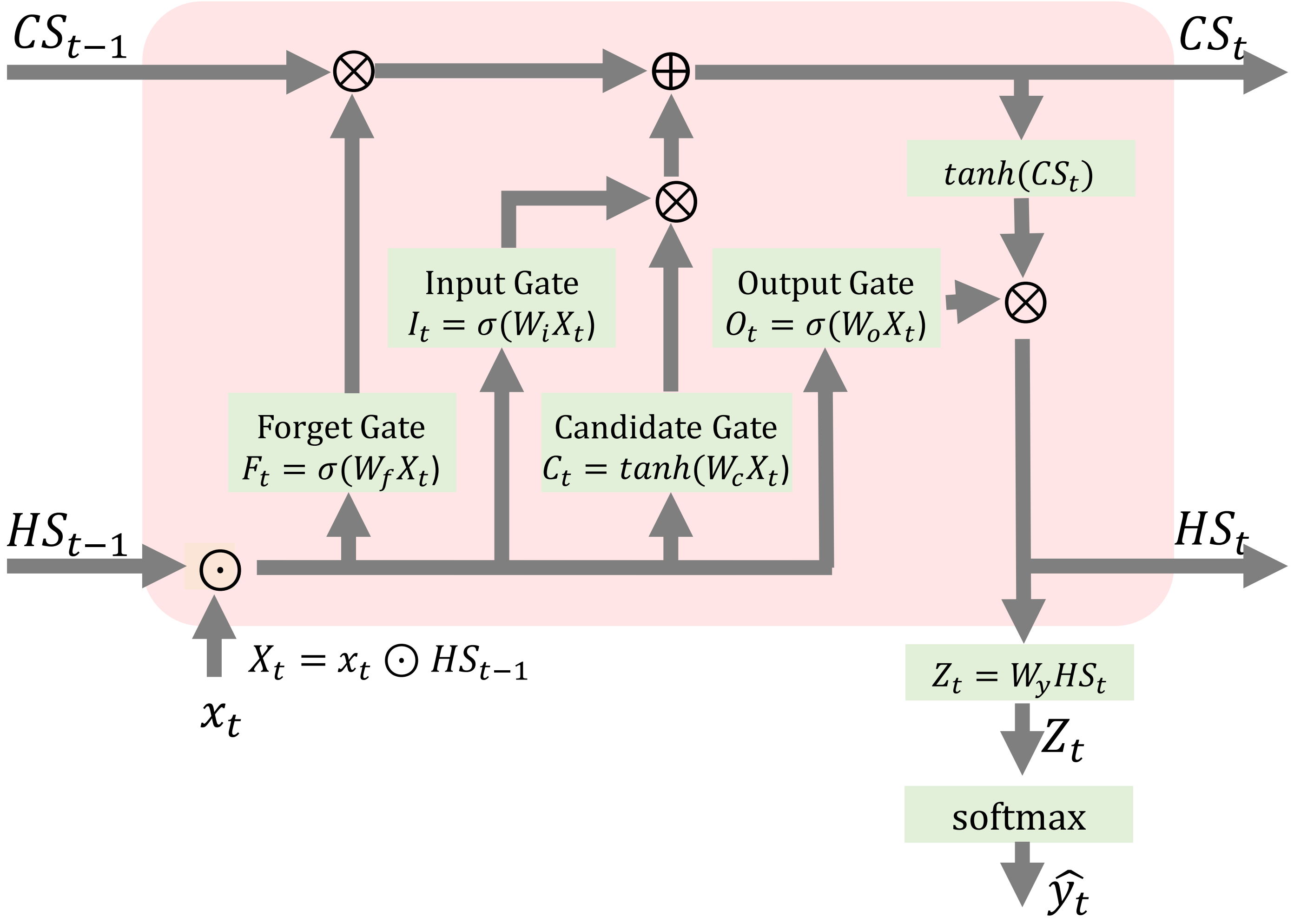
LSTM은 이 두개의 정보의 흐름을 이용하여 RNN보다 더 효율적으로



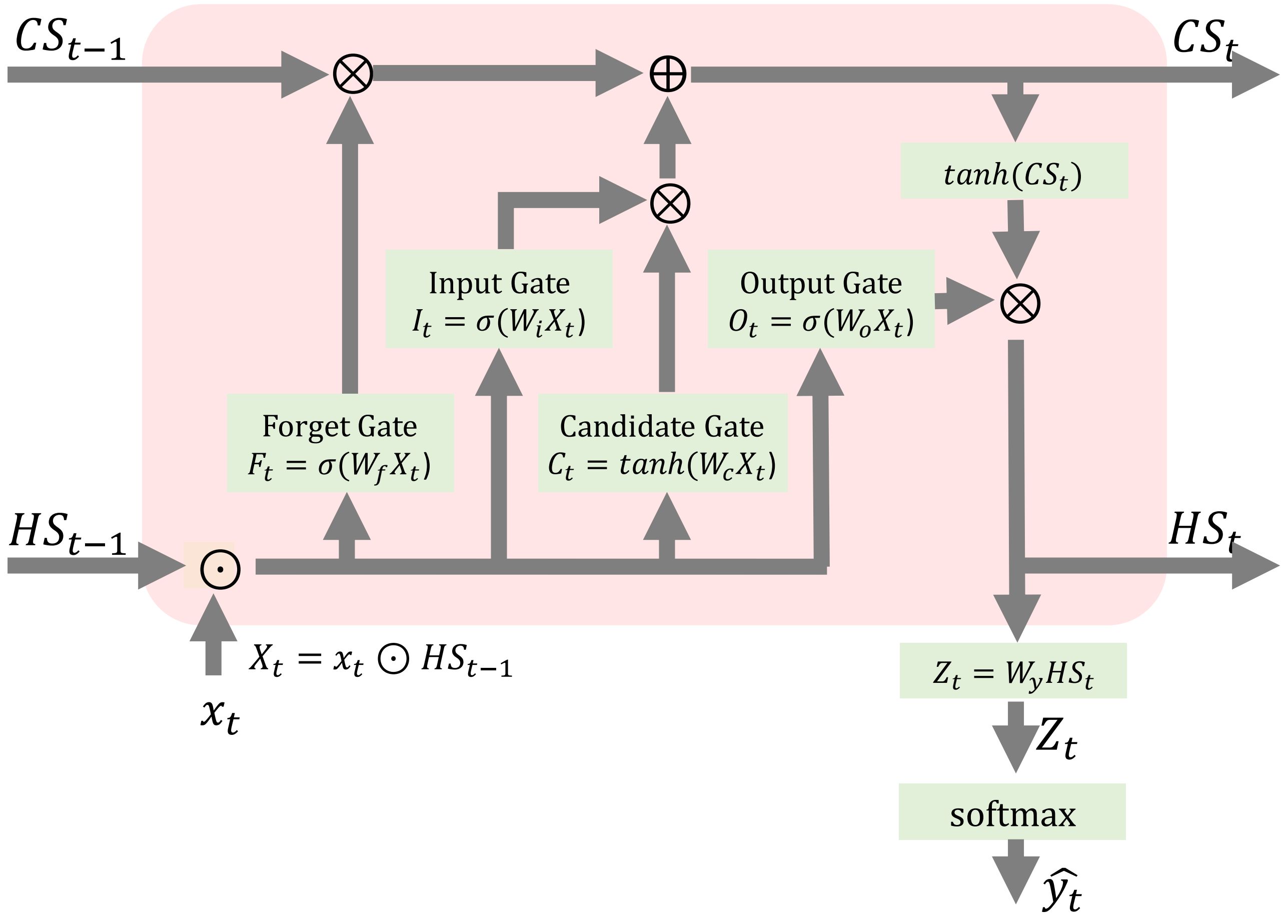
장기의존성(long-term dependency)문제를 다룰 수가 있는 것입니다



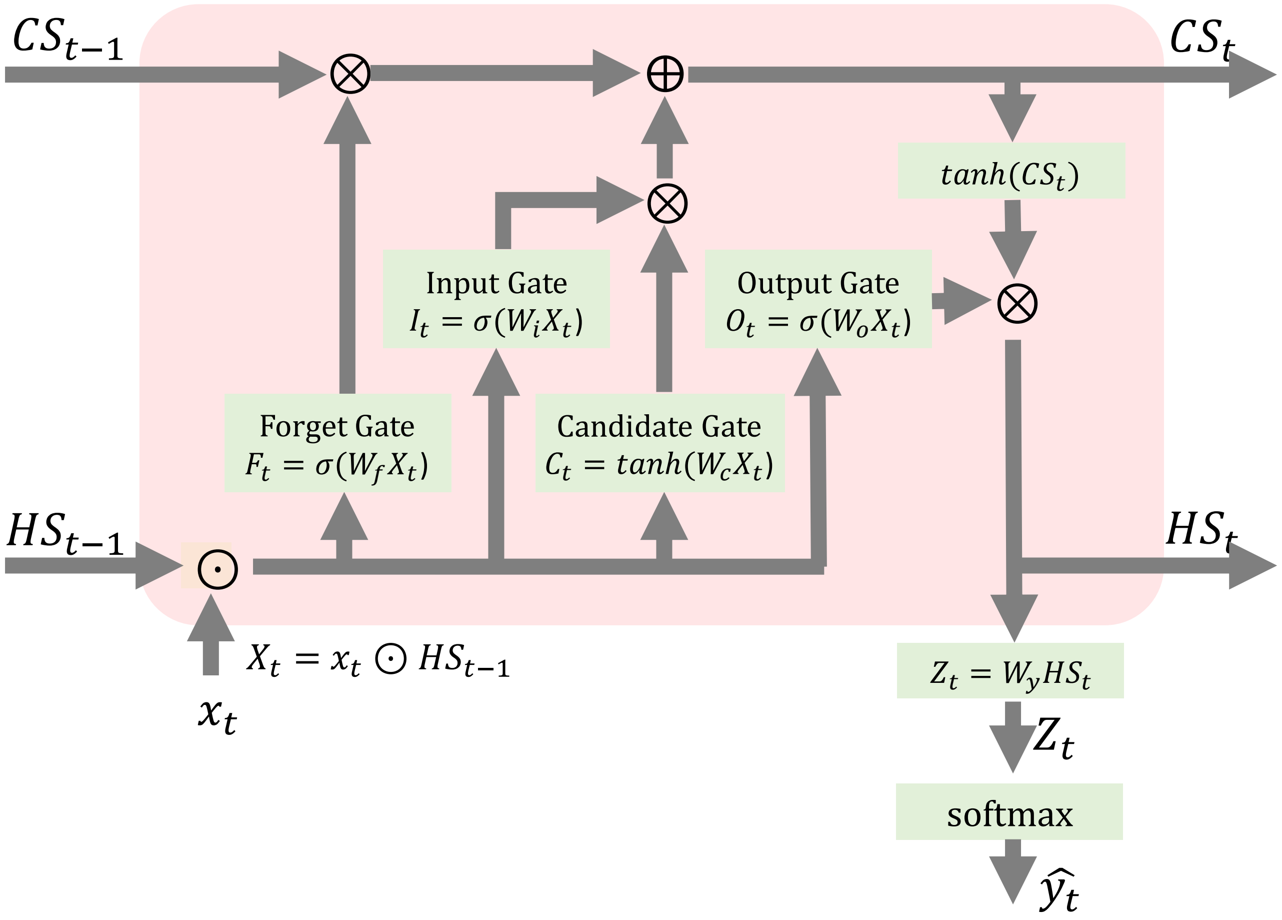
이제 그러면 숫자를 넣어서 순전파 feedforward 계산을 해보도록 하겠습니다



공간확보를 위해서 LSTM을 조금 옮겨보겠습니다

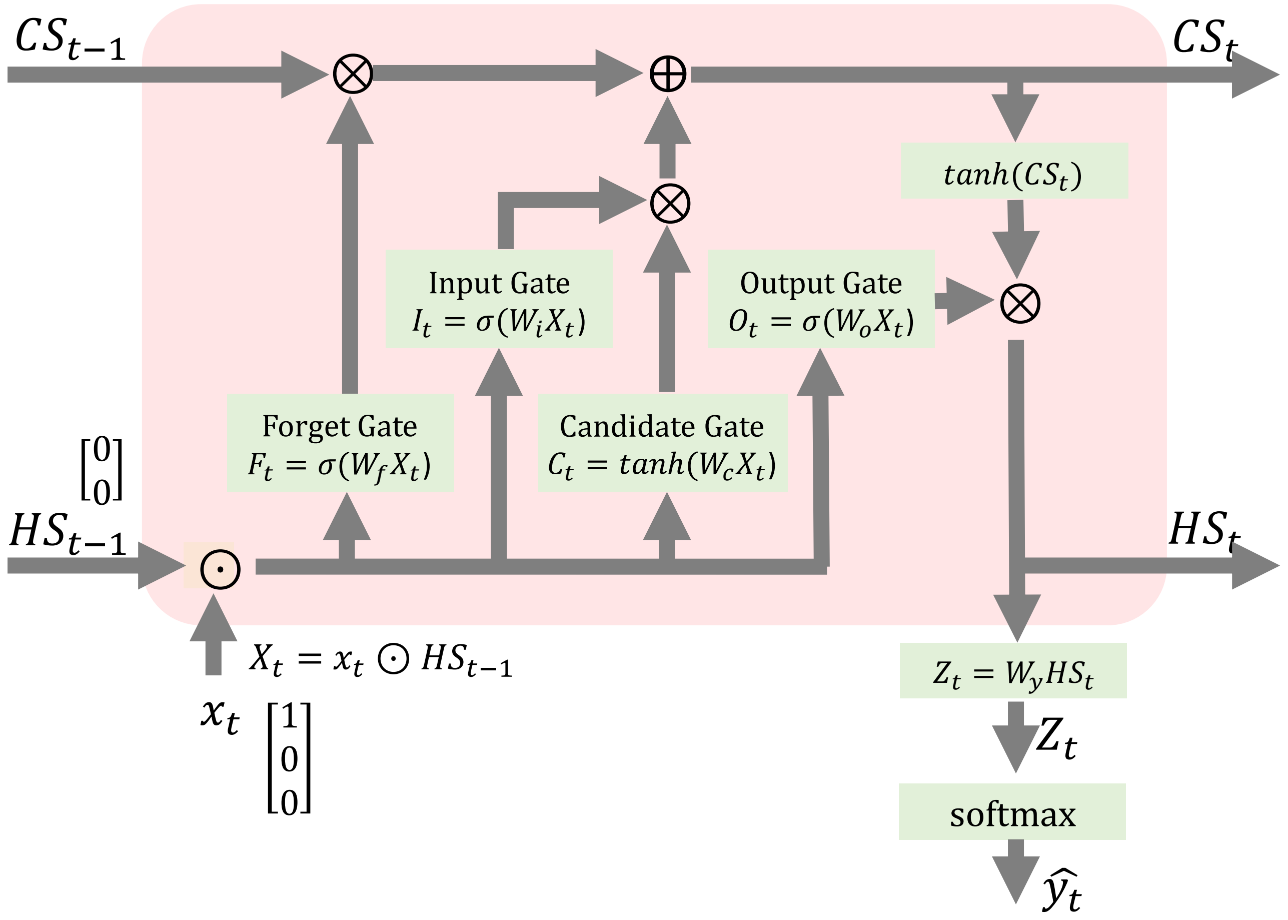


계산을 간단하게 하기 위해서 히든상태의 크기는 2, 입력x_t는 3으로 하도록 하겠습니다

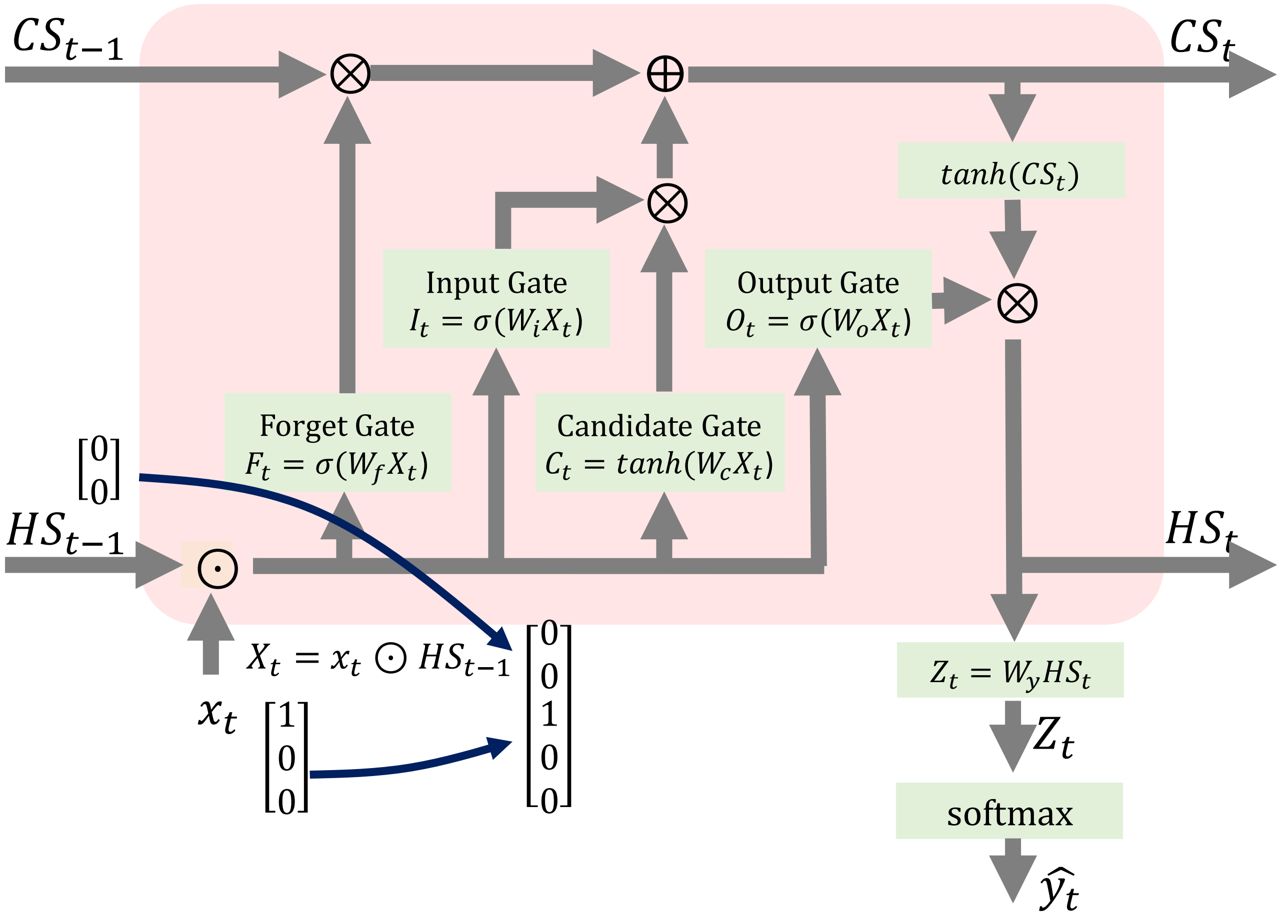


신박AI

그러면 다음과 같은 입력을 가정해 볼 수 있습니다



그러면 X_t는 두 행렬을 단순히 잇는 것이기 때문에 다음과 같습니다



그리고 내부 가중치들은 다음과 같이 초기화 하도록 하겠습니다

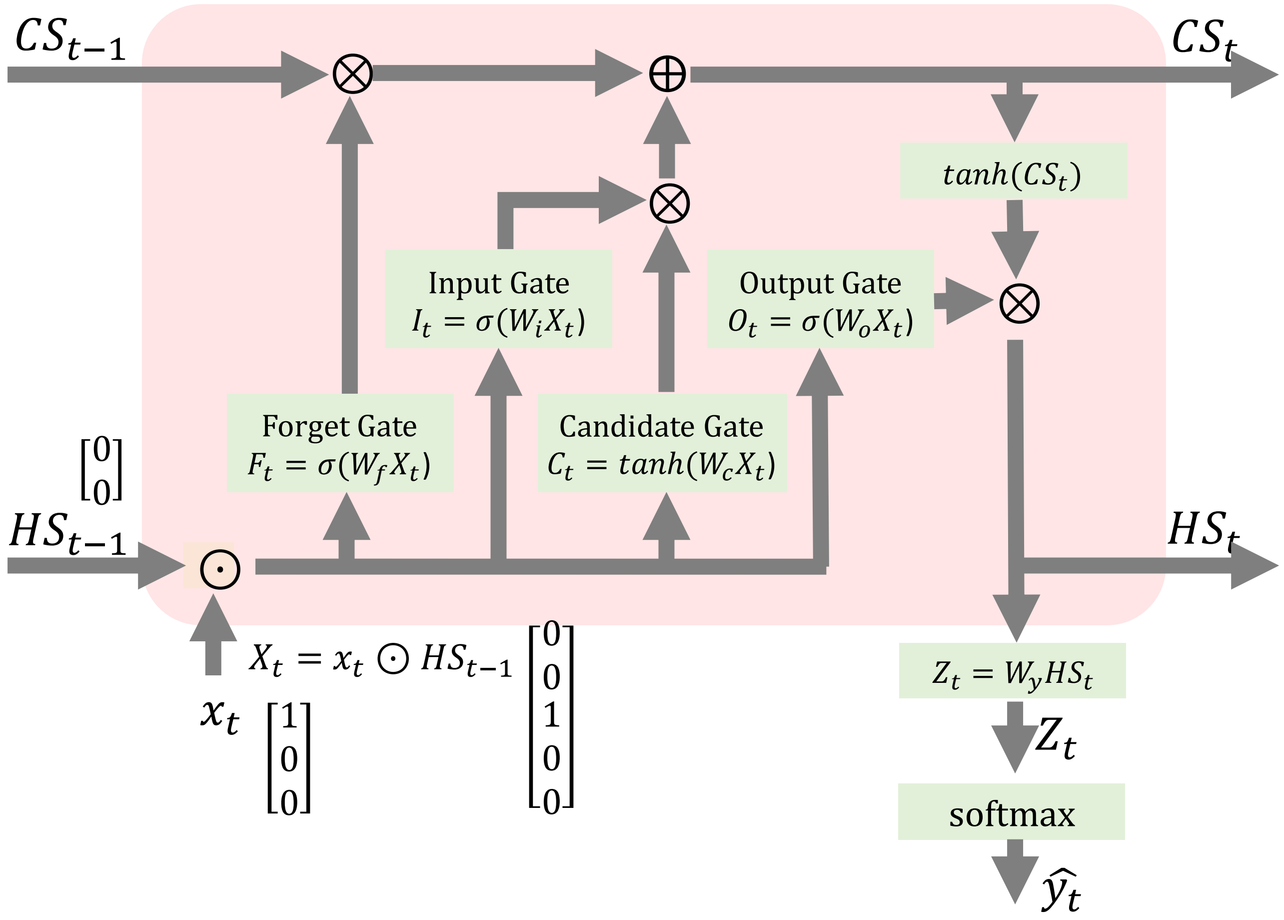
$$W_f = \begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix}$$

$$W_i = \begin{bmatrix} -0.209 & -0.14 & 0.031 & 0.226 & 0.696 \\ 0.101 & -0.435 & -0.406 & -0.796 & 0.324 \end{bmatrix}$$

$$W_c = \begin{bmatrix} -0.901 & -0.877 & -0.413 & 0.16 & -0.775 \\ -0.196 & 0.077 & 0.769 & -0.567 & -0.905 \end{bmatrix}$$

$$W_o = \begin{bmatrix} 0.668 & -0.605 & -0.402 & -0.691 & -0.486 \\ 0.613 & 0.875 & 0.549 & -0.623 & 0.262 \end{bmatrix}$$

$$W_y = \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix}$$



신박AI

그러면 Forget Gate의 출력값은 다음과 같습니다

$$W_f = \begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix}$$

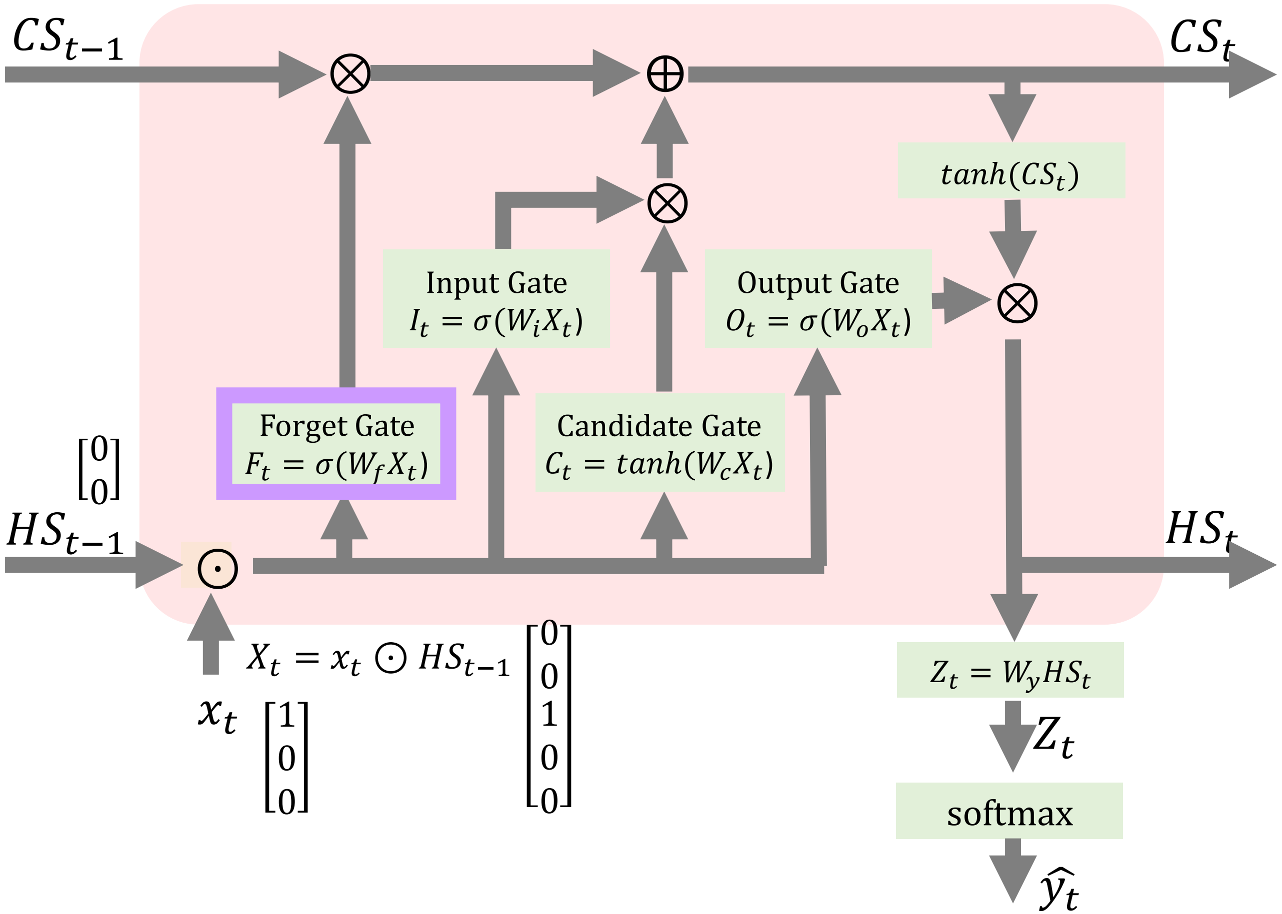
$$W_i = \begin{bmatrix} -0.209 & -0.14 & 0.031 & 0.226 & 0.696 \\ 0.101 & -0.435 & -0.406 & -0.796 & 0.324 \end{bmatrix}$$

$$W_c = \begin{bmatrix} -0.901 & -0.877 & -0.413 & 0.16 & -0.775 \\ -0.196 & 0.077 & 0.769 & -0.567 & -0.905 \end{bmatrix}$$

$$W_o = \begin{bmatrix} 0.668 & -0.605 & -0.402 & -0.691 & -0.486 \\ 0.613 & 0.875 & 0.549 & -0.623 & 0.262 \end{bmatrix}$$

$$W_y = \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix}$$

$$F_t = \sigma(W_f X_t)$$



그러면 Forget Gate의 출력값은 다음과 같습니다

$$W_f = \begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix}$$

$$W_i = \begin{bmatrix} -0.209 & -0.14 & 0.031 & 0.226 & 0.696 \\ 0.101 & -0.435 & -0.406 & -0.796 & 0.324 \end{bmatrix}$$

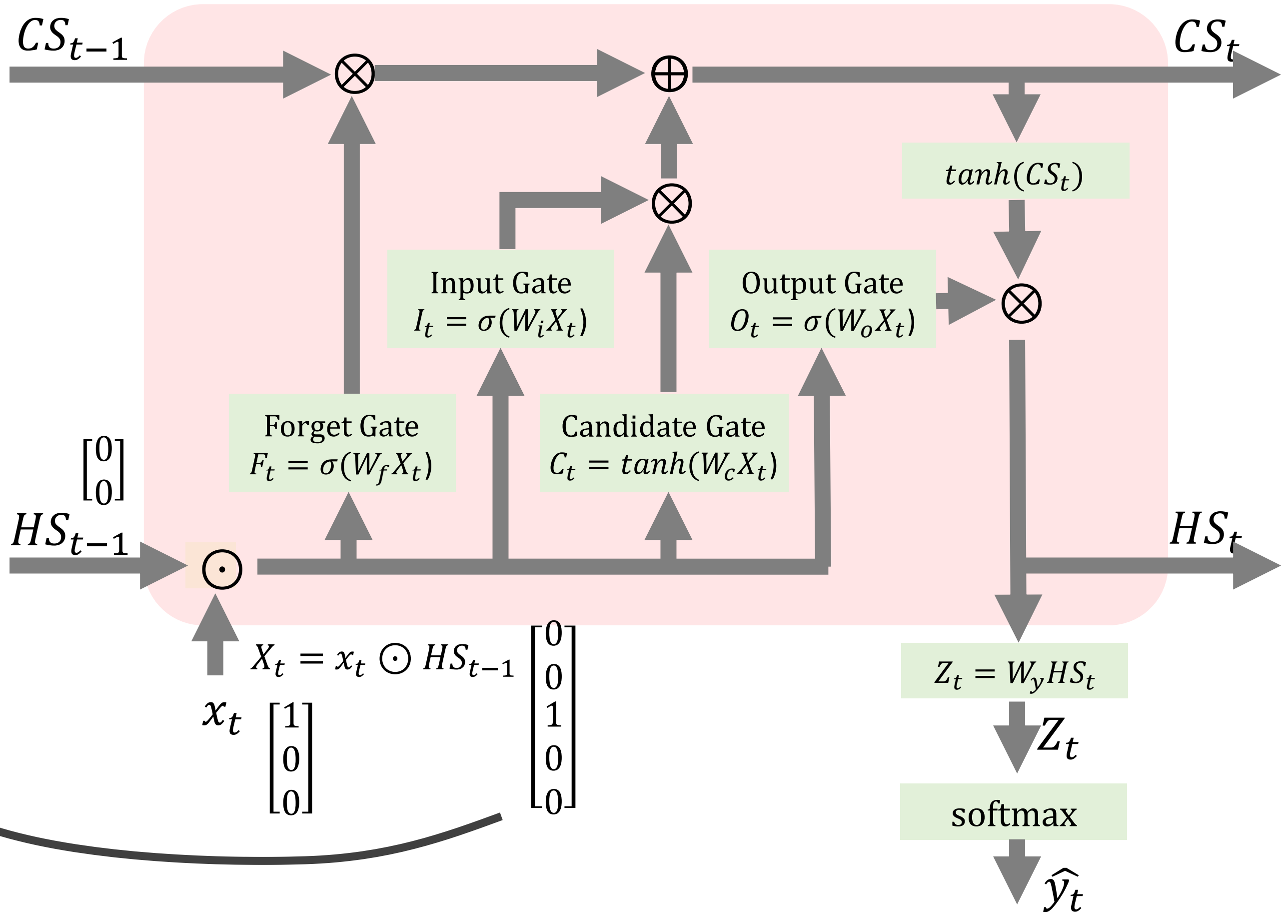
$$W_c = \begin{bmatrix} -0.901 & -0.877 & -0.413 & 0.16 & -0.775 \\ -0.196 & 0.077 & 0.769 & -0.567 & -0.905 \end{bmatrix}$$

$$W_o = \begin{bmatrix} 0.668 & -0.605 & -0.402 & -0.691 & -0.486 \\ 0.613 & 0.875 & 0.549 & -0.623 & 0.262 \end{bmatrix}$$

$$W_y = \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix}$$

$$F_t = \sigma(W_f X_t)$$

$$= \sigma\left(\begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}\right)$$



그러면 Forget Gate의 출력값은 다음과 같습니다

$$W_f = \begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix}$$

$$W_i = \begin{bmatrix} -0.209 & -0.14 & 0.031 & 0.226 & 0.696 \\ 0.101 & -0.435 & -0.406 & -0.796 & 0.324 \end{bmatrix}$$

$$W_c = \begin{bmatrix} -0.901 & -0.877 & -0.413 & 0.16 & -0.775 \\ -0.196 & 0.077 & 0.769 & -0.567 & -0.905 \end{bmatrix}$$

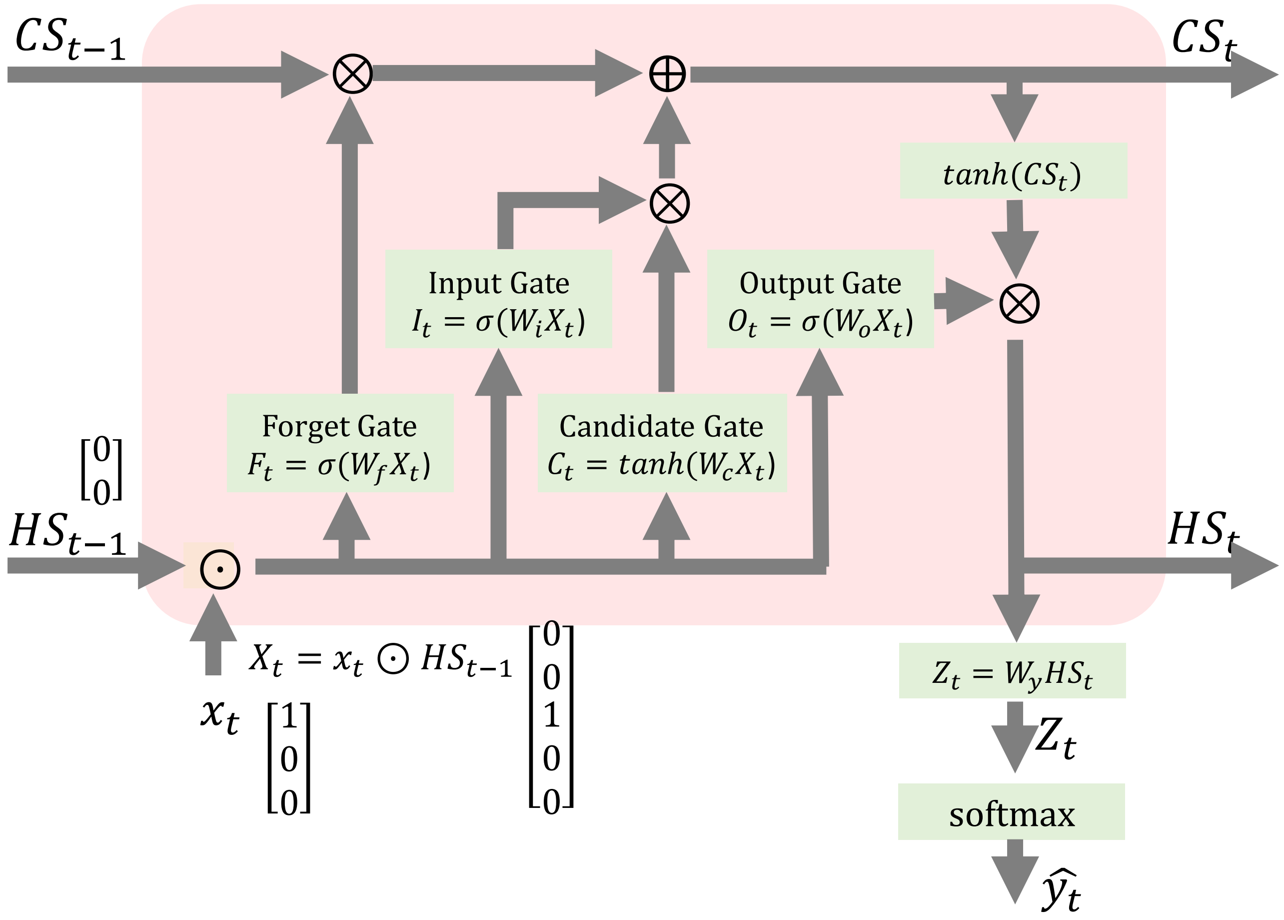
$$W_o = \begin{bmatrix} 0.668 & -0.605 & -0.402 & -0.691 & -0.486 \\ 0.613 & 0.875 & 0.549 & -0.623 & 0.262 \end{bmatrix}$$

$$W_y = \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix}$$

$$F_t = \sigma(W_f X_t)$$

$$= \sigma\left(\begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}\right)$$

$$= \sigma\left(\begin{bmatrix} 0.02 \\ 0.856 \end{bmatrix}\right)$$



신박AI

그러면 Forget Gate의 출력값은 다음과 같습니다

$$W_f = \begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix}$$

$$W_i = \begin{bmatrix} -0.209 & -0.14 & 0.031 & 0.226 & 0.696 \\ 0.101 & -0.435 & -0.406 & -0.796 & 0.324 \end{bmatrix}$$

$$W_c = \begin{bmatrix} -0.901 & -0.877 & -0.413 & 0.16 & -0.775 \\ -0.196 & 0.077 & 0.769 & -0.567 & -0.905 \end{bmatrix}$$

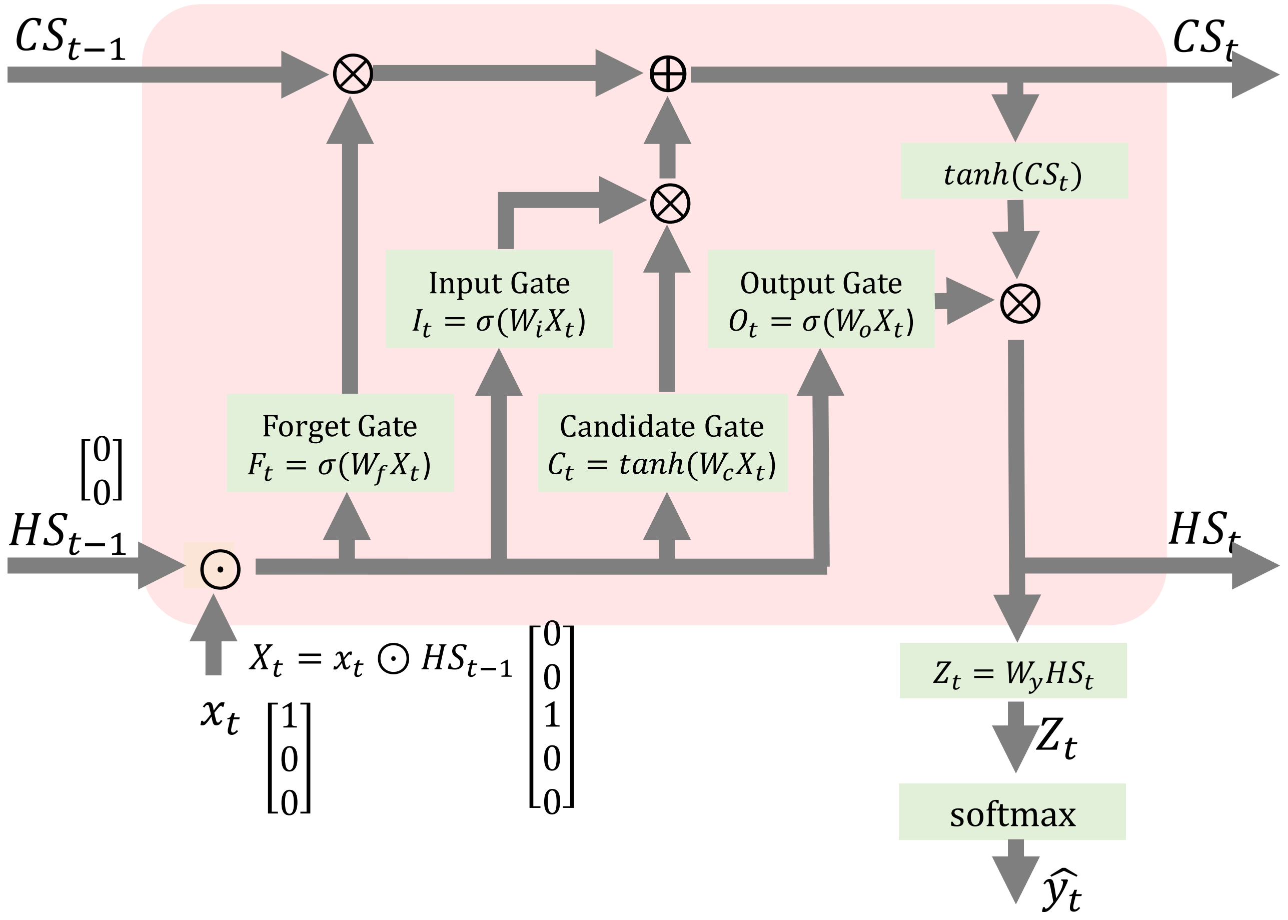
$$W_o = \begin{bmatrix} 0.668 & -0.605 & -0.402 & -0.691 & -0.486 \\ 0.613 & 0.875 & 0.549 & -0.623 & 0.262 \end{bmatrix}$$

$$W_y = \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix}$$

$$F_t = \sigma(W_f X_t)$$

$$= \sigma\left(\begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix}\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}\right)$$

$$= \sigma\left(\begin{bmatrix} 0.02 \\ 0.856 \end{bmatrix}\right) = \begin{bmatrix} 0.505 \\ 0.702 \end{bmatrix}$$



신박AI

그러면 Forget Gate의 출력값은 다음과 같습니다

$$W_f = \begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix}$$

$$W_i = \begin{bmatrix} -0.209 & -0.14 & 0.031 & 0.226 & 0.696 \\ 0.101 & -0.435 & -0.406 & -0.796 & 0.324 \end{bmatrix}$$

$$W_c = \begin{bmatrix} -0.901 & -0.877 & -0.413 & 0.16 & -0.775 \\ -0.196 & 0.077 & 0.769 & -0.567 & -0.905 \end{bmatrix}$$

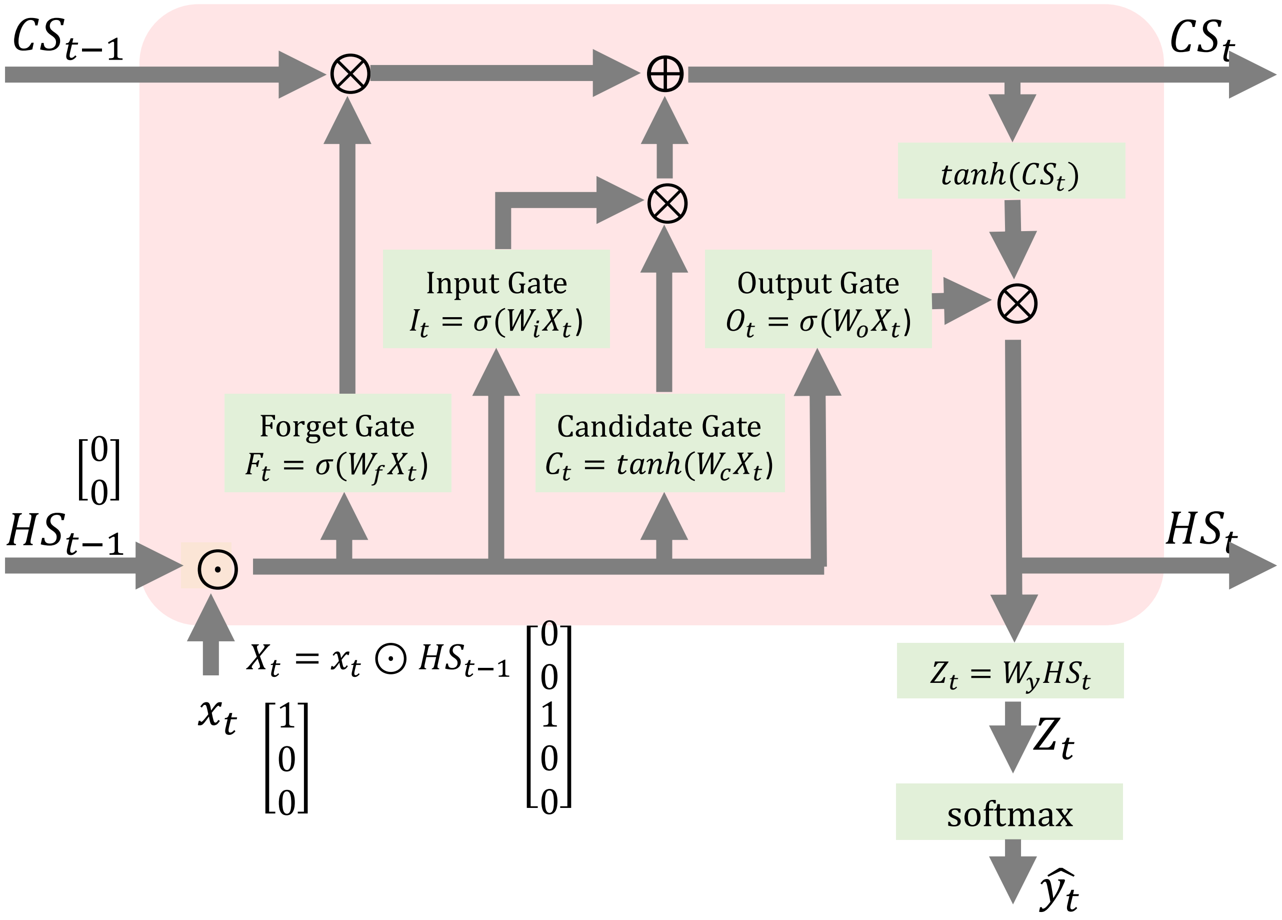
$$W_o = \begin{bmatrix} 0.668 & -0.605 & -0.402 & -0.691 & -0.486 \\ 0.613 & 0.875 & 0.549 & -0.623 & 0.262 \end{bmatrix}$$

$$W_y = \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix}$$

$$F_t = \sigma(W_f X_t)$$

$$= \sigma(\begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix})$$

$$= \sigma(\begin{bmatrix} 0.02 \\ 0.856 \end{bmatrix}) = \begin{bmatrix} 0.505 \\ 0.702 \end{bmatrix}$$



신박AI

그러면 Forget Gate의 출력값은 다음과 같습니다

$$W_f = \begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix}$$

$$W_i = \begin{bmatrix} -0.209 & -0.14 & 0.031 & 0.226 & 0.696 \\ 0.101 & -0.435 & -0.406 & -0.796 & 0.324 \end{bmatrix}$$

$$W_c = \begin{bmatrix} -0.901 & -0.877 & -0.413 & 0.16 & -0.775 \\ -0.196 & 0.077 & 0.769 & -0.567 & -0.905 \end{bmatrix}$$

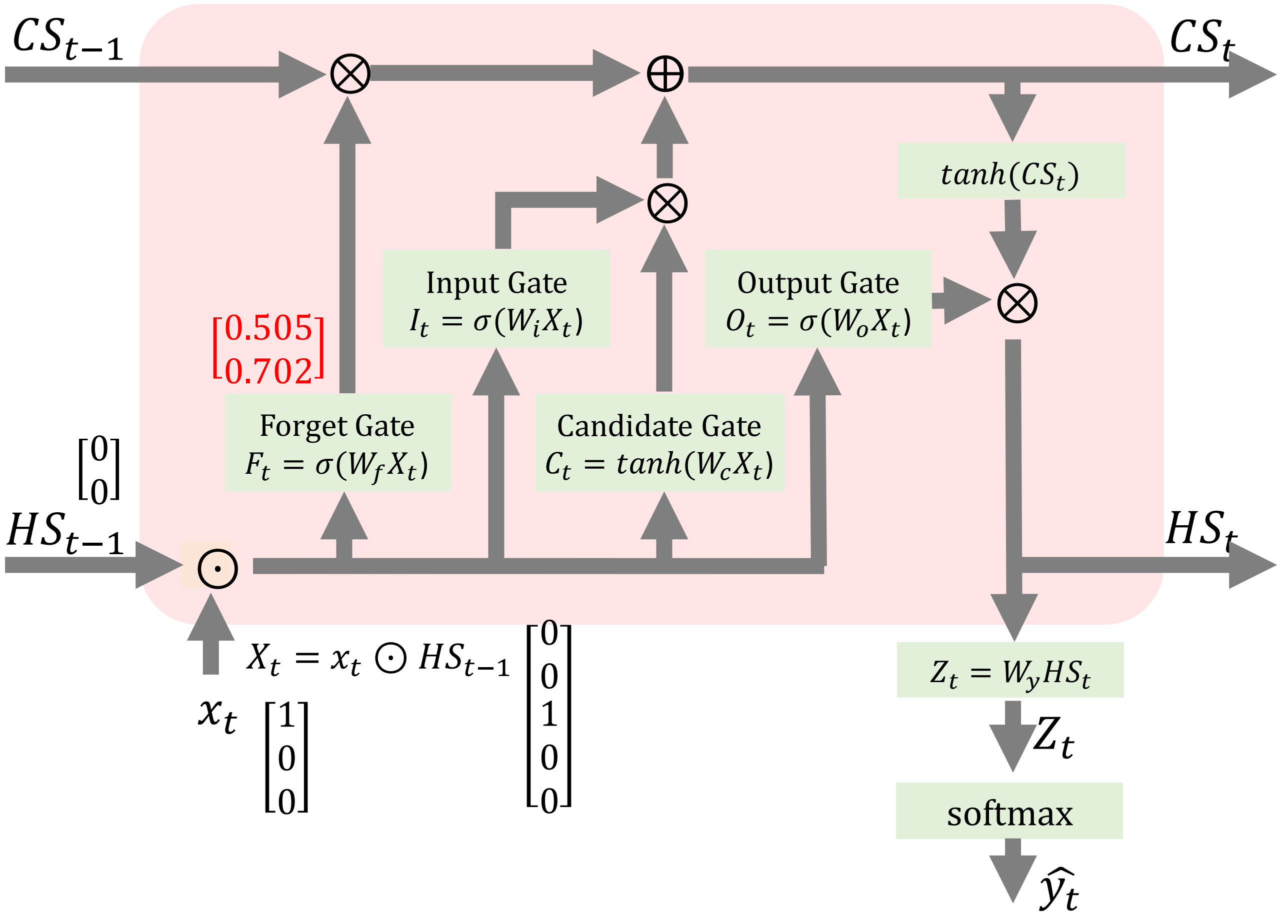
$$W_o = \begin{bmatrix} 0.668 & -0.605 & -0.402 & -0.691 & -0.486 \\ 0.613 & 0.875 & 0.549 & -0.623 & 0.262 \end{bmatrix}$$

$$W_y = \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix}$$

$$F_t = \sigma(W_f X_t)$$

$$= \sigma\left(\begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix}\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}\right)$$

$$= \sigma\left(\begin{bmatrix} 0.02 \\ 0.856 \end{bmatrix}\right) =$$



신박AI

똑같은 방식으로 Input Gate도 구할 수 있습니다

$$W_f = \begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix}$$

$$W_i = \begin{bmatrix} -0.209 & -0.14 & 0.031 & 0.226 & 0.696 \\ 0.101 & -0.435 & -0.406 & -0.796 & 0.324 \end{bmatrix}$$

$$W_c = \begin{bmatrix} -0.901 & -0.877 & -0.413 & 0.16 & -0.775 \\ -0.196 & 0.077 & 0.769 & -0.567 & -0.905 \end{bmatrix}$$

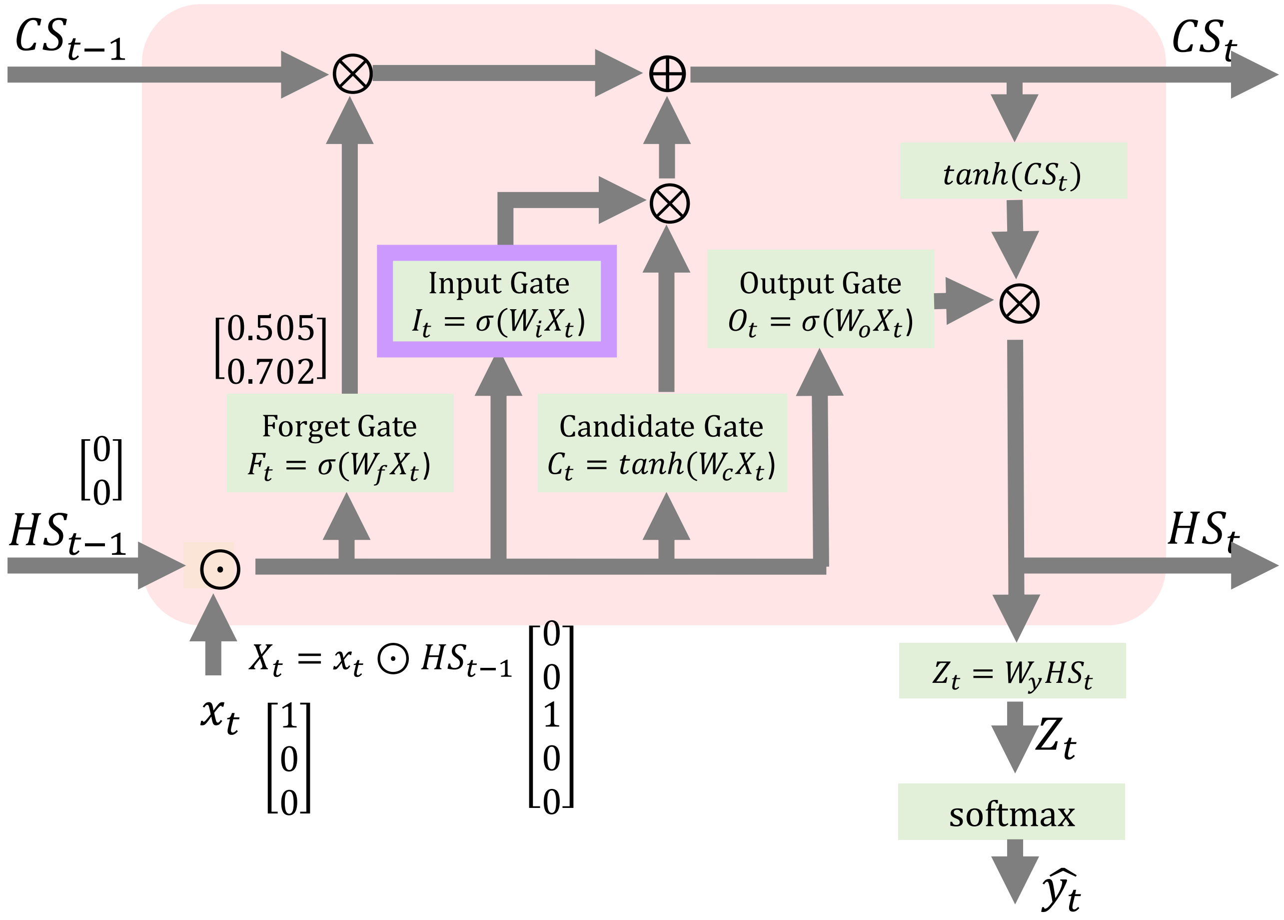
$$W_o = \begin{bmatrix} 0.668 & -0.605 & -0.402 & -0.691 & -0.486 \\ 0.613 & 0.875 & 0.549 & -0.623 & 0.262 \end{bmatrix}$$

$$W_y = \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix}$$

$$I_t = \sigma(W_i X_t)$$

$$= \sigma\left(\begin{bmatrix} -0.209 & -0.14 & 0.031 & 0.226 & 0.696 \\ 0.101 & -0.435 & -0.406 & -0.796 & 0.324 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}\right)$$

$$= \sigma\left(\begin{bmatrix} 0.31 \\ -0.406 \end{bmatrix}\right) = \begin{bmatrix} 0.508 \\ 0.4 \end{bmatrix}$$



신박AI

똑같은 방식으로 Candidate Gate도 구할 수 있습니다

$$W_f = \begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix}$$

$$W_i = \begin{bmatrix} -0.209 & -0.14 & 0.031 & 0.226 & 0.696 \\ 0.101 & -0.435 & -0.406 & -0.796 & 0.324 \end{bmatrix}$$

$$W_c = \begin{bmatrix} -0.901 & -0.877 & -0.413 & 0.16 & -0.775 \\ -0.196 & 0.077 & 0.769 & -0.567 & -0.905 \end{bmatrix}$$

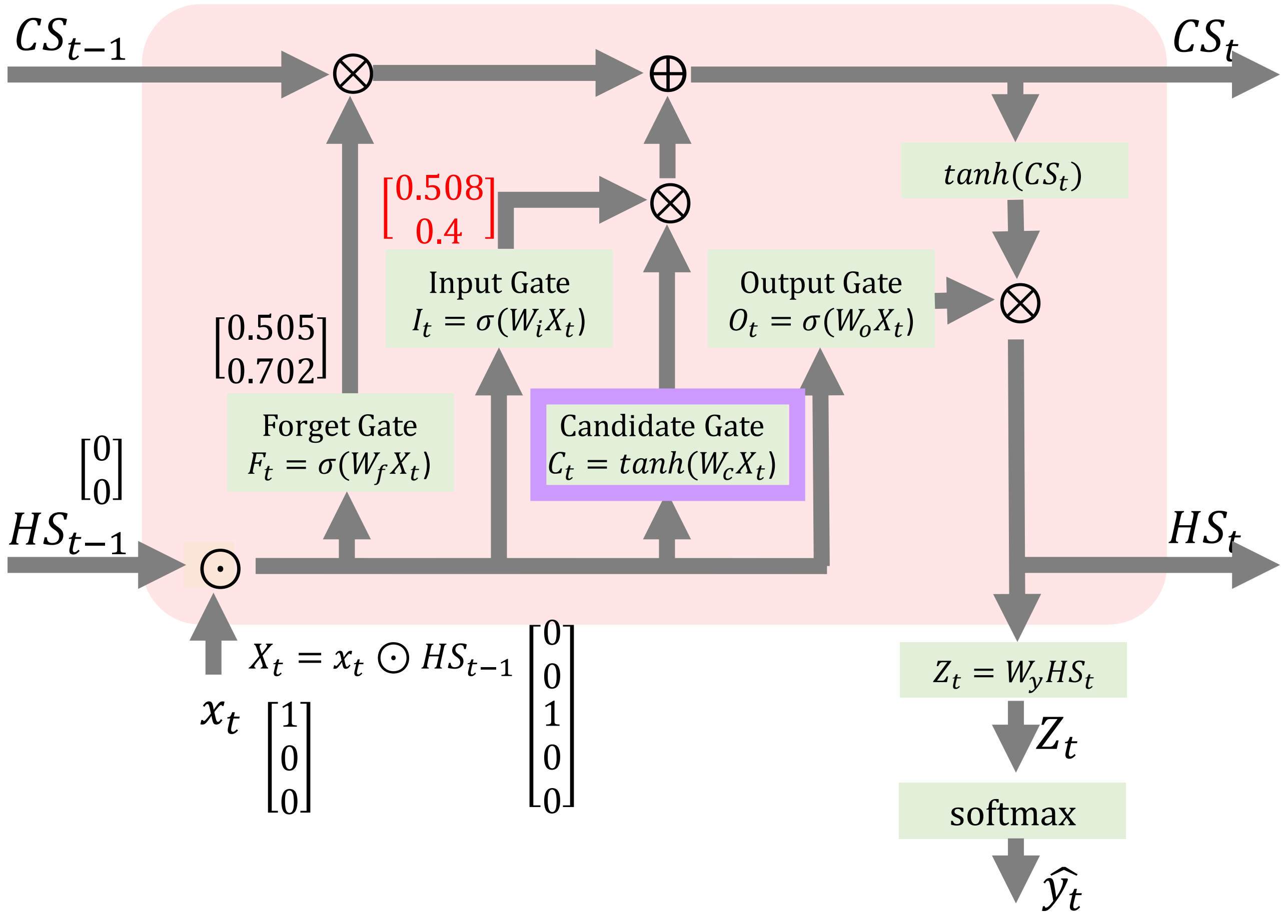
$$W_o = \begin{bmatrix} 0.668 & -0.605 & -0.402 & -0.691 & -0.486 \\ 0.613 & 0.875 & 0.549 & -0.623 & 0.262 \end{bmatrix}$$

$$W_y = \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix}$$

$$C_t = tanh(W_c X_t)$$

$$= tanh(\begin{bmatrix} -0.901 & -0.877 & -0.413 & 0.16 & -0.775 \\ -0.196 & 0.077 & 0.769 & -0.567 & -0.905 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix})$$

$$= tanh(\begin{bmatrix} -0.4121 \\ 0.769 \end{bmatrix}) = \begin{bmatrix} -0.391 \\ 0.646 \end{bmatrix}$$



그리고 Output Gate도 같은 방식으로 구해보았습니다

$$W_f = \begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix}$$

$$W_i = \begin{bmatrix} -0.209 & -0.14 & 0.031 & 0.226 & 0.696 \\ 0.101 & -0.435 & -0.406 & -0.796 & 0.324 \end{bmatrix}$$

$$W_c = \begin{bmatrix} -0.901 & -0.877 & -0.413 & 0.16 & -0.775 \\ -0.196 & 0.077 & 0.769 & -0.567 & -0.905 \end{bmatrix}$$

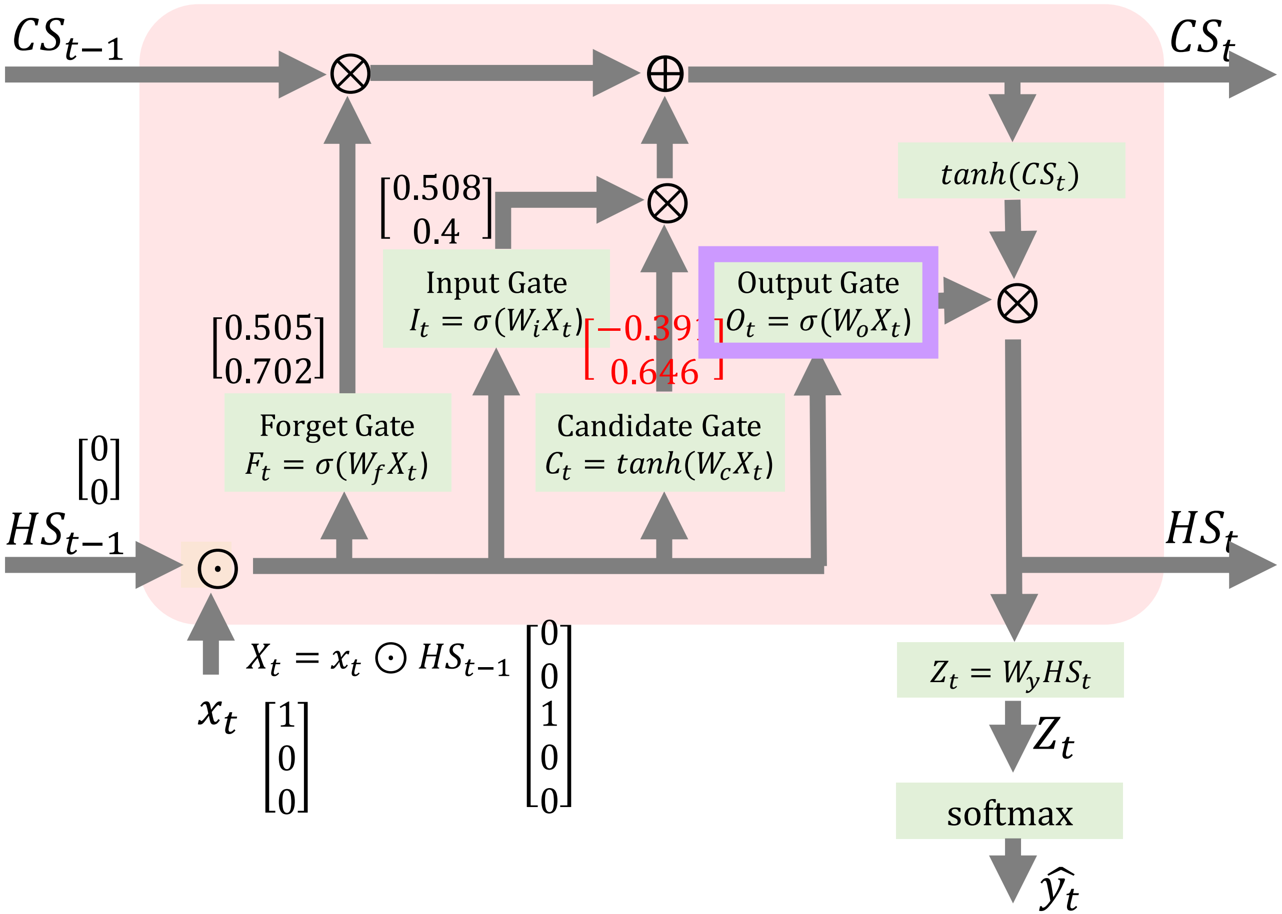
$$W_o = \begin{bmatrix} 0.668 & -0.605 & -0.402 & -0.691 & -0.486 \\ 0.613 & 0.875 & 0.549 & -0.623 & 0.262 \end{bmatrix}$$

$$W_y = \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix}$$

$$O_t = \sigma(W_o X_t)$$

$$= \sigma\left(\begin{bmatrix} 0.668 & -0.605 & -0.402 & -0.691 & -0.486 \\ 0.613 & 0.875 & 0.549 & -0.623 & 0.262 \end{bmatrix}\begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}\right)$$

$$= \sigma\left(\begin{bmatrix} -0.402 \\ 0.549 \end{bmatrix}\right) = \begin{bmatrix} 0.401 \\ 0.634 \end{bmatrix}$$



신박AI

이제는 셀상태 CS를 업데이트 해보도록 하겠습니다

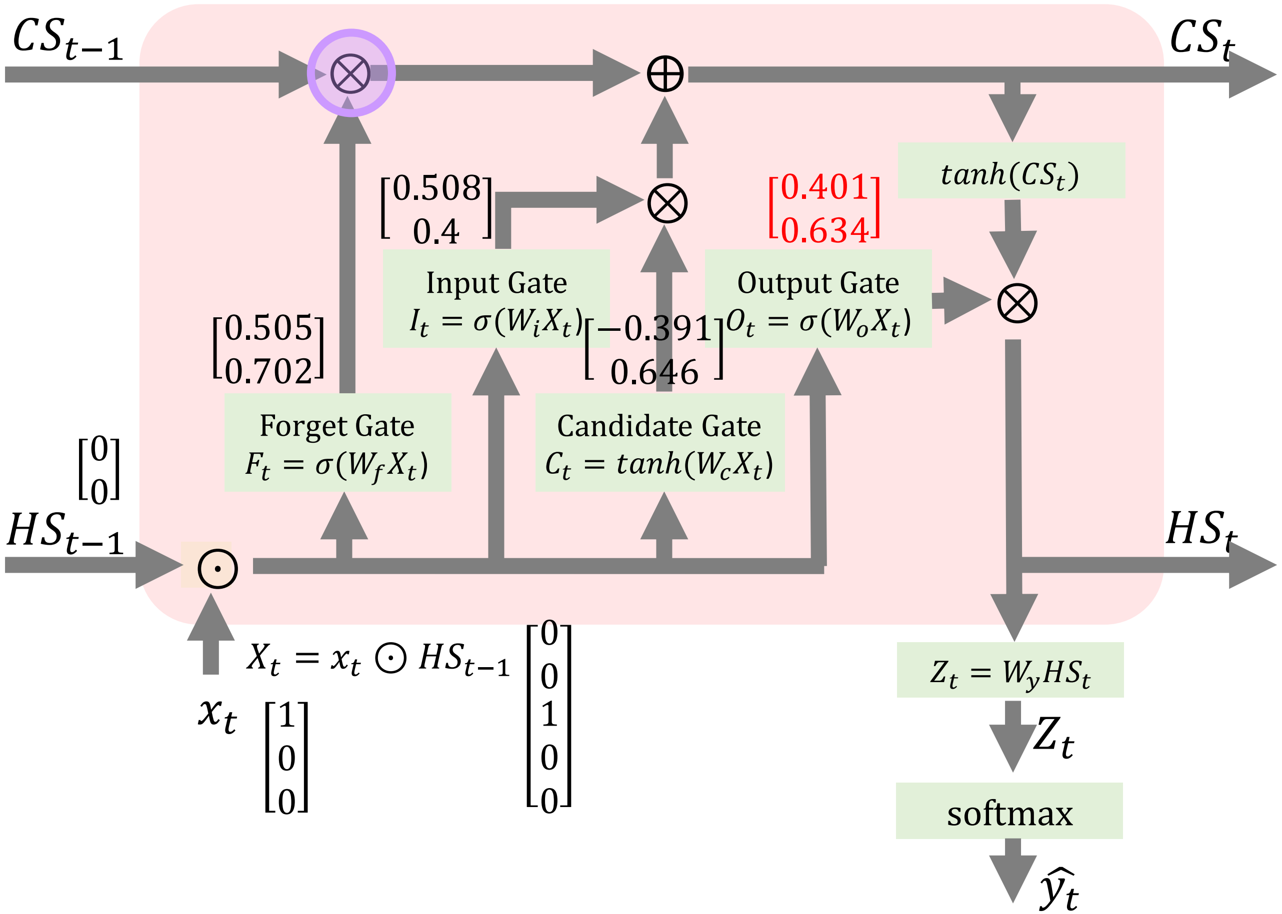
$$W_f = \begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix}$$

$$W_i = \begin{bmatrix} -0.209 & -0.14 & 0.031 & 0.226 & 0.696 \\ 0.101 & -0.435 & -0.406 & -0.796 & 0.324 \end{bmatrix}$$

$$W_c = \begin{bmatrix} -0.901 & -0.877 & -0.413 & 0.16 & -0.775 \\ -0.196 & 0.077 & 0.769 & -0.567 & -0.905 \end{bmatrix}$$

$$W_o = \begin{bmatrix} 0.668 & -0.605 & -0.402 & -0.691 & -0.486 \\ 0.613 & 0.875 & 0.549 & -0.623 & 0.262 \end{bmatrix}$$

$$W_y = \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix}$$



신박AI

이번에도 계산 편의상 CS_{t-1}는 [1,0]으로 하겠습니다

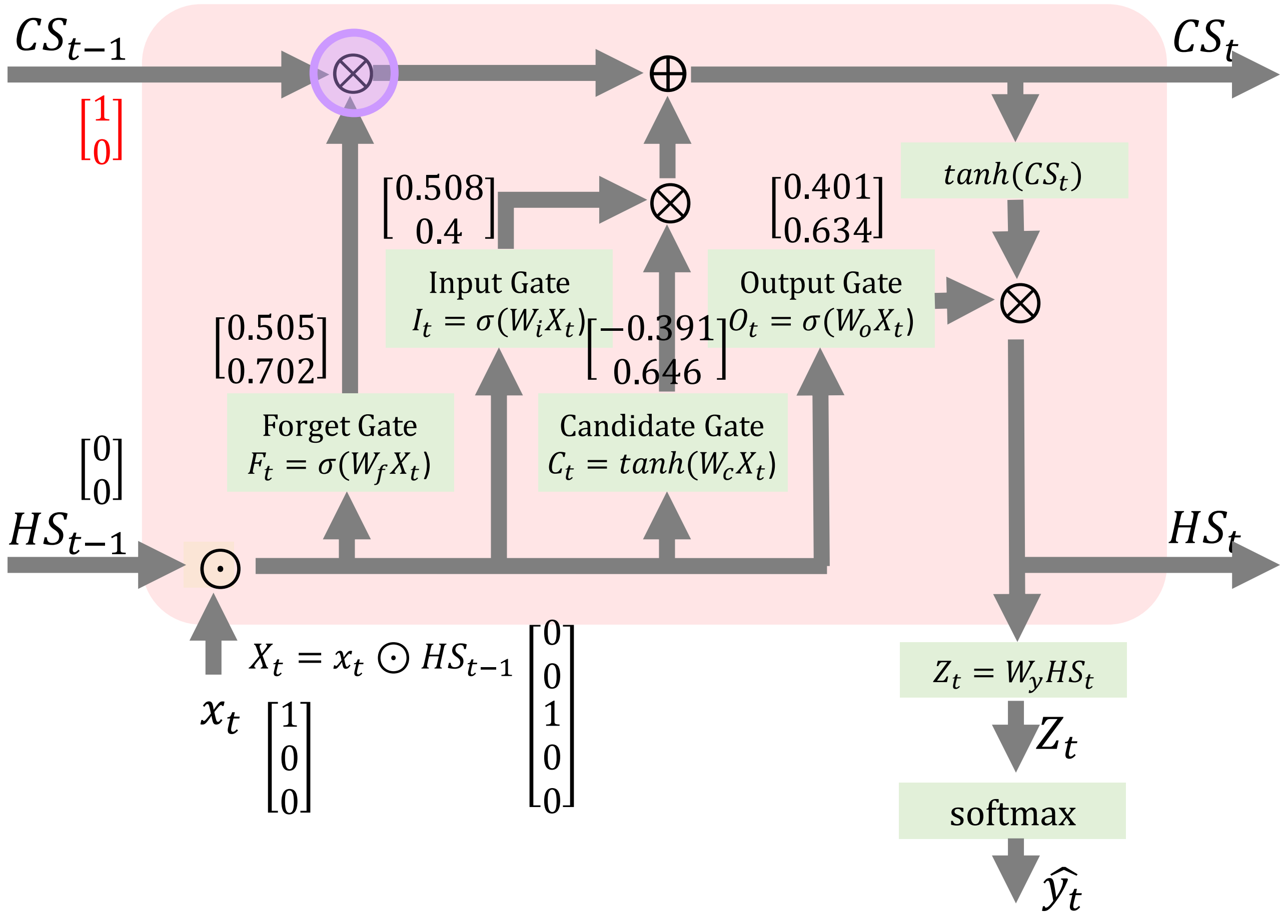
$$W_f = \begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix}$$

$$W_i = \begin{bmatrix} -0.209 & -0.14 & 0.031 & 0.226 & 0.696 \\ 0.101 & -0.435 & -0.406 & -0.796 & 0.324 \end{bmatrix}$$

$$W_c = \begin{bmatrix} -0.901 & -0.877 & -0.413 & 0.16 & -0.775 \\ -0.196 & 0.077 & 0.769 & -0.567 & -0.905 \end{bmatrix}$$

$$W_o = \begin{bmatrix} 0.668 & -0.605 & -0.402 & -0.691 & -0.486 \\ 0.613 & 0.875 & 0.549 & -0.623 & 0.262 \end{bmatrix}$$

$$W_y = \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix}$$



신박AI

그러면 이렇게 element-wise곱을 하게 되면 다음과 같이 됩니다

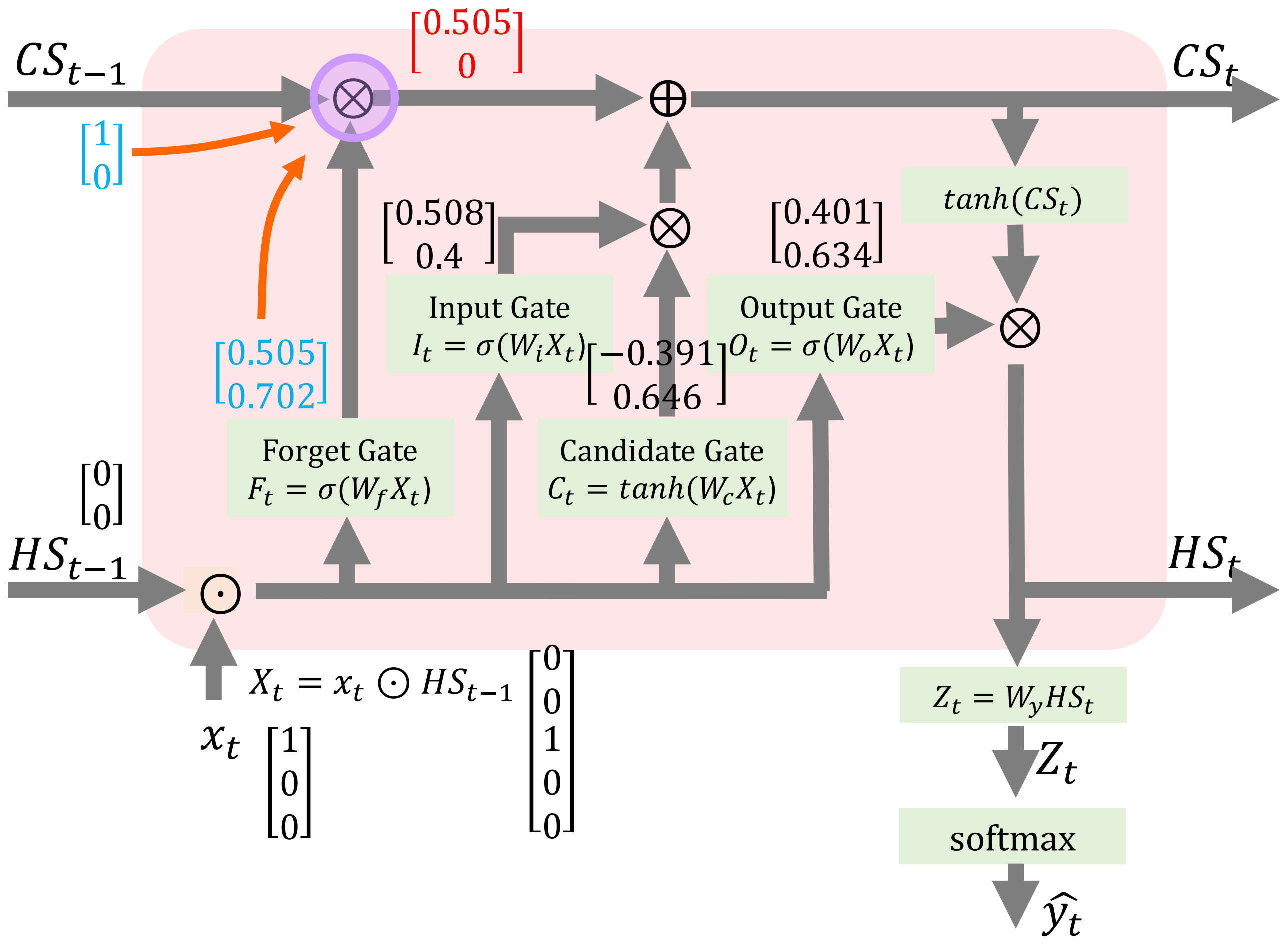
$$W_f = \begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix}$$

$$W_i = \begin{bmatrix} -0.209 & -0.14 & 0.031 & 0.226 & 0.696 \\ 0.101 & -0.435 & -0.406 & -0.796 & 0.324 \end{bmatrix}$$

$$W_c = \begin{bmatrix} -0.901 & -0.877 & -0.413 & 0.16 & -0.775 \\ -0.196 & 0.077 & 0.769 & -0.567 & -0.905 \end{bmatrix}$$

$$W_o = \begin{bmatrix} 0.668 & -0.605 & -0.402 & -0.691 & -0.486 \\ 0.613 & 0.875 & 0.549 & -0.623 & 0.262 \end{bmatrix}$$

$$W_y = \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix}$$



신박AI

그리고 이 둘을 또 element-wise곱을 하면 이렇게 됩니다

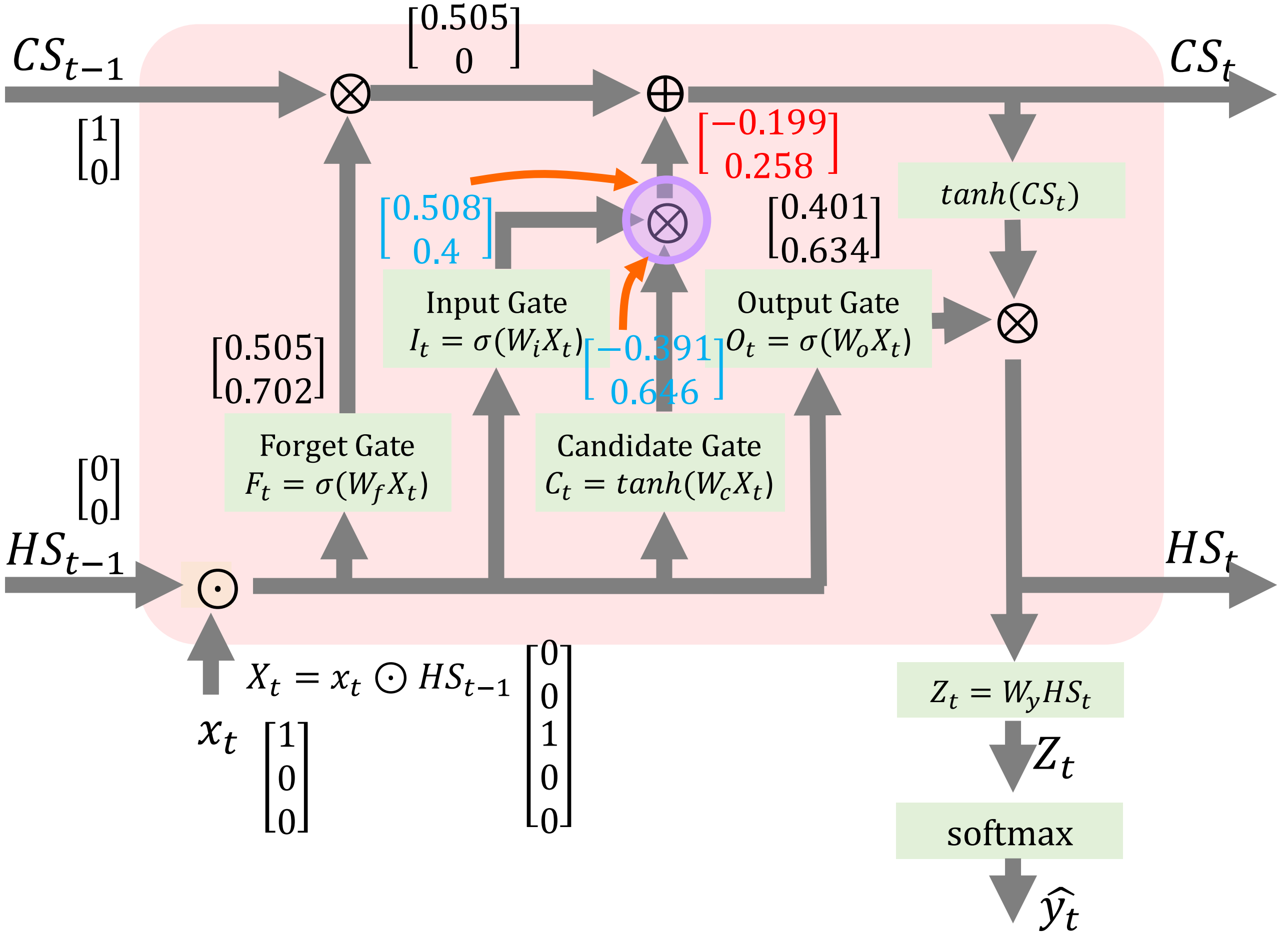
$$W_f = \begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix}$$

$$W_i = \begin{bmatrix} -0.209 & -0.14 & 0.031 & 0.226 & 0.696 \\ 0.101 & -0.435 & -0.406 & -0.796 & 0.324 \end{bmatrix}$$

$$W_c = \begin{bmatrix} -0.901 & -0.877 & -0.413 & 0.16 & -0.775 \\ -0.196 & 0.077 & 0.769 & -0.567 & -0.905 \end{bmatrix}$$

$$W_o = \begin{bmatrix} 0.668 & -0.605 & -0.402 & -0.691 & -0.486 \\ 0.613 & 0.875 & 0.549 & -0.623 & 0.262 \end{bmatrix}$$

$$W_y = \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix}$$



신박AI

그러면 이 둘을 더하면 다음과 같이 됩니다

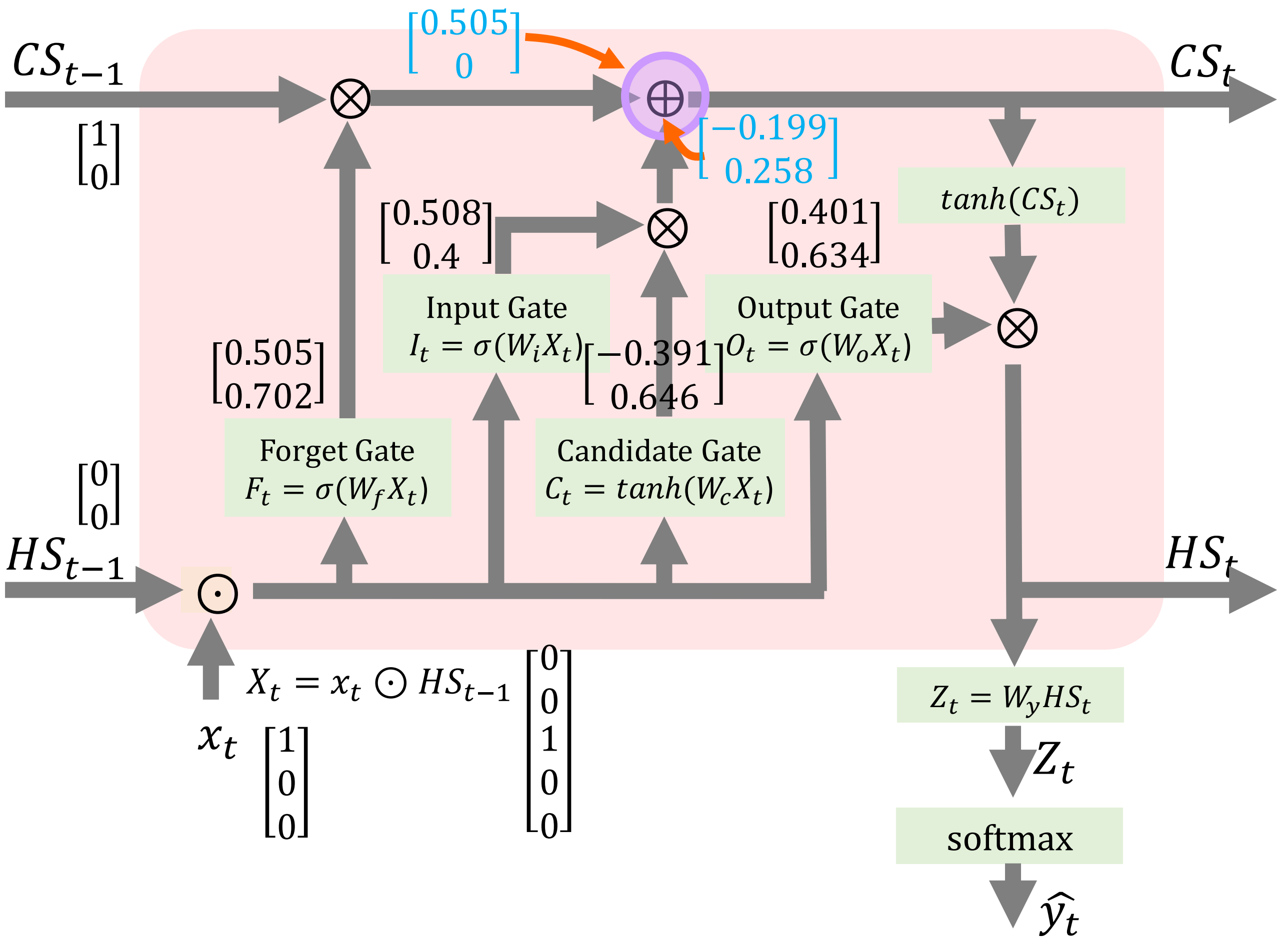
$$W_f = \begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix}$$

$$W_i = \begin{bmatrix} -0.209 & -0.14 & 0.031 & 0.226 & 0.696 \\ 0.101 & -0.435 & -0.406 & -0.796 & 0.324 \end{bmatrix}$$

$$W_c = \begin{bmatrix} -0.901 & -0.877 & -0.413 & 0.16 & -0.775 \\ -0.196 & 0.077 & 0.769 & -0.567 & -0.905 \end{bmatrix}$$

$$W_o = \begin{bmatrix} 0.668 & -0.605 & -0.402 & -0.691 & -0.486 \\ 0.613 & 0.875 & 0.549 & -0.623 & 0.262 \end{bmatrix}$$

$$W_y = \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix}$$



신박AI

새로운 셀 상태인 CS_t는 이렇게 계산이 됩니다

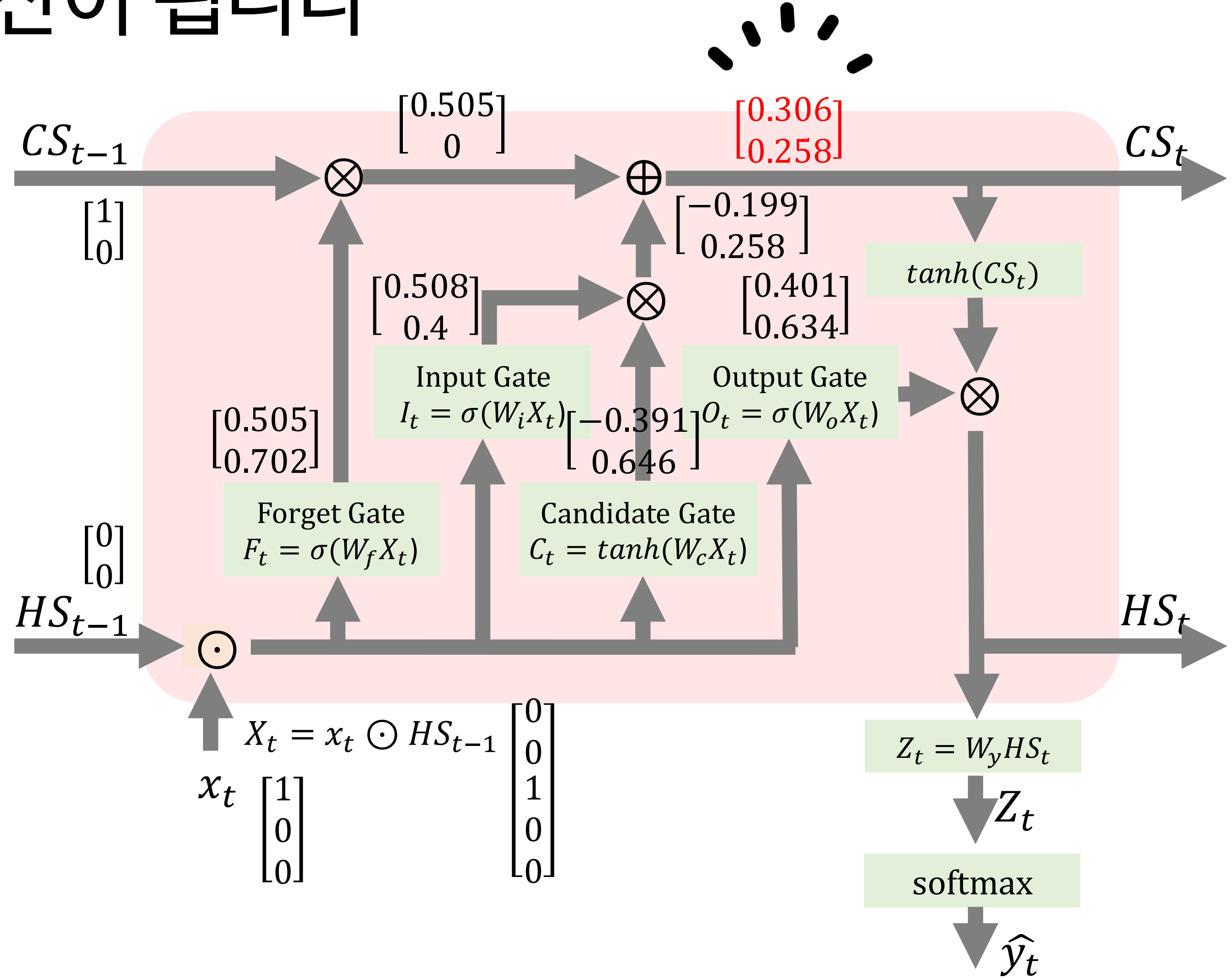
$$W_f = \begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix}$$

$$W_i = \begin{bmatrix} -0.209 & -0.14 & 0.031 & 0.226 & 0.696 \\ 0.101 & -0.435 & -0.406 & -0.796 & 0.324 \end{bmatrix}$$

$$W_c = \begin{bmatrix} -0.901 & -0.877 & -0.413 & 0.16 & -0.775 \\ -0.196 & 0.077 & 0.769 & -0.567 & -0.905 \end{bmatrix}$$

$$W_o = \begin{bmatrix} 0.668 & -0.605 & -0.402 & -0.691 & -0.486 \\ 0.613 & 0.875 & 0.549 & -0.623 & 0.262 \end{bmatrix}$$

$$W_y = \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix}$$



신박AI

새로운 CS_t를 tanh에 넣으면 이렇게 계산이 됩니다

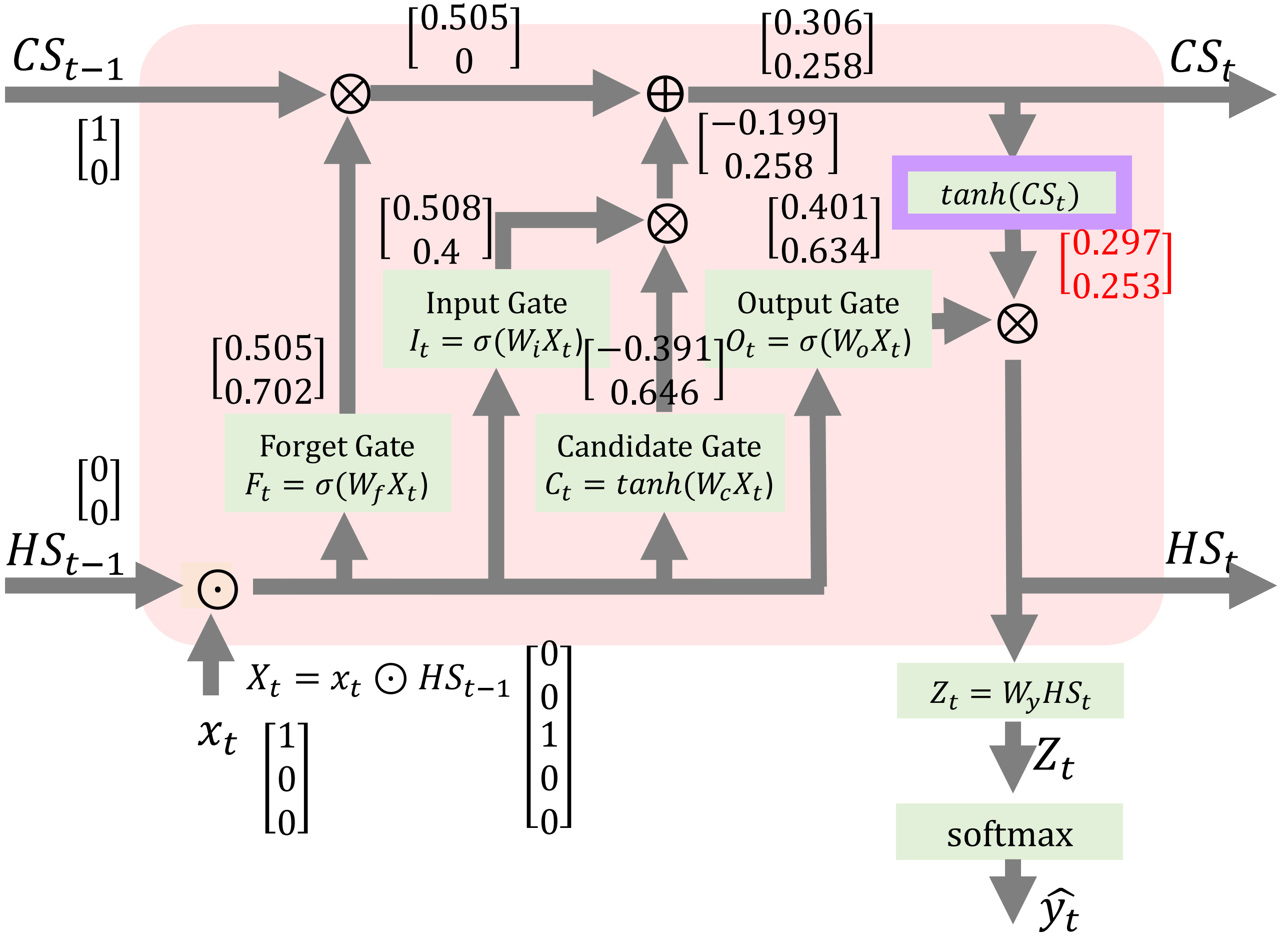
$$W_f = \begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix}$$

$$W_i = \begin{bmatrix} -0.209 & -0.14 & 0.031 & 0.226 & 0.696 \\ 0.101 & -0.435 & -0.406 & -0.796 & 0.324 \end{bmatrix}$$

$$W_c = \begin{bmatrix} -0.901 & -0.877 & -0.413 & 0.16 & -0.775 \\ -0.196 & 0.077 & 0.769 & -0.567 & -0.905 \end{bmatrix}$$

$$W_o = \begin{bmatrix} 0.668 & -0.605 & -0.402 & -0.691 & -0.486 \\ 0.613 & 0.875 & 0.549 & -0.623 & 0.262 \end{bmatrix}$$

$$W_y = \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix}$$



신박AI

이제 이 둘을 element-wise 곱할 차례입니다

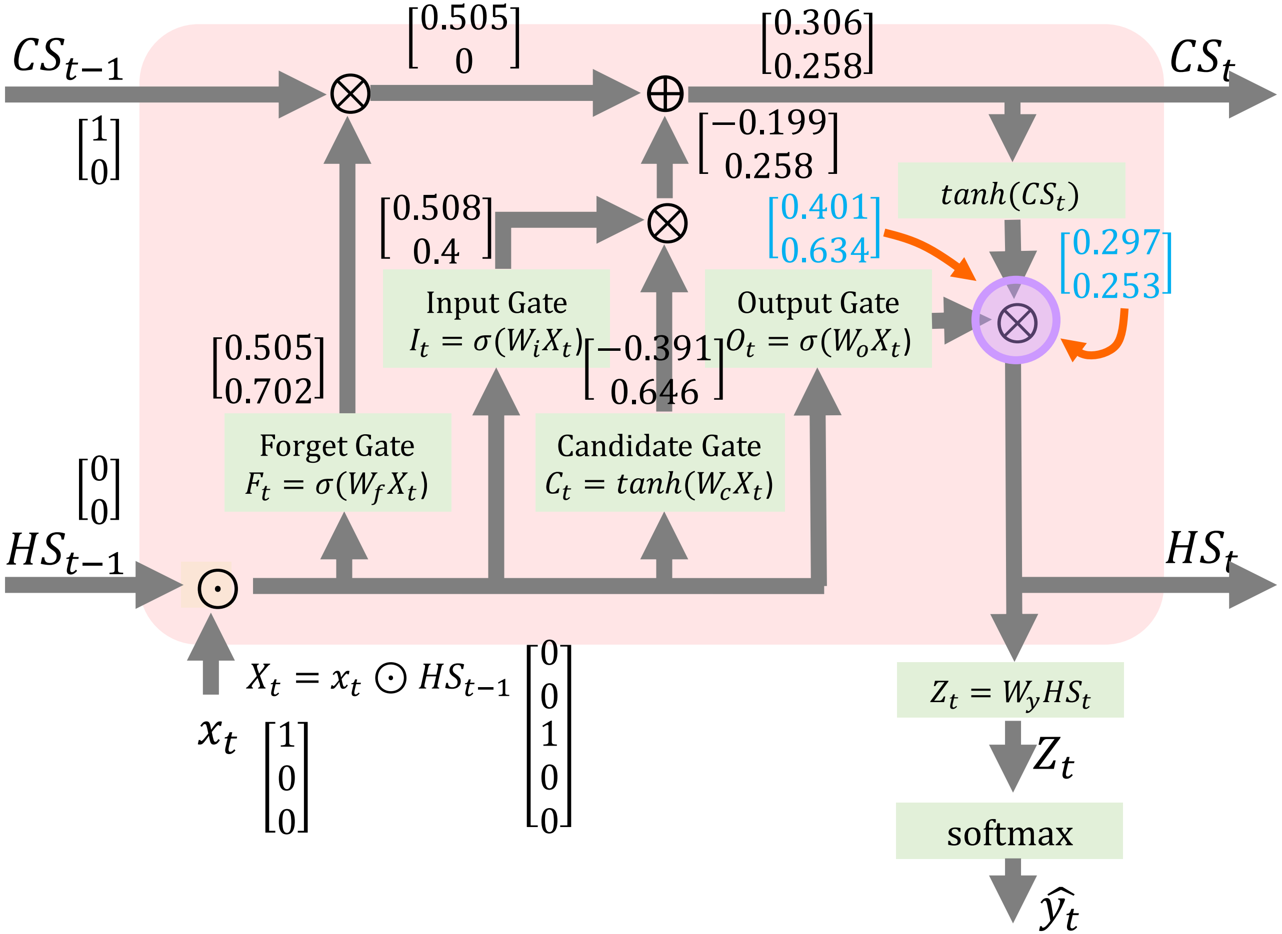
$$W_f = \begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix}$$

$$W_i = \begin{bmatrix} -0.209 & -0.14 & 0.031 & 0.226 & 0.696 \\ 0.101 & -0.435 & -0.406 & -0.796 & 0.324 \end{bmatrix}$$

$$W_c = \begin{bmatrix} -0.901 & -0.877 & -0.413 & 0.16 & -0.775 \\ -0.196 & 0.077 & 0.769 & -0.567 & -0.905 \end{bmatrix}$$

$$W_o = \begin{bmatrix} 0.668 & -0.605 & -0.402 & -0.691 & -0.486 \\ 0.613 & 0.875 & 0.549 & -0.623 & 0.262 \end{bmatrix}$$

$$W_y = \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix}$$



신박AI

그러면 새로운 히든 상태인 HS_t는 다음과 같이 계산이 됩니다

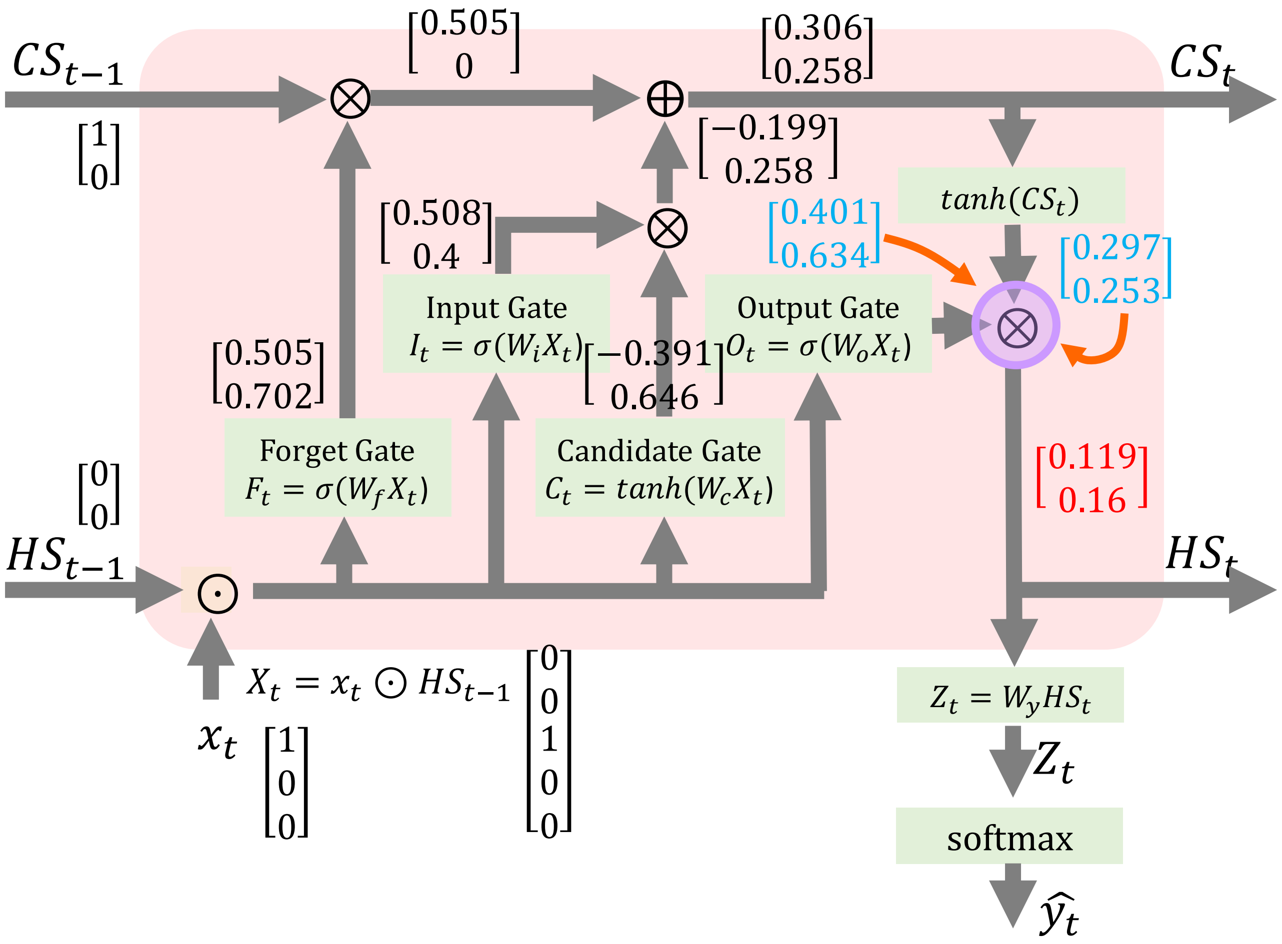
$$W_f = \begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix}$$

$$W_i = \begin{bmatrix} -0.209 & -0.14 & 0.031 & 0.226 & 0.696 \\ 0.101 & -0.435 & -0.406 & -0.796 & 0.324 \end{bmatrix}$$

$$W_c = \begin{bmatrix} -0.901 & -0.877 & -0.413 & 0.16 & -0.775 \\ -0.196 & 0.077 & 0.769 & -0.567 & -0.905 \end{bmatrix}$$

$$W_o = \begin{bmatrix} 0.668 & -0.605 & -0.402 & -0.691 & -0.486 \\ 0.613 & 0.875 & 0.549 & -0.623 & 0.262 \end{bmatrix}$$

$$W_y = \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix}$$



신박AI

이제는 최종 출력값을 계산할 차례입니다

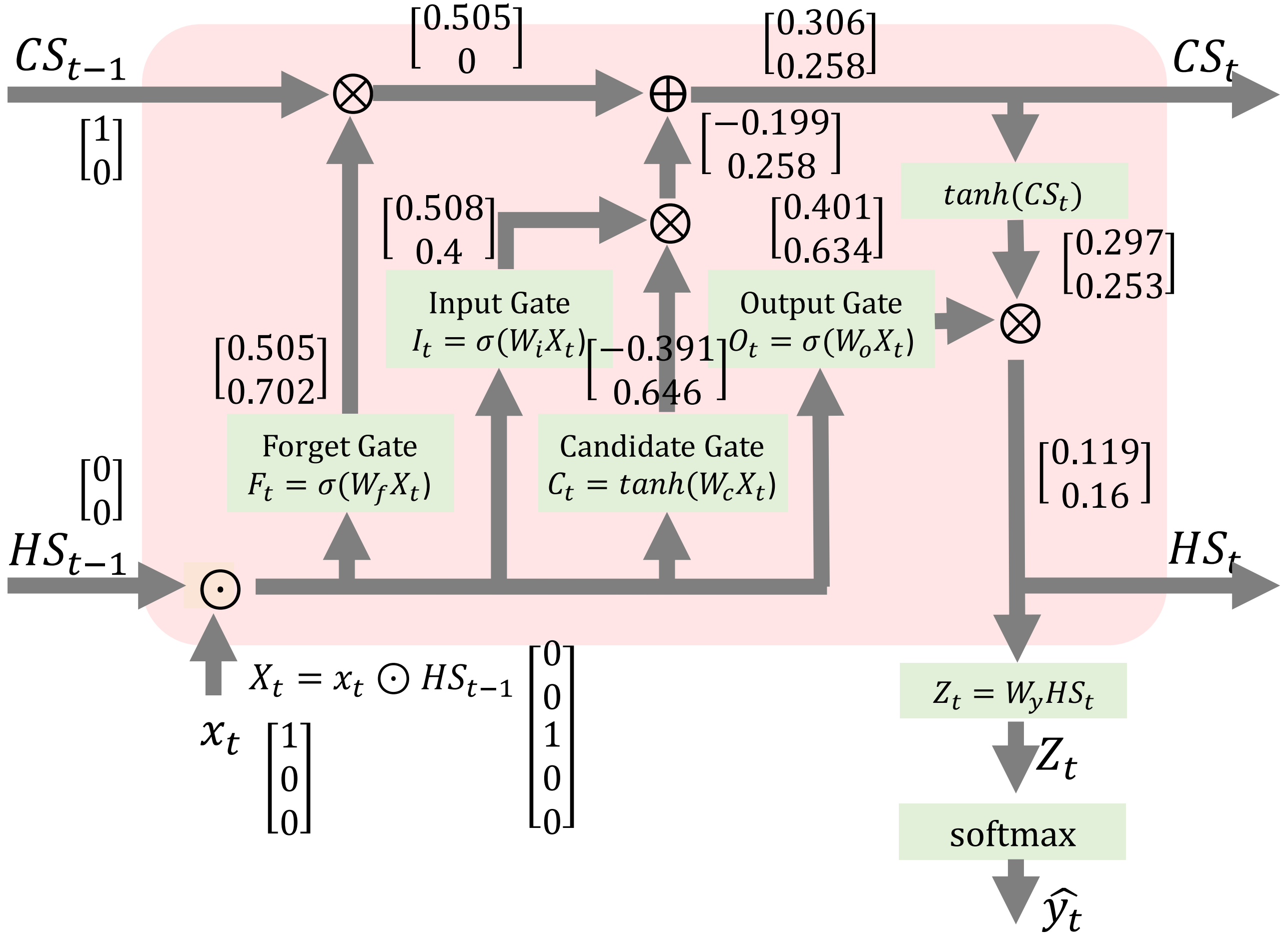
$$W_f = \begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix}$$

$$W_i = \begin{bmatrix} -0.209 & -0.14 & 0.031 & 0.226 & 0.696 \\ 0.101 & -0.435 & -0.406 & -0.796 & 0.324 \end{bmatrix}$$

$$W_c = \begin{bmatrix} -0.901 & -0.877 & -0.413 & 0.16 & -0.775 \\ -0.196 & 0.077 & 0.769 & -0.567 & -0.905 \end{bmatrix}$$

$$W_o = \begin{bmatrix} 0.668 & -0.605 & -0.402 & -0.691 & -0.486 \\ 0.613 & 0.875 & 0.549 & -0.623 & 0.262 \end{bmatrix}$$

$$W_y = \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix}$$



신박AI

마지막 Z_t층은 일종의 fully-connected 층처럼 사용되어서

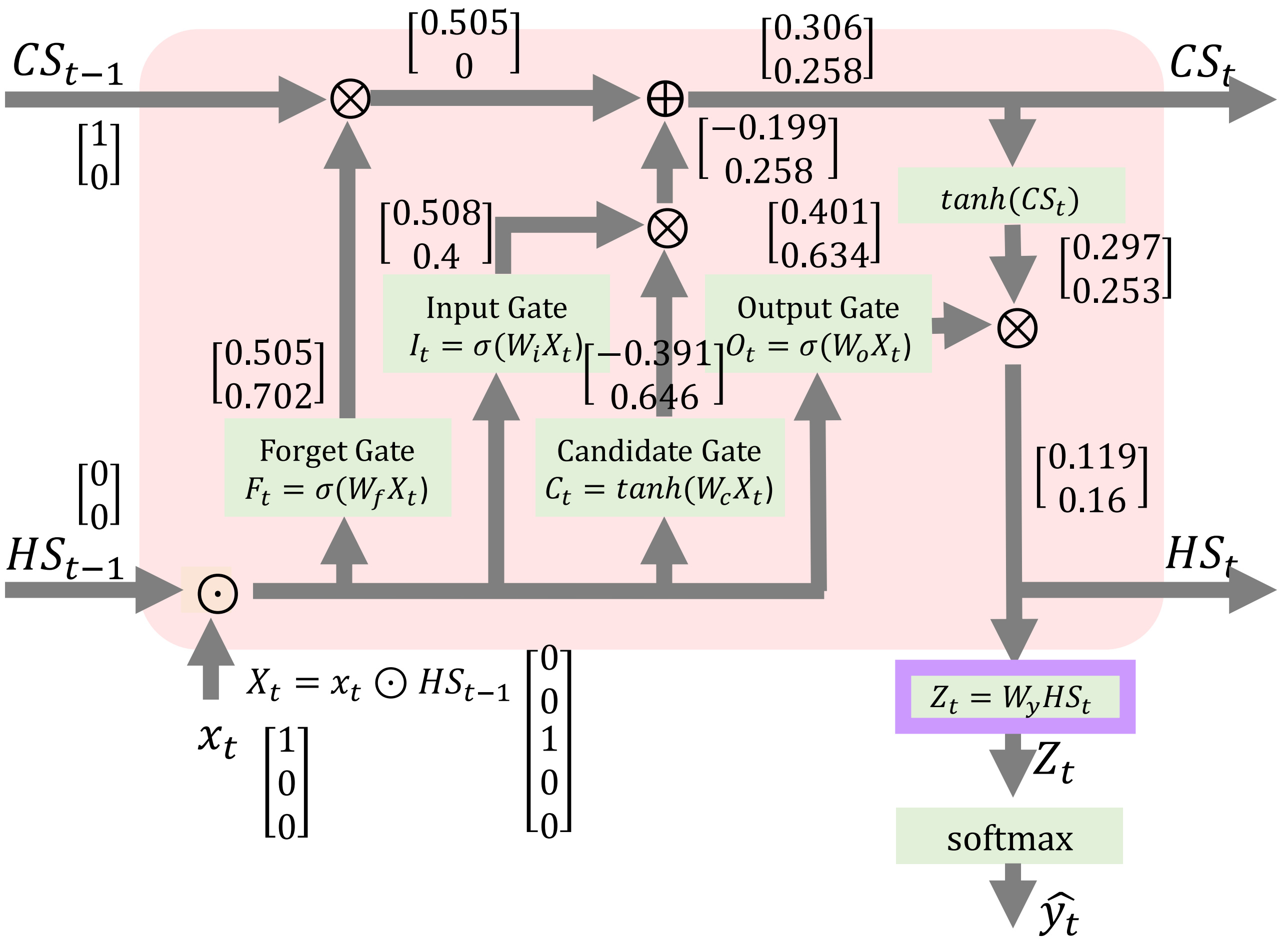
$$W_f = \begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix}$$

$$W_i = \begin{bmatrix} -0.209 & -0.14 & 0.031 & 0.226 & 0.696 \\ 0.101 & -0.435 & -0.406 & -0.796 & 0.324 \end{bmatrix}$$

$$W_c = \begin{bmatrix} -0.901 & -0.877 & -0.413 & 0.16 & -0.775 \\ -0.196 & 0.077 & 0.769 & -0.567 & -0.905 \end{bmatrix}$$

$$W_o = \begin{bmatrix} 0.668 & -0.605 & -0.402 & -0.691 & -0.486 \\ 0.613 & 0.875 & 0.549 & -0.623 & 0.262 \end{bmatrix}$$

$$W_y = \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix}$$



신박AI

LSTM의 내부상태의 길이가 몇이 되었든,

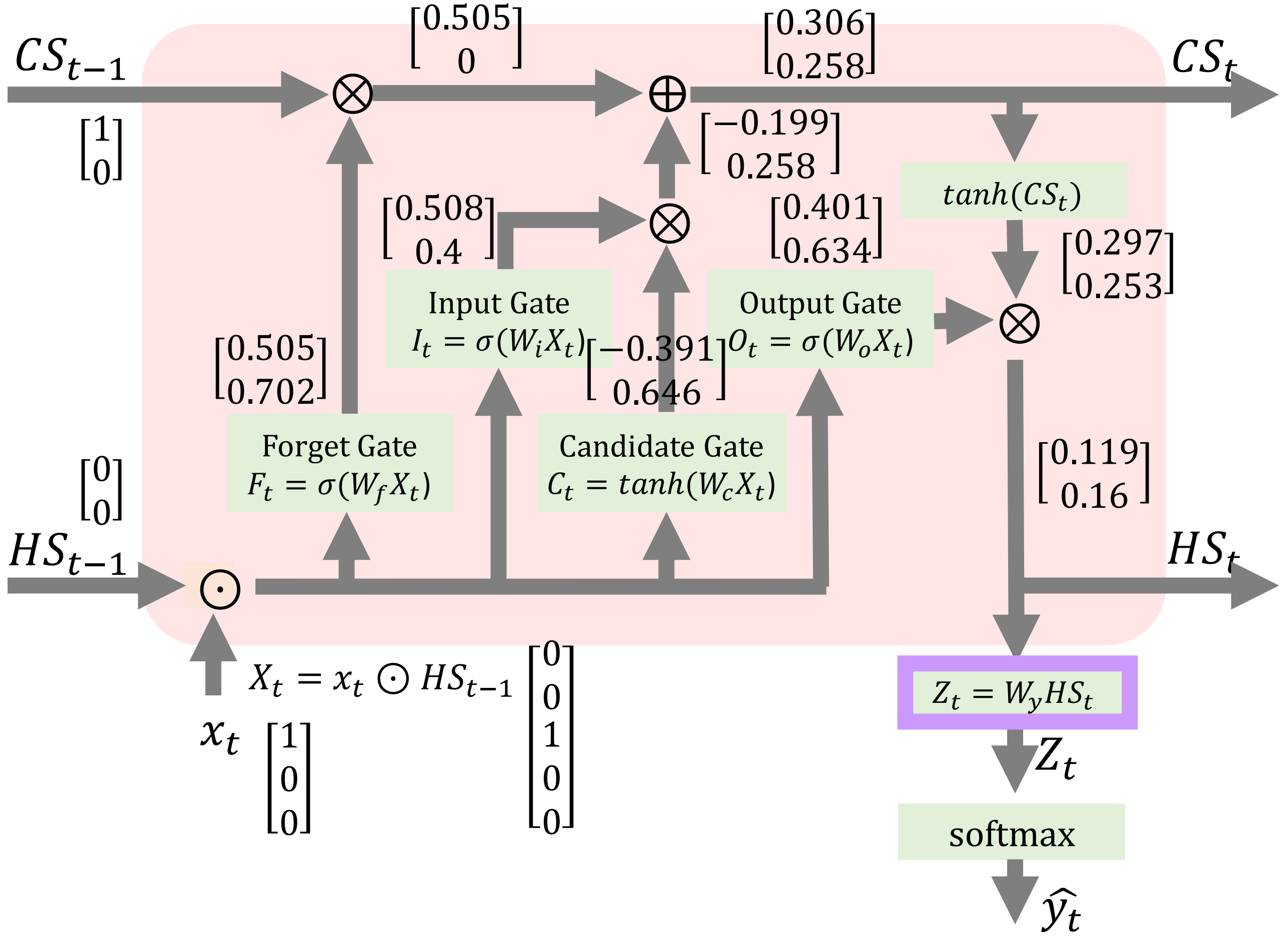
$$W_f = \begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix}$$

$$W_i = \begin{bmatrix} -0.209 & -0.14 & 0.031 & 0.226 & 0.696 \\ 0.101 & -0.435 & -0.406 & -0.796 & 0.324 \end{bmatrix}$$

$$W_c = \begin{bmatrix} -0.901 & -0.877 & -0.413 & 0.16 & -0.775 \\ -0.196 & 0.077 & 0.769 & -0.567 & -0.905 \end{bmatrix}$$

$$W_o = \begin{bmatrix} 0.668 & -0.605 & -0.402 & -0.691 & -0.486 \\ 0.613 & 0.875 & 0.549 & -0.623 & 0.262 \end{bmatrix}$$

$$W_y = \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix}$$



신박AI

최종 출력값의 길이로 (지금은 3)바꾸어 줍니다

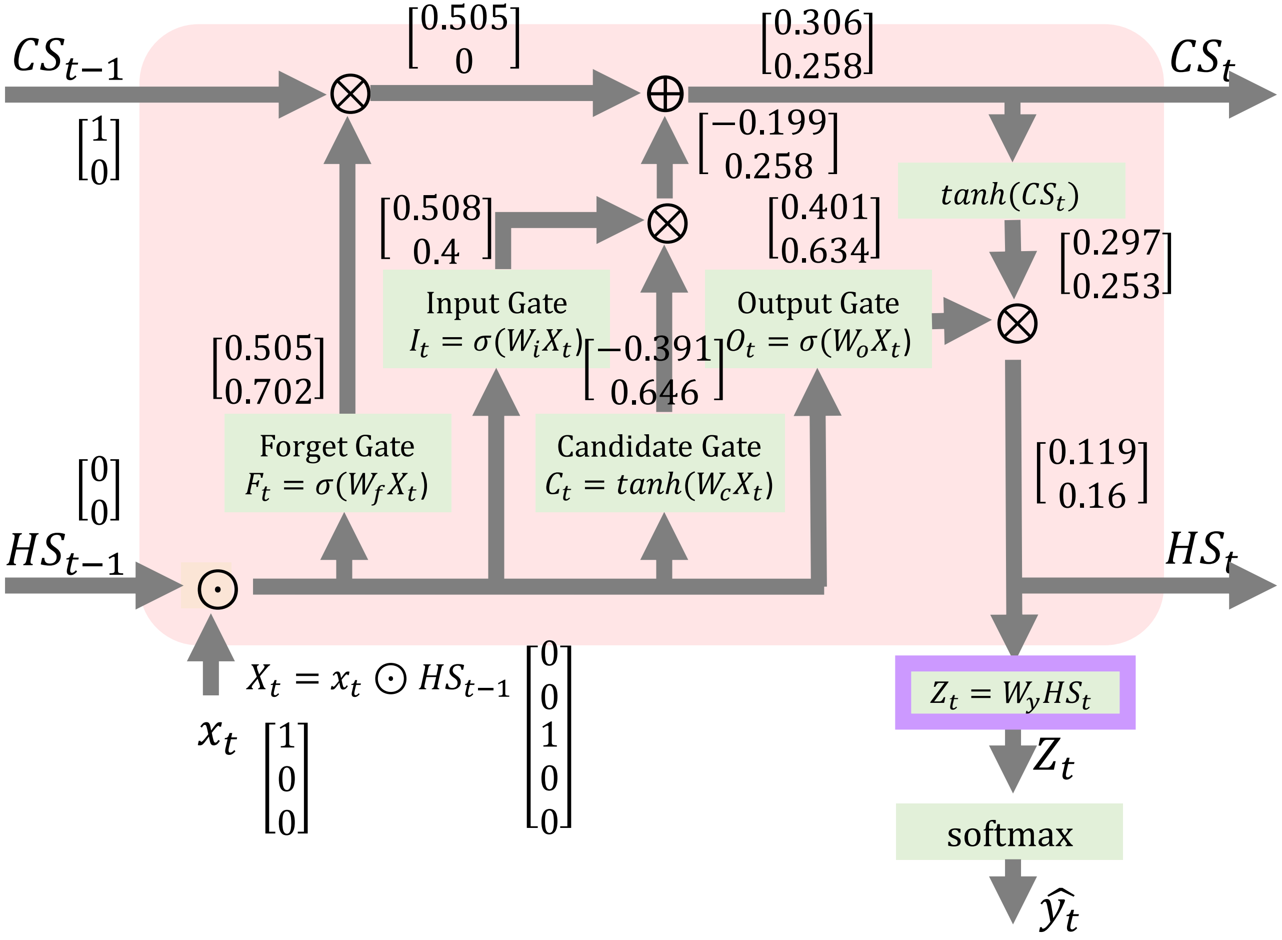
$$W_f = \begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix}$$

$$W_i = \begin{bmatrix} -0.209 & -0.14 & 0.031 & 0.226 & 0.696 \\ 0.101 & -0.435 & -0.406 & -0.796 & 0.324 \end{bmatrix}$$

$$W_c = \begin{bmatrix} -0.901 & -0.877 & -0.413 & 0.16 & -0.775 \\ -0.196 & 0.077 & 0.769 & -0.567 & -0.905 \end{bmatrix}$$

$$W_o = \begin{bmatrix} 0.668 & -0.605 & -0.402 & -0.691 & -0.486 \\ 0.613 & 0.875 & 0.549 & -0.623 & 0.262 \end{bmatrix}$$

$$W_y = \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix}$$



신박AI

그러면 최종 출력값을 계산해보도록 하겠습니다

$$W_f = \begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix}$$

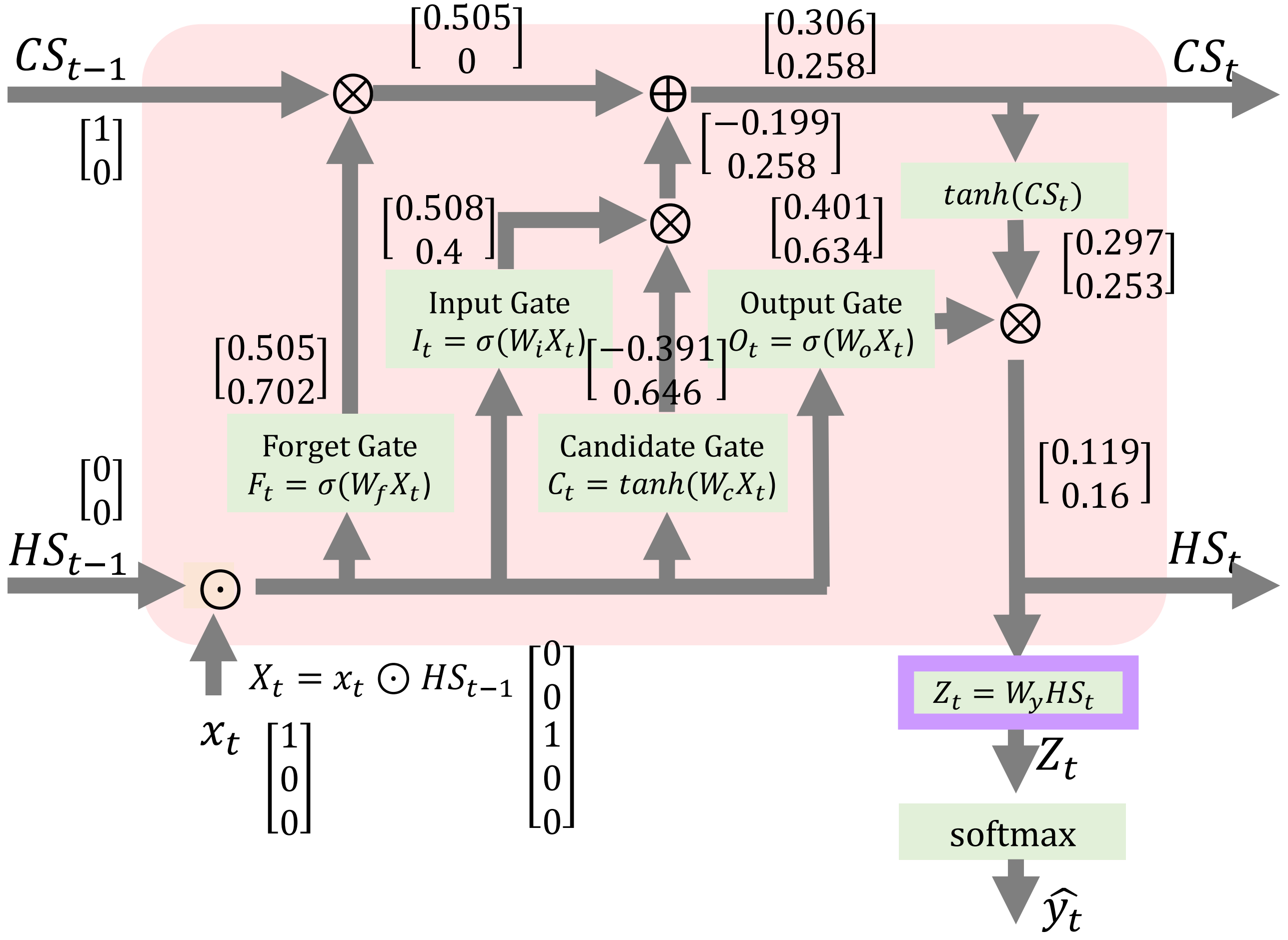
$$W_i = \begin{bmatrix} -0.209 & -0.14 & 0.031 & 0.226 & 0.696 \\ 0.101 & -0.435 & -0.406 & -0.796 & 0.324 \end{bmatrix}$$

$$W_c = \begin{bmatrix} -0.901 & -0.877 & -0.413 & 0.16 & -0.775 \\ -0.196 & 0.077 & 0.769 & -0.567 & -0.905 \end{bmatrix}$$

$$W_o = \begin{bmatrix} 0.668 & -0.605 & -0.402 & -0.691 & -0.486 \\ 0.613 & 0.875 & 0.549 & -0.623 & 0.262 \end{bmatrix}$$

$$W_y = \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix}$$

$$Z_t = W_y HS_t$$



신박AI

그러면 최종 출력값을 계산해보도록 하겠습니다

$$W_f = \begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix}$$

$$W_i = \begin{bmatrix} -0.209 & -0.14 & 0.031 & 0.226 & 0.696 \\ 0.101 & -0.435 & -0.406 & -0.796 & 0.324 \end{bmatrix}$$

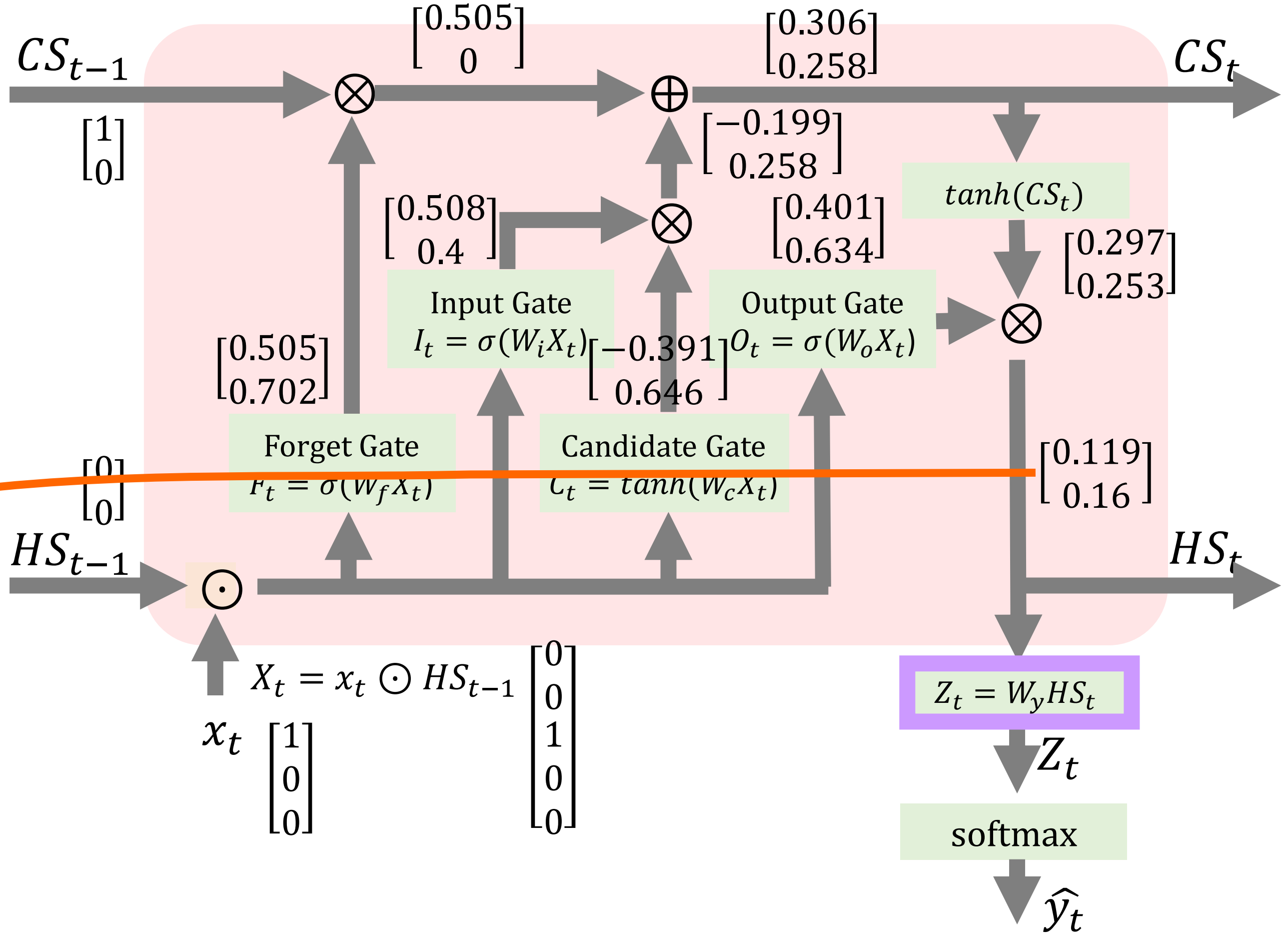
$$W_c = \begin{bmatrix} -0.901 & -0.877 & -0.413 & 0.16 & -0.775 \\ -0.196 & 0.077 & 0.769 & -0.567 & -0.905 \end{bmatrix}$$

$$W_o = \begin{bmatrix} 0.668 & -0.605 & -0.402 & -0.691 & -0.486 \\ 0.613 & 0.875 & 0.549 & -0.623 & 0.262 \end{bmatrix}$$

$$W_y = \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix}$$

$$Z_t = W_y HS_t$$

$$= \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix} \begin{bmatrix} 0.119 \\ 0.16 \end{bmatrix}$$



이렇게 길이가 3인 Z_t값을 계산하였습니다

$$W_f = \begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix}$$

$$W_i = \begin{bmatrix} -0.209 & -0.14 & 0.031 & 0.226 & 0.696 \\ 0.101 & -0.435 & -0.406 & -0.796 & 0.324 \end{bmatrix}$$

$$W_c = \begin{bmatrix} -0.901 & -0.877 & -0.413 & 0.16 & -0.775 \\ -0.196 & 0.077 & 0.769 & -0.567 & -0.905 \end{bmatrix}$$

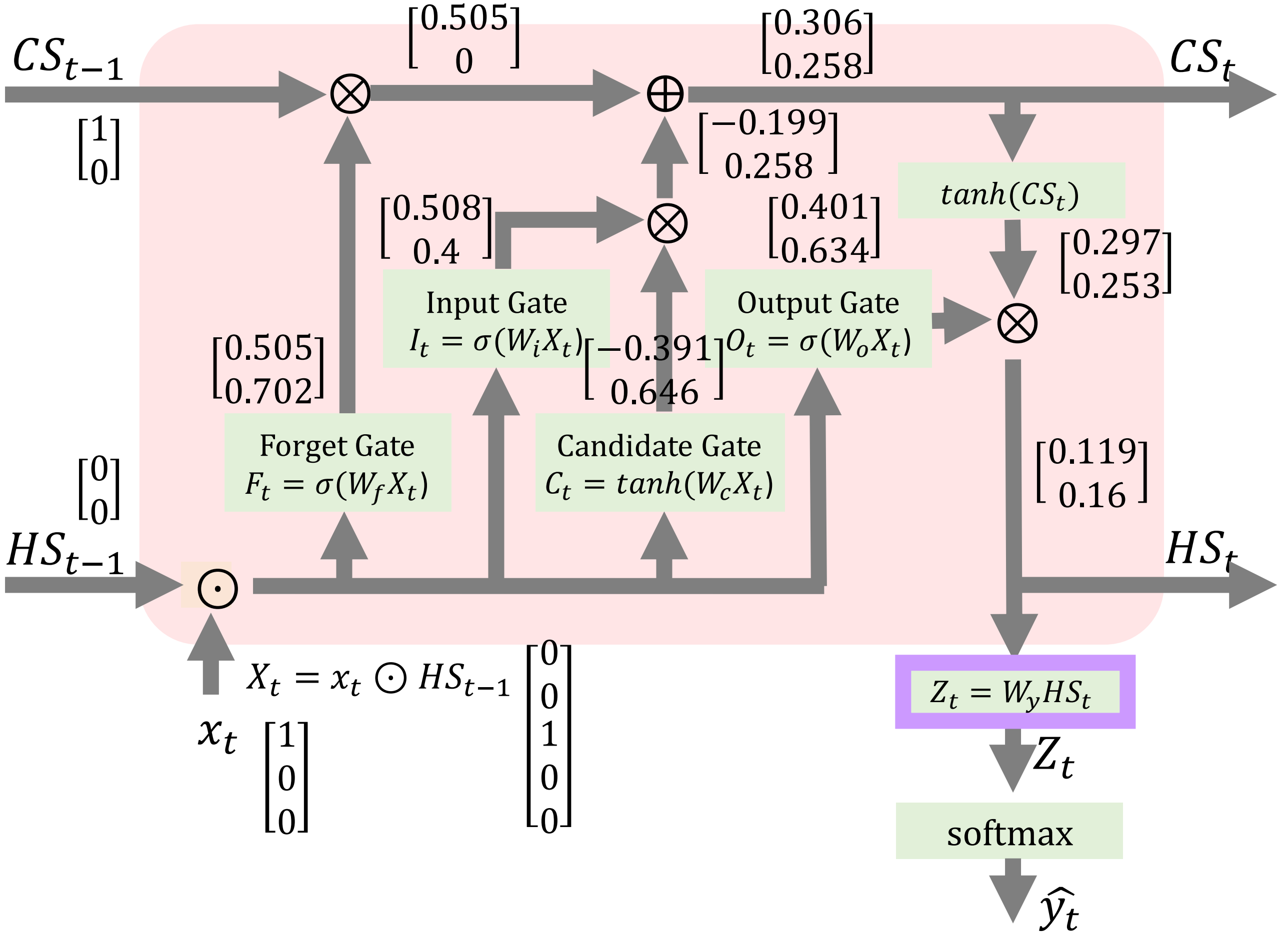
$$W_o = \begin{bmatrix} 0.668 & -0.605 & -0.402 & -0.691 & -0.486 \\ 0.613 & 0.875 & 0.549 & -0.623 & 0.262 \end{bmatrix}$$

$$W_y = \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix}$$

$$Z_t = W_y HS_t$$

$$= \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix} \begin{bmatrix} 0.119 \\ 0.16 \end{bmatrix}$$

$$= \begin{bmatrix} 0.011 \\ 0.109 \\ 0.168 \end{bmatrix}$$



신박AI

보통 Z_t값을 신경망의 raw output이라고도 부르고,

$$W_f = \begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix}$$

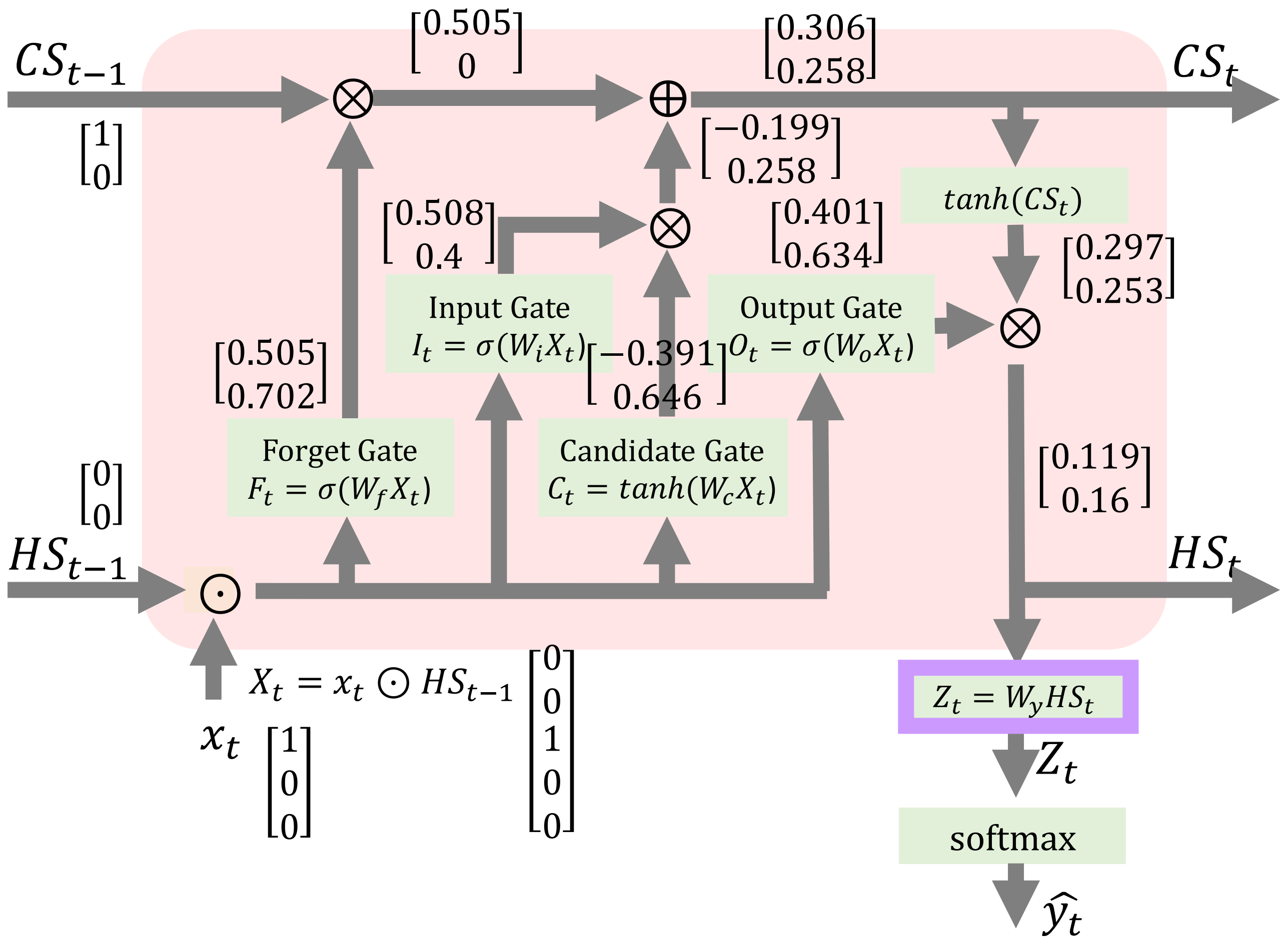
$$W_i = \begin{bmatrix} -0.209 & -0.14 & 0.031 & 0.226 & 0.696 \\ 0.101 & -0.435 & -0.406 & -0.796 & 0.324 \end{bmatrix}$$

$$W_c = \begin{bmatrix} -0.901 & -0.877 & -0.413 & 0.16 & -0.775 \\ -0.196 & 0.077 & 0.769 & -0.567 & -0.905 \end{bmatrix}$$

$$W_o = \begin{bmatrix} 0.668 & -0.605 & -0.402 & -0.691 & -0.486 \\ 0.613 & 0.875 & 0.549 & -0.623 & 0.262 \end{bmatrix}$$

$$W_y = \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix}$$

$$Z_t = W_y HS_t$$
$$= \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix} \begin{bmatrix} 0.119 \\ 0.16 \end{bmatrix}$$
$$= \begin{bmatrix} 0.011 \\ 0.109 \\ 0.168 \end{bmatrix}$$



Logit이라고도 부릅니다

$$W_f = \begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix}$$

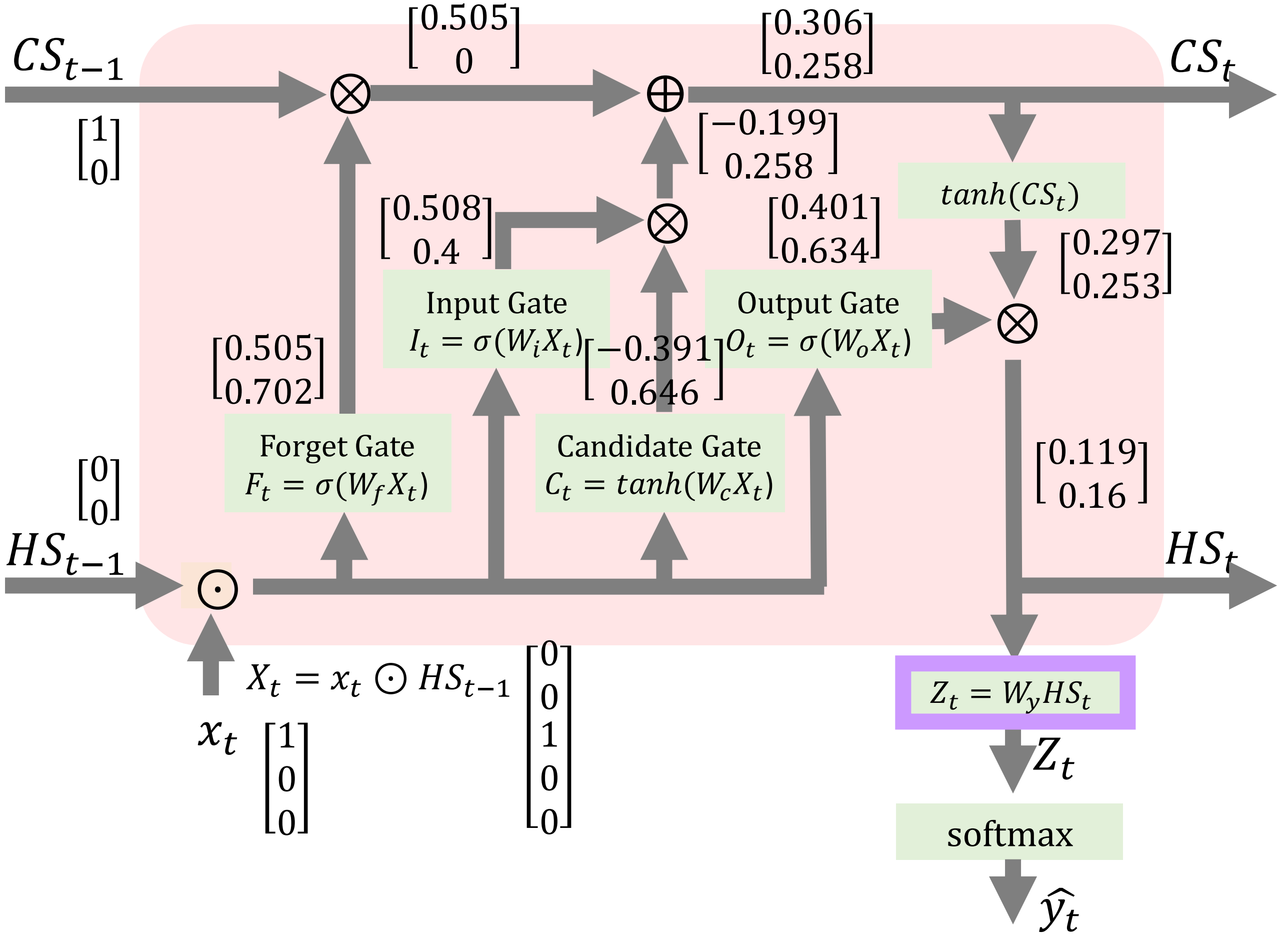
$$W_i = \begin{bmatrix} -0.209 & -0.14 & 0.031 & 0.226 & 0.696 \\ 0.101 & -0.435 & -0.406 & -0.796 & 0.324 \end{bmatrix}$$

$$W_c = \begin{bmatrix} -0.901 & -0.877 & -0.413 & 0.16 & -0.775 \\ -0.196 & 0.077 & 0.769 & -0.567 & -0.905 \end{bmatrix}$$

$$W_o = \begin{bmatrix} 0.668 & -0.605 & -0.402 & -0.691 & -0.486 \\ 0.613 & 0.875 & 0.549 & -0.623 & 0.262 \end{bmatrix}$$

$$W_y = \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix}$$

$$Z_t = W_y HS_t = \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix} \begin{bmatrix} 0.119 \\ 0.16 \end{bmatrix} = \begin{bmatrix} 0.011 \\ 0.109 \\ 0.168 \end{bmatrix}$$



신박AI

이 logit을 softmax함수에 넣어서 loss 계산에 필요한 확률로 바꿉니다

$$W_f = \begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix}$$

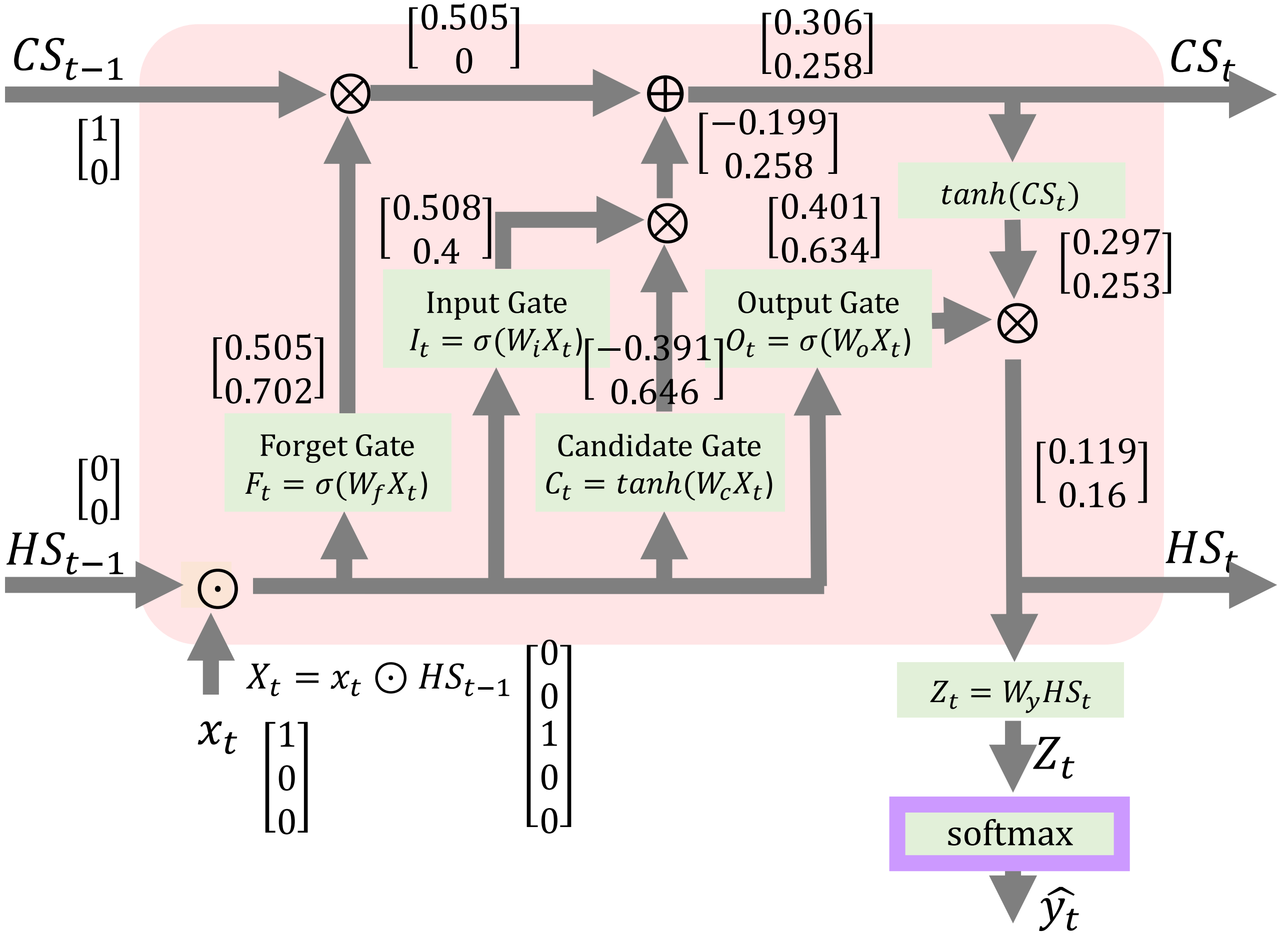
$$W_i = \begin{bmatrix} -0.209 & -0.14 & 0.031 & 0.226 & 0.696 \\ 0.101 & -0.435 & -0.406 & -0.796 & 0.324 \end{bmatrix}$$

$$W_c = \begin{bmatrix} -0.901 & -0.877 & -0.413 & 0.16 & -0.775 \\ -0.196 & 0.077 & 0.769 & -0.567 & -0.905 \end{bmatrix}$$

$$W_o = \begin{bmatrix} 0.668 & -0.605 & -0.402 & -0.691 & -0.486 \\ 0.613 & 0.875 & 0.549 & -0.623 & 0.262 \end{bmatrix}$$

$$W_y = \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix}$$

$$Z_t = W_y HS_t = \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix} \begin{bmatrix} 0.119 \\ 0.16 \end{bmatrix} = \begin{bmatrix} 0.011 \\ 0.109 \\ 0.168 \end{bmatrix}$$



신박AI

이 logit을 softmax함수에 넣어서 loss 계산에 필요한 확률로 바꿉니다

$$W_f = \begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix}$$

$$W_i = \begin{bmatrix} -0.209 & -0.14 & 0.031 & 0.226 & 0.696 \\ 0.101 & -0.435 & -0.406 & -0.796 & 0.324 \end{bmatrix}$$

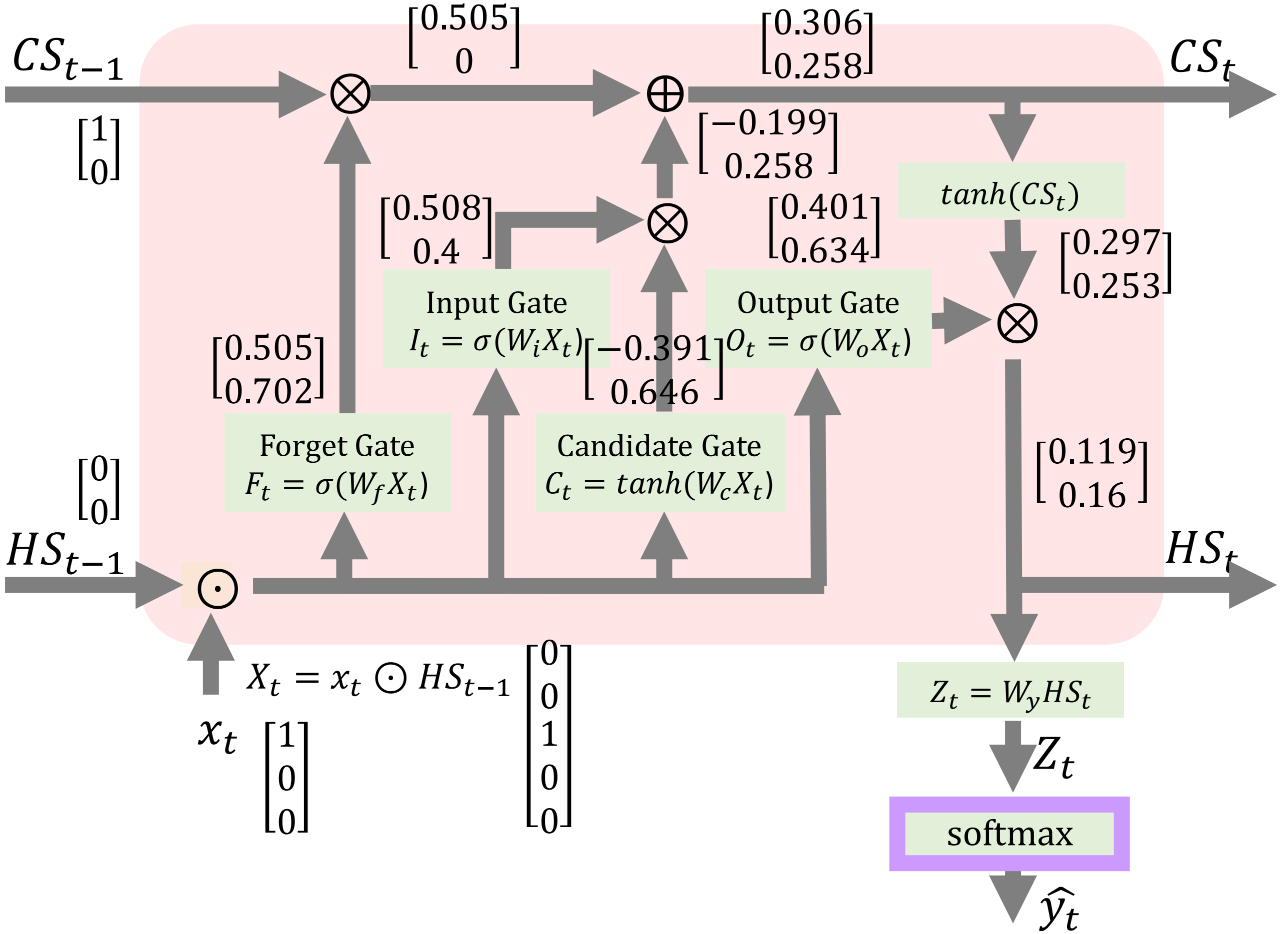
$$W_c = \begin{bmatrix} -0.901 & -0.877 & -0.413 & 0.16 & -0.775 \\ -0.196 & 0.077 & 0.769 & -0.567 & -0.905 \end{bmatrix}$$

$$W_o = \begin{bmatrix} 0.668 & -0.605 & -0.402 & -0.691 & -0.486 \\ 0.613 & 0.875 & 0.549 & -0.623 & 0.262 \end{bmatrix}$$

$$W_y = \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix}$$

$$Z_t = W_y HS_t = \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix} \begin{bmatrix} 0.119 \\ 0.16 \end{bmatrix} = \begin{bmatrix} 0.011 \\ 0.109 \\ 0.168 \end{bmatrix}$$

$$\hat{y}_t = softmax(Z_t) = \begin{bmatrix} 0.305 \\ 0.337 \\ 0.358 \end{bmatrix}$$



이 logit을 softmax함수에 넣어서 loss 계산에 필요한 확률로 바꿉니다

$$W_f = \begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix}$$

$$W_i = \begin{bmatrix} -0.209 & -0.14 & 0.031 & 0.226 & 0.696 \\ 0.101 & -0.435 & -0.406 & -0.796 & 0.324 \end{bmatrix}$$

$$W_c = \begin{bmatrix} -0.901 & -0.877 & -0.413 & 0.16 & -0.775 \\ -0.196 & 0.077 & 0.769 & -0.567 & -0.905 \end{bmatrix}$$

$$W_o = \begin{bmatrix} 0.668 & -0.605 & -0.402 & -0.691 & -0.486 \\ 0.613 & 0.875 & 0.549 & -0.623 & 0.262 \end{bmatrix}$$

$$W_y = \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix}$$

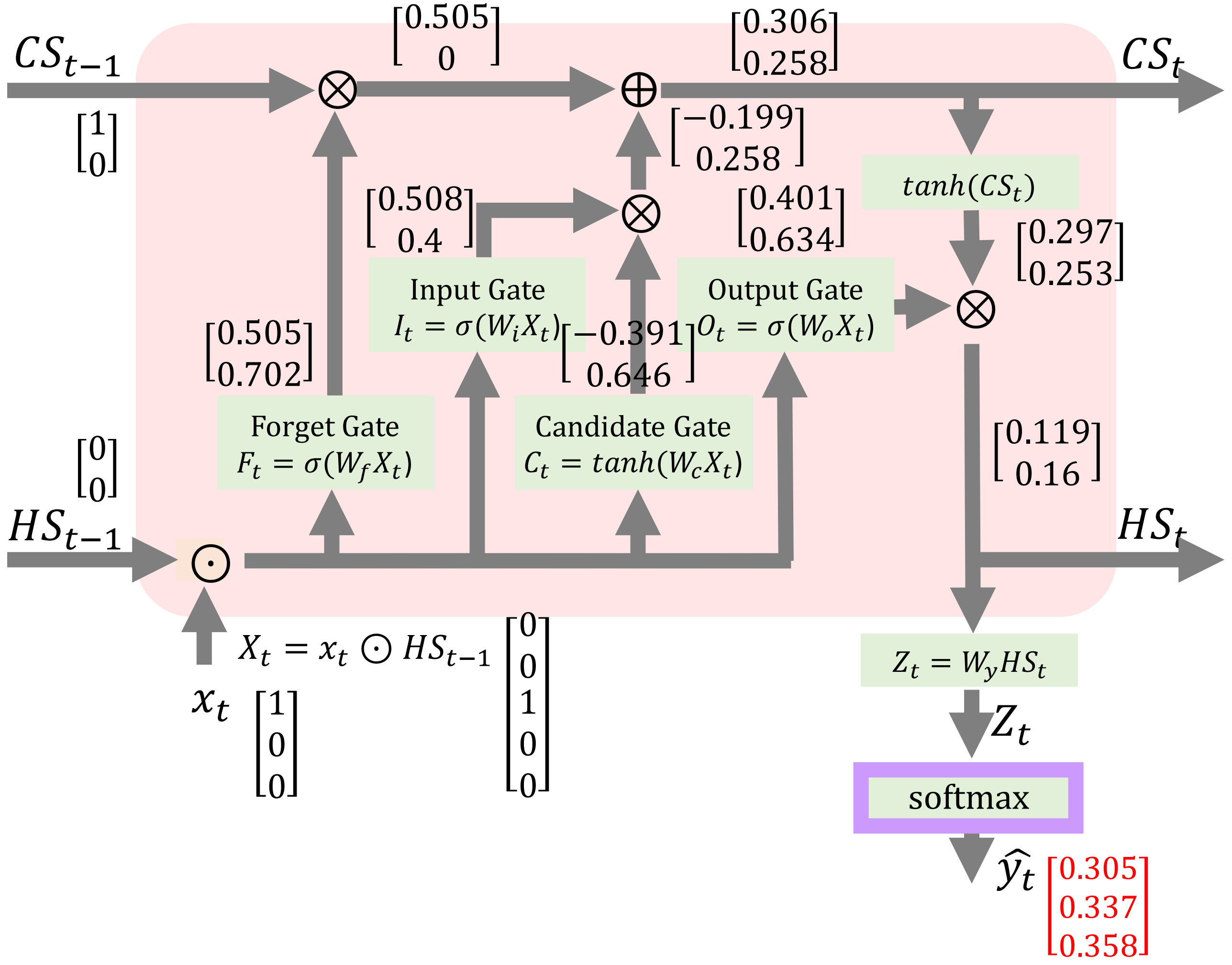
$$Z_t = W_y HS_t$$

$$= \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix} \begin{bmatrix} 0.119 \\ 0.16 \end{bmatrix}$$

$$= \begin{bmatrix} 0.011 \\ 0.109 \\ 0.168 \end{bmatrix}$$

$$\hat{y}_t = softmax(Z_t)$$

$$= \begin{bmatrix} 0.305 \\ 0.337 \\ 0.358 \end{bmatrix}$$



신박AI

여기까지 LSTM의 순전파 feedforward 계산을 알아보았습니다

$$W_f = \begin{bmatrix} 0.813 & -0.487 & 0.02 & -0.778 & 0.418 \\ -0.708 & 0.006 & 0.856 & -0.106 & -0.872 \end{bmatrix}$$

$$W_i = \begin{bmatrix} -0.209 & -0.14 & 0.031 & 0.226 & 0.696 \\ 0.101 & -0.435 & -0.406 & -0.796 & 0.324 \end{bmatrix}$$

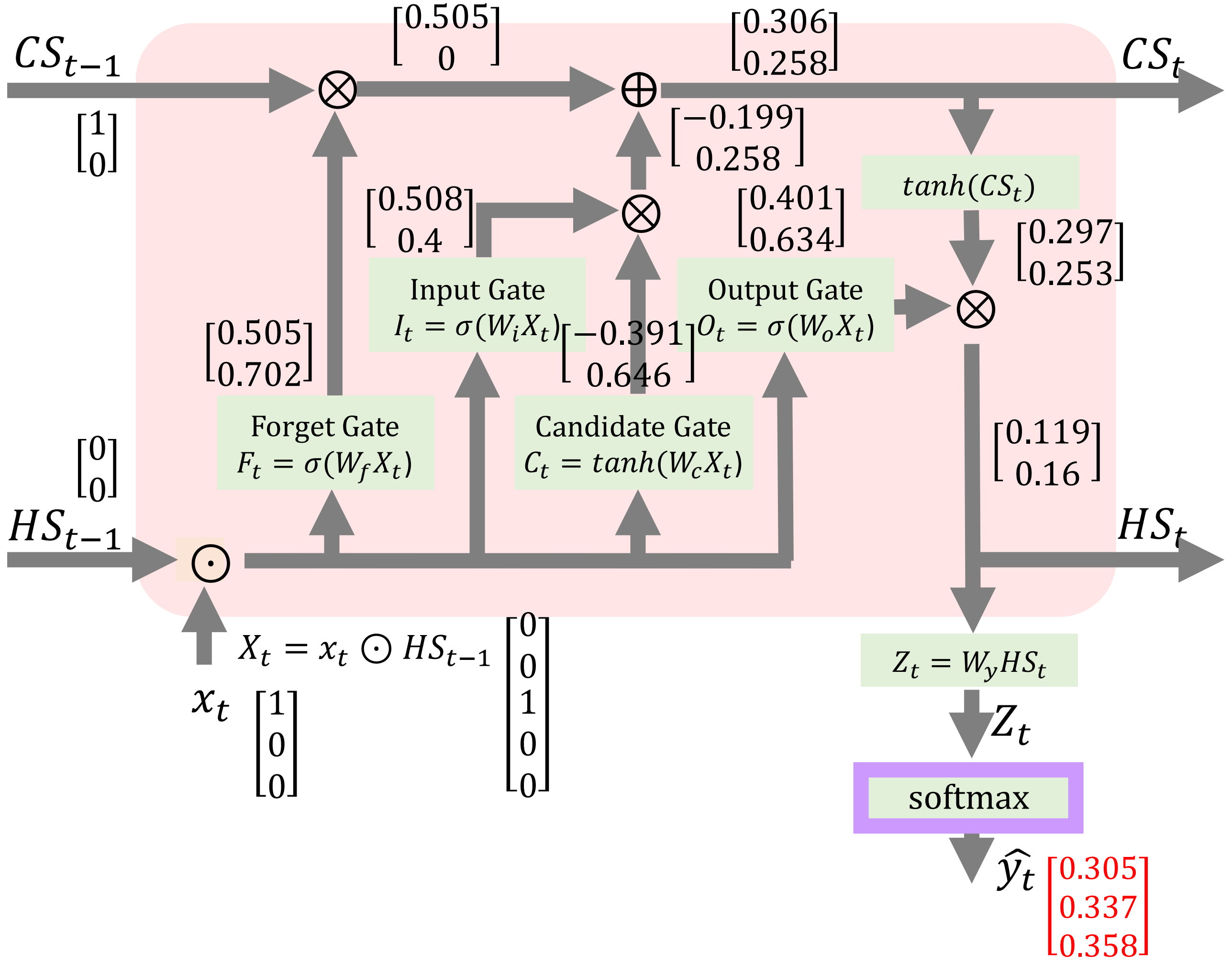
$$W_c = \begin{bmatrix} -0.901 & -0.877 & -0.413 & 0.16 & -0.775 \\ -0.196 & 0.077 & 0.769 & -0.567 & -0.905 \end{bmatrix}$$

$$W_o = \begin{bmatrix} 0.668 & -0.605 & -0.402 & -0.691 & -0.486 \\ 0.613 & 0.875 & 0.549 & -0.623 & 0.262 \end{bmatrix}$$

$$W_y = \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix}$$

$$\begin{aligned} Z_t &= W_y HS_t \\ &= \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix} \begin{bmatrix} 0.119 \\ 0.16 \end{bmatrix} \\ &= \begin{bmatrix} 0.011 \\ 0.109 \\ 0.168 \end{bmatrix} \end{aligned}$$

$$\begin{aligned} \hat{y}_t &= softmax(Z_t) \\ &= \begin{bmatrix} 0.305 \\ 0.337 \\ 0.358 \end{bmatrix} \end{aligned}$$



이제는 LSTM의 시간을 통한 역전파 BPTT를 알아볼 차례입니다

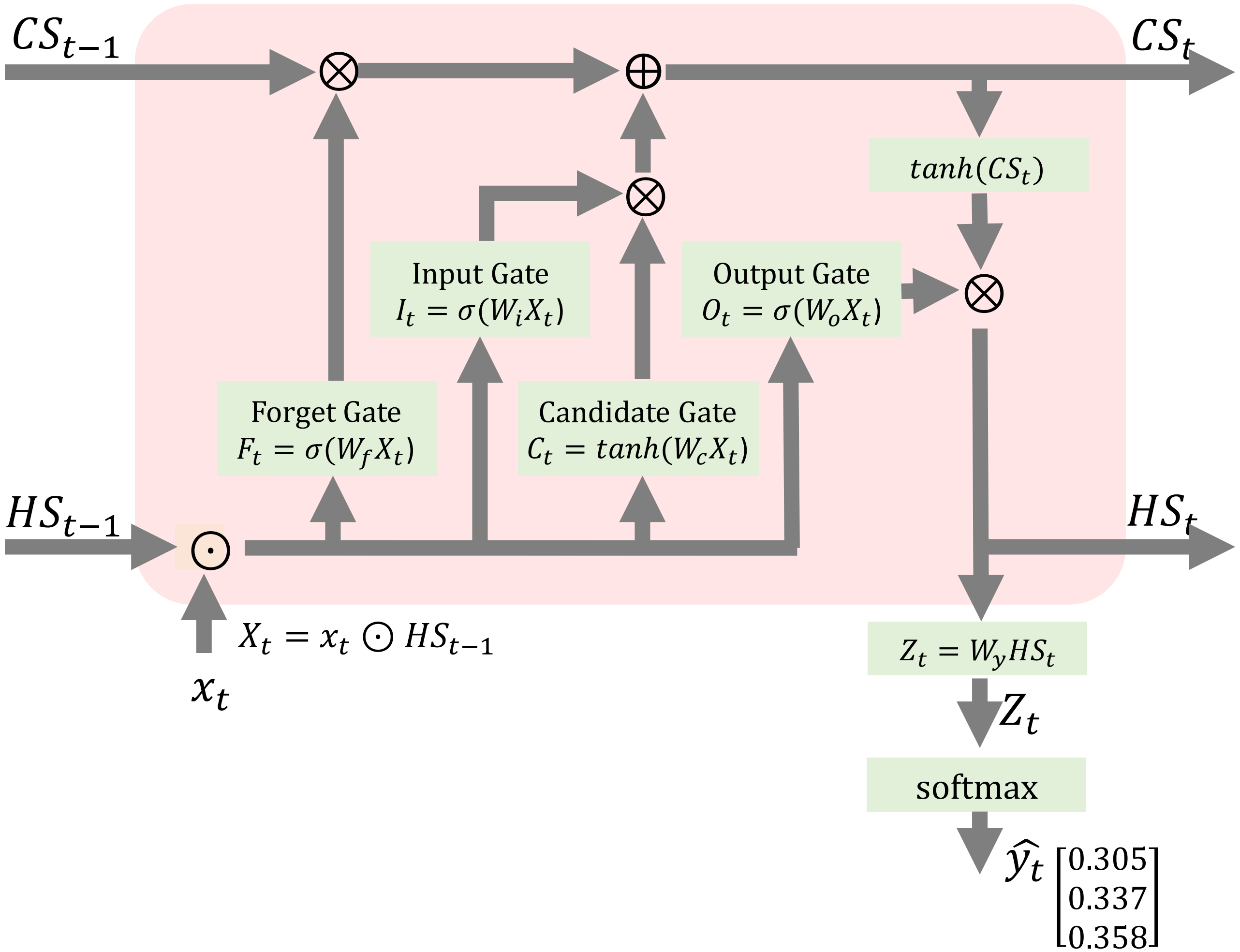
Forget Gate: $F_t = \sigma(W_f X_t)$

Input Gate: $I_t = \sigma(W_i X_t)$

Candidate Gate: $C_t = tanh(W_c X_t)$

Output Gate: $O_t = \sigma(W_o X_t)$

$Z_t = W_y HS_t$



LSTM의 학습이라는 것도 결국 역전파와 경사하강법을 통하여 가중치를 업데이트 하는 것입니다

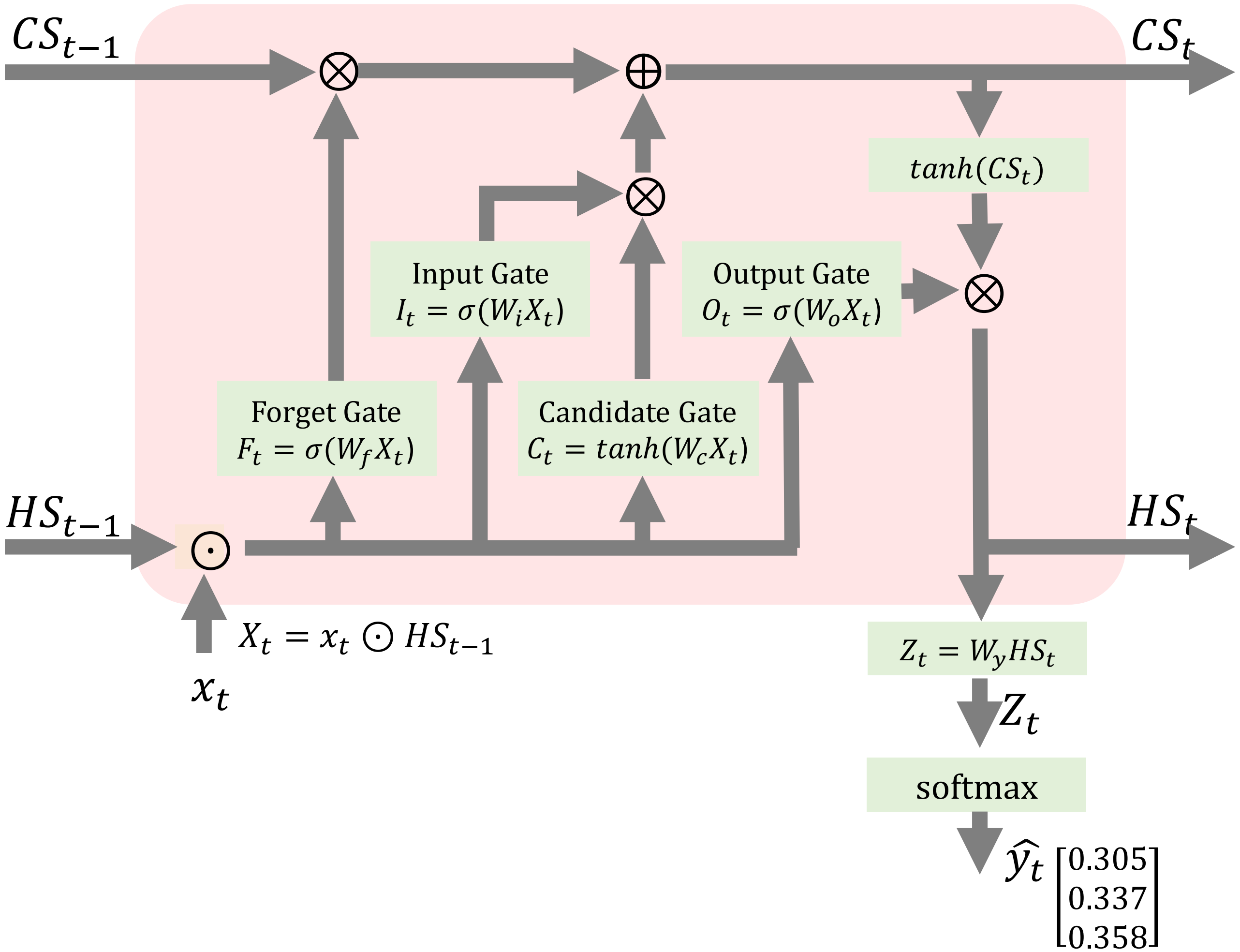
Forget Gate: $F_t = \sigma(W_f X_t)$

Input Gate: $I_t = \sigma(W_i X_t)$

Candidate Gate: $C_t = tanh(W_c X_t)$

Output Gate: $O_t = \sigma(W_o X_t)$

$Z_t = W_y HS_t$



신박AI

LSTM의 가중치는 다음 다섯개가 존재합니다

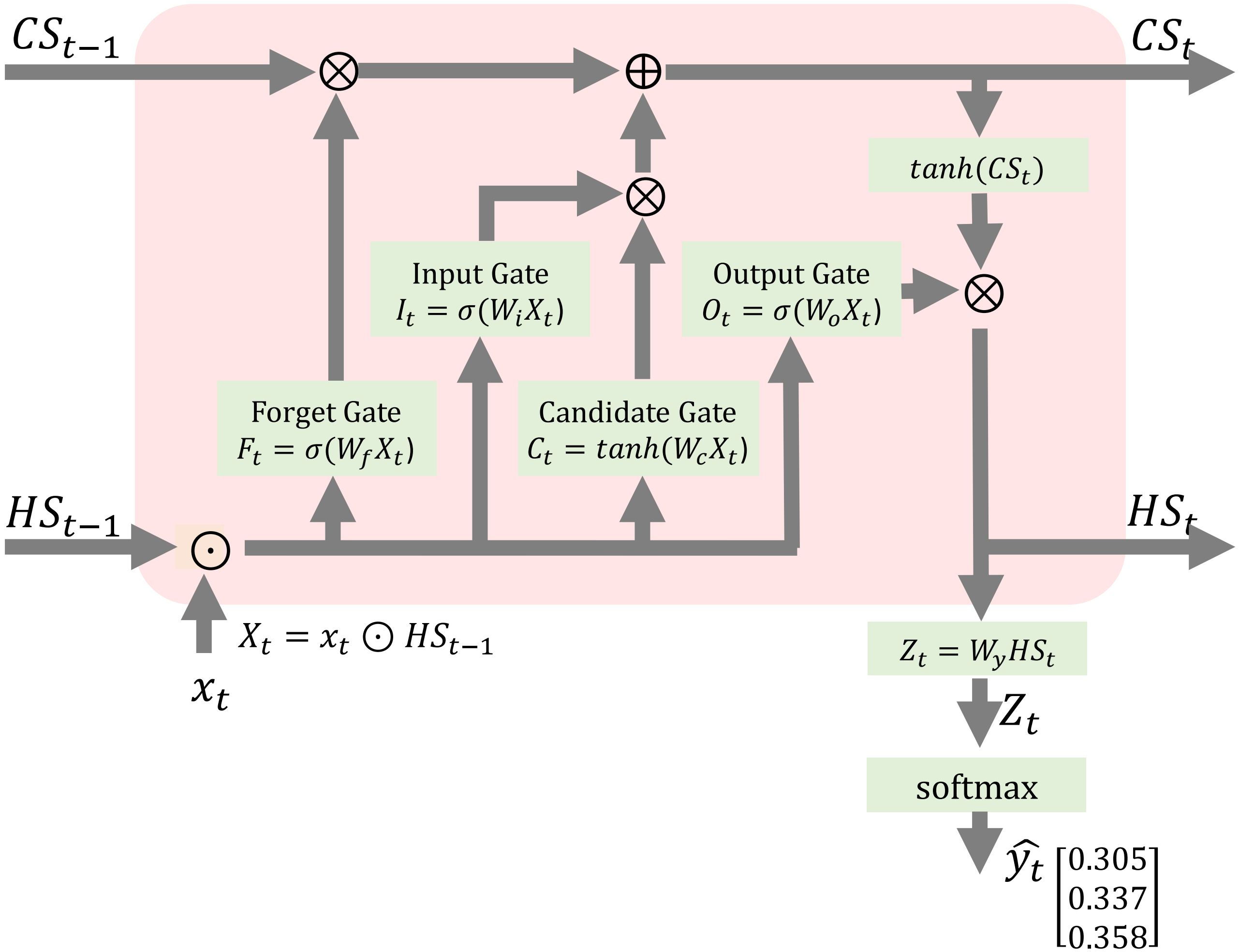
Forget Gate: $F_t = \sigma(W_f X_t)$

Input Gate: $I_t = \sigma(W_i X_t)$

Candidate Gate: $C_t = tanh(W_c X_t)$

Output Gate: $O_t = \sigma(W_o X_t)$

$Z_t = W_y HS_t$



신박AI

그러므로, 역전파는 손실함수에 대한 각각의 미분값을 구하면 되는데,

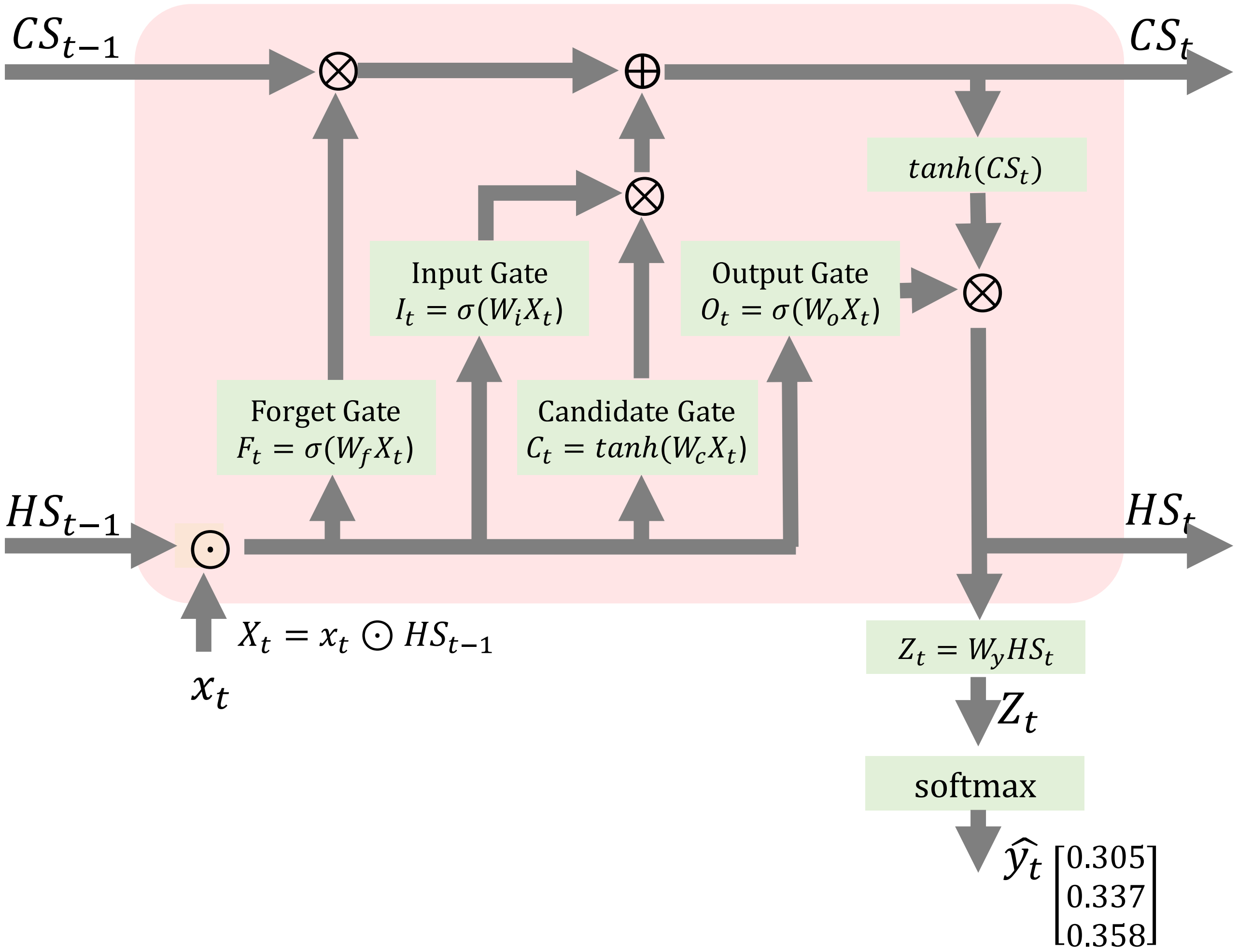
Forget Gate: $F_t = \sigma(W_f X_t)$

Input Gate: $I_t = \sigma(W_i X_t)$

Candidate Gate: $C_t = tanh(W_c X_t)$

Output Gate: $O_t = \sigma(W_o X_t)$

$Z_t = W_y HS_t$



신박AI

그러므로, 역전파는 손실함수에 대한 각각의 미분값을 구하면 되는데,

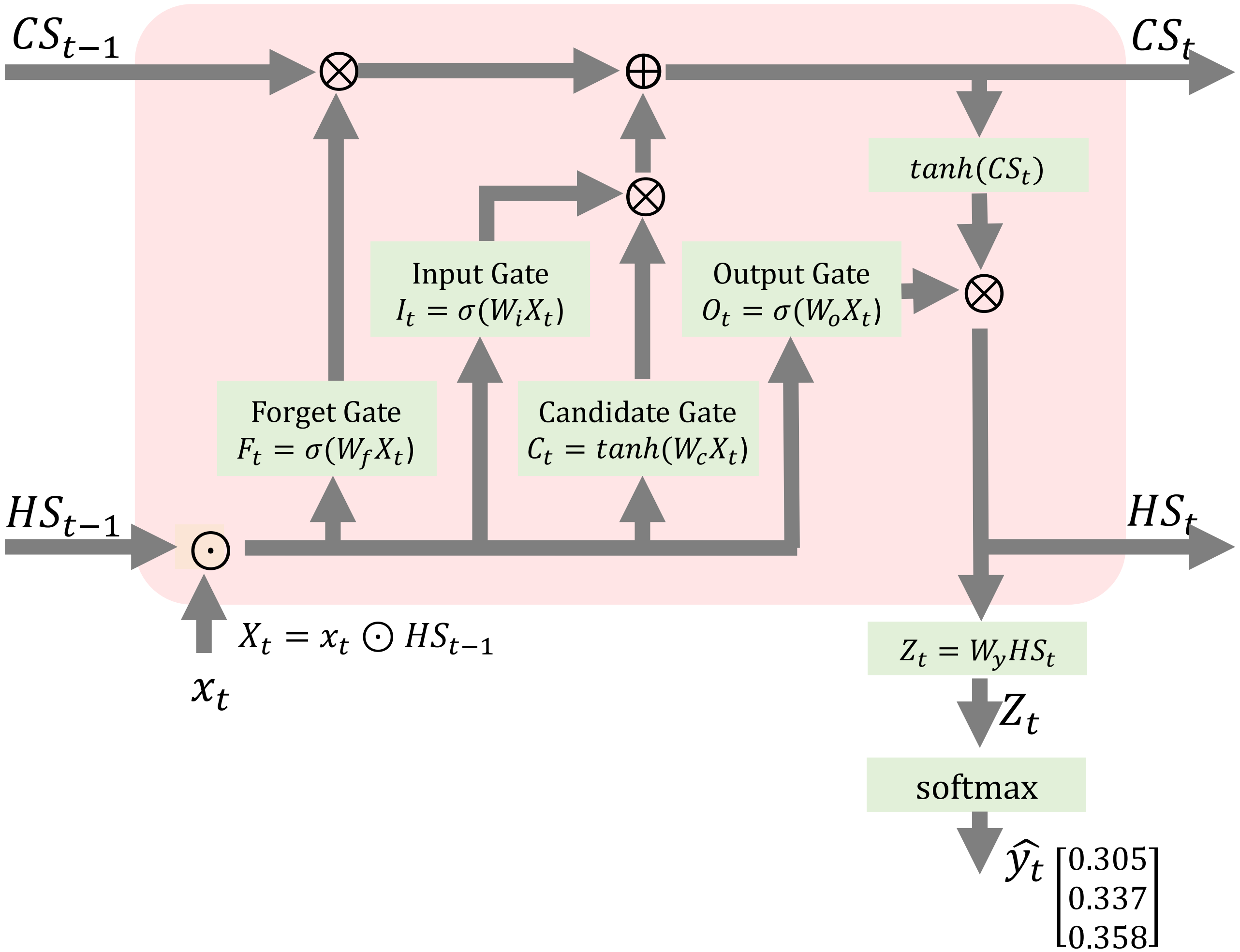
Forget Gate: $F_t = \sigma(W_f X_t)$ $\Longrightarrow$ $\frac{\partial L}{\partial W_f}$

Input Gate: $I_t = \sigma(W_i X_t)$

Candidate Gate: $C_t = tanh(W_c X_t)$

Output Gate: $O_t = \sigma(W_o X_t)$

$Z_t = W_y HS_t$



그러므로, 역전파는 손실함수에 대한 각각의 미분값을 구하면 되는데,

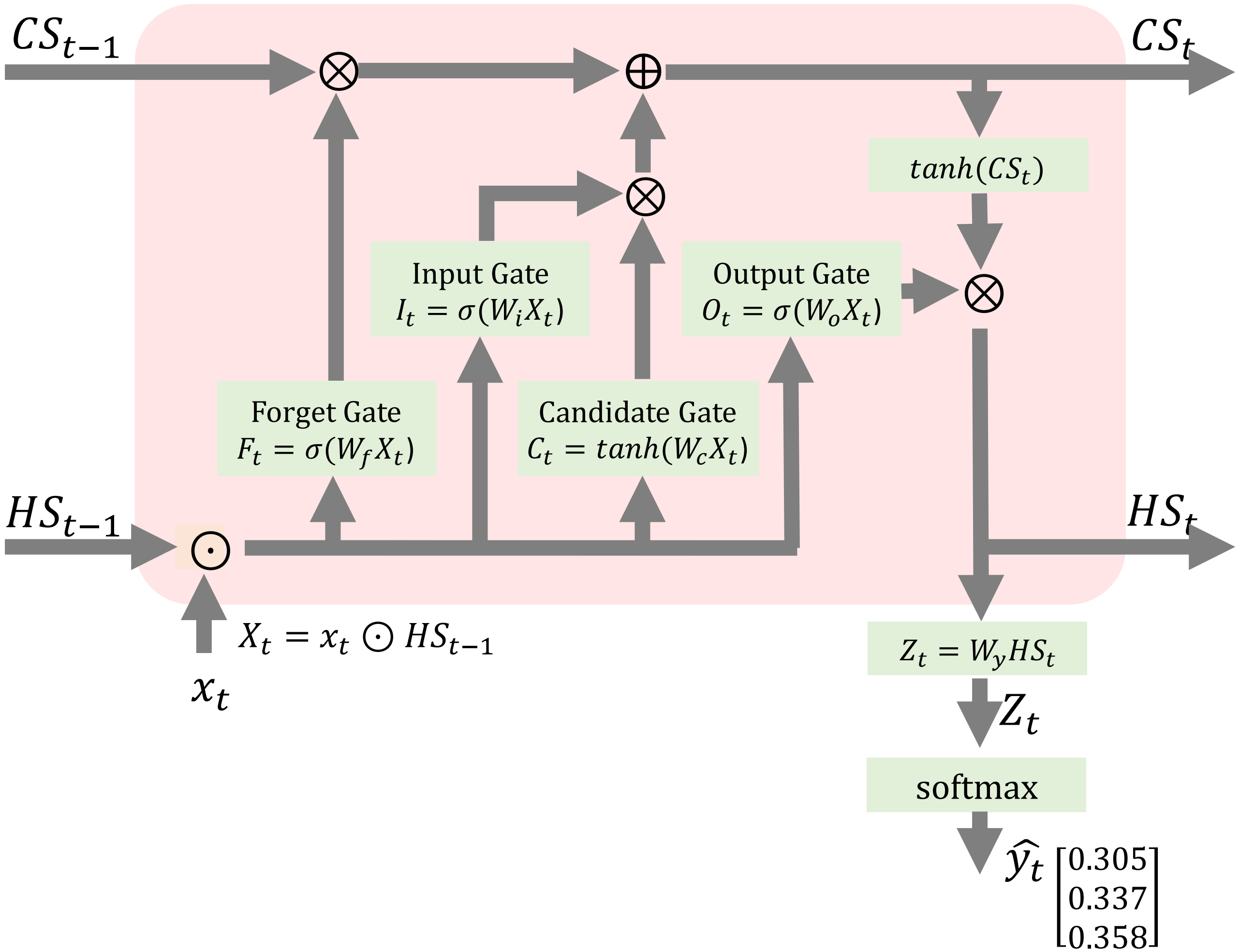
Forget Gate: $F_t = \sigma(W_f X_t)$ ⟹ $\dfrac{\partial L}{\partial W_f}$

Input Gate: $I_t = \sigma(W_i X_t)$ ⟹ $\dfrac{\partial L}{\partial W_i}$

Candidate Gate: $C_t = tanh(W_c X_t)$

Output Gate: $O_t = \sigma(W_o X_t)$

$Z_t = W_y HS_t$



그러므로, 역전파는 손실함수에 대한 각각의 미분값을 구하면 되는데,

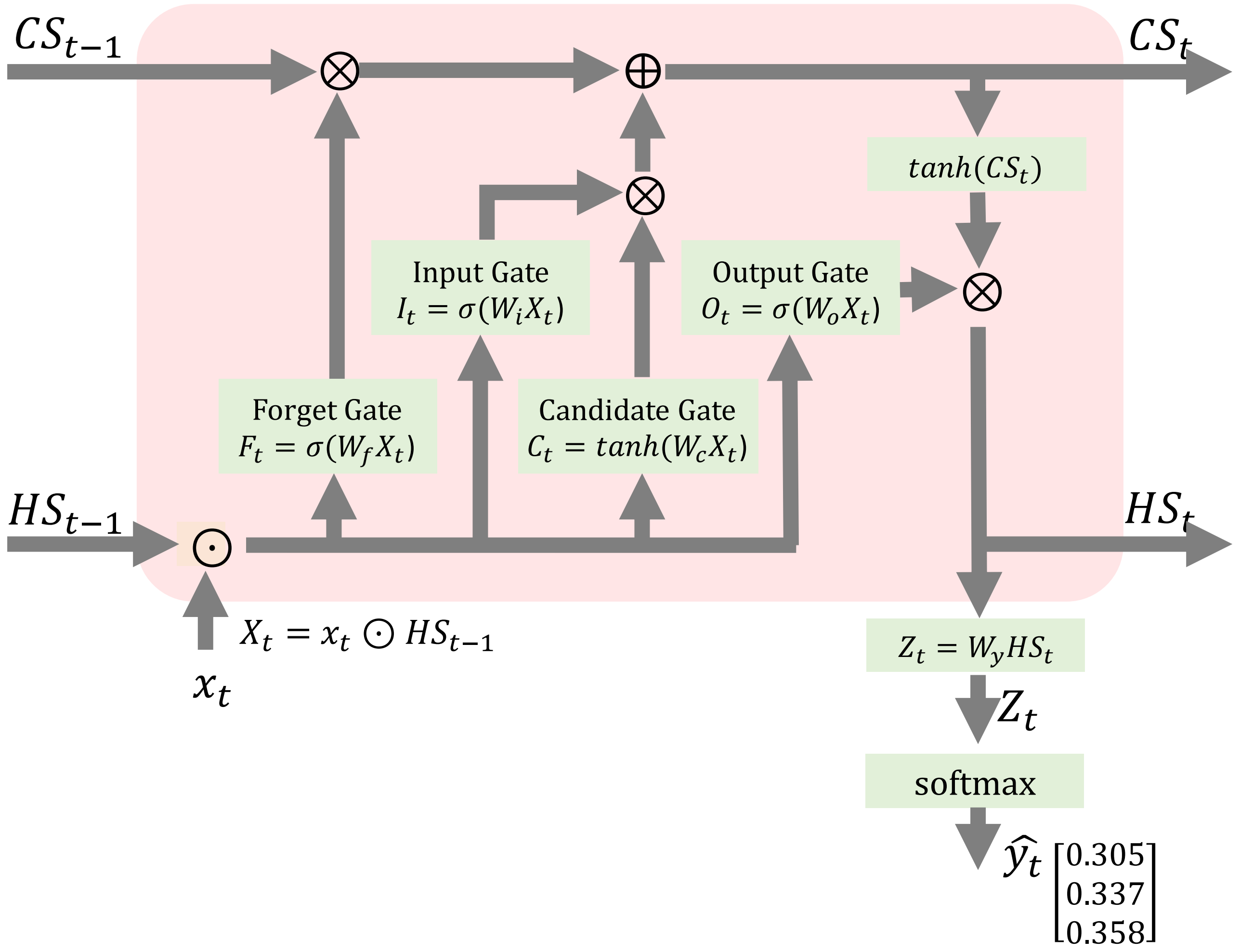
Forget Gate: $F_t = \sigma(W_f X_t)$ $\Longrightarrow$ $\dfrac{\partial L}{\partial W_f}$

Input Gate: $I_t = \sigma(W_i X_t)$ $\Longrightarrow$ $\dfrac{\partial L}{\partial W_i}$

Candidate Gate: $C_t = tanh(W_c X_t)$ $\Longrightarrow$ $\dfrac{\partial L}{\partial W_c}$

Output Gate: $O_t = \sigma(W_o X_t)$

$Z_t = W_y HS_t$



그러므로, 역전파는 손실함수에 대한 각각의 미분값을 구하면 되는데,

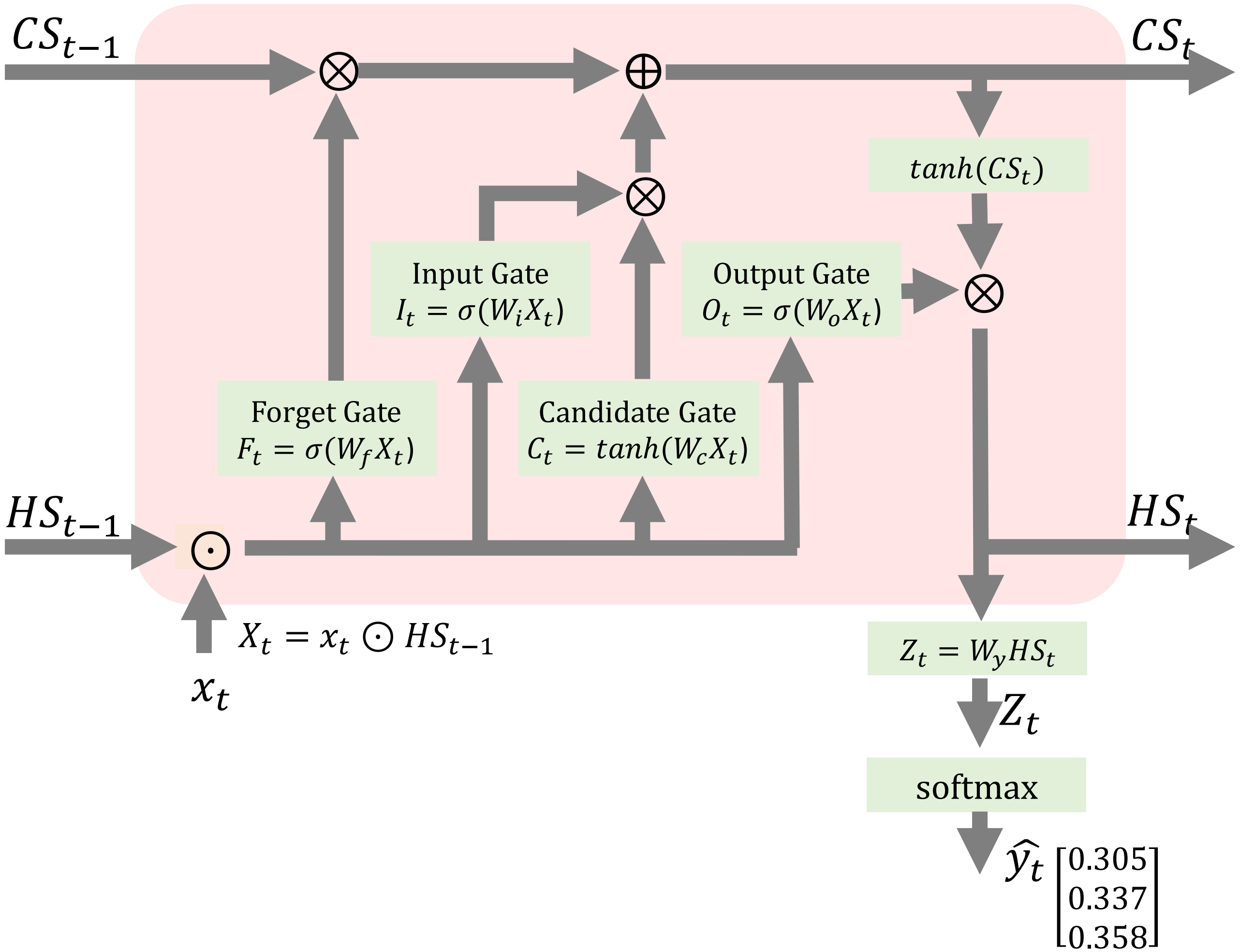
Forget Gate: $F_t = \sigma(W_f X_t)$ ⟹ $\frac{\partial L}{\partial W_f}$

Input Gate: $I_t = \sigma(W_i X_t)$ ⟹ $\frac{\partial L}{\partial W_i}$

Candidate Gate: $C_t = tanh(W_c X_t)$ ⟹ $\frac{\partial L}{\partial W_c}$

Output Gate: $O_t = \sigma(W_o X_t)$ ⟹ $\frac{\partial L}{\partial W_o}$

$Z_t = W_y HS_t$



그러므로, 역전파는 손실함수에 대한 각각의 미분값을 구하면 되는데,

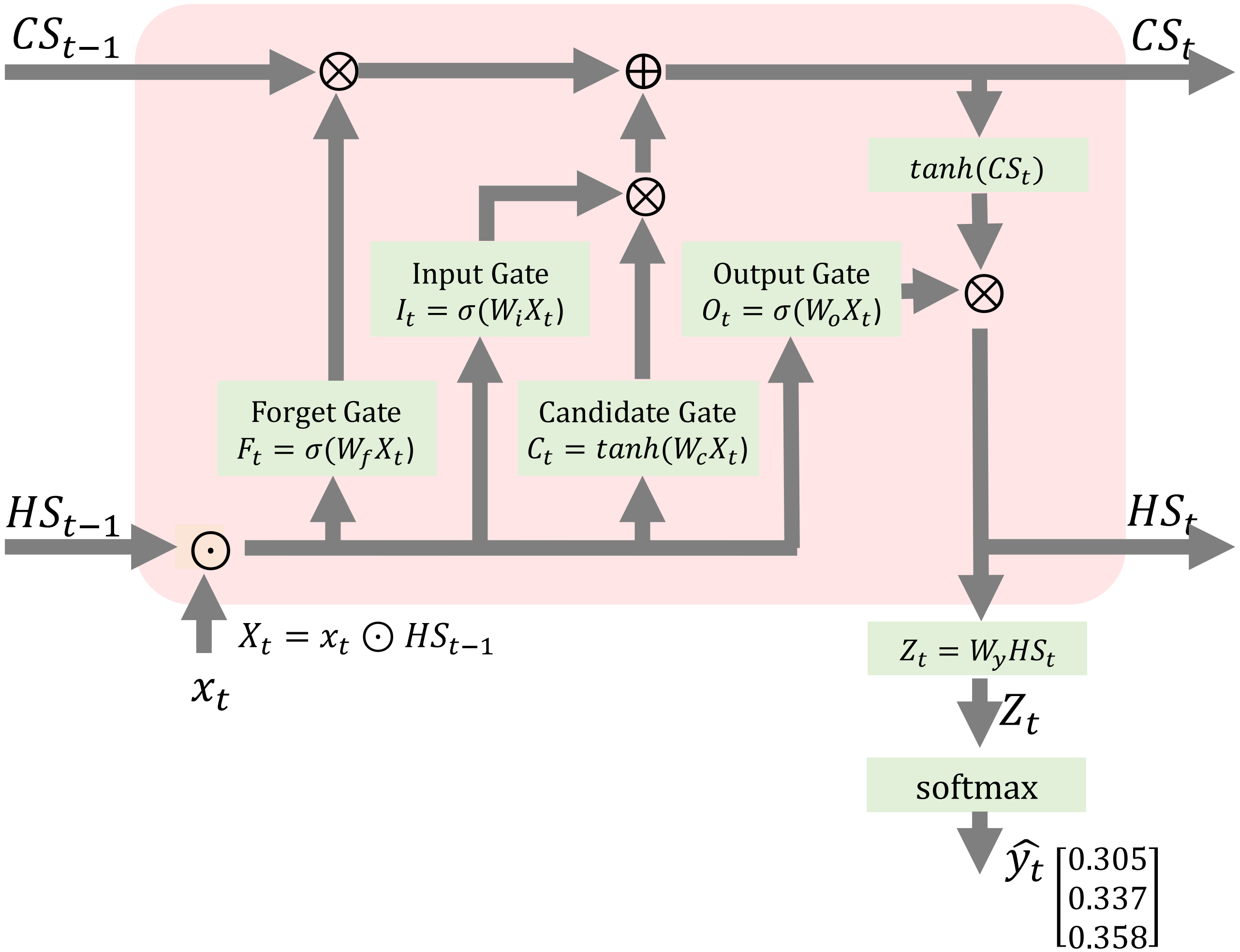
Forget Gate: $F_t = \sigma(W_f X_t)$ ⟹ $\dfrac{\partial L}{\partial W_f}$

Input Gate: $I_t = \sigma(W_i X_t)$ ⟹ $\dfrac{\partial L}{\partial W_i}$

Candidate Gate: $C_t = tanh(W_c X_t)$ ⟹ $\dfrac{\partial L}{\partial W_c}$

Output Gate: $O_t = \sigma(W_o X_t)$ ⟹ $\dfrac{\partial L}{\partial W_o}$

$Z_t = W_y HS_t$ ⟹ $\dfrac{\partial L}{\partial W_y}$



각각의 미분값을 구하는 방법은 역시나 체인룰입니다

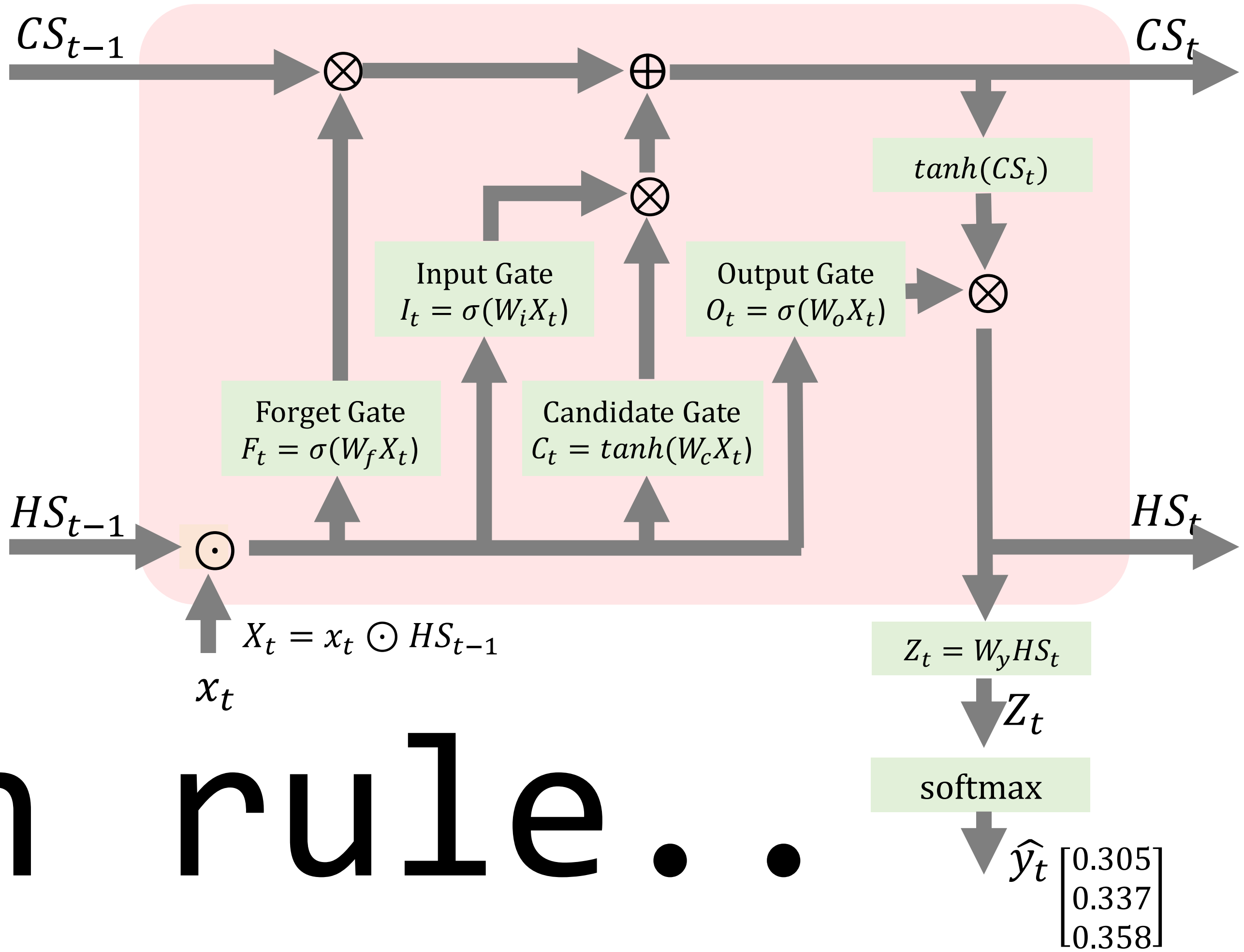
Forget Gate: $F_t = \sigma(W_f X_t)$ $\Rightarrow$ $\dfrac{\partial L}{\partial W_f}$

Input Gate: $I_t = \sigma(W_i X_t)$ $\Rightarrow$ $\dfrac{\partial L}{\partial W_i}$

Candidate Gate: $C_t = tanh(W_c X_t)$ $\Rightarrow$ $\dfrac{\partial L}{\partial W_c}$

Output Gate: $O_t = \sigma(W_o X_t)$ $\Rightarrow$ $\dfrac{\partial L}{\partial W_o}$

$Z_t = W_y HS_t$ $\Rightarrow$ $\dfrac{\partial L}{\partial W_y}$



Chain rule..

신박AI

역전파를 쉽게 계산하기 위해 LSTM 를 세분화 하여 그려보겠습니다

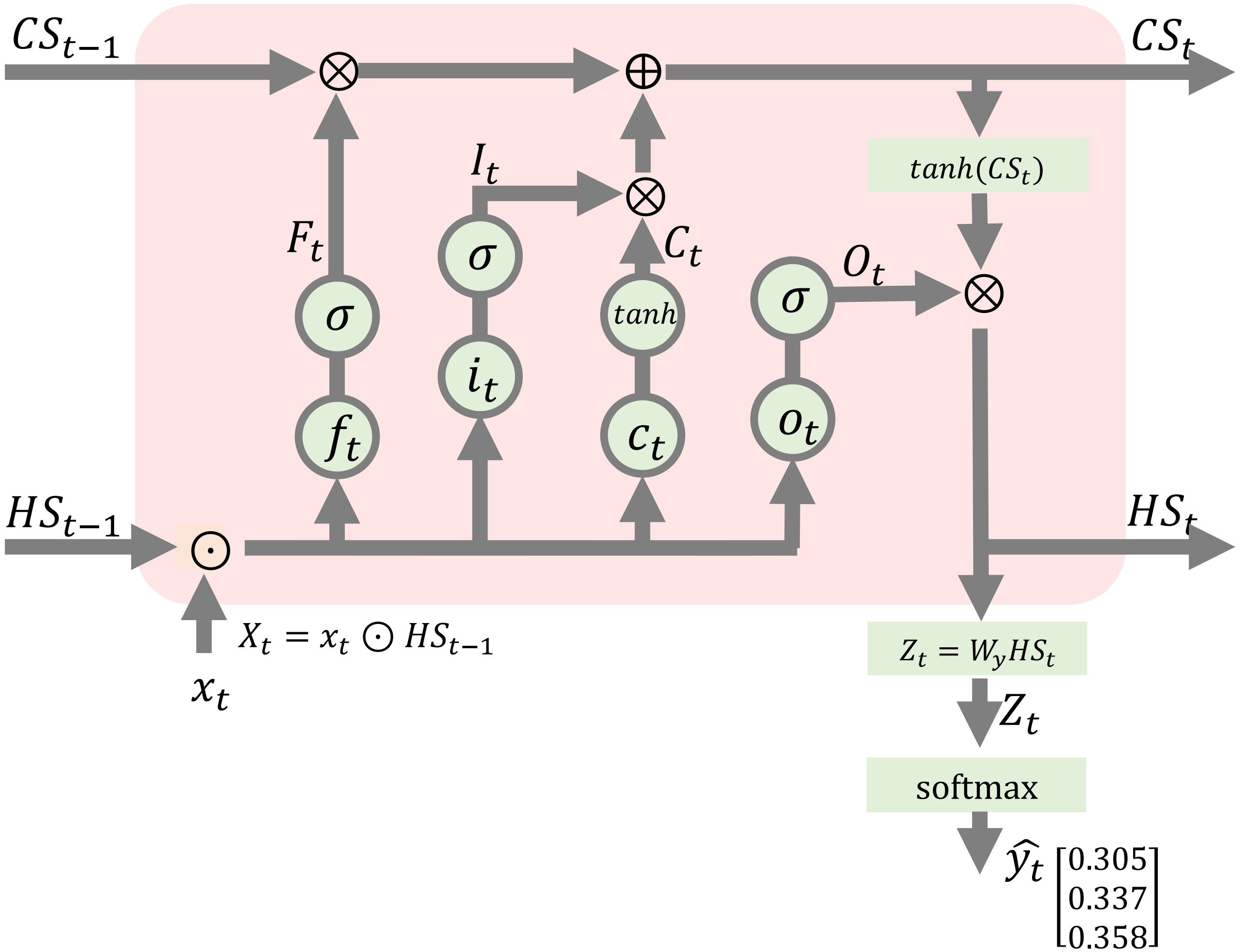
Forget Gate: $F_t = \sigma(W_f X_t)$

Input Gate: $I_t = \sigma(W_i X_t)$

Candidate Gate: $C_t = tanh(W_c X_t)$

Output Gate: $O_t = \sigma(W_o X_t)$

$Z_t = W_y HS_t$



그리고 각각의 게이트 식을 각각 두개의 식으로 나누어 표현하였습니다

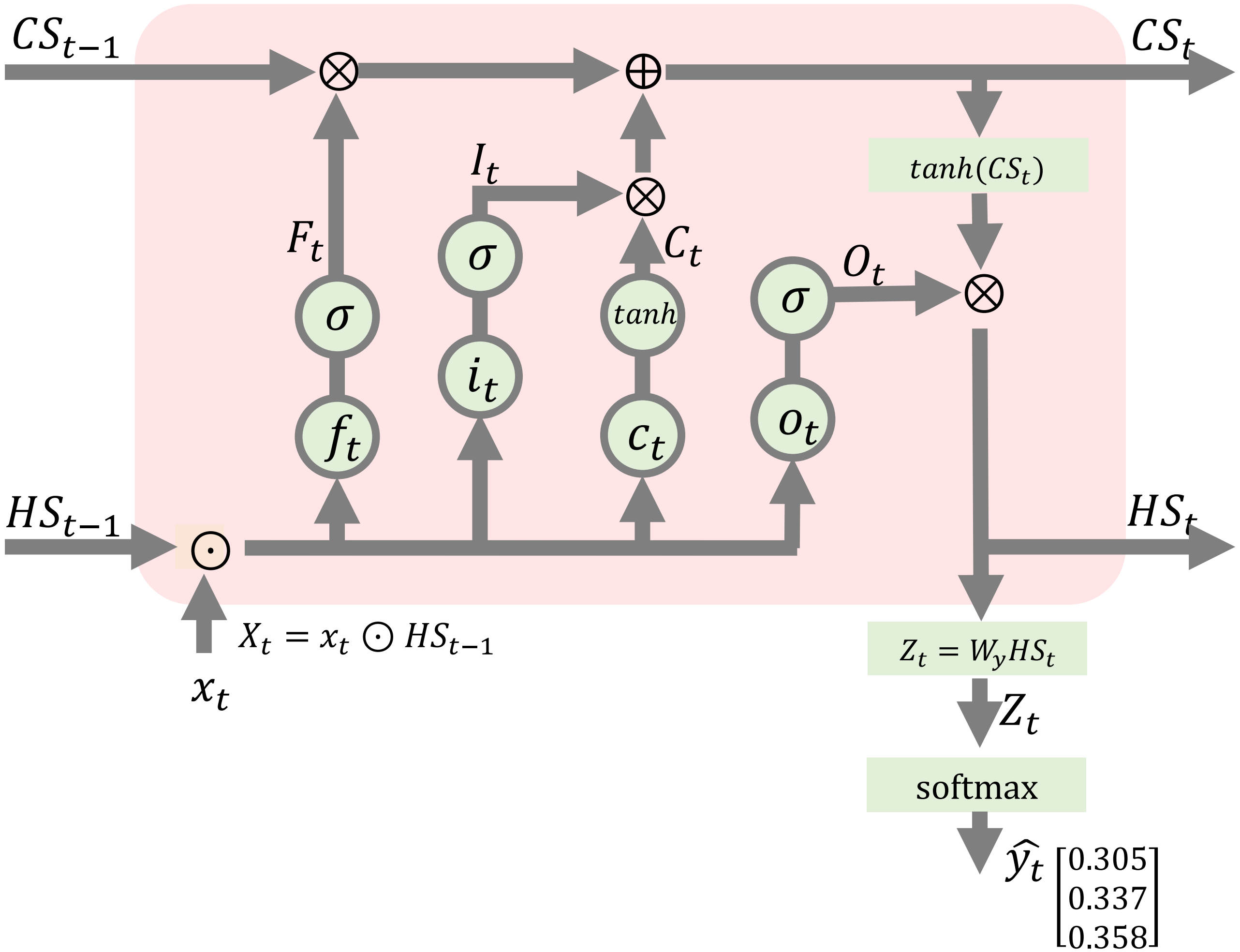
Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$



신박AI

계산 공간 확보를 위해서 식을 재배열하도록 하겠습니다

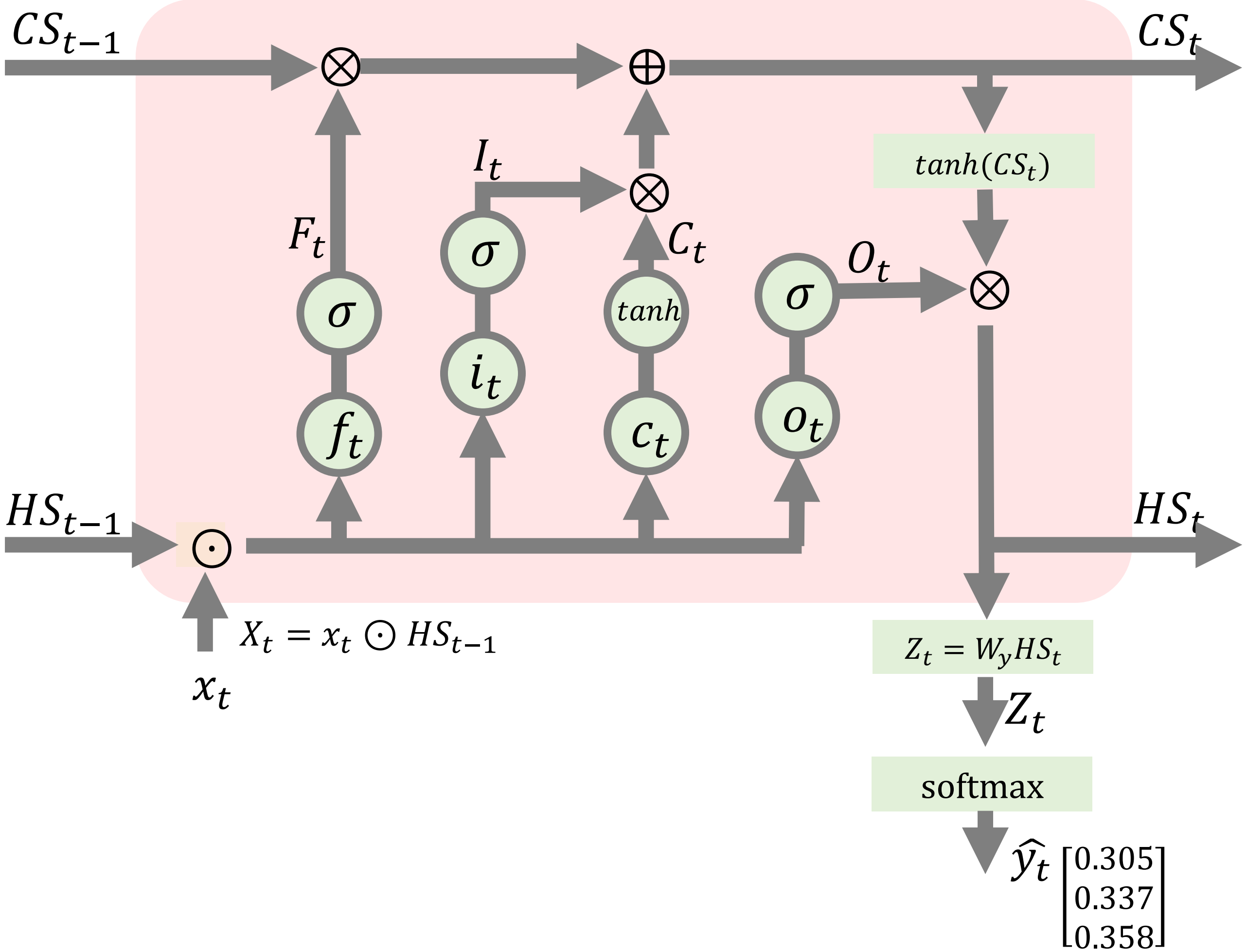
Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$



자 그러면 이제 역전파를 계산하기 위한 준비는 다 되었습니다

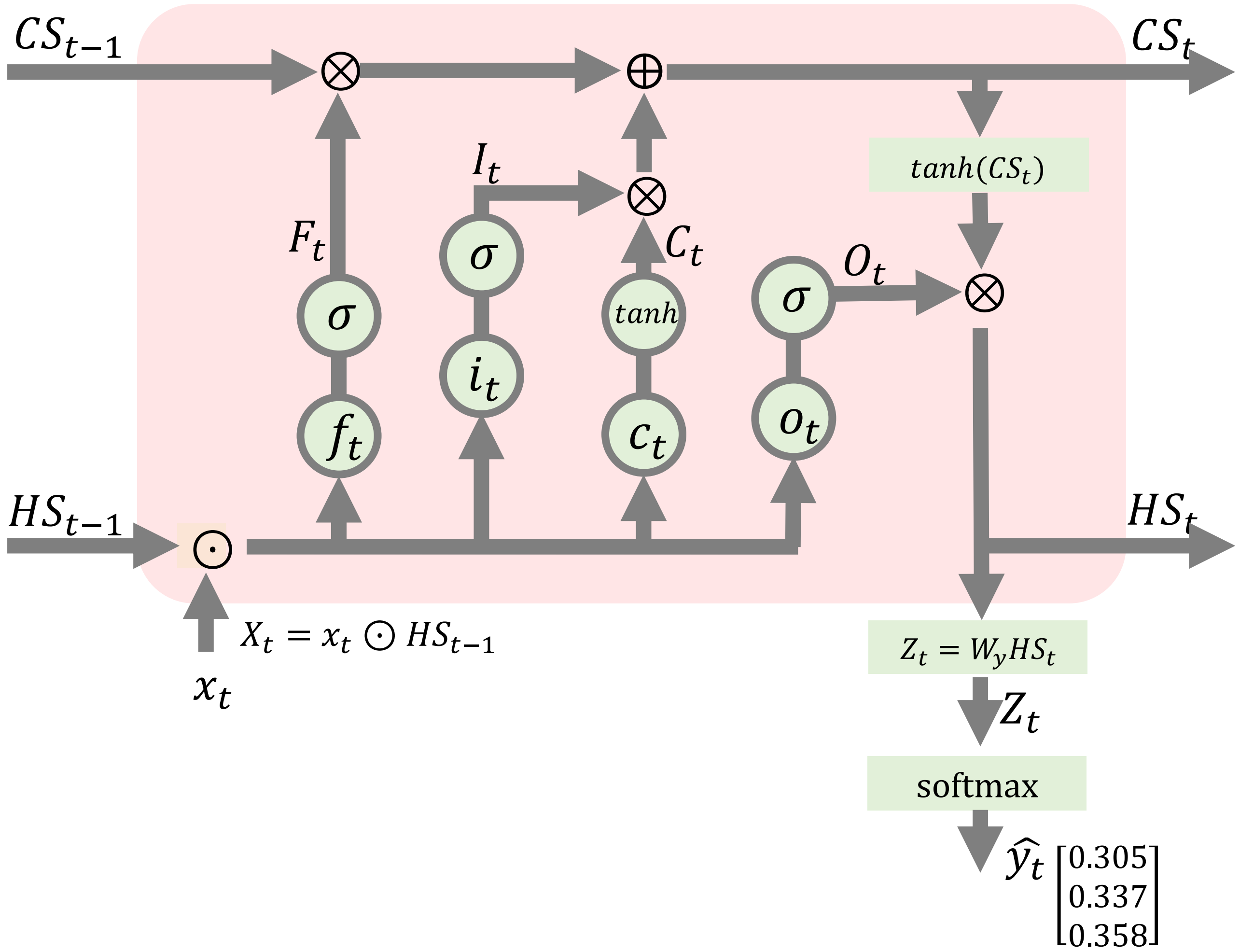
Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$



역전파는 먼저 순전파의 오차 error를 구해야 합니다

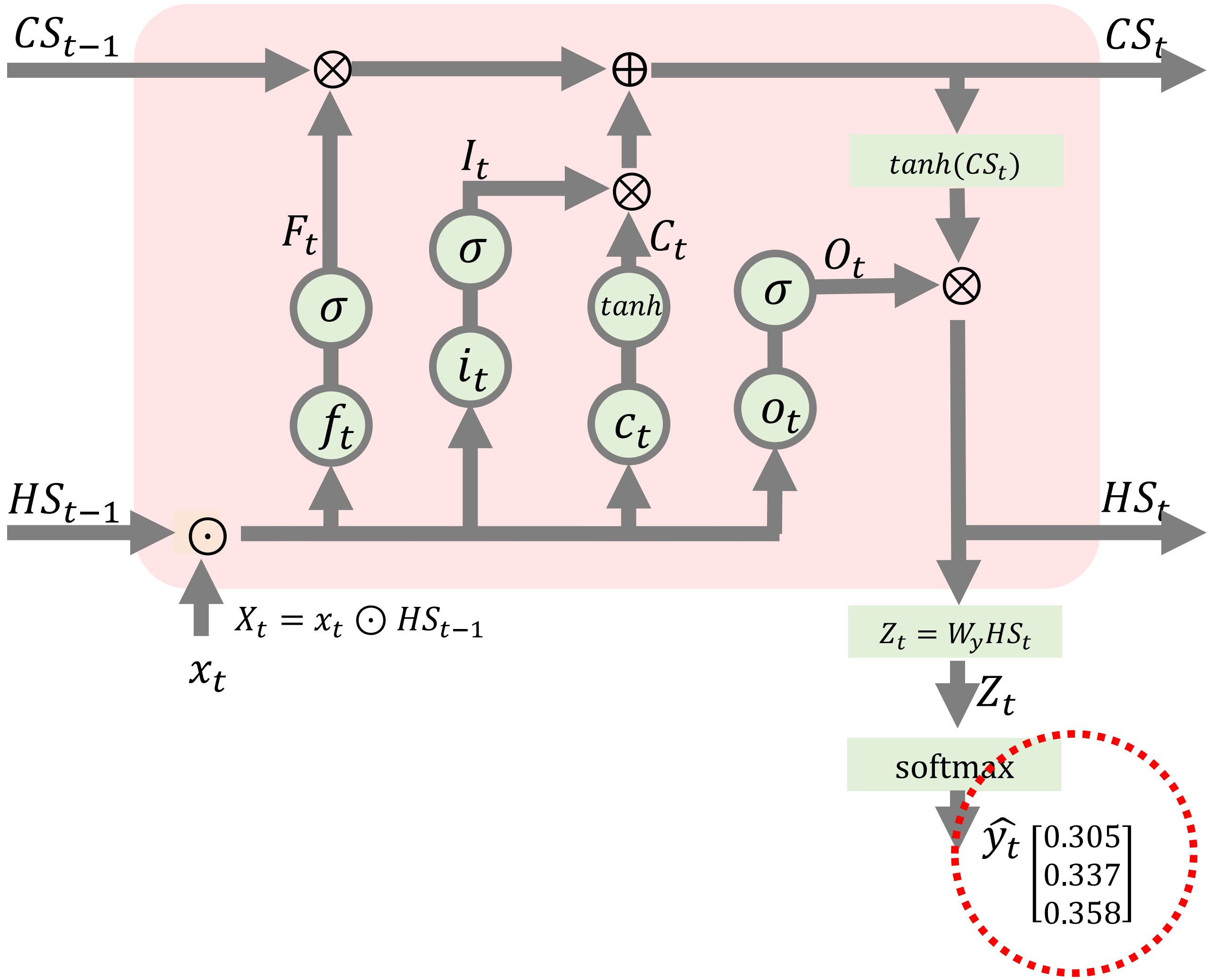
Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$



입력 [1, 0, 0]에 대한 출력값 y를 [0,1,0]이라 가정했을 때

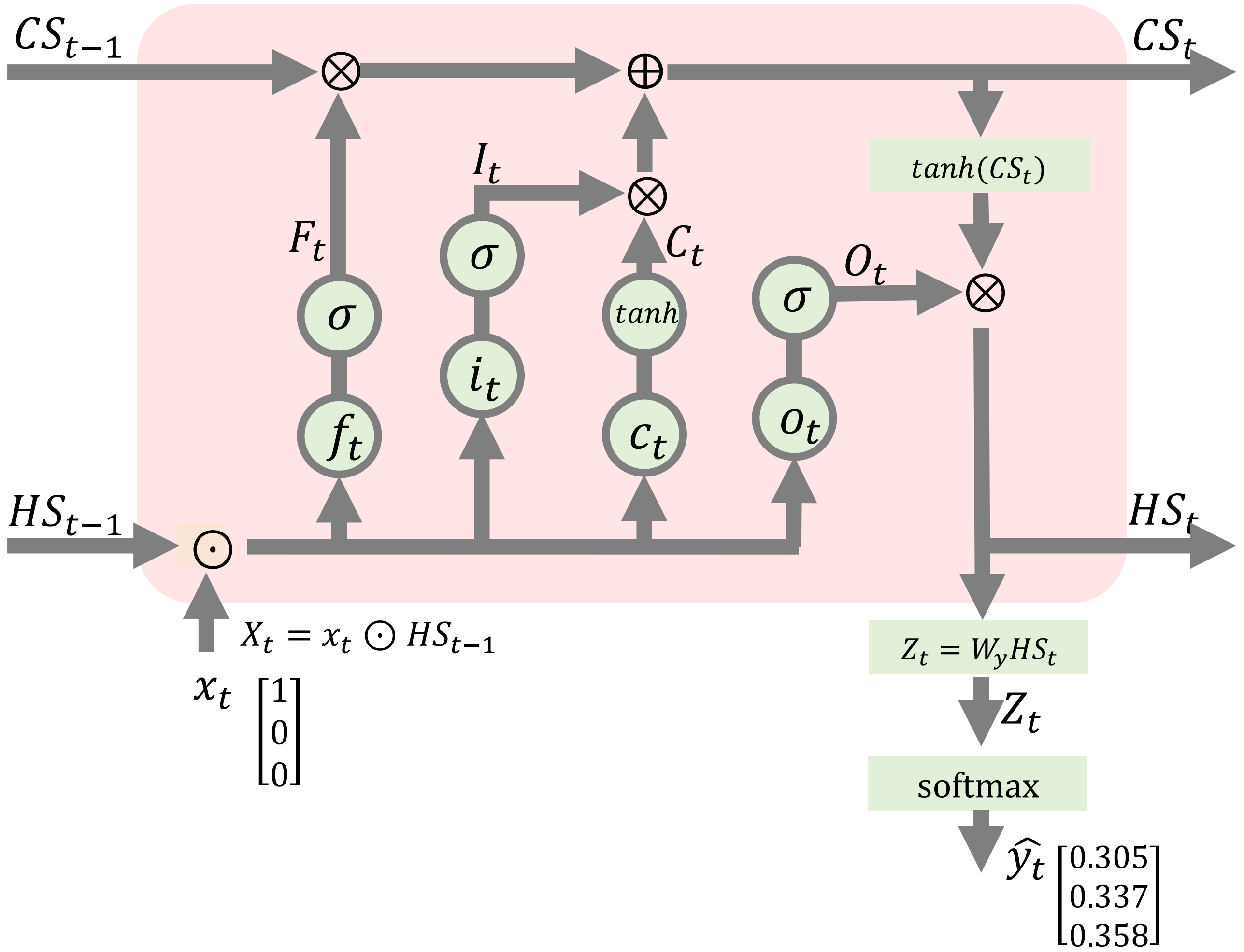
Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$



$$y = \begin{bmatrix}0\\1\\0\end{bmatrix}$$

신박AI

역전파 계산에 필요한 오차는 $\hat{y}_t - y$ 입니다

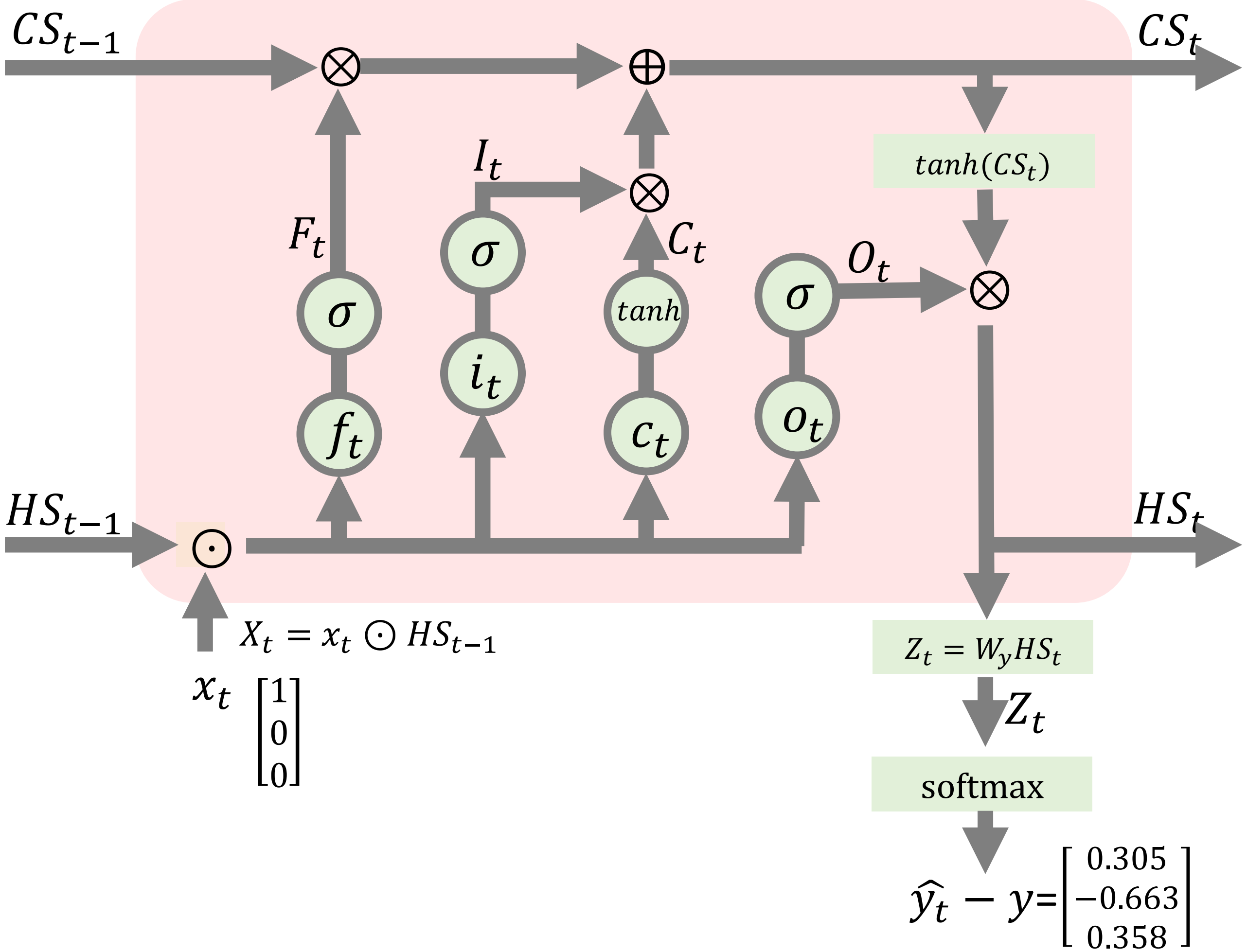
Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$



신박AI

왜냐하면 지난 RNN에서도 말씀드렸듯이, softmax와 cross-entropy를 사용할 경우,

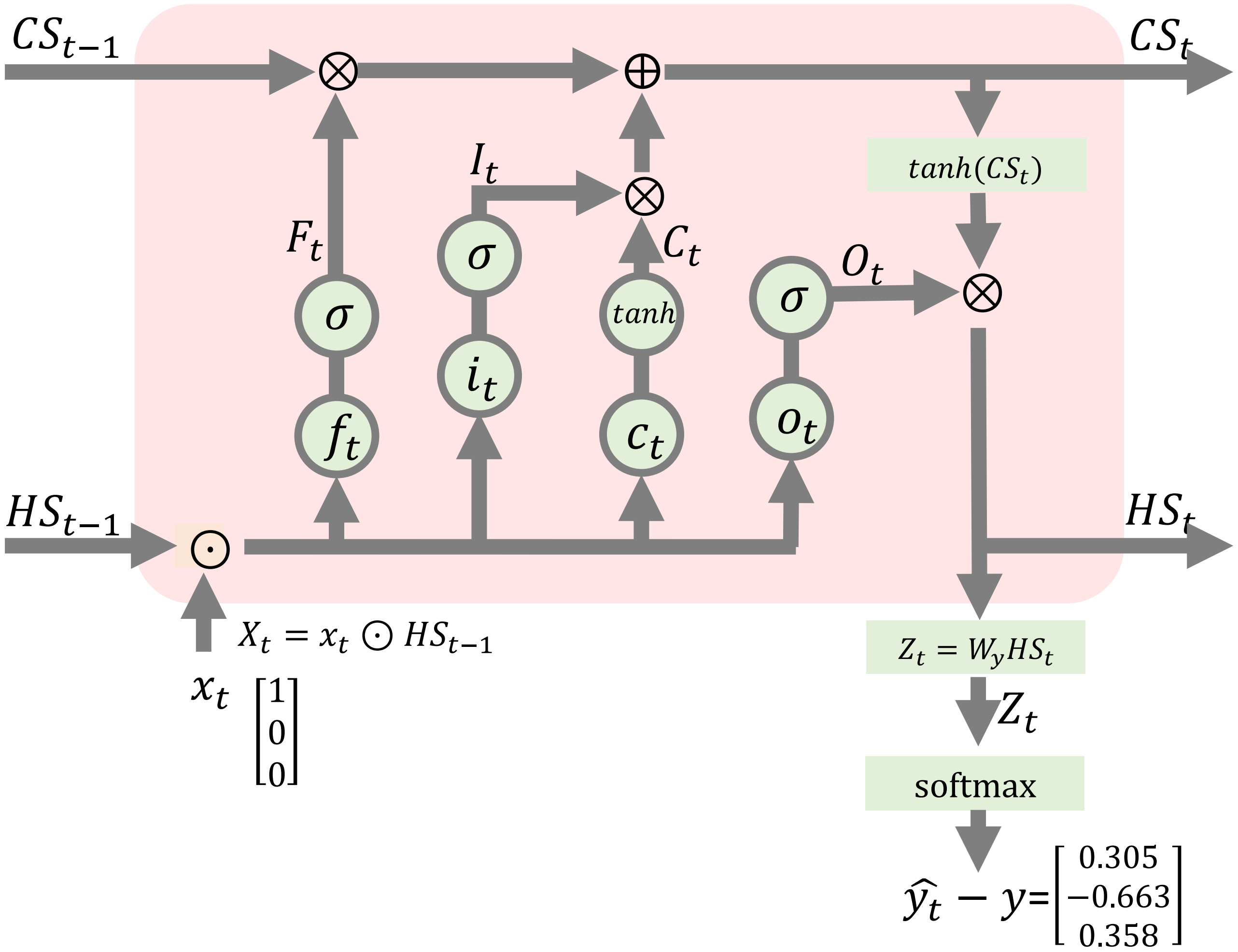
Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$



$\hat{y}_t - y$는 $\partial L/\partial Z_t$ 가 됩니다

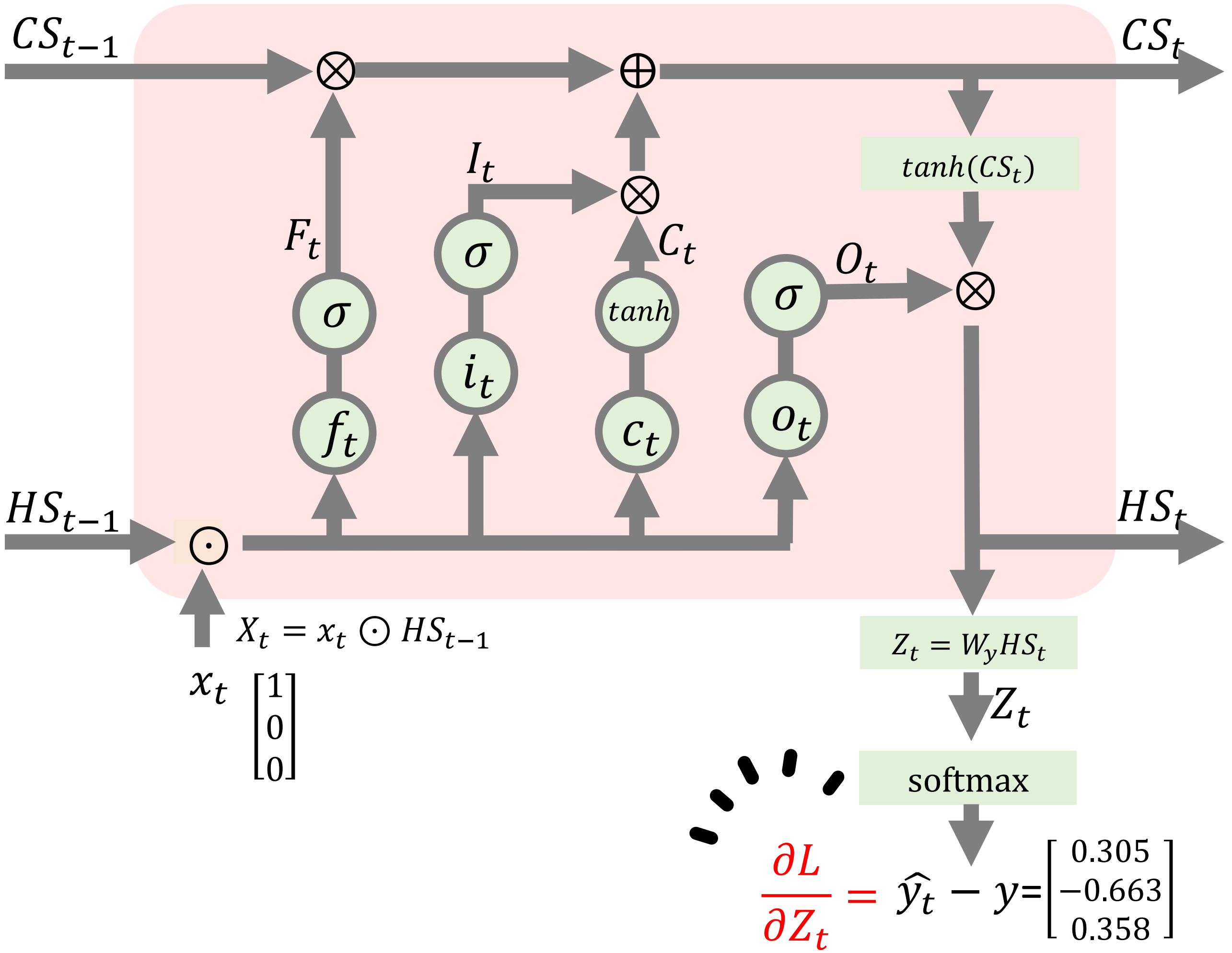
Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$



이 사실을 바탕으로 먼저 $\partial L/\partial W_y$을 구해보도록 하겠습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

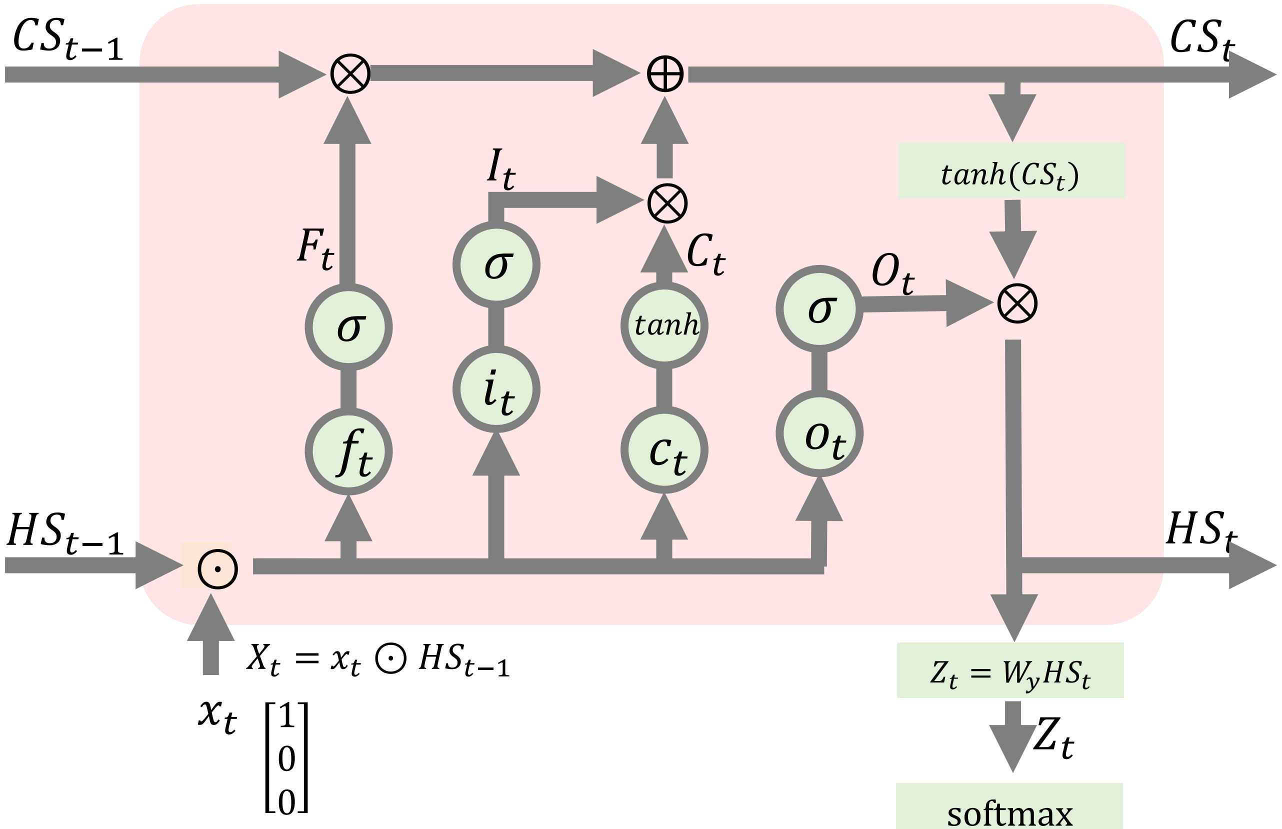
Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_y}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix}0.305\\-0.663\\0.358\end{bmatrix}$$

신박AI

$\partial L / \partial W_y$은 체인룰에 의해 다음과 같습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

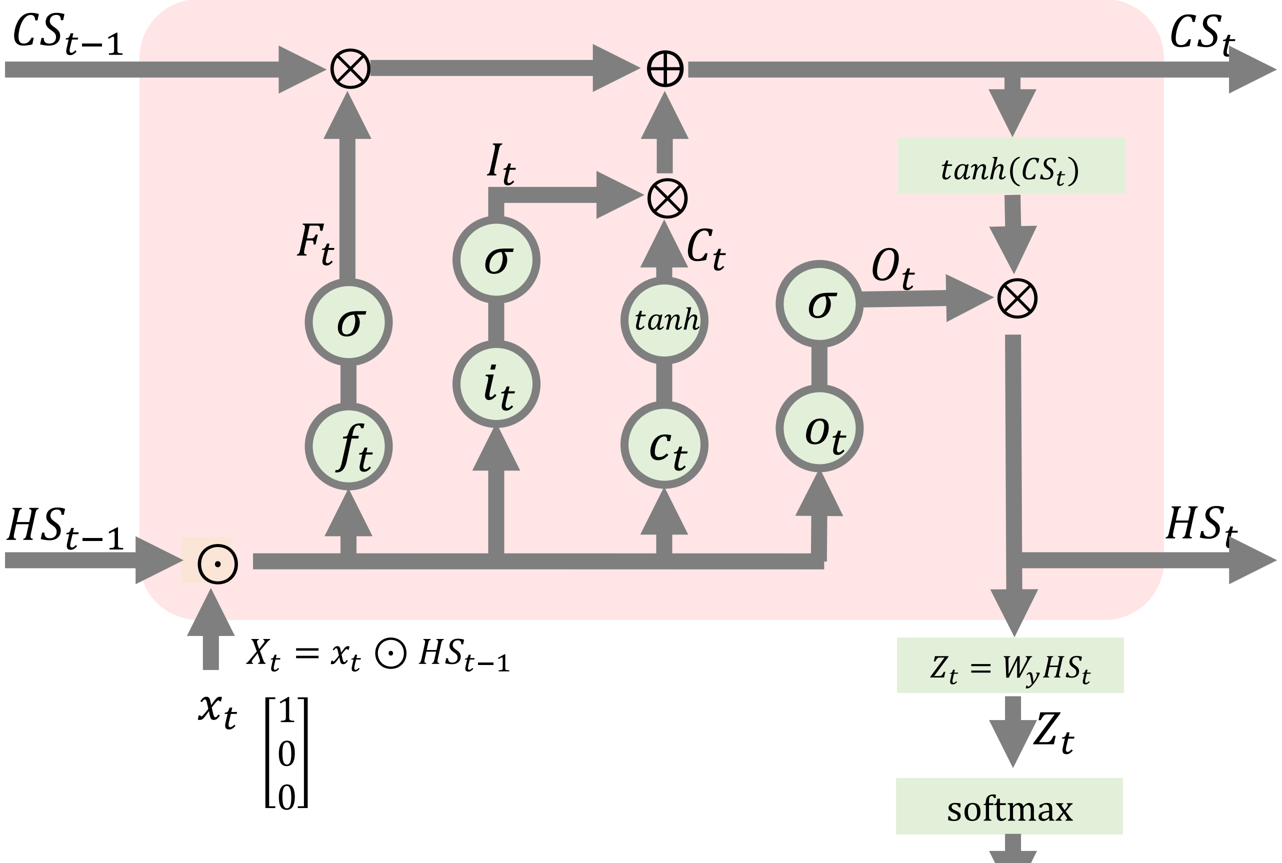
Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_y} = \frac{\partial L}{\partial Z_t}\frac{\partial Z_t}{\partial W_y}$$



$$\frac{\partial L}{\partial Z_t} = \widehat{y_t} - y = \begin{bmatrix}0.305\\-0.663\\0.358\end{bmatrix}$$

신박AI

그러면 $\partial L/\partial Z_t$는 $\widehat{y_t} - y$는 가 되고..

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

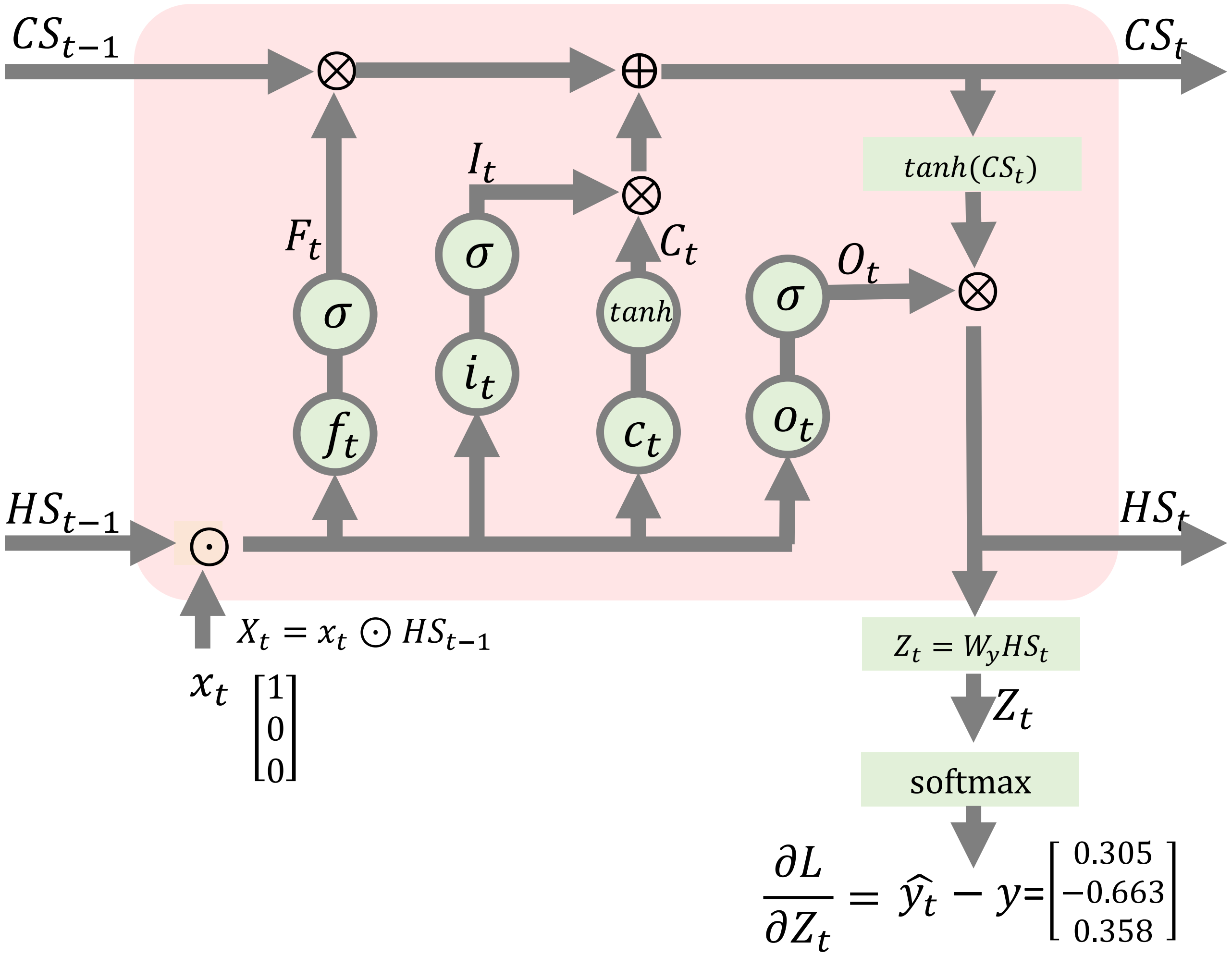
Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_y} = \frac{\partial L}{\partial Z_t}\frac{\partial Z_t}{\partial W_y}$$
$$= (\widehat{y_t} - y)\frac{\partial Z_t}{\partial W_y}$$



신박AI

$\partial Z_t / \partial W_y$는 공식에 의해서 HS_t 가 됩니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

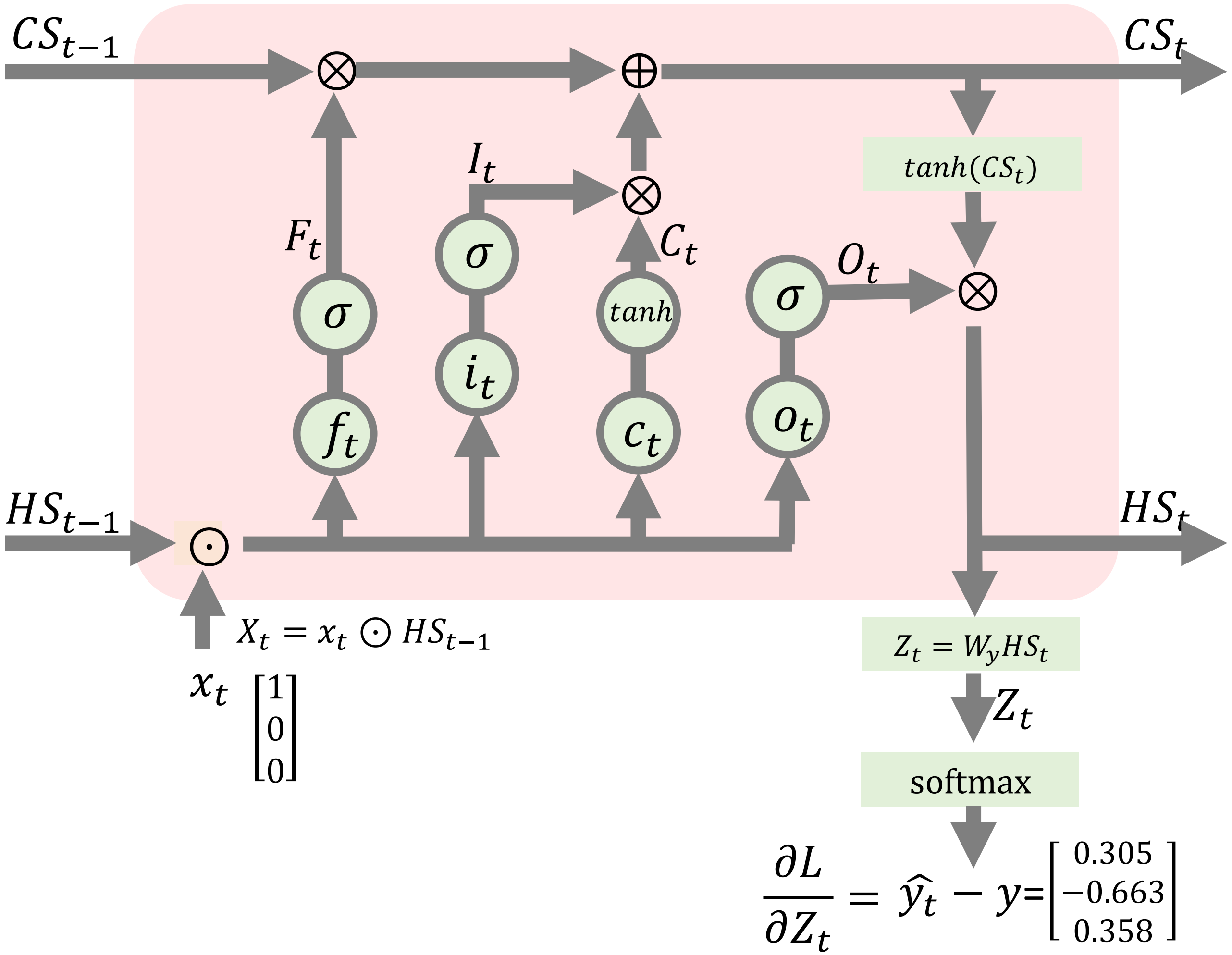
Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_y} = \frac{\partial L}{\partial Z_t}\frac{\partial Z_t}{\partial W_y}$$
$$= (\hat{y}_t - y) HS_t$$



신박AI

그러면 $\partial L/\partial W_y$는 다음과 같이 계산됩니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

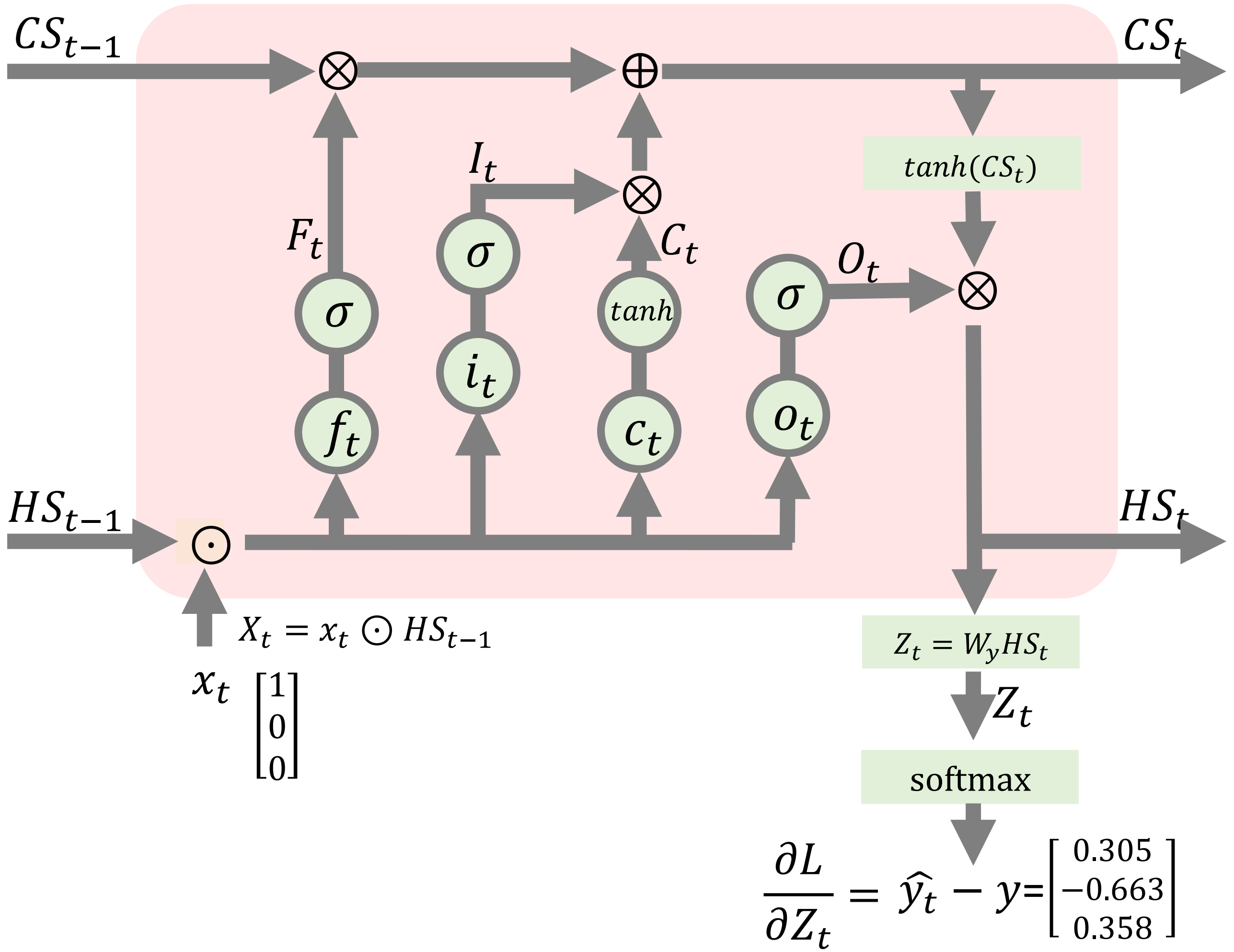
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_y} = \frac{\partial L}{\partial Z_t}\frac{\partial Z_t}{\partial W_y}$$

$$= (\hat{y}_t - y)HS_t$$

$$= \begin{bmatrix} 0.305 \\ -0.663 \\ 0.358 \end{bmatrix} \cdot \begin{bmatrix} 0.119 \\ 0.16 \end{bmatrix}^T = \begin{bmatrix} 0.036 & 0.049 \\ -0.079 & -0.106 \\ 0.043 & 0.057 \end{bmatrix}$$



신박AI

그러면 새로운 W_y인 W_y^*는 경사하강법에 의해 다음과 같이 계산할 수 있습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

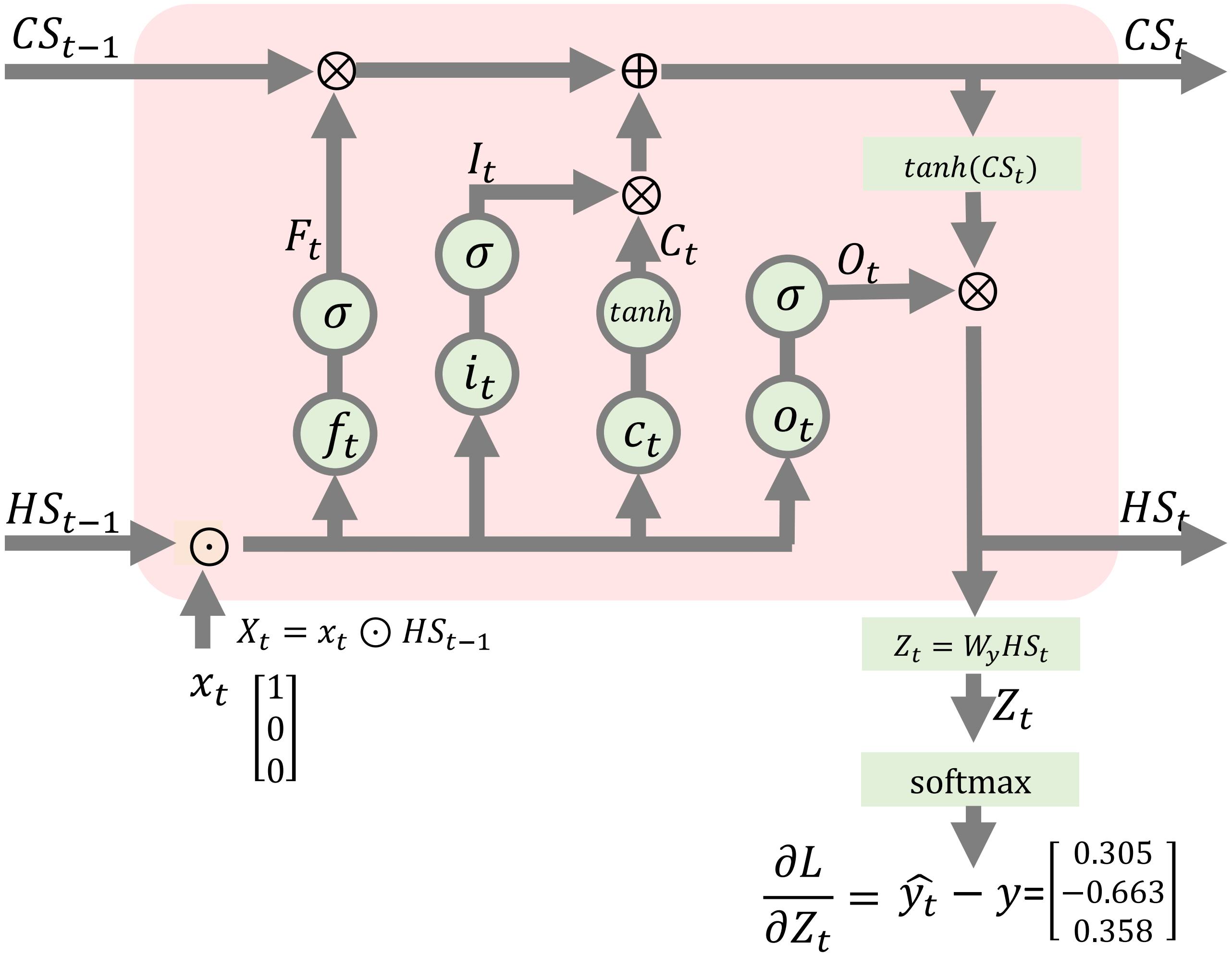
Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_y} = \frac{\partial L}{\partial Z_t}\frac{\partial Z_t}{\partial W_y} = (\hat{y}_t - y)HS_t = \begin{bmatrix} 0.305 \\ -0.663 \\ 0.358 \end{bmatrix} \cdot \begin{bmatrix} 0.119 \\ 0.16 \end{bmatrix}^T = \begin{bmatrix} 0.036 & 0.049 \\ -0.079 & -0.106 \\ 0.043 & 0.057 \end{bmatrix}$$

$$W_y^* = W_y - \alpha \cdot \frac{\partial L}{\partial W_y}$$



신박AI

지금의 경우는 입력 x_t의 길이가 1이고 출력 $\hat{y}_t$의 길이가 1인 경우입니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

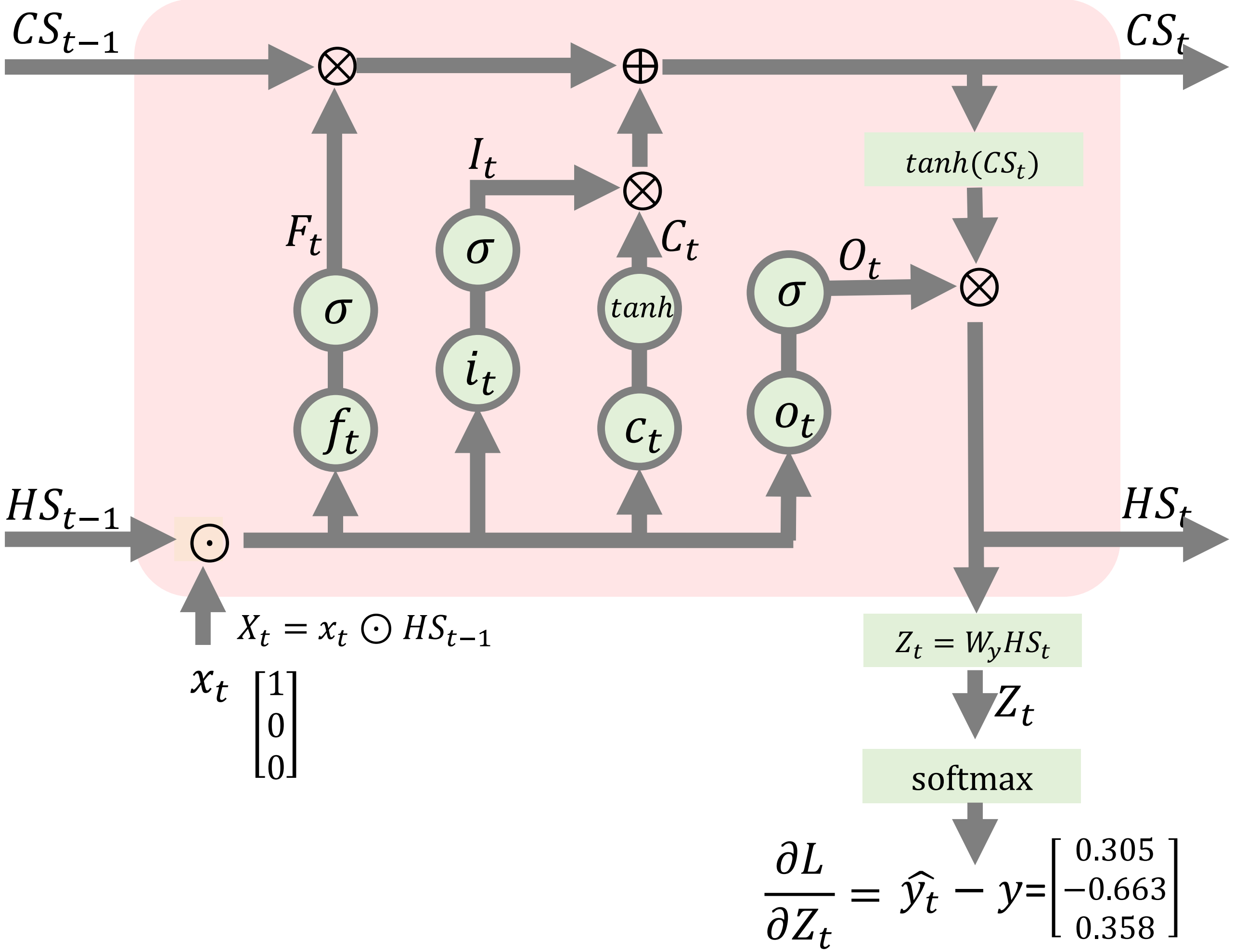
$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_y} = \frac{\partial L}{\partial Z_t}\frac{\partial Z_t}{\partial W_y}$$

$$= (\hat{y}_t - y)HS_t$$

$$= \begin{bmatrix} 0.305 \\ -0.663 \\ 0.358 \end{bmatrix} \cdot \begin{bmatrix} 0.119 \\ 0.16 \end{bmatrix}^T = \begin{bmatrix} 0.036 & 0.049 \\ -0.079 & -0.106 \\ 0.043 & 0.057 \end{bmatrix}$$

$$W_y^* = W_y - \alpha \cdot \frac{\partial L}{\partial W_y}$$



신박AI

즉 입력이 a➔b, b➔c 이렇게 하나씩 입력을 받아 오차를 계산하는 경우라고 가정할 경우,

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

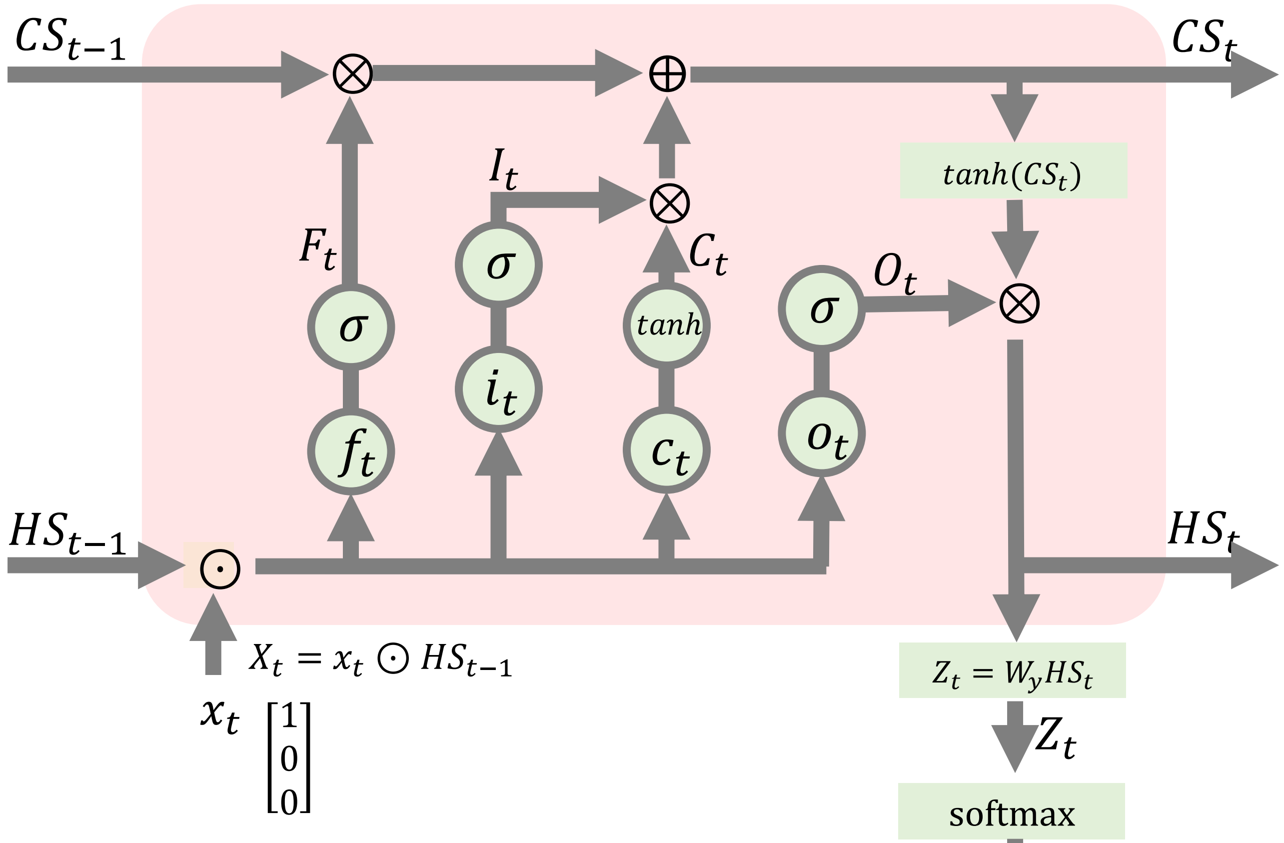
$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_y} = \frac{\partial L}{\partial Z_t}\frac{\partial Z_t}{\partial W_y}$$

$$= (\hat{y}_t - y)HS_t$$

$$= \begin{bmatrix} 0.305 \\ -0.663 \\ 0.358 \end{bmatrix} \cdot \begin{bmatrix} 0.119 \\ 0.16 \end{bmatrix}^T = \begin{bmatrix} 0.036 & 0.049 \\ -0.079 & -0.106 \\ 0.043 & 0.057 \end{bmatrix}$$

$$W_y^* = W_y - \alpha \cdot \frac{\partial L}{\partial W_y}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix} 0.305 \\ -0.663 \\ 0.358 \end{bmatrix}$$

신박AI

지금과 같이 계산을 해야하며,

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

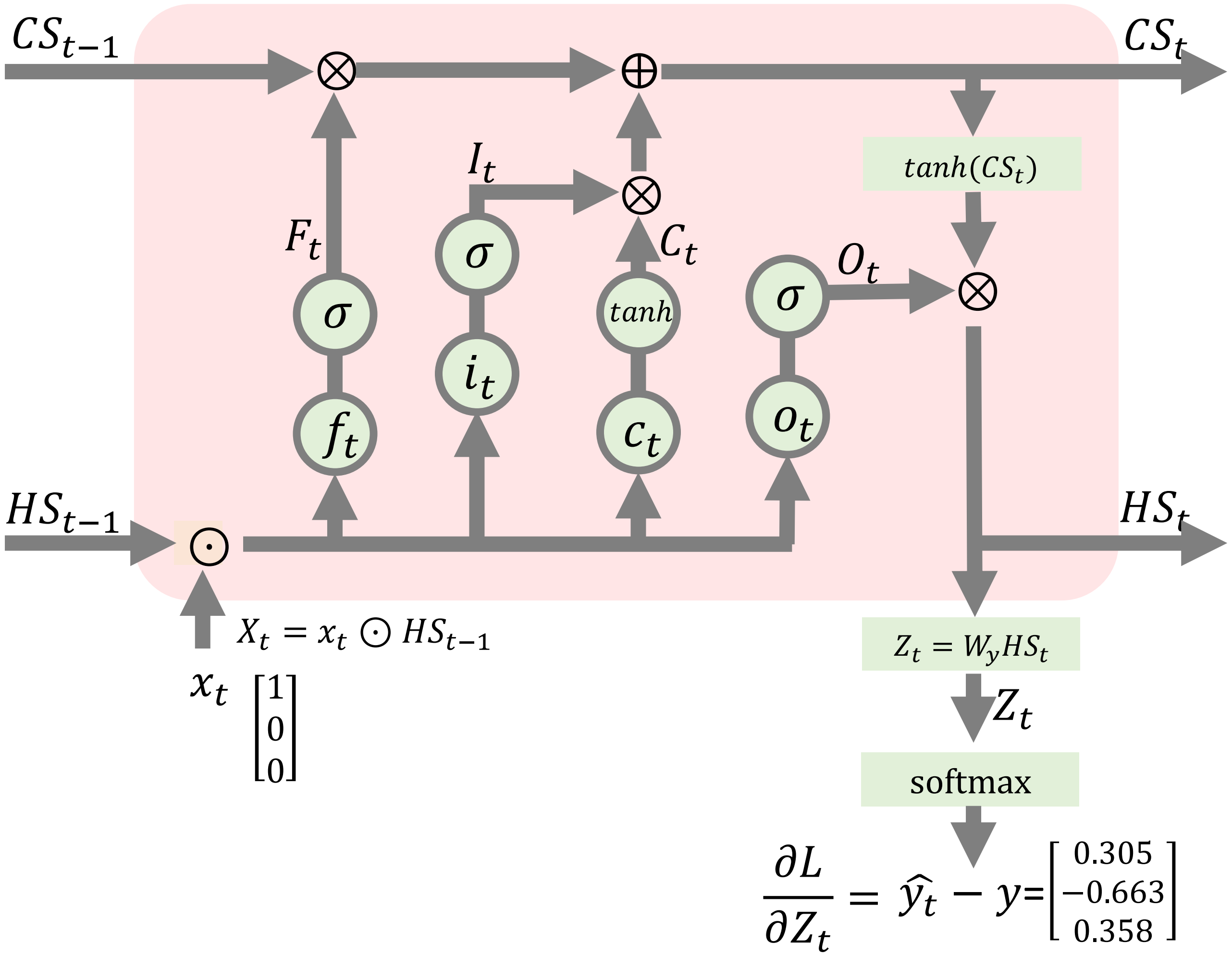
$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_y} = \frac{\partial L}{\partial Z_t}\frac{\partial Z_t}{\partial W_y}$$

$$= (\hat{y}_t - y)HS_t$$

$$= \begin{bmatrix} 0.305 \\ -0.663 \\ 0.358 \end{bmatrix} \cdot \begin{bmatrix} 0.119 \\ 0.16 \end{bmatrix}^T = \begin{bmatrix} 0.036 & 0.049 \\ -0.079 & -0.106 \\ 0.043 & 0.057 \end{bmatrix}$$

$$W_y^* = W_y - \alpha \cdot \frac{\partial L}{\partial W_y}$$



신박AI

만약 ab➔bc, bc➔ca 처럼 입력 길이가 길어질 수록..

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

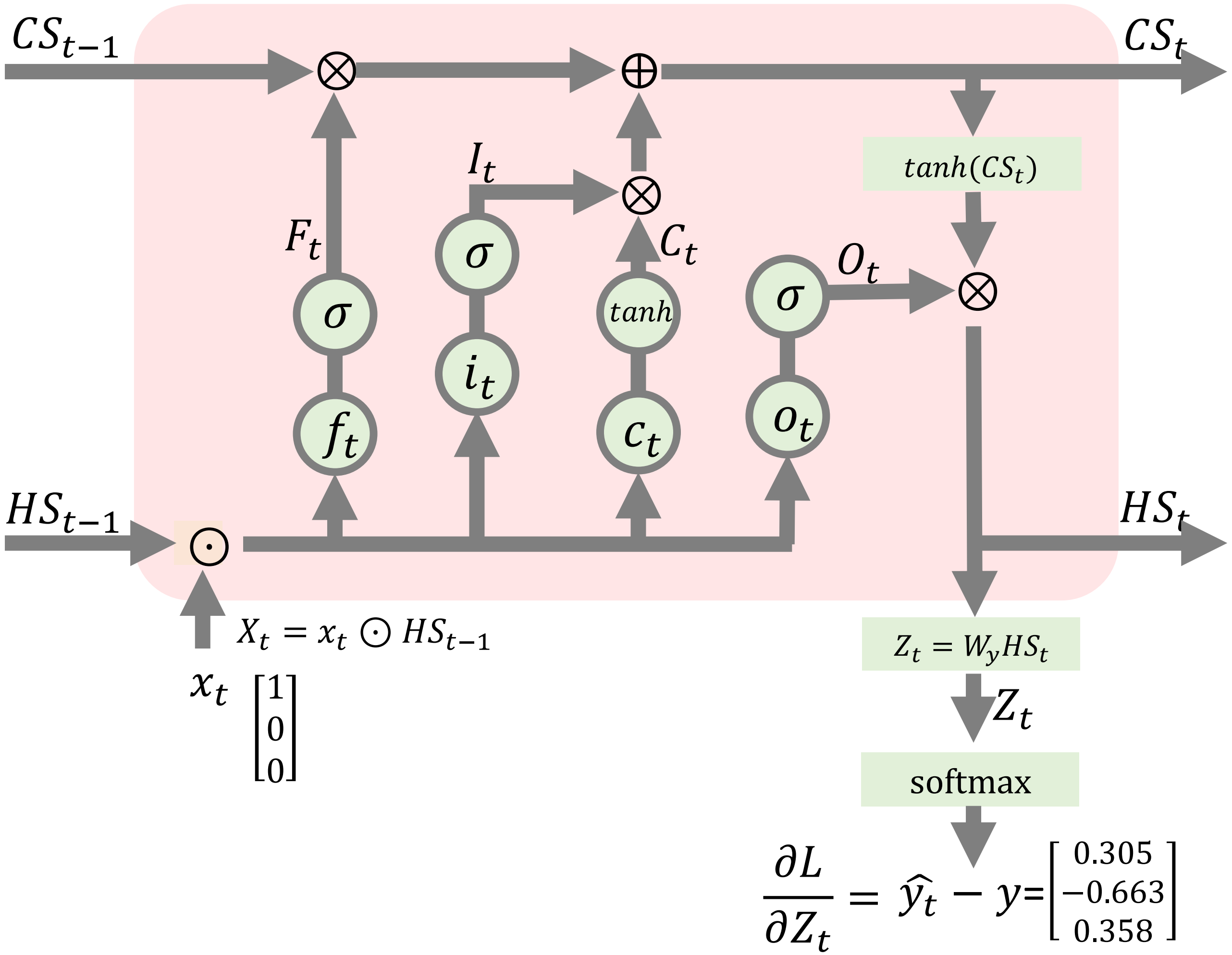
Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_y} = \frac{\partial L}{\partial Z_t}\frac{\partial Z_t}{\partial W_y}$$
$$= (\hat{y}_t - y)HS_t$$
$$= \begin{bmatrix} 0.305 \\ -0.663 \\ 0.358 \end{bmatrix} \cdot \begin{bmatrix} 0.119 \\ 0.16 \end{bmatrix}^T = \begin{bmatrix} 0.036 & 0.049 \\ -0.079 & -0.106 \\ 0.043 & 0.057 \end{bmatrix}$$

$$W_y^* = W_y - \alpha \cdot \frac{\partial L}{\partial W_y}$$



신박AI

다음처럼 에러를 더해주어야 한다는 점에 유의해주시길 바랍니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

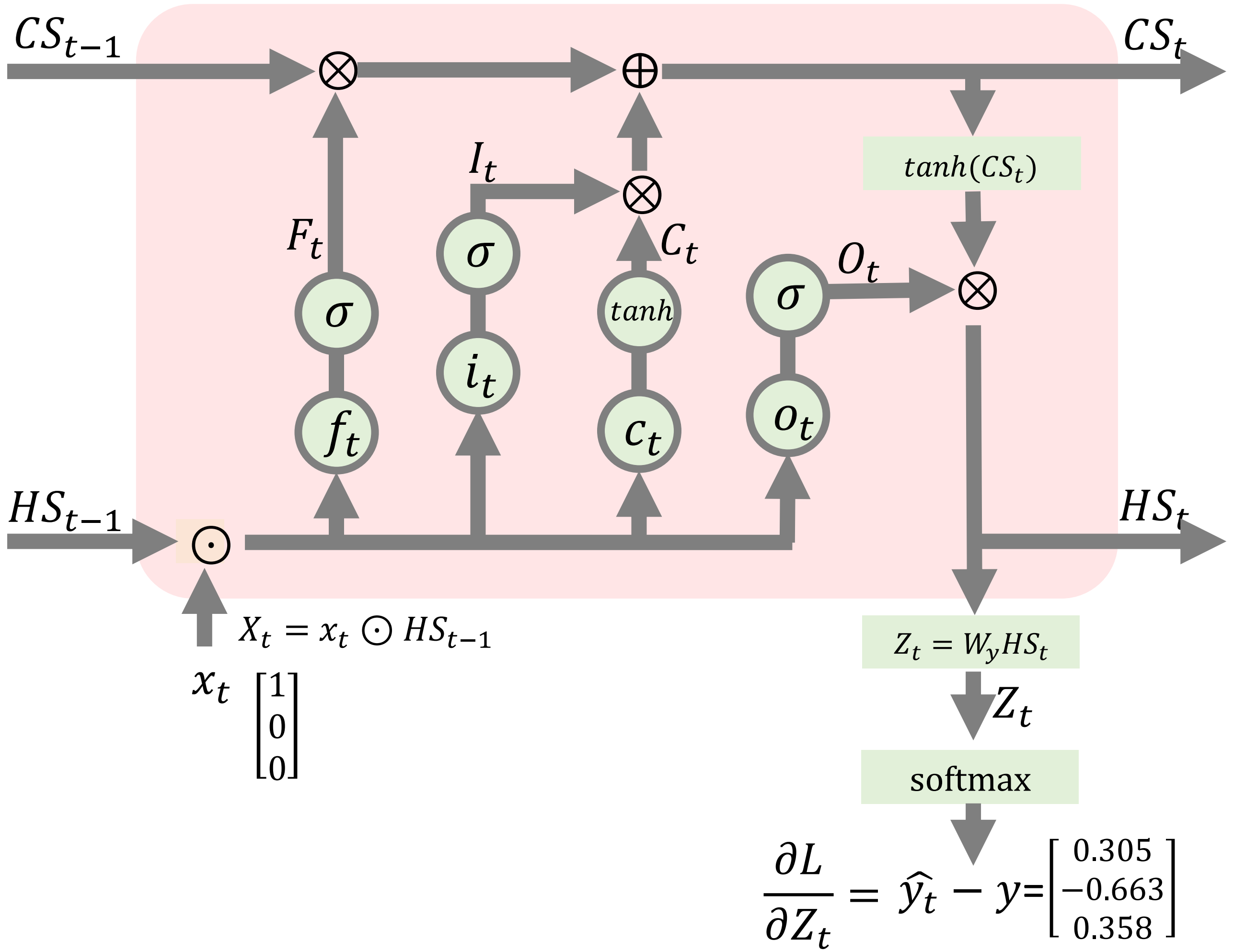
Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_y} = \frac{\partial L_1}{\partial W_y} + \frac{\partial L_2}{\partial W_y} + \frac{\partial L_3}{\partial W_y} + \cdots$$



신박AI

자 그러면 이제 게이트 쪽으로 넘어가도록 하겠습니다

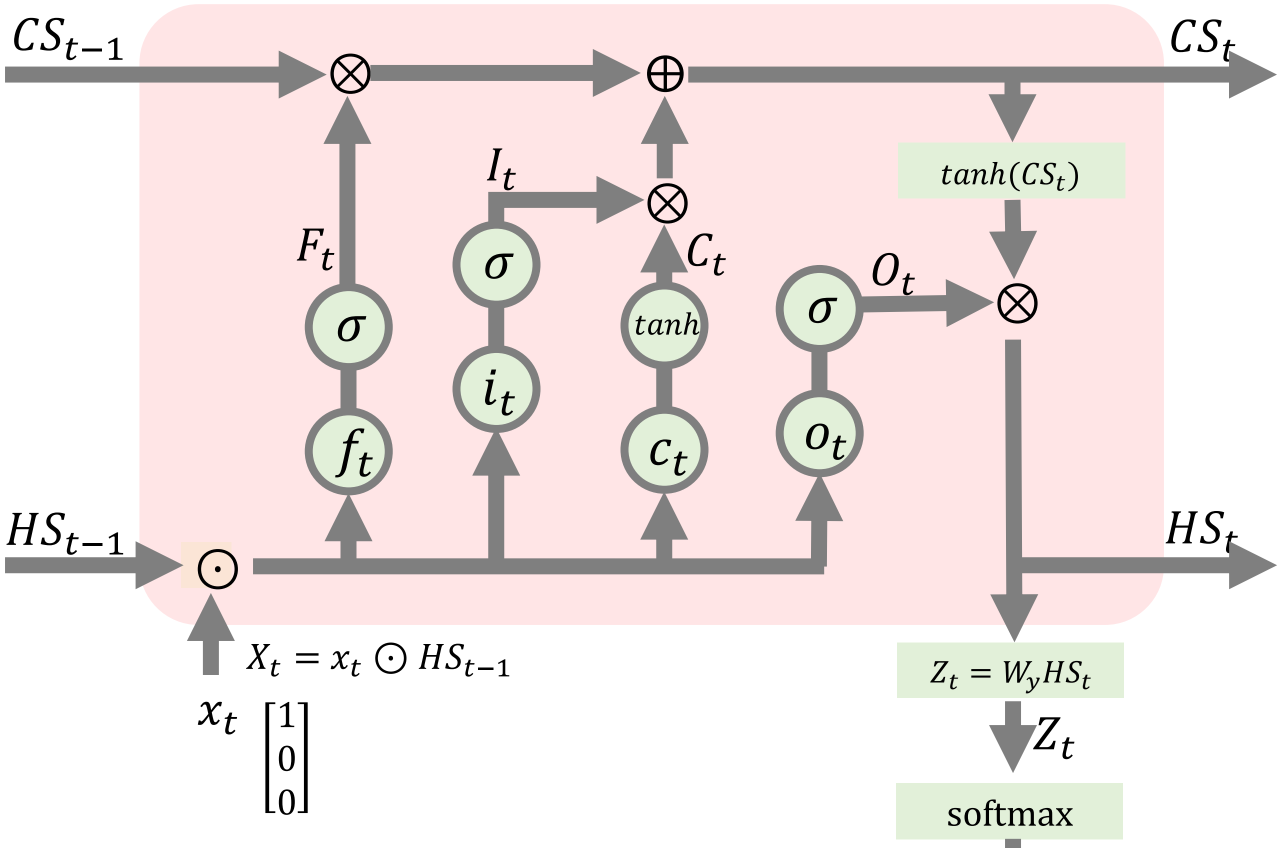
Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$



$x_t \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$

$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix} 0.305 \\ -0.663 \\ 0.358 \end{bmatrix}$$

먼저 Output Gate에 있는 가중치를 업데이트 해보도록 하겠습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

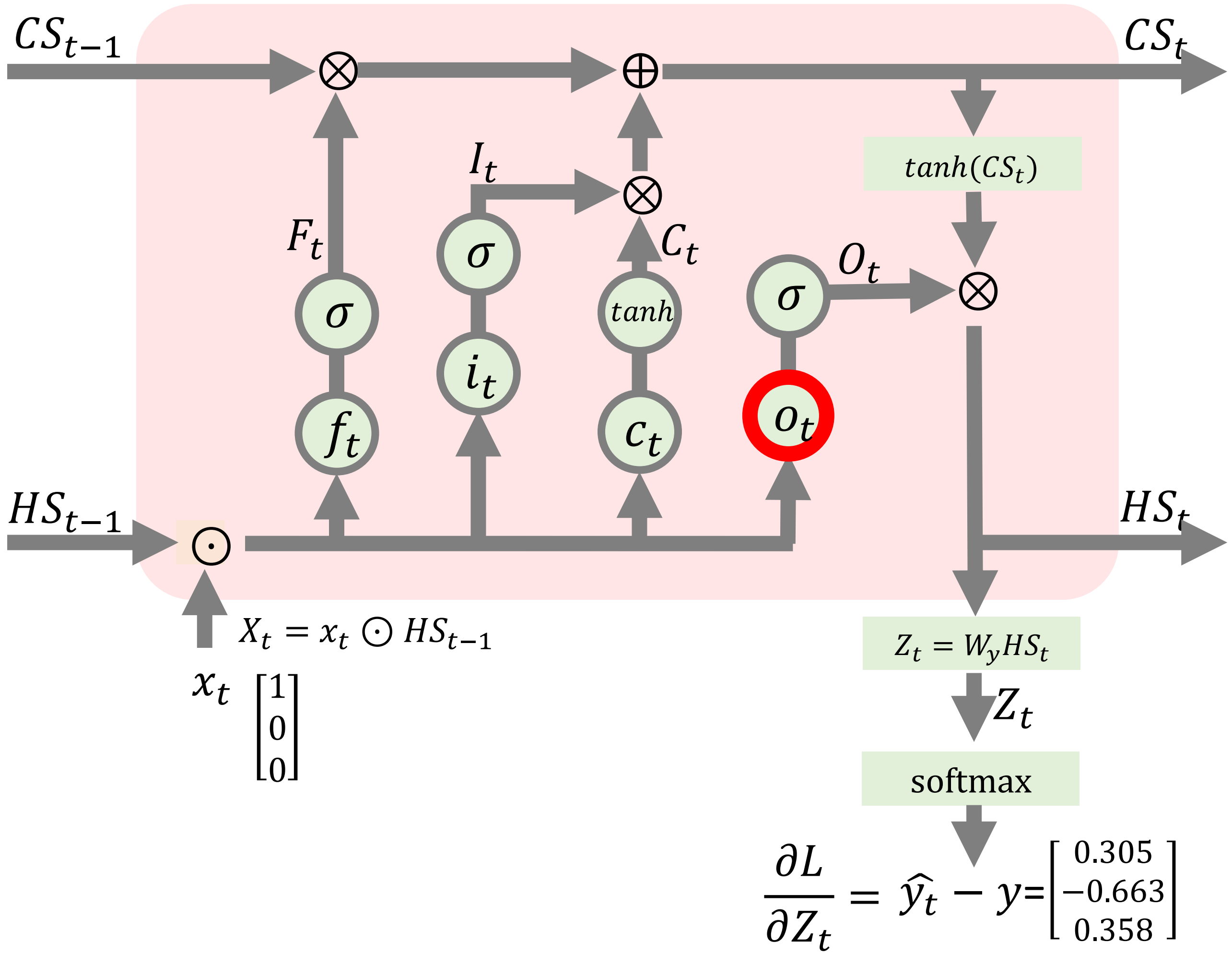
Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_o}$$



신박AI

$\partial L / \partial W_o$ 는 체인룰에 의해서 다음과 같이 전개할 수 있습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

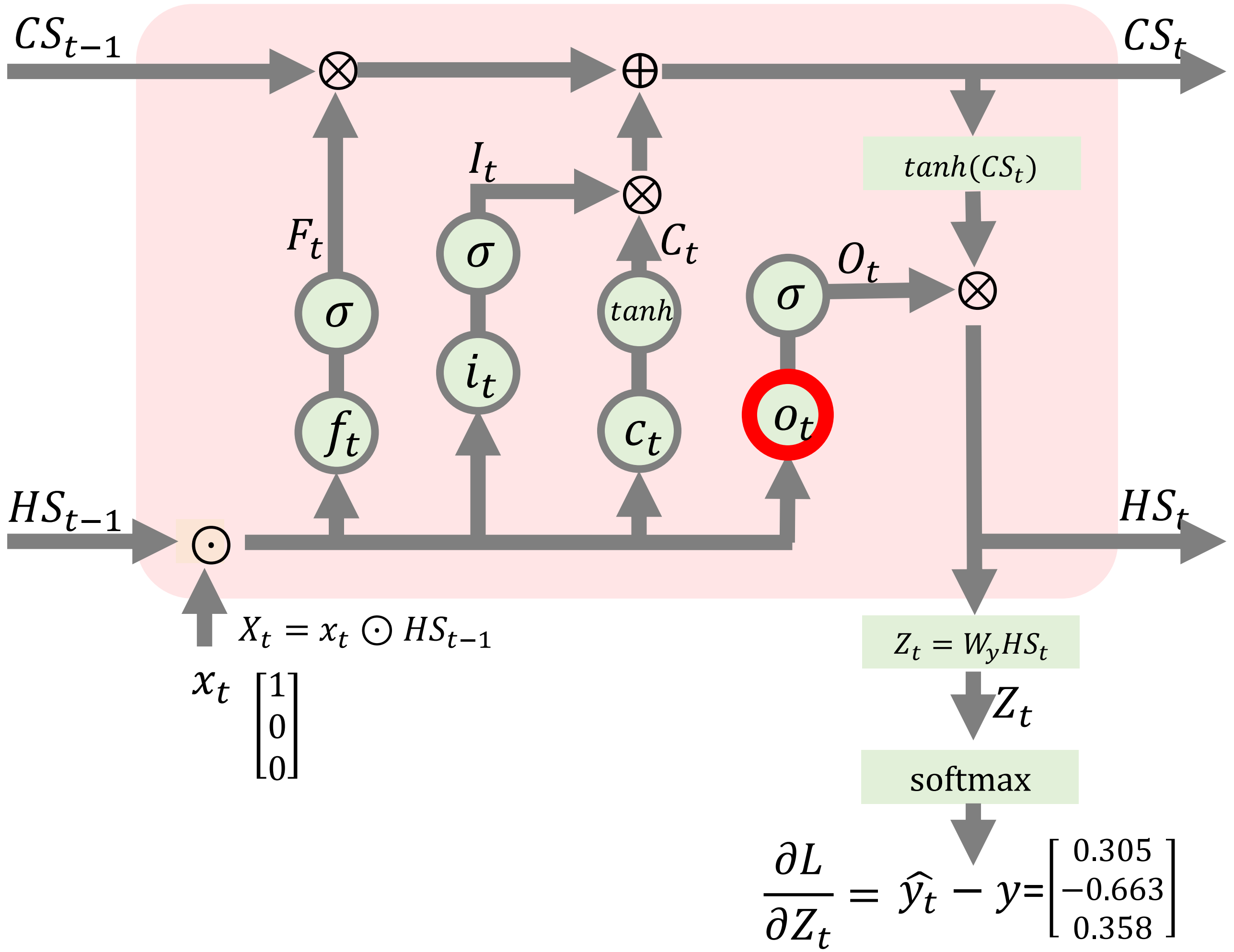
Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_o} = \frac{\partial L}{\partial HS_t} \frac{\partial HS_t}{\partial O_t} \frac{\partial O_t}{\partial o_t} \frac{\partial o_t}{\partial W_o}$$



신박AI

그러면 $\partial L/\partial HS_t$ 를 먼저 구해보도록 하겠습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

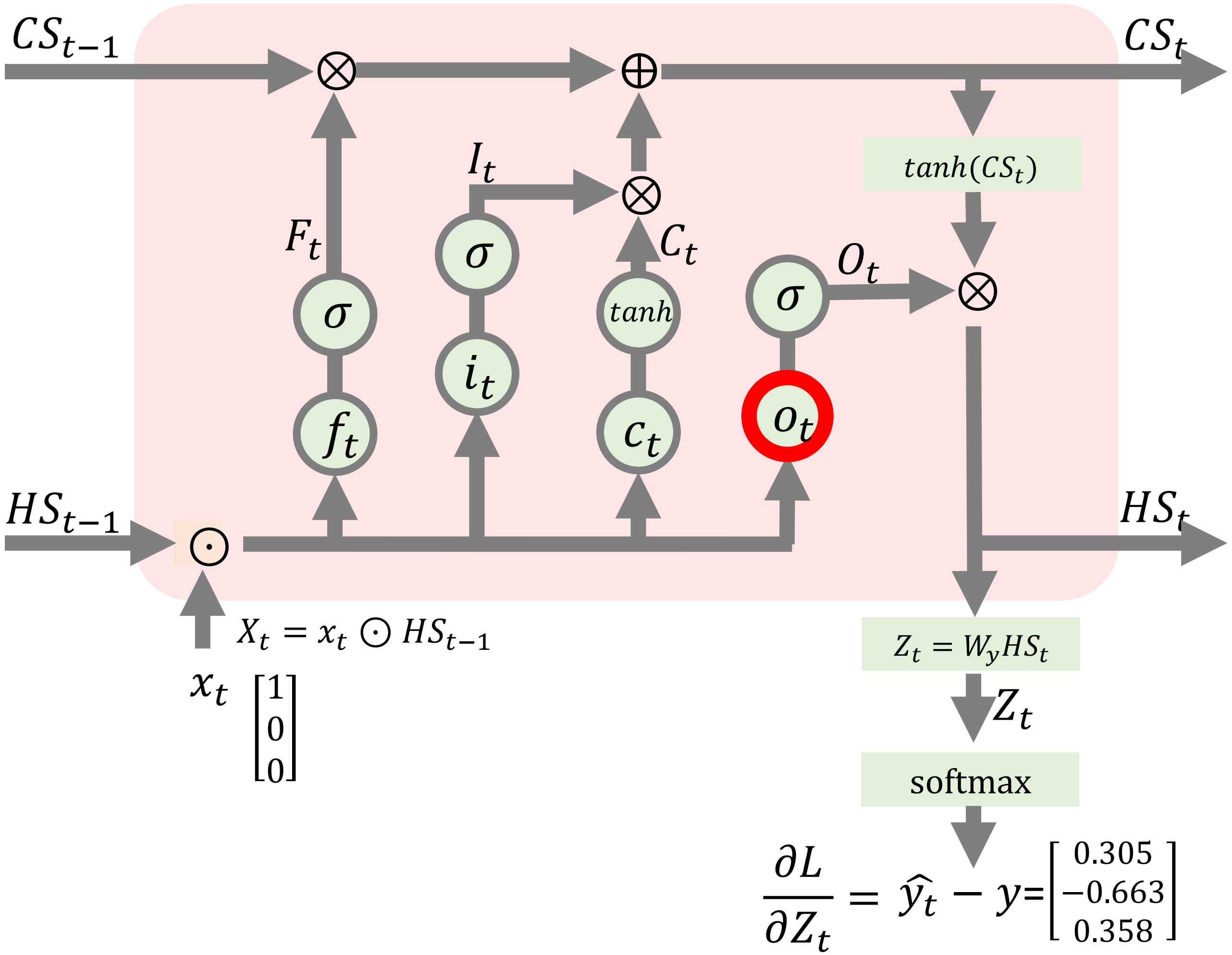
Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_o} = \boxed{\frac{\partial L}{\partial HS_t}} \frac{\partial HS_t}{\partial O_t} \frac{\partial O_t}{\partial o_t} \frac{\partial o_t}{\partial W_o}$$

$$\frac{\partial L}{\partial HS_t}$$



$\partial L/\partial HS_t$는 다음과 같이 전개할 수 있고

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

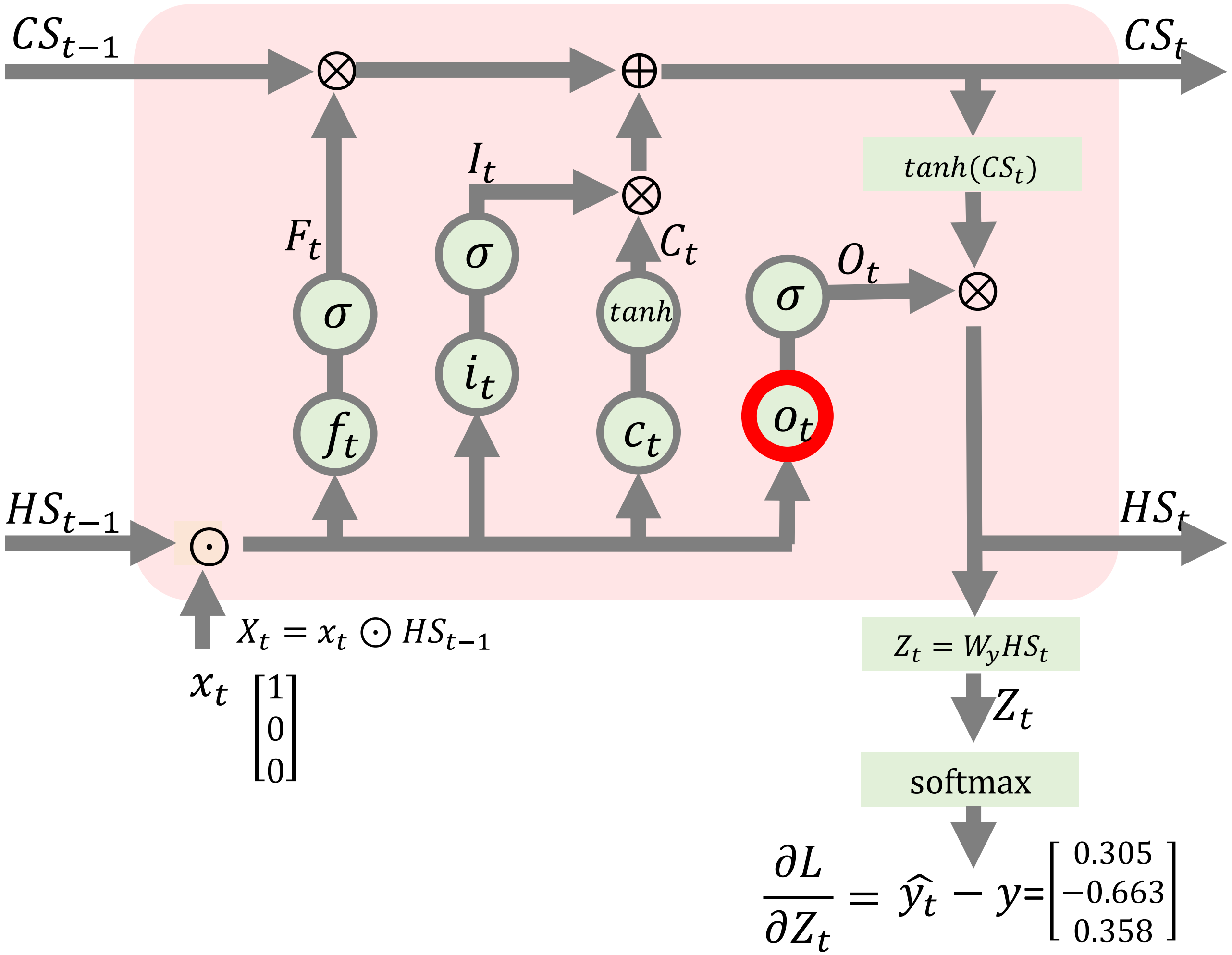
Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_o} = \frac{\partial L}{\partial HS_t}\frac{\partial HS_t}{\partial O_t}\frac{\partial O_t}{\partial o_t}\frac{\partial o_t}{\partial W_o}$$

$$\frac{\partial L}{\partial HS_t} = \frac{\partial L}{\partial Z_t}\frac{\partial Z_t}{\partial HS_t}$$



신박AI

$\partial L/\partial HS_t$는 계속해서 풀어보면 다음과 같습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

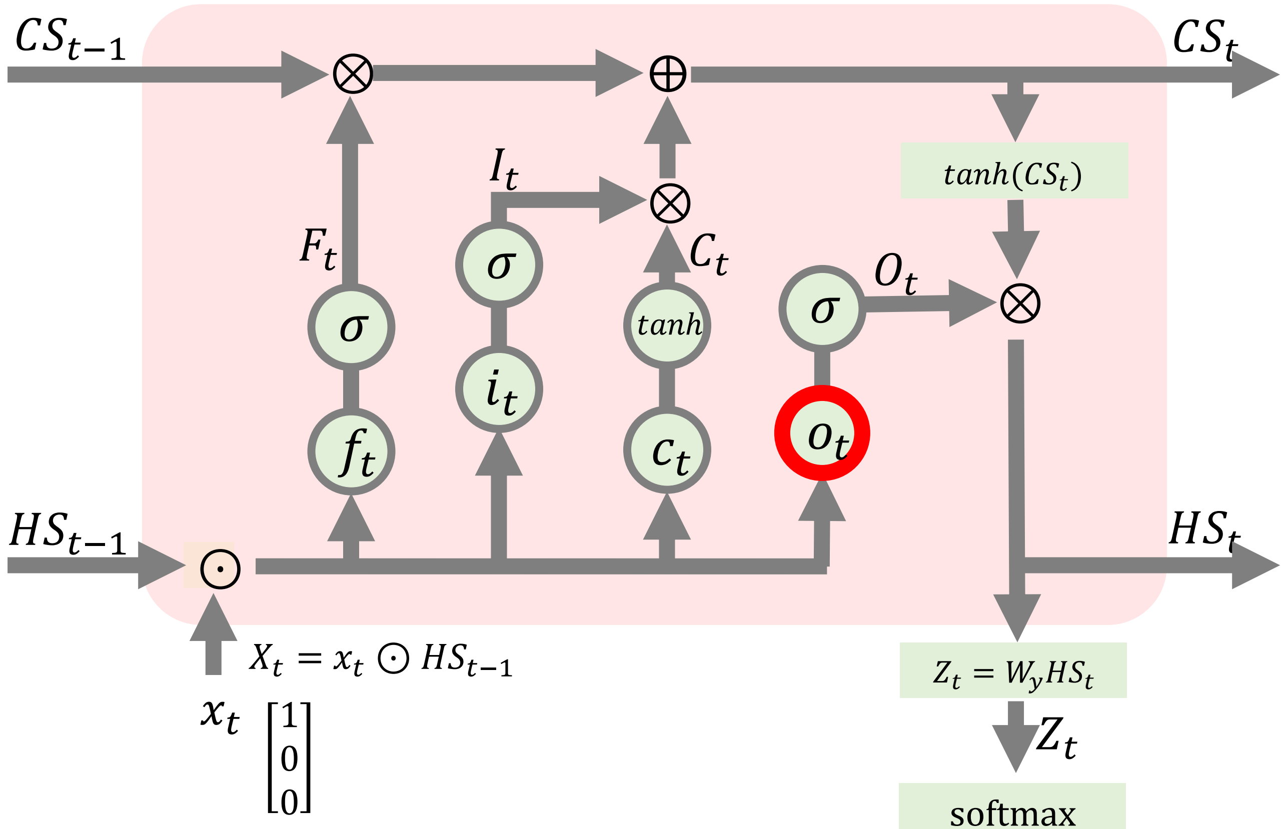
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$\boxed{Z_t = W_y HS_t}$

$$\frac{\partial L}{\partial W_o} = \frac{\partial L}{\partial HS_t}\frac{\partial HS_t}{\partial O_t}\frac{\partial O_t}{\partial o_t}\frac{\partial o_t}{\partial W_o}$$

$$\frac{\partial L}{\partial HS_t} = \frac{\partial L}{\partial Z_t}\frac{\partial Z_t}{\partial HS_t}$$

$$= (\hat{y}_t - y)W_y$$



$\boxed{\frac{\partial L}{\partial Z_t}} = \hat{y}_t - y = \begin{bmatrix}0.305\\-0.663\\0.358\end{bmatrix}$

신박AI

결과물을 바탕으로 식을 다시 정리하면 이렇게 됩니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

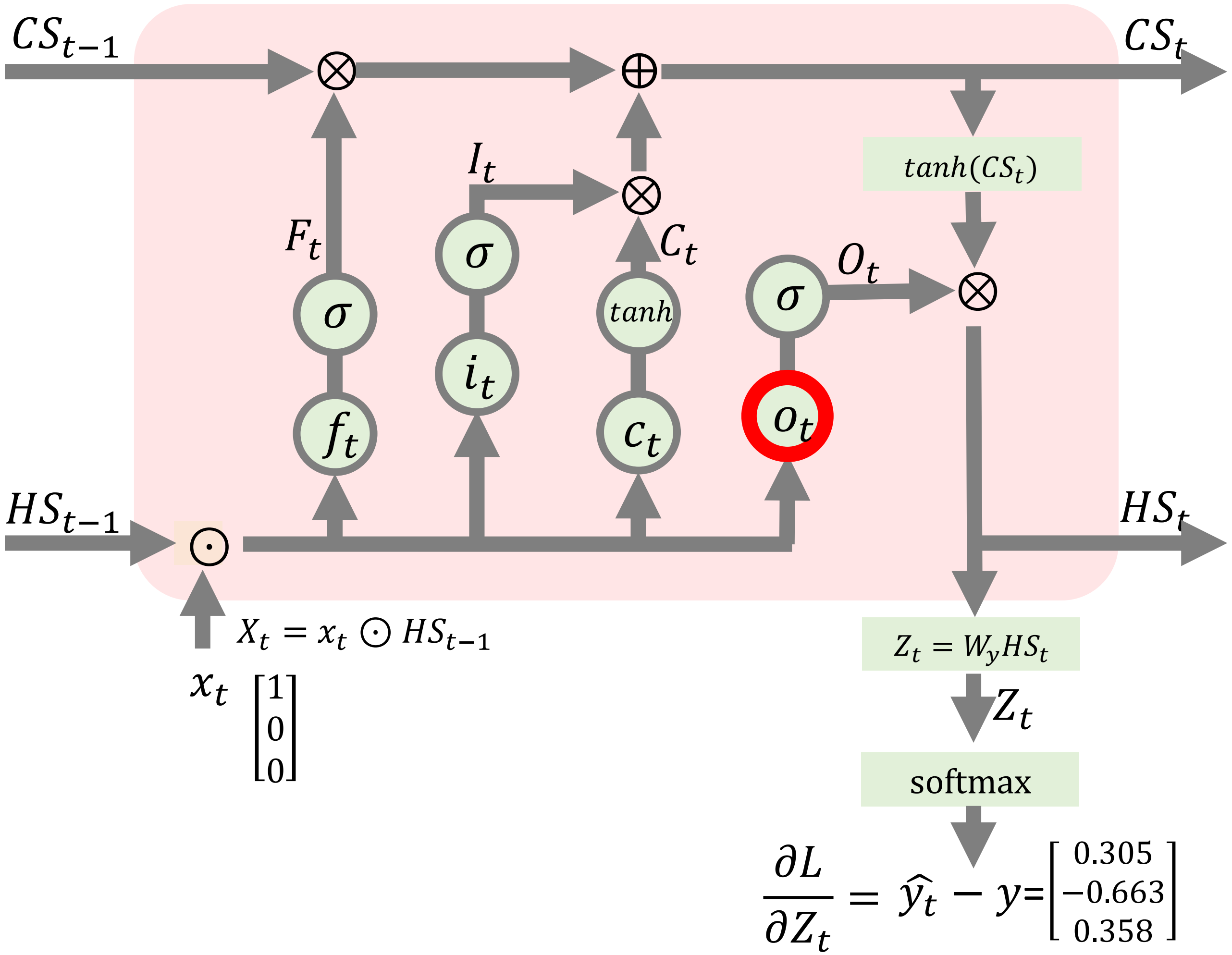
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_o} = (\hat{y}_t - y)W_y \frac{\partial HS_t}{\partial O_t}\frac{\partial O_t}{\partial o_t}\frac{\partial o_t}{\partial W_o}$$

$$\frac{\partial L}{\partial HS_t} = \frac{\partial L}{\partial Z_t}\frac{\partial Z_t}{\partial HS_t}$$

$$= (\hat{y}_t - y)W_y$$



신박AI

그런데 실제 $\partial L/\partial HS_t$ 는 이것보다는 좀 더 복잡합니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

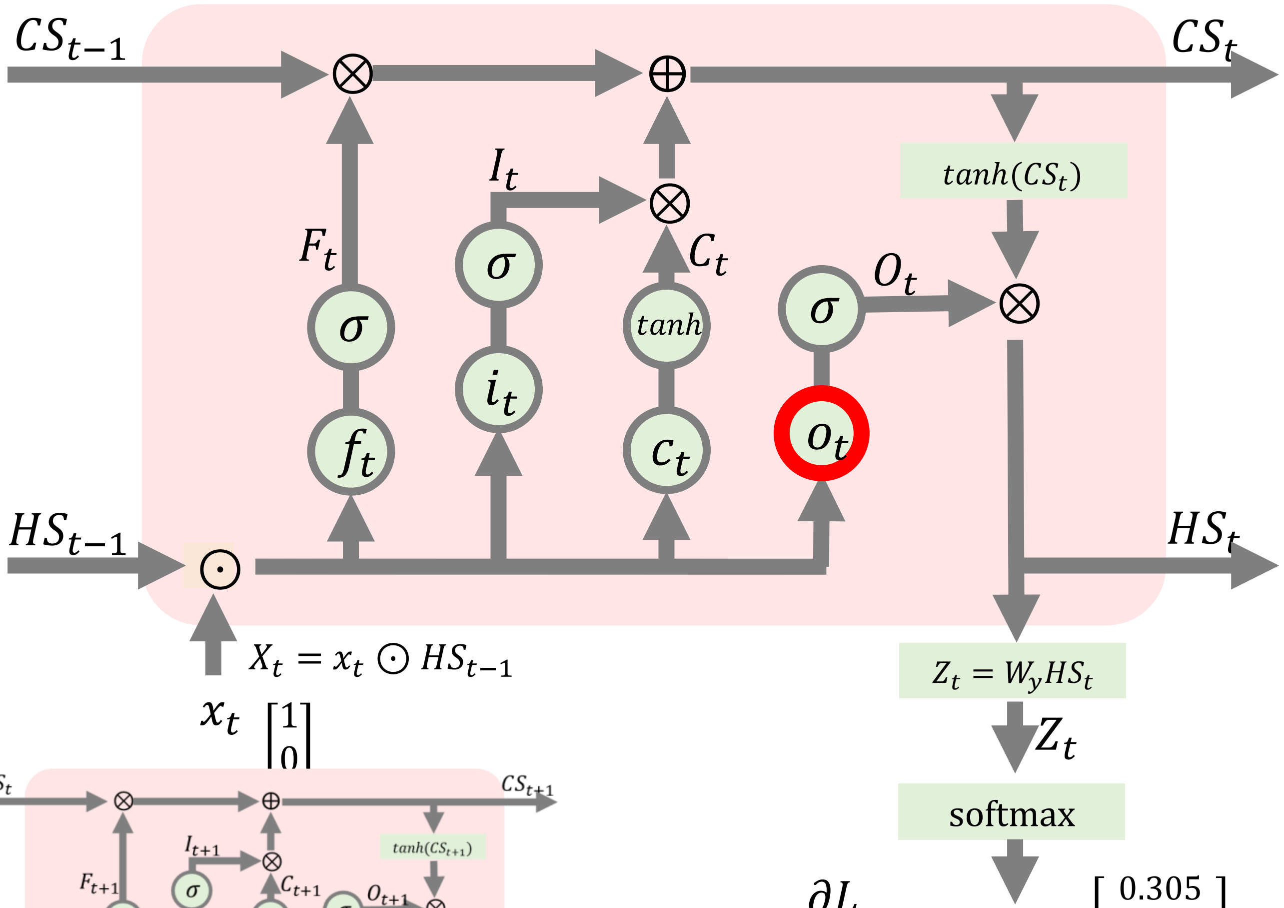
Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

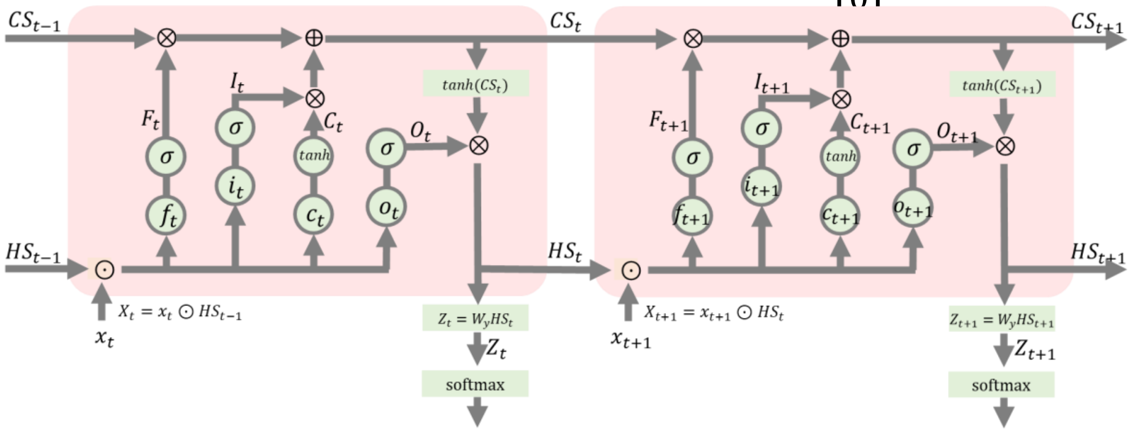
$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_o} = (\hat{y}_t - y)W_y \frac{\partial HS_t}{\partial O_t}\frac{\partial O_t}{\partial o_t}\frac{\partial o_t}{\partial W_o}$$

$$\frac{\partial L}{\partial HS_t} = \frac{\partial L}{\partial Z_t}\frac{\partial Z_t}{\partial HS_t}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix} 0.305 \\ -0.663 \\ 0.358 \end{bmatrix}$$



$\partial L/\partial HS_t$ 는 사실상 현재의 변화와 이전 단계에서의 변화가 다 함께 포함된 의미입니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

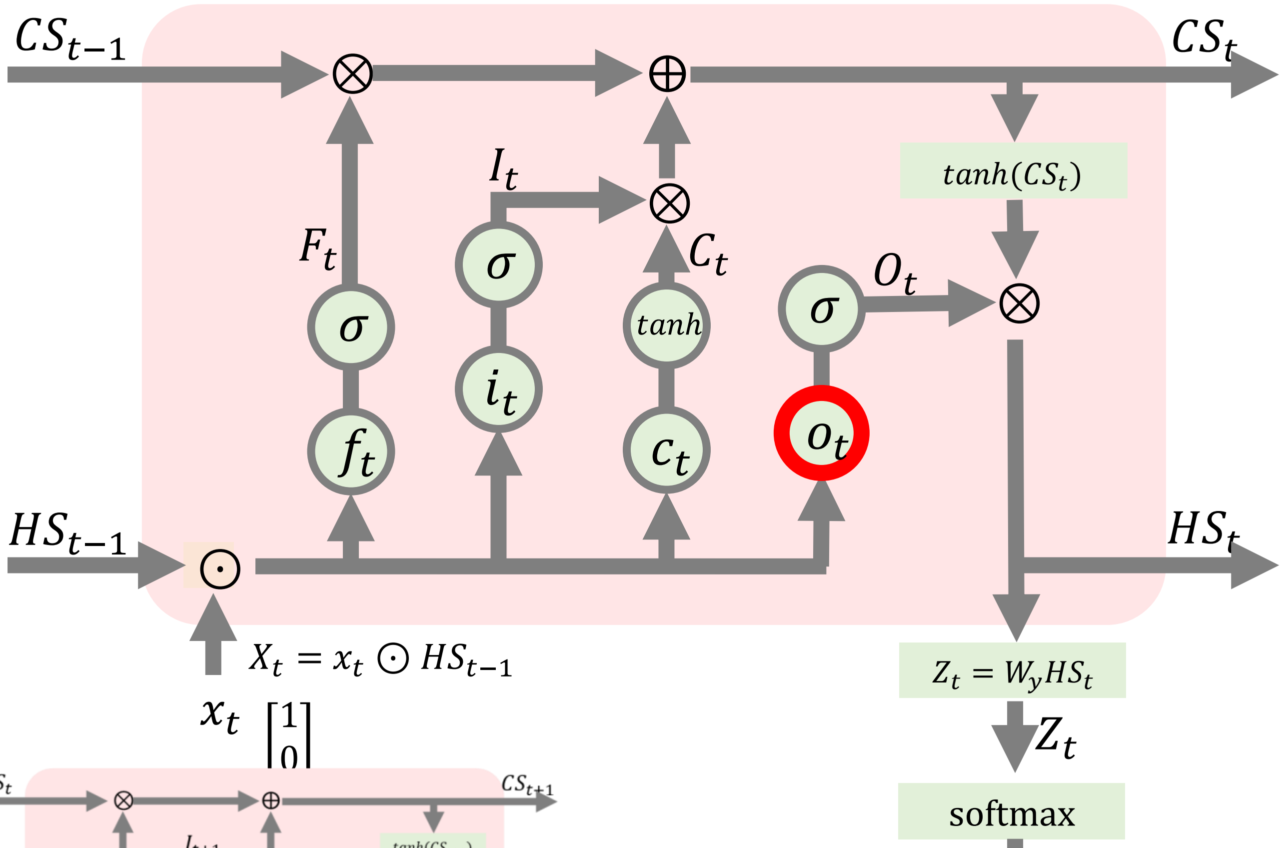
Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

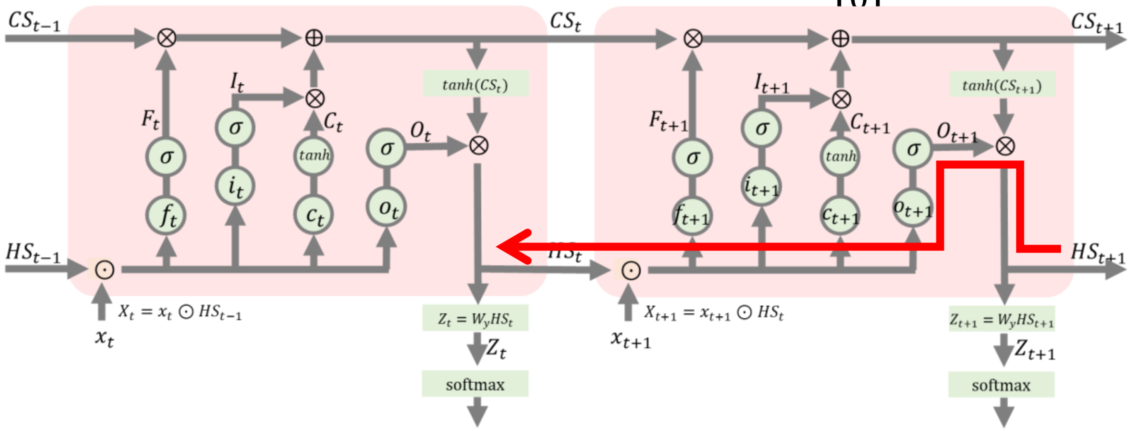
$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_o} = (\hat{y}_t - y) W_y \frac{\partial HS_t}{\partial O_t} \frac{\partial O_t}{\partial o_t} \frac{\partial o_t}{\partial W_o}$$

$$\frac{\partial L}{\partial HS_t} = \frac{\partial L}{\partial Z_t} \frac{\partial Z_t}{\partial HS_t} + dHS_{t+1}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix} 0.305 \\ -0.663 \\ 0.358 \end{bmatrix}$$



그래서 dHS_{t+1}를 $\partial L/\partial HS_t$계산에 포함해주어야 합니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

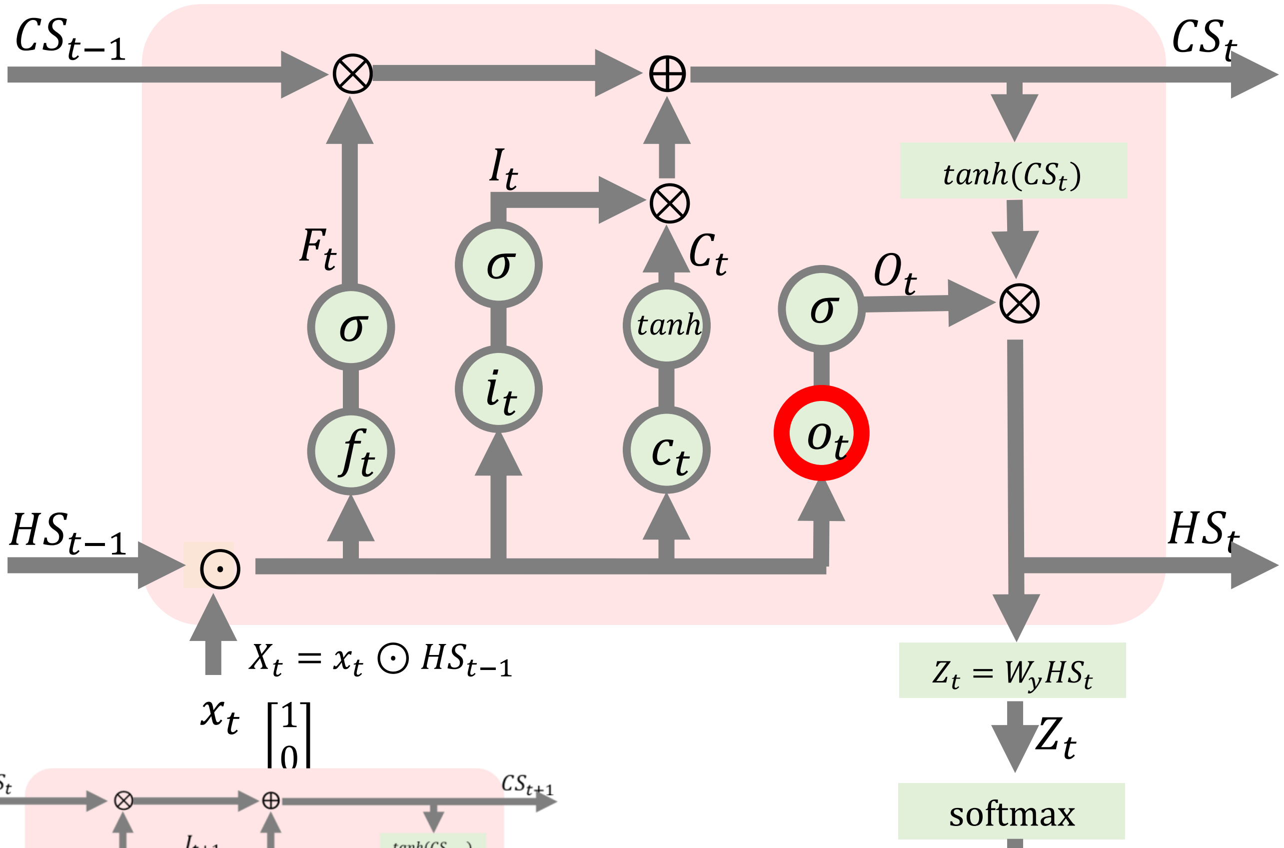
Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

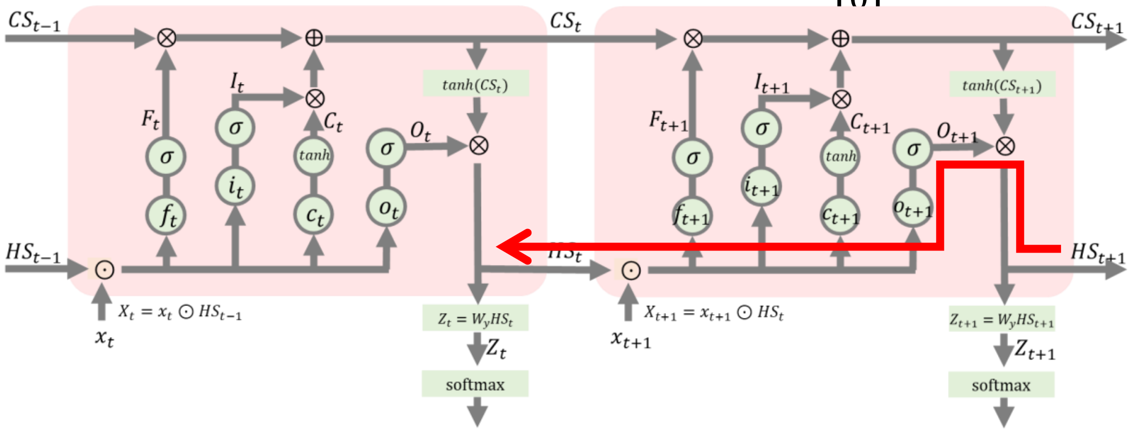
$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_o} = (\hat{y}_t - y)W_y \frac{\partial HS_t}{\partial O_t}\frac{\partial O_t}{\partial o_t}\frac{\partial o_t}{\partial W_o}$$

$$\frac{\partial L}{\partial HS_t} = \frac{\partial L}{\partial Z_t}\frac{\partial Z_t}{\partial HS_t} + dHS_{t+1}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix} 0.305 \\ -0.663 \\ 0.358 \end{bmatrix}$$



신박AI

그러나 지금처럼 숫자를 사용하여 BPTT를 확인하는 과정에서 이런 계산까지 포함하면 너무 복잡해지고

Forget Gate: $F_t = \sigma(W_f X_t)$

$F_t = \sigma(f_t)$

$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$

$I_t = \sigma(i_t)$

$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$

$C_t = tanh(c_t)$

$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$

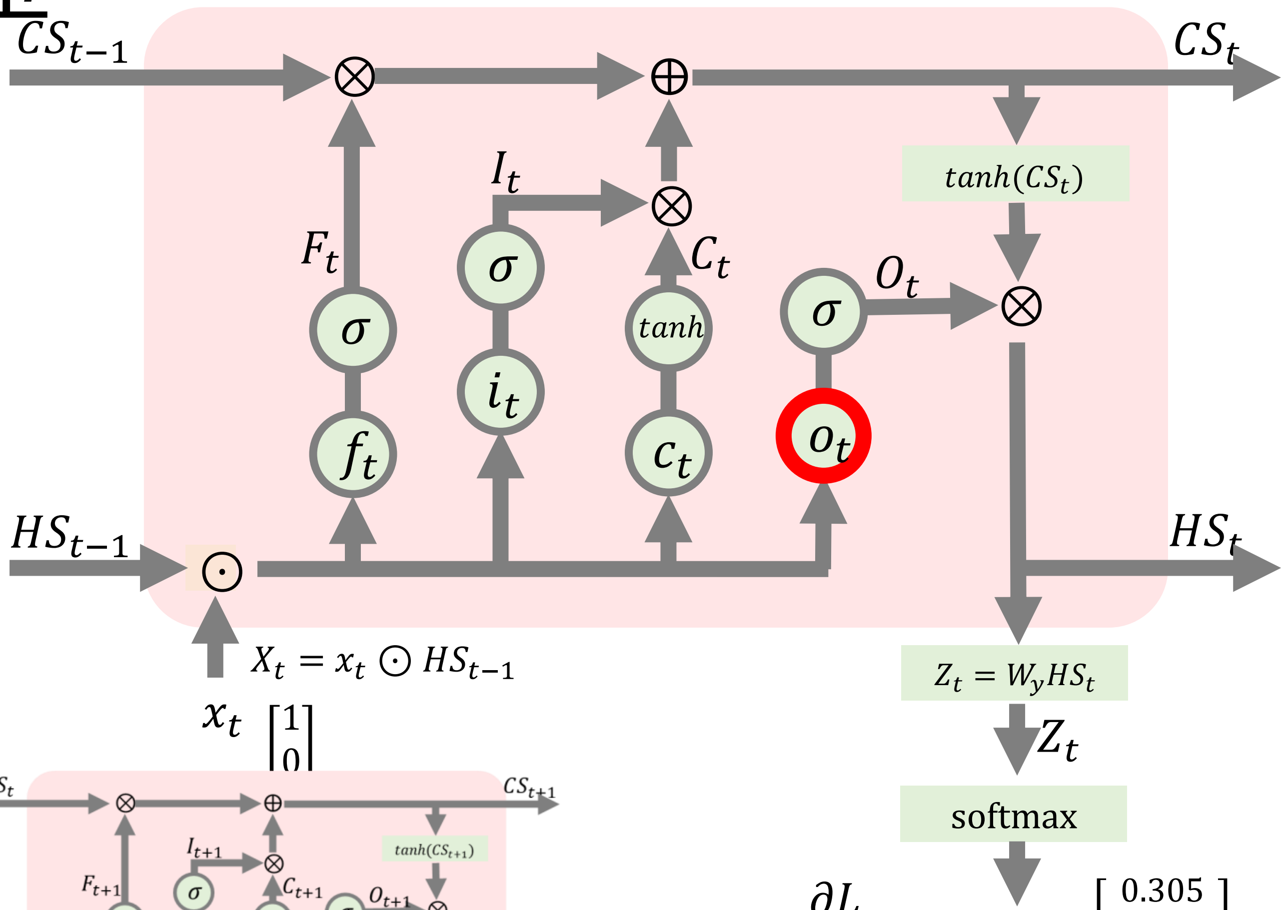
$O_t = \sigma(o_t)$

$o_t = W_o X_t$

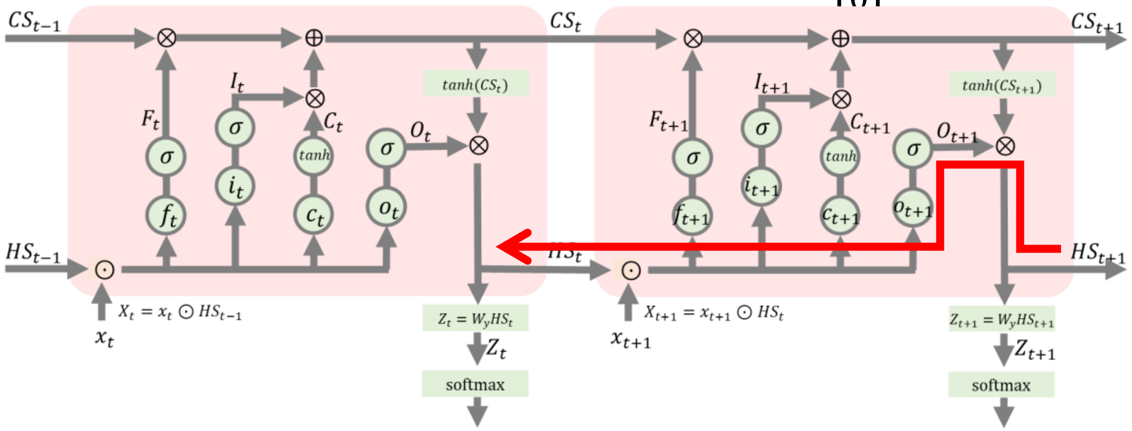
$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_o} = (\hat{y}_t - y) W_y \frac{\partial HS_t}{\partial O_t} \frac{\partial O_t}{\partial o_t} \frac{\partial o_t}{\partial W_o}$$

$$\frac{\partial L}{\partial HS_t} = \frac{\partial L}{\partial Z_t} \frac{\partial Z_t}{\partial HS_t} + dHS_{t+1}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix} 0.305 \\ -0.663 \\ 0.358 \end{bmatrix}$$



오히려 BPTT의 흐름을 이해하시는데 방해가 될 수도 있다는 생각이 듭니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

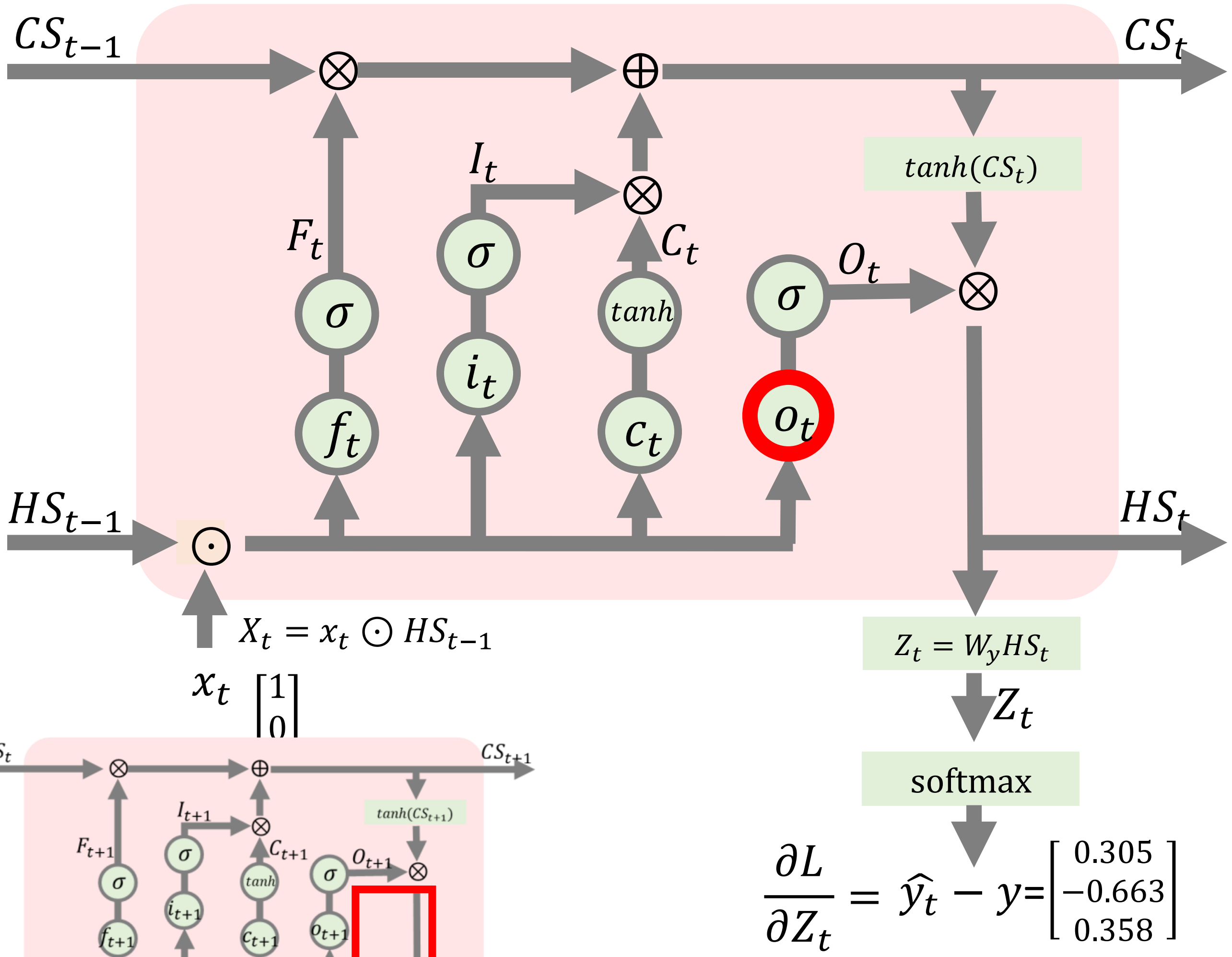
Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

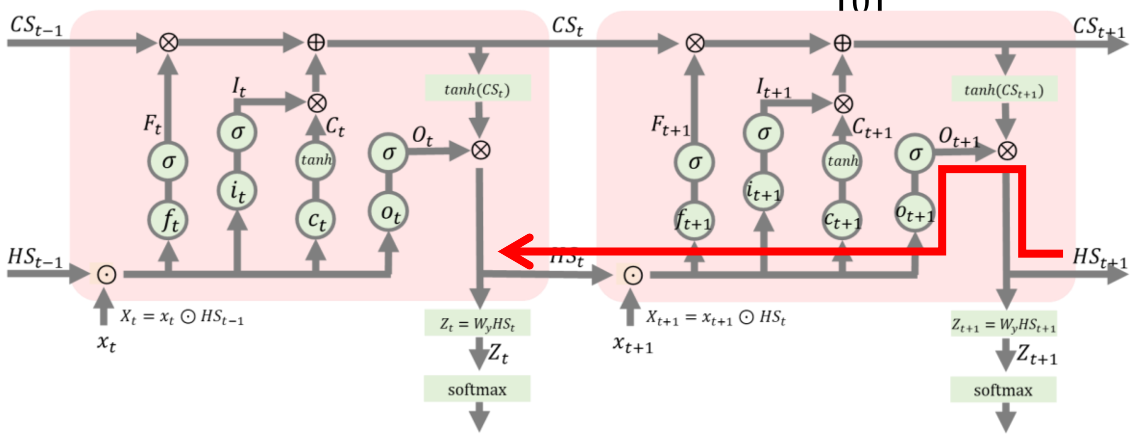
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_o} = (\hat{y}_t - y) W_y \frac{\partial HS_t}{\partial O_t} \frac{\partial O_t}{\partial o_t} \frac{\partial o_t}{\partial W_o}$$

$$\frac{\partial L}{\partial HS_t} = \frac{\partial L}{\partial Z_t} \frac{\partial Z_t}{\partial HS_t} + dHS_{t+1}$$





그래서 이 부분은 다음 영상인 실제 LSTM코드를 구현할 때 코드와 함께 설명을 드리도록 하겠습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

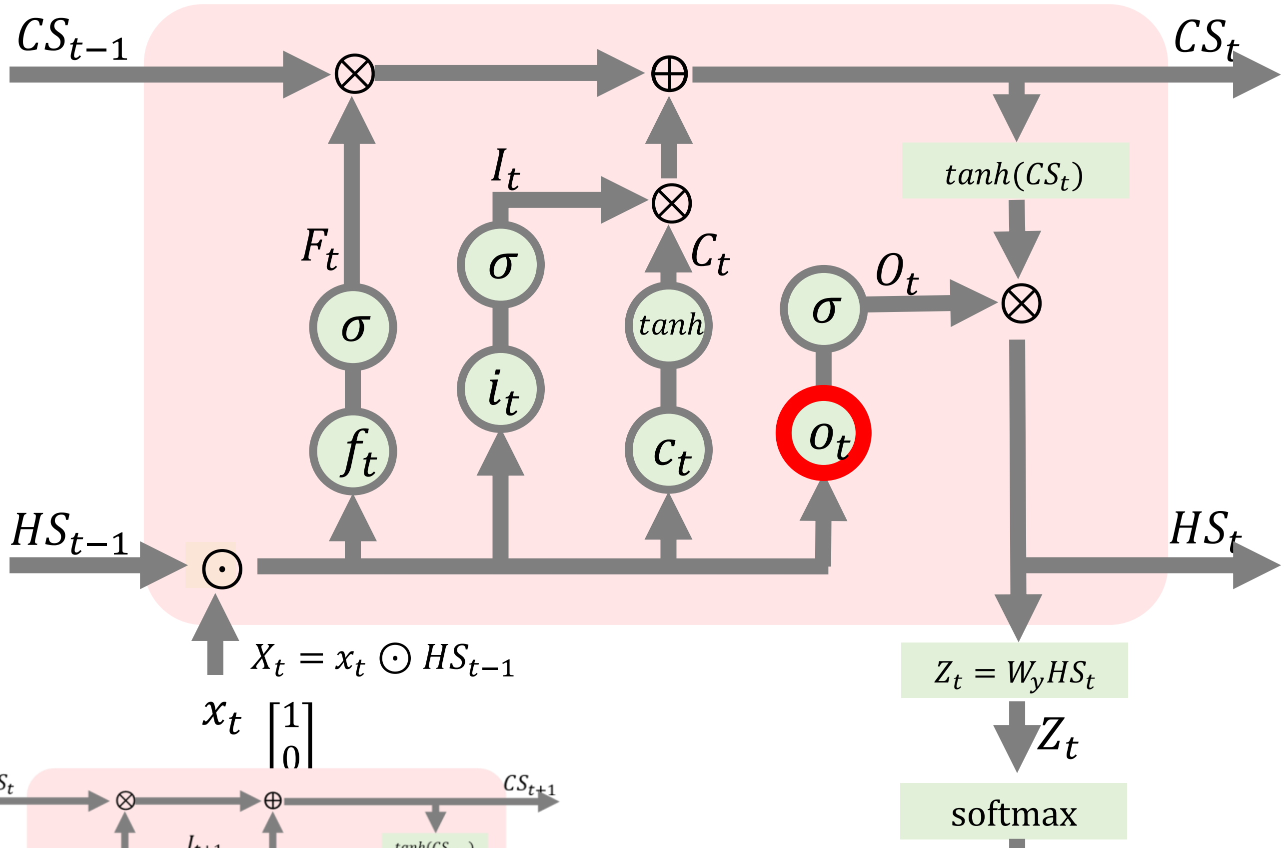
Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

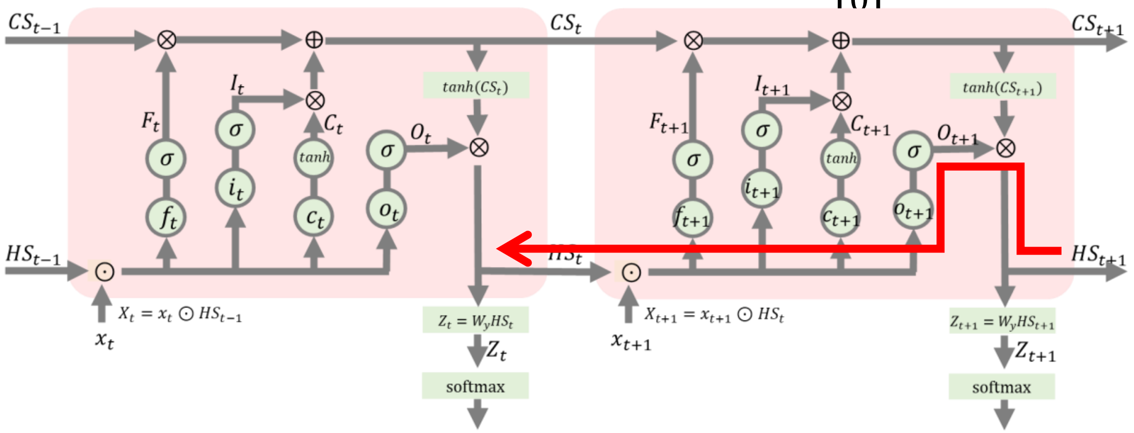
$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_o} = (\hat{y}_t - y) W_y \frac{\partial HS_t}{\partial O_t} \frac{\partial O_t}{\partial o_t} \frac{\partial o_t}{\partial W_o}$$

$$\frac{\partial L}{\partial HS_t} = \frac{\partial L}{\partial Z_t} \frac{\partial Z_t}{\partial HS_t} + dHS_{t+1}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix} 0.305 \\ -0.663 \\ 0.358 \end{bmatrix}$$



신박AI

이점 양해 부탁드립니다 ^^;;

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

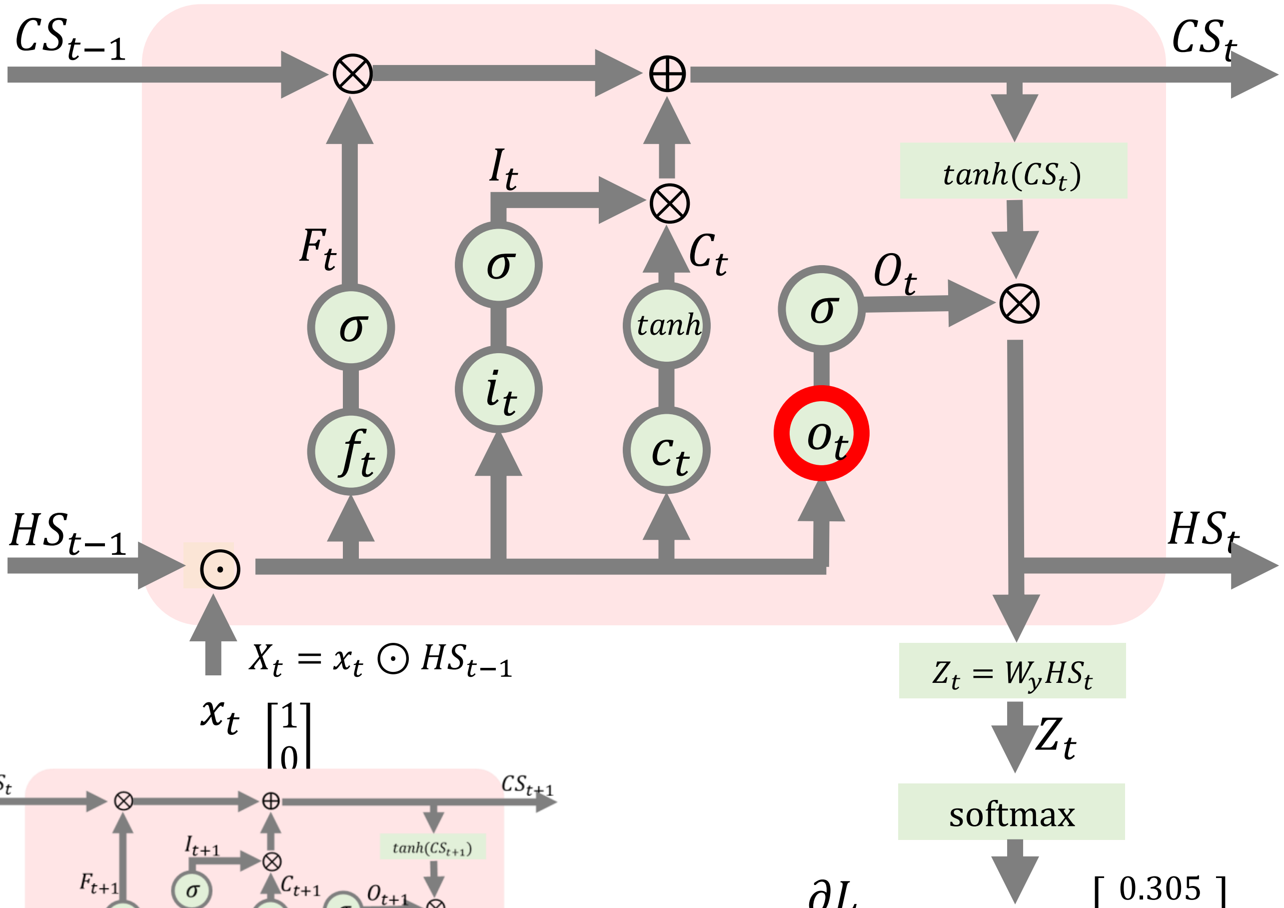
Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

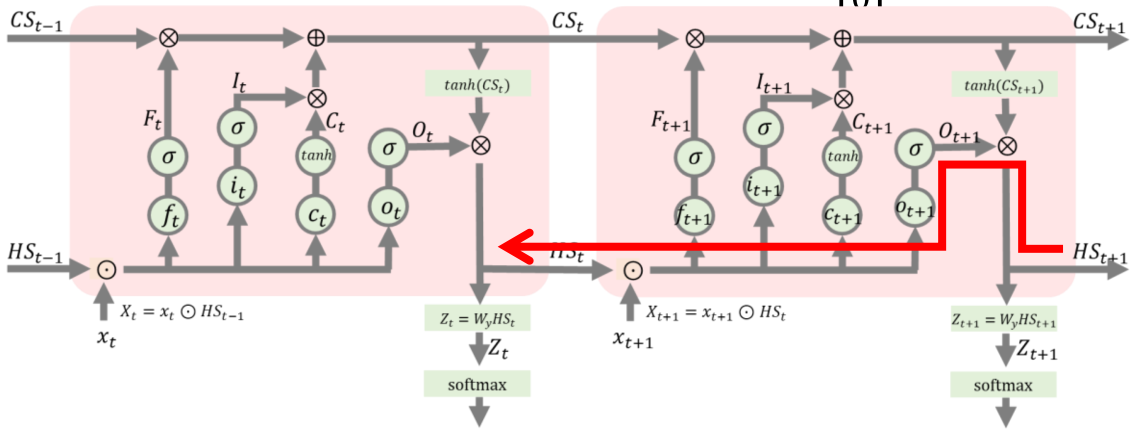
$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_o} = (\hat{y}_t - y) W_y \frac{\partial HS_t}{\partial O_t} \frac{\partial O_t}{\partial o_t} \frac{\partial o_t}{\partial W_o}$$

$$\frac{\partial L}{\partial HS_t} = \frac{\partial L}{\partial Z_t} \frac{\partial Z_t}{\partial HS_t} + dHS_{t+1}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix} 0.305 \\ -0.663 \\ 0.358 \end{bmatrix}$$



신박AI

자 그래서 이제는 $\partial HS_t / \partial O_t$을 구해보겠습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

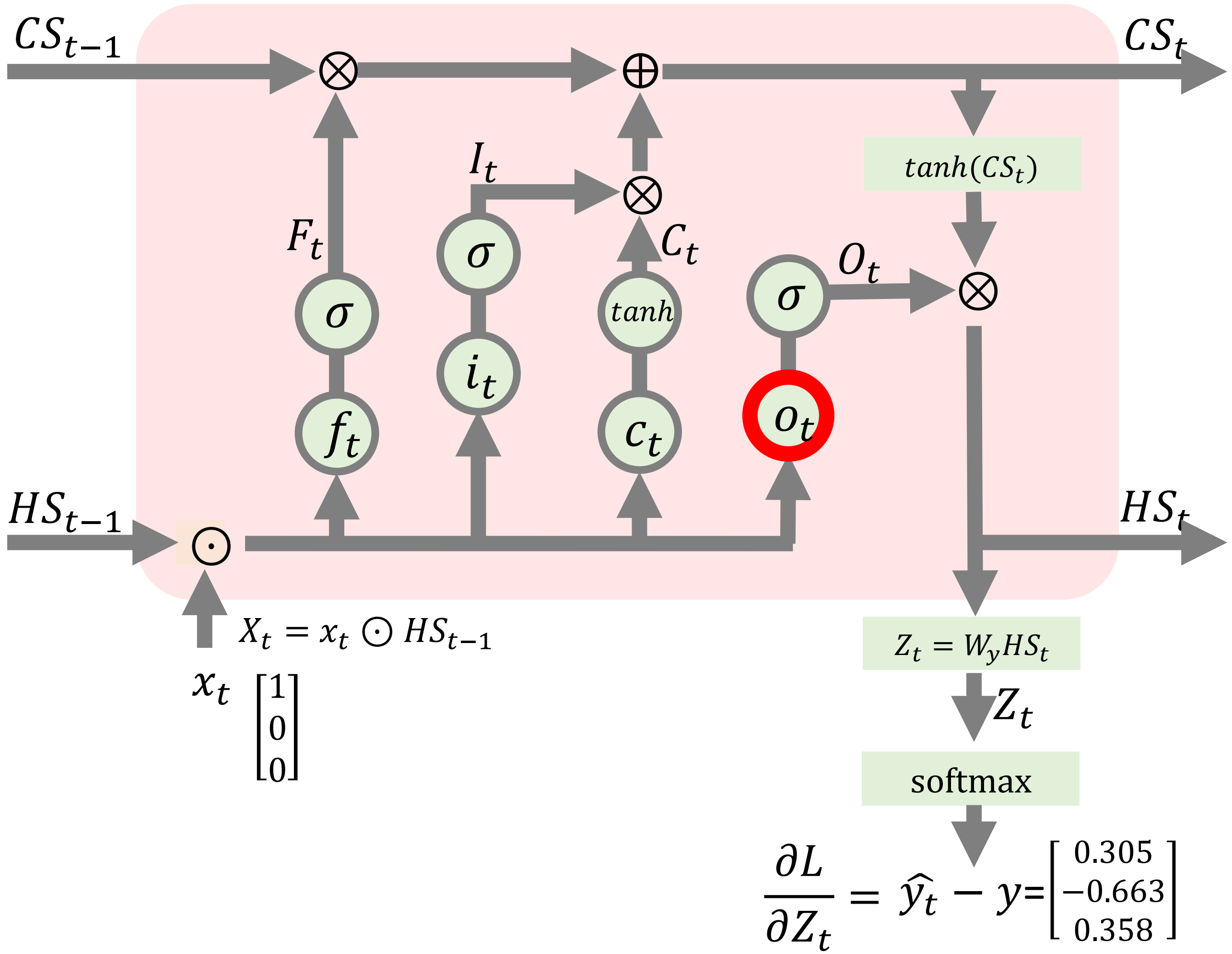
Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_o} = (\hat{y}_t - y)W_y \boxed{\frac{\partial HS_t}{\partial O_t}} \frac{\partial O_t}{\partial o_t} \frac{\partial o_t}{\partial W_o}$$



신박AI

$\partial HS_t/\partial O_t$을 구하는 공식은 다음과 같습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

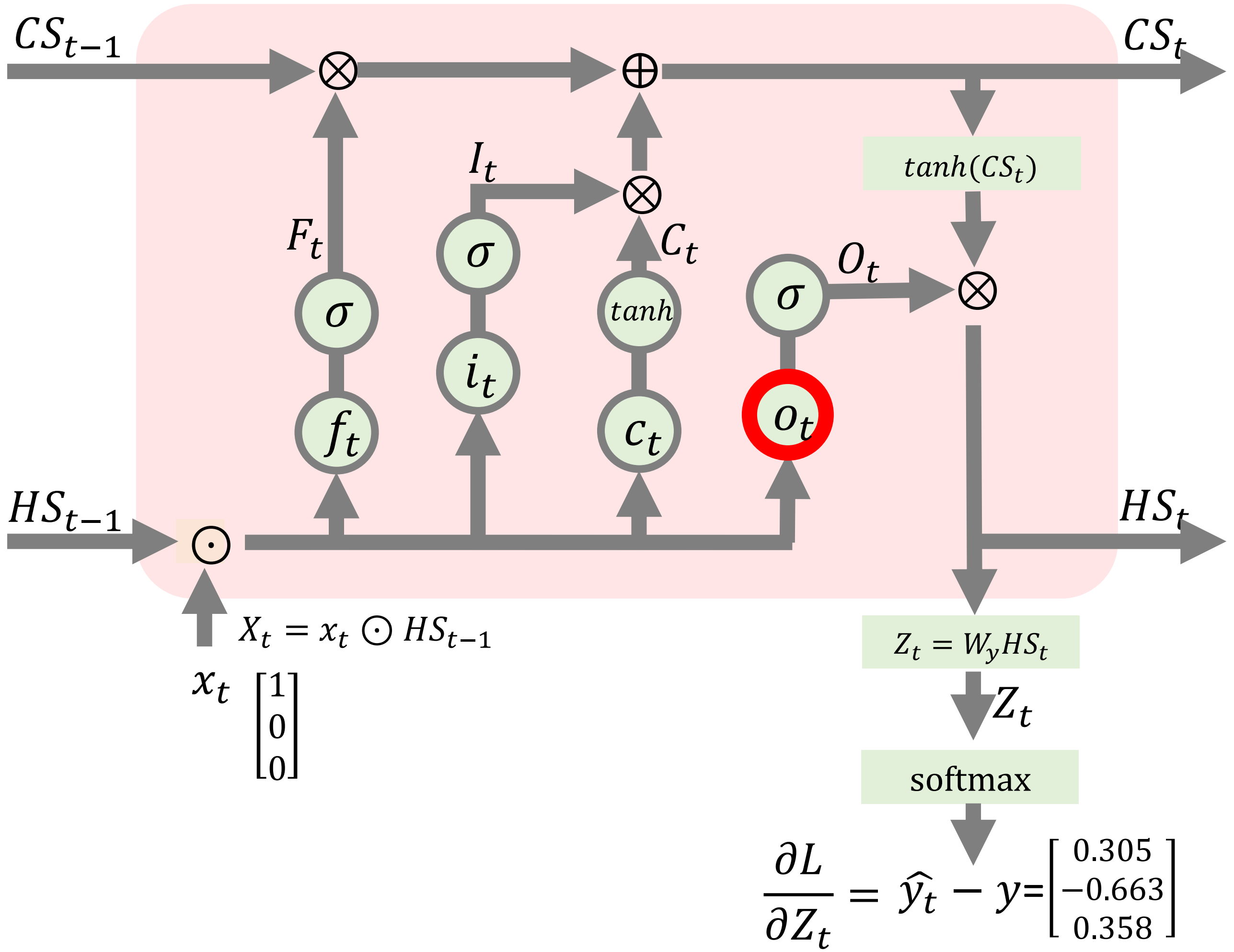
Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_o} = (\hat{y}_t - y)W_y \frac{\partial HS_t}{\partial O_t}\frac{\partial O_t}{\partial o_t}\frac{\partial o_t}{\partial W_o}$$

$$HS_t = O_t \otimes tanh(CS_t)$$



신박AI

여기서 ⊗는 element-wise곱입니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

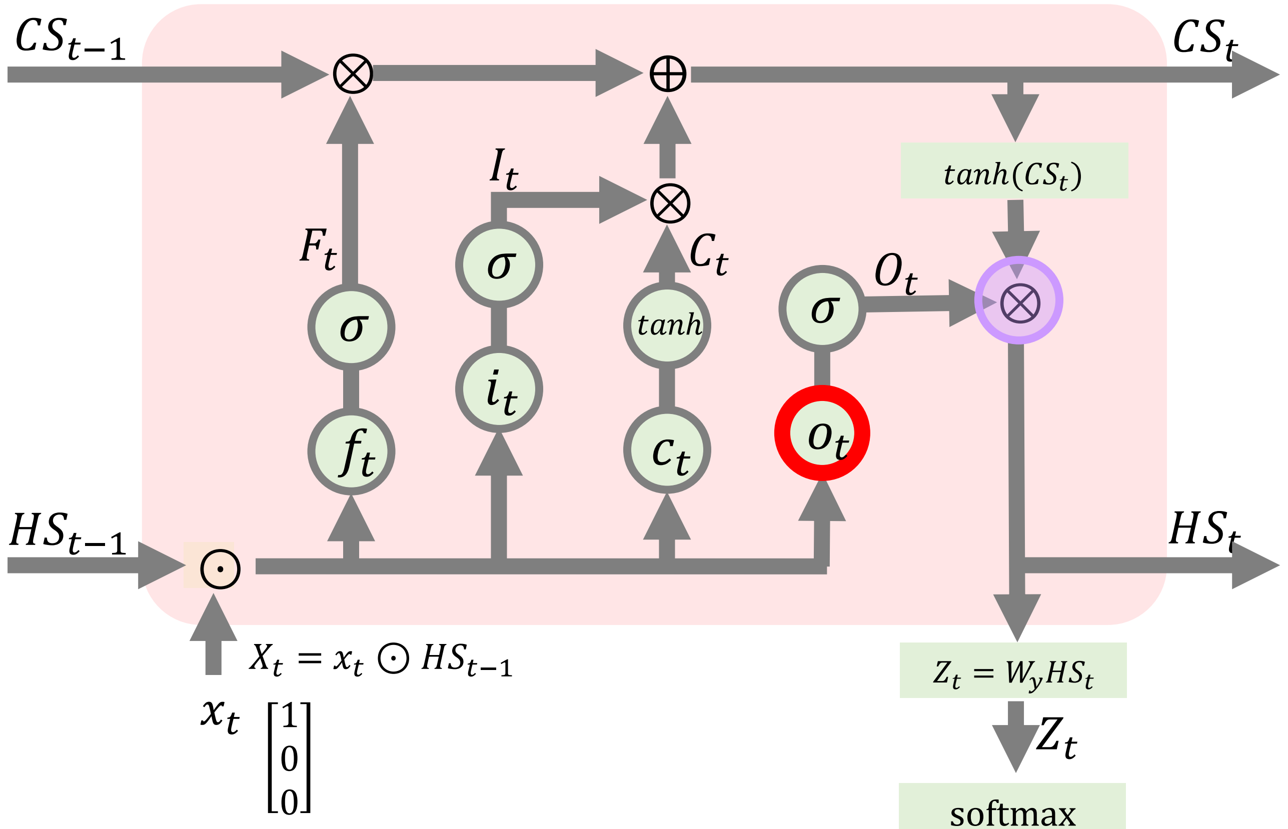
Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_o} = (\hat{y}_t - y)W_y \frac{\partial HS_t}{\partial O_t}\frac{\partial O_t}{\partial o_t}\frac{\partial o_t}{\partial W_o}$$

$$HS_t = O_t \otimes tanh(CS_t)$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix}0.305\\-0.663\\0.358\end{bmatrix}$$

신박AI

그래서 $\partial HS_t/\partial O_t$는 단순히 $tanh(CS_t)$가 됩니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

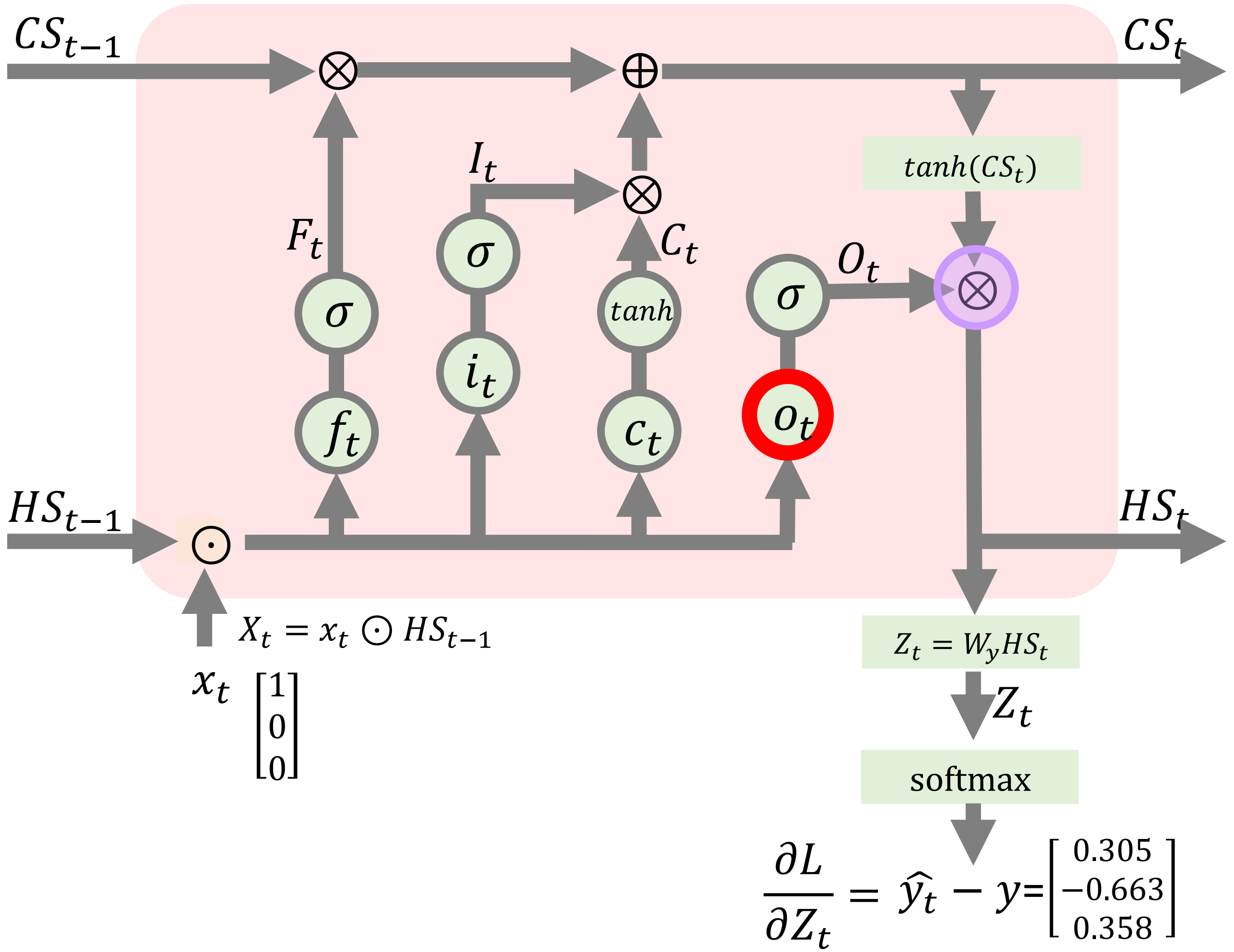
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_o} = (\hat{y}_t - y) W_y \frac{\partial HS_t}{\partial O_t} \frac{\partial O_t}{\partial o_t} \frac{\partial o_t}{\partial W_o}$$

$$HS_t = O_t \otimes tanh(CS_t)$$

$$\frac{\partial HS_t}{\partial O_t} = tanh(CS_t)$$



신박AI

왜냐하면 element-wise곱은 단순한 곱셈의 행렬 형태이기 때문입니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

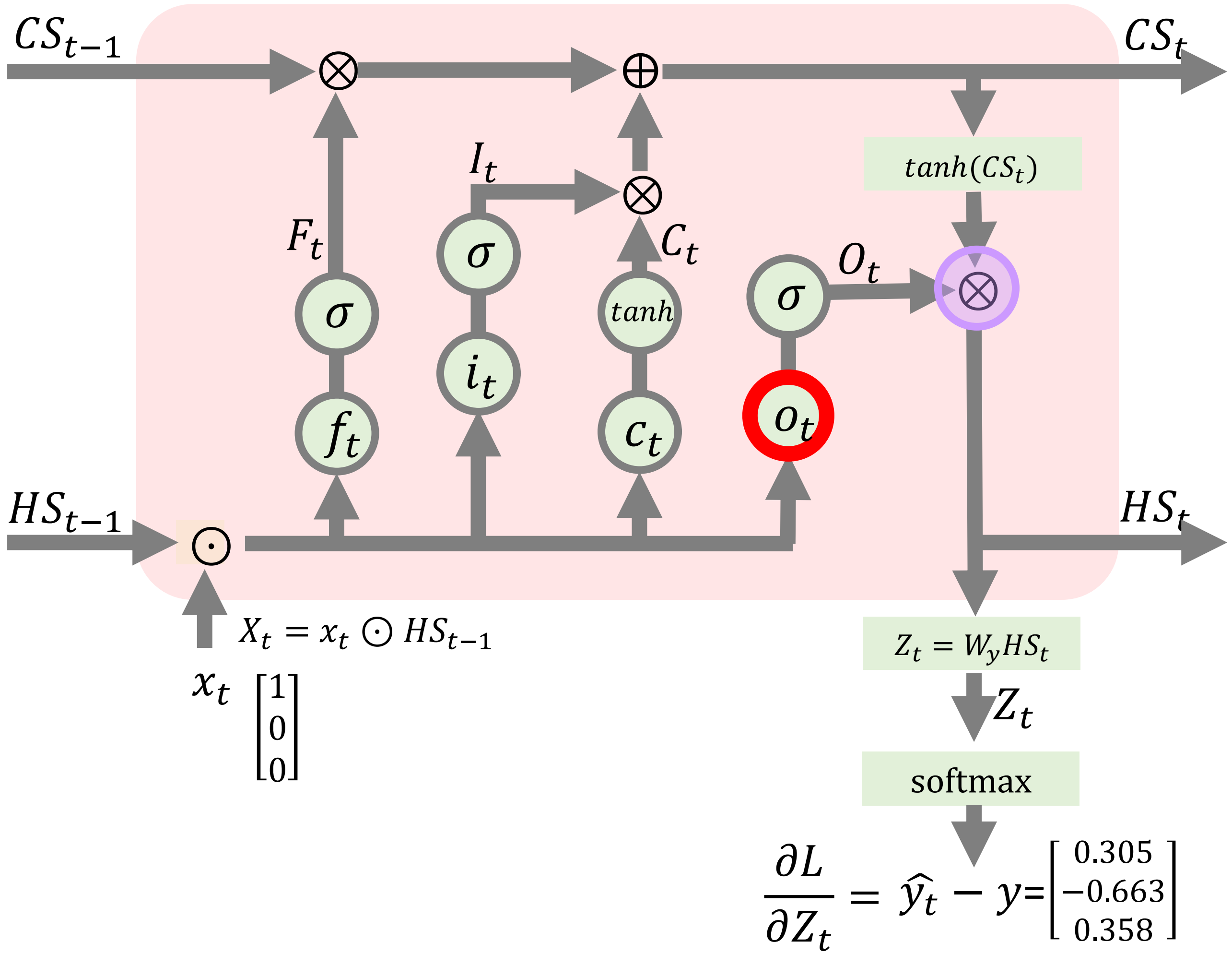
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_o} = (\hat{y}_t - y) W_y \frac{\partial HS_t}{\partial O_t} \frac{\partial O_t}{\partial o_t} \frac{\partial o_t}{\partial W_o}$$

$$HS_t = O_t \otimes tanh(CS_t)$$

$$\frac{\partial HS_t}{\partial O_t} = tanh(CS_t)$$



신박AI

예를들어 element-wise 곱을 원소별로 이렇게 표현할 수 있고

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

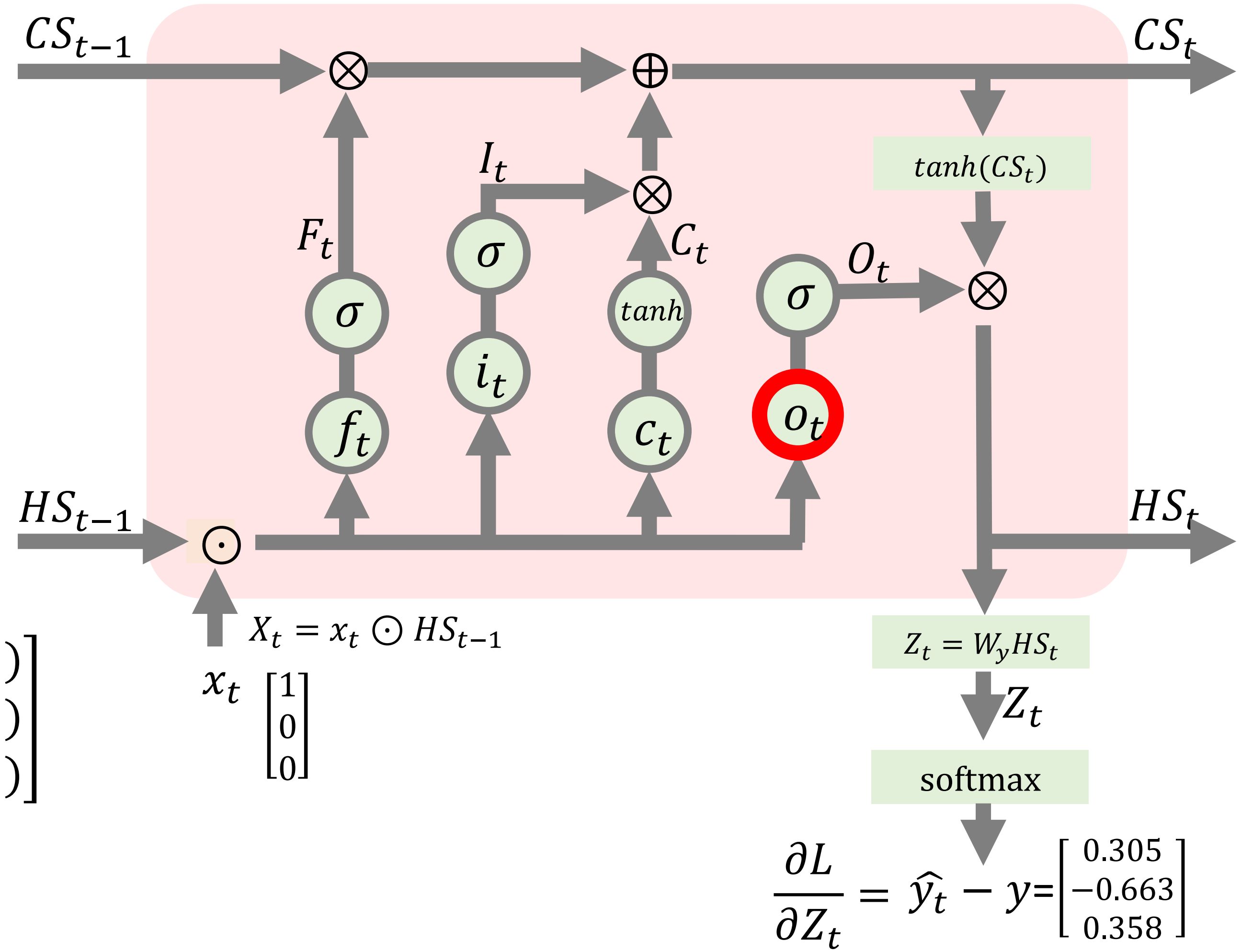
$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_o} = (\hat{y}_t - y) W_y \frac{\partial HS_t}{\partial O_t} \frac{\partial O_t}{\partial o_t} \frac{\partial o_t}{\partial W_o}$$

$$HS_t = O_t \otimes tanh(CS_t)$$

$$\frac{\partial HS_t}{\partial O_t} = tanh(CS_t)$$

$$\begin{bmatrix} HS_{t1} \\ HS_{t2} \\ HS_{t3} \end{bmatrix} = \begin{bmatrix} O_{t1} \times \tanh(CS_{t1}) \\ O_{t2} \times \tanh(CS_{t2}) \\ O_{t3} \times \tanh(CS_{t3}) \end{bmatrix}$$



신박AI

이 식을 $\partial/\partial O$로 편미분해보면 다음과 같이 되며,

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

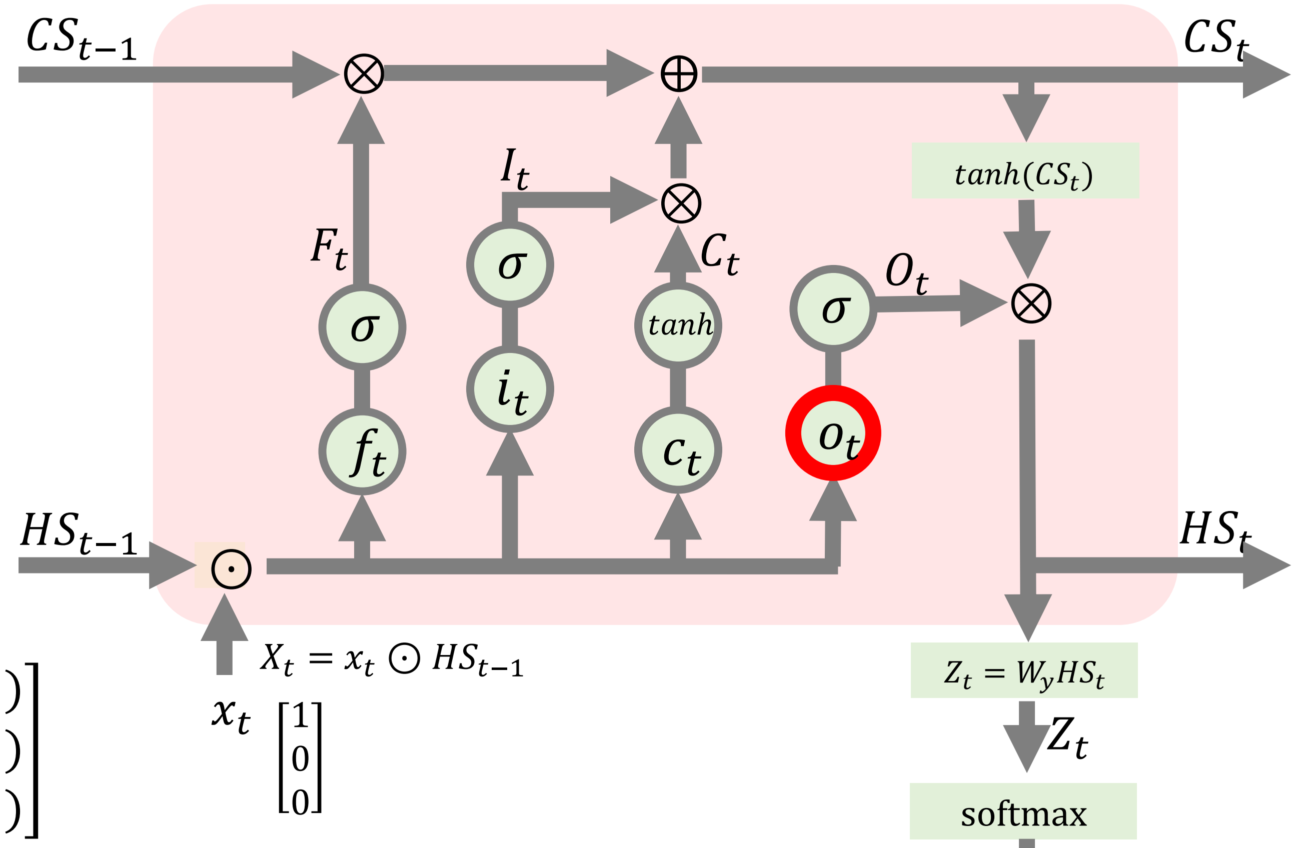
$$\frac{\partial L}{\partial W_o} = (\hat{y}_t - y)W_y \frac{\partial HS_t}{\partial O_t}\frac{\partial O_t}{\partial o_t}\frac{\partial o_t}{\partial W_o}$$

$$HS_t = O_t \otimes tanh(CS_t)$$

$$\frac{\partial HS_t}{\partial O_t} = tanh(CS_t)$$

$$\begin{bmatrix} HS_{t1} \\ HS_{t2} \\ HS_{t3} \end{bmatrix} = \begin{bmatrix} O_{t1} \times \tanh(CS_{t1}) \\ O_{t2} \times \tanh(CS_{t2}) \\ O_{t3} \times \tanh(CS_{t3}) \end{bmatrix}$$

$$\begin{bmatrix} HS_{t1}/\partial O_{t1} \\ HS_{t2}/\partial O_{t2} \\ HS_{t3}/\partial O_{t3} \end{bmatrix} = \begin{bmatrix} \partial/\partial O_{t1}(O_{t1} \times \tanh(CS_{t1})) \\ \partial/\partial O_{t2}(O_{t2} \times \tanh(CS_{t2})) \\ \partial/\partial O_{t3}(O_{t3} \times \tanh(CS_{t3})) \end{bmatrix}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix} 0.305 \\ -0.663 \\ 0.358 \end{bmatrix}$$

신박AI

그래서 다음과 같이 $\partial HS_t/\partial O_t$는 $tanh(CS_t)$가 됨을 확인하실 수 있습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_o} = (\hat{y}_t - y)W_y \frac{\partial HS_t}{\partial O_t}\frac{\partial O_t}{\partial o_t}\frac{\partial o_t}{\partial W_o}$$

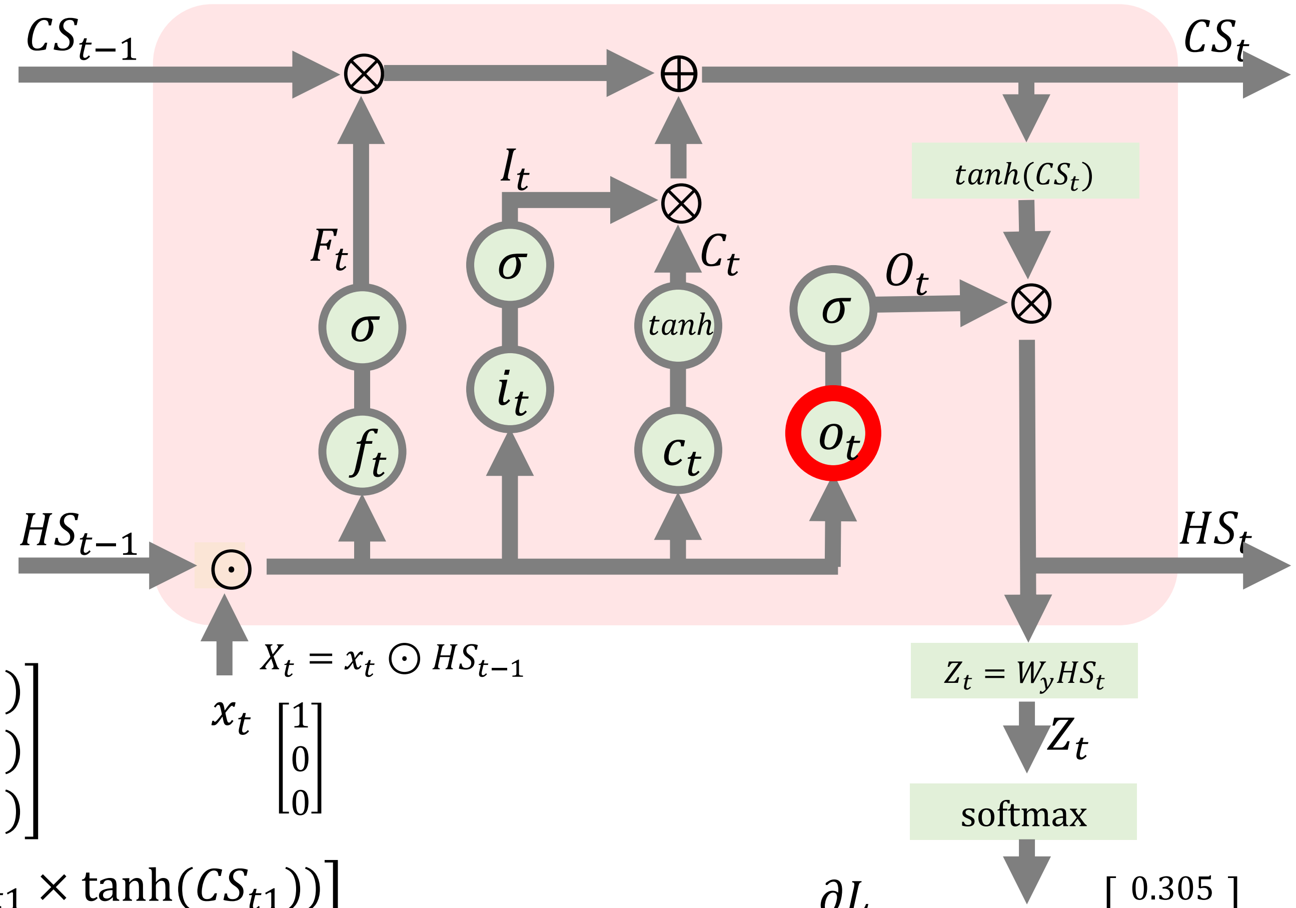
$$HS_t = O_t \otimes tanh(CS_t)$$

$$\boxed{\frac{\partial HS_t}{\partial O_t} = tanh(CS_t)}$$

$$\begin{bmatrix} HS_{t1} \\ HS_{t2} \\ HS_{t3} \end{bmatrix} = \begin{bmatrix} O_{t1} \times \tanh(CS_{t1}) \\ O_{t2} \times \tanh(CS_{t2}) \\ O_{t3} \times \tanh(CS_{t3}) \end{bmatrix}$$

$$\begin{bmatrix} HS_{t1}/\partial O_{t1} \\ HS_{t2}/\partial O_{t2} \\ HS_{t3}/\partial O_{t3} \end{bmatrix} = \begin{bmatrix} \partial/\partial O_{t1}(O_{t1} \times \tanh(CS_{t1})) \\ \partial/\partial O_{t2}(O_{t2} \times \tanh(CS_{t2})) \\ \partial/\partial O_{t3}(O_{t3} \times \tanh(CS_{t3})) \end{bmatrix}$$

$$\begin{bmatrix} HS_{t1}/\partial O_{t1} \\ HS_{t2}/\partial O_{t2} \\ HS_{t3}/\partial O_{t3} \end{bmatrix} = \begin{bmatrix} \tanh(CS_{t1}) \\ \tanh(CS_{t2}) \\ \tanh(CS_{t3}) \end{bmatrix}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix} 0.305 \\ -0.663 \\ 0.358 \end{bmatrix}$$

신박AI

그래서 식을 다시 정리하면 다음과 같이 되고,

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

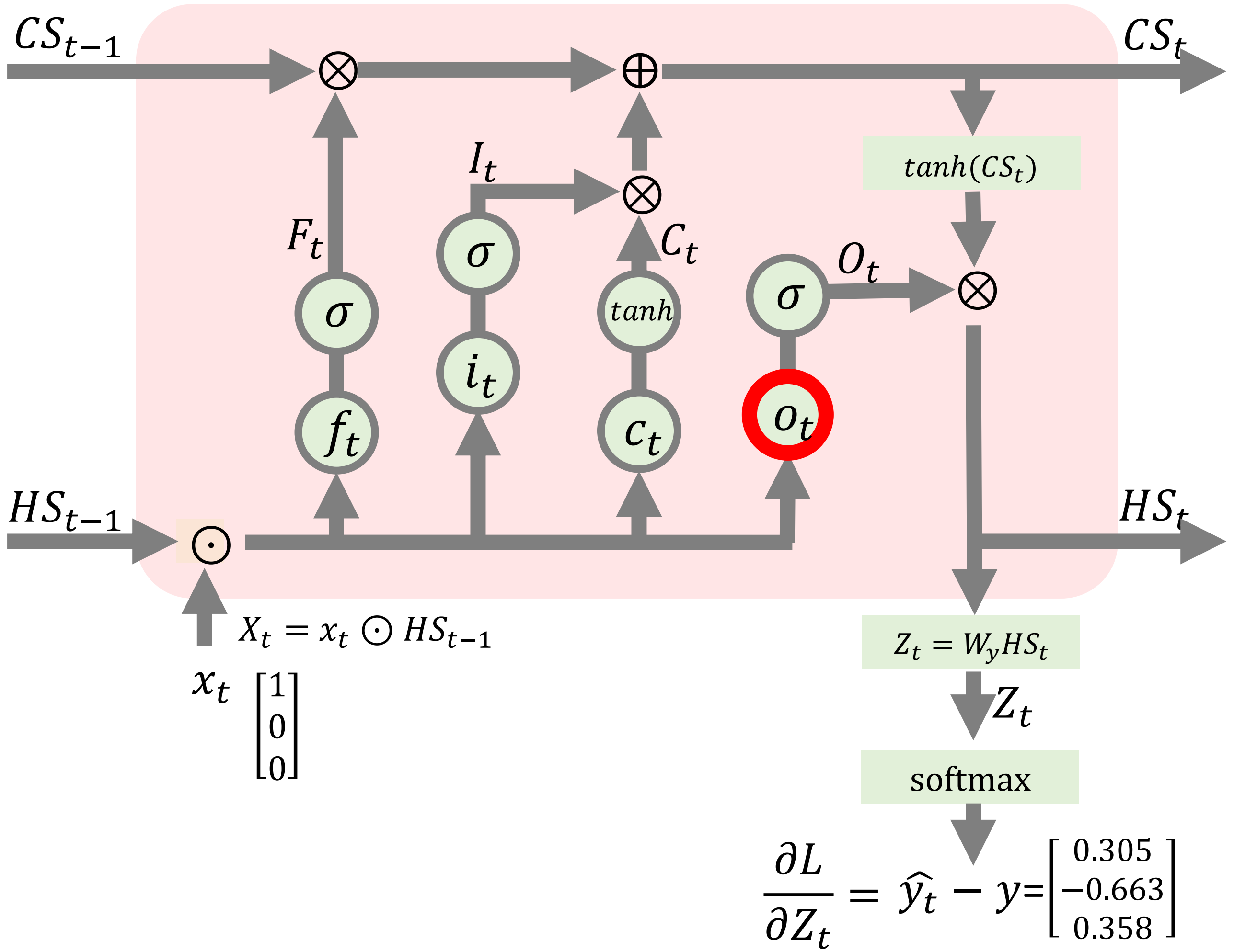
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_o} = (\hat{y}_t - y) W_y tanh(CS_t) \frac{\partial O_t}{\partial o_t} \frac{\partial o_t}{\partial W_o}$$

$$HS_t = O_t \otimes tanh(CS_t)$$

$$\boxed{\frac{\partial HS_t}{\partial O_t} = tanh(CS_t)}$$



신박AI

계속해서 $\partial O_t/\partial o_t$를 구해보도록 하겠습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

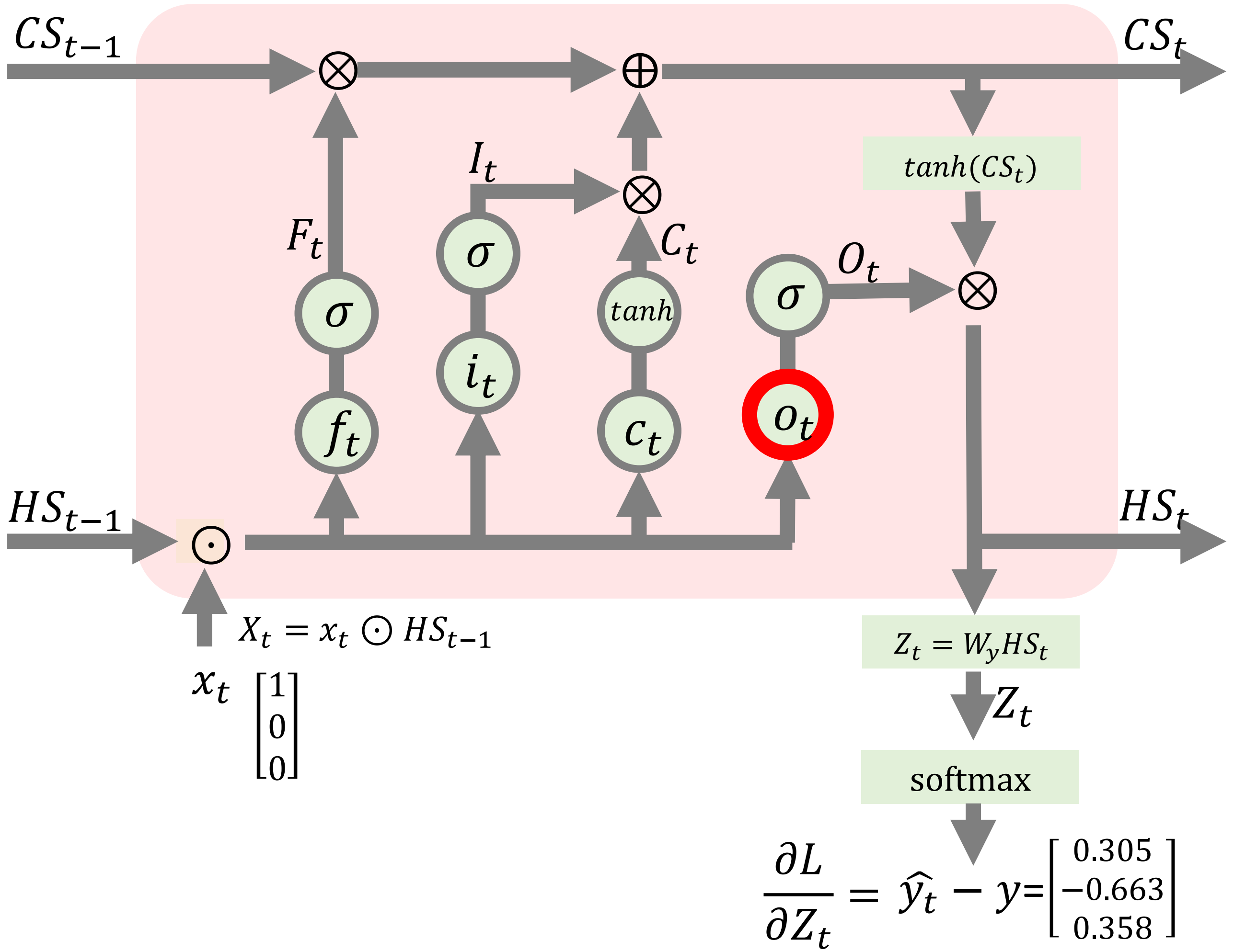
Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_o} = (\hat{y}_t - y) W_y tanh(CS_t) \frac{\partial O_t}{\partial o_t} \frac{\partial o_t}{\partial W_o}$$



신박AI

O_t 와 o_t의 관계는 이미 공식에 나와 있습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

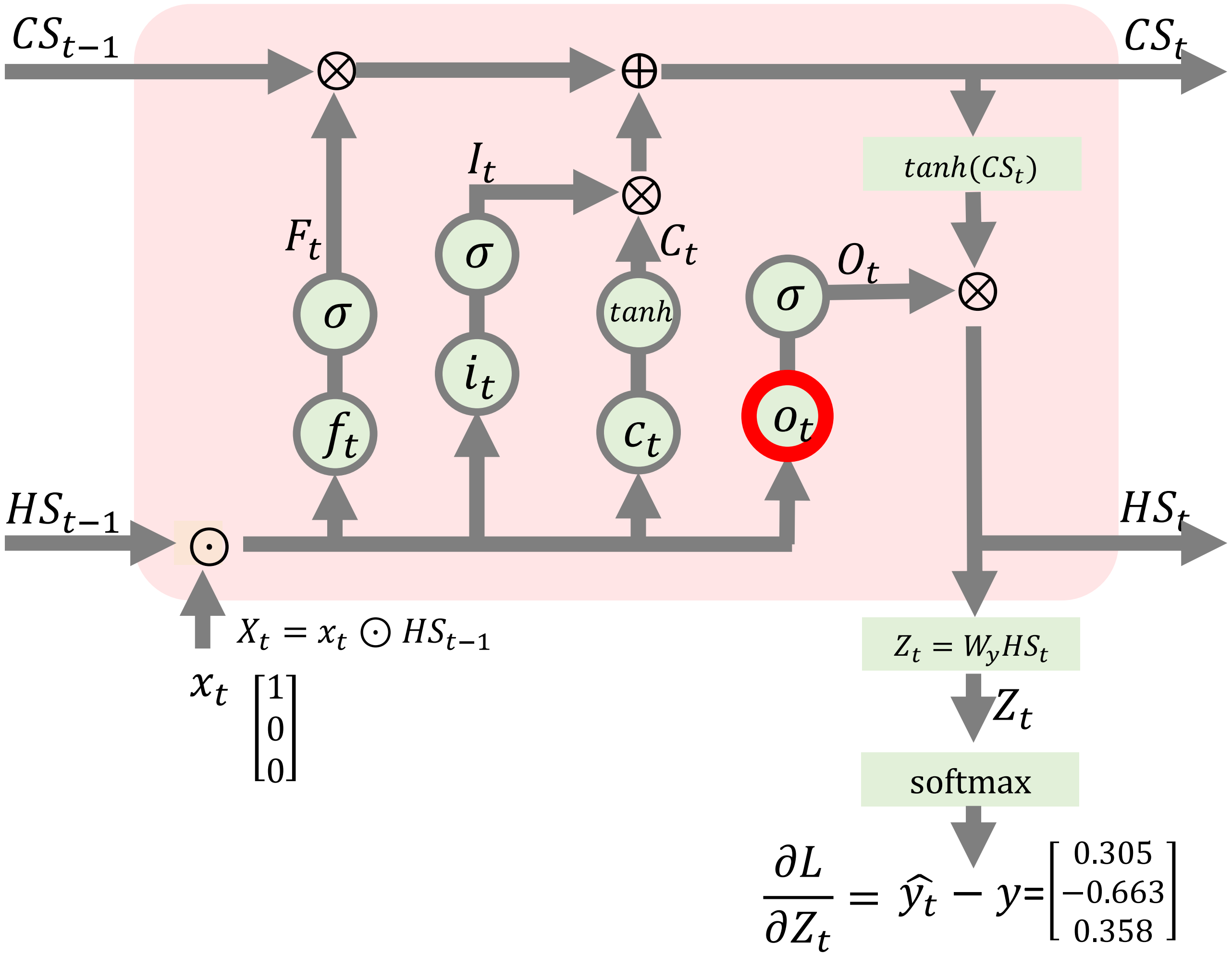
Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_o} = (\hat{y}_t - y) W_y tanh(CS_t) \frac{\partial O_t}{\partial o_t} \frac{\partial o_t}{\partial W_o}$$



신박AI

그러므로 $\partial O_t/\partial o_t$는 시그모이드 미분함수에 의해서 다음과 같습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

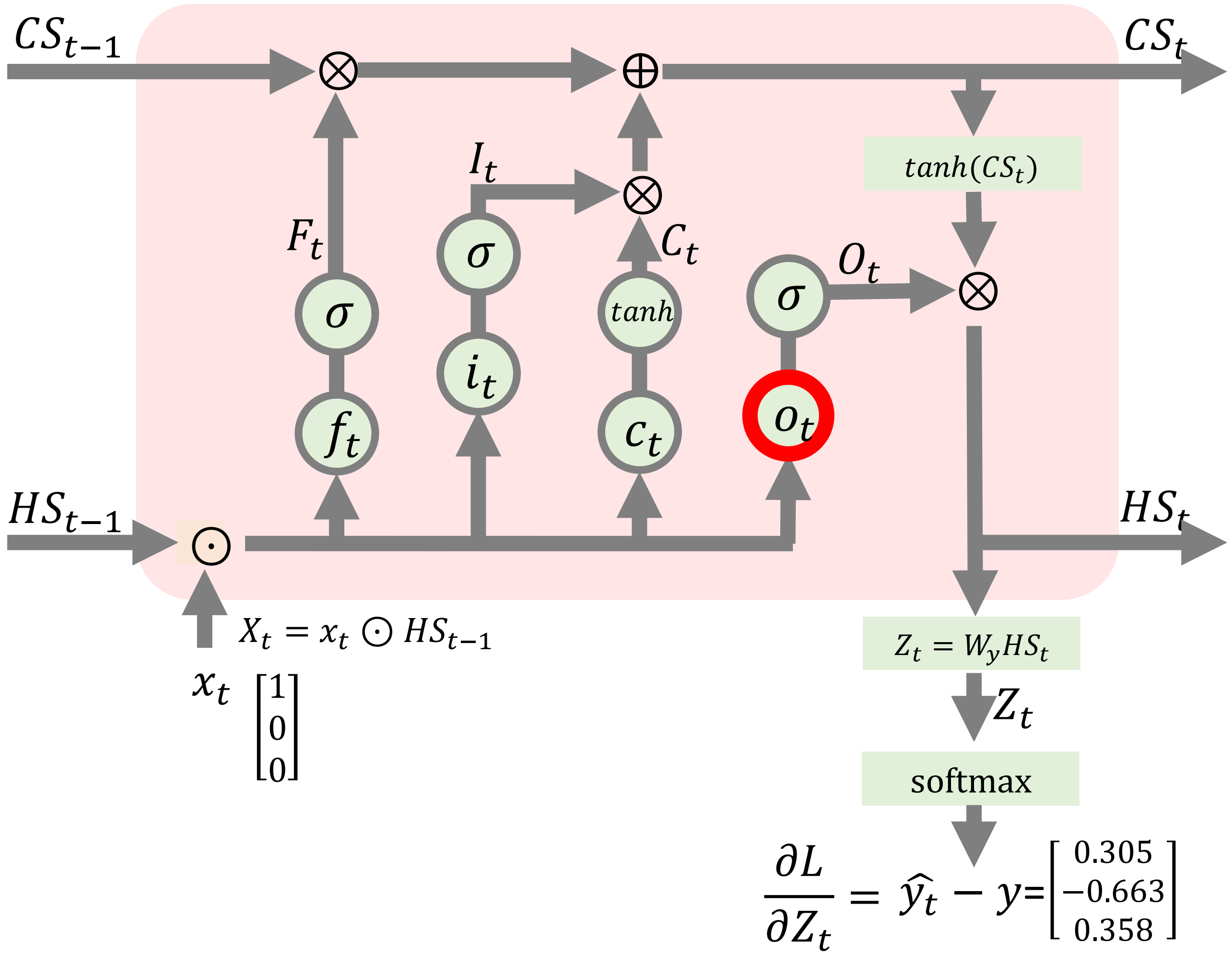
Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$\boxed{O_t = \sigma(o_t)}$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_o} = (\hat{y}_t - y) W_y tanh(CS_t) \frac{\partial O_t}{\partial o_t} \frac{\partial o_t}{\partial W_o}$$

$$\frac{\partial O_t}{\partial o_t} = O_t(1 - O_t)$$



신박AI

그래서 식을 다시 정리하면 다음과 같이 되고,

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

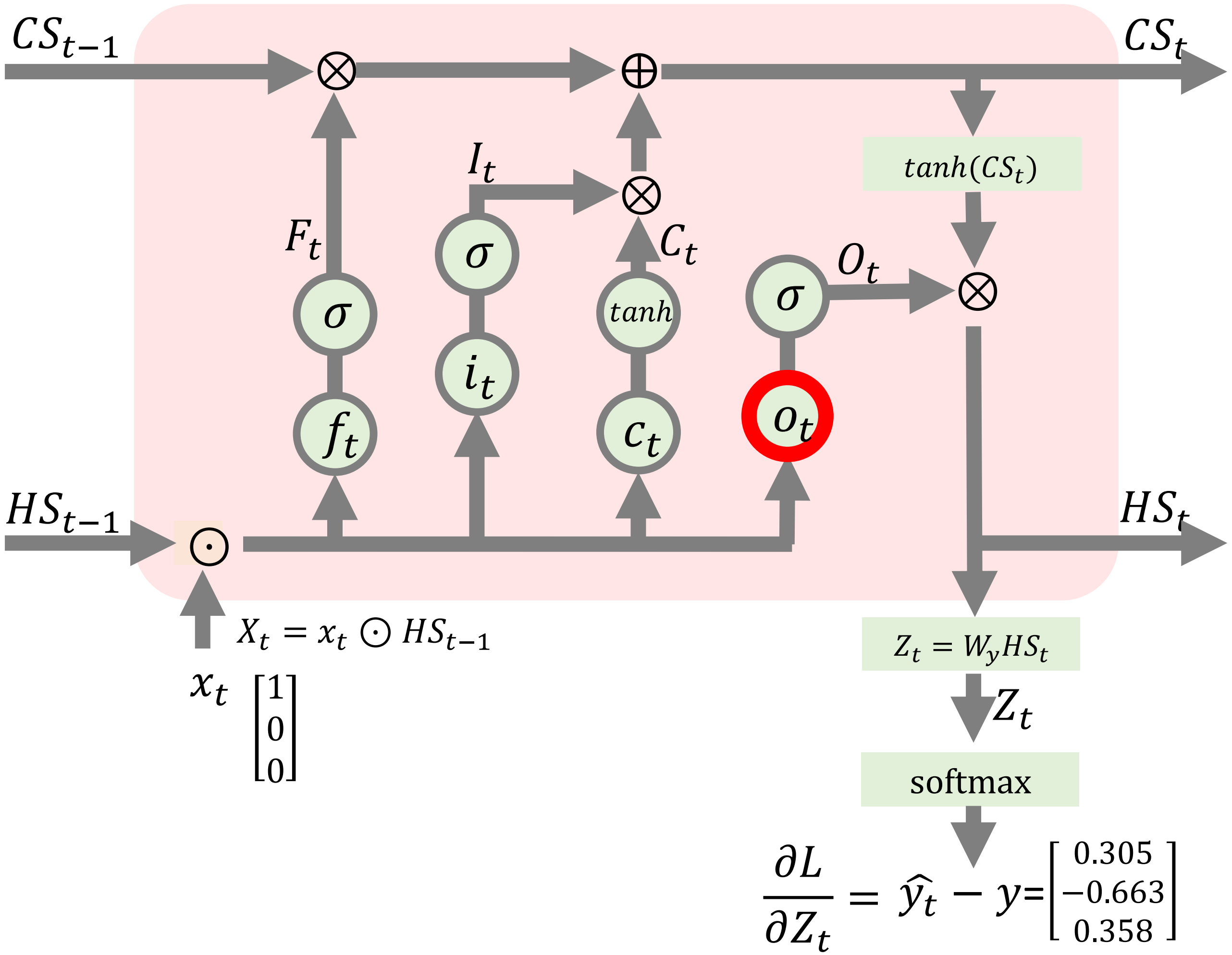
Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_o} = (\hat{y}_t - y) W_y tanh(CS_t) \boxed{O_t(1 - O_t)} \frac{\partial o_t}{\partial W_o}$$

$$\frac{\partial O_t}{\partial o_t} = O_t(1 - O_t)$$



신박AI

이제 남은 것은 $\partial o_t / \partial W_o$ 입니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

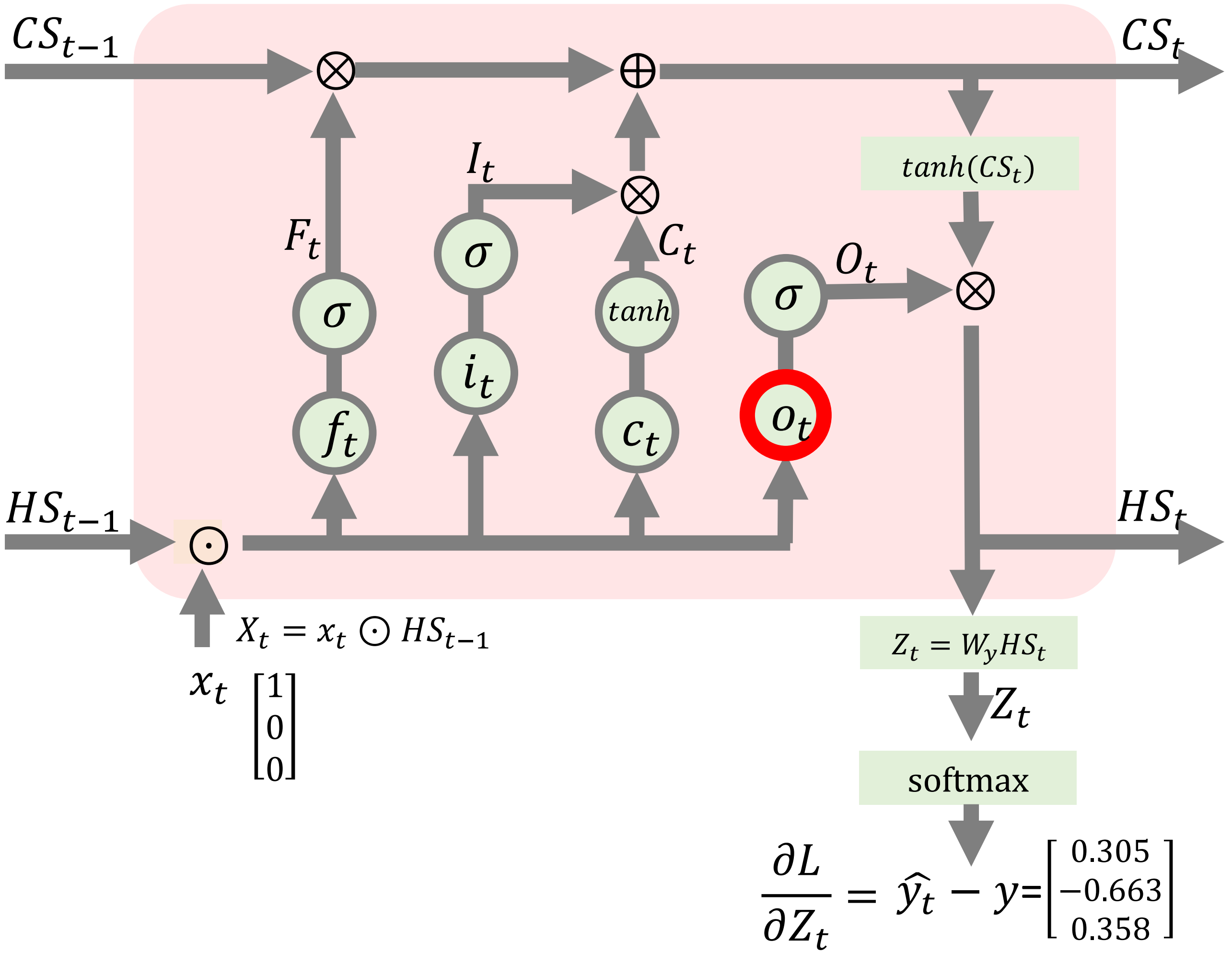
Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_o} = (\hat{y}_t - y) W_y tanh(CS_t) O_t (1 - O_t) \frac{\partial o_t}{\partial W_o}$$



신박AI

o_t 와 W_o의 관계도 공식에 나와 있습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

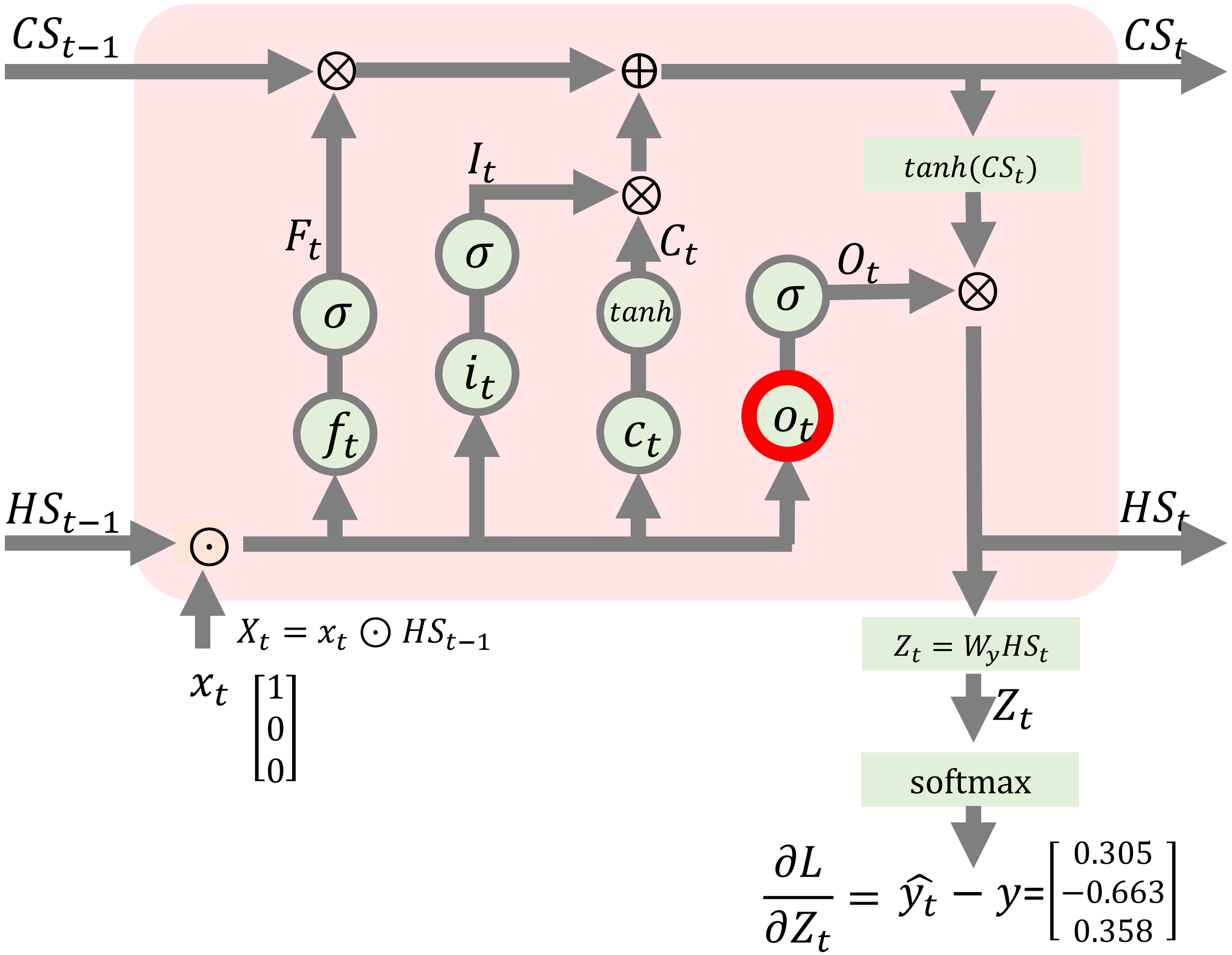
Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$\boxed{o_t = W_o X_t}$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_o} = (\hat{y}_t - y) W_y tanh(CS_t) O_t (1 - O_t) \frac{\partial o_t}{\partial W_o}$$



신박AI

그러므로 $\partial o_t/\partial W_o$는 다음과 같이 구할 수가 있습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

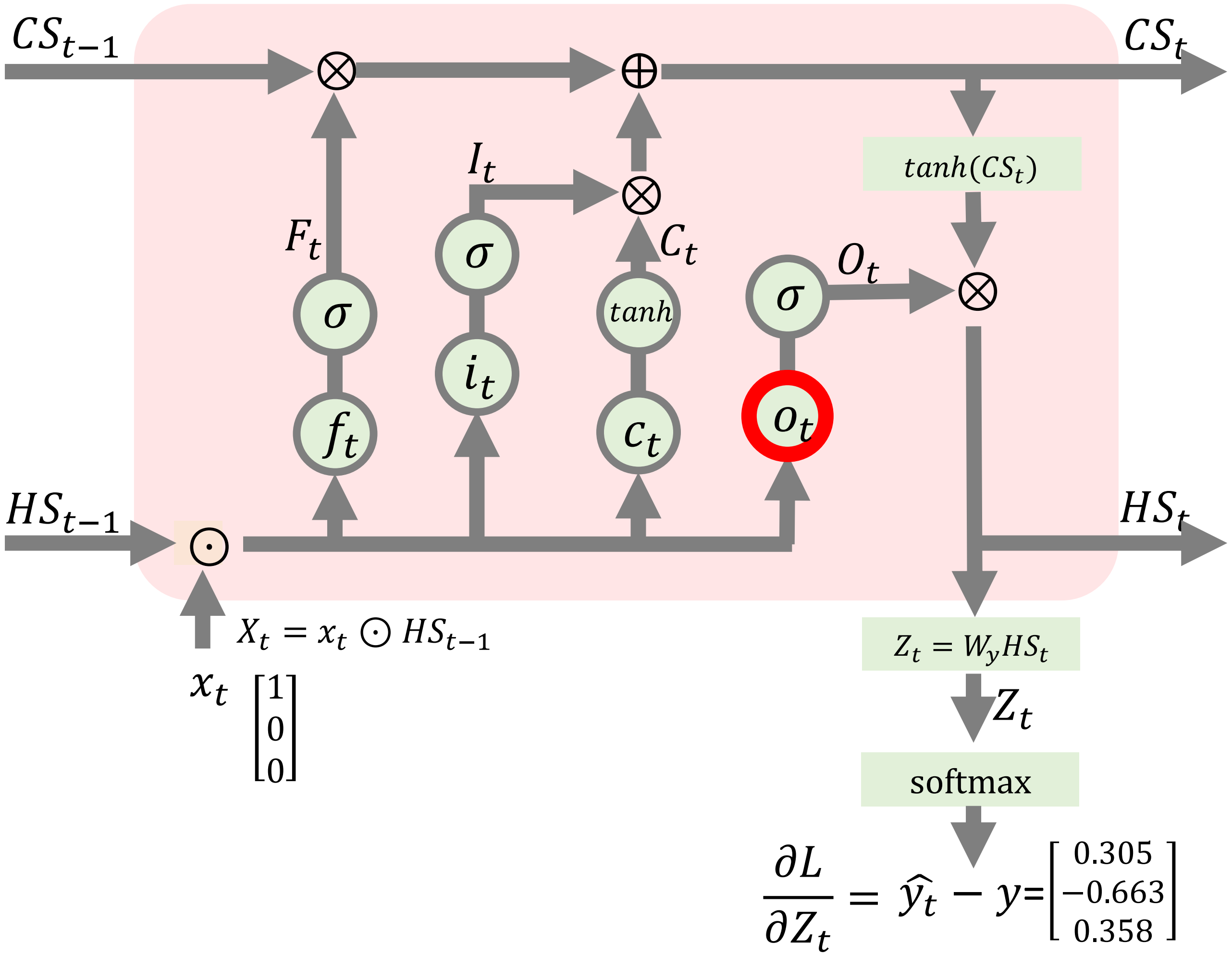
Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$\boxed{o_t = W_o X_t}$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_o} = (\hat{y}_t - y) W_y tanh(CS_t) O_t (1 - O_t) \frac{\partial o_t}{\partial W_o}$$

$$\frac{\partial o_t}{\partial W_o} = X_t$$



그래서 식을 다시 정리하면 다음과 같이 됩니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

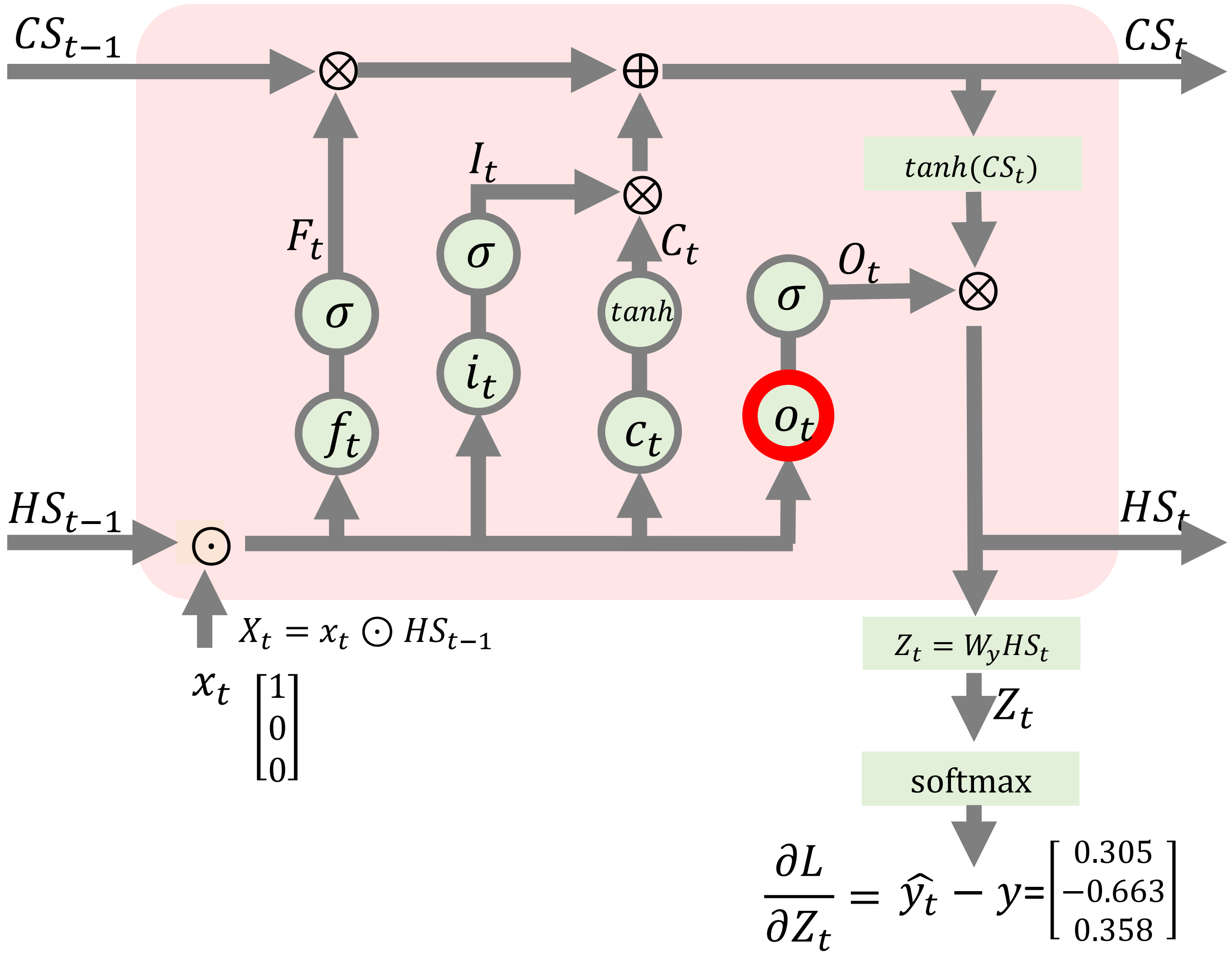
Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_o} = (\hat{y}_t - y) W_y tanh(CS_t) O_t (1 - O_t) \boxed{X_t}$$

$$\frac{\partial o_t}{\partial W_o} = X_t$$



신박AI

드디어 숫자를 넣어서 계산해 보도록 하겠습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

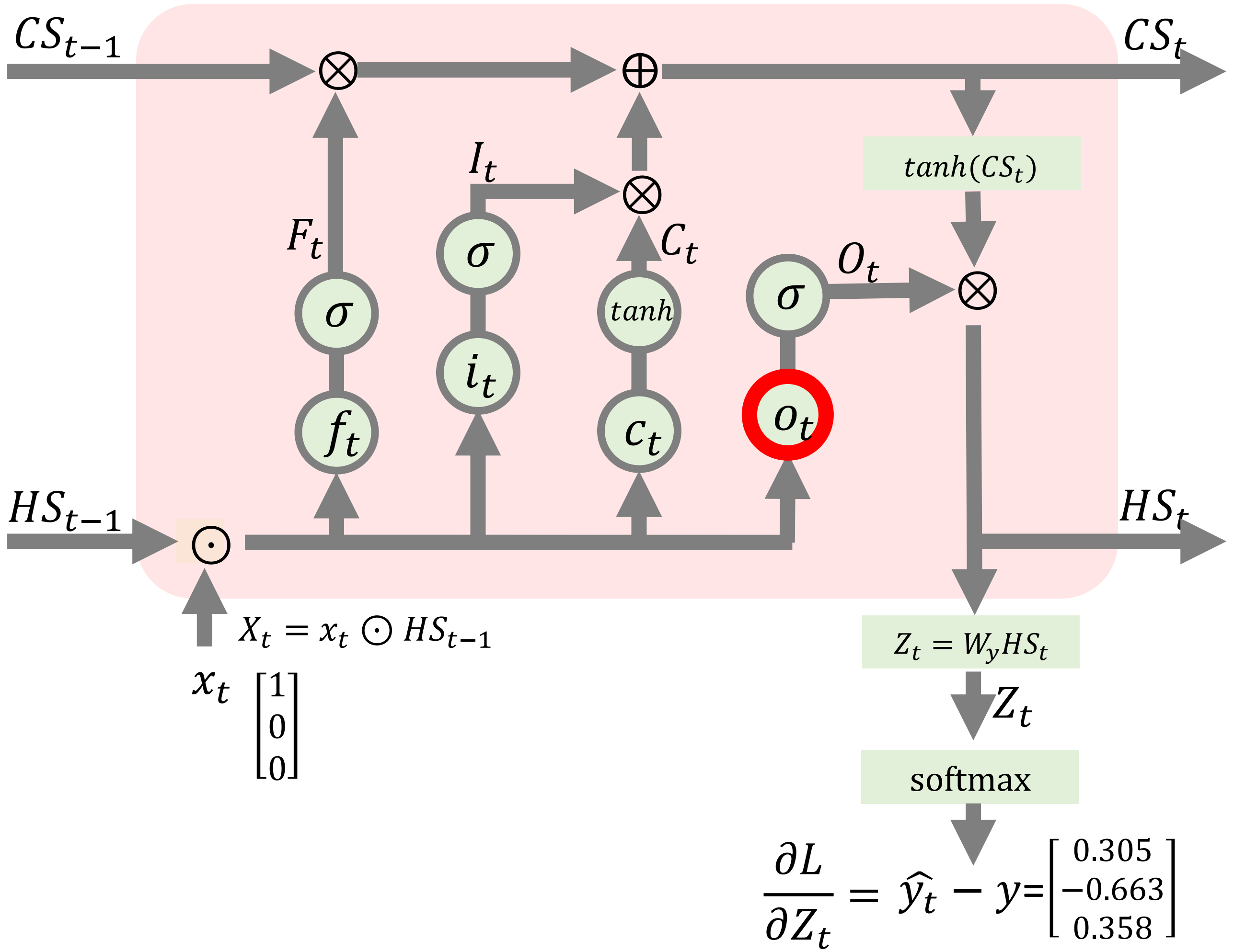
Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_o} = (\hat{y}_t - y) W_y tanh(CS_t) O_t (1 - O_t)\, X_t$$



신박AI

행렬의 차원을 맞추기 위해 숫자를 대입할 때는 약간의 변화(transpose)가 필요합니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

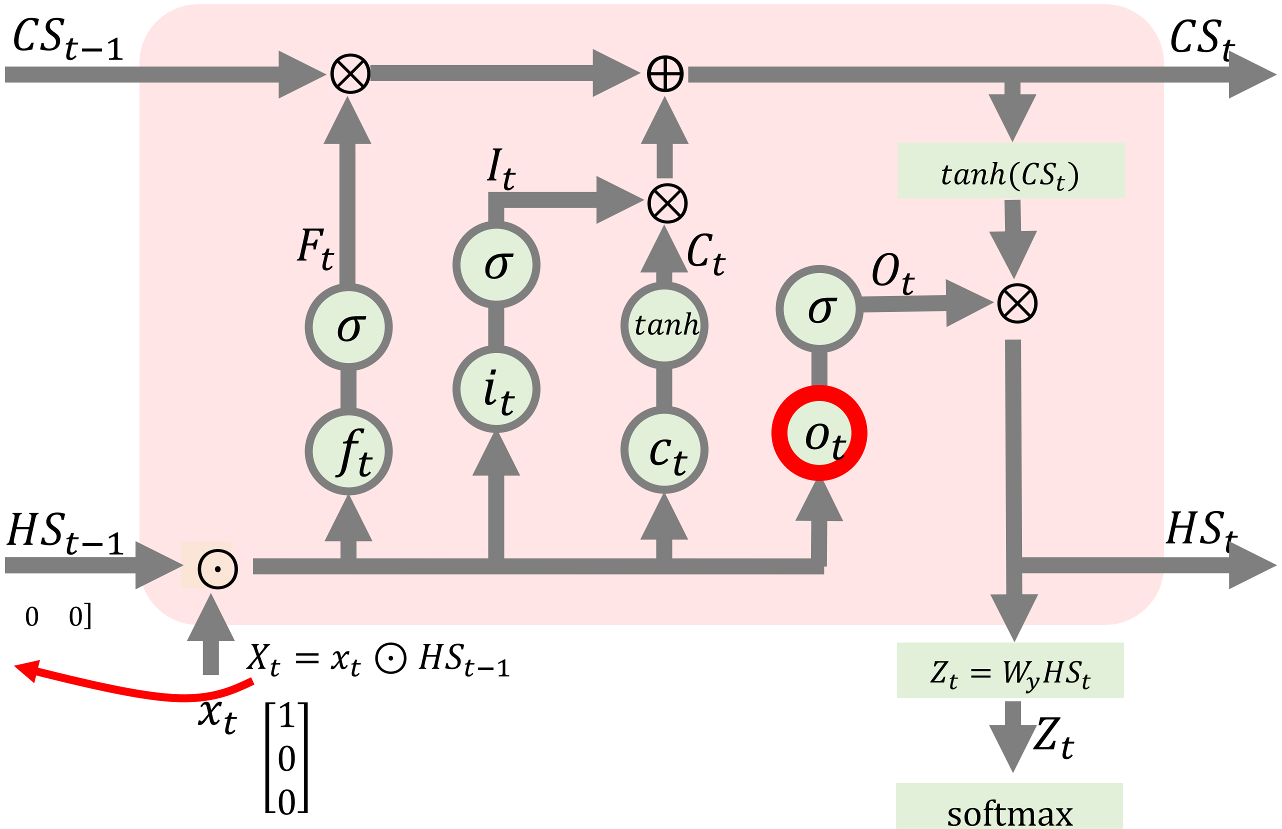
Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_o} = (\hat{y}_t - y) W_y tanh(CS_t) O_t (1 - O_t)\, X_t$$

$$= \left(\begin{bmatrix} 0.305 & -0.663 & 0.358 \end{bmatrix} \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix} \right)^T \begin{bmatrix} 0.05 \\ -0.039 \end{bmatrix} \begin{bmatrix} 0.221 \\ 0.24 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix} 0.305 \\ -0.663 \\ 0.358 \end{bmatrix}$$

드디어 $\partial L/\partial W_o$ 를 계산해보았습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

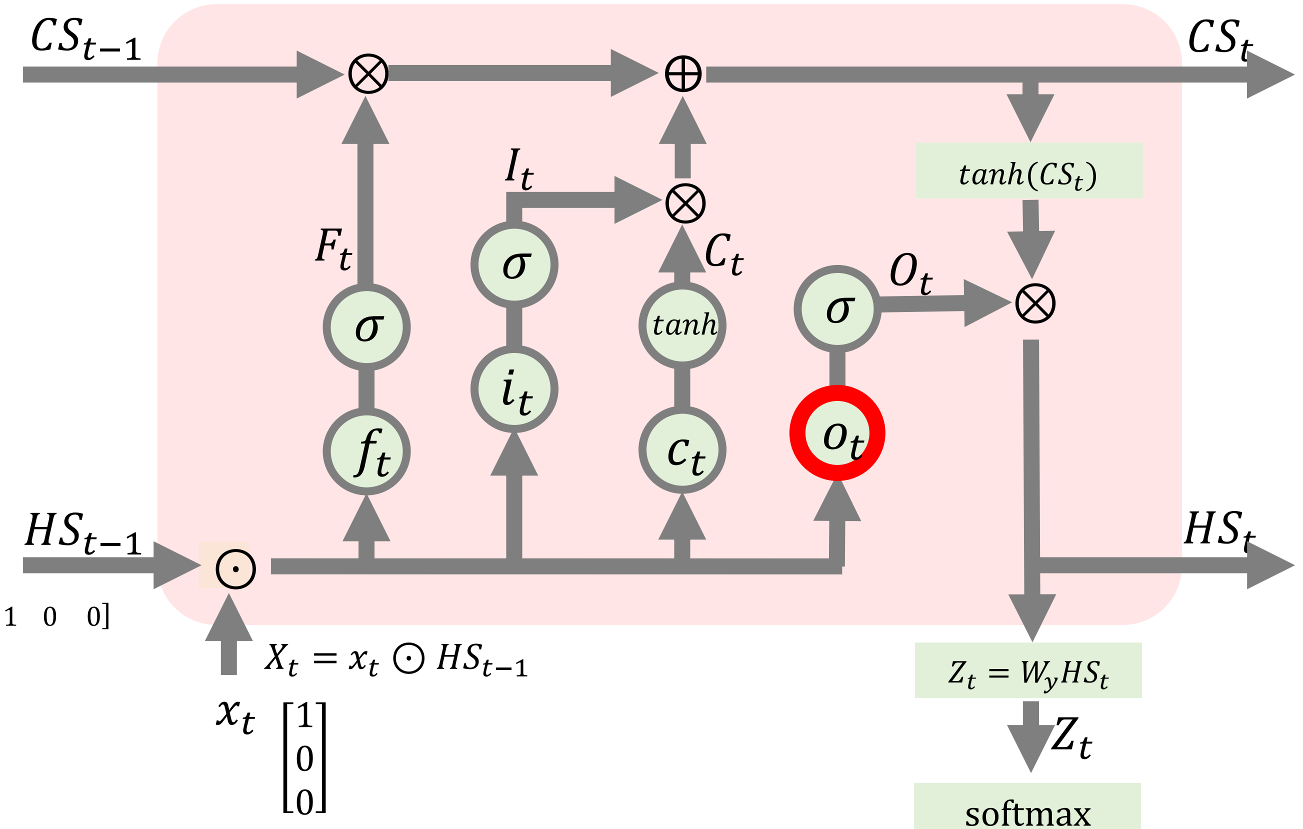
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_o} = (\hat{y}_t - y) W_y tanh(CS_t) O_t (1 - O_t)\, X_t$$

$$= \left(\begin{bmatrix} 0.305 & -0.663 & 0.358 \end{bmatrix} \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix} \right)^T \begin{bmatrix} 0.05 \\ -0.039 \end{bmatrix} \begin{bmatrix} 0.221 \\ 0.24 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 0 & 0.009 & 0 & 0 \\ 0 & 0 & -0.009 & 0 & 0 \end{bmatrix}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix} 0.305 \\ -0.663 \\ 0.358 \end{bmatrix}$$

신박AI

이젠 $\partial L/\partial W_f$ 차례입니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

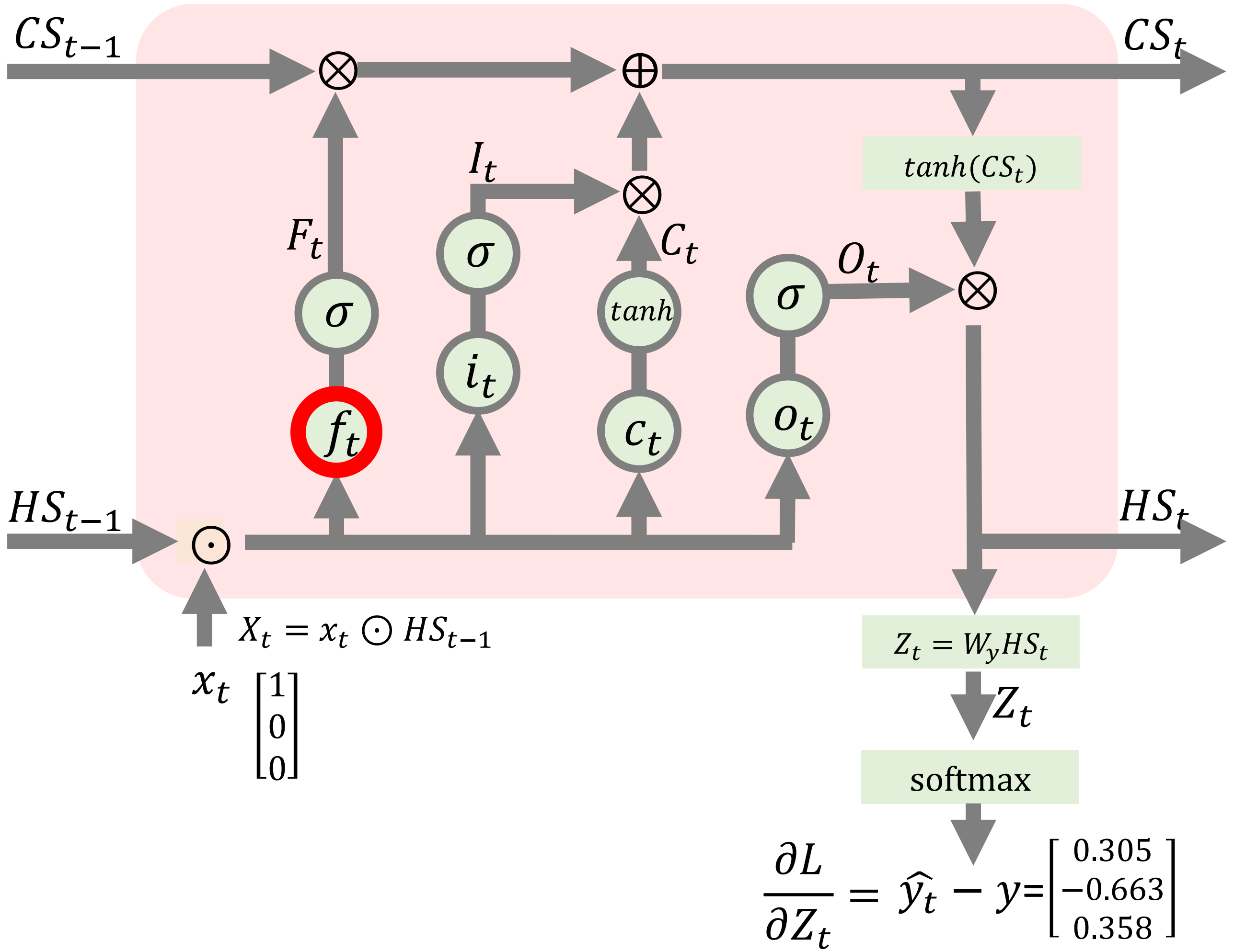
Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_f} =$$



$\partial L/\partial W_f$ 는 체인룰에 의해서 다음과 같이 전개할 수 있습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

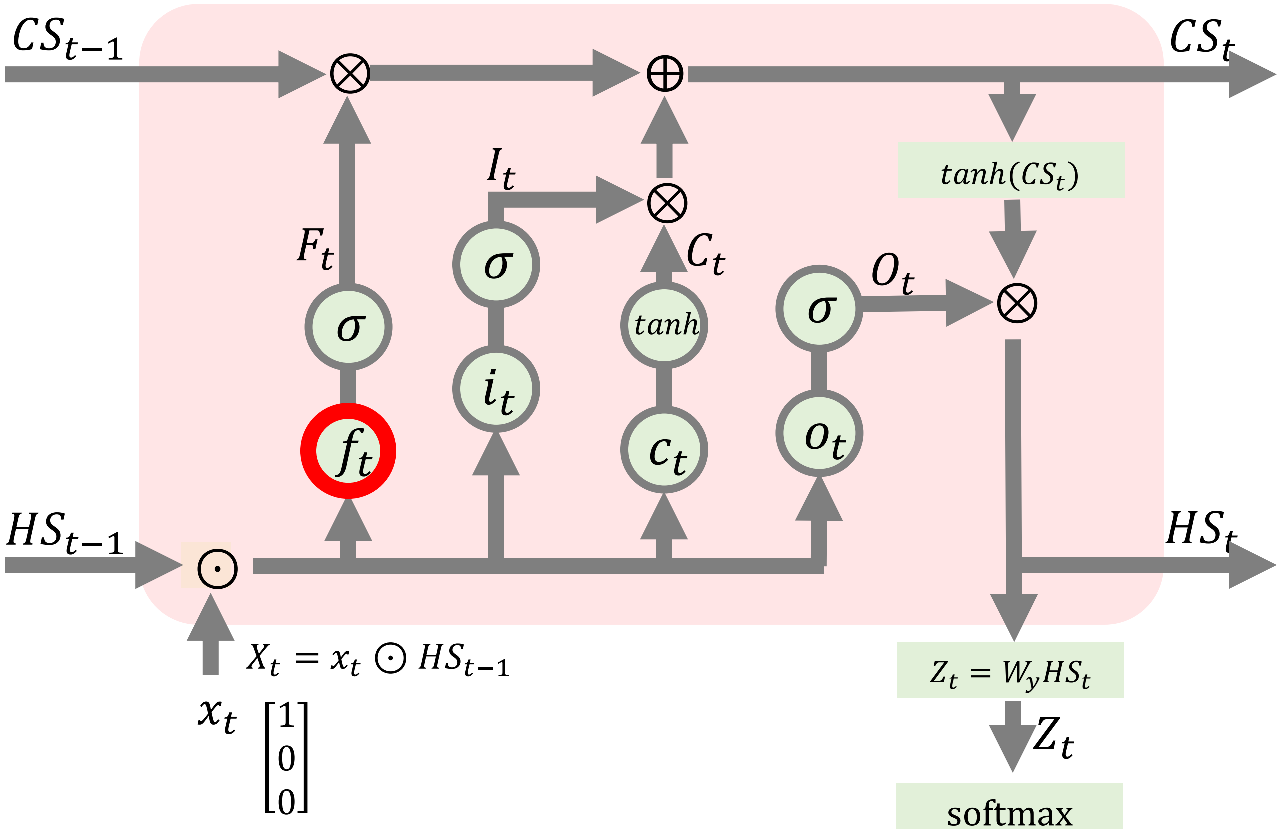
Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_f} = \frac{\partial L}{\partial CS_t}\frac{\partial CS_t}{\partial F_t}\frac{\partial F_t}{\partial f_t}\frac{\partial f_t}{\partial W_f}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix}0.305\\-0.663\\0.358\end{bmatrix}$$

$\partial L/\partial CS_t$ 부터 구해보도록 하겠습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

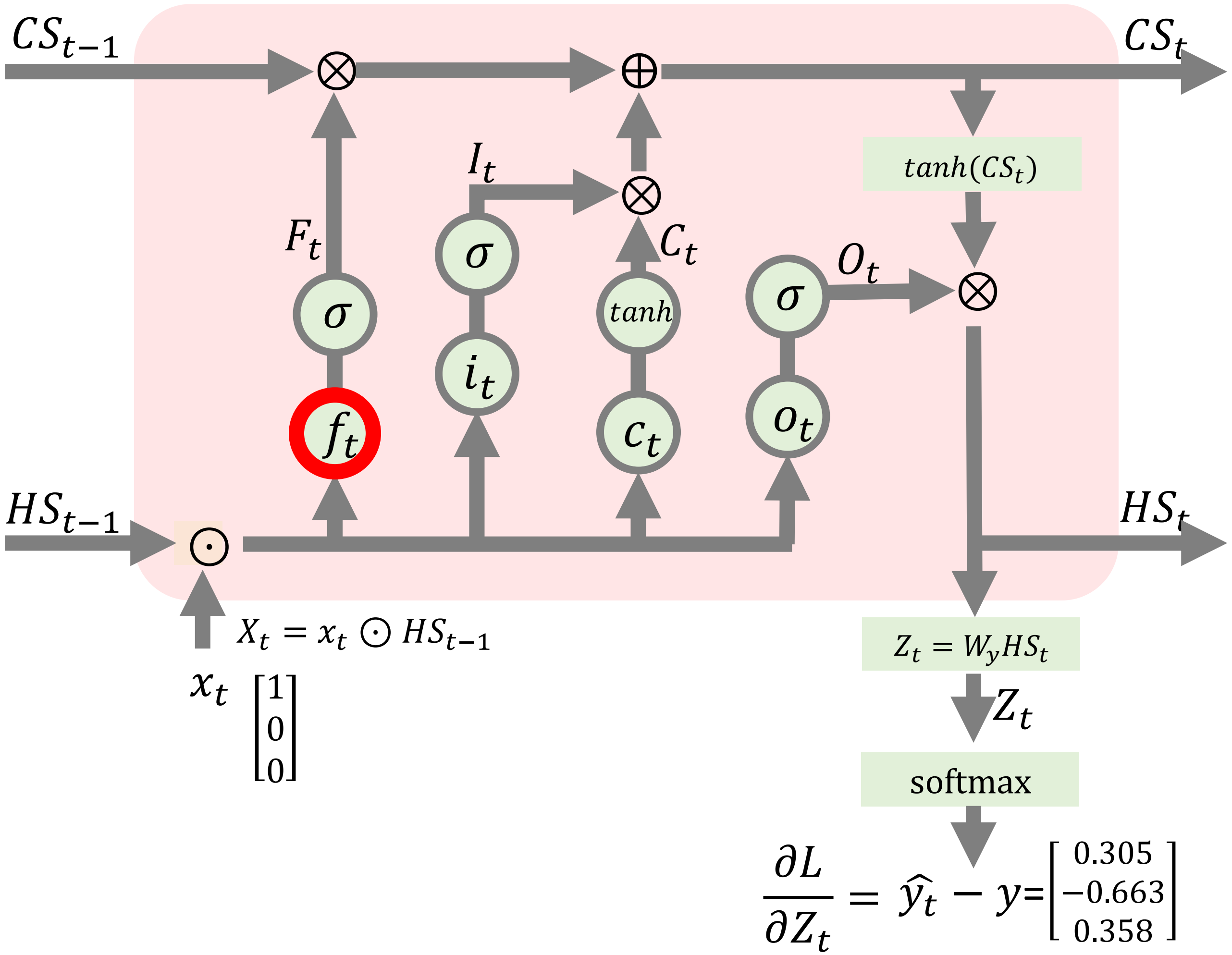
Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_f} = \boxed{\frac{\partial L}{\partial CS_t}} \frac{\partial CS_t}{\partial F_t} \frac{\partial F_t}{\partial f_t} \frac{\partial f_t}{\partial W_f}$$

$$\frac{\partial L}{\partial CS_t} =$$



신박AI

$\partial L/\partial CS_t$ 는 체인룰에 의해서 다음과 같이 전개할 수 있습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

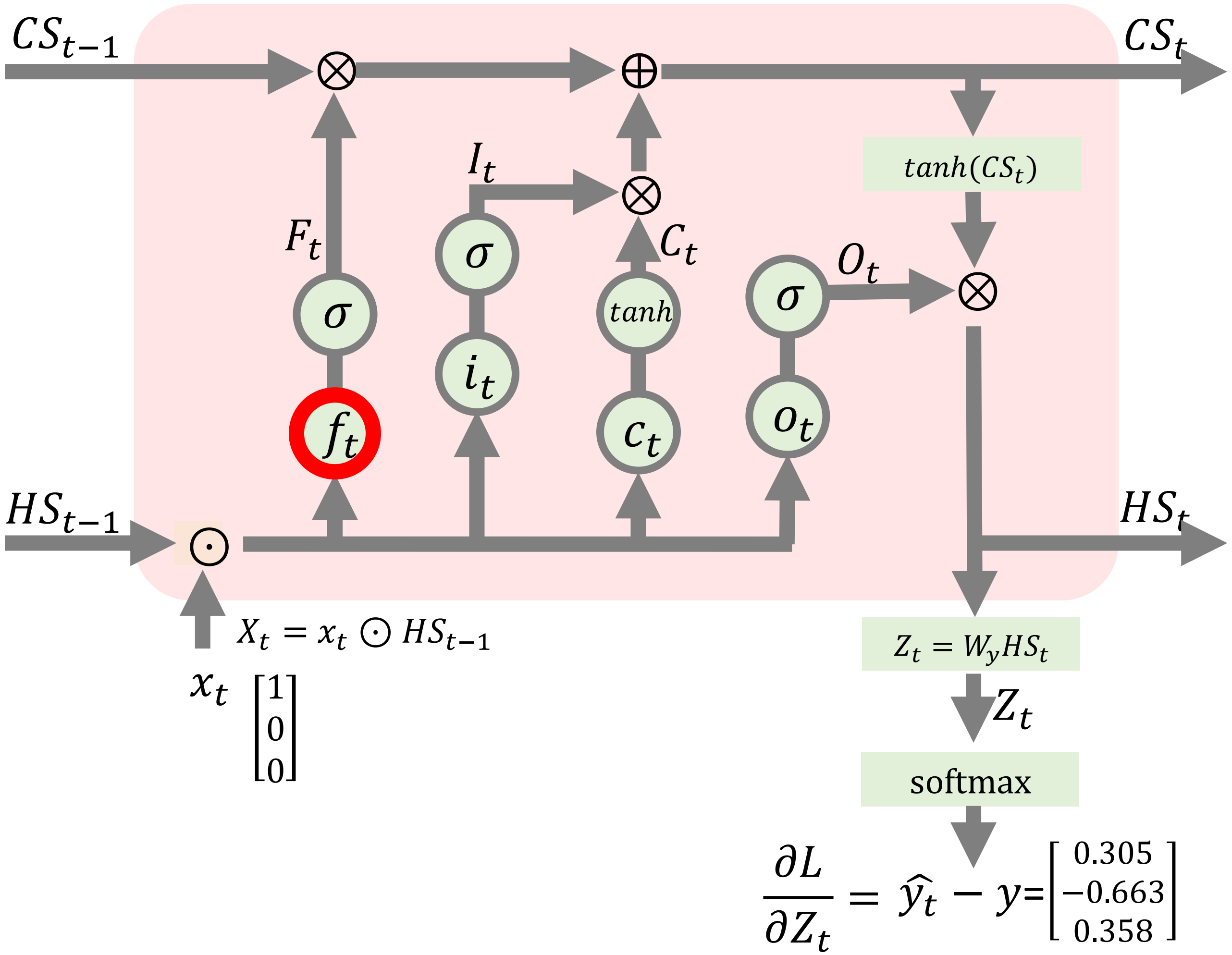
Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_f} = \boxed{\frac{\partial L}{\partial CS_t}} \frac{\partial CS_t}{\partial F_t} \frac{\partial F_t}{\partial f_t} \frac{\partial f_t}{\partial W_f}$$

$$\frac{\partial L}{\partial CS_t} =$$



신박AI

$\partial L/\partial CS_t$ 는 체인룰에 의해서 다음과 같이 전개할 수 있습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

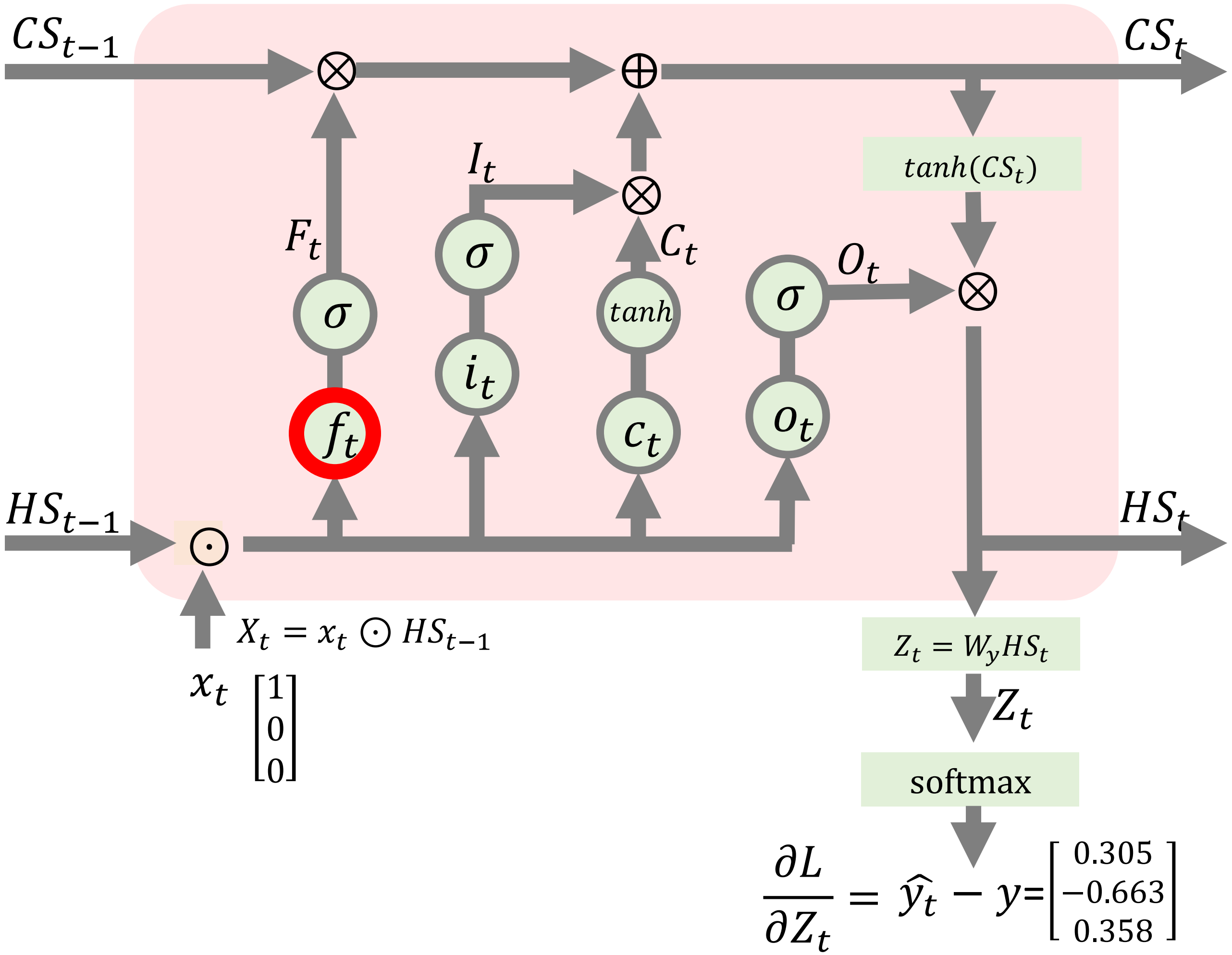
Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_f} = \frac{\partial L}{\partial CS_t}\frac{\partial CS_t}{\partial F_t}\frac{\partial F_t}{\partial f_t}\frac{\partial f_t}{\partial W_f}$$

$$\frac{\partial L}{\partial CS_t} = \frac{\partial L}{\partial Z_t}\frac{\partial Z_t}{\partial HS_t}\frac{\partial HS_t}{\partial CS_t}$$



신박AI

그러면 이 부분은 앞에서 구해 본 바 대로

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

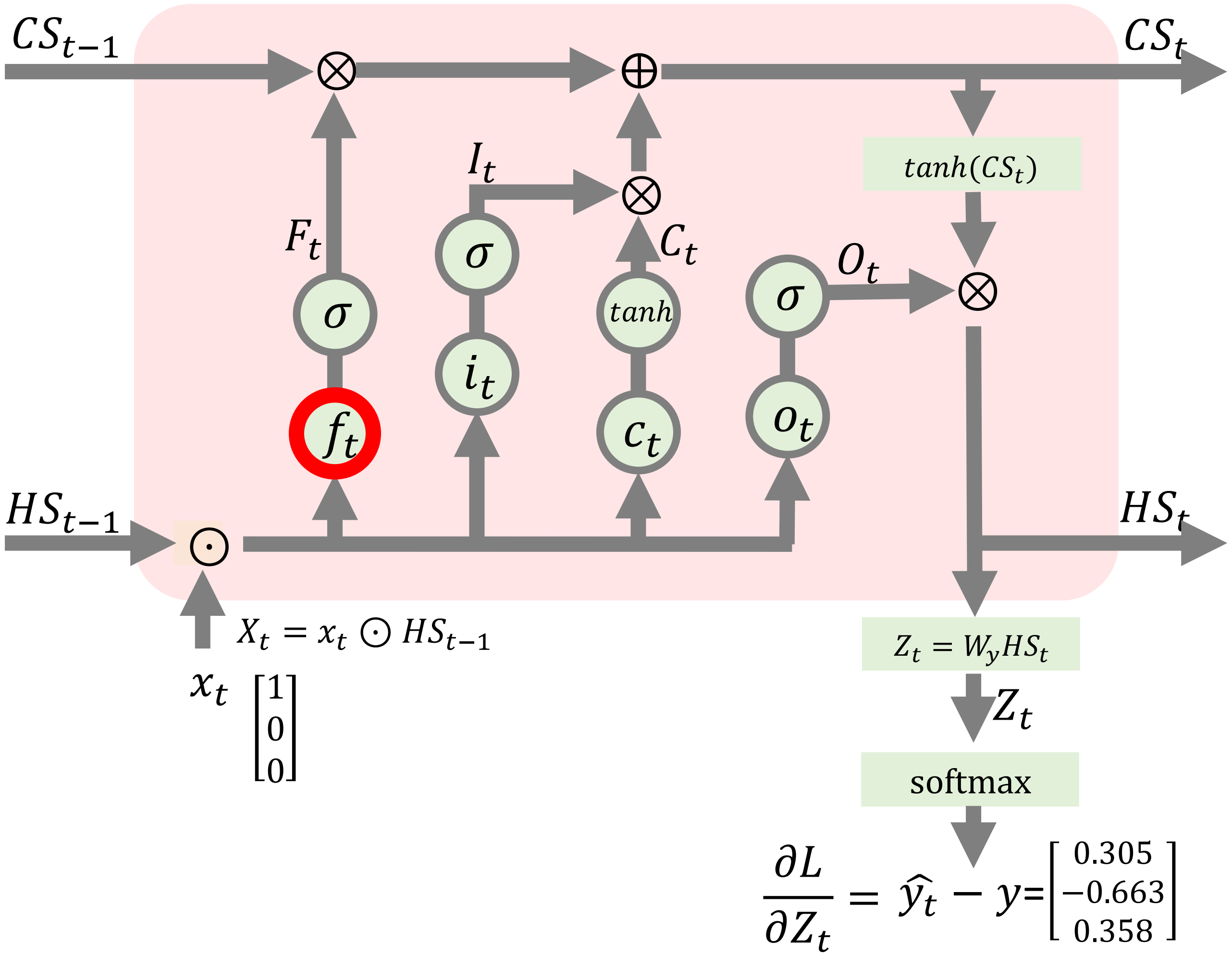
Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_f} = \frac{\partial L}{\partial CS_t}\frac{\partial CS_t}{\partial F_t}\frac{\partial F_t}{\partial f_t}\frac{\partial f_t}{\partial W_f}$$

$$\frac{\partial L}{\partial CS_t} = \boxed{\frac{\partial L}{\partial Z_t}\frac{\partial Z_t}{\partial HS_t}}\frac{\partial HS_t}{\partial CS_t}$$



이렇게 됩니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

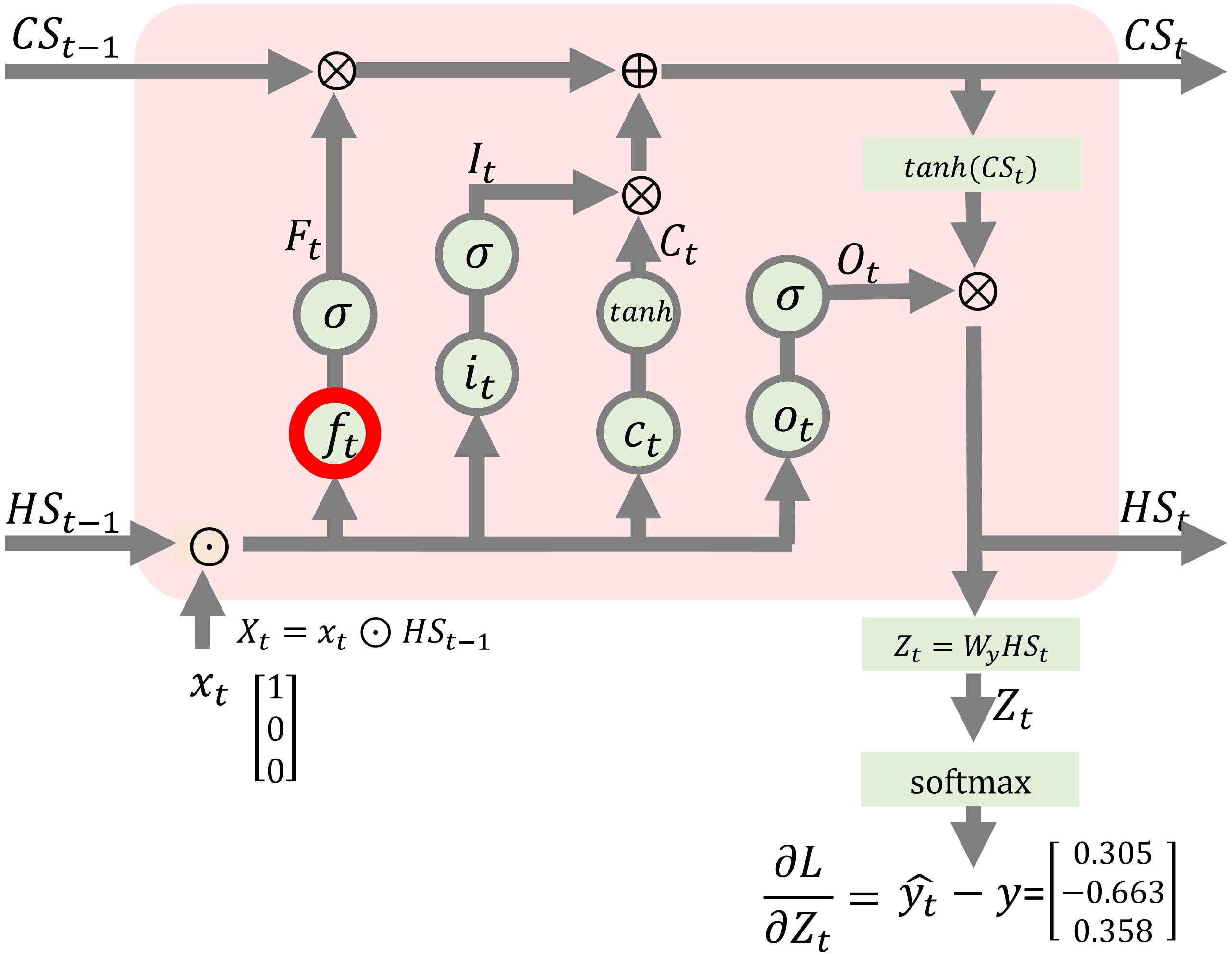
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_f} = \frac{\partial L}{\partial CS_t}\frac{\partial CS_t}{\partial F_t}\frac{\partial F_t}{\partial f_t}\frac{\partial f_t}{\partial W_f}$$

$$\frac{\partial L}{\partial CS_t} = \frac{\partial L}{\partial Z_t}\frac{\partial Z_t}{\partial HS_t}\frac{\partial HS_t}{\partial CS_t}$$

$$= (\hat{y}_t - y)W_y \frac{\partial HS_t}{\partial CS_t}$$



신박AI

그러면 이 부분은

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

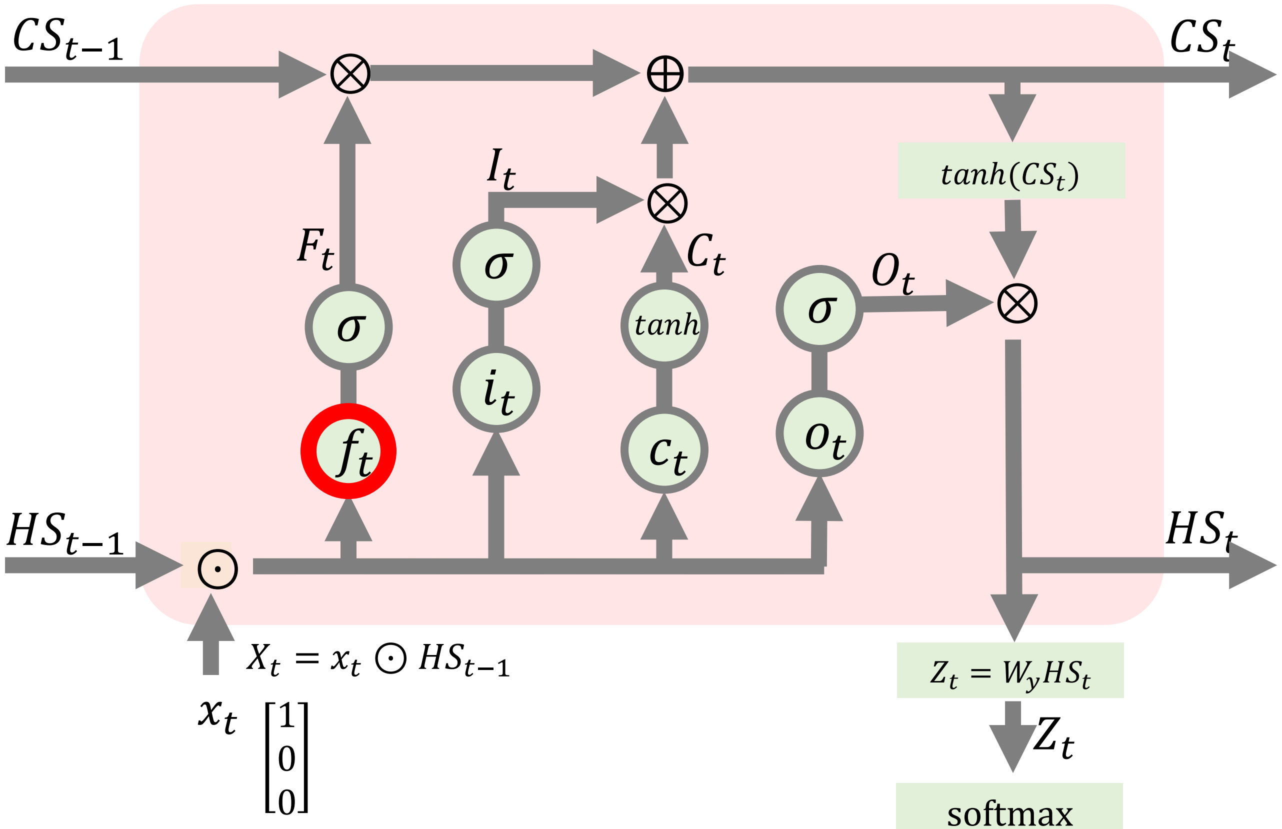
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_f} = \frac{\partial L}{\partial CS_t}\frac{\partial CS_t}{\partial F_t}\frac{\partial F_t}{\partial f_t}\frac{\partial f_t}{\partial W_f}$$

$$\frac{\partial L}{\partial CS_t} = \frac{\partial L}{\partial Z_t}\frac{\partial Z_t}{\partial HS_t}\frac{\partial HS_t}{\partial CS_t}$$

$$= (\hat{y}_t - y)W_y\boxed{\frac{\partial HS_t}{\partial CS_t}}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix}0.305\\-0.663\\0.358\end{bmatrix}$$

이 공식에 의해서

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

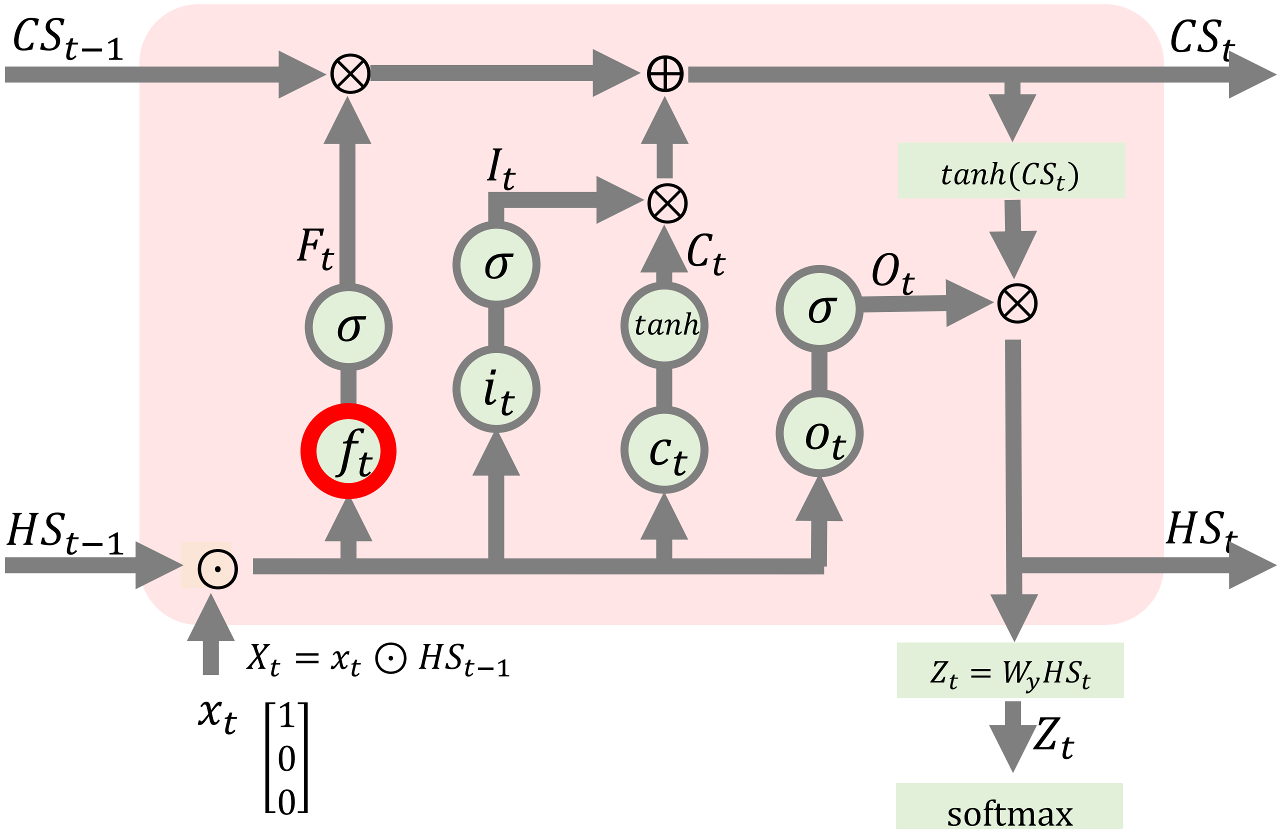
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_f} = \frac{\partial L}{\partial CS_t}\frac{\partial CS_t}{\partial F_t}\frac{\partial F_t}{\partial f_t}\frac{\partial f_t}{\partial W_f}$$

$$\frac{\partial L}{\partial CS_t} = \frac{\partial L}{\partial Z_t}\frac{\partial Z_t}{\partial HS_t}\frac{\partial HS_t}{\partial CS_t}$$

$$= (\hat{y}_t - y)W_y \boxed{\frac{\partial HS_t}{\partial CS_t}}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix}0.305\\-0.663\\0.358\end{bmatrix}$$

$$HS_t = O_t \otimes tanh(CS_t) \rightarrow \frac{\partial HS_t}{\partial CS_t} = O_t(1 - tanh^2(CS_t))$$

신박AI

이렇게 바꾸어 쓸수가 있습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

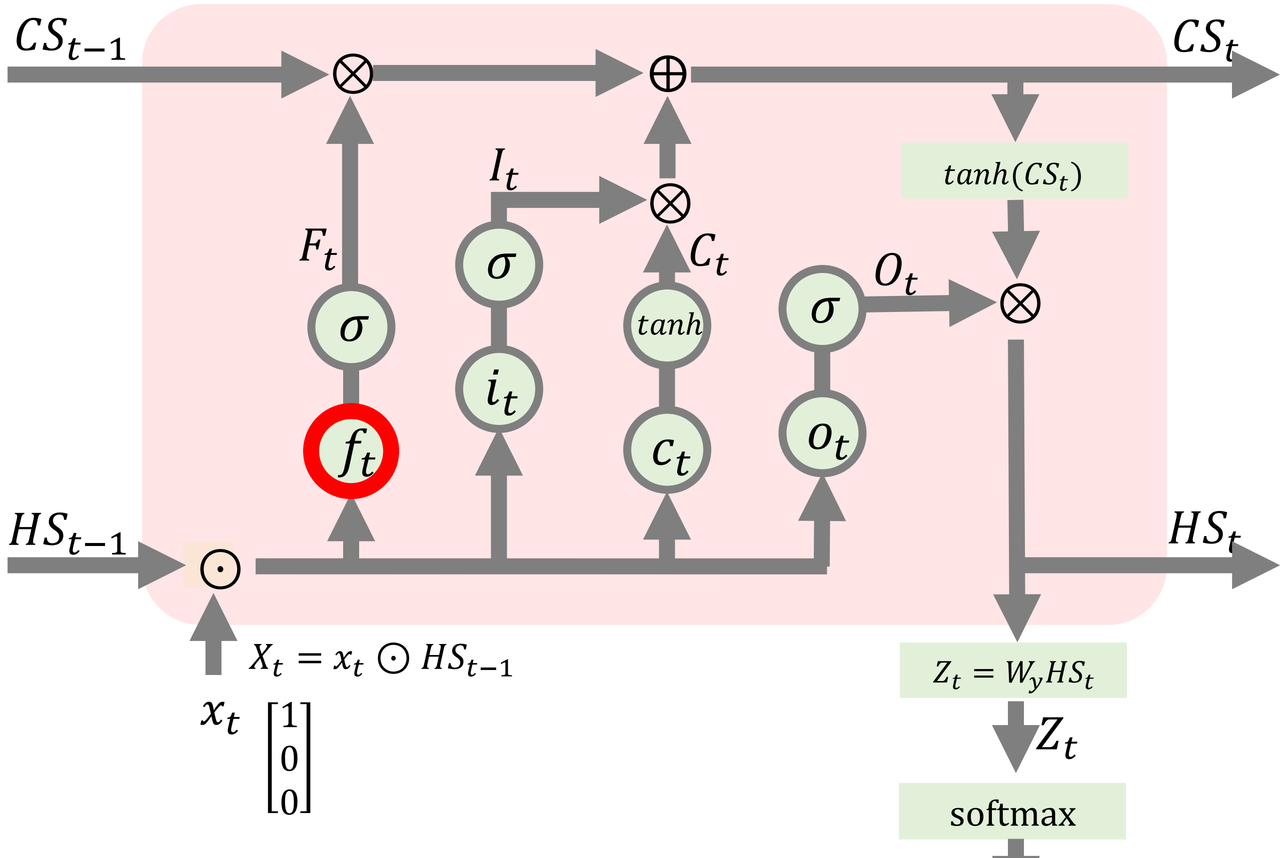
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_f} = \frac{\partial L}{\partial CS_t}\frac{\partial CS_t}{\partial F_t}\frac{\partial F_t}{\partial f_t}\frac{\partial f_t}{\partial W_f}$$

$$\frac{\partial L}{\partial CS_t} = \frac{\partial L}{\partial Z_t}\frac{\partial Z_t}{\partial HS_t}\frac{\partial HS_t}{\partial CS_t}$$

$$= (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix}0.305\\-0.663\\0.358\end{bmatrix}$$

$$HS_t = O_t \otimes tan\,h(CS_t) \rightarrow \frac{\partial HS_t}{\partial CS_t} = O_t(1 - tanh^2(CS_t))$$

신박AI

물론 $\partial L/\partial CS_t$ 도 이전 상태에서 전달받는 것 까지 고려를 해야합니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

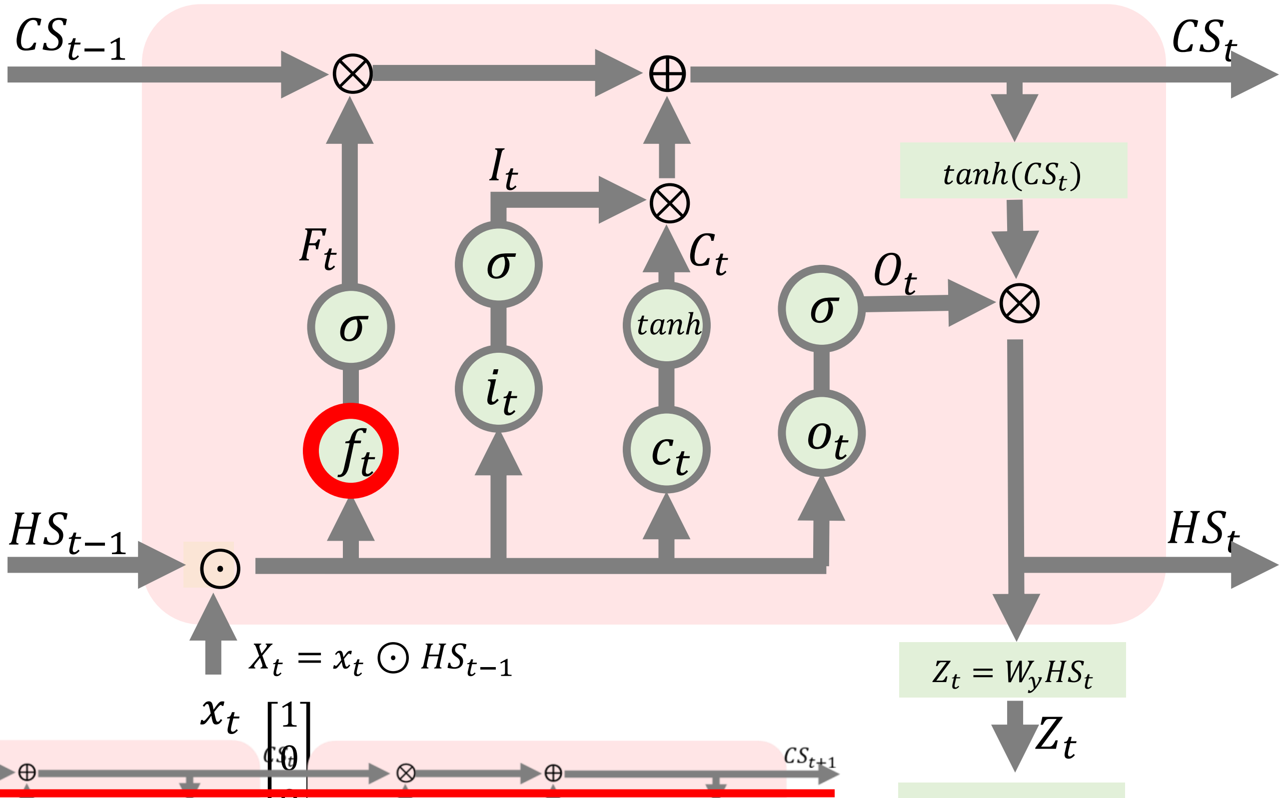
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

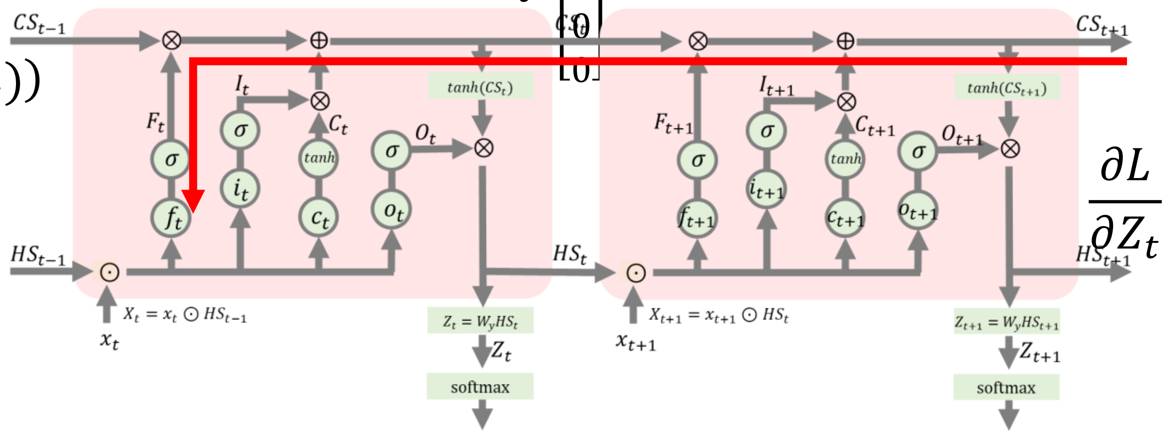
$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_f} = \frac{\partial L}{\partial CS_t}\frac{\partial CS_t}{\partial F_t}\frac{\partial F_t}{\partial f_t}\frac{\partial f_t}{\partial W_f}$$

$$\frac{\partial L}{\partial CS_t} = \frac{\partial L}{\partial Z_t}\frac{\partial Z_t}{\partial HS_t}\frac{\partial HS_t}{\partial CS_t} + dCS_{t+1}$$

$$= (\hat{y}_t - y)W_y O_t(1 - tanh^2(CS_t))$$





$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix}0.305\\-0.663\\0.358\end{bmatrix}$$

이 부분도 실제 LSTM코드를 구현할 때 코드와 함께 설명을 드리도록 하겠습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

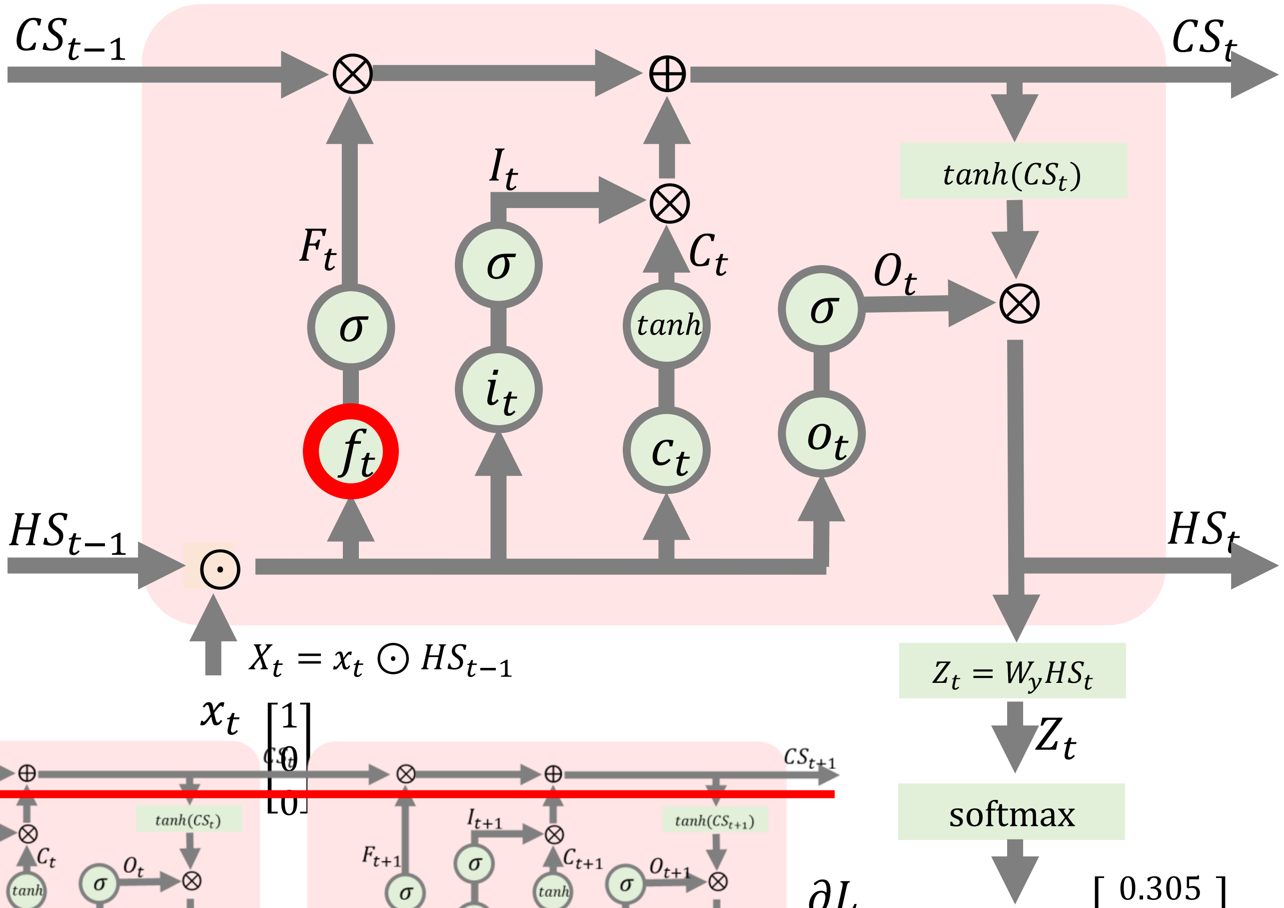
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

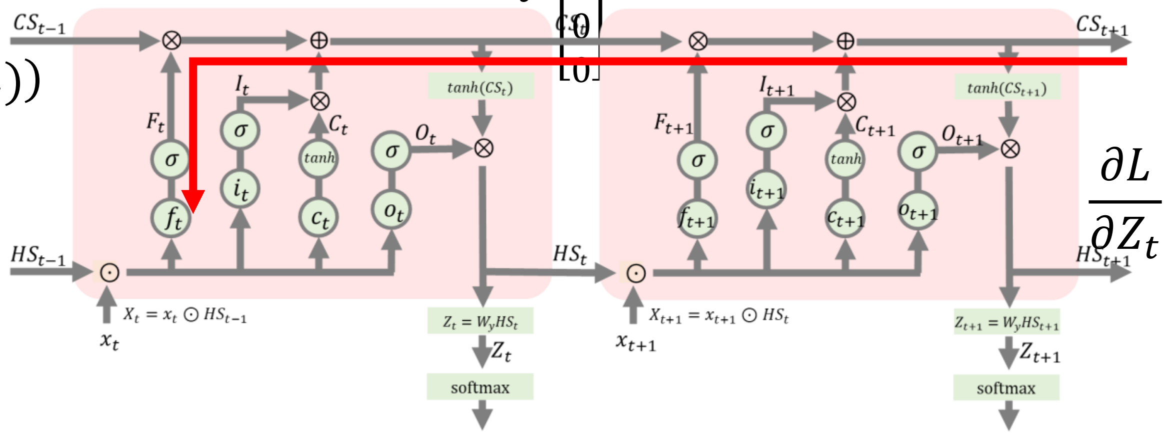
$$\frac{\partial L}{\partial W_f} = \frac{\partial L}{\partial CS_t}\frac{\partial CS_t}{\partial F_t}\frac{\partial F_t}{\partial f_t}\frac{\partial f_t}{\partial W_f}$$

$$\frac{\partial L}{\partial CS_t} = \frac{\partial L}{\partial Z_t}\frac{\partial Z_t}{\partial HS_t}\frac{\partial HS_t}{\partial CS_t} + dCS_{t+1}$$

$$= (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix} 0.305 \\ -0.663 \\ 0.358 \end{bmatrix}$$



자 어쨌든, $\partial L/\partial CS_t$까지 전개해 보았습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

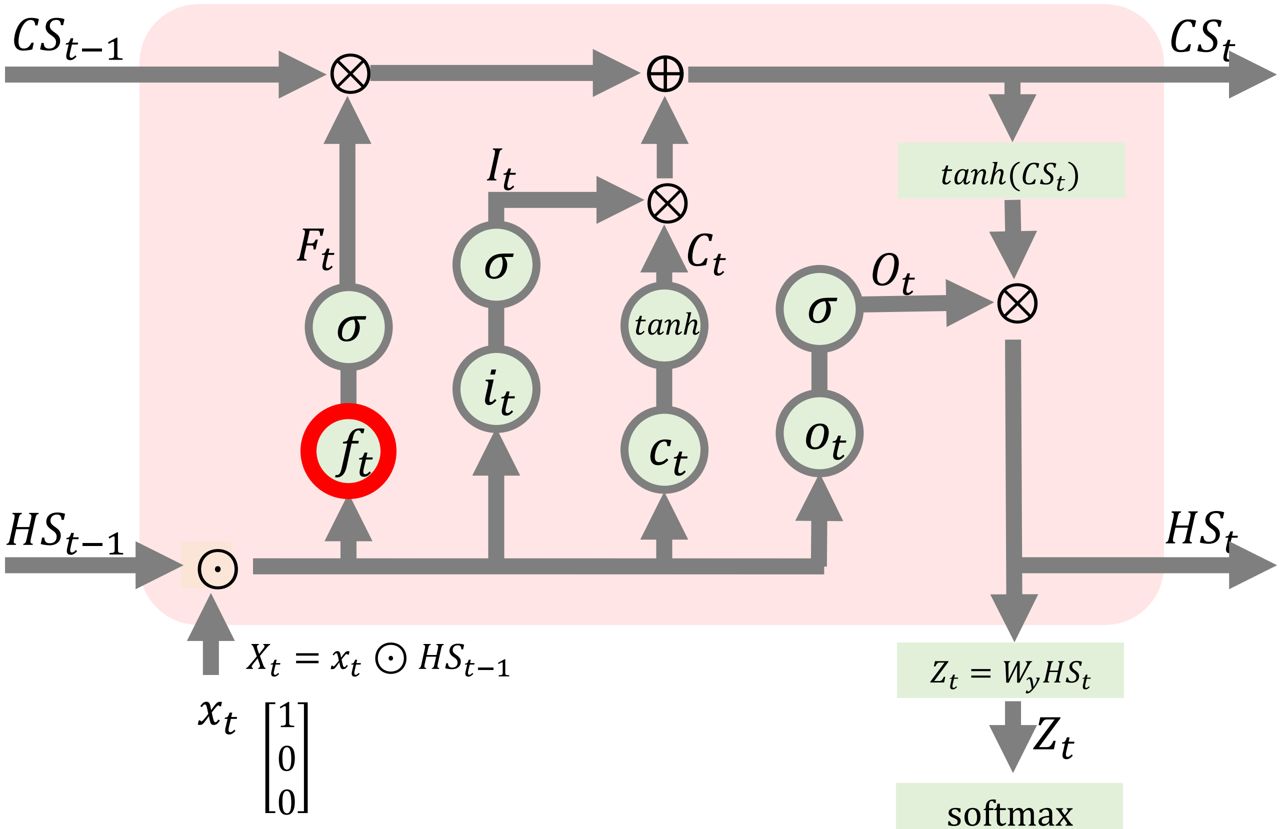
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_f} = \frac{\partial L}{\partial CS_t}\frac{\partial CS_t}{\partial F_t}\frac{\partial F_t}{\partial f_t}\frac{\partial f_t}{\partial W_f}$$

$$\frac{\partial L}{\partial CS_t} = \frac{\partial L}{\partial Z_t}\frac{\partial Z_t}{\partial HS_t}\frac{\partial HS_t}{\partial CS_t}$$

$$= \boxed{(\hat{y}_t - y)W_y O_t(1 - tanh^2(CS_t))}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix}0.305\\-0.663\\0.358\end{bmatrix}$$

신박AI

이 식은 앞으로도 계속 쓰이니까 귀퉁이에 잘 기록해 두겠습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

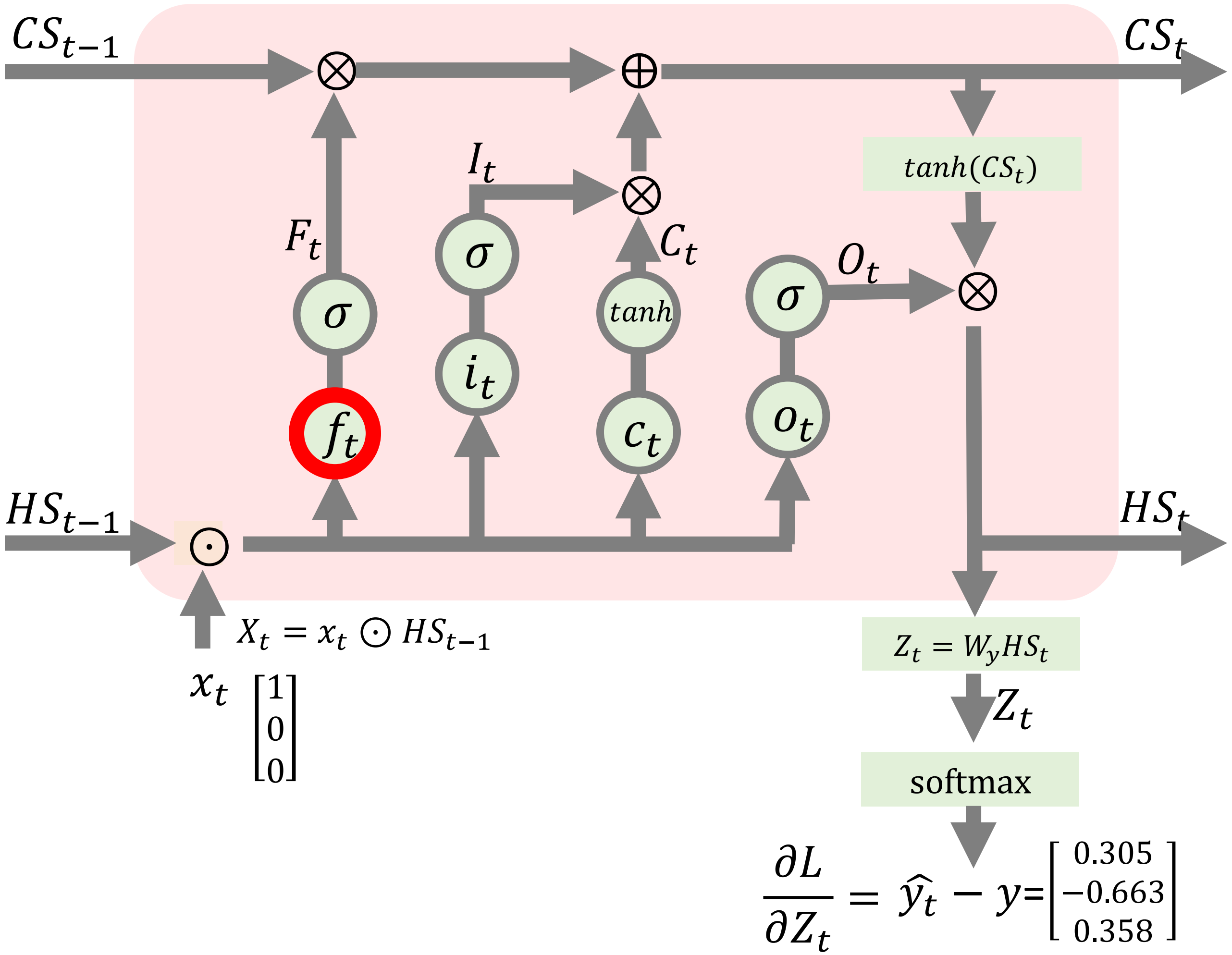
$\boxed{\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))}$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_f} = \frac{\partial L}{\partial CS_t} \frac{\partial CS_t}{\partial F_t} \frac{\partial F_t}{\partial f_t} \frac{\partial f_t}{\partial W_f}$$

$$\frac{\partial L}{\partial CS_t} = \frac{\partial L}{\partial Z_t} \frac{\partial Z_t}{\partial HS_t} \frac{\partial HS_t}{\partial CS_t}$$

$$(\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$$



신박AI

그리고 이렇게 식을 다시 써보았습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

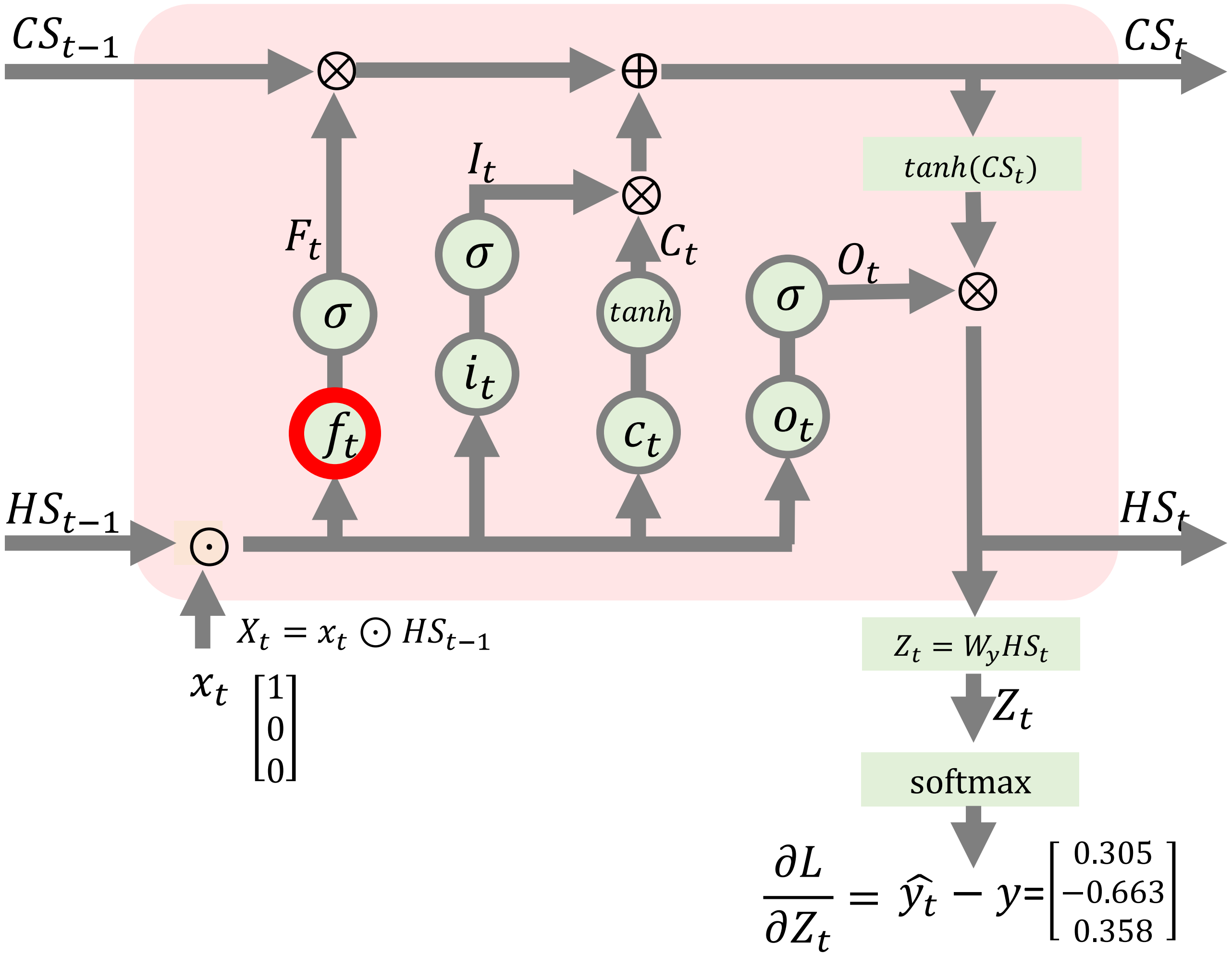
Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y)W_y O_t(1 - tanh^2(CS_t))$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_f} = (\hat{y}_t - y)W_y O_t(1 - tanh^2(CS_t))\frac{\partial CS_t}{\partial F_t}\frac{\partial F_t}{\partial f_t}\frac{\partial f_t}{\partial W_f}$$



그리고 다음은 $\partial CS_t/\partial F_t$를 구해볼 차례입니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

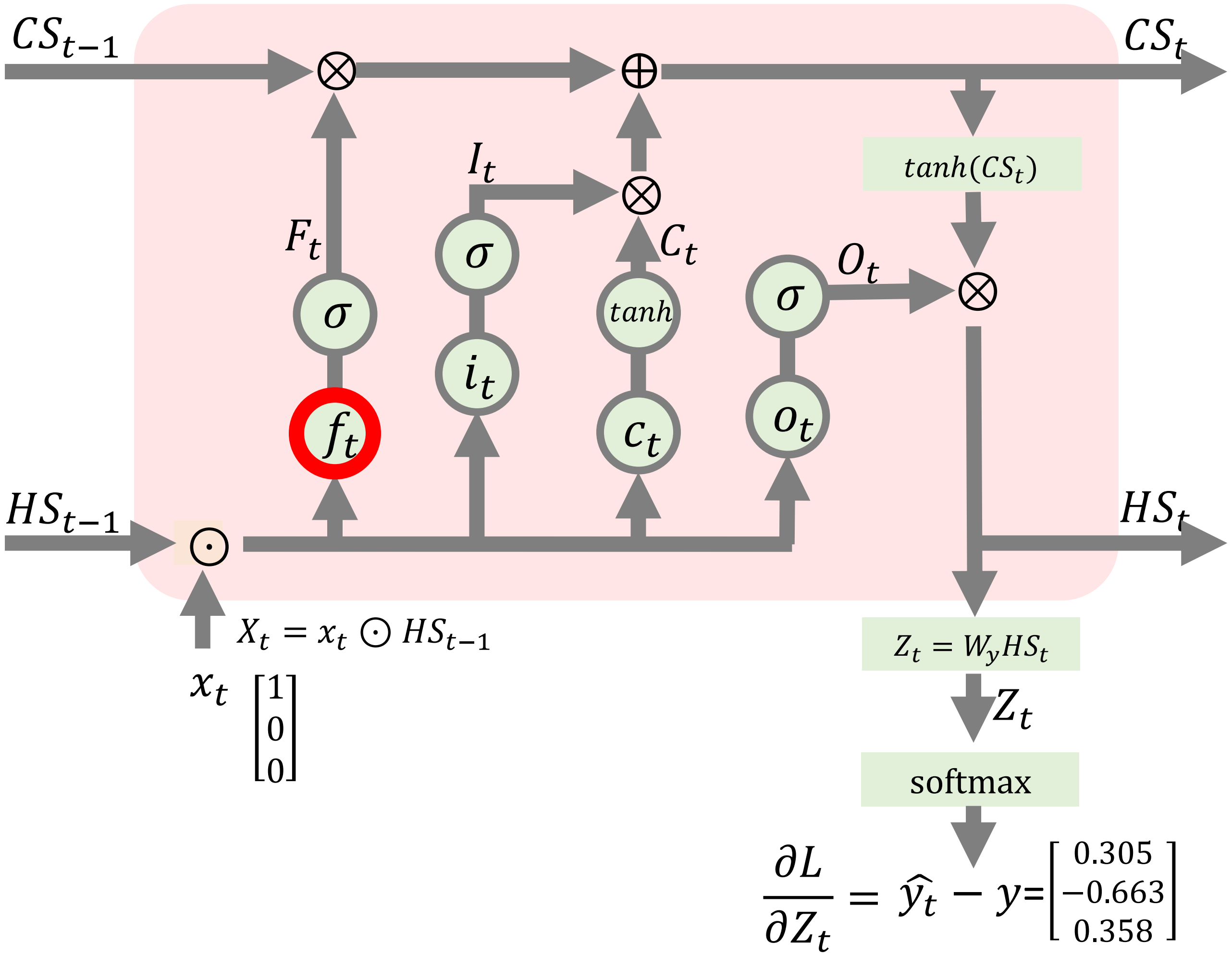
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y)W_y O_t(1 - tanh^2(CS_t))$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_f} = (\hat{y}_t - y)W_y O_t(1 - tanh^2(CS_t))\frac{\partial CS_t}{\partial F_t}\frac{\partial F_t}{\partial f_t}\frac{\partial f_t}{\partial W_f}$$

$$\frac{\partial CS_t}{\partial F_t}$$



신박AI

우선 셀 상태 CS_t 공식은 다음과 같습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

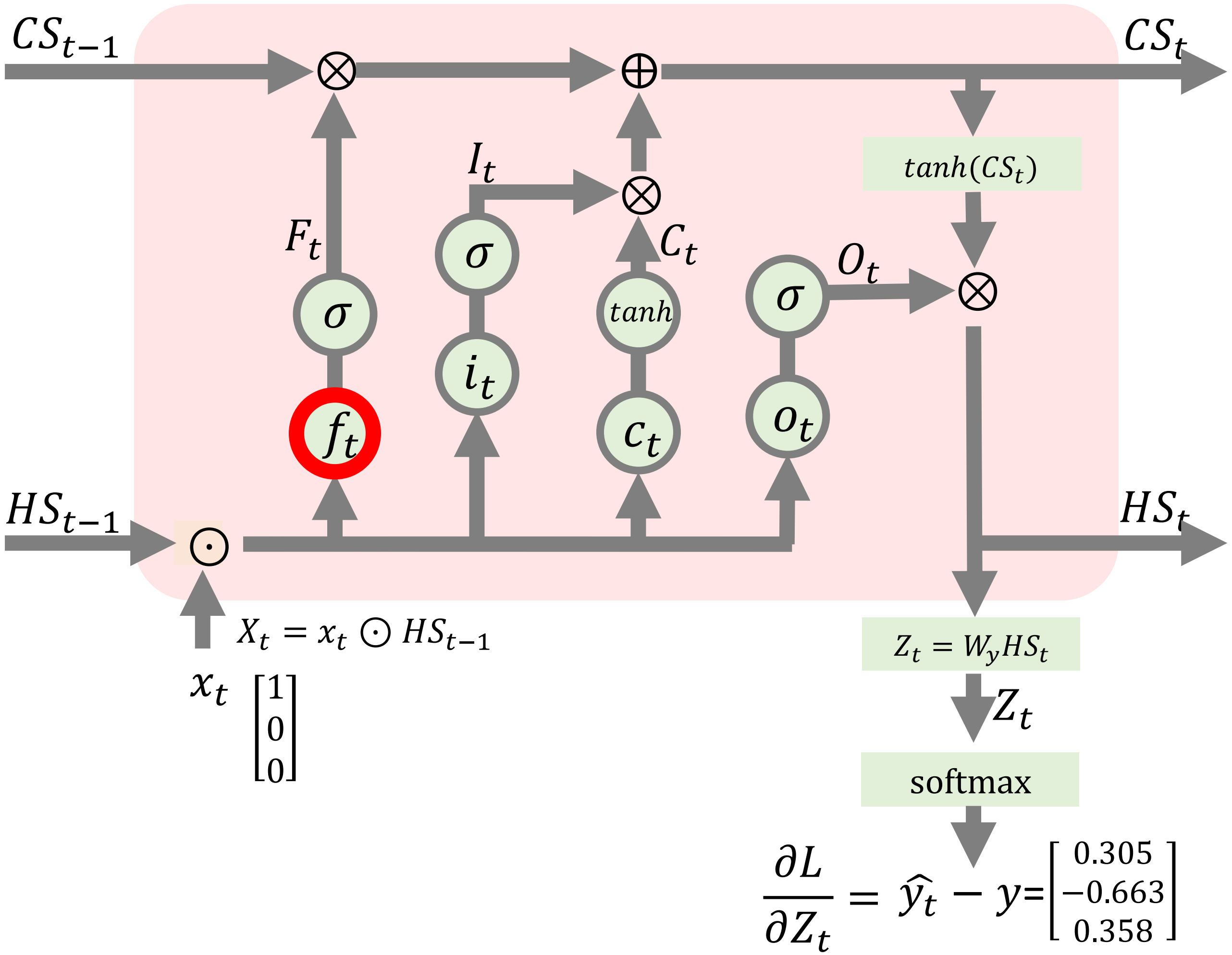
$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$

$Z_t = W_y HS_t$

$$\frac{\partial L}{\partial W_f} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t)) \frac{\partial CS_t}{\partial F_t} \frac{\partial F_t}{\partial f_t} \frac{\partial f_t}{\partial W_f}$$

$$\frac{\partial CS_t}{\partial F_t}$$

$$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$$



셀 상태 CS_t 공식도 자주 쓰게 되니 귀퉁이에 담아두겠습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$

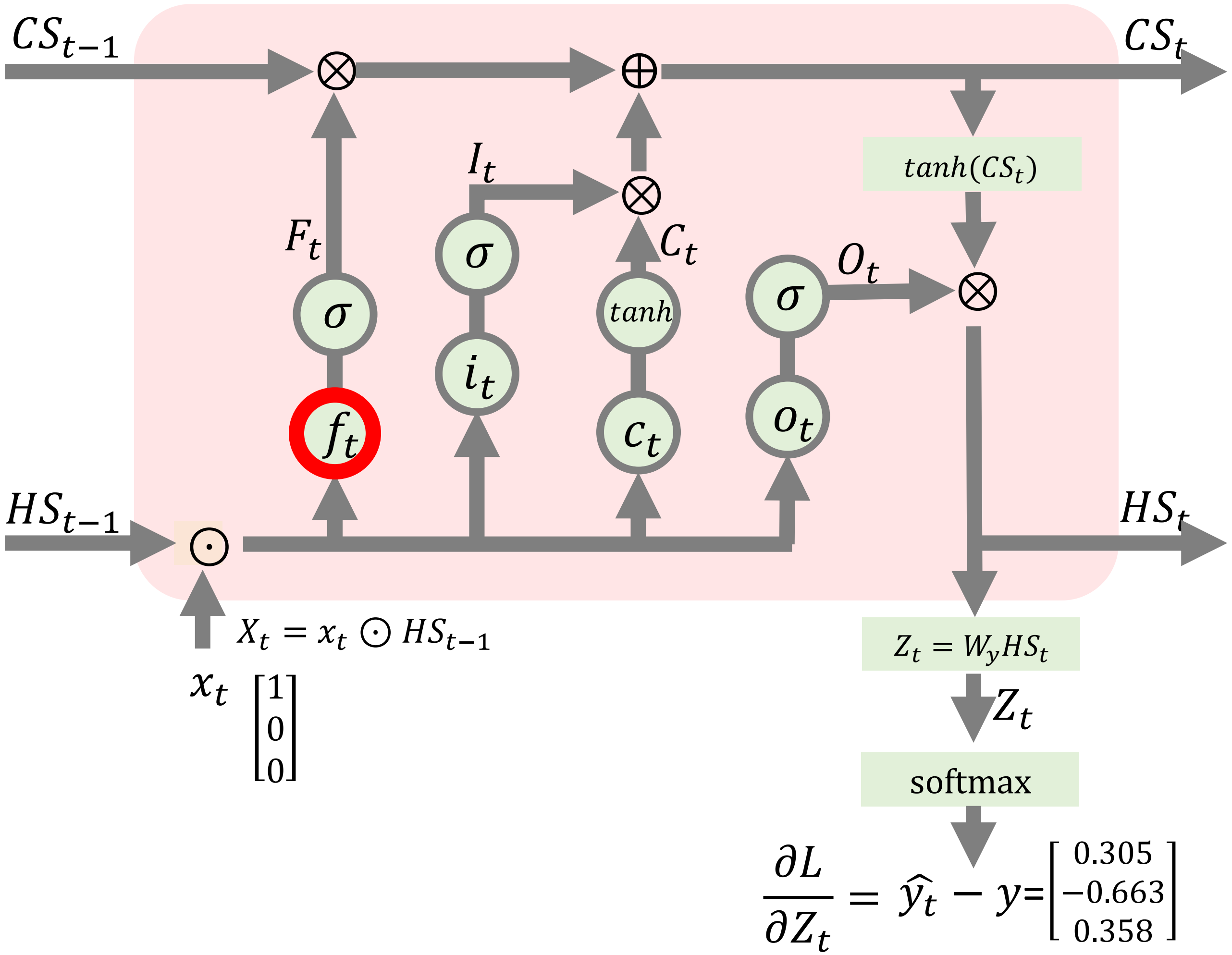
$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial W_f} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t)) \frac{\partial CS_t}{\partial F_t} \frac{\partial F_t}{\partial f_t} \frac{\partial f_t}{\partial W_f}$$

$$\frac{\partial CS_t}{\partial F_t}$$

$$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$$



신박AI

그러면 $\partial CS_t / \partial F_t$ 는 어렵지 않게 구할 수 있습니다. CS_{t-1} 입니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$$

$Z_t = W_y HS_t$

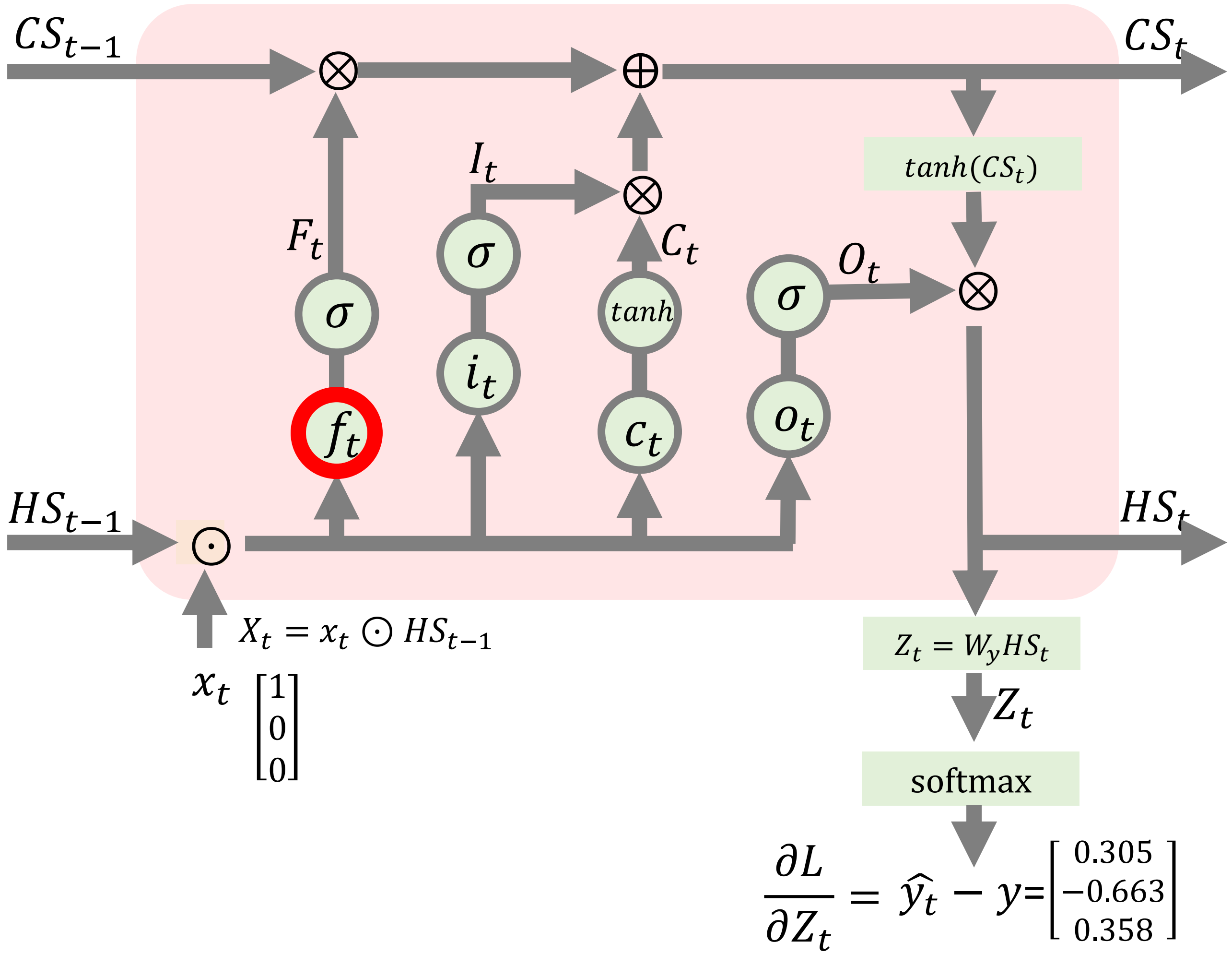
$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial W_f} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t)) \frac{\partial CS_t}{\partial F_t} \frac{\partial F_t}{\partial f_t} \frac{\partial f_t}{\partial W_f}$$

$$\frac{\partial CS_t}{\partial F_t}$$

$$CS_t = CS_{t-1} \otimes \cancel{F_t} + I_t \cancel{\otimes} C_t$$

$$\frac{\partial CS_t}{\partial F_t} = CS_{t-1}$$



이렇게 식을 업데이트 할 수가 있습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$$

$$Z_t = W_y HS_t$$

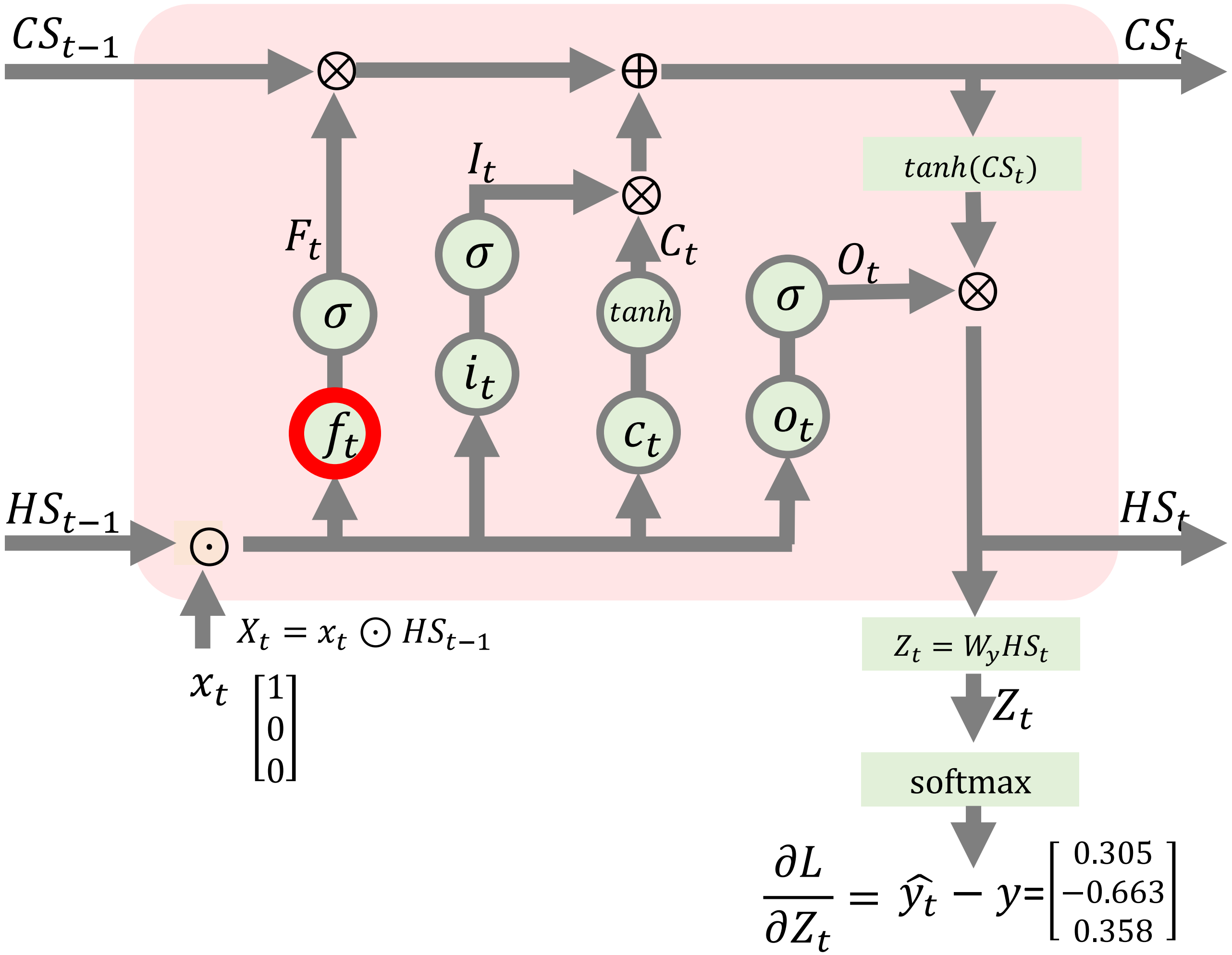
$$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$$

$$\frac{\partial L}{\partial W_f} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t)) CS_{t-1} \frac{\partial F_t}{\partial f_t} \frac{\partial f_t}{\partial W_f}$$

$$\frac{\partial CS_t}{\partial F_t}$$

$$CS_t = CS_{t-1} \otimes \cancel{F_t} + I_t \cancel{\otimes} C_t$$

$$\frac{\partial CS_t}{\partial F_t} = CS_{t-1}$$



그 다음은 $\partial F_t / \partial f_t$를 구해보겠습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

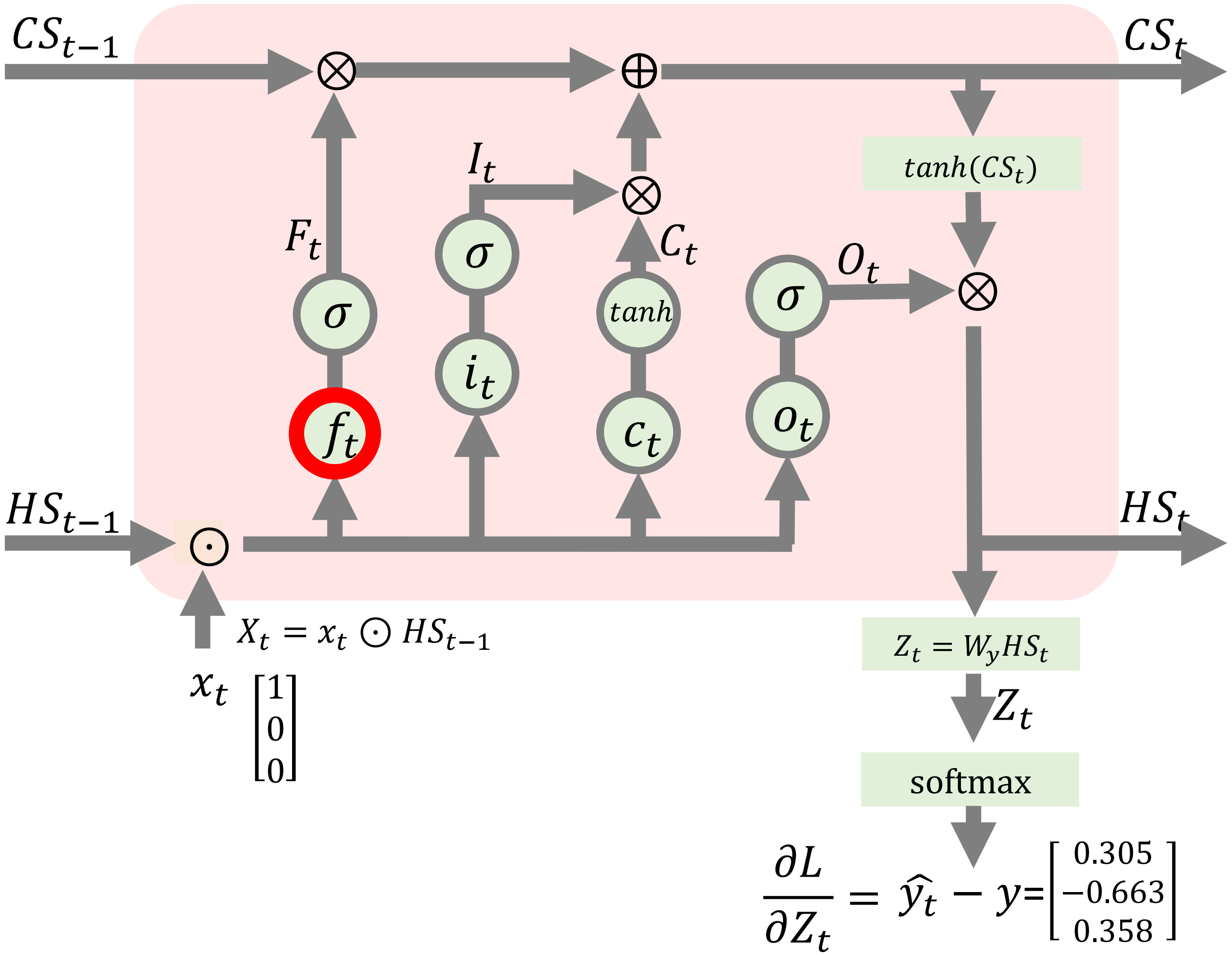
$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial W_f} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t)) CS_{t-1} \boxed{\frac{\partial F_t}{\partial f_t}} \frac{\partial f_t}{\partial W_f}$$

$$\frac{\partial F_t}{\partial f_t}$$



신박AI

$\partial F_t/\partial f_t$은 시그모이드 미분 함수에 의해서

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

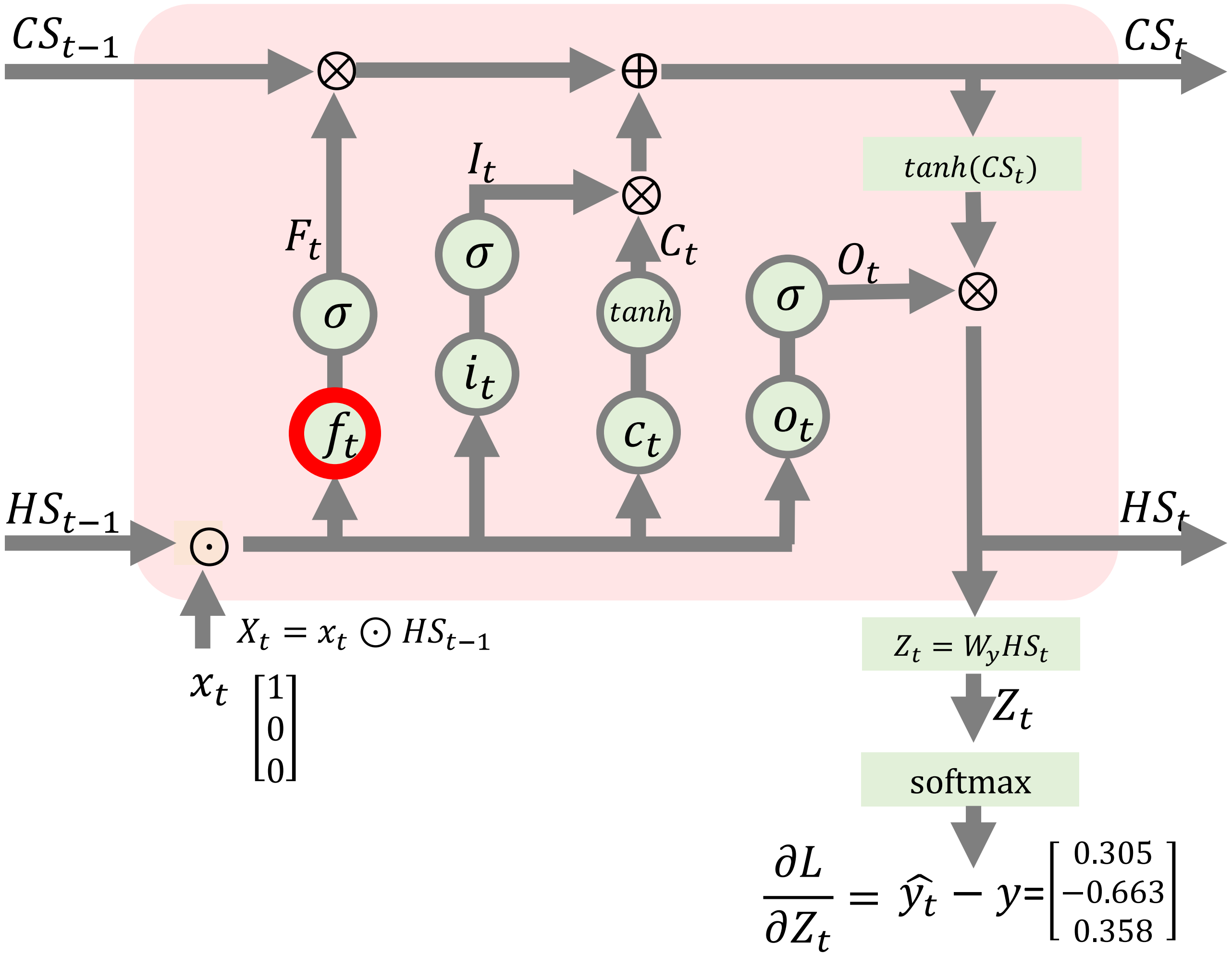
$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y)W_y O_t(1 - tanh^2(CS_t))$$

$$Z_t = W_y HS_t$$

$$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$$

$$\frac{\partial L}{\partial W_f} = (\hat{y}_t - y)W_y O_t(1 - tanh^2(CS_t))CS_{t-1}\frac{\partial F_t}{\partial f_t}\frac{\partial f_t}{\partial W_f}$$

$$\frac{\partial F_t}{\partial f_t}$$



신박AI

이렇게 구할 수가 있고,

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

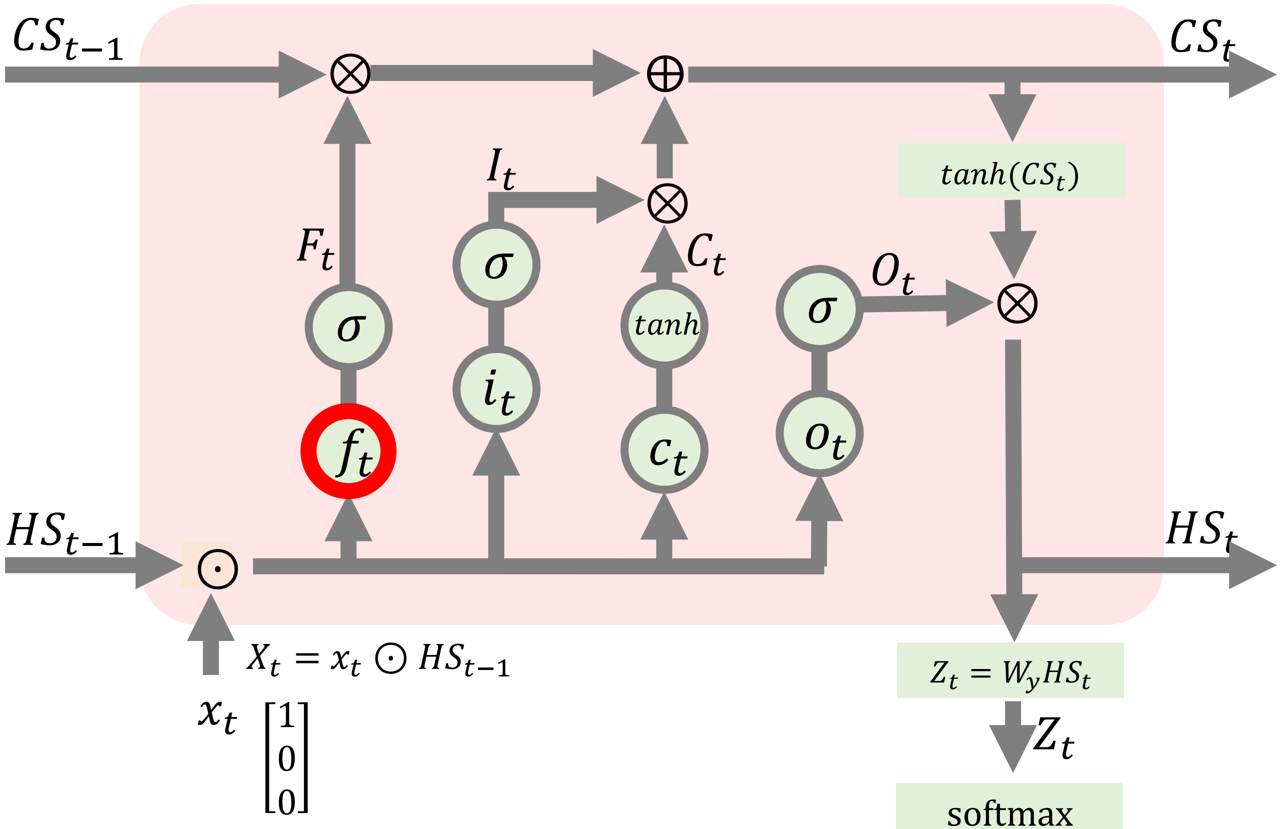
$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y)W_y O_t(1 - tanh^2(CS_t))$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial W_f} = (\hat{y}_t - y)W_y O_t(1 - tanh^2(CS_t))CS_{t-1}\frac{\partial F_t}{\partial f_t}\frac{\partial f_t}{\partial W_f}$$

$$\boxed{\frac{\partial F_t}{\partial f_t} = F_t(1 - F_t)}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix}0.305\\-0.663\\0.358\end{bmatrix}$$

이어서 $\partial f_t/\partial W_f$는 이 공식에 의해서

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$\boxed{f_t = W_f X_t}$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

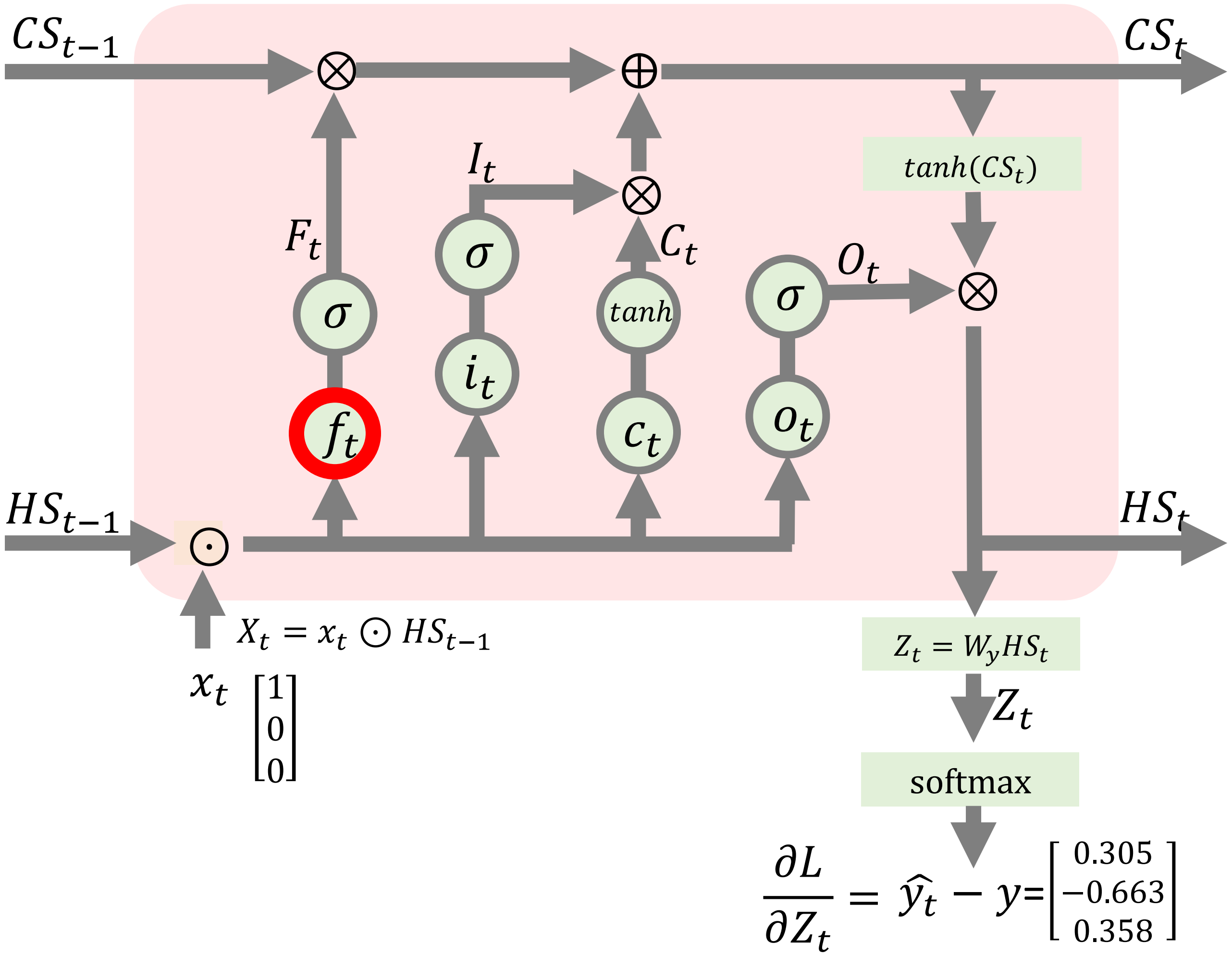
$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial W_f} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t)) CS_{t-1} \frac{\partial F_t}{\partial f_t} \frac{\partial f_t}{\partial W_f}$$

$$\frac{\partial F_t}{\partial f_t} = F_t(1 - F_t)$$

$$\frac{\partial f_t}{\partial W_f}$$



신박AI

X_t로 구할 수 있습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$

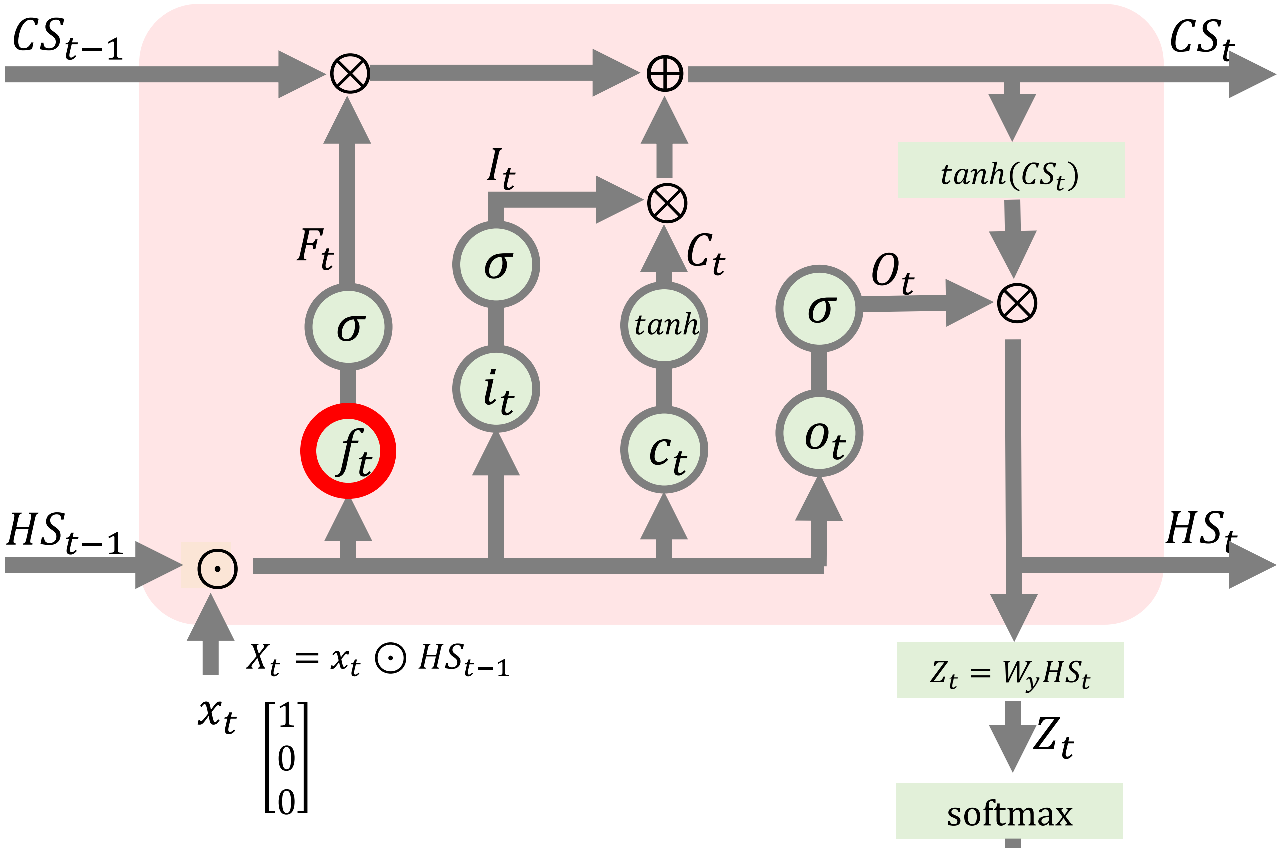
$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial W_f} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t)) CS_{t-1} \frac{\partial F_t}{\partial f_t} \frac{\partial f_t}{\partial W_f}$$

$$\frac{\partial F_t}{\partial f_t} = F_t(1 - F_t)$$

$$\boxed{\frac{\partial f_t}{\partial W_f} = X_t}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix}0.305\\-0.663\\0.358\end{bmatrix}$$

신박AI

그러면 이 두 식을 $\partial L/\partial W_f$ 식에 넣으면,

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$$

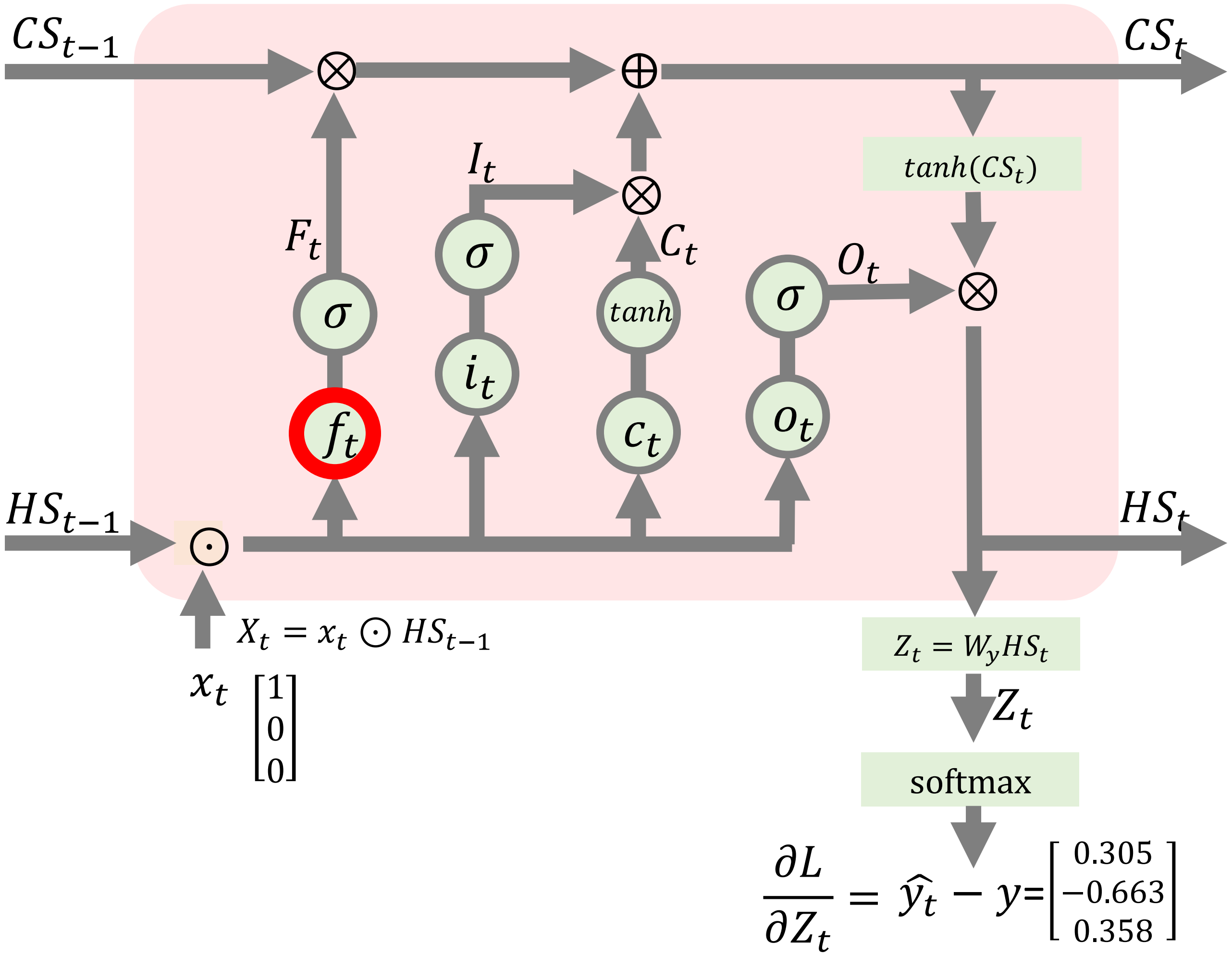
$$Z_t = W_y HS_t$$

$$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$$

$$\frac{\partial L}{\partial W_f} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t)) CS_{t-1} \frac{\partial F_t}{\partial f_t} \frac{\partial f_t}{\partial W_f}$$

$$\frac{\partial F_t}{\partial f_t} = F_t(1 - F_t)$$

$$\frac{\partial f_t}{\partial W_f} = X_t$$



신박AI

$\partial L/\partial W_f$ 식이 완성 되었습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y)W_y O_t(1 - tanh^2(CS_t))$

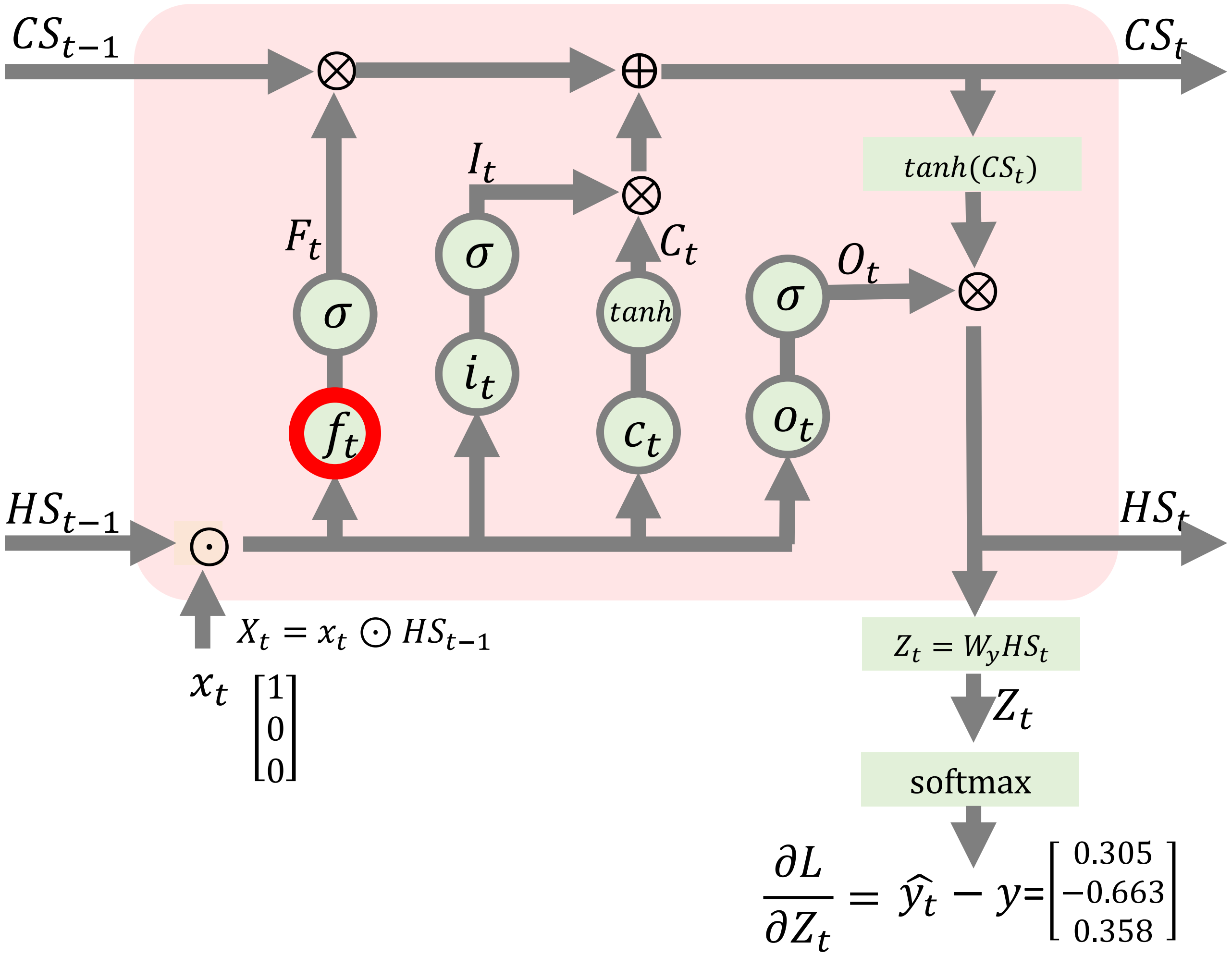
Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial W_f} = (\hat{y}_t - y)W_y O_t(1 - tanh^2(CS_t))CS_{t-1}F_t(1 - F_t)X_t$$



신박AI

자 이제 숫자를 넣어보도록 하겠습니다

Forget Gate: $F_t = \sigma(W_f X_t)$

$F_t = \sigma(f_t)$

$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$

$I_t = \sigma(i_t)$

$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$

$C_t = tanh(c_t)$

$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$

$O_t = \sigma(o_t)$

$o_t = W_o X_t$

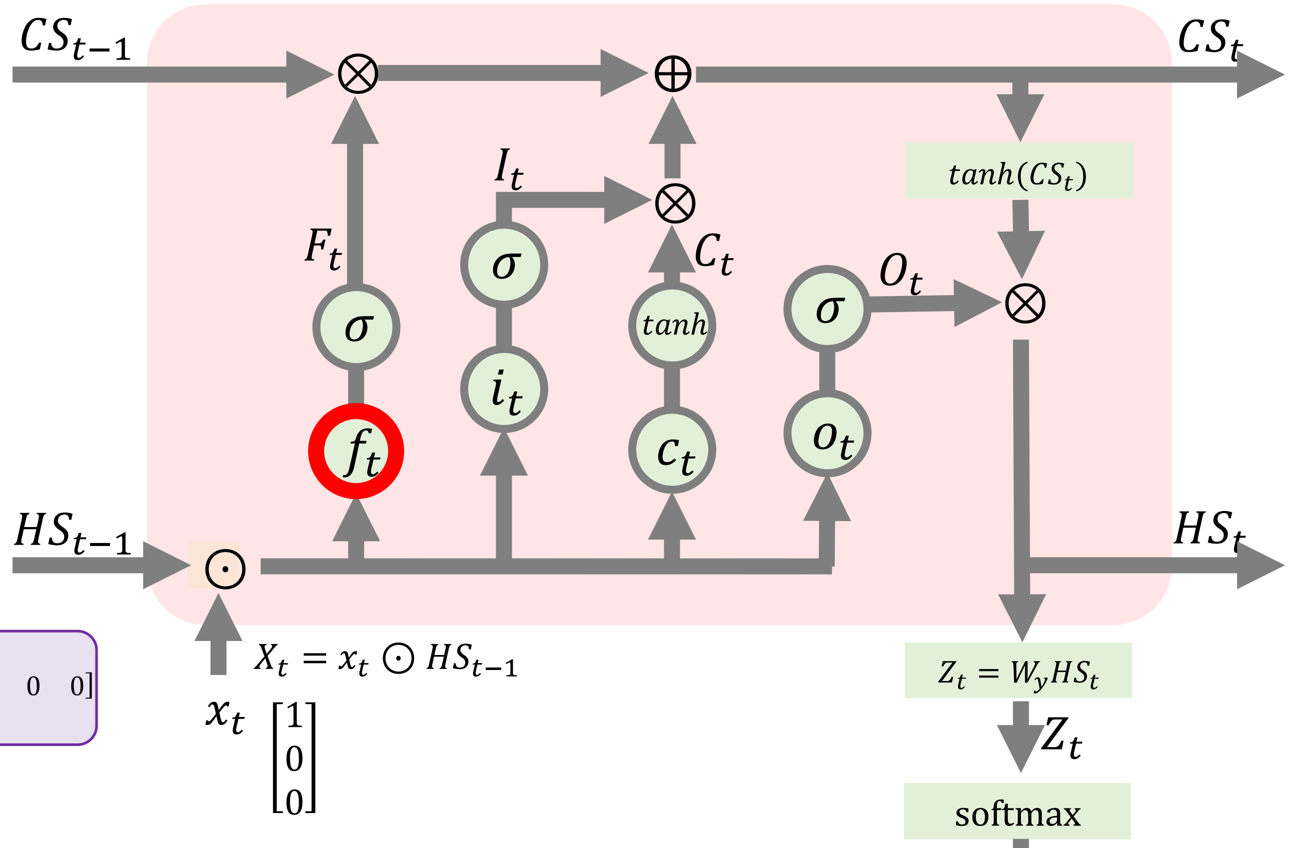
$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial W_f} = \boxed{(\hat{y}_t - y) W_y} \boxed{O_t} \boxed{(1 - tanh^2(CS_t))} \boxed{CS_{t-1}} \boxed{F_t(1 - F_t)} \boxed{X_t}$$

$$= \boxed{\left(\begin{bmatrix} 0.305 & -0.663 & 0.358 \end{bmatrix} \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix} \right)^T} \boxed{\begin{bmatrix} 0.401 \\ 0.634 \end{bmatrix}} \boxed{\begin{bmatrix} 0.912 \\ 0.936 \end{bmatrix}} \boxed{\begin{bmatrix} 1 \\ 0 \end{bmatrix}} \boxed{\begin{bmatrix} 0.25 \\ 0.209 \end{bmatrix}} \boxed{\begin{bmatrix} 0 & 0 & 1 & 0 & 0 \end{bmatrix}}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix} 0.305 \\ -0.663 \\ 0.358 \end{bmatrix}$$

이렇게 $\partial L/\partial W_f$을 계산해보았습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$$

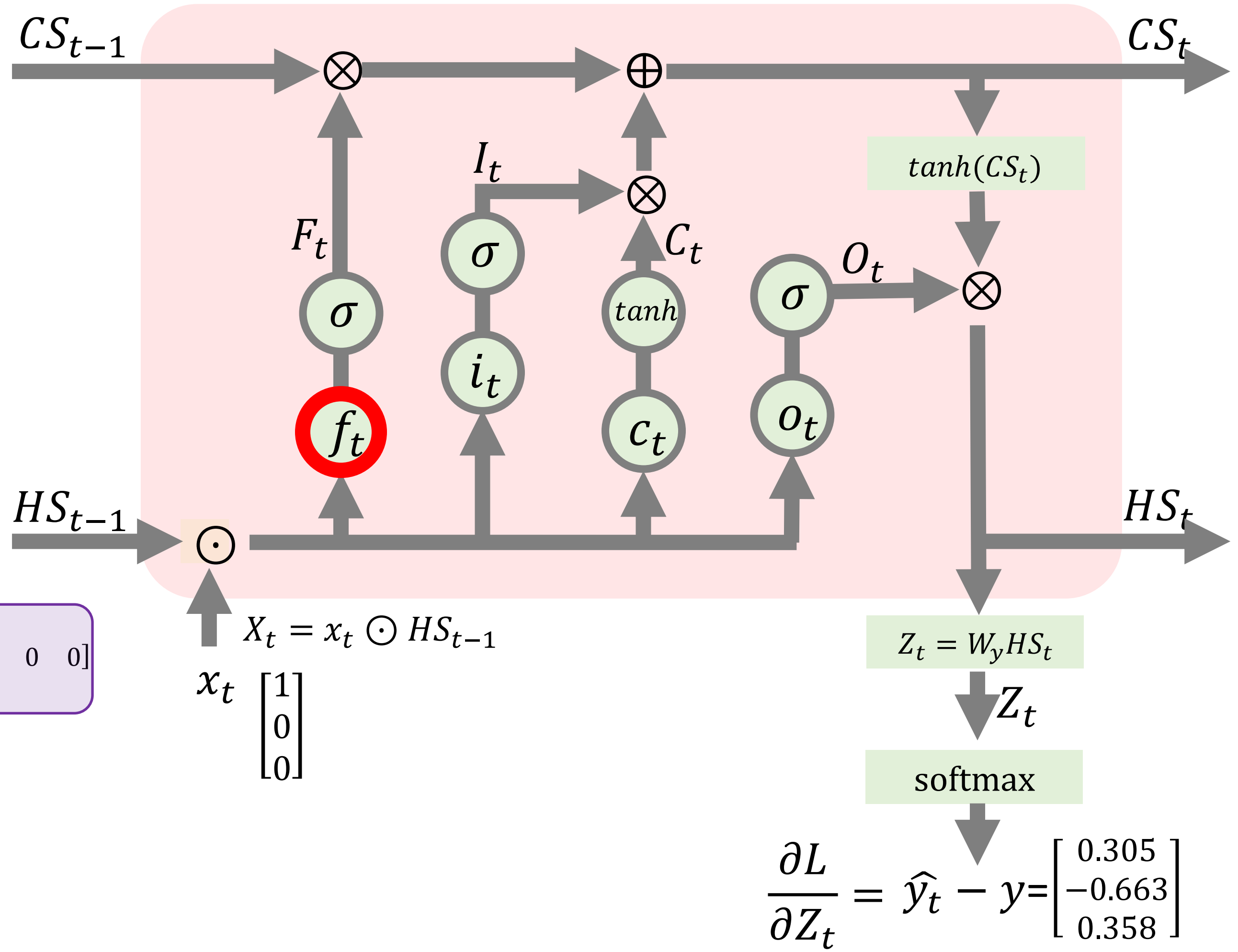
$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial W_f} = \boxed{(\hat{y}_t - y)W_y}\boxed{O_t}\boxed{(1 - tanh^2(CS_t))}\boxed{CS_{t-1}}\boxed{F_t(1-F_t)}\boxed{X_t}$$

$$= \boxed{\left([0.305 \quad -0.663 \quad 0.358]\begin{bmatrix}0.32 & -0.172\\0.449 & 0.349\\0.914 & 0.371\end{bmatrix}\right)^T}\boxed{\begin{bmatrix}0.401\\0.634\end{bmatrix}}\boxed{\begin{bmatrix}0.912\\0.936\end{bmatrix}}\boxed{\begin{bmatrix}1\\0\end{bmatrix}}\boxed{\begin{bmatrix}0.25\\0.209\end{bmatrix}}\boxed{[0 \quad 0 \quad 1 \quad 0 \quad 0]}$$

$$= \begin{bmatrix}0 & 0 & 0.012 & 0 & 0\\0 & 0 & 0 & 0 & 0\end{bmatrix}$$



신박AI

이젠 $\partial L/\partial W_i$ 를 계산할 차례입니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$$

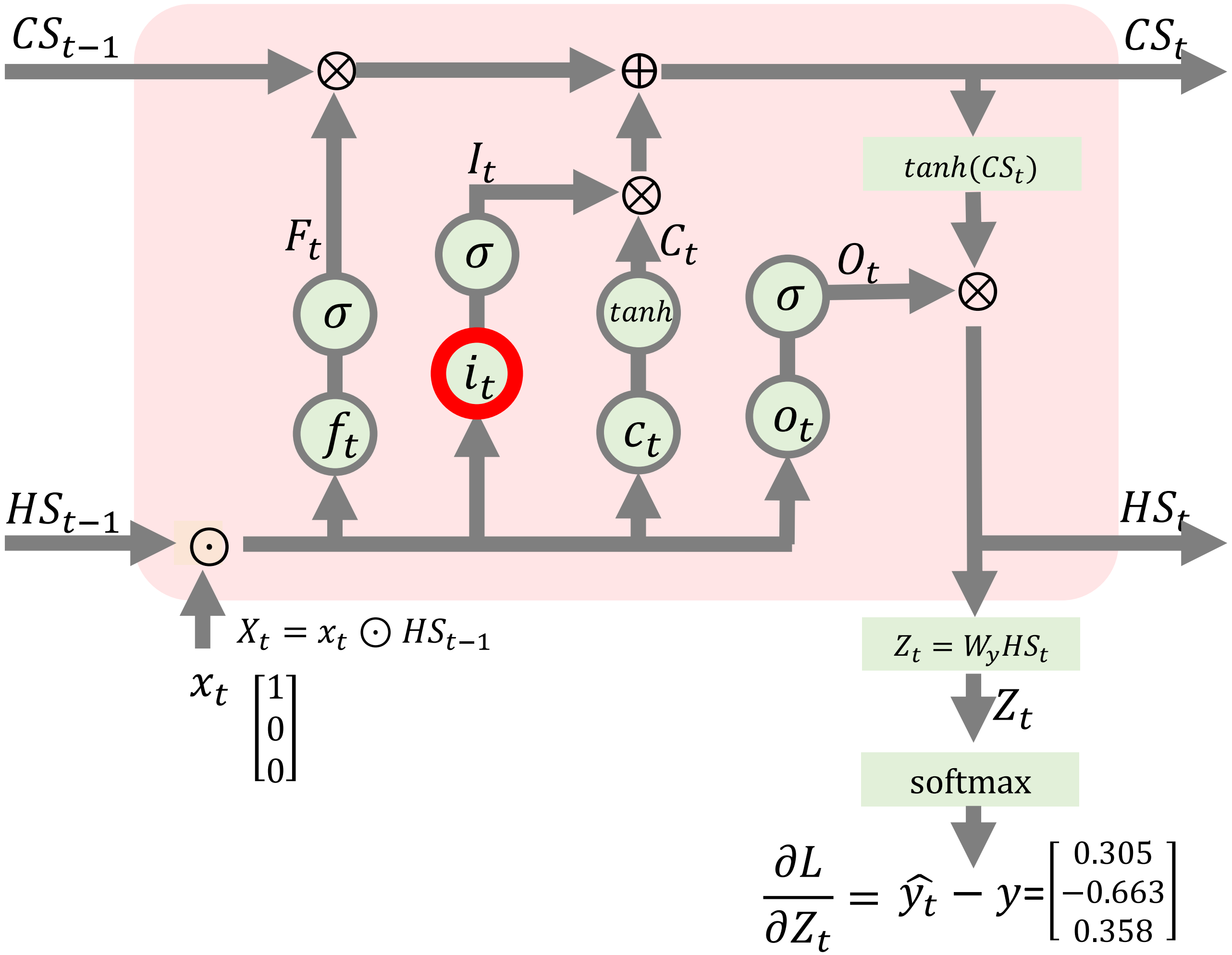
Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial W_i} =$$



신박AI

$\partial L/\partial W_i$ 는 체인룰에 의해서 다음과 같이 전개할 수 있습니다

Forget Gate: $F_t = \sigma(W_f X_t)$

$F_t = \sigma(f_t)$

$f_t = W_f X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$

$C_t = tanh(c_t)$

$c_t = W_c X_t$

Input Gate: $I_t = \sigma(W_i X_t)$

$I_t = \sigma(i_t)$

$i_t = W_i X_t$

Output Gate: $O_t = \sigma(W_o X_t)$

$O_t = \sigma(o_t)$

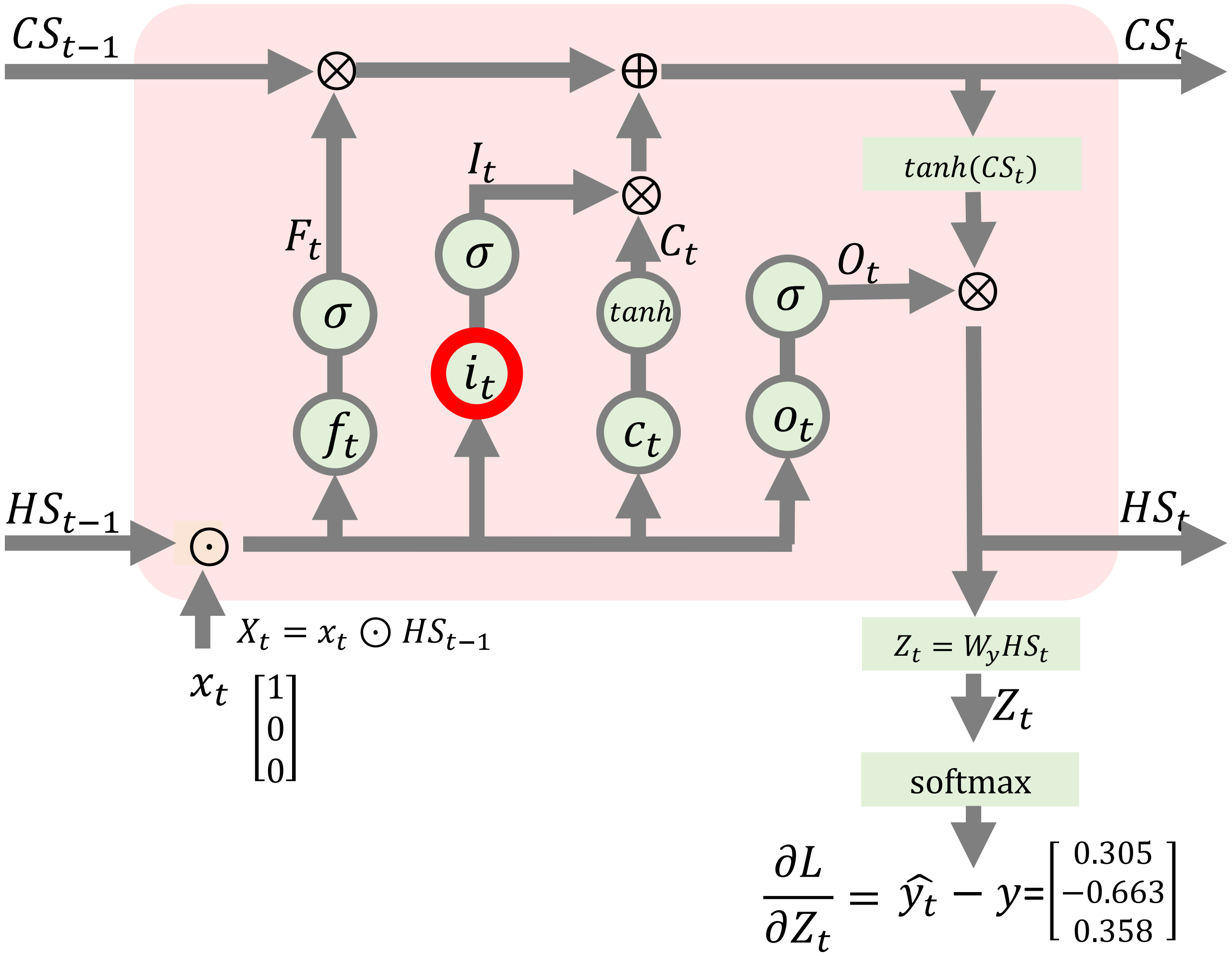
$o_t = W_o X_t$

$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y)W_y O_t(1 - tanh^2(CS_t))$$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial W_i} = \frac{\partial L}{\partial CS_t}\frac{\partial CS_t}{\partial I_t}\frac{\partial I_t}{\partial i_t}\frac{\partial i_t}{\partial W_i}$$



신박AI

$\partial L/\partial CS_t$ 는 앞서 전개한 이 공식을 사용할 수 있습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

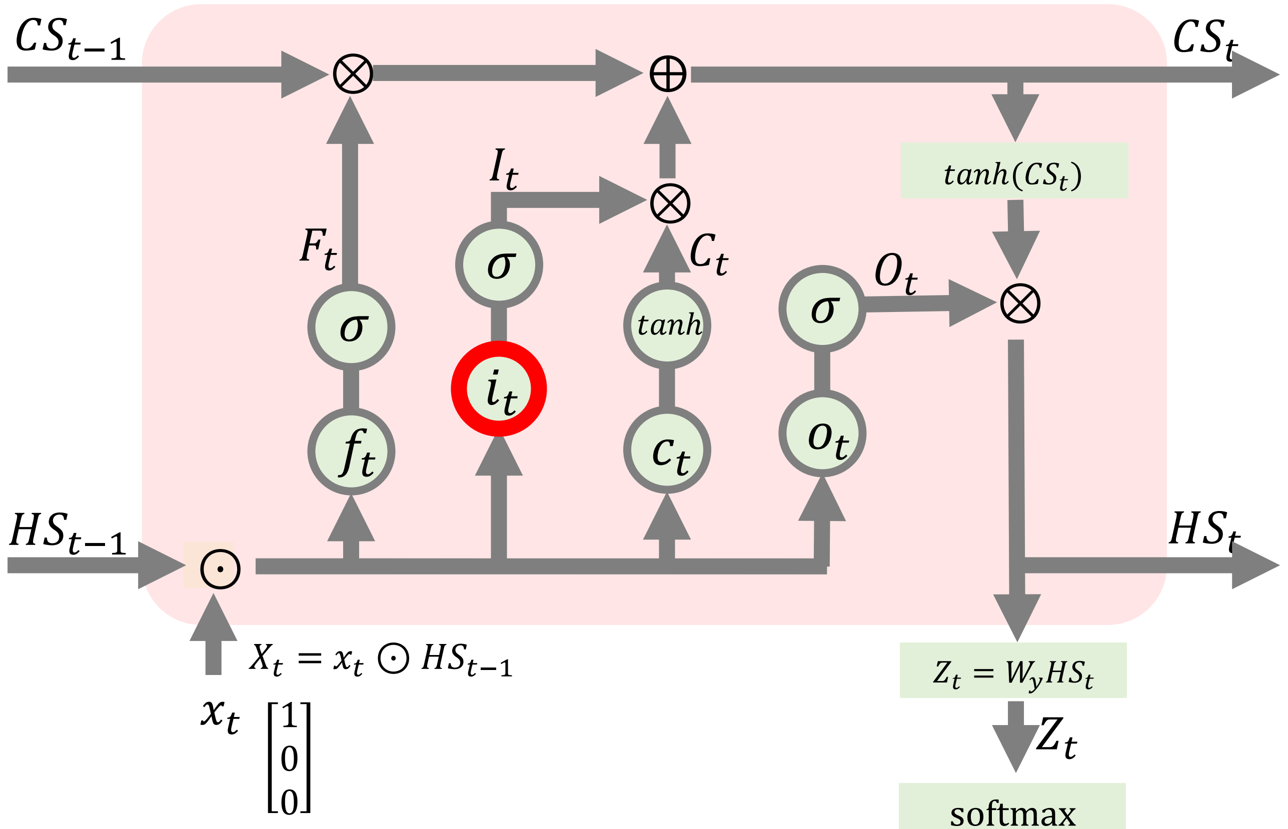
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y)W_y O_t(1 - tanh^2(CS_t))$$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial W_i} = \frac{\partial L}{\partial CS_t}\frac{\partial CS_t}{\partial I_t}\frac{\partial I_t}{\partial i_t}\frac{\partial i_t}{\partial W_i}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix}0.305\\-0.663\\0.358\end{bmatrix}$$

신박AI

$\partial CS_t/\partial I_t$ 는 앞서 보여드렸던 CS_t 공식을 미분하면 됩니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$

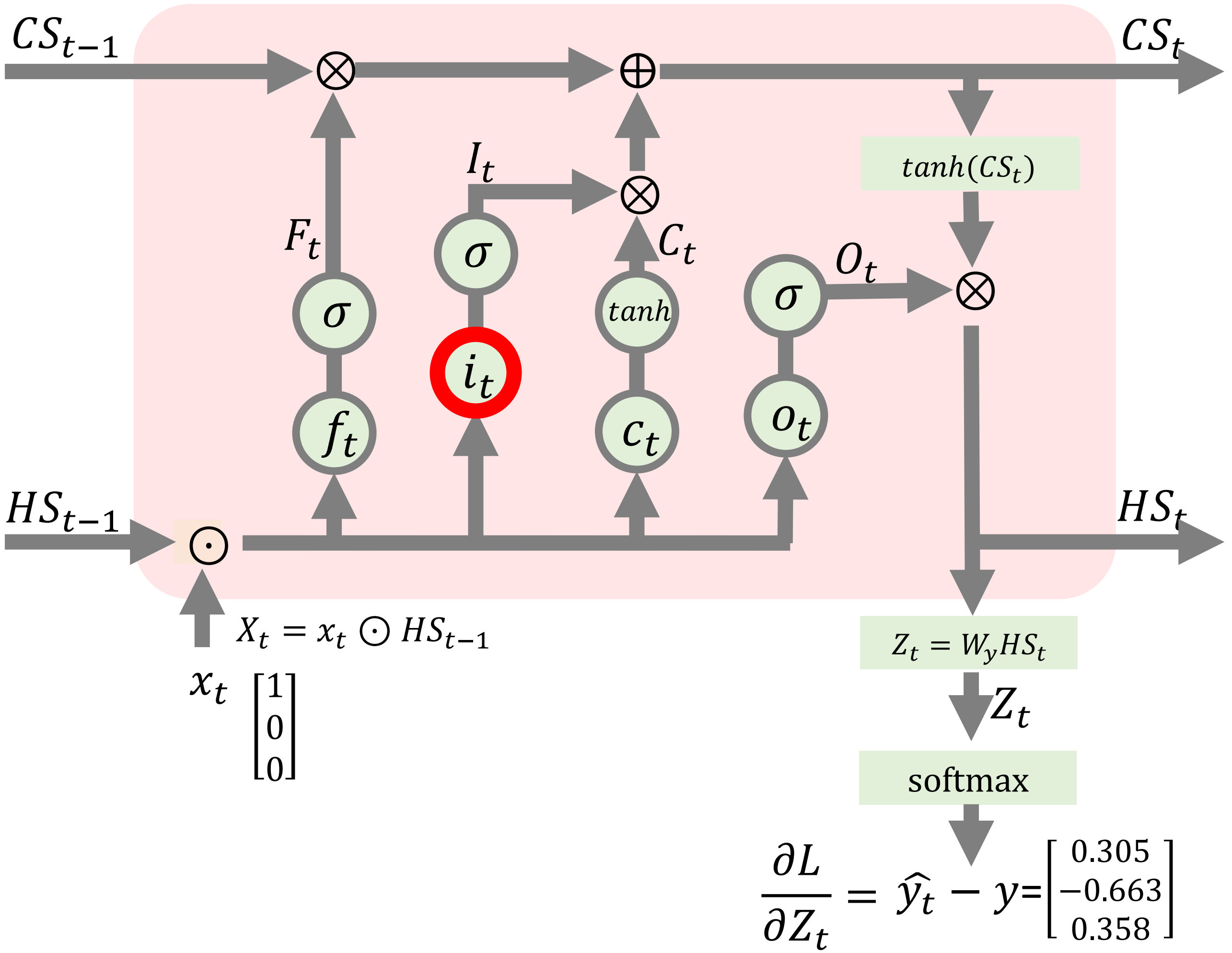
$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial W_i} = \frac{\partial L}{\partial CS_t} \frac{\partial CS_t}{\partial I_t} \frac{\partial I_t}{\partial i_t} \frac{\partial i_t}{\partial W_i}$$

$$\frac{\partial CS_t}{\partial I_t}$$

$$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$$



신박AI

그러면 $\partial CS_t/\partial I_t$ 는 C_t가 됩니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$

$Z_t = W_y HS_t$

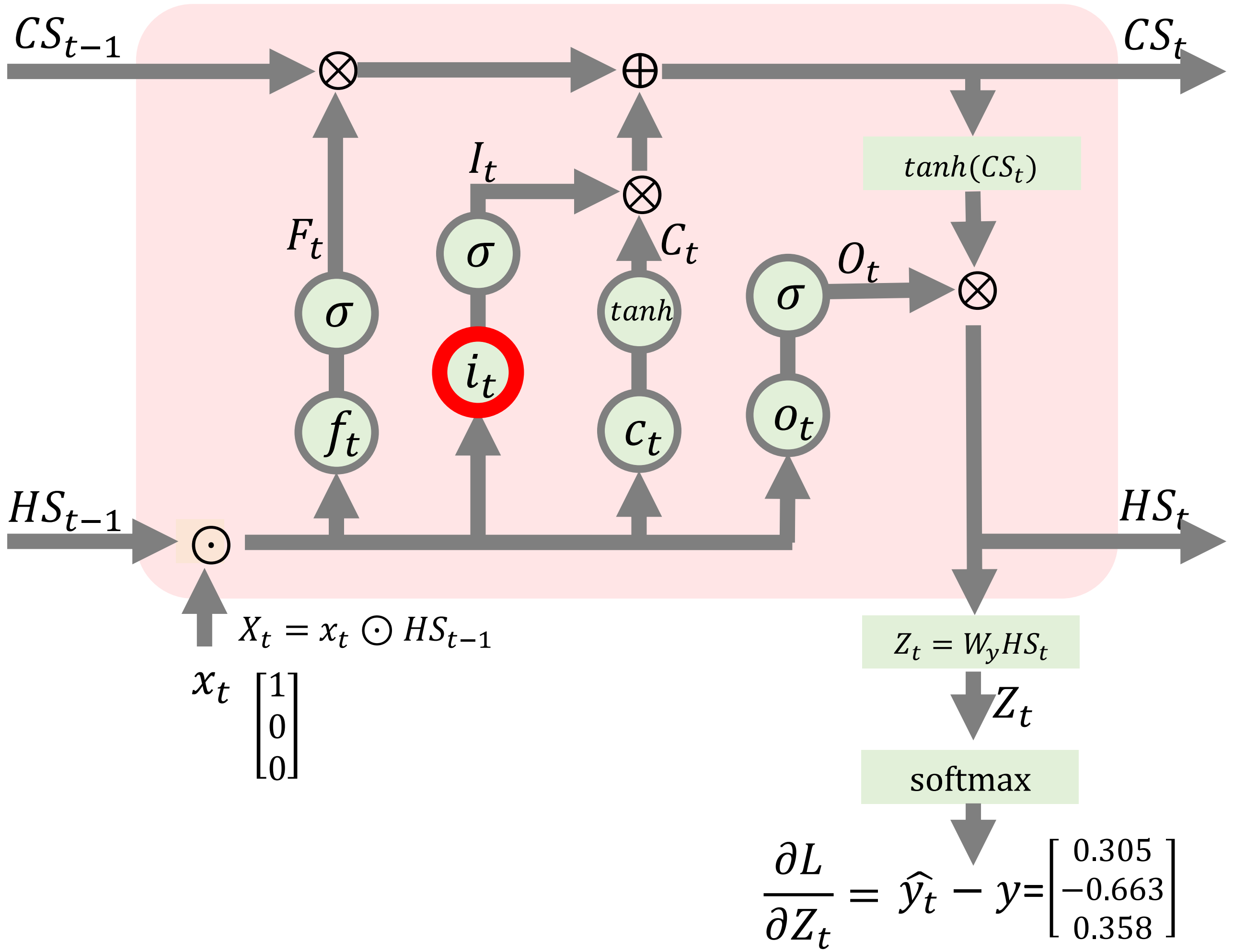
$\boxed{CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t}$

$$\frac{\partial L}{\partial W_i} = \frac{\partial L}{\partial CS_t} \frac{\partial CS_t}{\partial I_t} \frac{\partial I_t}{\partial i_t} \frac{\partial i_t}{\partial W_i}$$

$$\frac{\partial CS_t}{\partial I_t}$$

$$CS_t = \cancel{CS_{t-1}} \otimes \cancel{F_t} + \cancel{I_t} \otimes C_t$$

$$\frac{\partial CS_t}{\partial I_t} = C_t$$



신박AI

그러면 도출한 식들을 $\partial L/\partial W_i$에 넣고 다시 식을 작성해보겠습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$$\frac{\partial L}{\partial CS_t} = \boxed{(\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))}$$

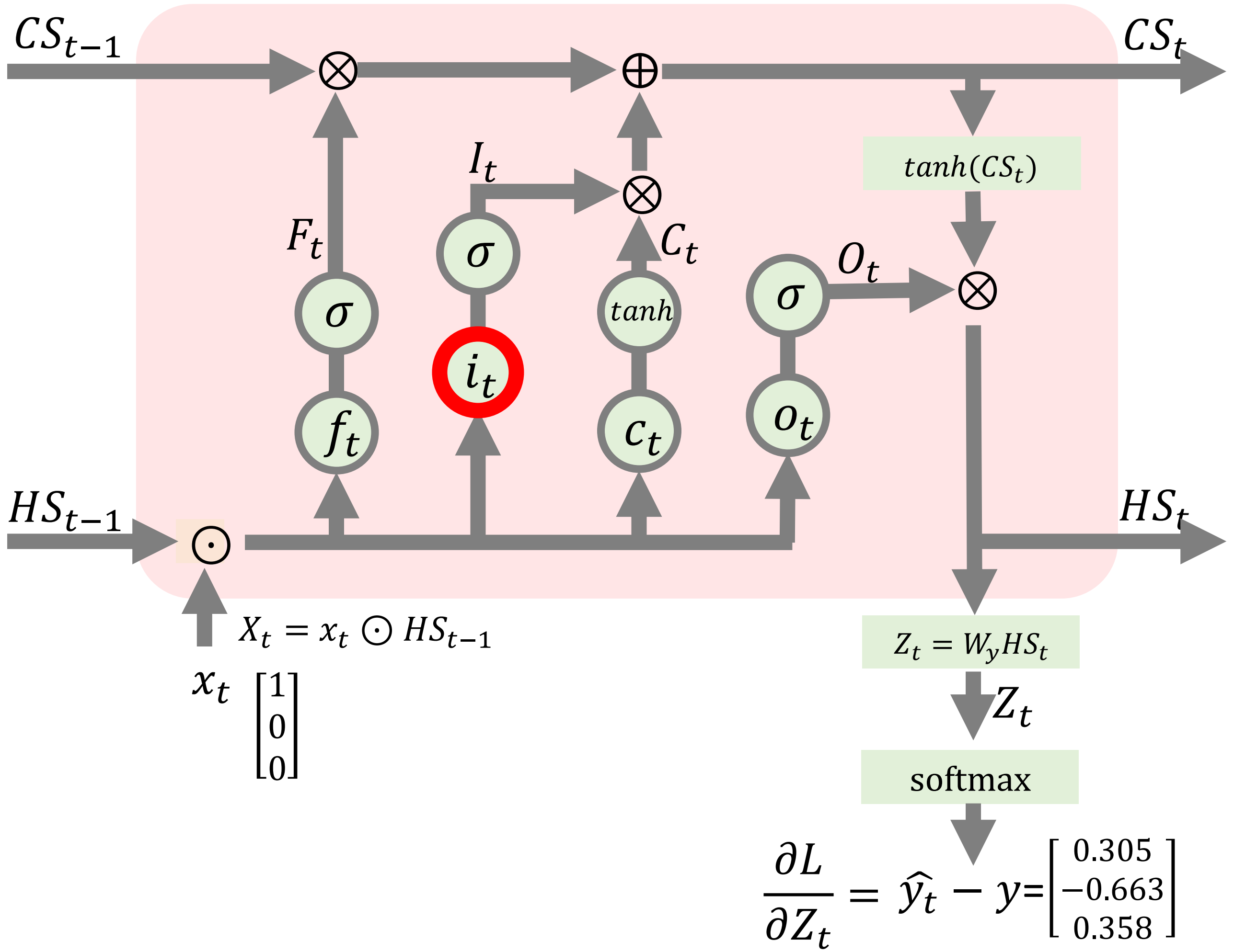
$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial W_i} = \frac{\partial L}{\partial CS_t}\frac{\partial CS_t}{\partial I_t}\frac{\partial I_t}{\partial i_t}\frac{\partial i_t}{\partial W_i}$$

$$\frac{\partial CS_t}{\partial I_t}$$

$$\frac{\partial CS_t}{\partial I_t} = \boxed{C_t}$$



그러면 $\partial L/\partial W_i$은 다음처럼 정리가 됩니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$

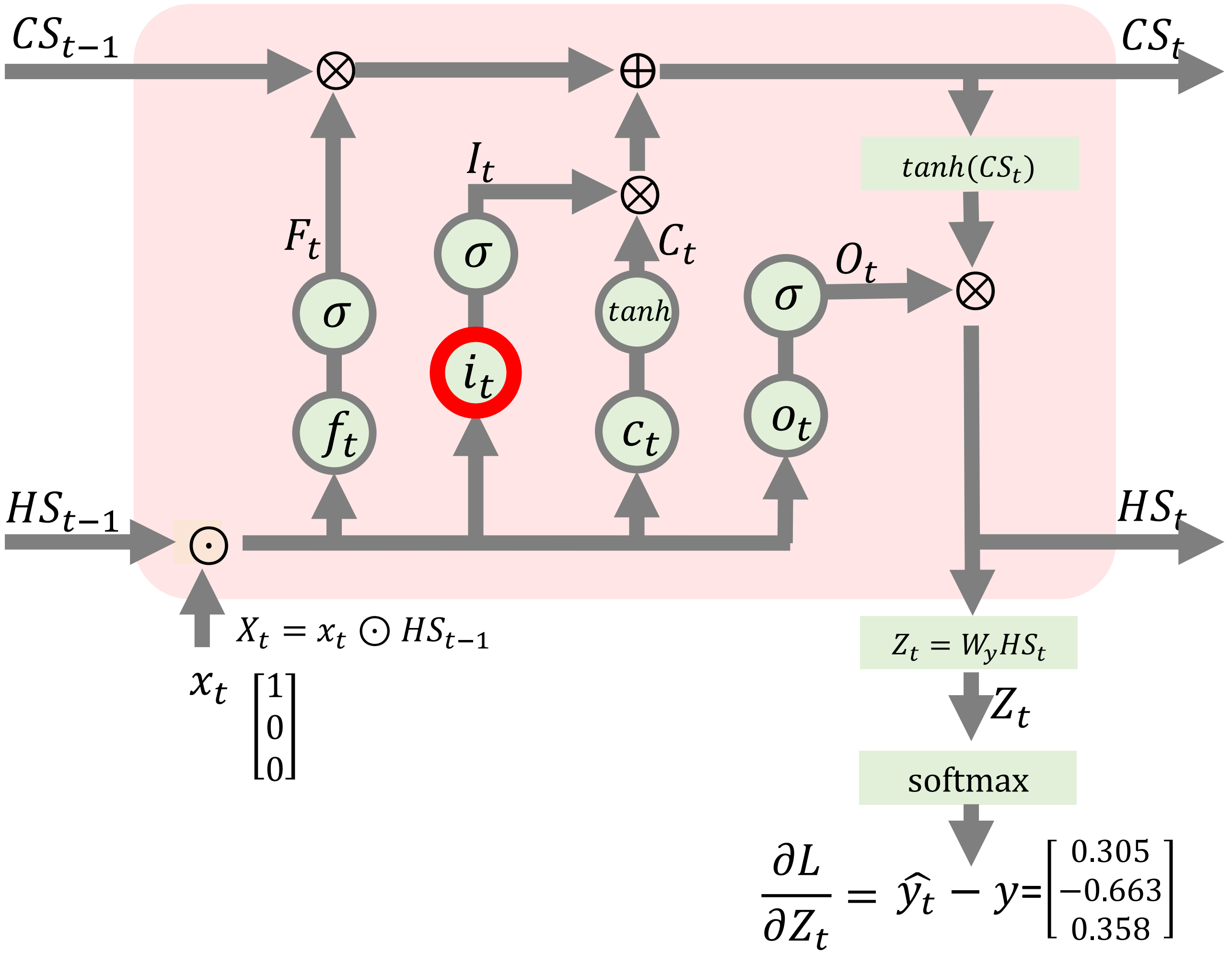
Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial W_i} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t)) C_t \frac{\partial I_t}{\partial i_t} \frac{\partial i_t}{\partial W_i}$$



신박AI

그 다음은 $\partial I_t / \partial i_t$를 구해보겠습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

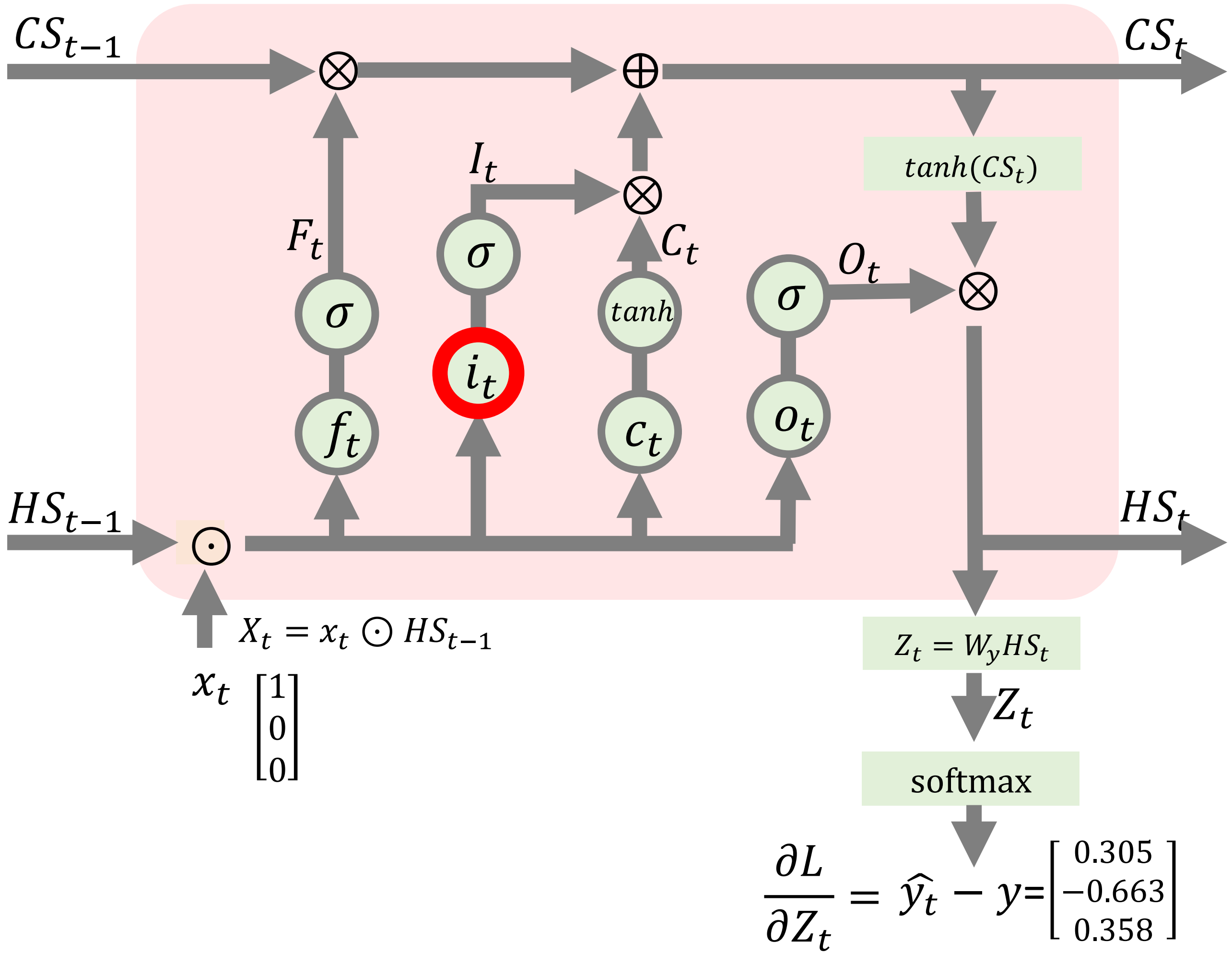
$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial W_i} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t)) C_t \boxed{\frac{\partial I_t}{\partial i_t}} \frac{\partial i_t}{\partial W_i}$$

$$\frac{\partial I_t}{\partial i_t}$$



신박AI

$\partial I_t/\partial i_t$은 시그모이드 미분 함수에 의해서

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$\boxed{I_t = \sigma(i_t)}$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

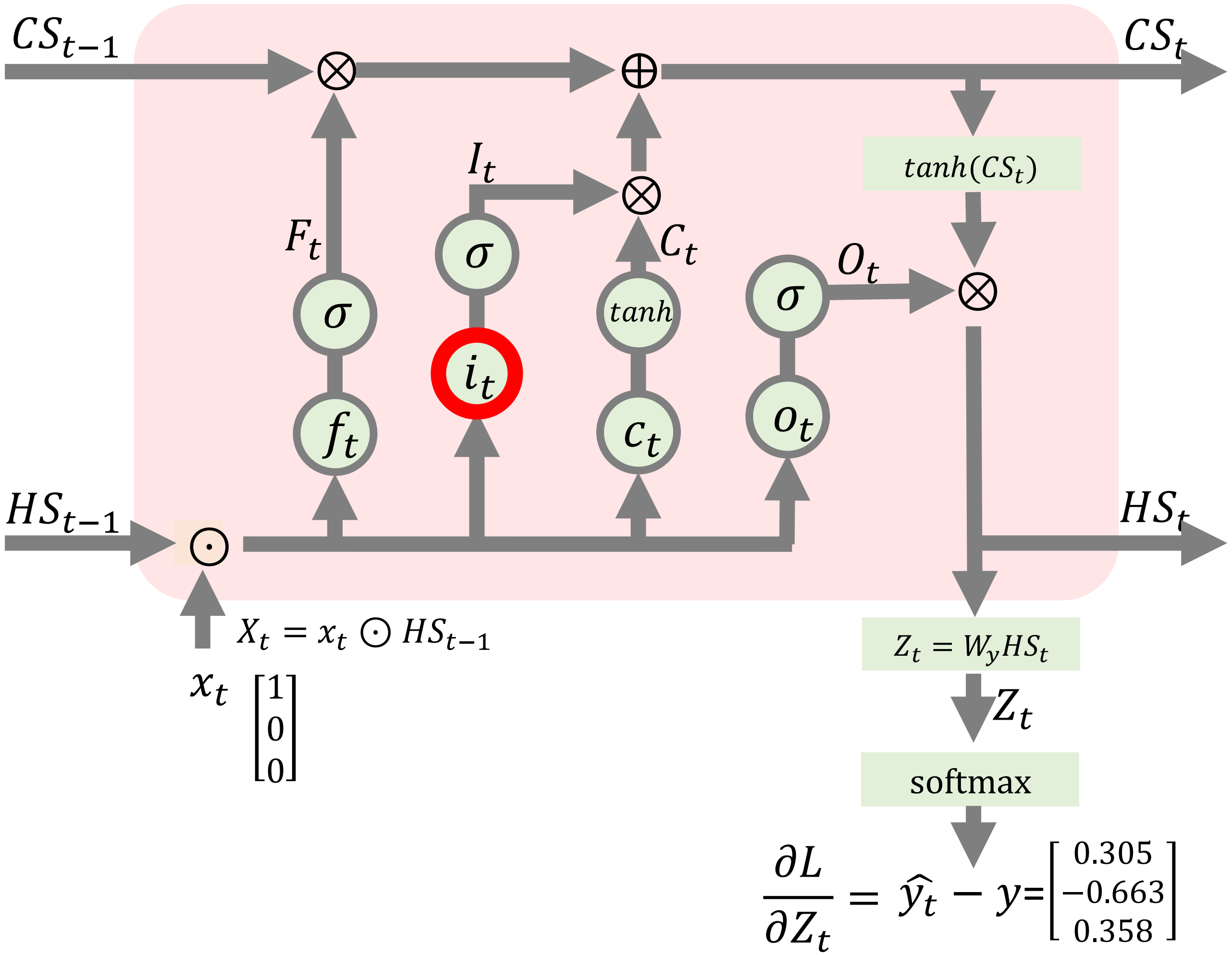
$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y)W_y O_t(1 - tanh^2(CS_t))$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial W_i} = (\hat{y}_t - y)W_y O_t(1 - tanh^2(CS_t))C_t \boxed{\frac{\partial I_t}{\partial i_t}} \frac{\partial i_t}{\partial W_i}$$

$$\frac{\partial I_t}{\partial i_t}$$



신박AI

이렇게 구할 수가 있고,

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

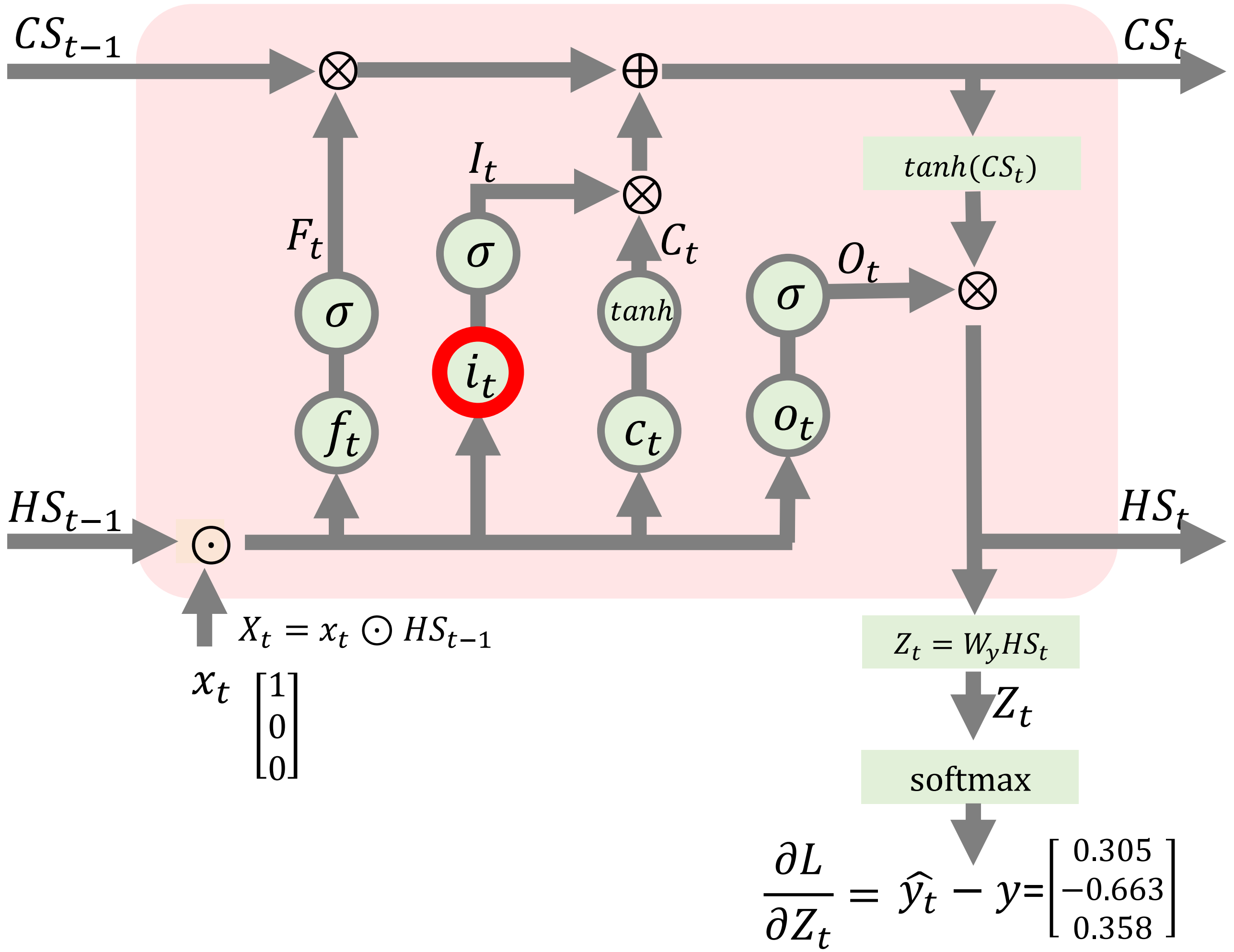
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial W_i} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t)) C_t \frac{\partial I_t}{\partial i_t} \frac{\partial i_t}{\partial W_i}$$

$$\boxed{\frac{\partial I_t}{\partial i_t} = I_t(1 - I_t)}$$



이어서 $\partial i_t / \partial W_i$는 이 공식에 의해서

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$\boxed{i_t = W_i X_t}$

$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

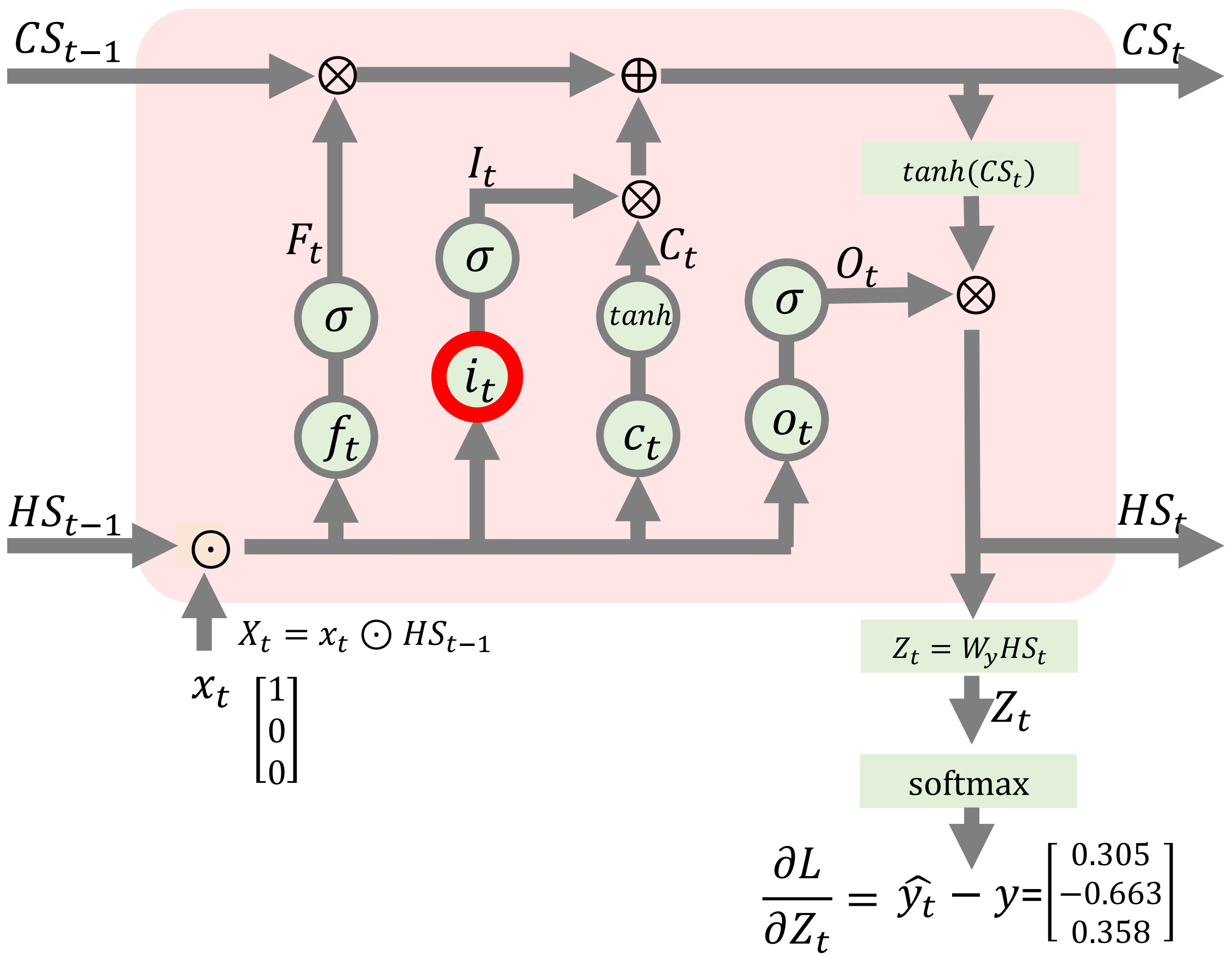
$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial W_i} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t)) C_t \frac{\partial I_t}{\partial i_t} \boxed{\frac{\partial i_t}{\partial W_i}}$$

$$\frac{\partial I_t}{\partial i_t} = I_t (1 - I_t)$$

$$\frac{\partial i_t}{\partial W_i}$$



신박AI

X_t로 구할 수 있습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$\boxed{i_t = W_i X_t}$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$

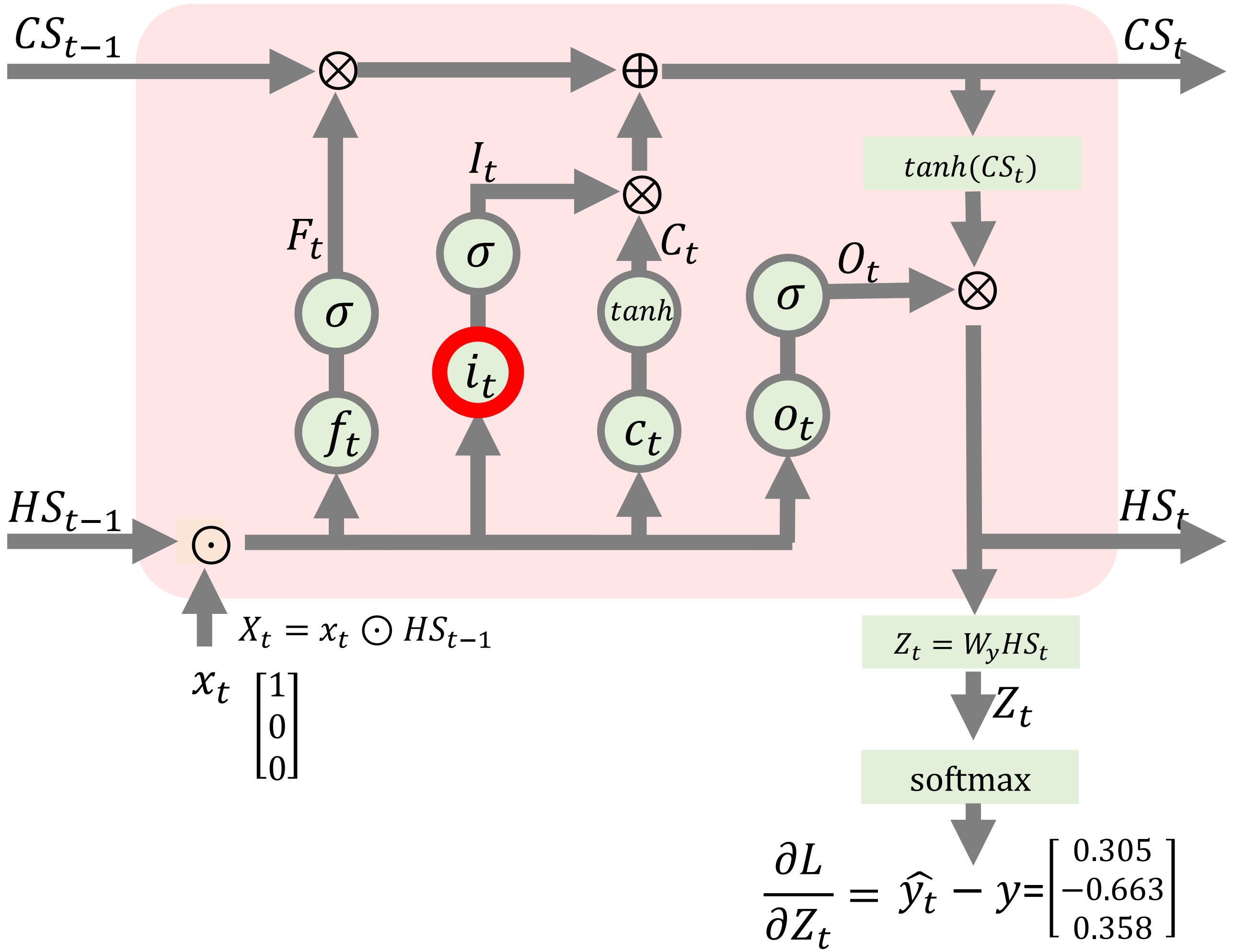
$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial W_i} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t)) C_t \frac{\partial I_t}{\partial i_t} \boxed{\frac{\partial i_t}{\partial W_i}}$$

$$\frac{\partial I_t}{\partial i_t} = I_t(1 - I_t)$$

$$\boxed{\frac{\partial i_t}{\partial W_i} = X_t}$$



신박AI

그러면 이 두 식을 $\partial L/\partial W_i$식에 넣으면,

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$

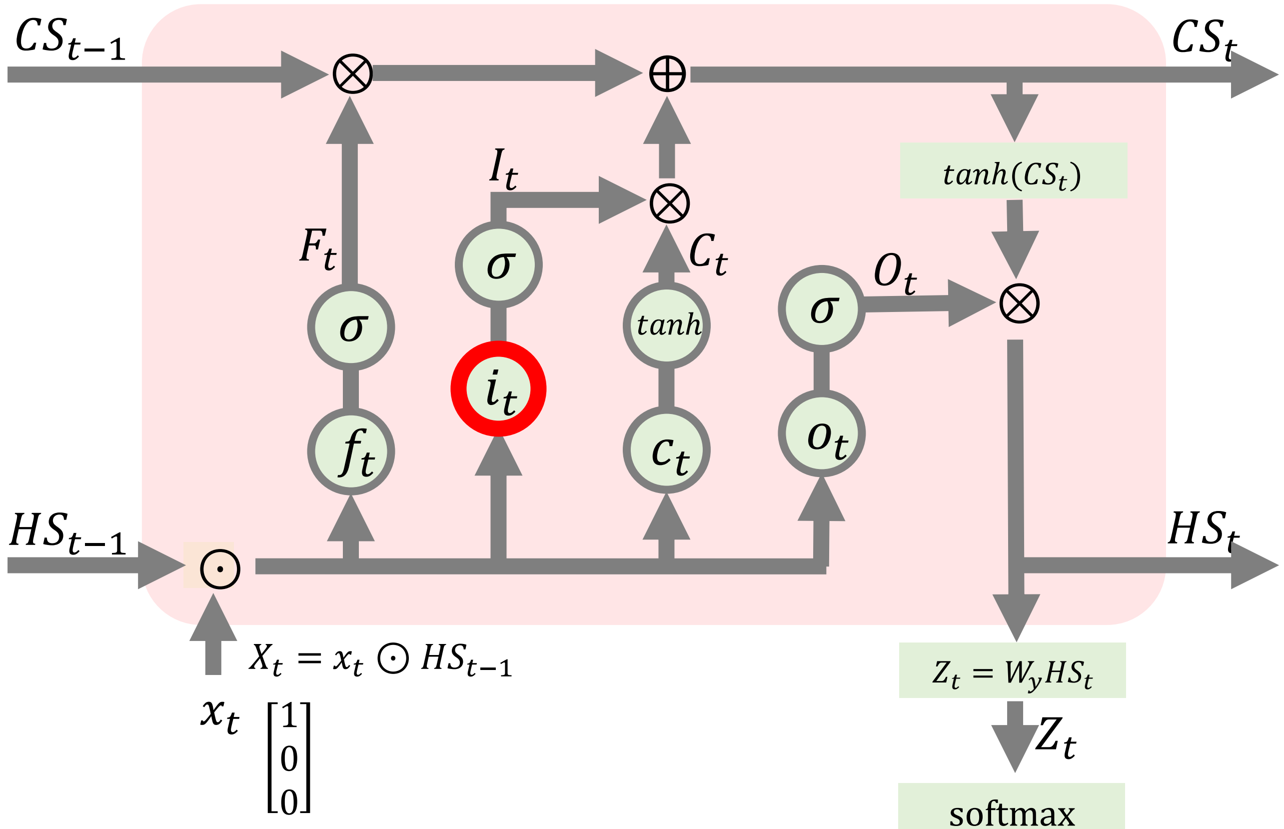
$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial W_i} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t)) C_t \frac{\partial I_t}{\partial i_t} \frac{\partial i_t}{\partial W_i}$$

$$\frac{\partial I_t}{\partial i_t} = I_t(1 - I_t)$$

$$\frac{\partial i_t}{\partial W_i} = X_t$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix}0.305\\-0.663\\0.358\end{bmatrix}$$

신박AI

$\partial L/\partial W_i$식이 완성 되었습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$$

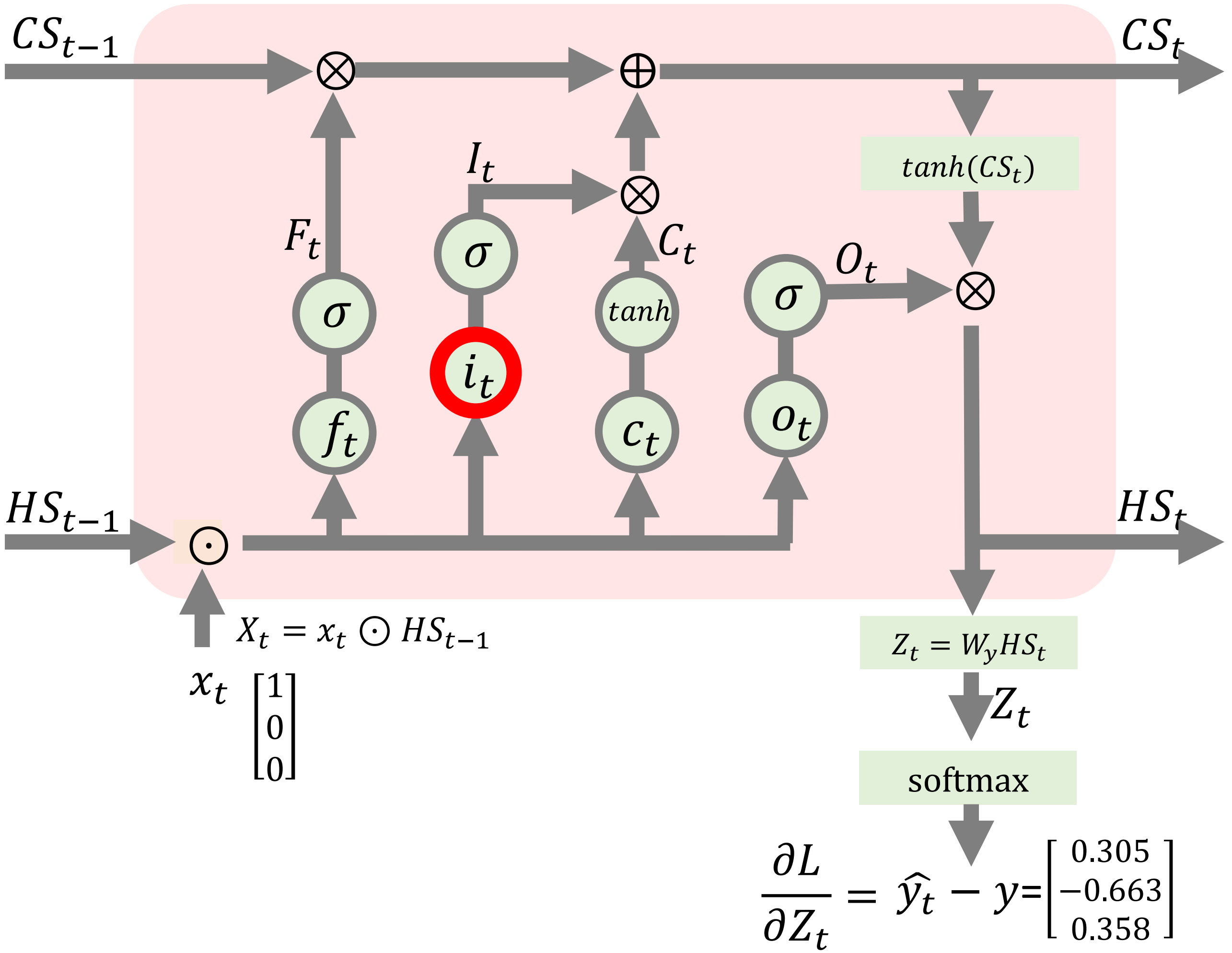
Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial W_i} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t)) C_t I_t (1 - I_t) X_t$$



자 이제 숫자를 넣어보도록 하겠습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

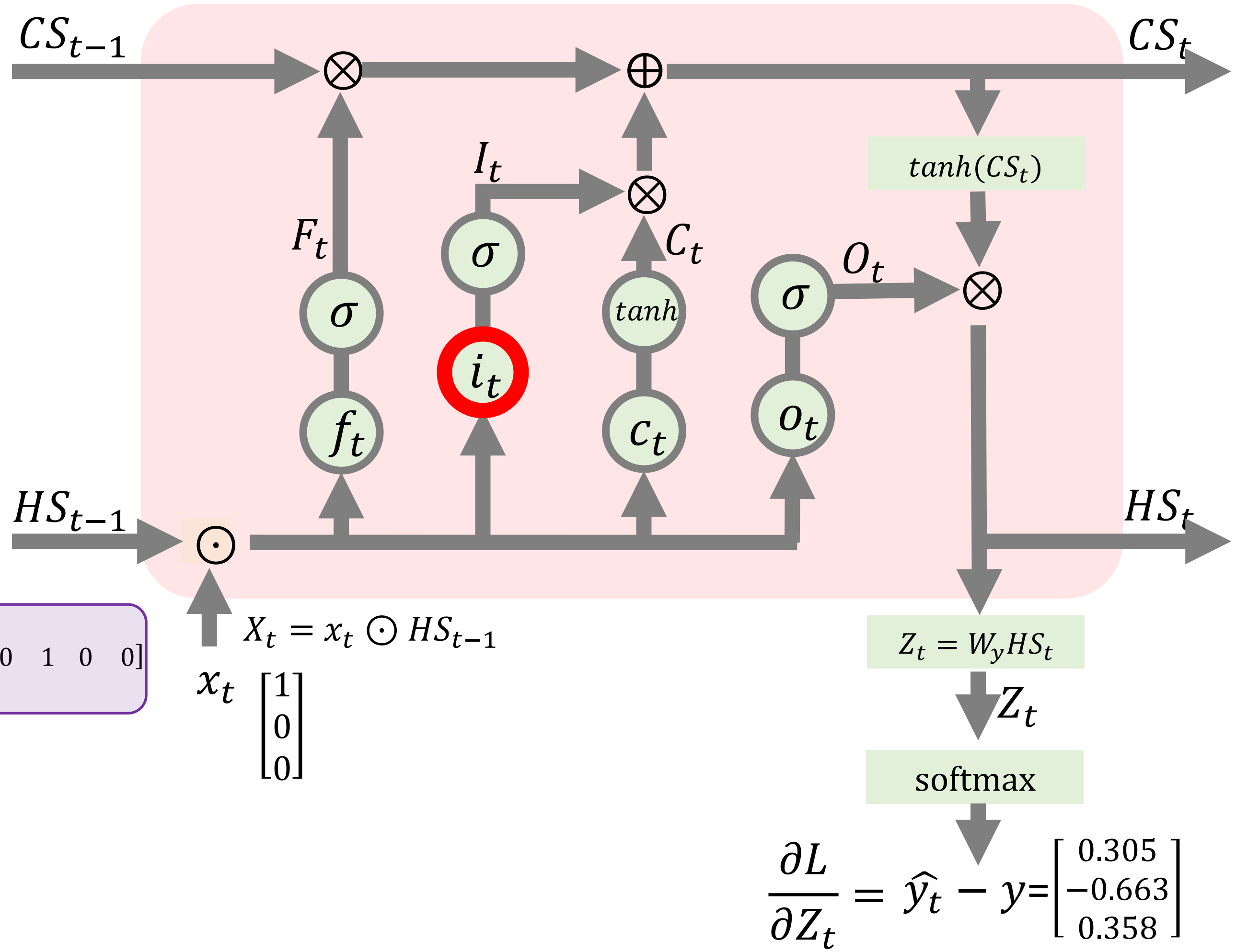
$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$$

$$Z_t = W_y HS_t$$

$$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$$

$$\frac{\partial L}{\partial W_i} = \boxed{(\hat{y}_t - y) W_y} \boxed{O_t} \boxed{(1 - tanh^2(CS_t))} \boxed{C_t} \boxed{I_t (1 - I_t)} \boxed{X_t}$$

$$= \left(\begin{bmatrix} 0.305 & -0.663 & 0.358 \end{bmatrix} \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix} \right)^T \begin{bmatrix} 0.401 \\ 0.634 \end{bmatrix} \begin{bmatrix} 0.912 \\ 0.936 \end{bmatrix} \begin{bmatrix} -0.391 \\ 0.646 \end{bmatrix} \begin{bmatrix} 0.25 \\ 0.24 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$



신박AI

이렇게 $\partial L/\partial W_i$을 계산해보았습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$

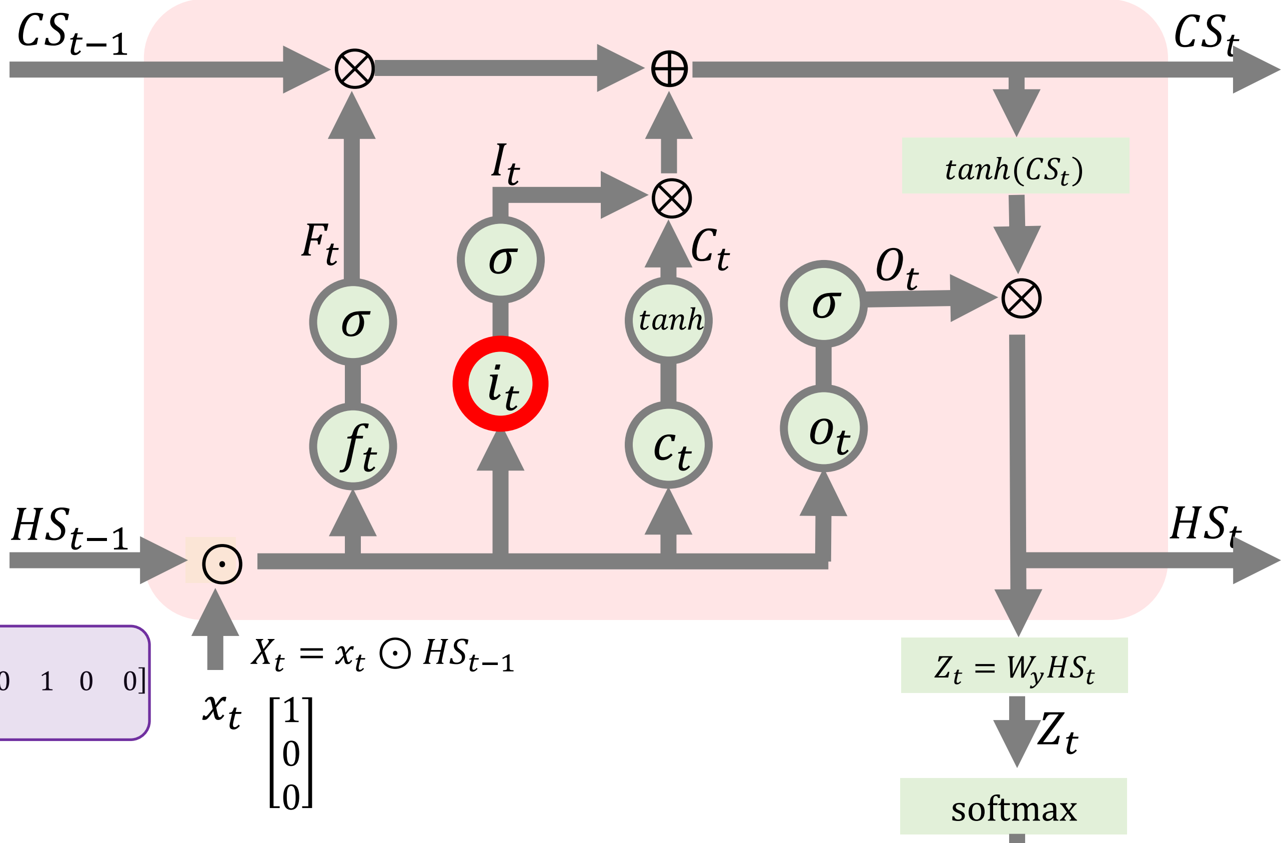
$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial W_i} = (\hat{y}_t - y)W_y O_t (1 - tanh^2(CS_t)) C_t I_t (1 - I_t) X_t$$

$$= \left(\begin{bmatrix} 0.305 & -0.663 & 0.358 \end{bmatrix} \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix} \right)^T \begin{bmatrix} 0.401 \\ 0.634 \end{bmatrix} \begin{bmatrix} 0.912 \\ 0.936 \end{bmatrix} \begin{bmatrix} -0.391 \\ 0.646 \end{bmatrix} \begin{bmatrix} 0.25 \\ 0.24 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 0 & -0.005 & 0 & 0 \\ 0 & 0 & -0.014 & 0 & 0 \end{bmatrix}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix} 0.305 \\ -0.663 \\ 0.358 \end{bmatrix}$$

이젠 $\partial L/\partial W_c$ 를 계산할 차례입니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

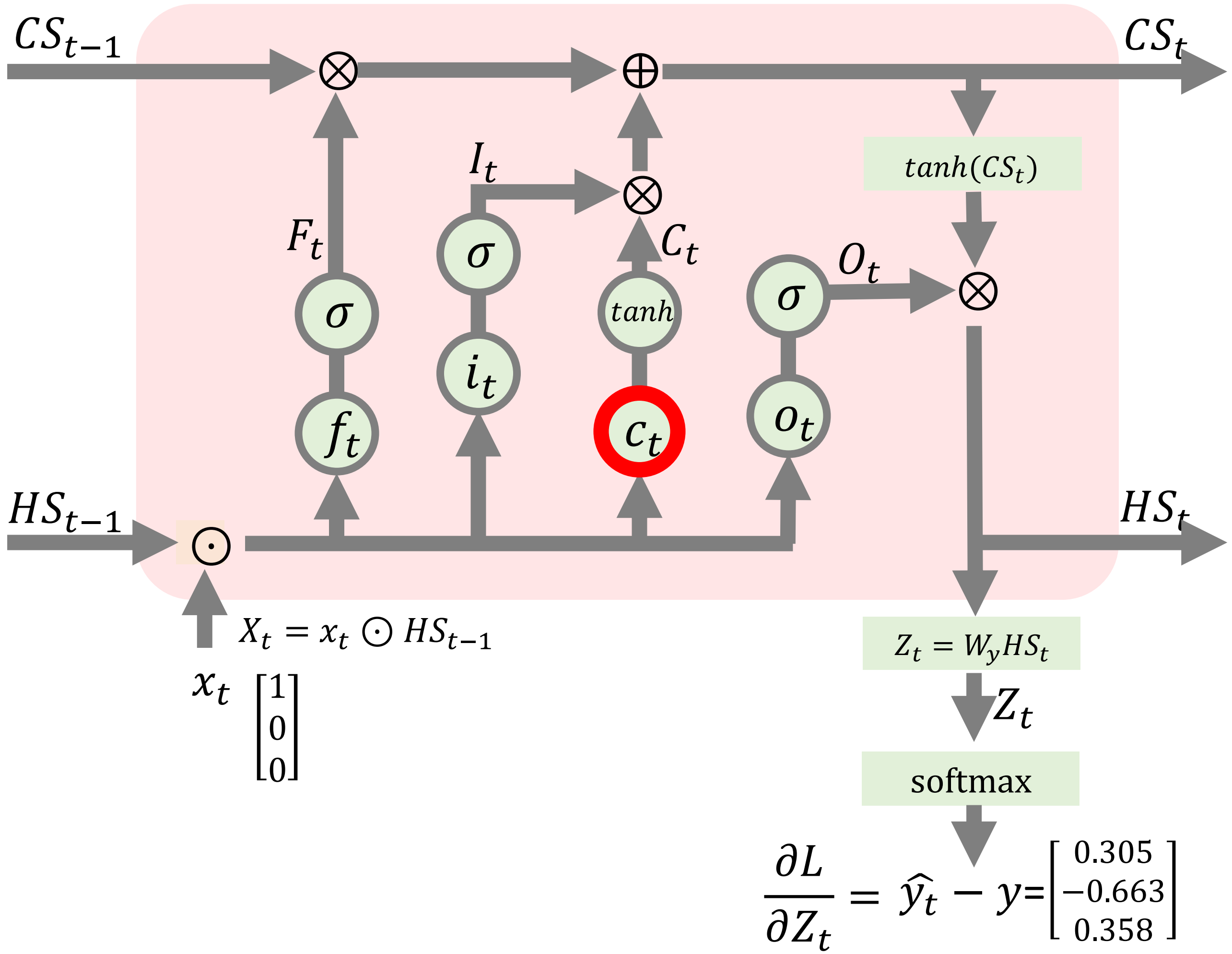
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial W_c} =$$



신박AI

$\partial L/\partial W_c$ 는 체인룰에 의해서 다음과 같이 전개할 수 있습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

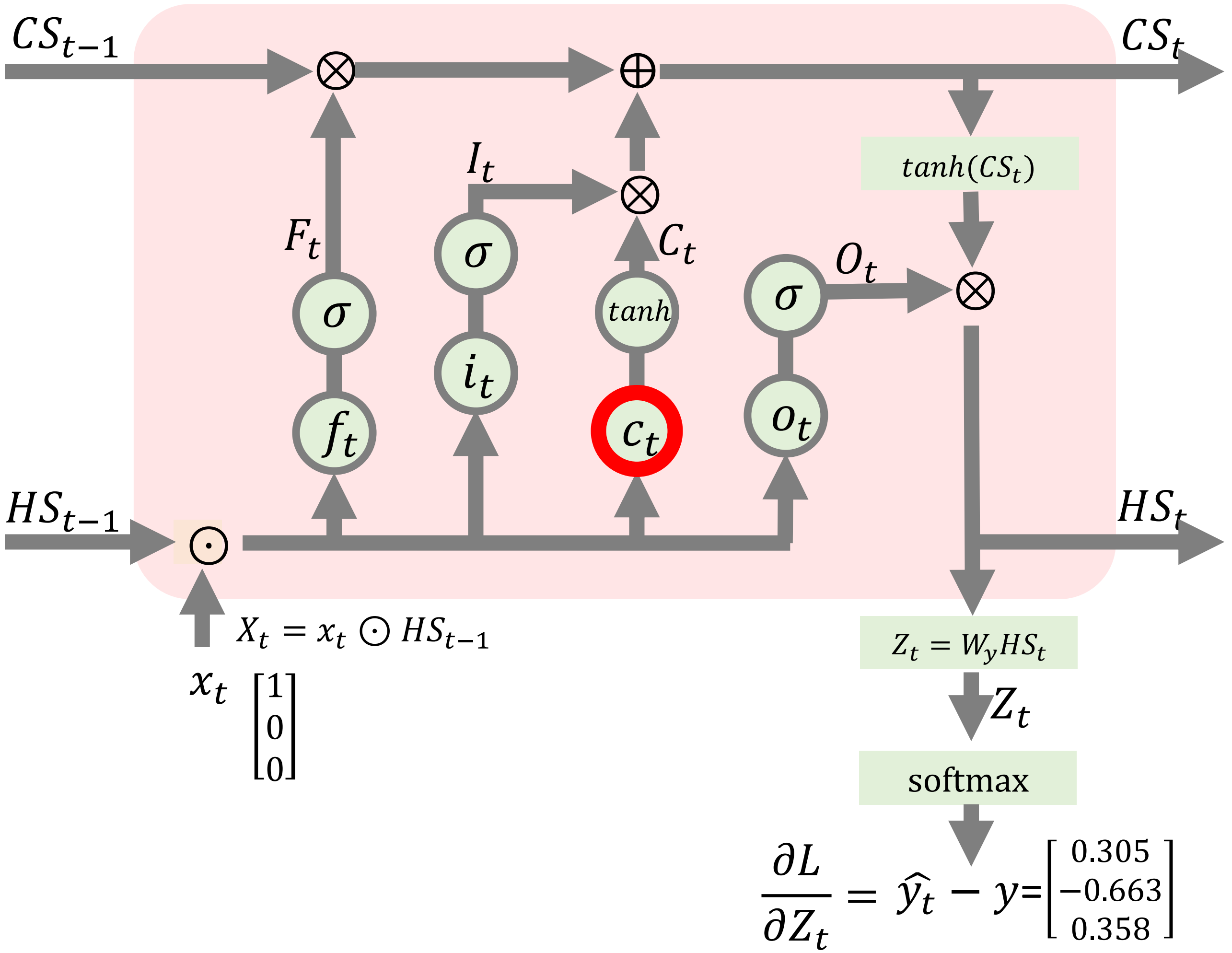
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y)W_y O_t(1 - tanh^2(CS_t))$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial W_c} = \frac{\partial L}{\partial CS_t}\frac{\partial CS_t}{\partial C_t}\frac{\partial C_t}{\partial c_t}\frac{\partial c_t}{\partial W_c}$$



$\partial L/\partial CS_t$ 는 앞서 전개한 이 공식을 사용할 수 있습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

$$\boxed{\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))}$$

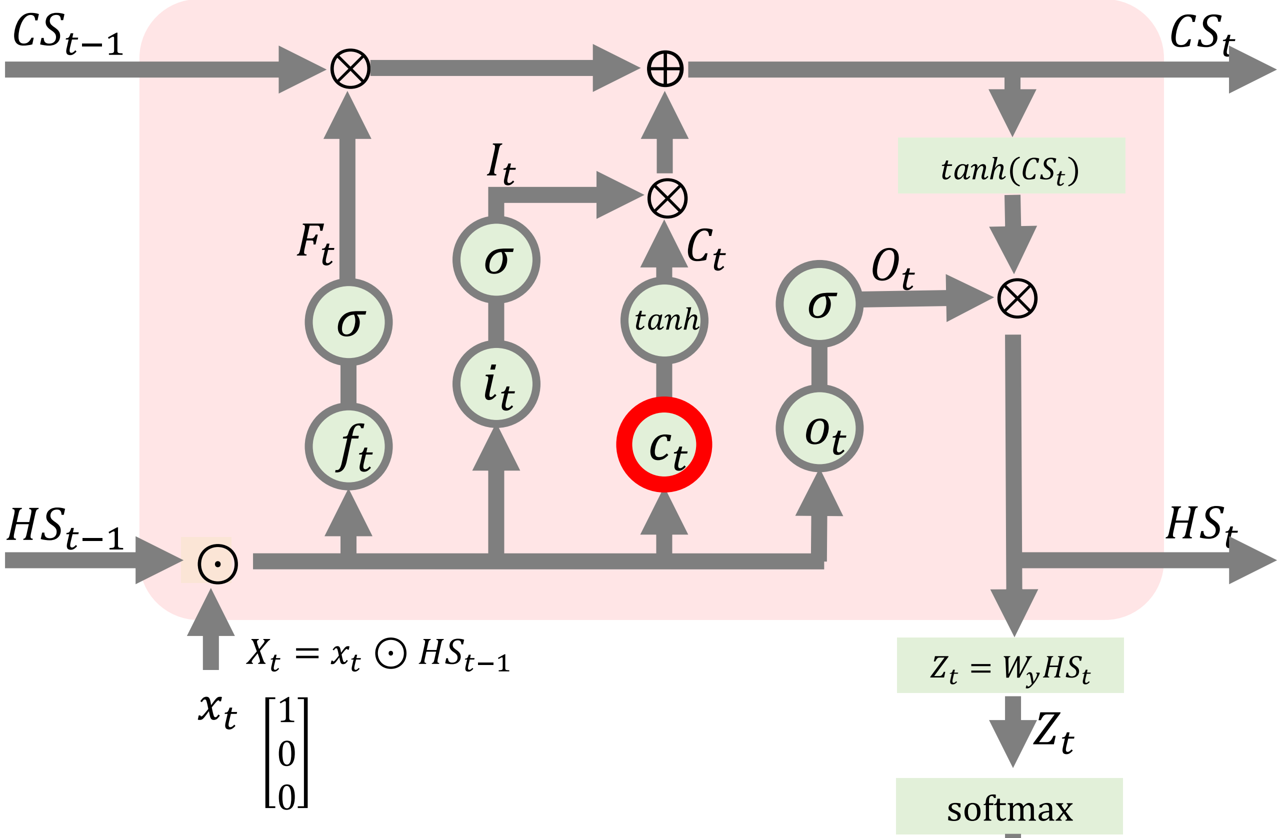
Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial W_c} = \frac{\partial L}{\partial CS_t} \frac{\partial CS_t}{\partial C_t} \frac{\partial C_t}{\partial c_t} \frac{\partial c_t}{\partial W_c}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix}0.305\\-0.663\\0.358\end{bmatrix}$$

$\partial CS_t / \partial C_t$ 는 앞서 보여드렸던 CS_t 공식을 미분하면 됩니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$

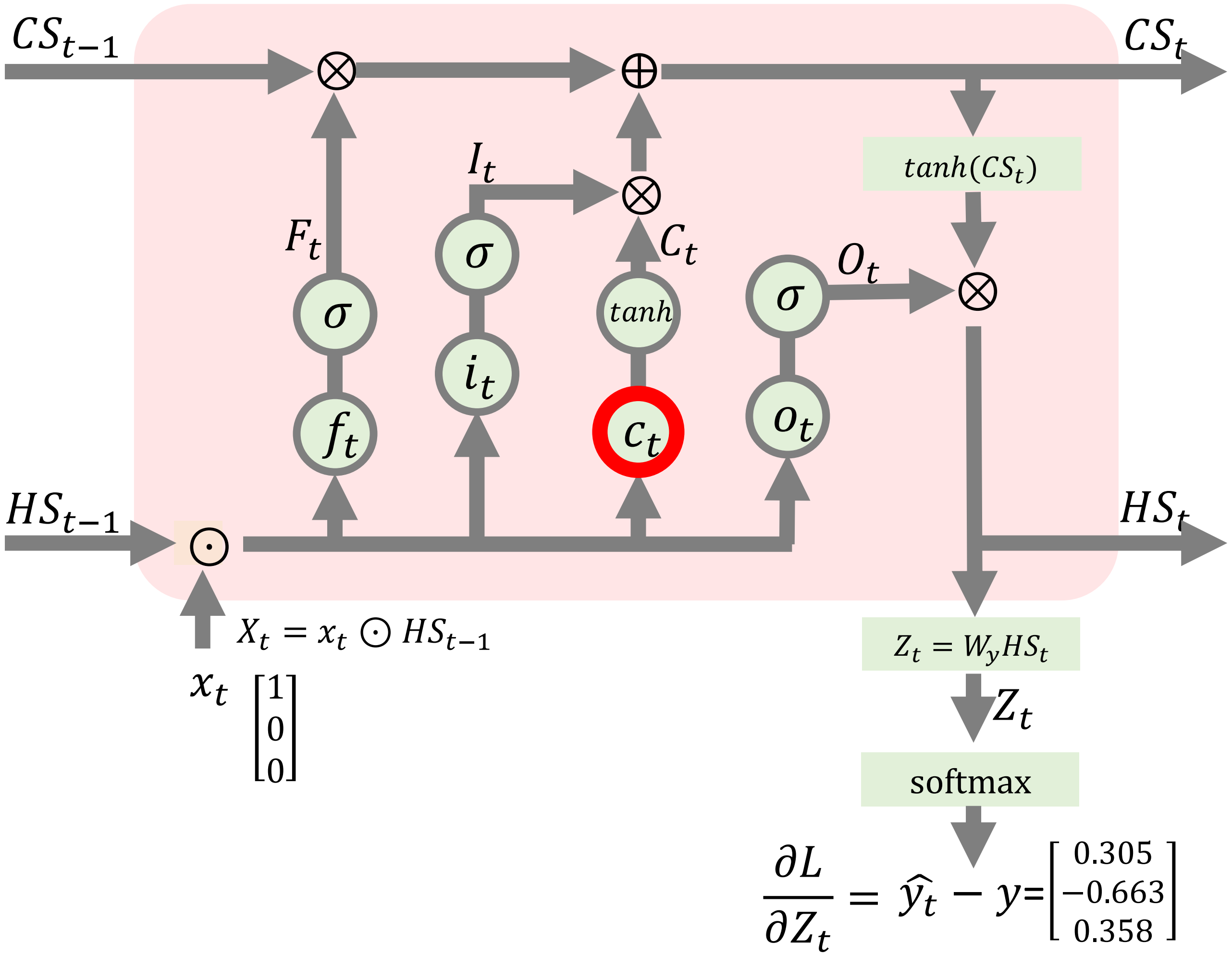
$Z_t = W_y HS_t$

$\boxed{CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t}$

$$\frac{\partial L}{\partial W_c} = \frac{\partial L}{\partial CS_t} \frac{\partial CS_t}{\partial C_t} \frac{\partial C_t}{\partial c_t} \frac{\partial c_t}{\partial W_c}$$

$$\frac{\partial CS_t}{\partial C_t}$$

$$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$$



신박AI

그러면 $\partial CS_t / \partial C_t$ 는 I_t가 됩니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$$

$Z_t = W_y HS_t$

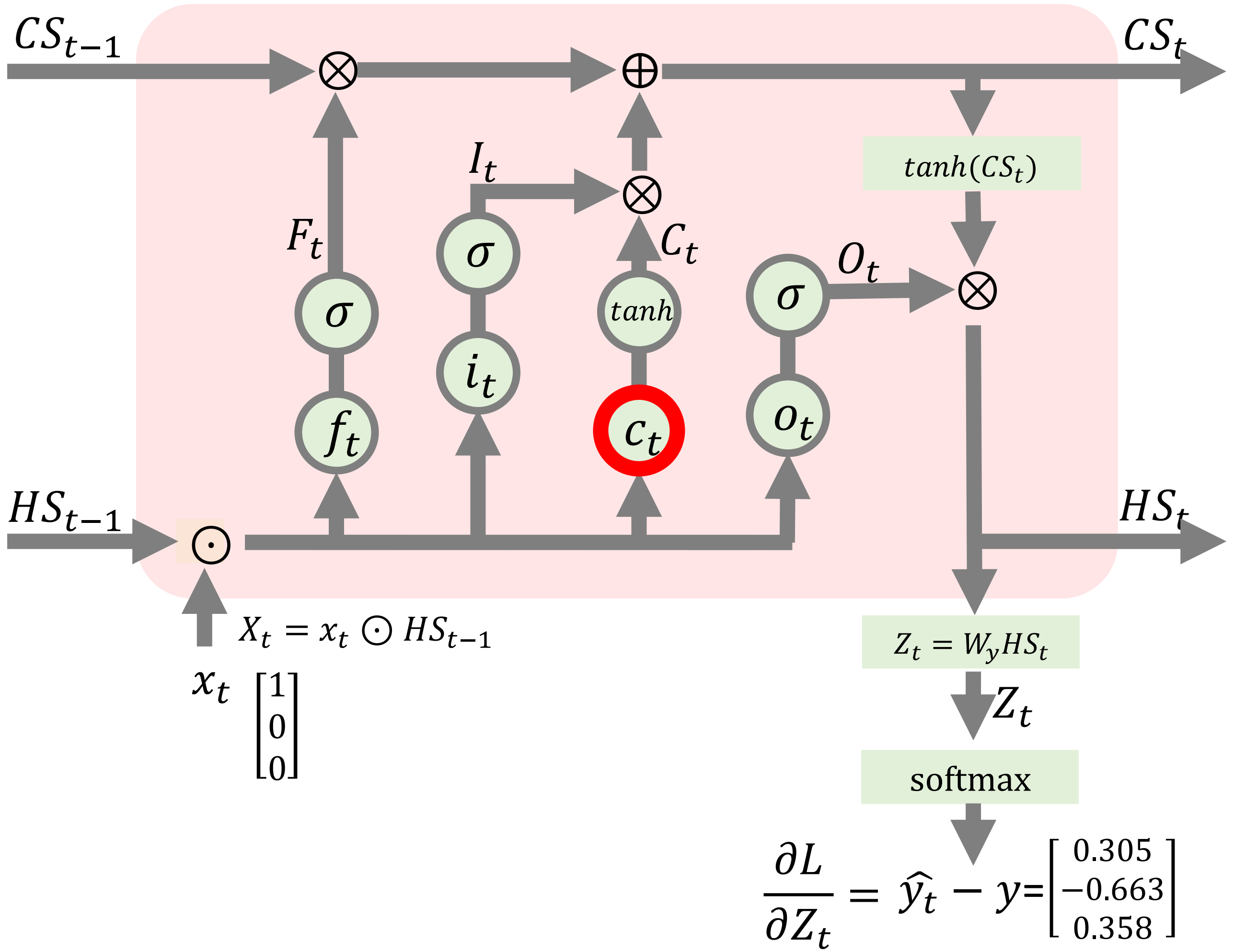
$\boxed{CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t}$

$$\frac{\partial L}{\partial W_c} = \frac{\partial L}{\partial CS_t} \frac{\partial CS_t}{\partial C_t} \frac{\partial C_t}{\partial c_t} \frac{\partial c_t}{\partial W_c}$$

$$\frac{\partial CS_t}{\partial C_t}$$

$$CS_t = \cancel{CS_{t-1}} \otimes \cancel{F_t} + I_t \otimes \cancel{C_t}$$

$$\frac{\partial CS_t}{\partial C_t} = I_t$$



그러면 도출한 식들을 $\partial L/\partial W_C$에 넣고 다시 식을 작성해보겠습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$$\frac{\partial L}{\partial CS_t} = \boxed{(\hat{y}_t - y)W_y O_t(1 - tanh^2(CS_t))}$$

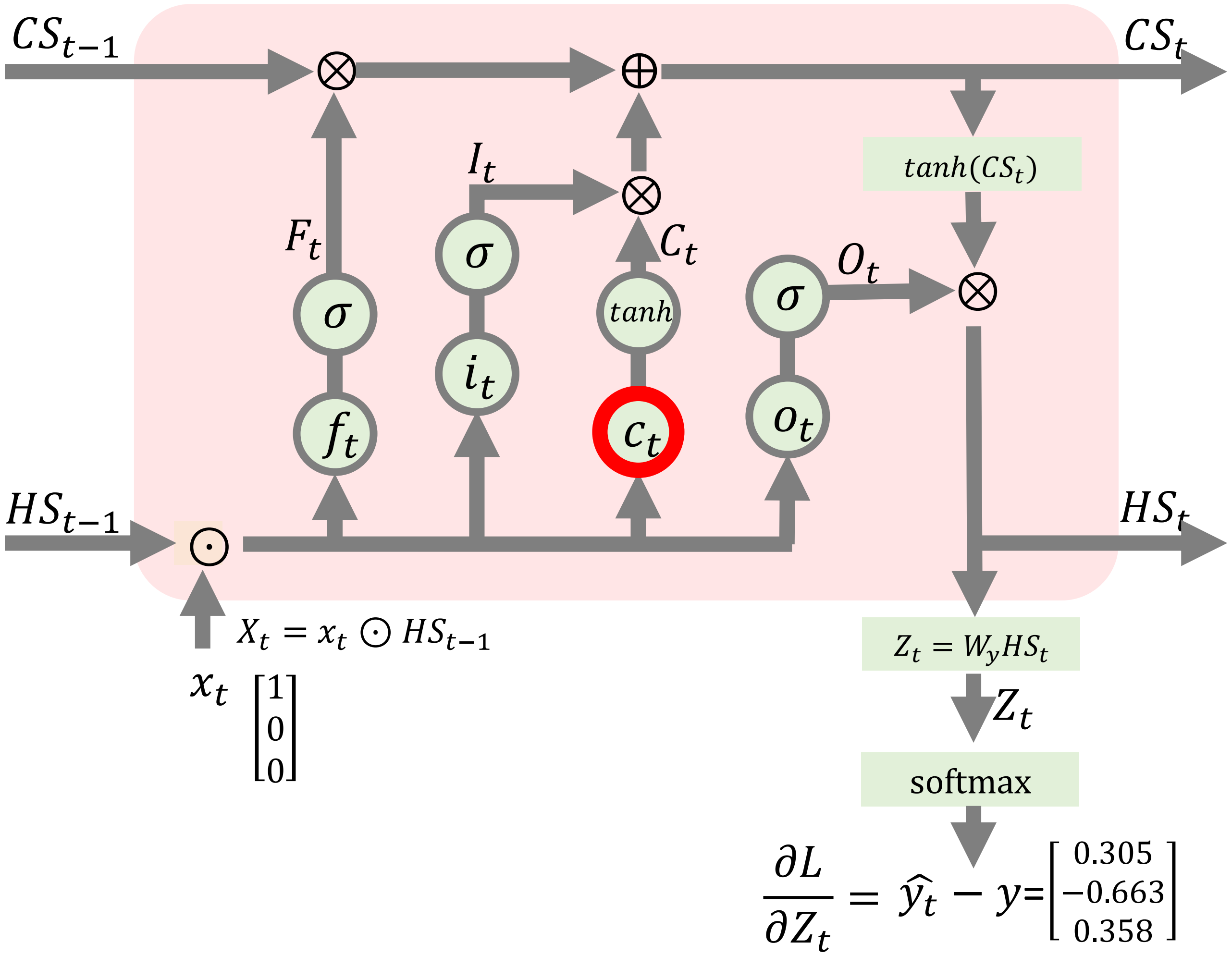
$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial W_c} = \frac{\partial L}{\partial CS_t}\frac{\partial CS_t}{\partial C_t}\frac{\partial C_t}{\partial c_t}\frac{\partial c_t}{\partial W_c}$$

$$\frac{\partial CS_t}{\partial C_t}$$

$$\frac{\partial CS_t}{\partial C_t} = \boxed{I_t}$$



신박AI

그러면 $\partial L/\partial W_C$은 다음처럼 정리가 됩니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

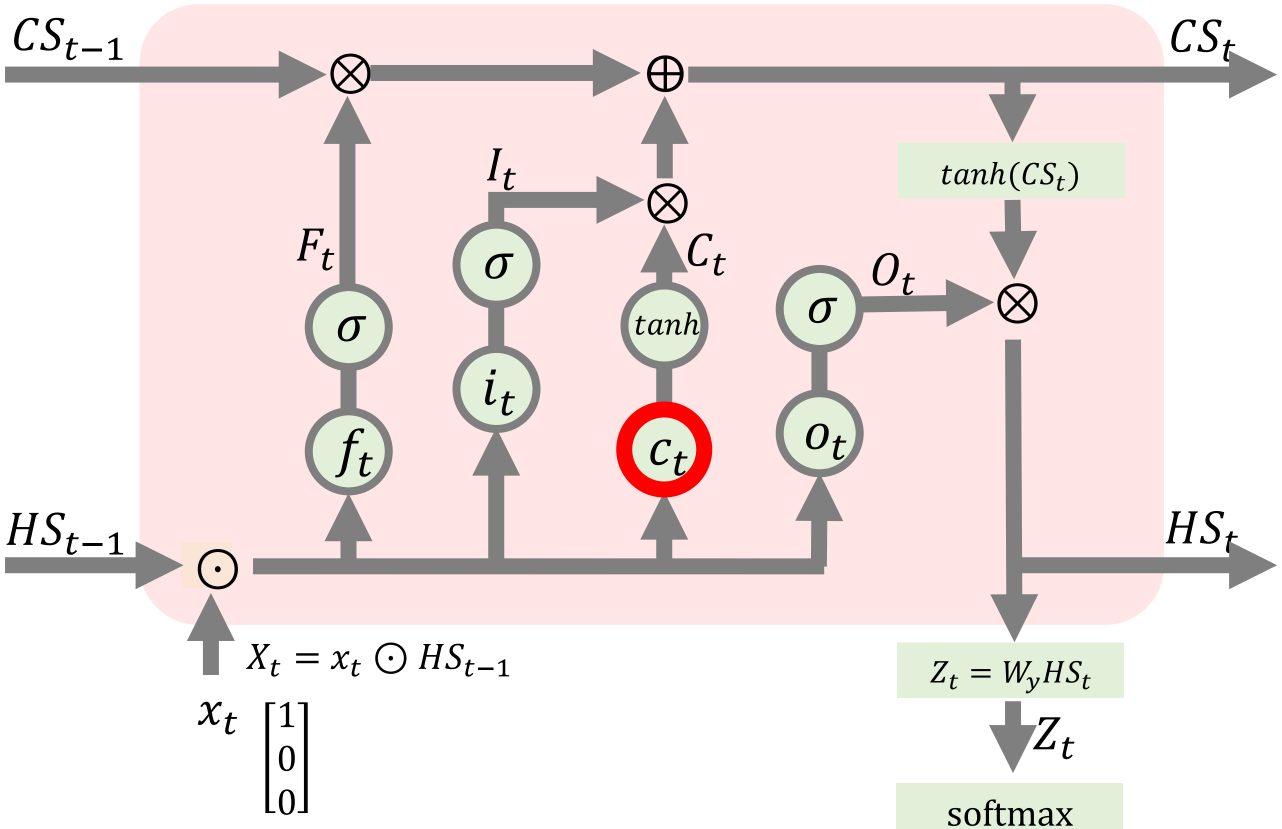
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial W_c} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t)) I_t \frac{\partial C_t}{\partial c_t} \frac{\partial c_t}{\partial W_c}$$



$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix}0.305\\-0.663\\0.358\end{bmatrix}$

신박AI

그 다음은 $\partial C_t/\partial c_t$를 구해보겠습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

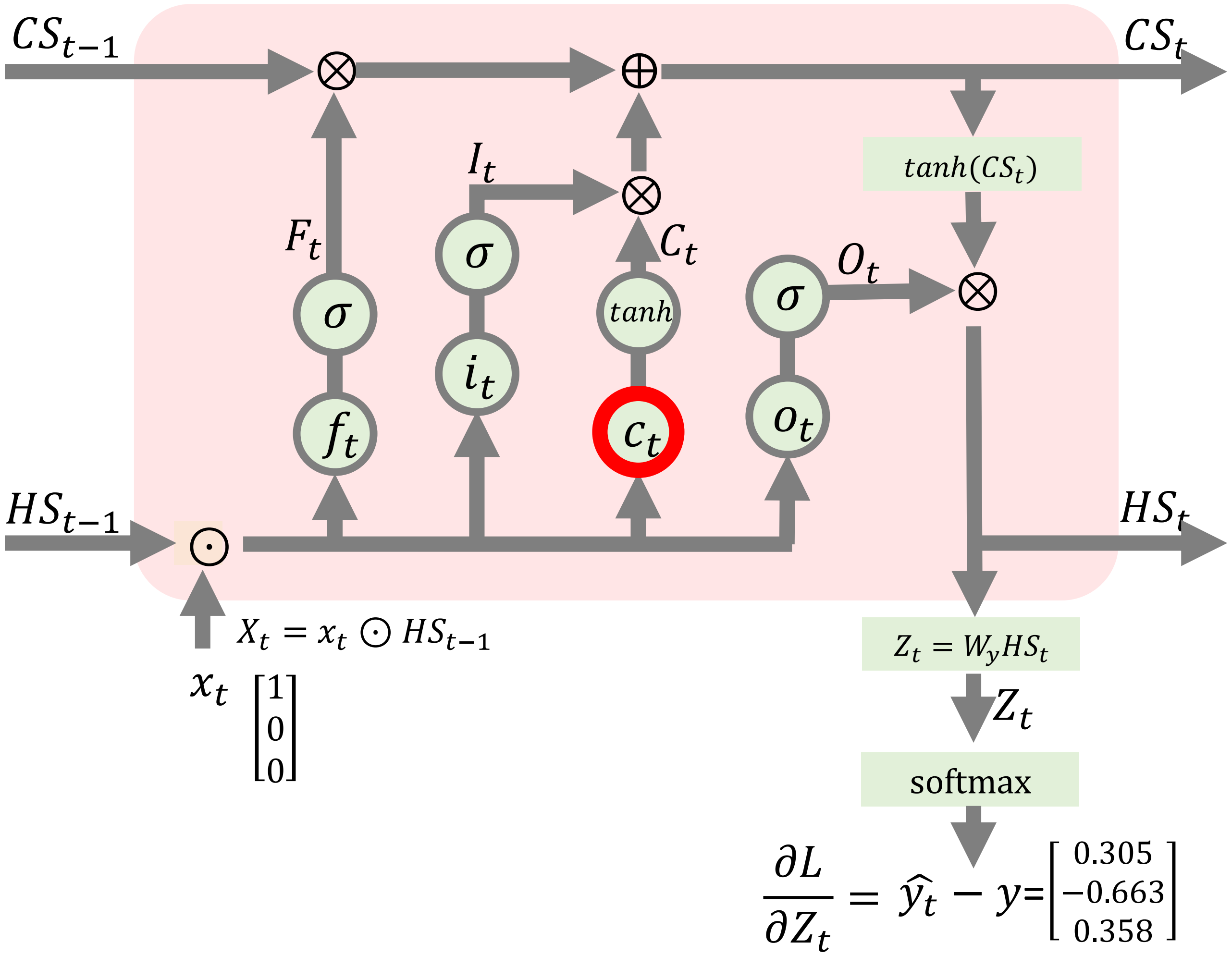
$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y)W_y O_t(1 - tanh^2(CS_t))$$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial W_c} = (\hat{y}_t - y)W_y O_t(1 - tanh^2(CS_t))I_t \boxed{\frac{\partial C_t}{\partial c_t}} \frac{\partial c_t}{\partial W_c}$$

$$\frac{\partial C_t}{\partial c_t}$$



신박AI

$\partial C_t / \partial c_t$은 $tanh$ 미분 함수에 의해서

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$$

Candidate Gate: $C_t = tanh(W_c X_t)$
$\boxed{C_t = tanh(c_t)}$
$c_t = W_c X_t$

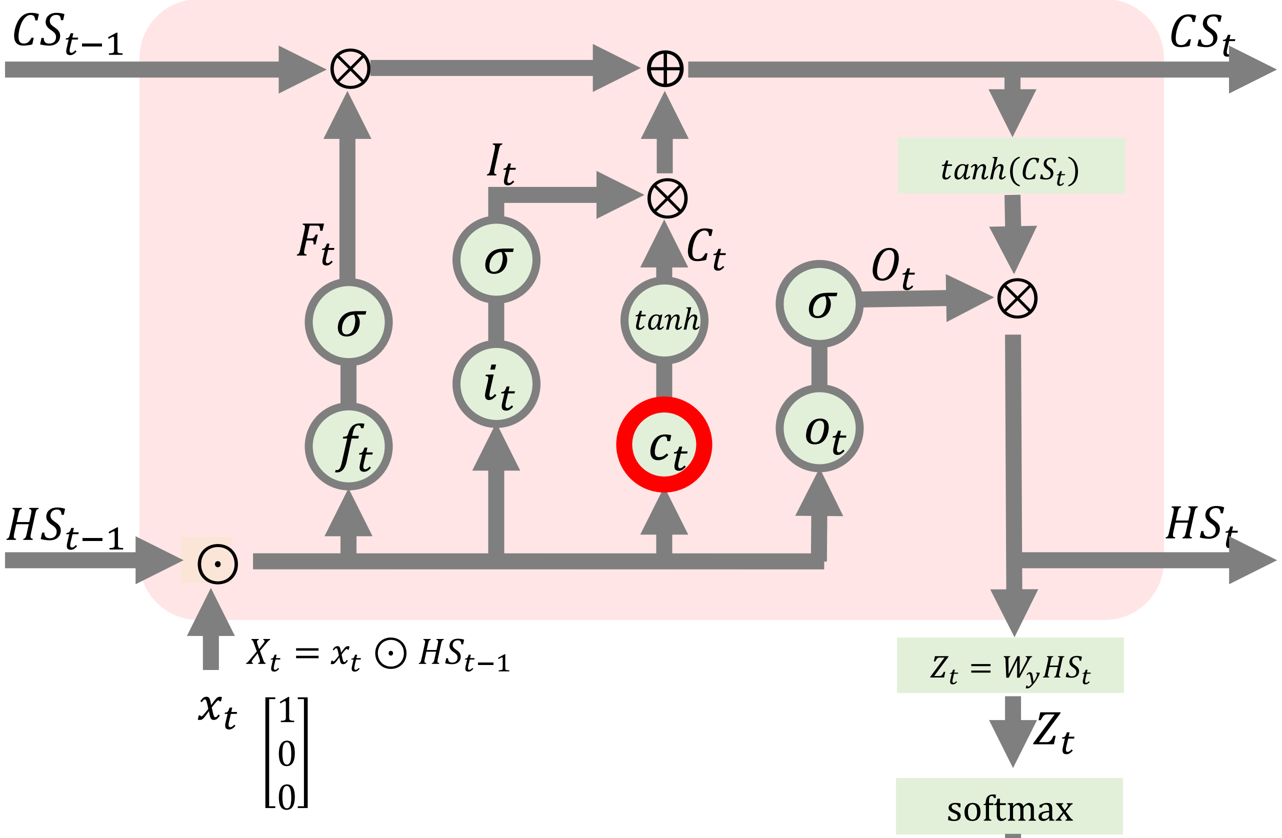
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial W_c} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t)) I_t \boxed{\frac{\partial C_t}{\partial c_t}} \frac{\partial c_t}{\partial W_c}$$

$$\frac{\partial C_t}{\partial c_t}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix} 0.305 \\ -0.663 \\ 0.358 \end{bmatrix}$$

이렇게 구할 수가 있고,

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$\boxed{C_t = tanh(c_t)}$
$c_t = W_c X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

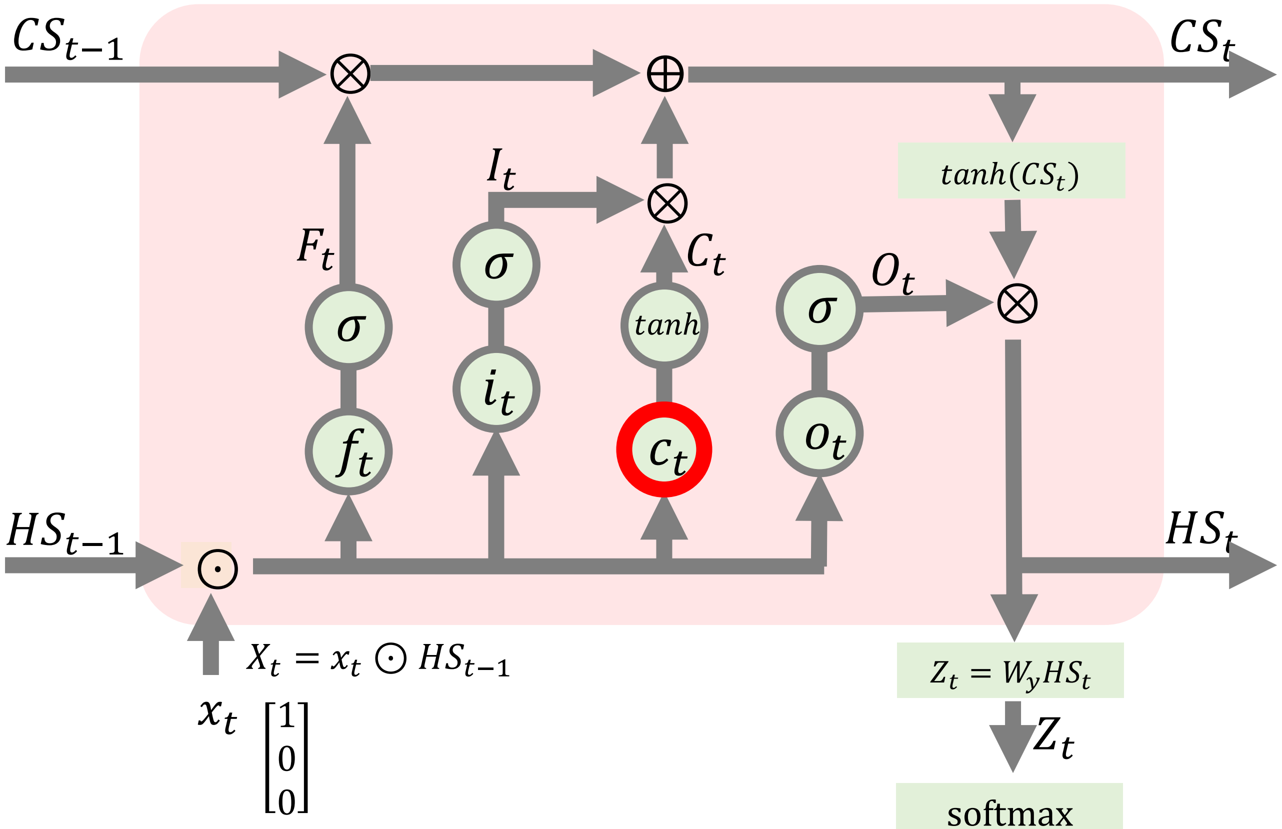
$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y)W_y O_t(1 - tanh^2(CS_t))$$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial W_c} = (\hat{y}_t - y)W_y O_t(1 - tanh^2(CS_t))I_t \boxed{\frac{\partial C_t}{\partial c_t}} \frac{\partial c_t}{\partial W_c}$$

$$\boxed{\frac{\partial C_t}{\partial c_t} = (1 - tanh^2(C_t))}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix}0.305\\-0.663\\0.358\end{bmatrix}$$

신박AI

이어서 $\partial c_t / \partial W_c$는 이 공식에 의해서

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$\boxed{c_t = W_c X_t}$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$$

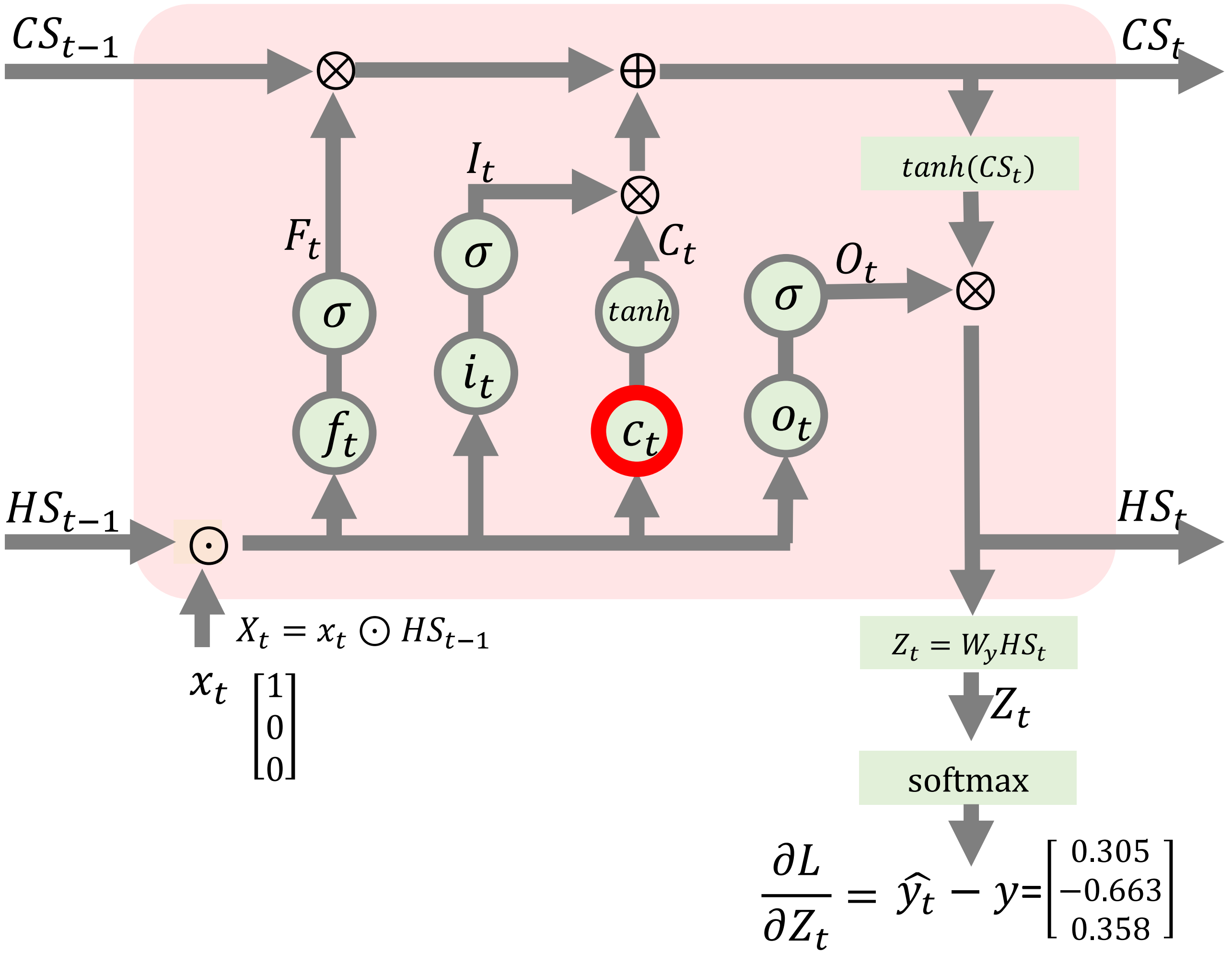
$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial W_c} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t)) I_t \frac{\partial C_t}{\partial c_t} \boxed{\frac{\partial c_t}{\partial W_c}}$$

$$\frac{\partial C_t}{\partial c_t} = (1 - tanh^2(C_t))$$

$$\frac{\partial c_t}{\partial W_c}$$



신박AI

X_t로 구할 수 있습니다

Forget Gate: $F_t = \sigma(W_f X_t)$

$F_t = \sigma(f_t)$

$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$

$I_t = \sigma(i_t)$

$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$

$C_t = tanh(c_t)$

$\boxed{c_t = W_c X_t}$

Output Gate: $O_t = \sigma(W_o X_t)$

$O_t = \sigma(o_t)$

$o_t = W_o X_t$

$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$$

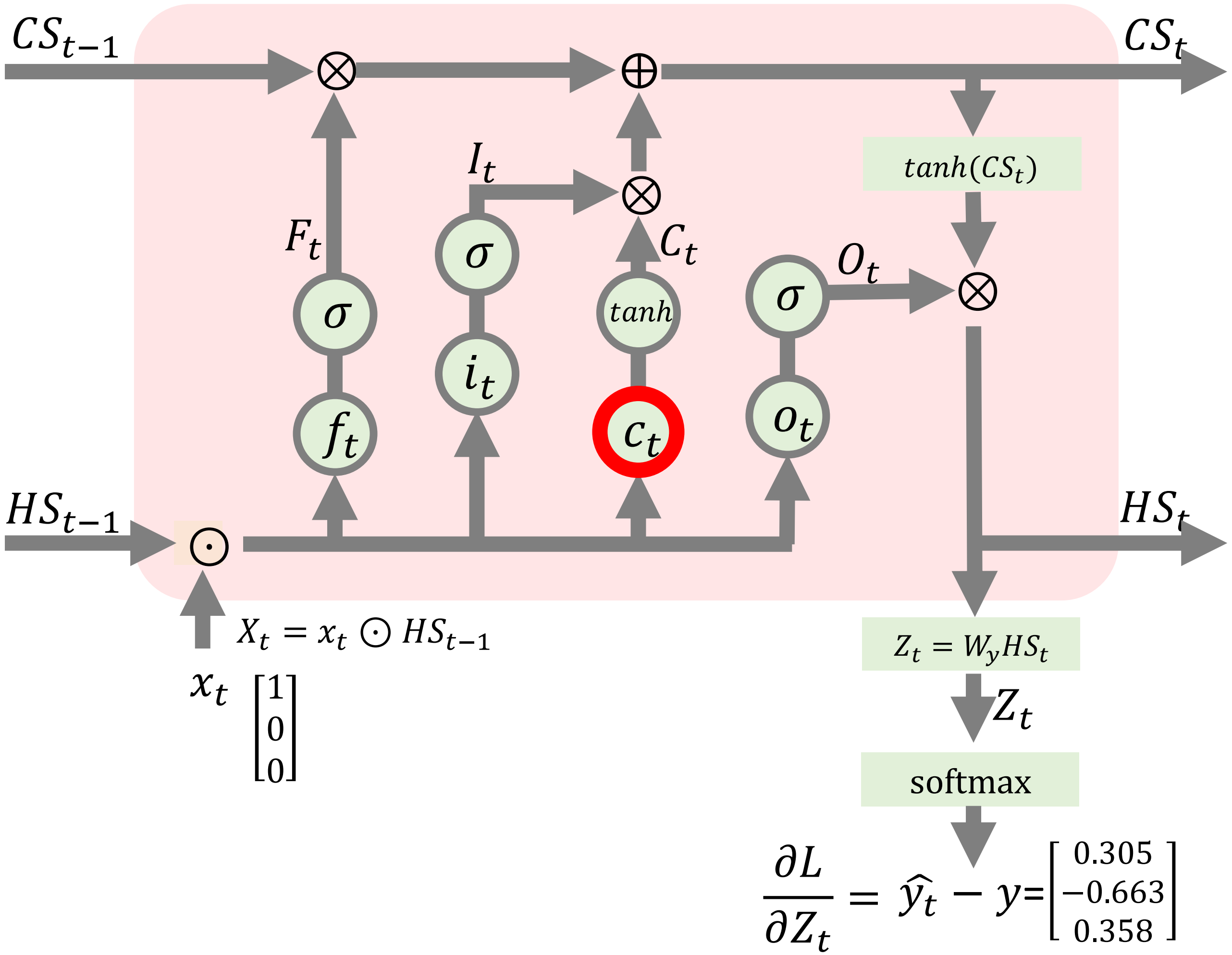
$$Z_t = W_y HS_t$$

$$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$$

$$\frac{\partial L}{\partial W_c} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t)) I_t \frac{\partial C_t}{\partial c_t} \boxed{\frac{\partial c_t}{\partial W_c}}$$

$$\frac{\partial C_t}{\partial c_t} = (1 - tanh^2(C_t))$$

$$\boxed{\frac{\partial c_t}{\partial W_c} = X_t}$$



그러면 이 두 식을 $\partial L/\partial W_c$식에 넣으면,

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$$

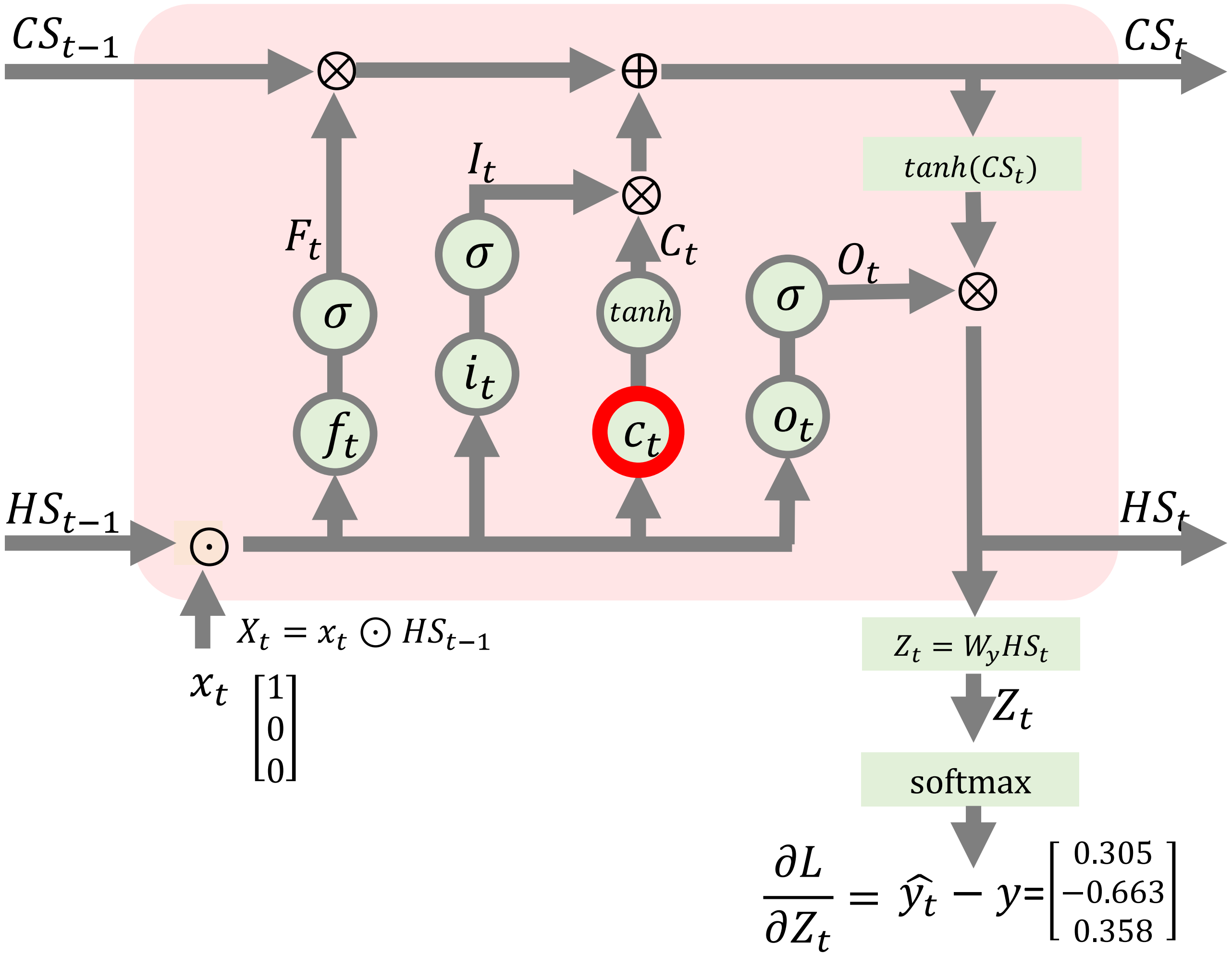
$$Z_t = W_y HS_t$$

$$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$$

$$\frac{\partial L}{\partial W_c} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t)) I_t \frac{\partial C_t}{\partial c_t} \frac{\partial c_t}{\partial W_c}$$

$$\frac{\partial C_t}{\partial c_t} = (1 - tanh^2(C_t))$$

$$\frac{\partial c_t}{\partial W_c} = X_t$$



$\partial L/\partial W_c$식이 완성 되었습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$

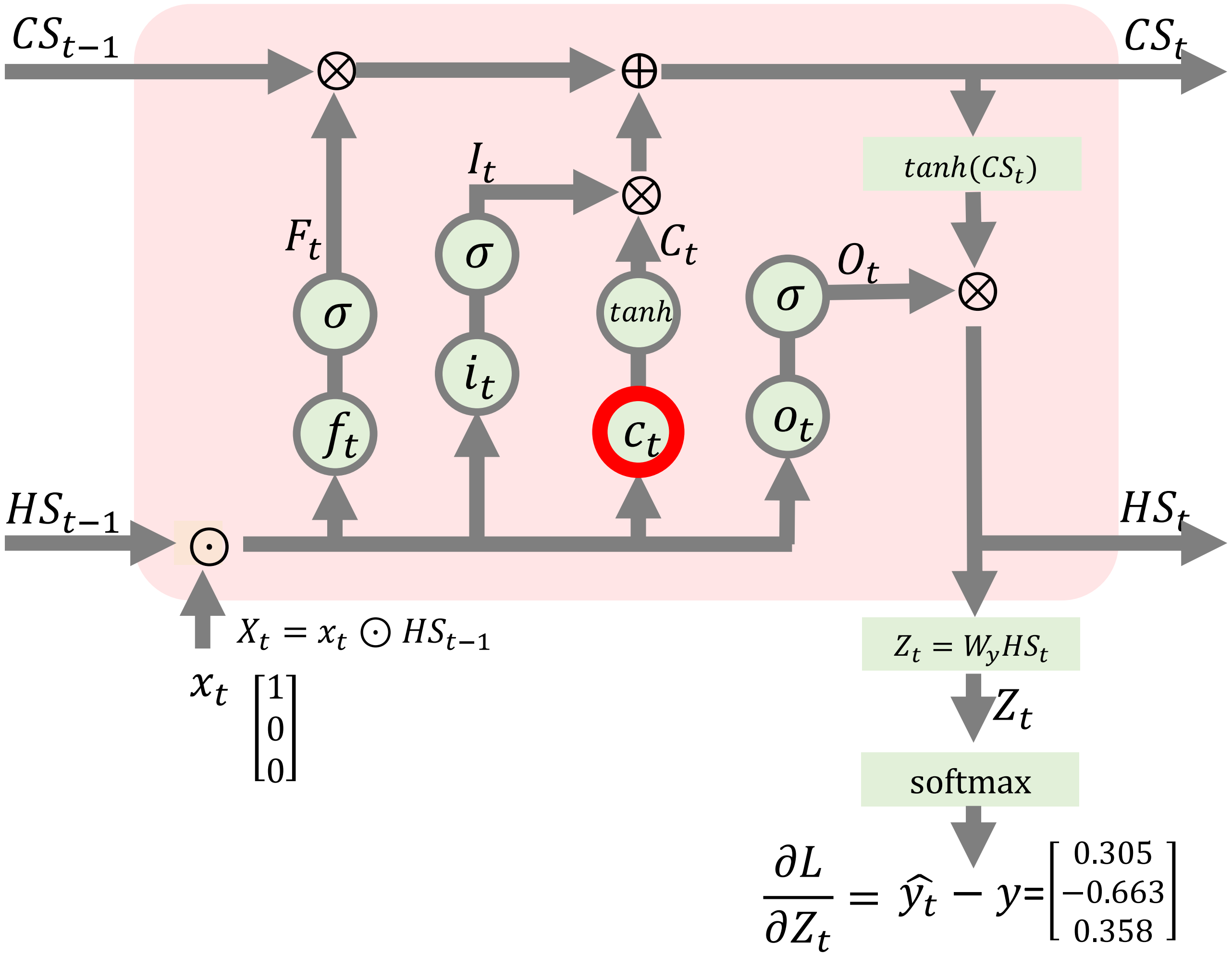
Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial W_c} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t)) I_t (1 - tanh^2(C_t)) X_t$$



자 이제 숫자를 넣어보도록 하겠습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

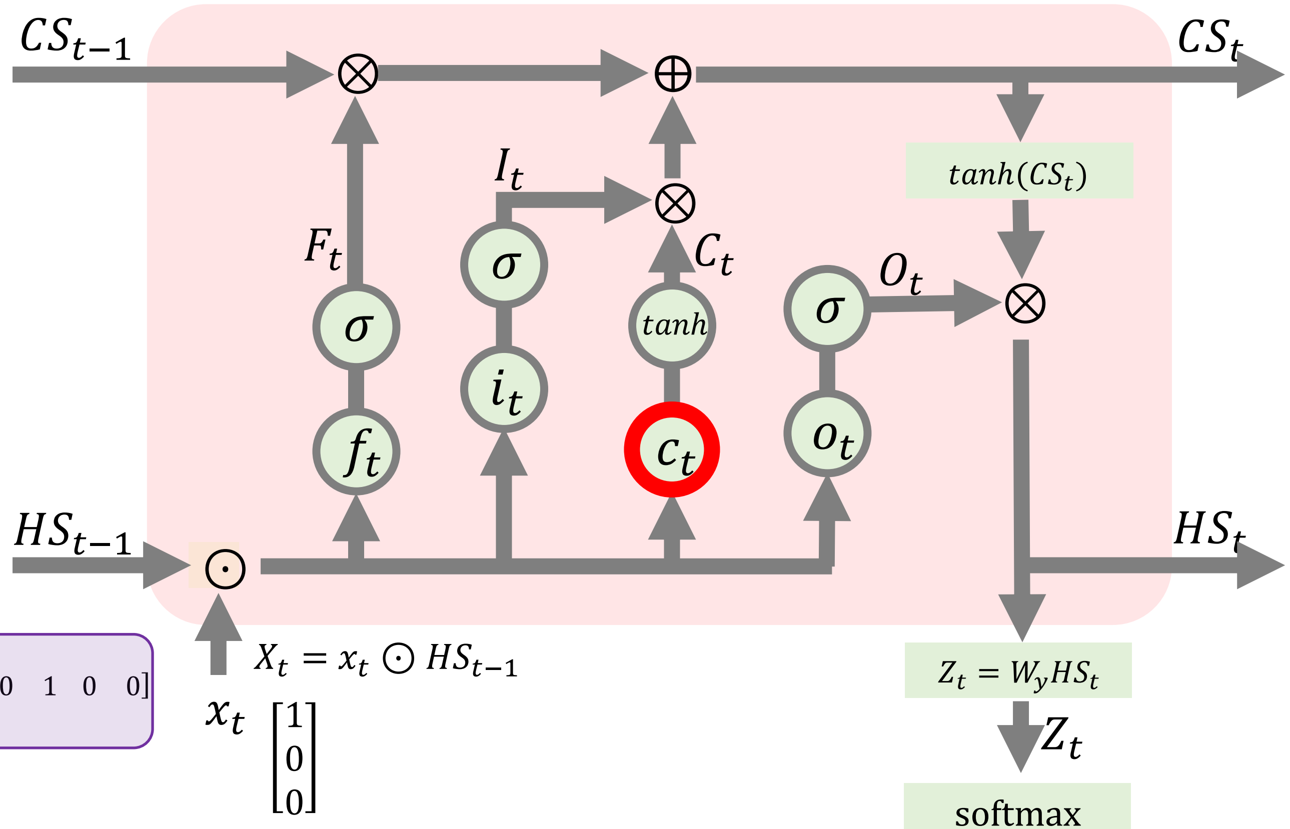
$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial W_c} = \boxed{(\hat{y}_t - y) W_y}\boxed{O_t}\boxed{(1 - tanh^2(CS_t))}\boxed{I_t}\boxed{(1 - tanh^2(C_t))}\boxed{X_t}$$

$$= \boxed{\left(\begin{bmatrix} 0.305 & -0.663 & 0.358 \end{bmatrix} \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix} \right)^T} \boxed{\begin{bmatrix} 0.401 \\ 0.634 \end{bmatrix}} \boxed{\begin{bmatrix} 0.912 \\ 0.936 \end{bmatrix}} \boxed{\begin{bmatrix} 0.508 \\ 0.4 \end{bmatrix}} \boxed{\begin{bmatrix} 0.847 \\ 0.582 \end{bmatrix}} \boxed{\begin{bmatrix} 0 & 0 & 1 & 0 & 0 \end{bmatrix}}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix} 0.305 \\ -0.663 \\ 0.358 \end{bmatrix}$$

신박AI

이렇게 $\partial L/\partial W_c$을 계산해보았습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y)W_y O_t(1 - tanh^2(CS_t))$$

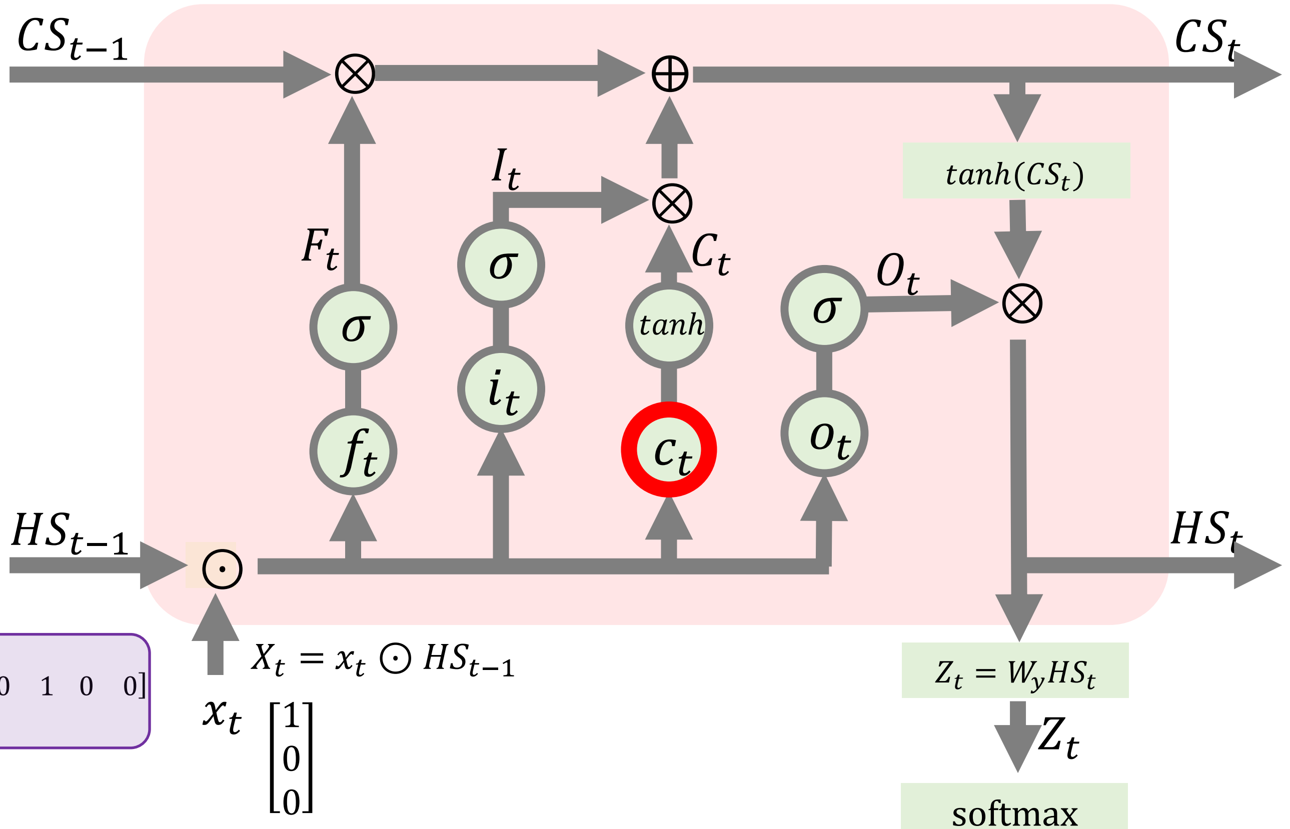
$$Z_t = W_y HS_t$$

$$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$$

$$\frac{\partial L}{\partial W_c} = (\hat{y}_t - y)W_y\, O_t\, (1 - tanh^2(CS_t))\, I_t\, (1 - tanh^2(C_t))\, X_t$$

$$= \left(\begin{bmatrix} 0.305 & -0.663 & 0.358 \end{bmatrix} \begin{bmatrix} 0.32 & -0.172 \\ 0.449 & 0.349 \\ 0.914 & 0.371 \end{bmatrix}\right)^T \begin{bmatrix} 0.401 \\ 0.634 \end{bmatrix} \begin{bmatrix} 0.912 \\ 0.936 \end{bmatrix} \begin{bmatrix} 0.508 \\ 0.4 \end{bmatrix} \begin{bmatrix} 0.847 \\ 0.582 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & 0 & 0.02 & 0 & 0 \\ 0 & 0 & -0.021 & 0 & 0 \end{bmatrix}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix} 0.305 \\ -0.663 \\ 0.358 \end{bmatrix}$$

신박AI

여기까지 LSTM의 가중치의 변화량을 다 구해보았습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

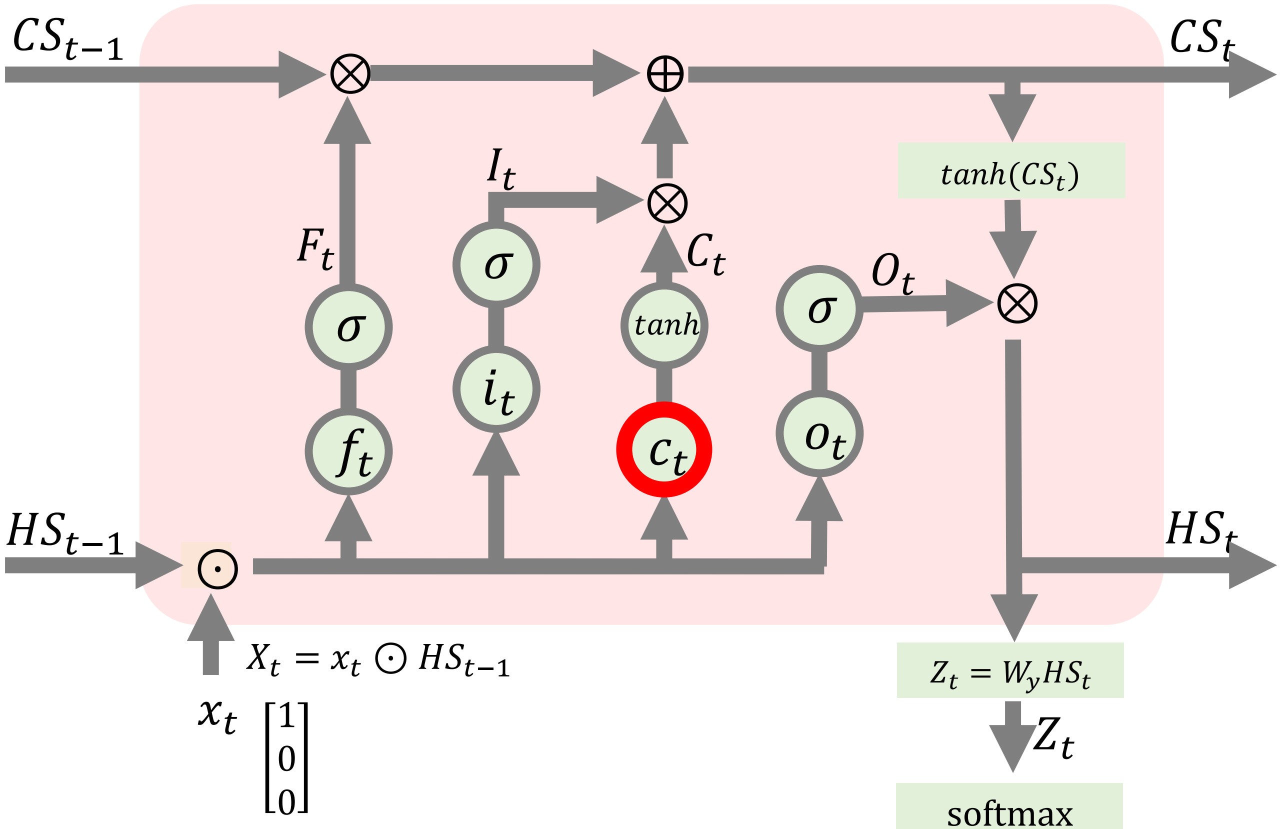
$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial W_f} = \begin{bmatrix} 0 & 0 & 0.012 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \qquad \frac{\partial L}{\partial W_i} = \begin{bmatrix} 0 & 0 & -0.005 & 0 & 0 \\ 0 & 0 & -0.014 & 0 & 0 \end{bmatrix}$$

$$\frac{\partial L}{\partial W_o} = \begin{bmatrix} 0 & 0 & 0.009 & 0 & 0 \\ 0 & 0 & -0.009 & 0 & 0 \end{bmatrix} \qquad \frac{\partial L}{\partial W_c} = \begin{bmatrix} 0 & 0 & 0.02 & 0 & 0 \\ 0 & 0 & -0.021 & 0 & 0 \end{bmatrix}$$

$$\frac{\partial L}{\partial W_y} = \begin{bmatrix} 0.036 & 0.049 \\ -0.079 & -0.106 \\ 0.043 & 0.057 \end{bmatrix}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix} 0.305 \\ -0.663 \\ 0.358 \end{bmatrix}$$

신박AI

그러면 경사하강법을 통해서 가중치를 업데이트 할 수가 있을 것입니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

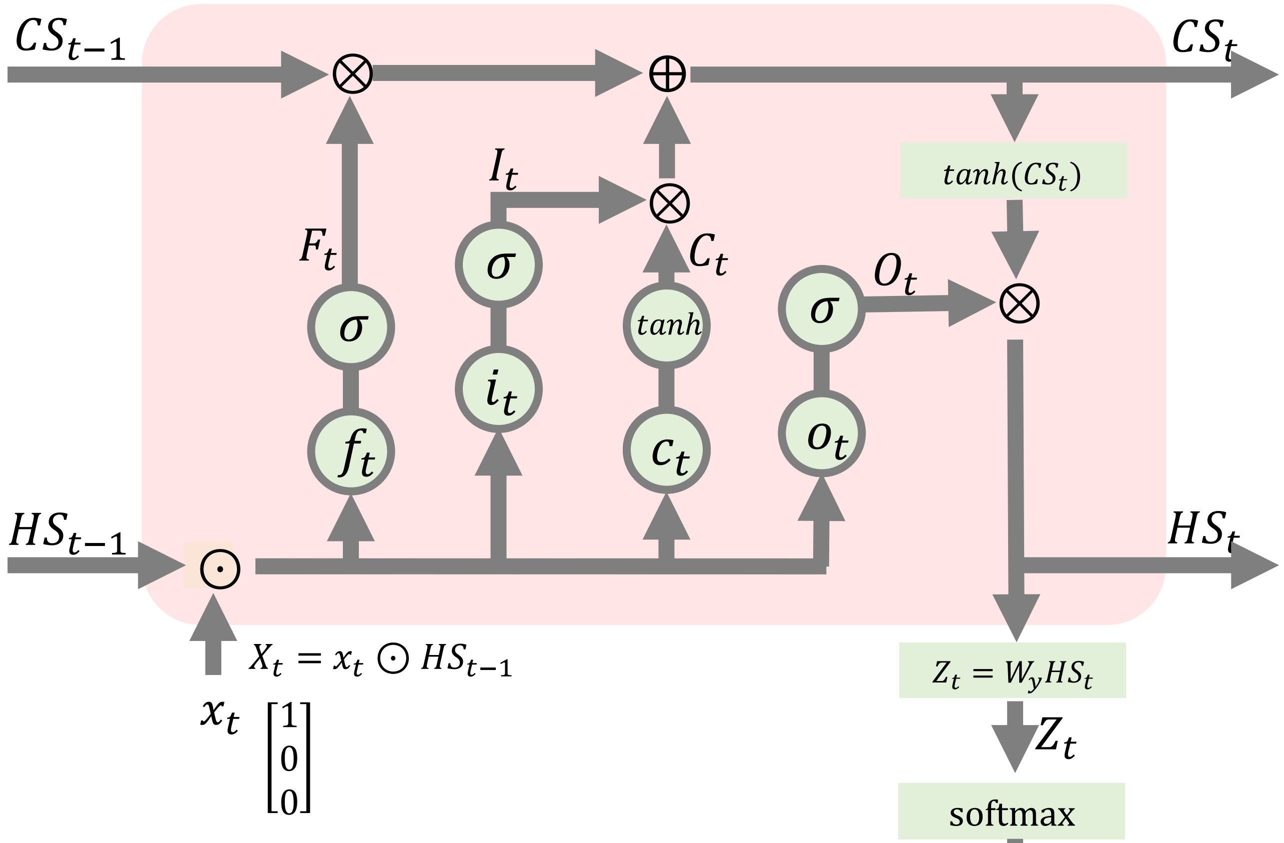
$$\frac{\partial L}{\partial W_f} = \begin{bmatrix} 0 & 0 & 0.012 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\frac{\partial L}{\partial W_i} = \begin{bmatrix} 0 & 0 & -0.005 & 0 & 0 \\ 0 & 0 & -0.014 & 0 & 0 \end{bmatrix}$$

$$\frac{\partial L}{\partial W_o} = \begin{bmatrix} 0 & 0 & 0.009 & 0 & 0 \\ 0 & 0 & -0.009 & 0 & 0 \end{bmatrix}$$

$$\frac{\partial L}{\partial W_c} = \begin{bmatrix} 0 & 0 & 0.02 & 0 & 0 \\ 0 & 0 & -0.021 & 0 & 0 \end{bmatrix}$$

$$\frac{\partial L}{\partial W_y} = \begin{bmatrix} 0.036 & 0.049 \\ -0.079 & -0.106 \\ 0.043 & 0.057 \end{bmatrix}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix} 0.305 \\ -0.663 \\ 0.358 \end{bmatrix}$$

$$W^* = W + \alpha\left(-\frac{\partial L}{\partial W}\right)$$

그리고 아직 끝나지 않았습니다

Forget Gate: $F_t = \sigma(W_f X_t)$

$F_t = \sigma(f_t)$

$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$

$I_t = \sigma(i_t)$

$i_t = W_i X_t$

$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$$

Candidate Gate: $C_t = tanh(W_c X_t)$

$C_t = tanh(c_t)$

$c_t = W_c X_t$

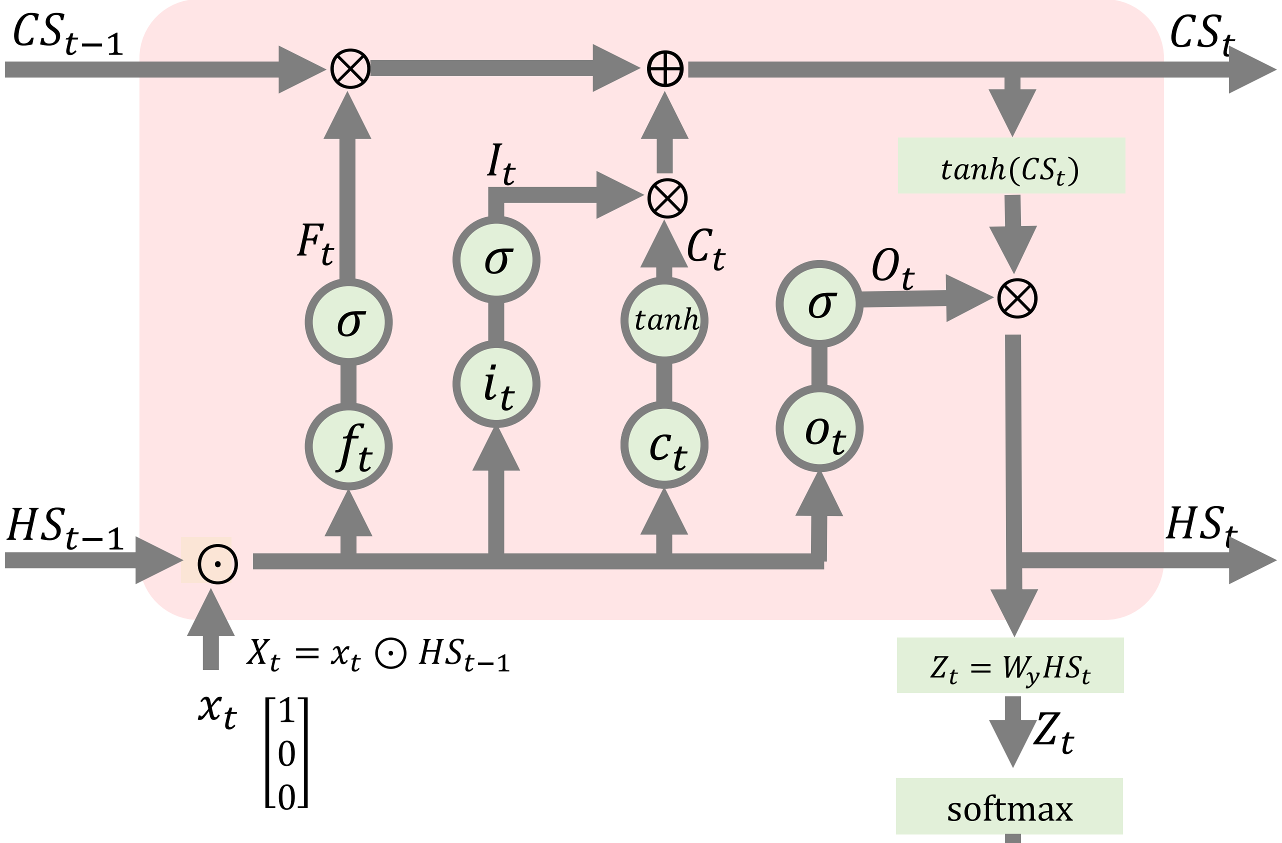
Output Gate: $O_t = \sigma(W_o X_t)$

$O_t = \sigma(o_t)$

$o_t = W_o X_t$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$



$x_t \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$

$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix} 0.305 \\ -0.663 \\ 0.358 \end{bmatrix}$$

신박AI

현 시점 t에서 발생한 Loss가 이전 시간t-1에도 전달 되는 것을 알아보겠습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

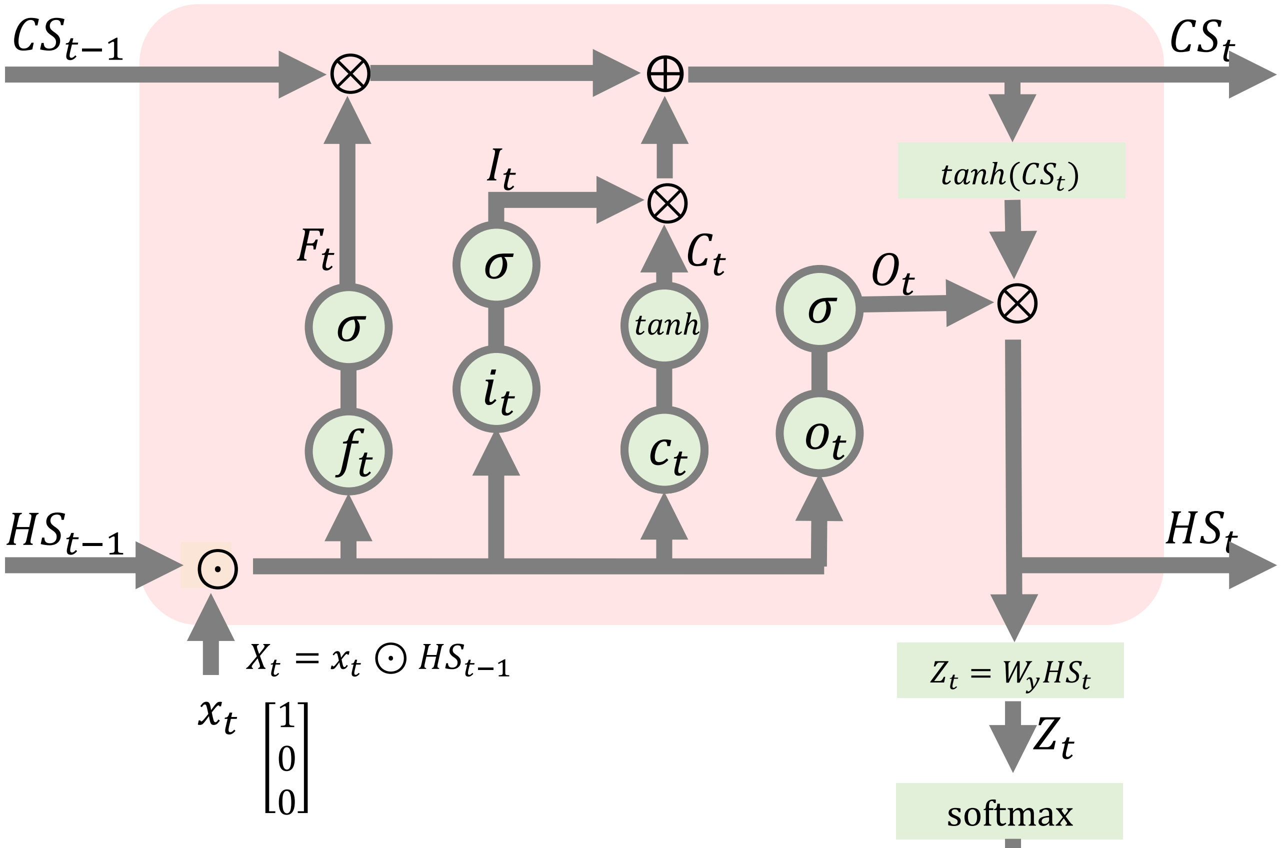
$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$



$$x_t \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix} 0.305 \\ -0.663 \\ 0.358 \end{bmatrix}$$

현재의 Loss가 이전 셀상태에 준 영향을 표현하기 위해서

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

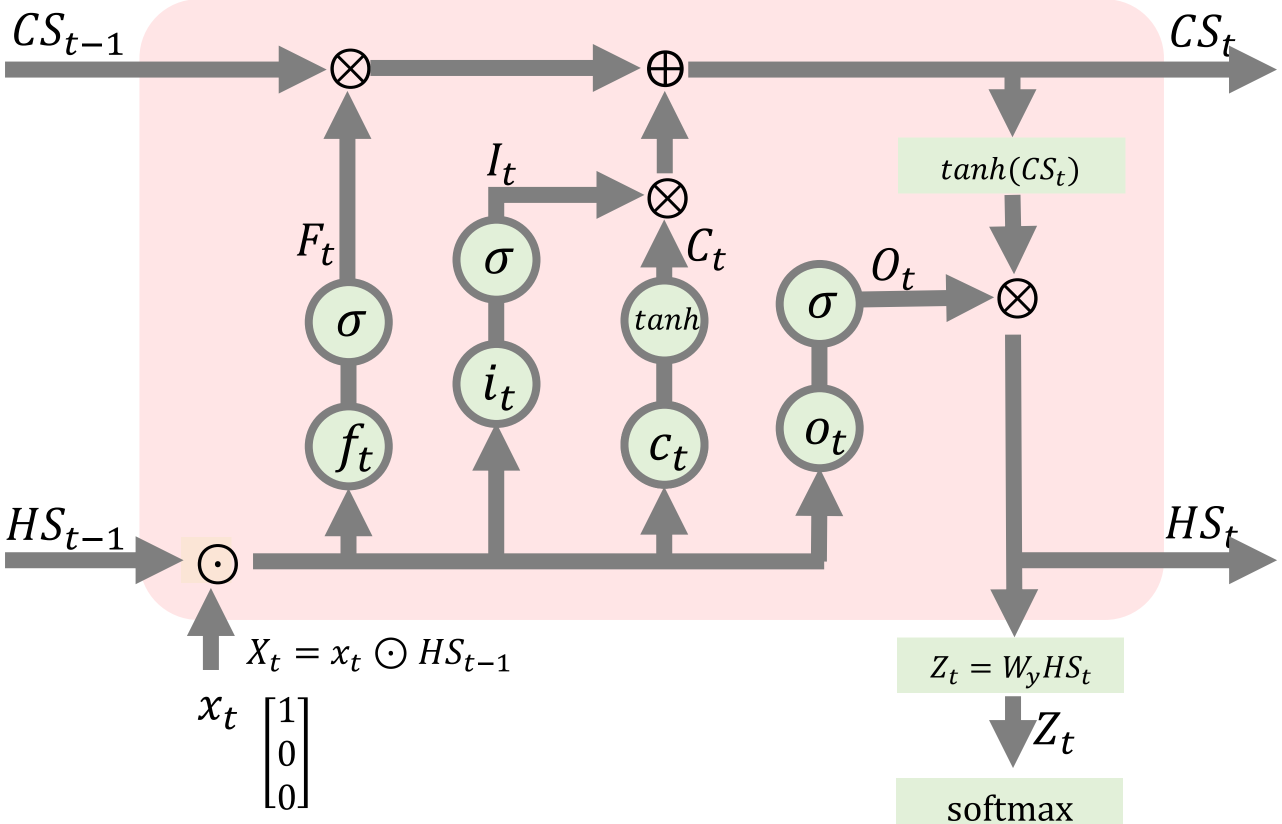
$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix} 0.305 \\ -0.663 \\ 0.358 \end{bmatrix}$$

신박AI

$\partial L/\partial CS_{t-1}$을 구해보도록 하겠습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$

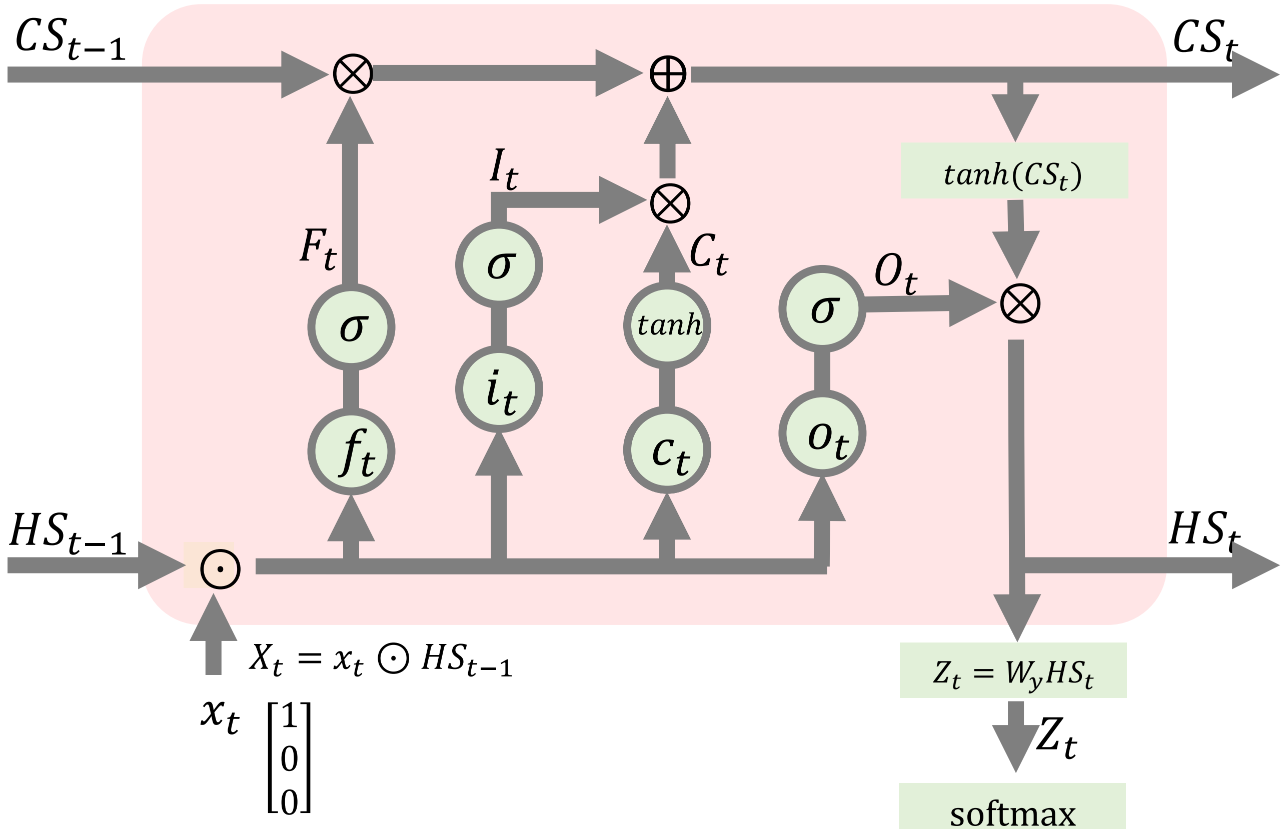
Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial CS_{t-1}} =$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix}0.305\\-0.663\\0.358\end{bmatrix}$$

신박AI

$\partial L/\partial CS_{t-1}$을 이렇게 전개해 볼 수 있고

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

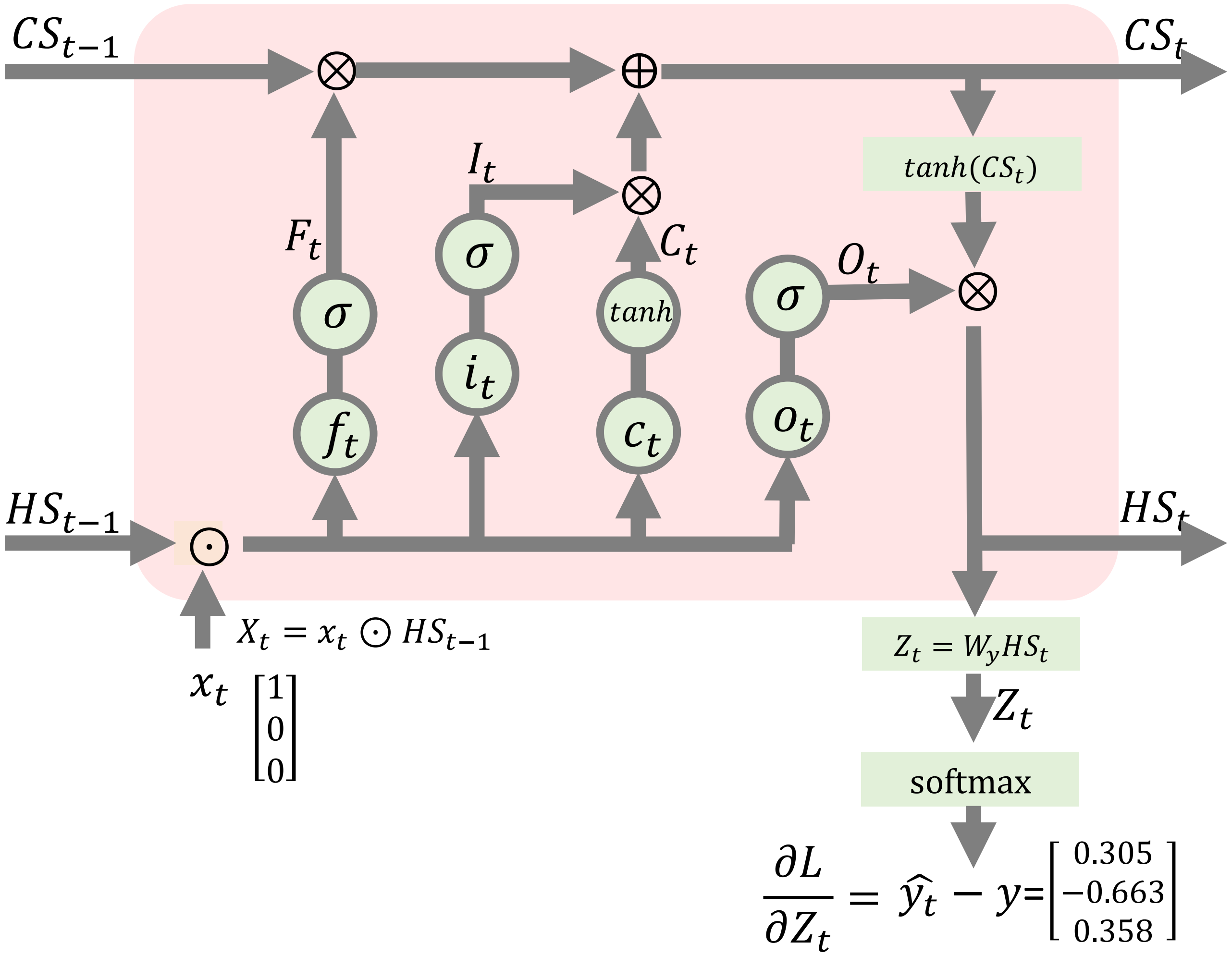
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial CS_{t-1}} = \frac{\partial L}{\partial CS_t} \frac{\partial CS_t}{\partial CS_{t-1}}$$



신박AI

$\partial L / \partial CS_t$은 이미 우리가 전개해본 바가 있습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

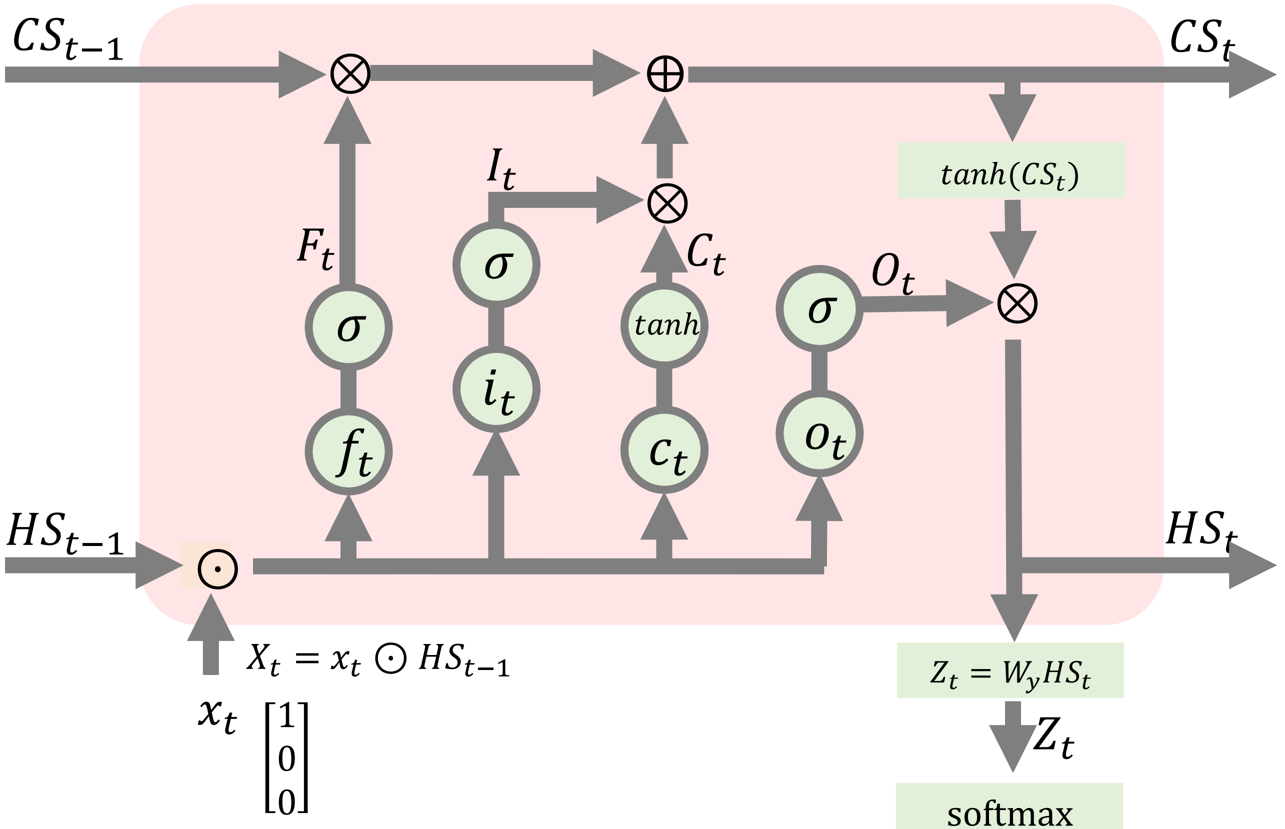
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\boxed{\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))}$$

$$\frac{\partial L}{\partial CS_{t-1}} = \frac{\partial L}{\partial CS_t} \frac{\partial CS_t}{\partial CS_{t-1}}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix}0.305\\-0.663\\0.358\end{bmatrix}$$

신박AI

그리고 $\partial CS_t/\partial CS_{t-1}$은 이 식으로 구해볼 수 있습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

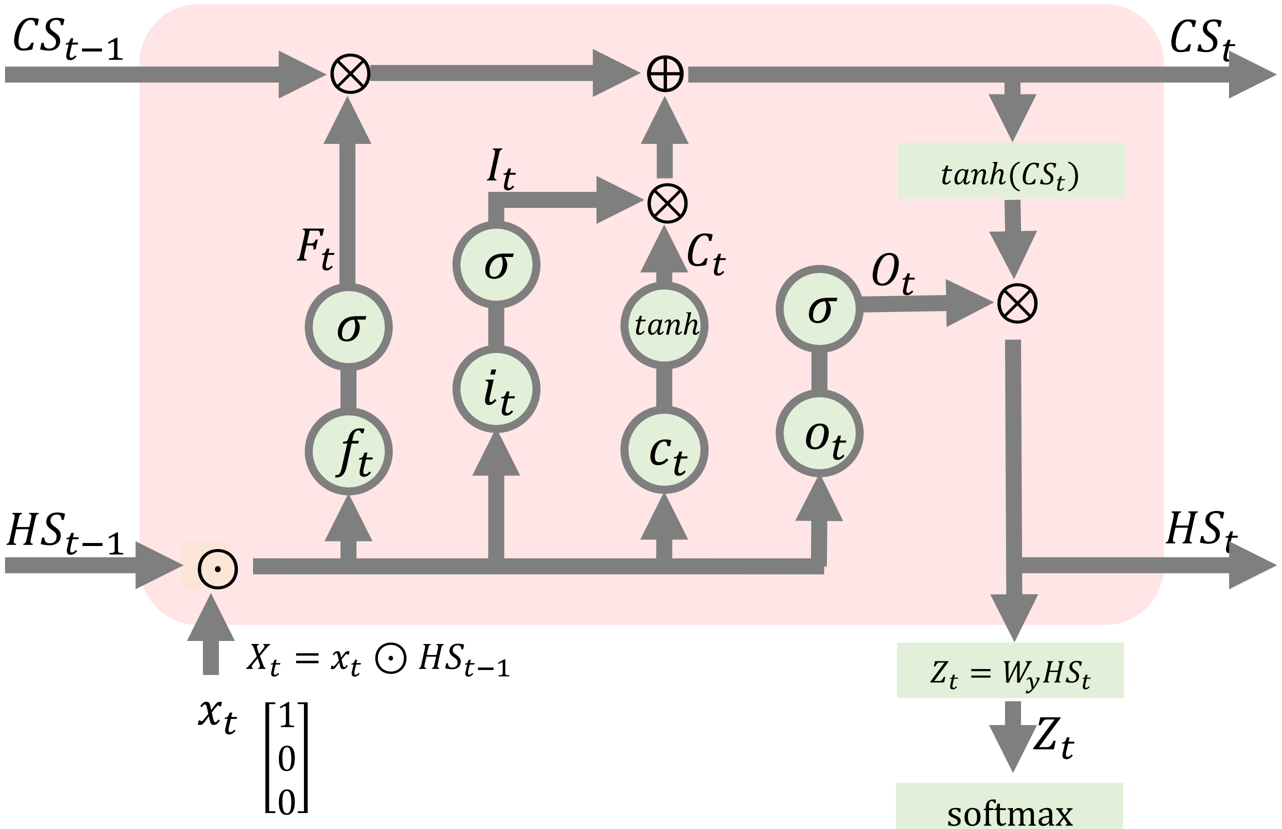
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$$

$Z_t = W_y HS_t$

$$\boxed{CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t}$$

$$\frac{\partial L}{\partial CS_{t-1}} = \frac{\partial L}{\partial CS_t} \frac{\partial CS_t}{\partial CS_{t-1}}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix}0.305\\-0.663\\0.358\end{bmatrix}$$

신박AI

그러면 $\partial CS_t/\partial CS_{t-1}$은 F_t가 됨을 볼수가 있습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

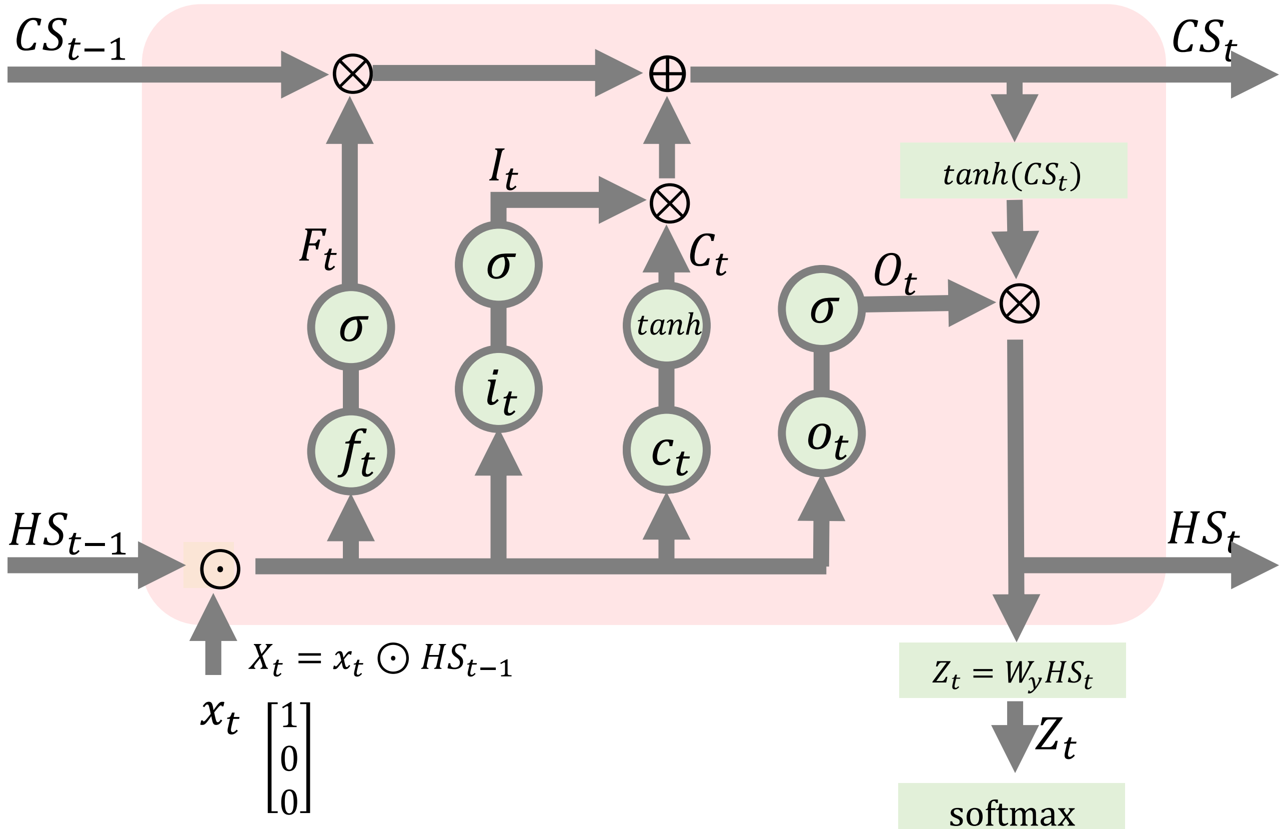
$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$$

$Z_t = W_y HS_t$

$$\boxed{CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t}$$

$$\frac{\partial L}{\partial CS_{t-1}} = \frac{\partial L}{\partial CS_t} \frac{\partial CS_t}{\partial CS_{t-1}}$$

$$\frac{\partial CS_t}{\partial CS_{t-1}} = F_t$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix}0.305\\-0.663\\0.358\end{bmatrix}$$

즉 $\partial L/\partial CS_{t-1}$은 다음과 같이 구할 수 있습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$$

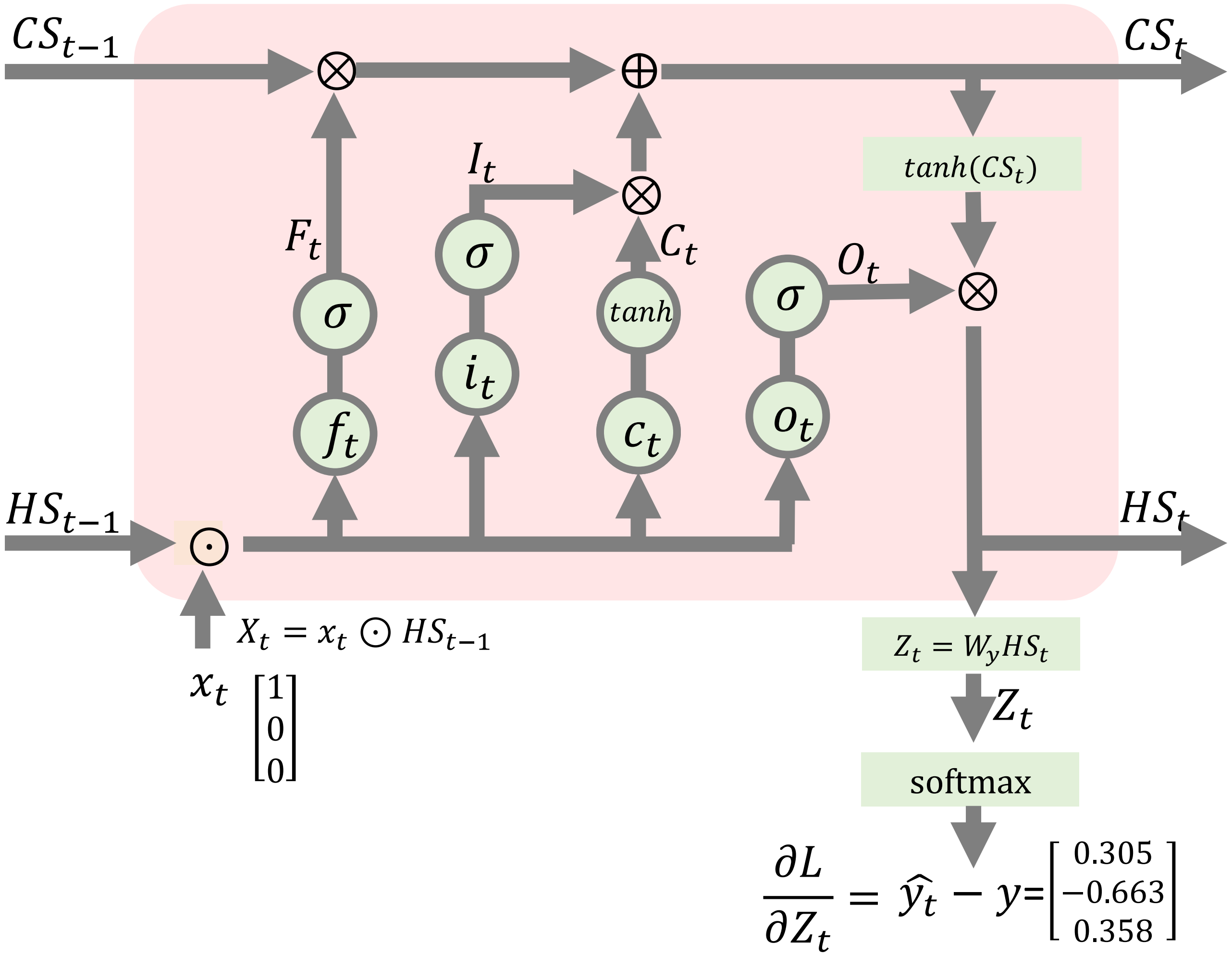
$Z_t = W_y HS_t$

$\boxed{CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t}$

$$\frac{\partial L}{\partial CS_{t-1}} = \frac{\partial L}{\partial CS_t} \frac{\partial CS_t}{\partial CS_{t-1}}$$

$$\frac{\partial CS_t}{\partial CS_{t-1}} = F_t$$

$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))\, F_t$$



신박AI

그리고 $\partial L/\partial HS_{t-1}$도 구해볼 수 있습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y)W_y O_t(1 - tanh^2(CS_t))$

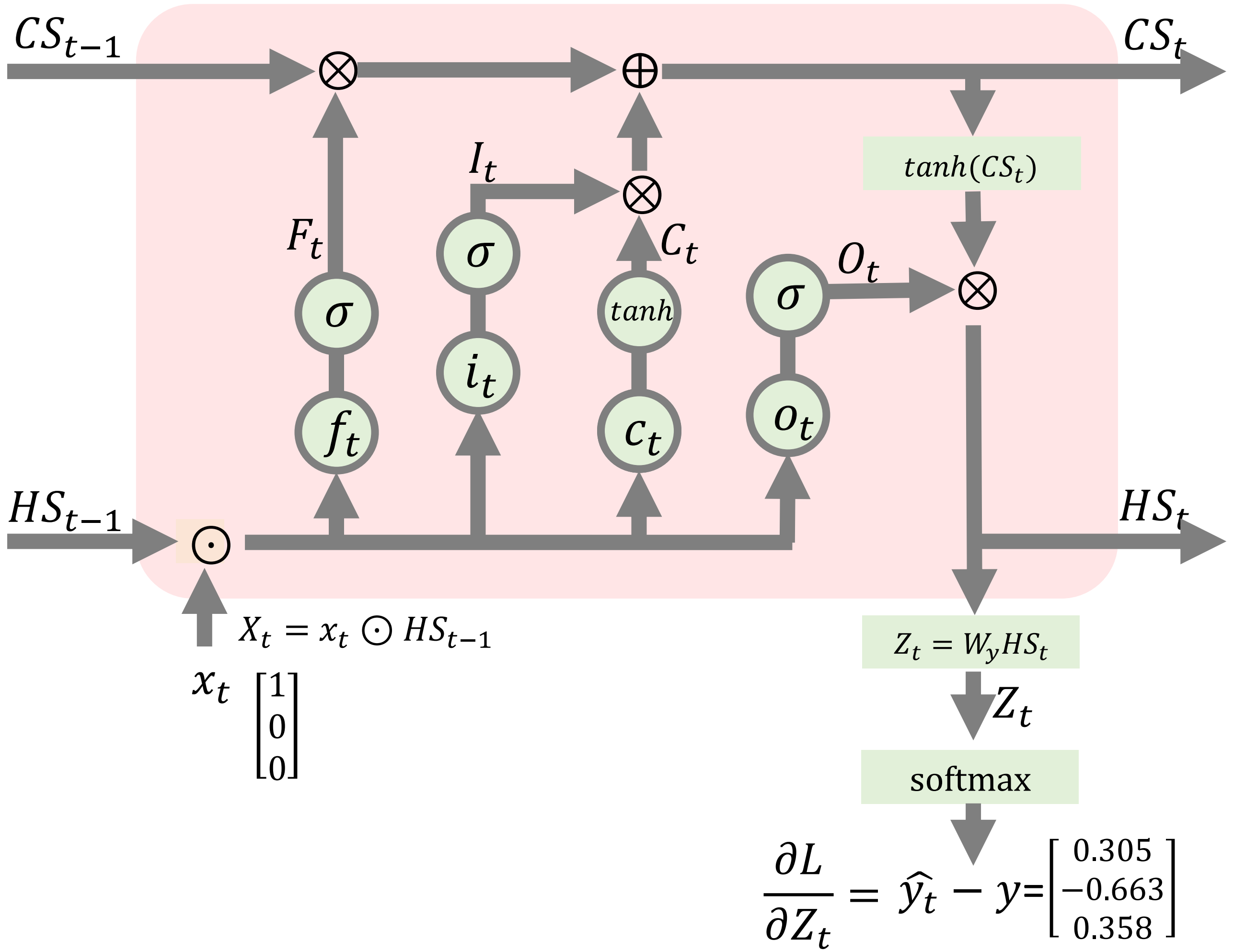
Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial HS_{t-1}} =$$



신박AI

여기서 보시듯, $\partial L/\partial HS_{t-1}$에 영향을 주는 루트는 네 곳입니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

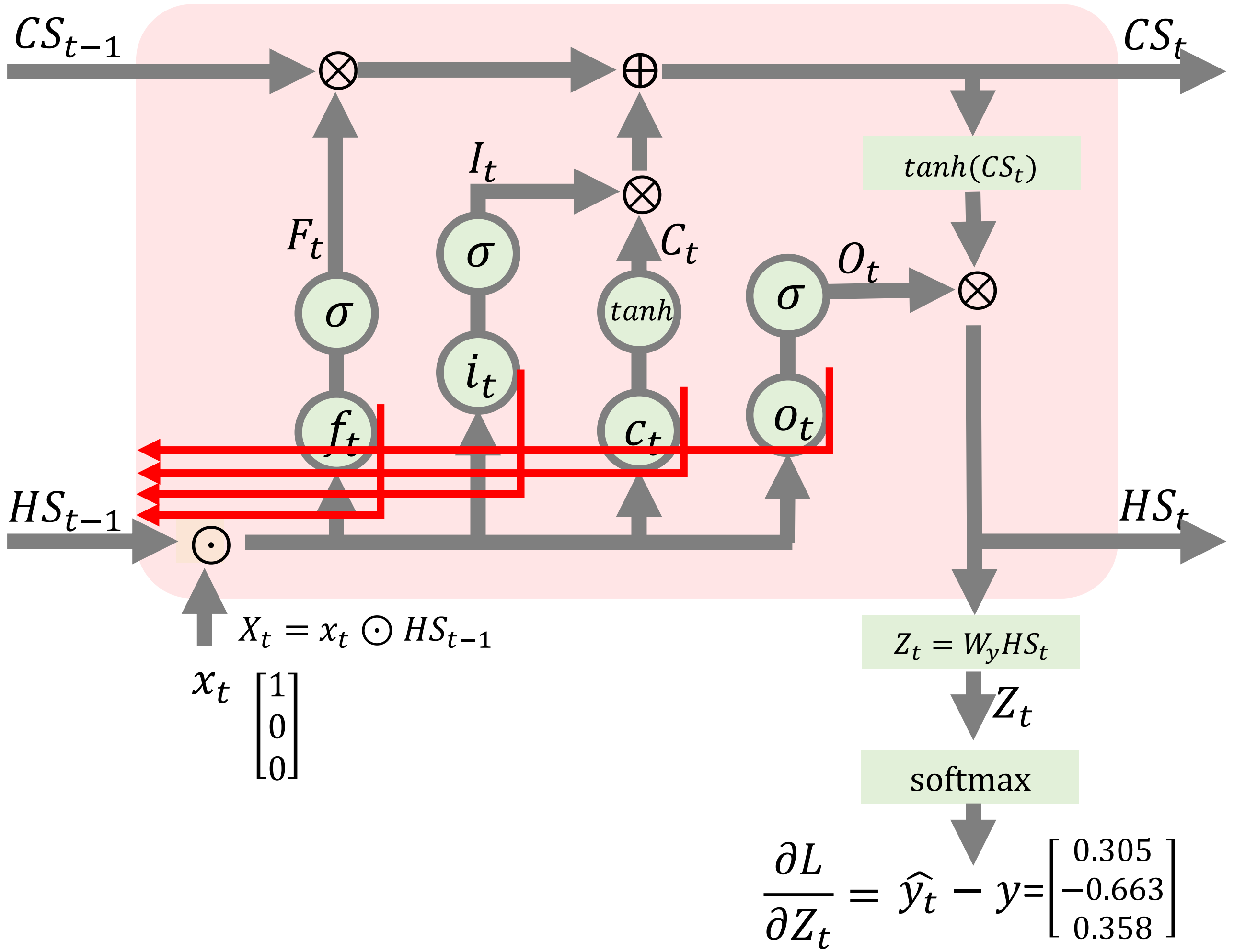
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial HS_{t-1}} =$$



신박AI

그래서 이렇게 네개의 항으로 나눌수 있습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

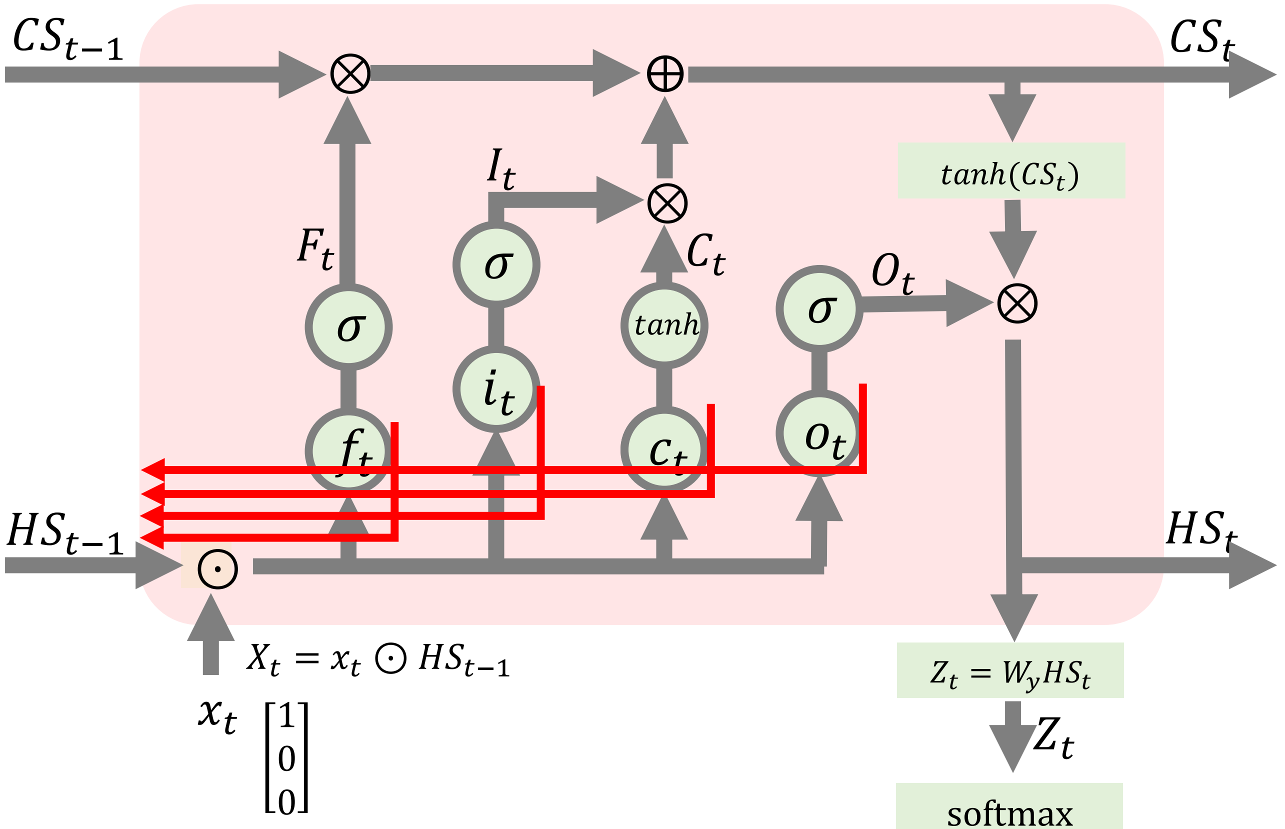
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial HS_{t-1}} = \frac{\partial L}{\partial f_t}\frac{\partial f_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial i_t}\frac{\partial i_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial c_t}\frac{\partial c_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial o_t}\frac{\partial o_t}{\partial HS_{t-1}}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix}0.305\\-0.663\\0.358\end{bmatrix}$$

신박AI

제가 이렇게 네 개의 항을 더하는 식을 보여드린 이유는

Forget Gate: $F_t = \sigma(W_f X_t)$

$F_t = \sigma(f_t)$

$f_t = W_f X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$

$C_t = tanh(c_t)$

$c_t = W_c X_t$

Input Gate: $I_t = \sigma(W_i X_t)$

$I_t = \sigma(i_t)$

$i_t = W_i X_t$

Output Gate: $O_t = \sigma(W_o X_t)$

$O_t = \sigma(o_t)$

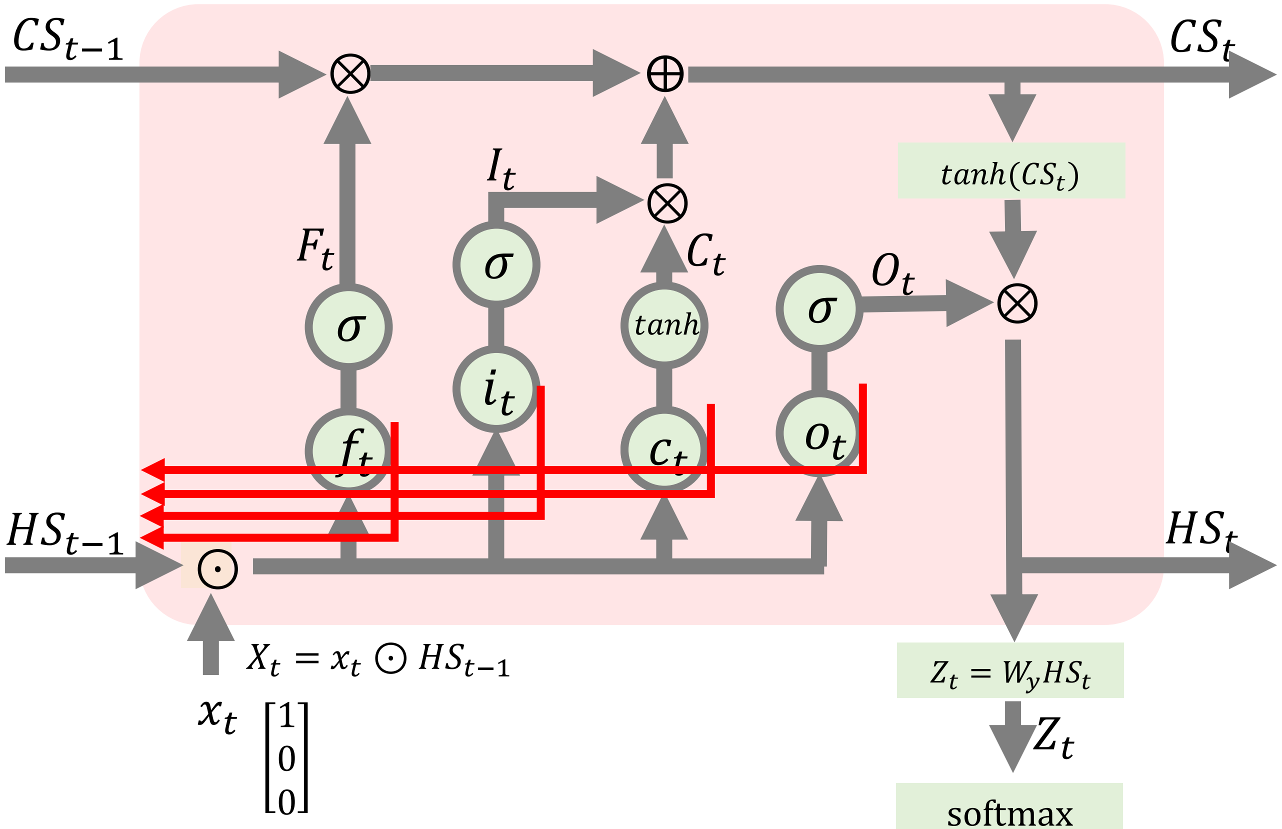
$o_t = W_o X_t$

$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial HS_{t-1}} = \frac{\partial L}{\partial f_t}\frac{\partial f_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial i_t}\frac{\partial i_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial c_t}\frac{\partial c_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial o_t}\frac{\partial o_t}{\partial HS_{t-1}}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix}0.305\\-0.663\\0.358\end{bmatrix}$$

신박AI

시간을 따라 전달되는 은닉상태의 기울기가 RNN에 비해서 쉽게 0으로 근접해지지 않는다는 것을

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

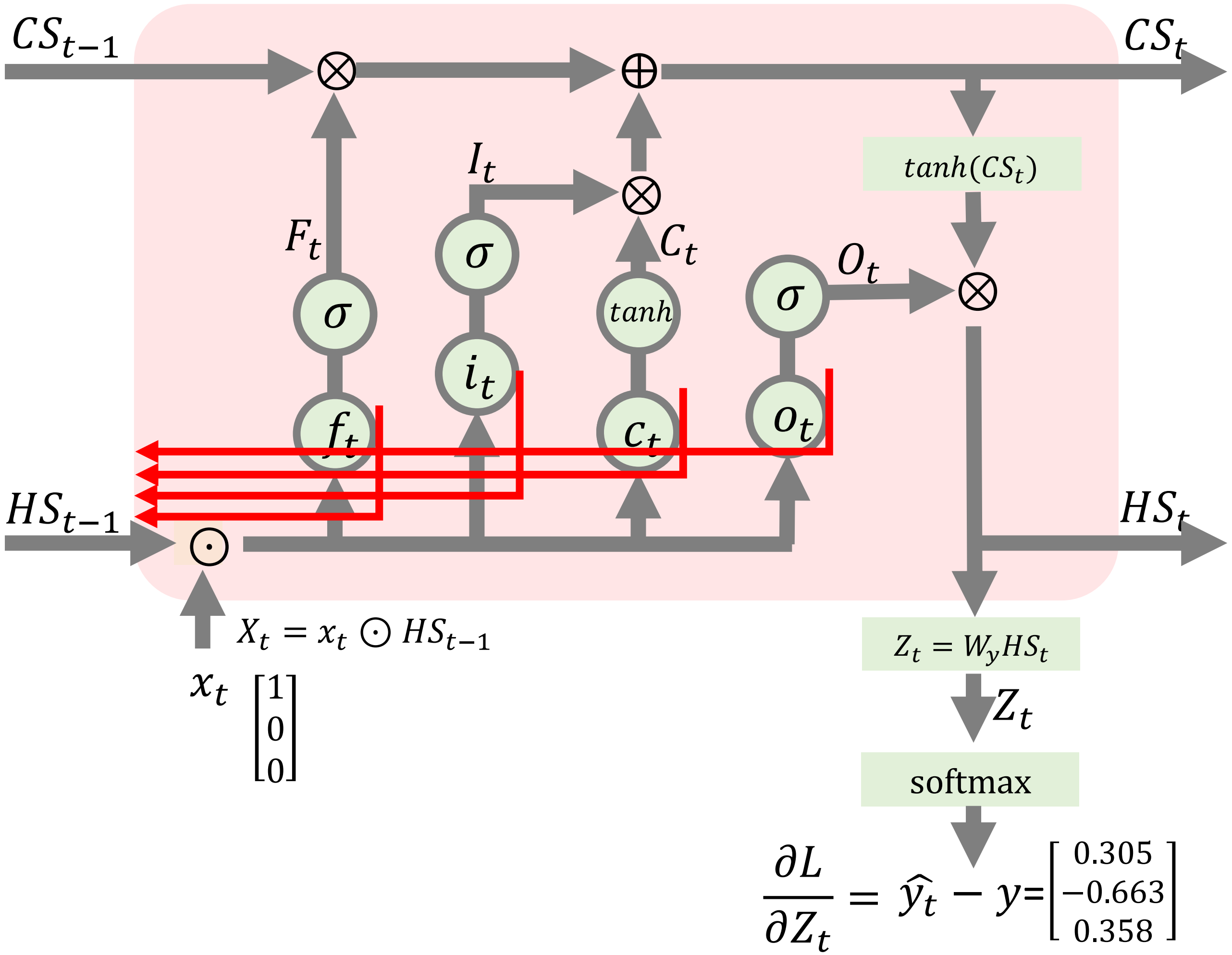
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial HS_{t-1}} = \frac{\partial L}{\partial f_t}\frac{\partial f_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial i_t}\frac{\partial i_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial c_t}\frac{\partial c_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial o_t}\frac{\partial o_t}{\partial HS_{t-1}}$$



보여드리고 싶었습니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

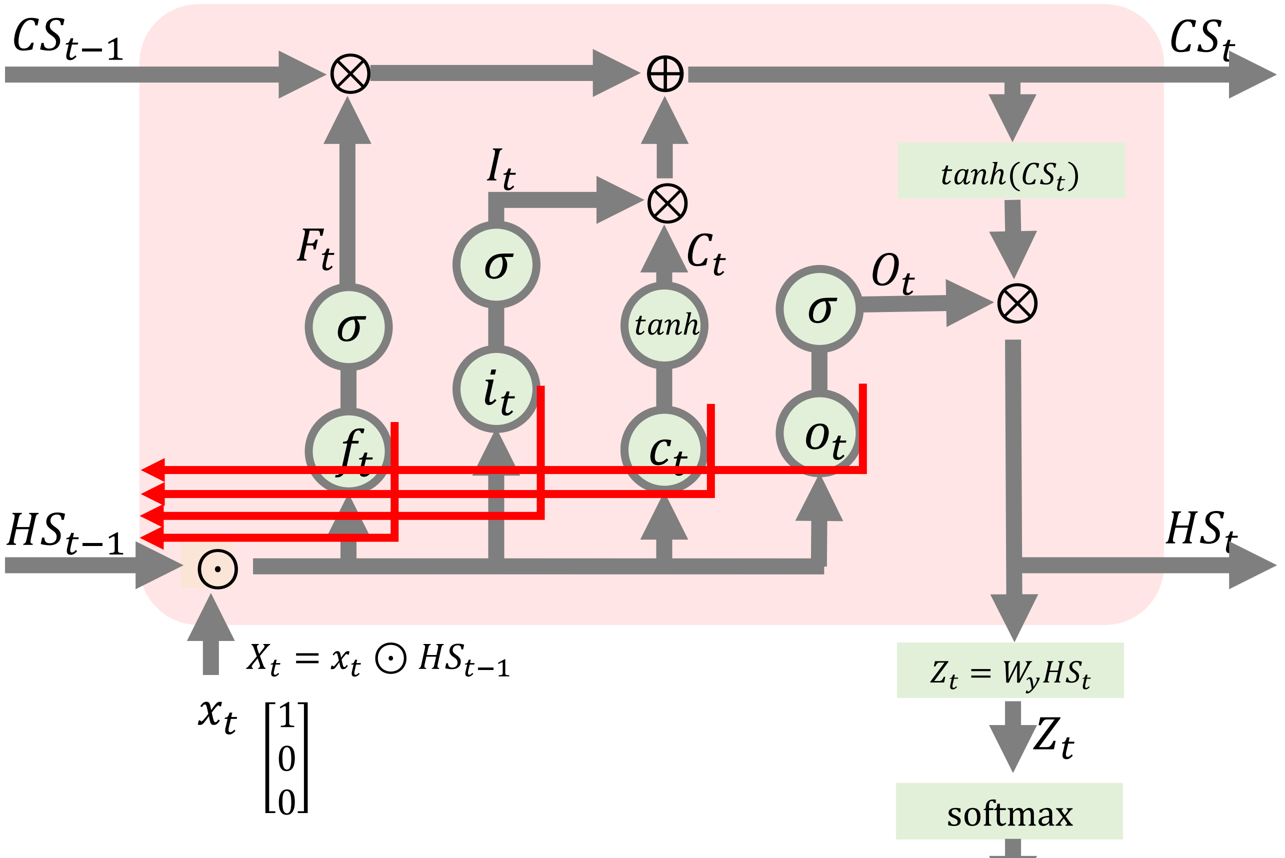
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial HS_{t-1}} = \frac{\partial L}{\partial f_t}\frac{\partial f_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial i_t}\frac{\partial i_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial c_t}\frac{\partial c_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial o_t}\frac{\partial o_t}{\partial HS_{t-1}}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix}0.305\\-0.663\\0.358\end{bmatrix}$$

왜냐하면, 만약 1보다 작은 수를 계속 곱하는 형태로 전달이 된다면,

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

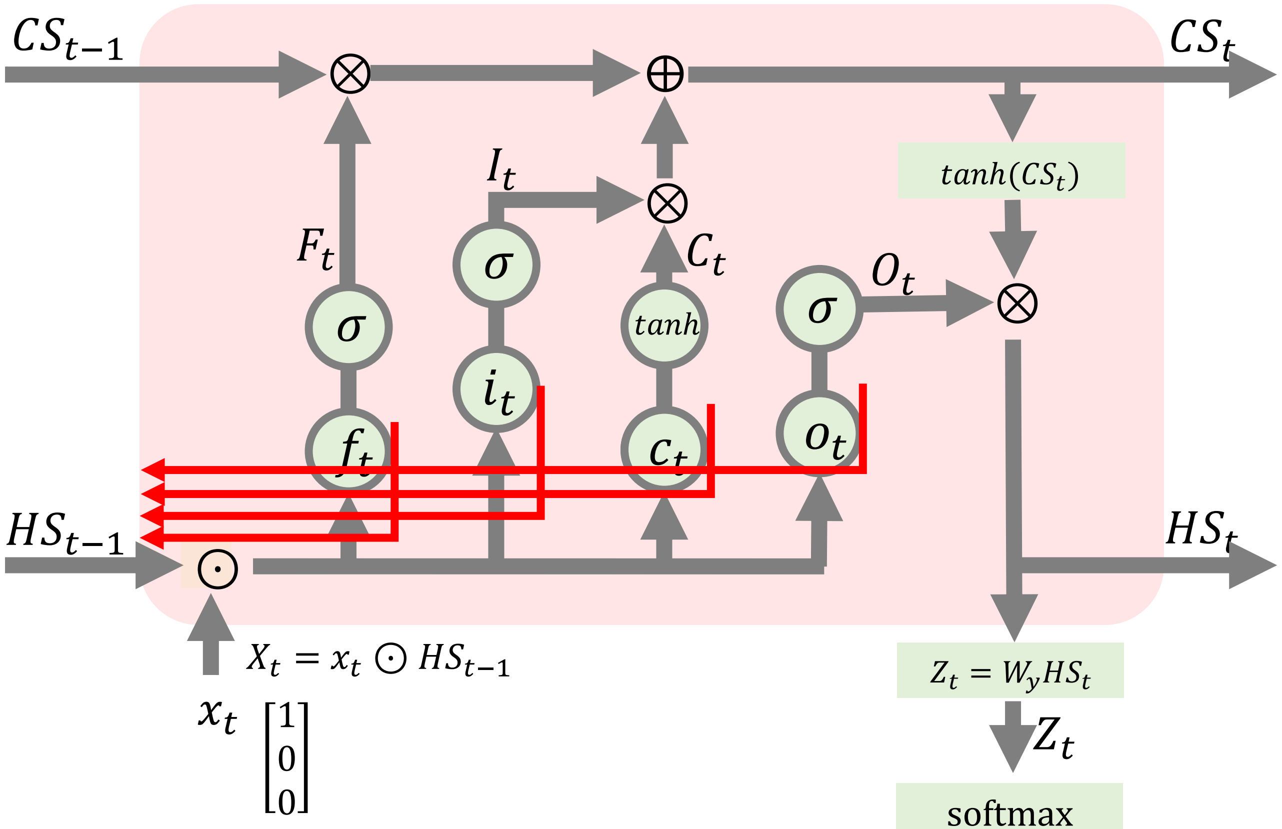
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial HS_{t-1}} = \frac{\partial L}{\partial f_t}\frac{\partial f_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial i_t}\frac{\partial i_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial c_t}\frac{\partial c_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial o_t}\frac{\partial o_t}{\partial HS_{t-1}}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix}0.305\\-0.663\\0.358\end{bmatrix}$$

언젠가는 0에 수렴하여 장기 의존성 문제가 발생할 수 있지만,

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

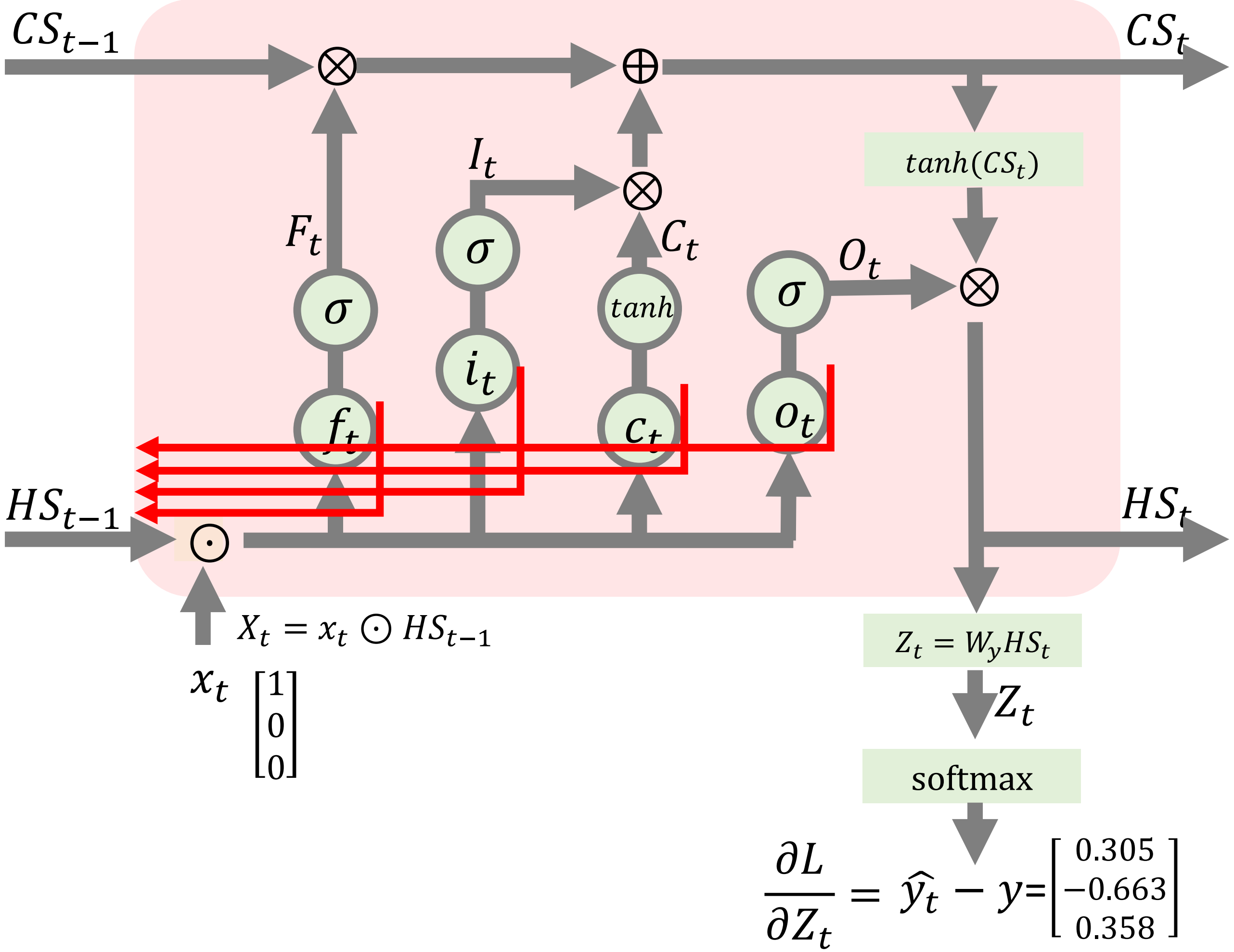
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial HS_{t-1}} = \frac{\partial L}{\partial f_t}\frac{\partial f_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial i_t}\frac{\partial i_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial c_t}\frac{\partial c_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial o_t}\frac{\partial o_t}{\partial HS_{t-1}}$$



이 LSTM의 경우는 곱하는 것이 아닌 더하는 형태로 전달되기 때문에

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

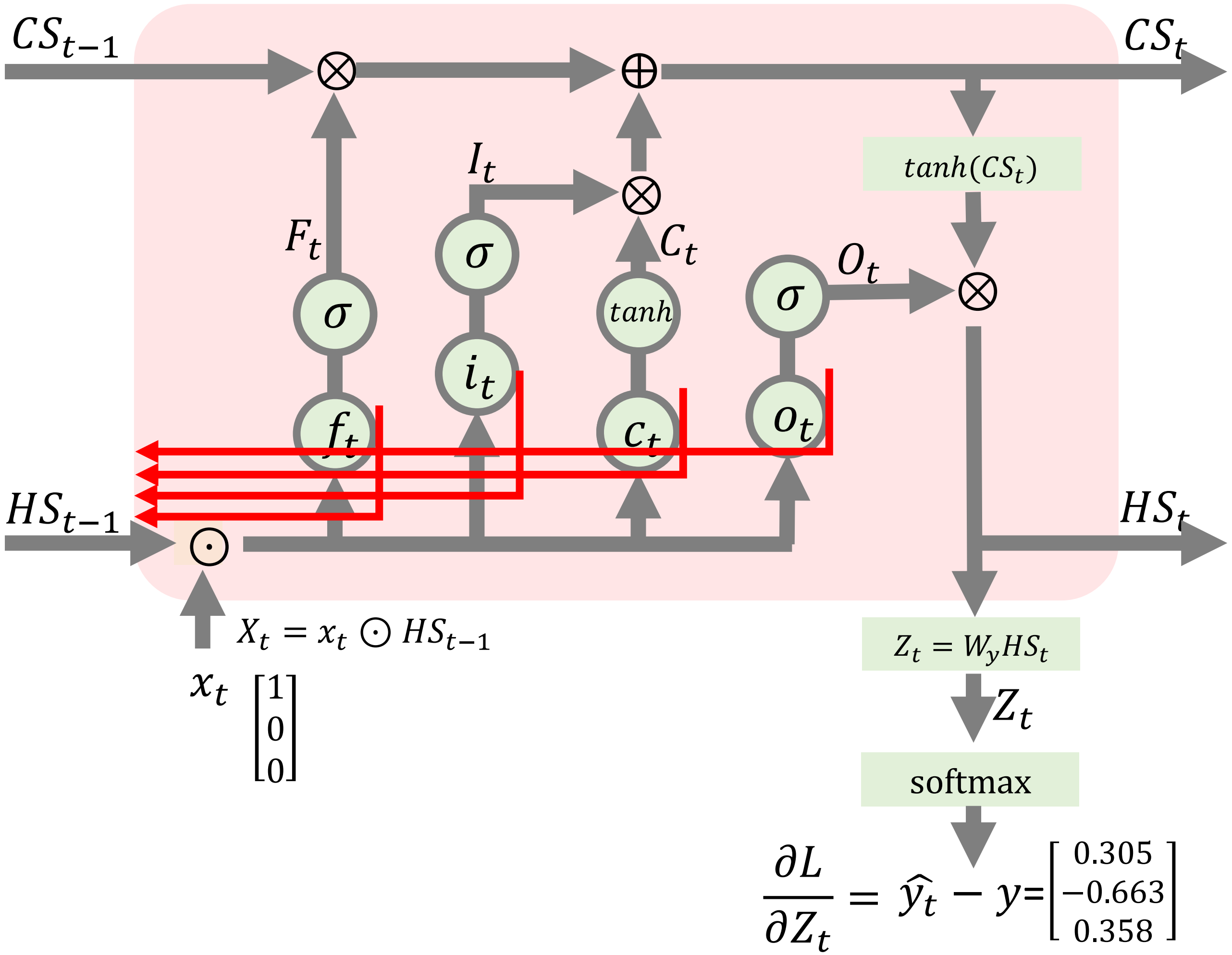
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial HS_{t-1}} = \frac{\partial L}{\partial f_t}\frac{\partial f_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial i_t}\frac{\partial i_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial c_t}\frac{\partial c_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial o_t}\frac{\partial o_t}{\partial HS_{t-1}}$$



신박AI

RNN에 비해서 장기 의존성 문제가 잘 발생하지 않을 수 있고

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$

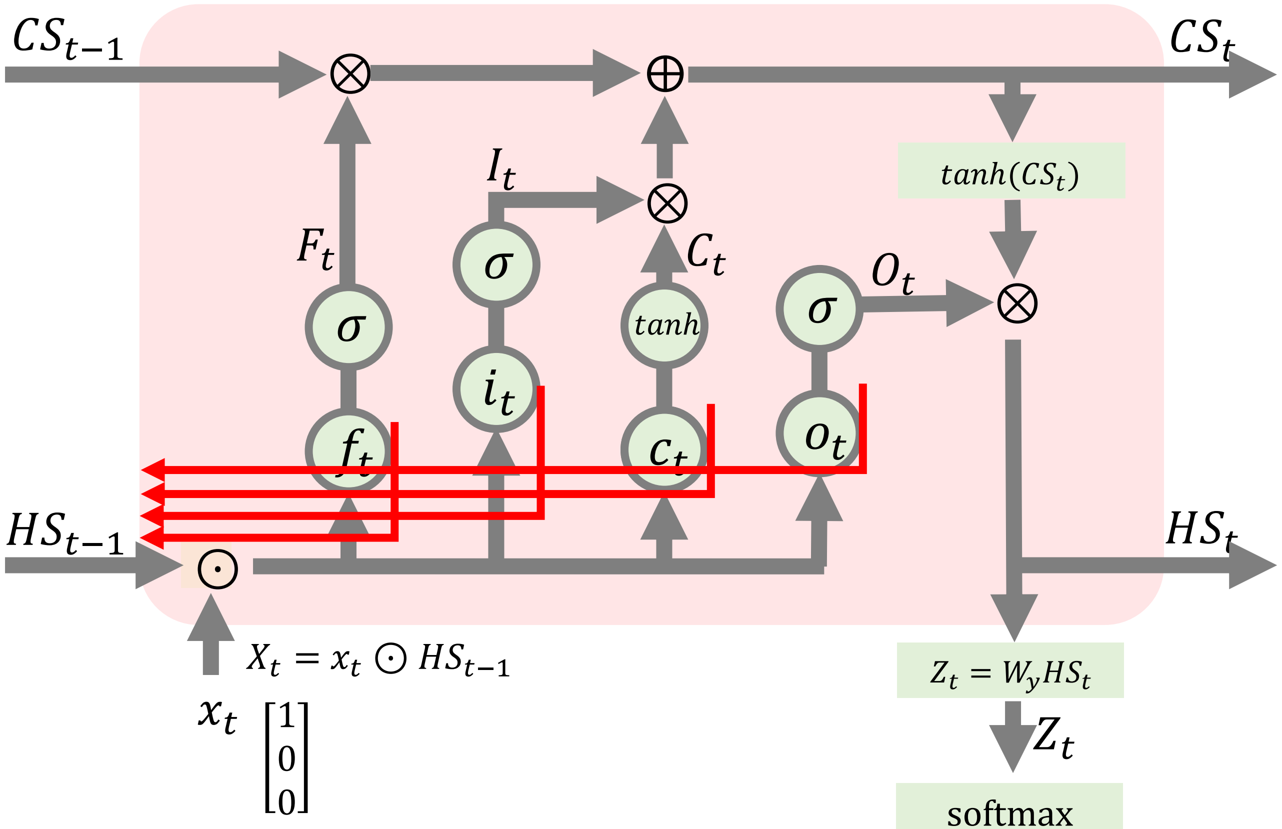
Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial HS_{t-1}} = \frac{\partial L}{\partial f_t}\frac{\partial f_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial i_t}\frac{\partial i_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial c_t}\frac{\partial c_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial o_t}\frac{\partial o_t}{\partial HS_{t-1}}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix}0.305\\-0.663\\0.358\end{bmatrix}$$

신박AI

또 이와 같은 셀 상태의 변화량도

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

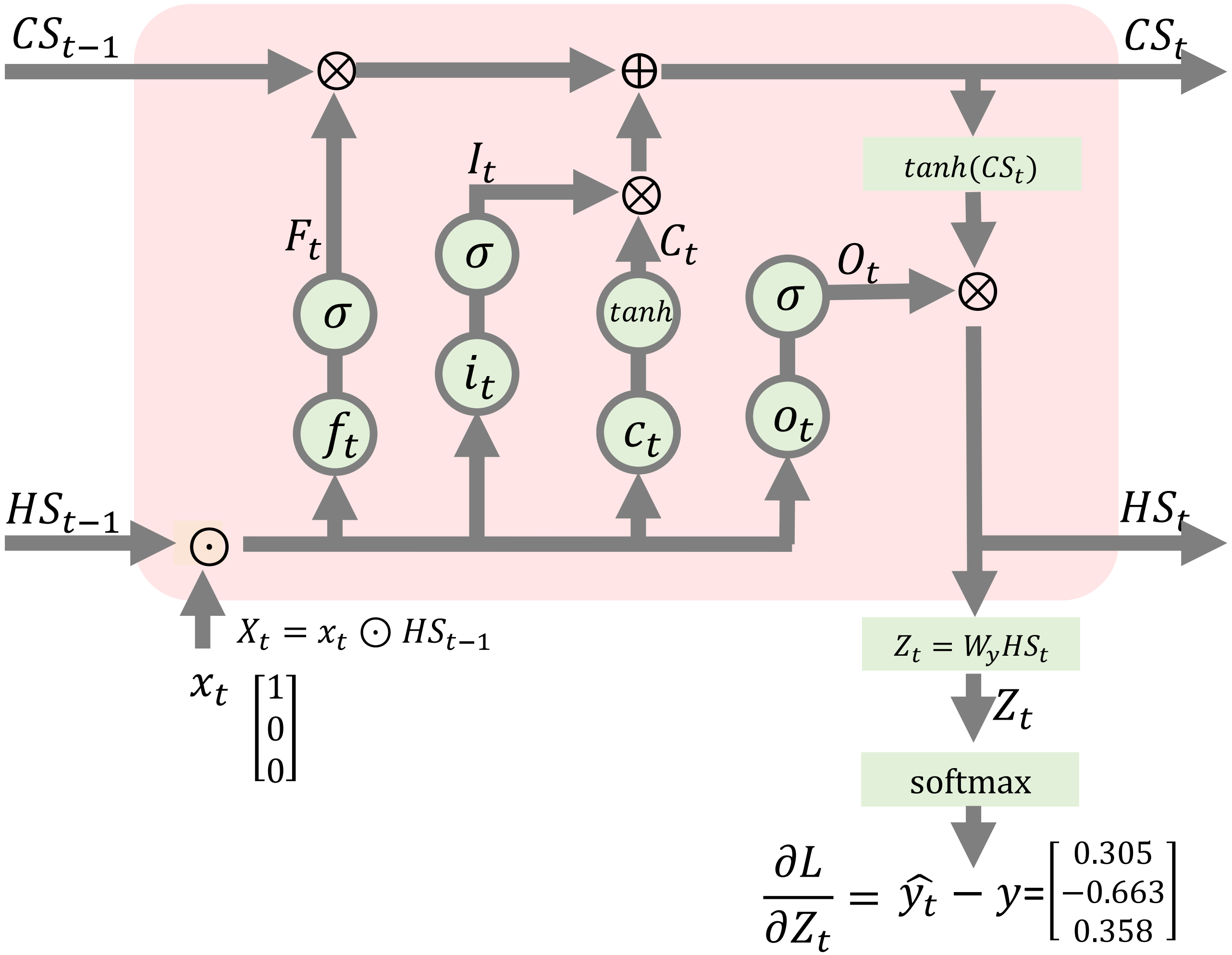
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial HS_{t-1}} = \frac{\partial L}{\partial f_t}\frac{\partial f_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial i_t}\frac{\partial i_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial c_t}\frac{\partial c_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial o_t}\frac{\partial o_t}{\partial HS_{t-1}}$$

$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))\, F_t$$



비록 앞 항들은 곱하기로 연결이 되어 있지만..

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

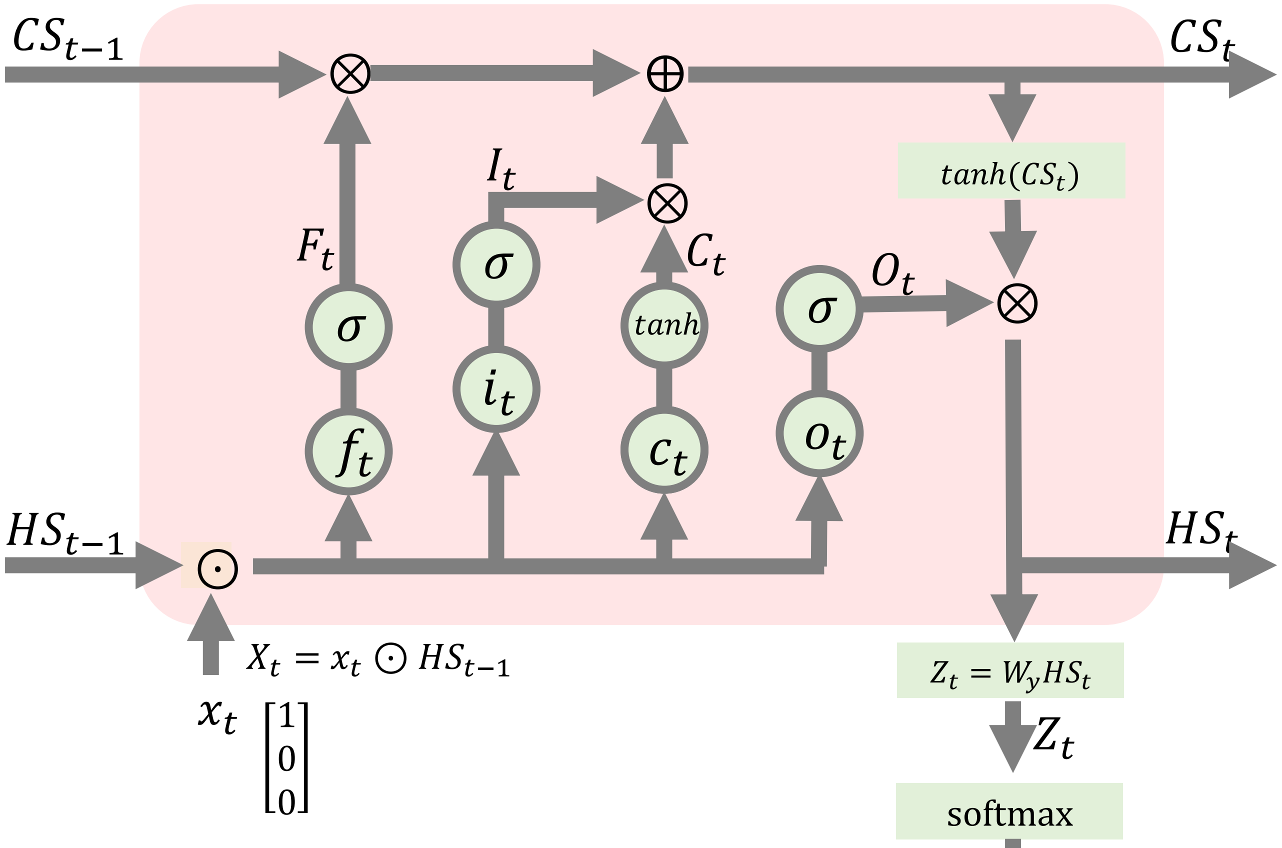
$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y)W_y O_t(1 - tanh^2(CS_t))$$

$$Z_t = W_y HS_t$$

$$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$$

$$\frac{\partial L}{\partial HS_{t-1}} = \frac{\partial L}{\partial f_t}\frac{\partial f_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial i_t}\frac{\partial i_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial c_t}\frac{\partial c_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial o_t}\frac{\partial o_t}{\partial HS_{t-1}}$$

$$\frac{\partial L}{\partial CS_t} = \boxed{(\hat{y}_t - y)W_y O_t(1 - tanh^2(CS_t))} F_t$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix}0.305\\-0.663\\0.358\end{bmatrix}$$

셀상태의 변화량 계산에는 이 forget gate가 있어서

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

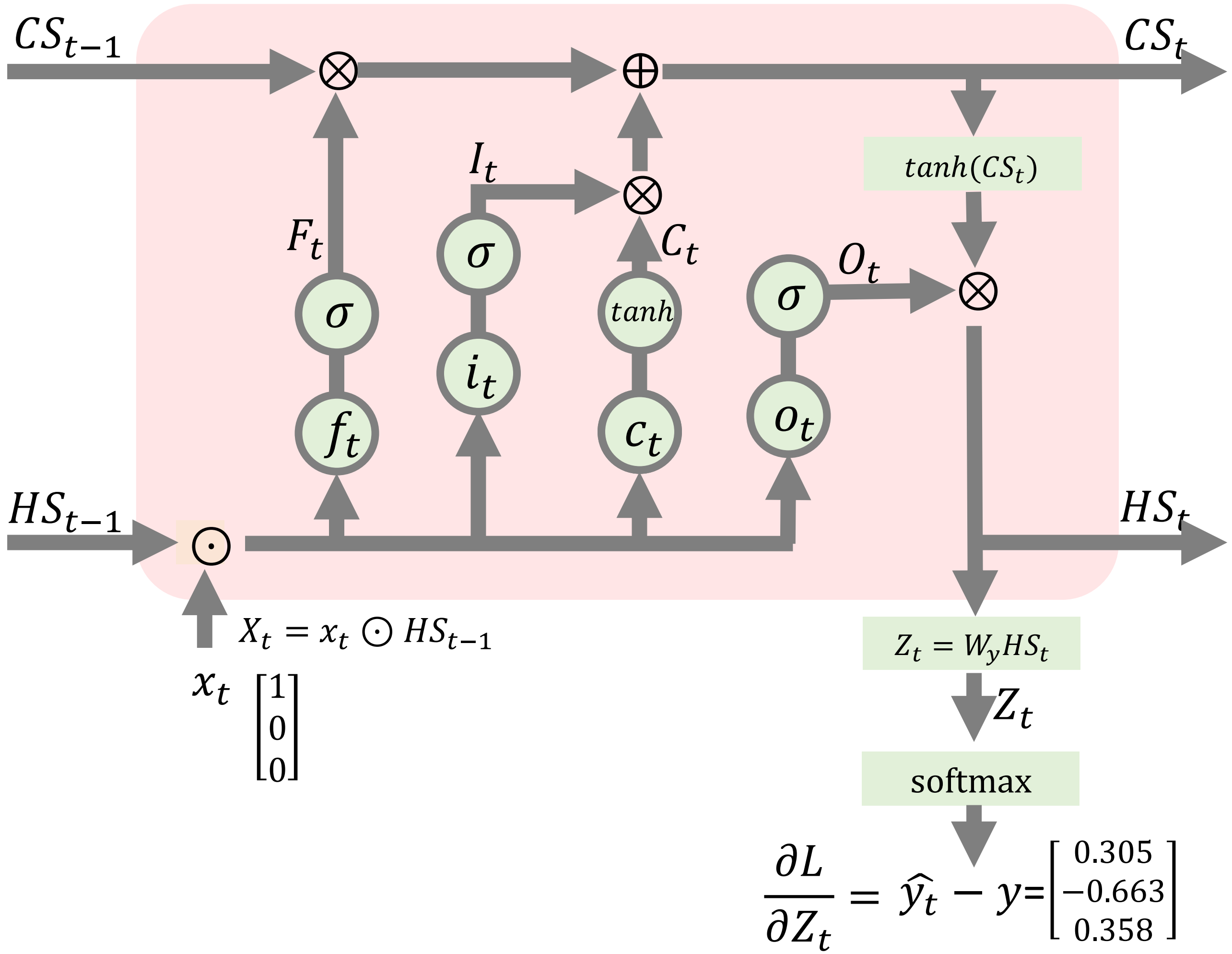
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial HS_{t-1}} = \frac{\partial L}{\partial f_t}\frac{\partial f_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial i_t}\frac{\partial i_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial c_t}\frac{\partial c_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial o_t}\frac{\partial o_t}{\partial HS_{t-1}}$$

$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t)) \boxed{F_t}$$



변화량이 0으로 수렴해지는 것을 막아줍니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

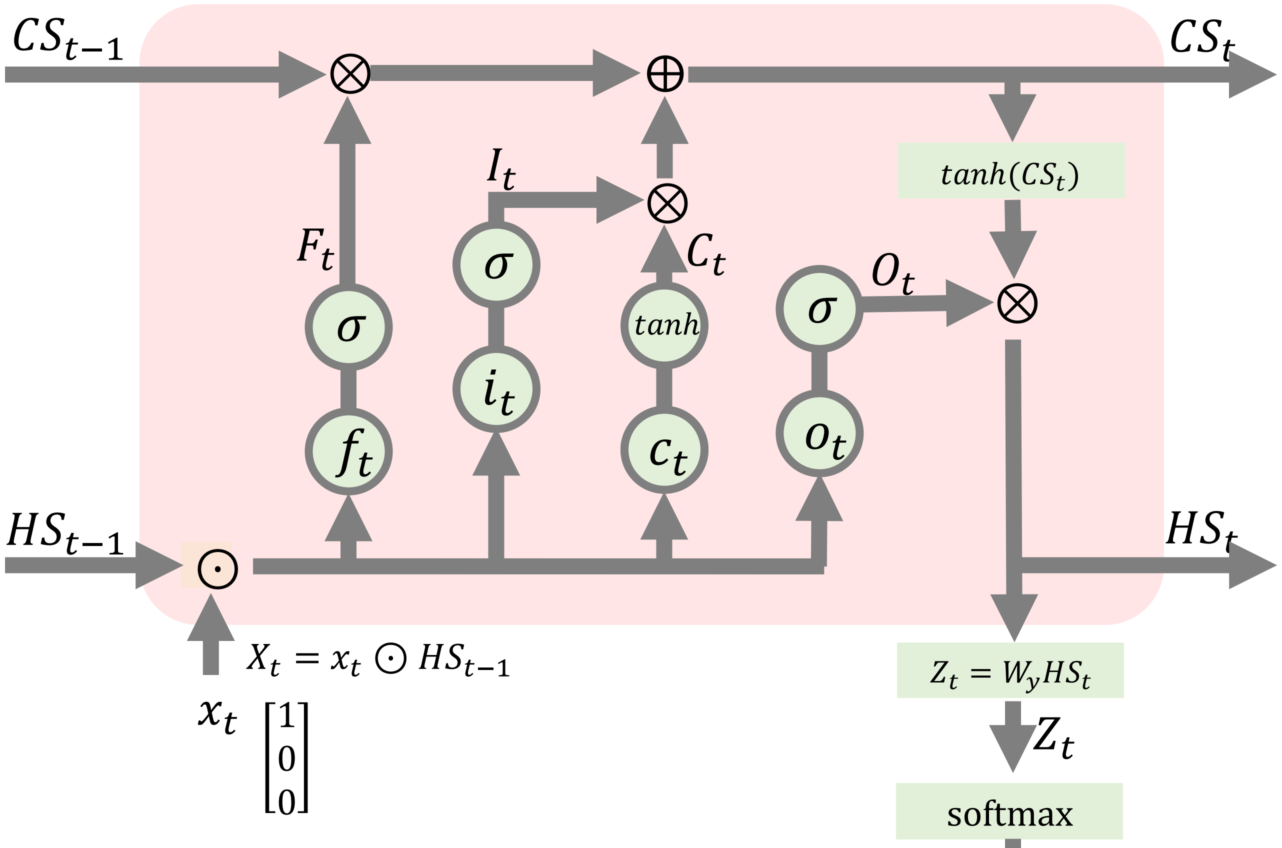
$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y)W_y O_t(1 - tanh^2(CS_t))$

$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial HS_{t-1}} = \frac{\partial L}{\partial f_t}\frac{\partial f_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial i_t}\frac{\partial i_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial c_t}\frac{\partial c_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial o_t}\frac{\partial o_t}{\partial HS_{t-1}}$$

$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y)W_y O_t(1 - tanh^2(CS_t))\boxed{F_t}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix} 0.305 \\ -0.663 \\ 0.358 \end{bmatrix}$$

왜냐하면 Forget Gate는 내부가 시그모이드 함수로 되어 있어서

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

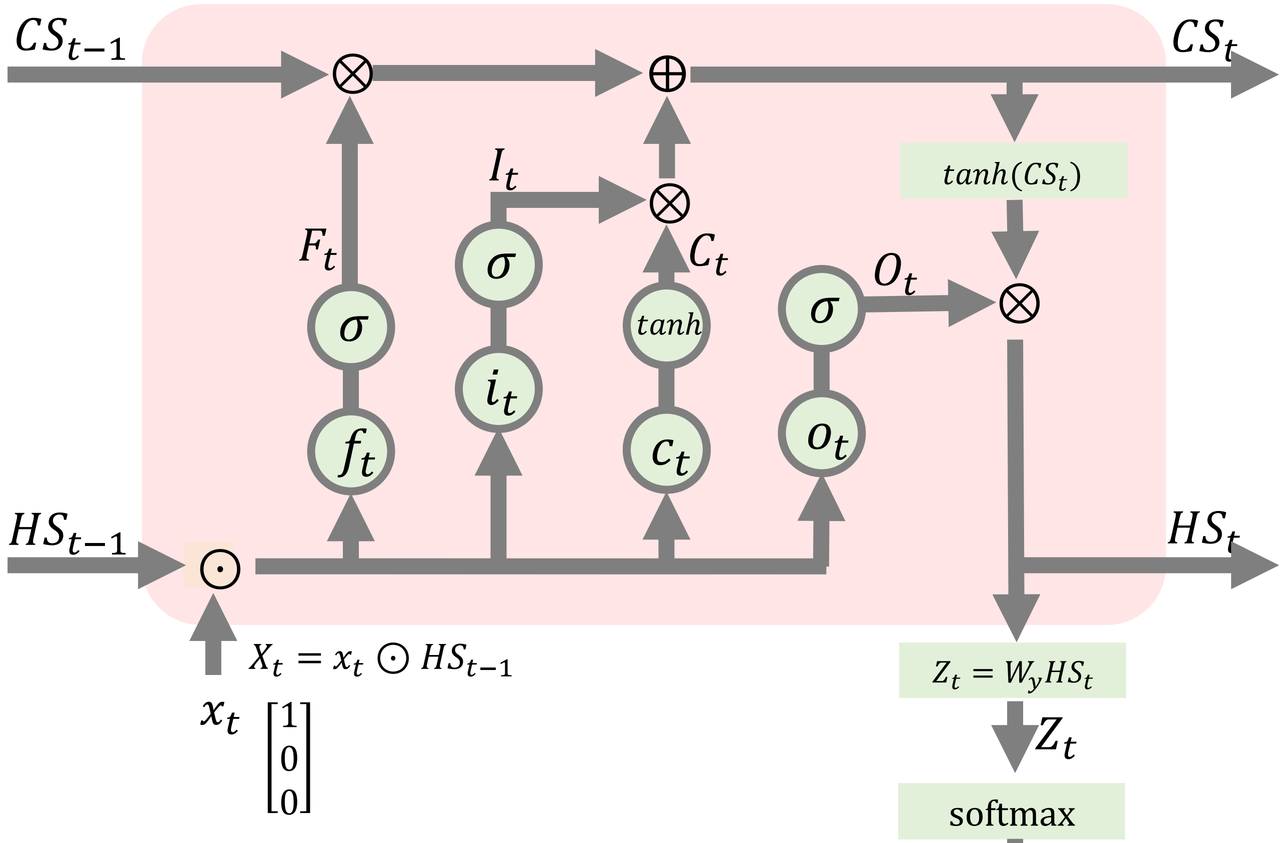
$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$

$Z_t = W_y HS_t$

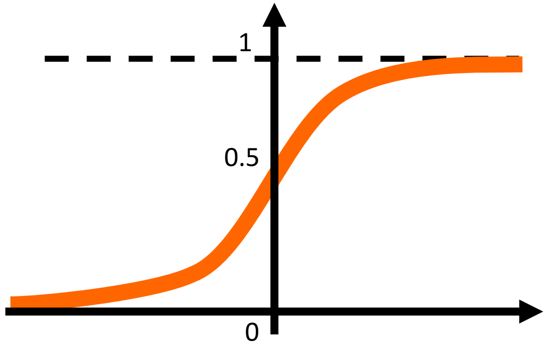
$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial HS_{t-1}} = \frac{\partial L}{\partial f_t}\frac{\partial f_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial i_t}\frac{\partial i_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial c_t}\frac{\partial c_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial o_t}\frac{\partial o_t}{\partial HS_{t-1}}$$

$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t)) \boxed{F_t}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix}0.305\\-0.663\\0.358\end{bmatrix}$$



앞의 곱셈항들이 0으로 수렴하려는 경향을 보여도

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

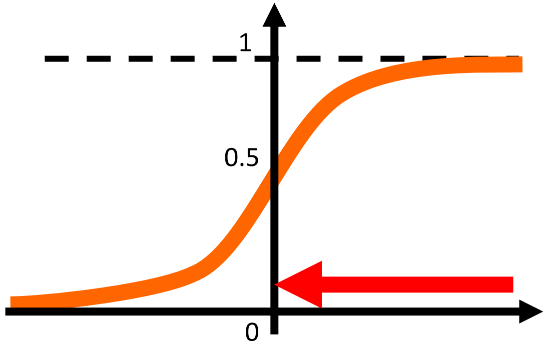
$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$

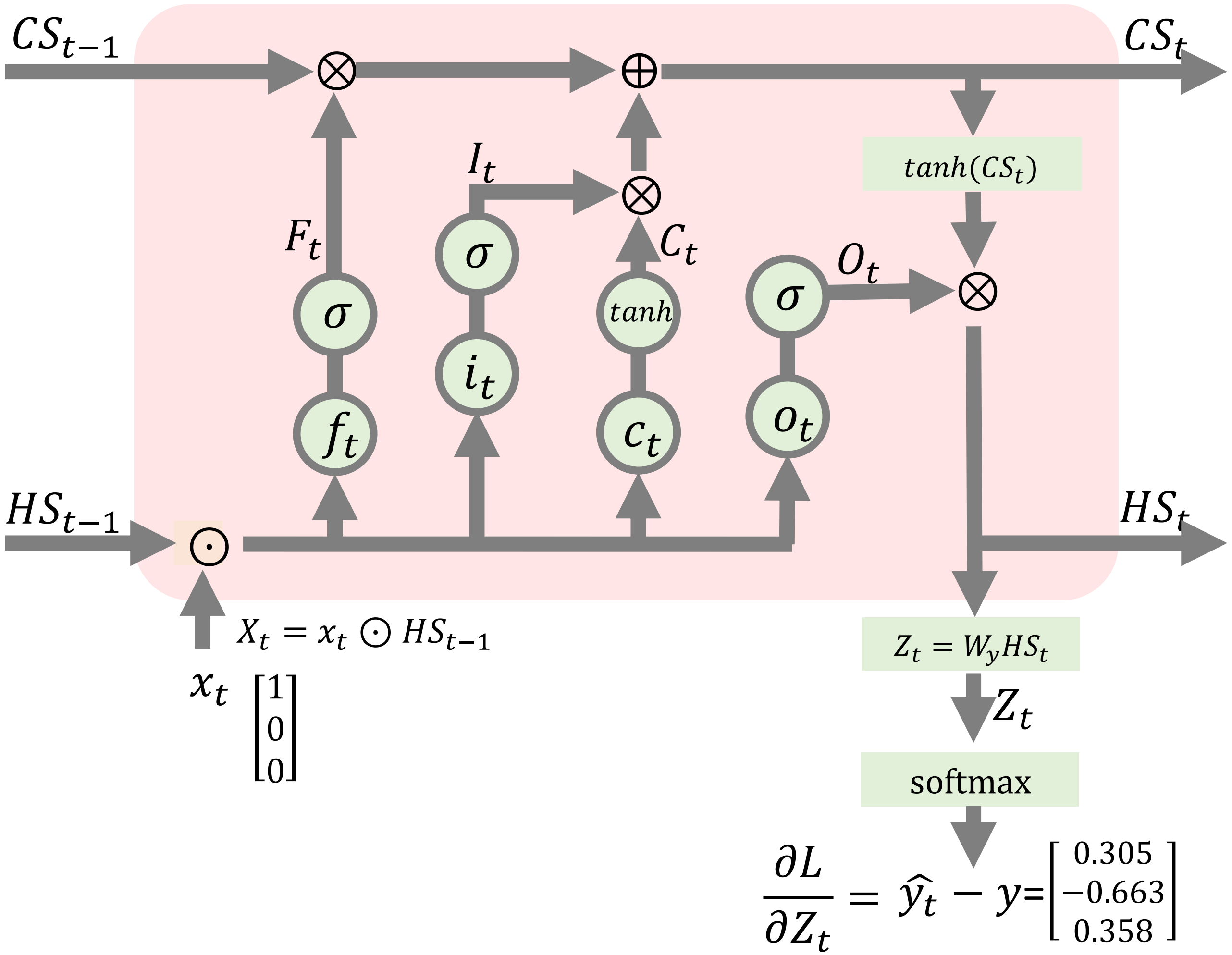
$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial HS_{t-1}} = \frac{\partial L}{\partial f_t}\frac{\partial f_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial i_t}\frac{\partial i_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial c_t}\frac{\partial c_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial o_t}\frac{\partial o_t}{\partial HS_{t-1}}$$

$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t)) \boxed{F_t}$$





신박AI

이 Forget Gate가 0에 수렴하지 않고 0.5 이상으로 변화량을 높여주기 때문에

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

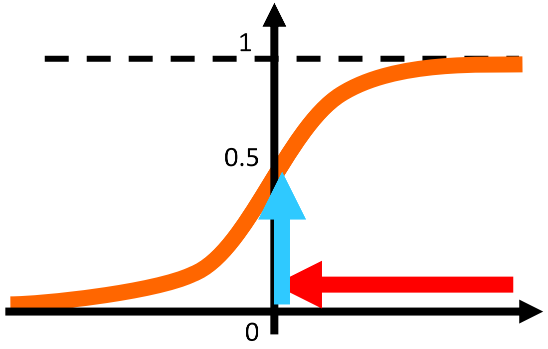
$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$

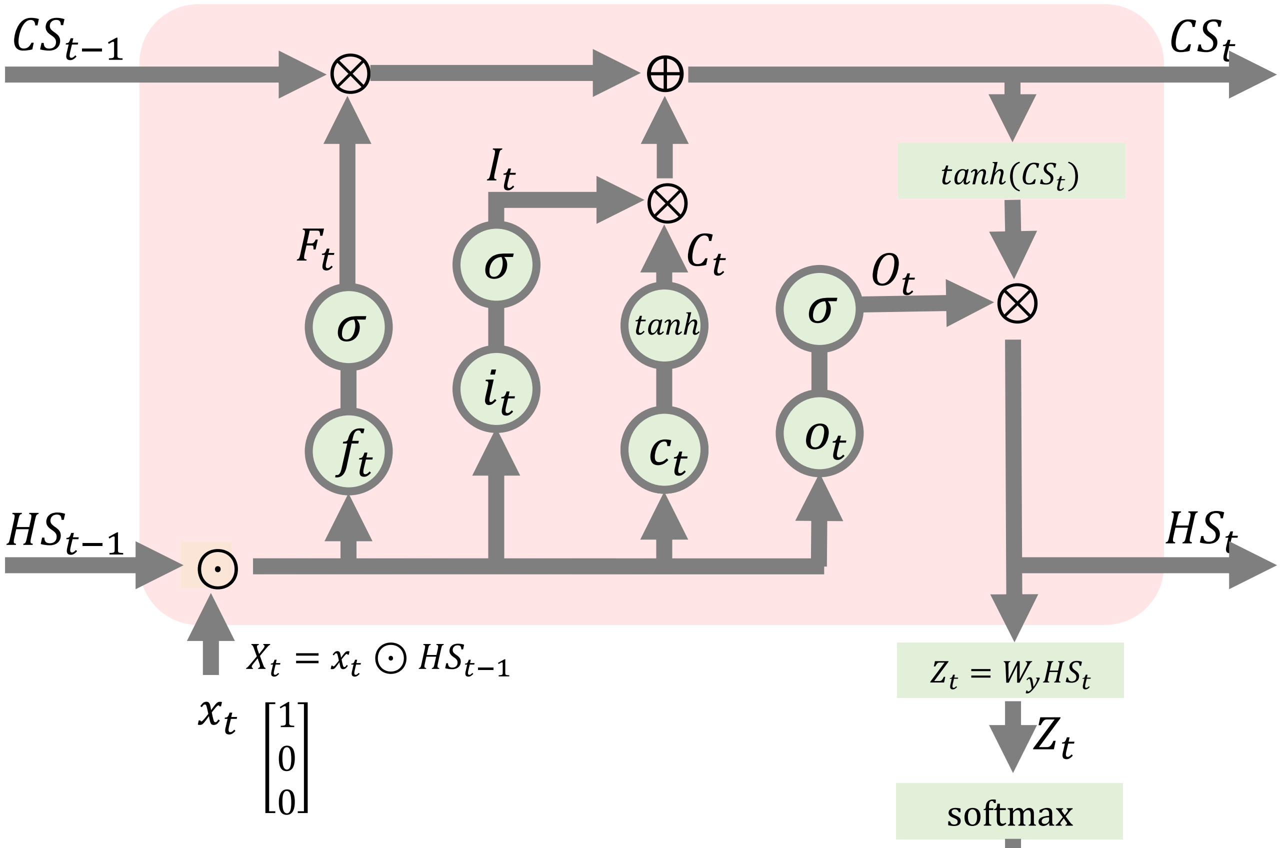
$Z_t = W_y HS_t$

$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial HS_{t-1}} = \frac{\partial L}{\partial f_t}\frac{\partial f_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial i_t}\frac{\partial i_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial c_t}\frac{\partial c_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial o_t}\frac{\partial o_t}{\partial HS_{t-1}}$$

$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t)) \boxed{F_t}$$





$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix}0.305\\-0.663\\0.358\end{bmatrix}$$

신박AI

기울기가 시간을 거슬러 갈 수록 0에 가까워지는 RNN에 비하면

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

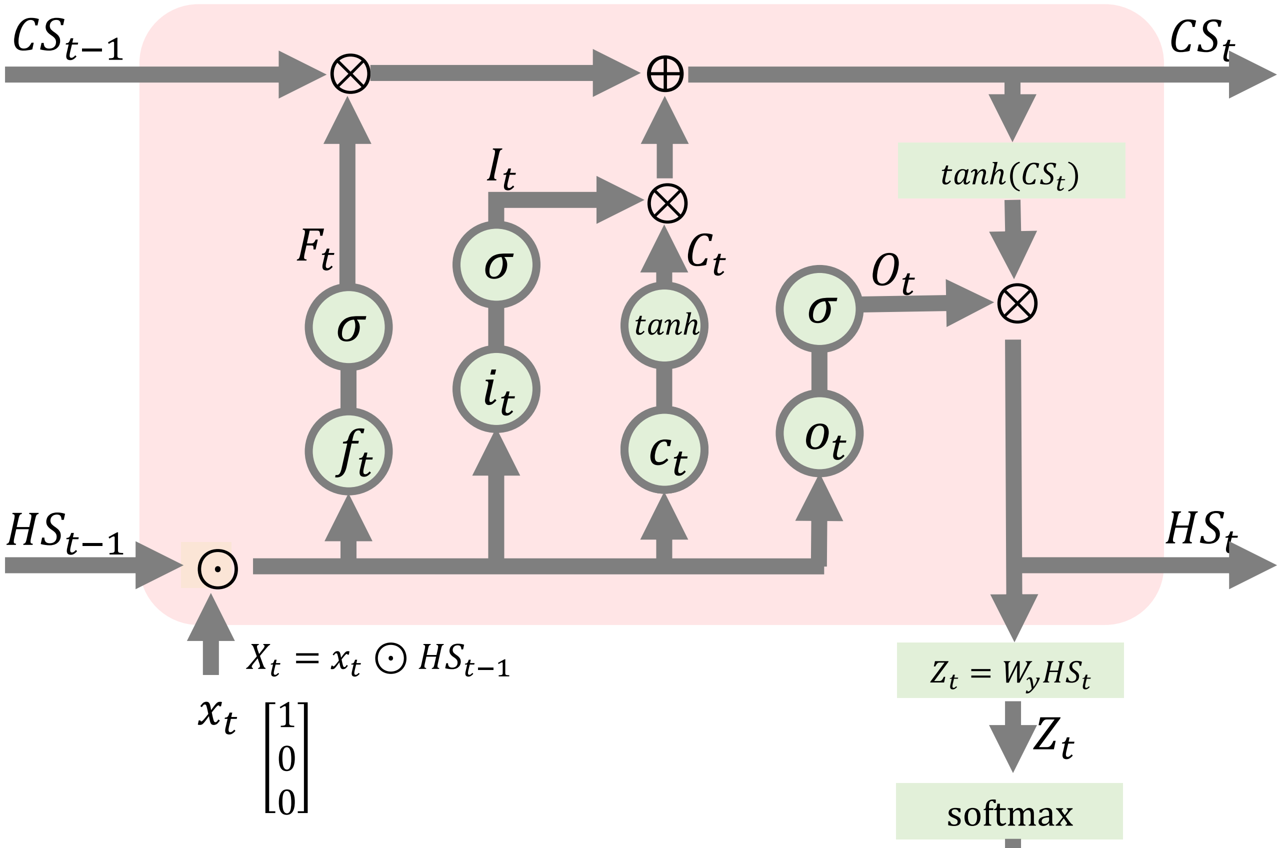
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$

$Z_t = W_y HS_t$

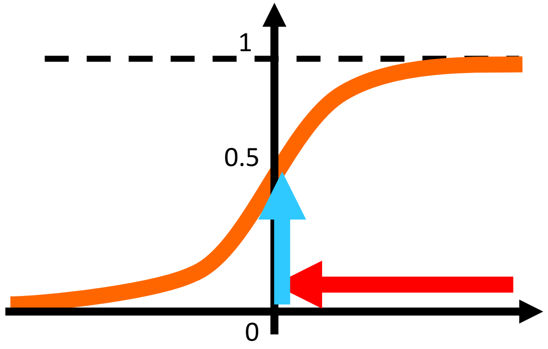
$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial HS_{t-1}} = \frac{\partial L}{\partial f_t}\frac{\partial f_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial i_t}\frac{\partial i_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial c_t}\frac{\partial c_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial o_t}\frac{\partial o_t}{\partial HS_{t-1}}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix}0.305\\-0.663\\0.358\end{bmatrix}$$

$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t)) \boxed{F_t}$$



신박AI

LSTM은 이러한 장치들로 인해 장기 의존성 문제가 쉽게 발생하지 않게 되는 것입니다

Forget Gate: $F_t = \sigma(W_f X_t)$
$F_t = \sigma(f_t)$
$f_t = W_f X_t$

Candidate Gate: $C_t = tanh(W_c X_t)$
$C_t = tanh(c_t)$
$c_t = W_c X_t$

Input Gate: $I_t = \sigma(W_i X_t)$
$I_t = \sigma(i_t)$
$i_t = W_i X_t$

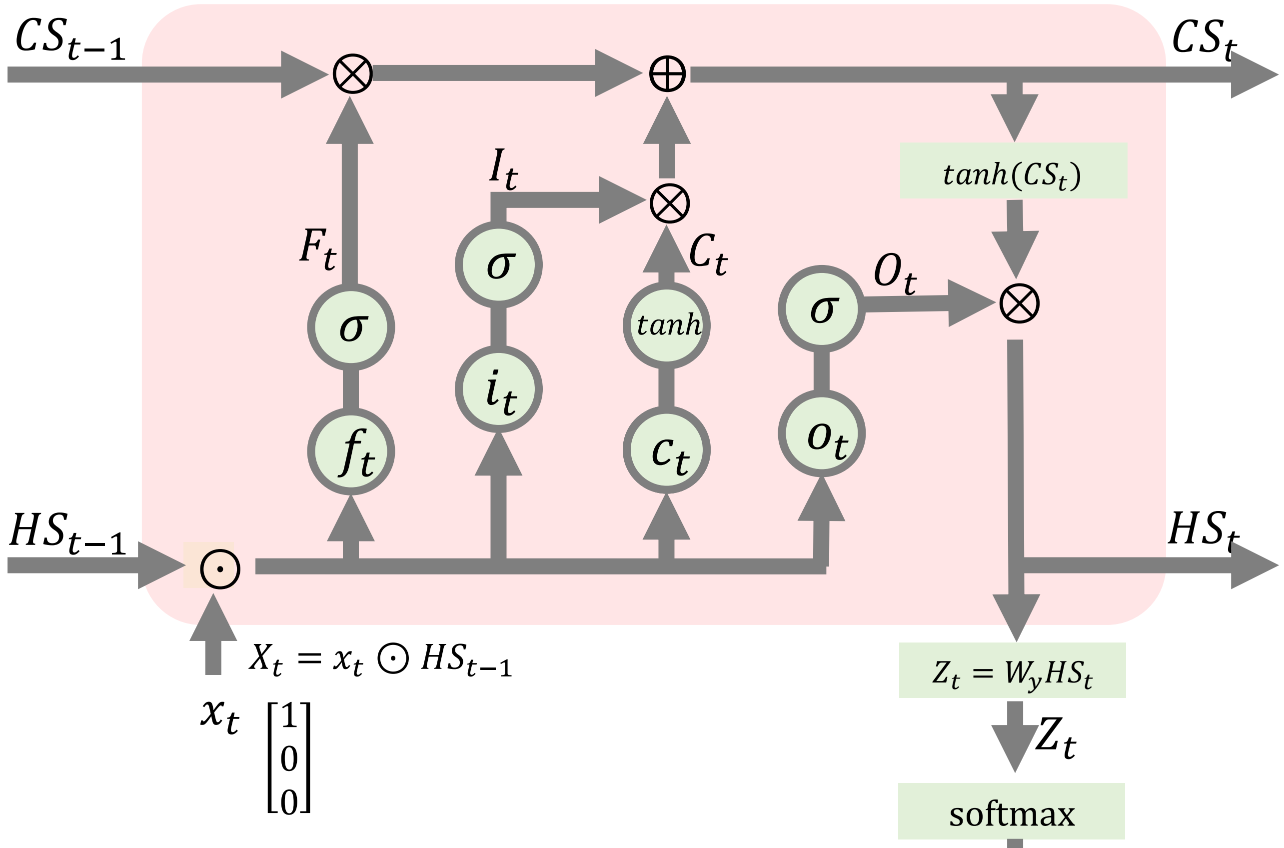
Output Gate: $O_t = \sigma(W_o X_t)$
$O_t = \sigma(o_t)$
$o_t = W_o X_t$

$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t))$$

$Z_t = W_y HS_t$

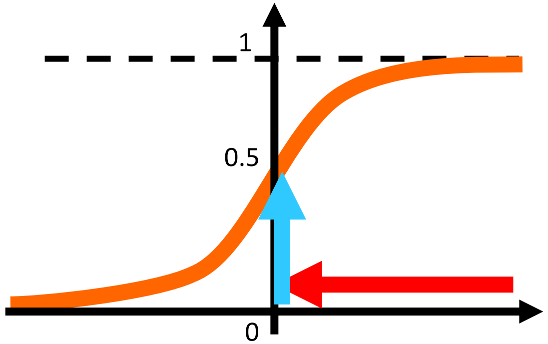
$CS_t = CS_{t-1} \otimes F_t + I_t \otimes C_t$

$$\frac{\partial L}{\partial HS_{t-1}} = \frac{\partial L}{\partial f_t}\frac{\partial f_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial i_t}\frac{\partial i_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial c_t}\frac{\partial c_t}{\partial HS_{t-1}} + \frac{\partial L}{\partial o_t}\frac{\partial o_t}{\partial HS_{t-1}}$$



$$\frac{\partial L}{\partial Z_t} = \hat{y}_t - y = \begin{bmatrix}0.305\\-0.663\\0.358\end{bmatrix}$$

$$\frac{\partial L}{\partial CS_t} = (\hat{y}_t - y) W_y O_t (1 - tanh^2(CS_t)) \boxed{F_t}$$



오늘 제가 준비한 LSTM영상은
여기까지 입니다

신박AI

딥러닝 공부하시는데 도움이
되셨기를 바라는 마음이 큽니다

다음시간에는 오늘 배운 이론을 바탕으로
어떻게 실제로 구현하는지

LSTM 구현을 해보는
시간을 갖도록 하겠습니다

오늘 긴 시간 시청해 주셔서..

신박AI

감사합니다!

좋은 하루 되세요!!

이 채널은 여러분의 관심과 사랑이 필요합니다

신박AI

‘좋아요’와 ‘구독’버튼은 강의 준비에 큰 힘이 됩니다!

신박AI

그리고 영상 자료를 사용하실때는
출처 '신박AI'를 밝혀주세요

신박AI