5. RNN

0. 전통적인 TimeSeries Data Prediction

RNN

Sequence Data 의 표현법

$$x = (\begin{pmatrix}0.1\\0.2\\0.3\end{pmatrix}, \begin{pmatrix}0.2\\0.3\\0.4\end{pmatrix}, \ldots, \begin{pmatrix}0.1\\0.2\\0.3\end{pmatrix})^T$$

0.4, 0.5, 0.6 으로 수정

- 일반화
  - $x^{(i)}$ 와 $x^{(j)}$ 사이에 의존성이 존재 한다.

$$x = (x^{(1)}, x^{(2)}, \ldots, x^{(\mathrm{T})})^T$$

특성

- 순서가 중요하다.
  - “i am”
  - “am i ?”
- 샘플마다 길이가 다르다.
  - 순환 신경망은 은닉층에 순환 에지를 부여하여 가변 길이 수용
  - 길이가 τ 인 데이터를 처리 하기 위해서는 은닉층이 τ번 나타나야 한다.
  - τ는 가변적이다.
  - “낮엔 파란 하는 별이보이는밤”
  - “기분 좋은 날 오랜만에 모일까?”
- 이전 데이터의 의존성

- 이전에 나타난 결과를 저장하여 이후에 나타나는 데이터에 적용한다.
- 주어 → 동사
- 23일의 주가 → 24일의 주가

순환 신경망의 구조

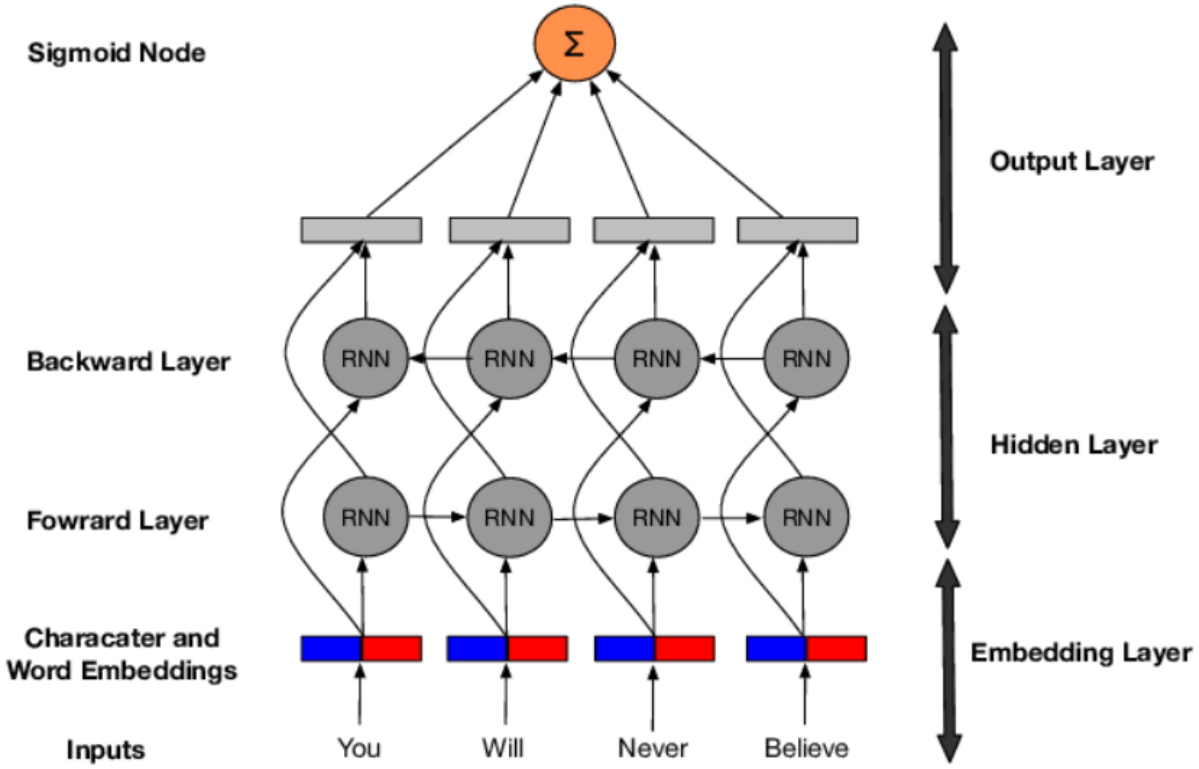
기능

위에서 나타난 sequence data의 해석에 필요한 기능들이 필요하다.

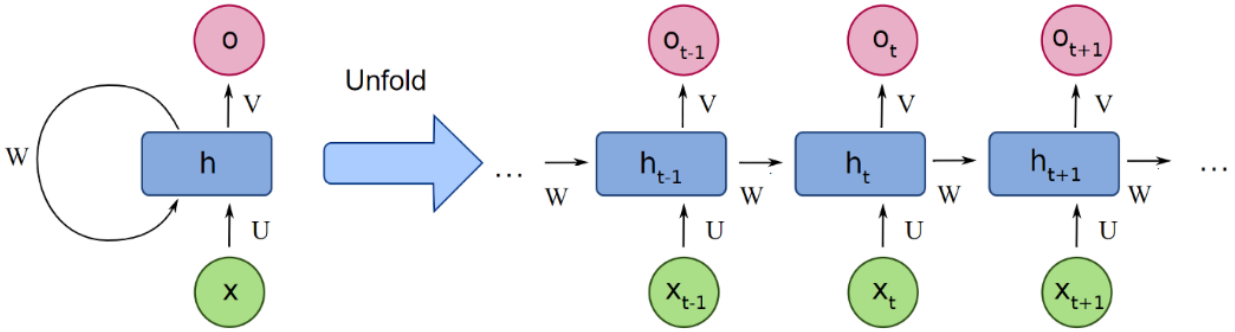
- 입력층, 은닉층, 출력층을 가진다.
- 순환 에지(recurrent edge)

구조

전체 구조



각요소별 구조



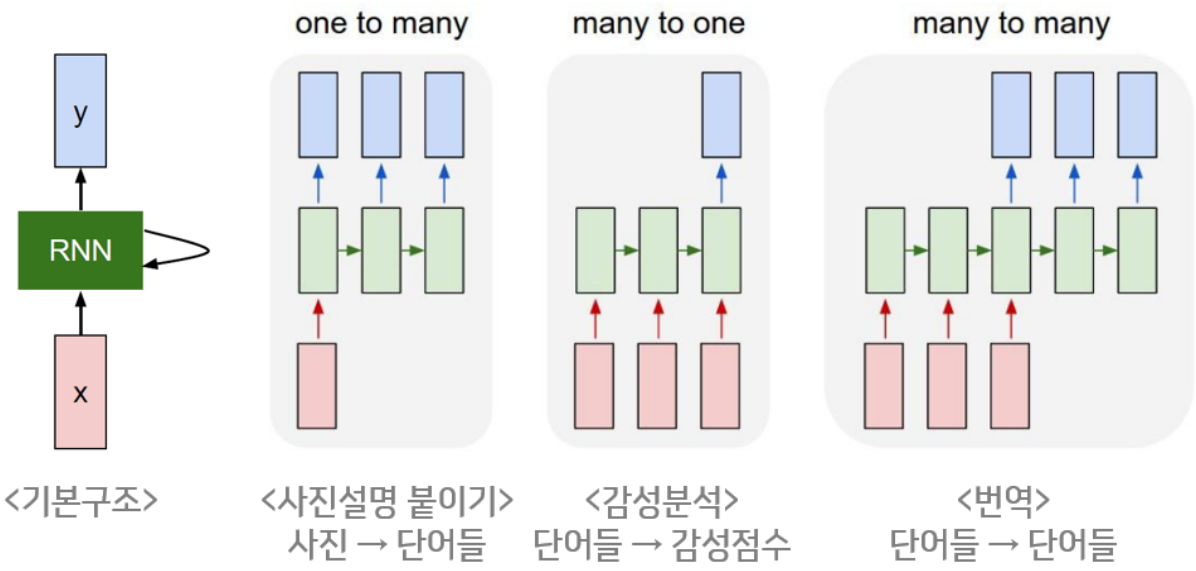
- x 입력크기 P
- y 출력 출력 Q
- h 은닉층의 값
- U 입력층과 은닉층 간의가중치 $P \times D$

- W은닉층에서 은닉층으로의 가중치 $D \times D$(칼만필터에서의 A행렬과 같다.)
- V 은닉층에서 출력층으로의 가중치 $D \times Q$

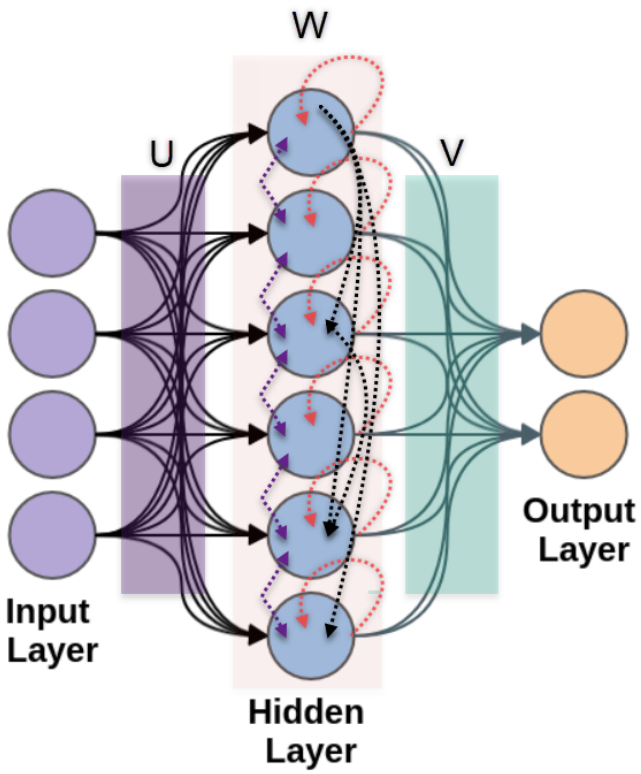
가중치 공유

- 매개변수는 서로 다른값을 이 아닌 같은 값을 공유한다. 즉 동일한 네트워크가 스퀀스의 각 요소에 적용되는 구조
    - 학습시 추정할 매개변수의 수가 줄어든다. (엄청)

RNN 의 종류

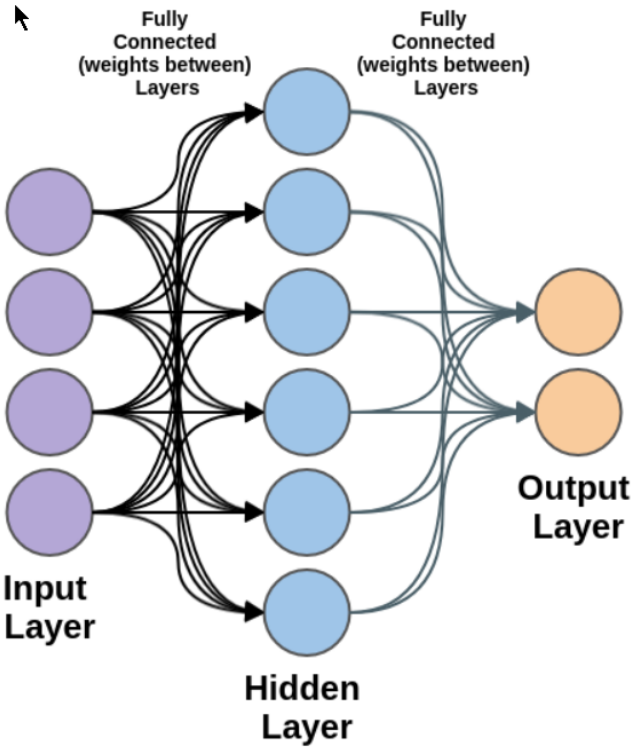


행렬값



- 가중치 U 의 크기는 입력 3개, 출력 6개 이므로 크기 6×3인 행렬
- 가중치 V 의 크기는 입력 6개, 출력 2개 이므로 크기 2×6인 행렬
- 은닉층에서 은닉층으로의 가중치 W 의 크기는 입력 6개, 출력 6개 이므로 크기 6×6인 행렬

FNN과의 차이점



- 위와 같은 Feed Forward Network 인 경우 은닉층에서 은닉층으로의 가중치가 필요없다.

수학적 표현

Recurrent

- Recurrence
  - $s^{(t)} = f(s^{(t-1)}; \theta)$
    - s : State
    - t : Time
    - t에서의 s를 알기 위해 t-1의 s를 이용해 알아내는 것을 의미 한다.
-

Vector 연산을 위한 Notation

notation

- RNN 은 현재 시점 t의 은닉층 값 h_t을 결정할 때, 직전 시점 $t-1$의 은닉층 값 h_{t-1}과 현재 시점의 입력값 x_t를 사용한다.
  - 마찬가지로 현재 시점 $t-1$의 은닉층 값 h을 결정할 때, 직전 시점 $t-2$의 은닉층 값 h_{t-2}과 현재 시점의 입력값 x_{t-1}를 사용한다.
- RNN 에서는 현재 시점의 값 h_t는 과거의 모든 입력과 은닉층의 값에 영향을 받는다.
- 위의 RNN의 구조를 Unfold 한 구조를 보면서 생각해보자.
- RNN은 매 시점 t마다 다음의 연산을 수행한다.
  - b_s, b_z : 편향, f : 활성화 함수
  - $s_t \ = \ Ux_t \ + \ Wh_{t-1} \ + \ b_s$
  - $h_t \ = \ f(s_t)$
  - $z_t \ = \ Vh_t \ + \ b_z$
  - $y_t \ = \ g(z_t)$
- 활성화 함수는 보통 $tanh$를 사용한다.(기울기 소실 문제 때문에)

notation2

$u_j \ = \ (u_{j1}, u_{j2}, \ \ldots, u_{jd})$ 는 U 행렬의 j번째 행 (h_j에 연결된 에지의 가중치들)

$$U = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1d} \\ u_{21} & u_{22} & \cdots & u_{2d} \\ \vdots & \vdots & \vdots & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pd} \end{pmatrix}$$

$$V = \begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1p} \\ v_{21} & v_{22} & \cdots & v_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ v_{q1} & v_{q2} & \cdots & v_{qp} \end{pmatrix}$$

$$W = \begin{pmatrix} w_{11} & w_{12} & \dots & w_{1d} \\ w_{21} & w_{22} & \dots & w_{2d} \\ \vdots & \vdots & \vdots & \vdots \\ w_{d1} & w_{d2} & \dots & w_{dd} \end{pmatrix}$$

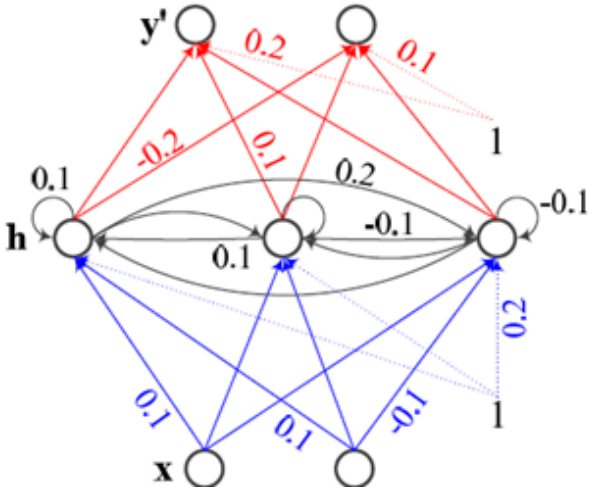
예제

1. 입력

$$U = \begin{pmatrix} 0.1 & 0.1 \\ 0.0 & 0.0 \\ 0.0 & -0.1 \end{pmatrix}, W = \begin{pmatrix} 0.1 & 0.1 & 0.0 \\ 0.0 & 0.0 & 0.0 \\ 0.2 & -0.1 & -0.1 \end{pmatrix}$$

$$V = \begin{pmatrix} 0.0 & 0.1 & 0.0 \\ -0.2 & 0.0 & 0.0 \end{pmatrix}, b = \begin{pmatrix} 0.0 \\ 0.0 \\ 0.2 \end{pmatrix}, c = \begin{pmatrix} 0.2 \\ 0.1 \end{pmatrix}$$

일때 RNN 구조를 그려보시오.



2. RNN에 샘플

$$x = (\begin{pmatrix} 0.0 \\ 1.0 \end{pmatrix}, \begin{pmatrix} 0.0 \\ 0.1 \end{pmatrix}, \begin{pmatrix} 0.1 \\ -0.2 \end{pmatrix}, \begin{pmatrix} 0.5 \\ 0.0 \end{pmatrix})^T, y = (\begin{pmatrix} 0.7 \\ 0.4 \end{pmatrix}, \begin{pmatrix} 0.3 \\ 0.4 \end{pmatrix}, \begin{pmatrix} 0.2 \\ 0.4 \end{pmatrix}, \begin{pmatrix} 0.4 \\ 0.5 \end{pmatrix})^T$$

가 주어졌다고 가정 하면 다음과 같은 연산이 일어 난다.

$t=1$일 때, 식 (8.7)과 식 (8.8)에 값을 대입하면 다음과 같다. 활성함수로 tanh를 사용한다고 가정하였다. 은닉층의 초깃값 $\mathbf{h}^{(0)} = (0\ 0\ 0)^{\mathrm{T}}$라고 가정한다.

$$\mathbf{a}^{(1)} = \mathbf{W}\mathbf{h}^{(0)} + \mathbf{U}\mathbf{x}^{(1)} + \mathbf{b} = \begin{pmatrix} 0.1 & 0.1 & 0.0 \\ 0.0 & 0.0 & 0.0 \\ 0.2 & -0.1 & -0.1 \end{pmatrix}\begin{pmatrix} 0.0 \\ 0.0 \\ 0.0 \end{pmatrix} + \begin{pmatrix} 0.1 & 0.1 \\ 0.0 & 0.0 \\ 0.0 & -0.1 \end{pmatrix}\begin{pmatrix} 0.0 \\ 1.0 \end{pmatrix} + \begin{pmatrix} 0.0 \\ 0.0 \\ 0.2 \end{pmatrix} = \begin{pmatrix} 0.1 \\ 0.0 \\ 0.1 \end{pmatrix}$$

$$\mathbf{h}^{(1)} = \tau(\mathbf{a}^{(1)}) = \begin{pmatrix} 0.0997 \\ 0.0 \\ 0.0997 \end{pmatrix}$$

$$\mathbf{y}'^{(1)} = \mathrm{softmax}(\mathbf{V}\mathbf{h}^{(1)} + \mathbf{c}) = \mathrm{softmax}\left(\begin{pmatrix} 0.0 & 0.1 & 0.0 \\ -0.2 & 0.0 & 0.0 \end{pmatrix}\begin{pmatrix} 0.0997 \\ 0.0 \\ 0.0997 \end{pmatrix} + \begin{pmatrix} 0.2 \\ 0.1 \end{pmatrix}\right) = \begin{pmatrix} 0.5299 \\ 0.4701 \end{pmatrix}$$

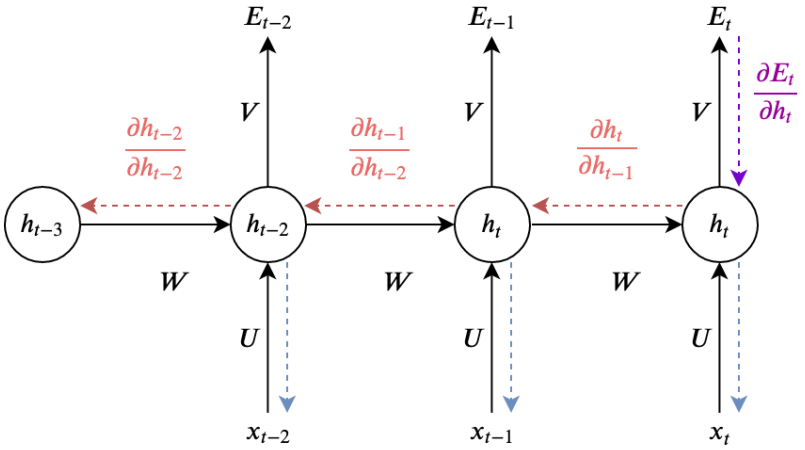
비슷한 방식으로 $t=2,3,4$일 때 계산 결과는 다음과 같다.

$$\mathbf{y}'^{(2)} = \begin{pmatrix} 0.5260 \\ 0.4740 \end{pmatrix},\ \mathbf{y}'^{(3)} = \begin{pmatrix} 0.5246 \\ 0.4754 \end{pmatrix},\ \mathbf{y}'^{(4)} = \begin{pmatrix} 0.5274 \\ 0.4726 \end{pmatrix}$$

이 샘플의 레이블, 즉 기대 출력이 $\mathbf{y} = \left(\begin{pmatrix} 0.7 \\ 0.4 \end{pmatrix}, \begin{pmatrix} 0.3 \\ 0.4 \end{pmatrix}, \begin{pmatrix} 0.2 \\ 0.4 \end{pmatrix}, \begin{pmatrix} 0.4 \\ 0.5 \end{pmatrix}\right)^{\mathrm{T}}$ 인데, 출력이 $\mathbf{y}' = \left(\begin{pmatrix} 0.5299 \\ 0.4701 \end{pmatrix}, \begin{pmatrix} 0.5260 \\ 0.4740 \end{pmatrix}, \begin{pmatrix} 0.5246 \\ 0.4754 \end{pmatrix}, \begin{pmatrix} 0.5274 \\ 0.4726 \end{pmatrix}\right)^{\mathrm{T}}$ 이므로 현재 가중치, 즉 매개변수 Θ는 상당한 오차를 발생시켰다고 판단할 수 있다. 8.2.3절에서는 매개변수 Θ의 값을 반복적으로 개선하여 최적해를 구하는 RNN의 학습 알고리즘을 학습한다.

BPTT(Back Propagation Throught Time))

- RNN 의 학습에서는 기본적으로 오차역전파 알고리즘을 사용한다.
- RNN에서의 오차 역전파 알고리즘은 현재 시점에서 과거 시점으로 시간을 거슬러 가며 오차 정보가 전달되어 간다.
- 따라서, BPTT(Back Propagation Throught Time) 알고리즘 이라고 부른다.
- 학습시에는 각 시점의 그레디언트를 구한 다음, 그 평균값을 해당 변수에 대한 그레디언트로 사용한다.



- 다음 그림은 시간 t에서 각 이전 시전으로 전파되는 오차 정보의 흐름을 나타낸다.
- $E(y_t, y\prime_t)$ 를 시점 *t*에서의 오차 라고 하자.
- $E(y, y) = \sum E_t(y_t, y\prime_t)$
- $\frac{\partial E}{\partial W} = \sum_i \frac{\partial E_t(y_t, y_t')}{\partial W}$
- $\frac{\partial E_t}{\partial W} = \frac{\partial E_t}{\partial y_t}\frac{\partial y_t}{\partial h_t}\frac{\partial h_t}{\partial W}$
- 이때, $h_t = f(Ux_t + Wh_{h-1} + b_s)$ 이다. 따라서 $\frac{\partial h_{t-1}}{\partial W}$ 는 다음과 같이 전개 된다.
- $\frac{\partial E_t}{\partial W} = \sum_{k=0}^{t} \frac{\partial E_t}{\partial y}\frac{\partial y_t}{\partial h_t}\frac{\partial h_t}{\partial h_k}\frac{\partial h_k}{\partial W} = \frac{\partial E_t}{\partial y}\frac{\partial y_t}{\partial h_t}[\frac{\partial h_t}{\partial h_k}\frac{\partial h_k}{\partial W} + \frac{\partial h_t}{\partial h_{k-1}}\frac{\partial h_{k-1}}{\partial W} + \frac{\partial h_t}{\partial h_{k-2}}\frac{\partial h_{k-2}}{\partial W} + \cdots + \frac{\partial h_t}{\partial h_0}\frac{\partial h_0}{\partial W}]$
- 즉, 현재 시점*t*부터 *t* = 0까지 모든 시점에서 오차 정보를 역전파시켜서 경사 하각법에 따라 가중치를 수정한다.
- RNN 의 가중치 *V*에 대한 *E*의 그레디언트 $\frac{\partial E_3}{\partial V}$는 다음과 같이 계산 될수 있다.
  - $\frac{\partial E_3}{\partial V} = \frac{\partial E_3}{\partial y_t}\frac{\partial y_t}{\partial V} = \frac{\partial E_3}{\partial y_t}\frac{\partial y_t}{\partial z_t}\frac{\partial z_t}{\partial V}$
- RNN의 가중치 *U*에 대한 *E*의 그레디언트 $\frac{\partial E_3}{\partial U}$는 다음과 같이 계산 될 수 있다.
  - $\frac{\partial E_t}{\partial U} = \sum_{k=0}^{t} \frac{\partial E_t}{\partial y}\frac{\partial y_t}{\partial h_t}\frac{\partial h_t}{\partial h_k}\frac{\partial h_k}{\partial U}$

2. RNN 의 문제점과 해결법