# Backpropagation in RNNs

dohye

July 2019

이 글은 공부를 위해 Gang Chen의 "A Gentel Tutorial of Recurrent Neural Network with Error Backpropagation" 논문을 일부 번역 및 정리한 내용을 바탕으로 한다. 본 논문에서는 RNN과 LSTM의 역전파에 대해 소개하고 있는데, 먼저 RNN의 역전파에 대한 내용을 정리해보도록 하며, 두개의 레이어를 가진 RNN 모델을 예시로 역전파를 통한 손실함수의 기울기를 직접 구해볼 것이다.

## 1 abstract

본 논문에서는 손글씨 인식, 음성인식 및 이미지와 텍스트와 같은 순차적인 작업(Sequential tasks)에 관심을 모으는 recurrent neural networks(RNNs)에 대해 설명할 것이다. RNNs는 feedback loop를 가지고 있으므로 역전파 단계를 이해하는 것이 조금 어렵다. 따라서 모델 매개변수와 관련하여 기울기를 계산하기 위한 오차 역전파와 같은 기본에 중점을 둔다.

## 2 Sequential data

순차적 데이터는 자연어 처리, 음성 인식 등 다양한 영역에서 일반적이며 시계열 데이터와 정렬 된 데이터 구조로 구분된다. 이 작업의 목표는 시퀀스 라벨링(sequence labeling)을 하는 것이다. Sequential data를 모델링하고 feedback loop를 통한 역전파에 대한 세부정보를 소개할 것이다.

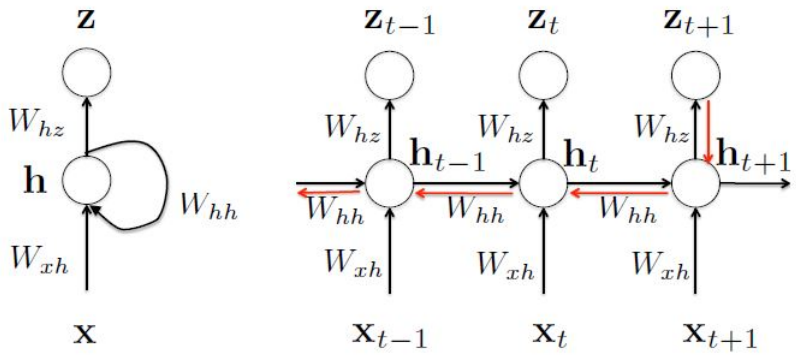


Figure 1: It is a RNN example: the left recursive description for RNNs, and the right is the corresponding extended RNN model in a time sequential manner.

# 3 Recurrent neural networks

RNN 모델은 동적시스템인데, $h_t$가 $x_t$(현재 관찰값)에만 의존하는 것이 아니라 이전 시점인 $h_{t-1}$에도 의존한다. $h_t$는 다음과 같이 표현할 수 있으며 여기서 $f$는 비선형함수를 사용한다.

$$h_t = f(h_{t-1}, x_t) \tag{1}$$

그러므로 $h_t$는 전체 시퀀스에 대한 정보를 포함한다. RNN은 hidden 변수를 메모리로 사용하여 시퀀스의 장기정보를 가지고 있을 수 있다.

다음과 같은 RNN 모델을 가정하자,

$$h_t = tanh(W_{hh}h_{t-1} + W_{xh}x_t + b_h) \tag{2}$$
$$z_t = softmax(W_{hz}h_t + b_z) \tag{3}$$

여기서 $z_t$는 time t에서의 예측값이고, $tanh(x)$는 다음과 같이 정의된다.

$$tanh(x) = \frac{sinh(x)}{cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{e^{2x} - 1}{e^{2x} + 1}$$

위의 RNN모델은 Fig.1에 묘사된것처럼 하나의 hidden layer를 가진다. 참고로 하나의 hidden case를 multiple layer로 확장하는 것은 쉽다. 확장할 때 각 시퀀스 데이터의 길이가 변하는 것을 고려하여, 각 순차 단계의 매개변수가 전체 시퀀스 분석에서 동일하다고 가정한다. 그렇지 않으면 기울기(gradient)를 계산하기가 어려울 것이다. 또한 가중치를 어떤 시퀀스 길이(Sequential length)에 대해서도 모두 공유하는 것은 모델을 잘 일반화 할 수 있게 한다. 그리고 시퀀스 라벨링(Sequential labeling)에 대해서는 모델 매개변수를 추정하기 위해 MLE를 사용한다. 다시 말해, 목적함수인 음의 로그가능도 함수(the negative log likelihood)를 최소로 하는 학습방법을 사용한다. 여기서 $\mathcal{L}$를 단순히 목적함수로 정의하며, $\mathcal{L}(t+1)$은 t+1 타임 스텝에서의 아웃풋을 의미한다. $\mathcal{L}(t+1) = -y_{t+1} \log z_{t+1}$

$$\mathcal{L}(y, z) = -\sum_t y_t \log z_t \tag{4}$$

참고로, $\mathcal{L}(y_t, z_t) = -y_t \log z_t$ 이고 $\mathcal{L}(y, z) = -\sum_t \mathcal{L}(y_t, z_t) = -\sum_t y_t \log z_t$ 이다. 예를 들어, 단어를 예측할 때 보통 전체 시퀀스(문장)를 하나의 학습 데이터(샘플)로 생각하고, 총 error는 매 시간 스텝(단어)마다의 에러의 총합으로 구한다. [1]

여기서 $\alpha_t = W_{hz}h_t + b_z$라 두고, $z_t = softmax(\alpha_t)$라 하자. $\alpha_t$와 관련하여 미분을 취함으로써 다음을 얻을 수 있다.

$$\frac{\partial \mathcal{L}}{\partial \alpha_t} = -(y_t - z_t) \tag{5}$$

먼저 $W_{hz}$에 대한 기울기를 구한다. $W_{hz}$는 모든 시간 순서(time sequence)에서 공유되기(shared) 때문에 각 시간 스텝에서의 미분을 하고 다 더할 수 있다.

$$\frac{\partial \mathcal{L}}{\partial W_{hz}} = \sum_t \frac{\partial \mathcal{L}}{\partial z_t} \frac{\partial z_t}{\partial W_{hz}} \tag{6}$$

[1] https://aikorea.org/blog/rnn-tutorial-3/

비슷하게 bias $b_z$ 도 구할 수 있다.

$$\frac{\partial \mathcal{L}}{\partial b_z} = \sum_t \frac{\partial \mathcal{L}}{\partial z_t} \cdot \frac{\partial z_t}{\partial b_z} \tag{7}$$

이제 $W_{hh}$에 대한 기울기를 구해본다. 이 때 $t \to t+1$ 을 고려한다.

$$\frac{\partial \mathcal{L}(t+1)}{\partial W_{hh}} = \frac{\partial \mathcal{L}(t+1)}{\partial z_{t+1}} \frac{\partial z_{t+1}}{\partial h_{t+1}} \frac{\partial h_{t+1}}{\partial W_{hh}} \tag{8}$$

위의 식은 $t \to t+1$로 가는 단계만 생각한다. 그러나 $h_{t+1}$가 부분적으로 $h_t$에 의존하기 하기 때문에 아래와 같이 $h_t$에 대한 항을 추가한다. (다른 $W_{hh}$가 전체 시간 순서에 걸쳐 공유되는 것을 고려함)

$$\frac{\partial \mathcal{L}(t+1)}{\partial W_{hh}} = \frac{\partial \mathcal{L}(t+1)}{\partial z_{t+1}} \frac{\partial z_{t+1}}{\partial h_{t+1}} \frac{\partial h_{t+1}}{\partial h_t} \frac{\partial h_t}{\partial W_{hh}} \tag{9}$$

그러므로 BPTT를 이용해서 $W_{hh}$에 대한 기울기를 계산해보면 다음과 같으며,

$$\frac{\partial \mathcal{L}(t+1)}{\partial W_{hh}} = \sum_{k=1}^{t} \frac{\partial \mathcal{L}(t+1)}{\partial z_{t+1}} \frac{\partial z_{t+1}}{\partial h_{t+1}} \frac{\partial h_{t+1}}{\partial h_k} \frac{\partial h_k}{\partial W_{hh}} \tag{10}$$

전체 시간에 걸친 $W_{hh}$에 대한 기울기를 합산하면 다음과 같다.

$$\frac{\partial \mathcal{L}}{\partial W_{hh}} = \sum_t \sum_{k=1}^{t+1} \frac{\partial \mathcal{L}(t+1)}{\partial z_{t+1}} \frac{\partial z_{t+1}}{\partial h_{t+1}} \frac{\partial h_{t+1}}{\partial h_k} \frac{\partial h_k}{\partial W_{hh}} \tag{11}$$

이제 $W_{xh}$에 대한 기울기를 구해본다. 비슷하게 $t+1$ 시간 스텝을 고려한다.

$$\frac{\partial \mathcal{L}(t+1)}{\partial W_{xh}} = \frac{\partial \mathcal{L}(t+1)}{\partial h_{t+1}} \frac{\partial h_{t+1}}{\partial W_{xh}}$$

$h_t$와 $x_{t+1}$ 모두가 $h_{t+1}$에 영향을 주기 때문에, $h_t$에 대한 역전파도 필요하다. 따라서 다음을 얻을 수 있다.

$$\begin{aligned} \frac{\partial \mathcal{L}(t+1)}{\partial W_{xh}} &= \frac{\partial \mathcal{L}(t+1)}{\partial h_{t+1}} \frac{\partial h_{t+1}}{\partial W_{xh}} + \frac{\partial \mathcal{L}(t+1)}{\partial h_t} \frac{\partial h_t}{\partial W_{xh}} \\ &= \frac{\partial \mathcal{L}(t+1)}{\partial h_{t+1}} \frac{\partial h_{t+1}}{\partial W_{xh}} + \frac{\partial \mathcal{L}(t+1)}{\partial h_{t+1}} \frac{\partial h_{t+1}}{\partial h_t} \frac{\partial h_t}{\partial W_{xh}} \end{aligned} \tag{12}$$

이는 다음과 같이 표현할 수 있다.

$$\frac{\partial \mathcal{L}(t+1)}{\partial W_{xh}} = \sum_{k=1}^{t+1} \frac{\partial \mathcal{L}(t+1)}{\partial h_{t+1}} \frac{\partial h_{t+1}}{\partial h_k} \frac{\partial h_k}{\partial W_{xh}} \tag{13}$$

그리고 전체 시간에 걸친 $W_{xh}$에 대한 기울기를 합산하면 다음과 같다.

$$\frac{\partial \mathcal{L}}{\partial W_{xh}} = \sum_t \sum_{k=1}^{t+1} \frac{\partial \mathcal{L}(t+1)}{\partial z_{t+1}} \frac{\partial z_{t+1}}{\partial h_{t+1}} \frac{\partial h_{t+1}}{\partial h_k} \frac{\partial h_k}{\partial W_{xh}} \tag{14}$$

그러나 RNN에서는 gradient vanishing이나 exploding 문제가 발생한다. (14)번식 $\frac{\partial h_{t+1}}{\partial h_k}$은 시퀀스에 대한 행렬 곱을 나타낸다. 따라서 RNN은 긴 시퀀스에서 기울기를 역전파해야 하므로 이런 문제가 발생한다. 그래서 이런 약점을 고려하여 long short term memory(LSTM)가 제안되었다. RNN은 $tanh$ 함수를 사용하여 $x_t$, $h_{t-1}$, $h_t$와의 관계를 통합하는 반면에, LSTM은 메모리 단위와의 상관관계를 모델링한다.

# 4 Example

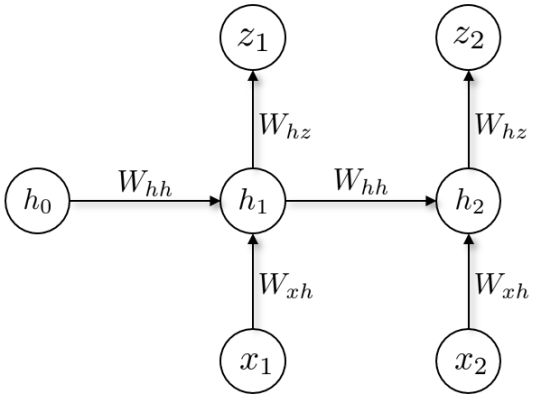


Figure 1: RNN model with 2 layer

위와 같이 $t = 2$인 RNN 모델을 가정하면 $h_1$, $\alpha_1$, $z_1$, $h_2$, $\alpha_2$, $z_2$ 와 목적함수 $\mathcal{L}$ 은 다음과 같다.

$$h_1 = tanh(W_{hh}h_0 + W_{xh}x_1 + b_h)$$
$$\alpha_1 = W_{hz}h_1 + b_z$$
$$z_1 = softmax(\alpha_1)$$

$$h_2 = tanh(W_{hh}h_1 + W_{xh}x_2 + b_h)$$
$$\alpha_2 = W_{hz}h_2 + b_z$$
$$z_2 = softmax(\alpha_2)$$

$$L = -\sum_{t=1}^{2} y_t log z_t$$

$$= -y_1 log z_1 - y_2 log z_2$$

$\mathcal{L}(t)$는 $t$번째 손실함수이며 $\mathcal{L}(t) = -y_t log z_t$ 로 정의된다.

**(1) $\mathcal{L}(1)$에 대한 기울기**

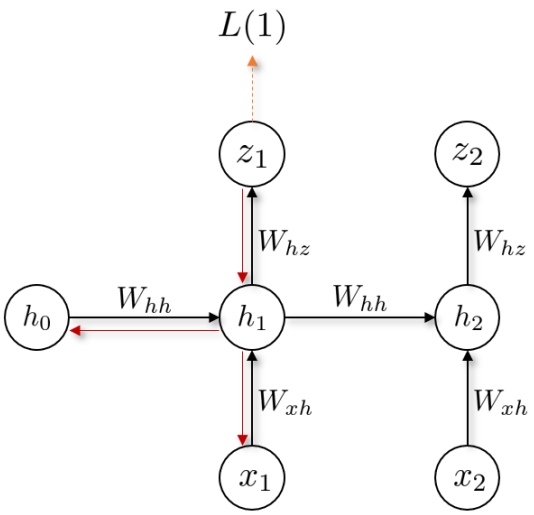


Figure 2: Backpropagation in L(1)

$$\frac{\partial \mathcal{L}(1)}{\partial \alpha_1} = -(y_1 - z_1) \tag{15}$$

$$\begin{aligned} \frac{\partial \mathcal{L}(1)}{\partial W_{hz}} &= \frac{\partial \mathcal{L}(1)}{\partial z_1} \frac{\partial z_1}{\partial \alpha_1} \frac{\partial \alpha_1}{\partial W_{hz}} \\ &= -(y_1 - z_1) h_1 \end{aligned} \tag{16}$$

$$\begin{aligned} \frac{\partial \mathcal{L}(1)}{\partial W_{hh}} &= \frac{\partial \mathcal{L}(1)}{\partial z_1} \frac{\partial z_1}{\partial \alpha_1} \frac{\partial \alpha_1}{\partial h_1} \frac{\partial h_1}{\partial W_{hh}} \\ &= -(y_1 - z_1) W_{hz} (1 - {h_1}^2) h_0 \end{aligned} \tag{17}$$

$$\begin{aligned} \frac{\partial \mathcal{L}(1)}{\partial W_{xh}} &= \frac{\partial \mathcal{L}(1)}{\partial z_1} \frac{\partial z_1}{\partial \alpha_1} \frac{\partial \alpha_1}{\partial h_1} \frac{\partial h_1}{\partial W_{xh}} \\ &= -(y_1 - z_1) W_{hz} (1 - {h_1}^2) x_1 \end{aligned} \tag{18}$$

**(2)** $\mathcal{L}(2)$**에 대한 기울기**

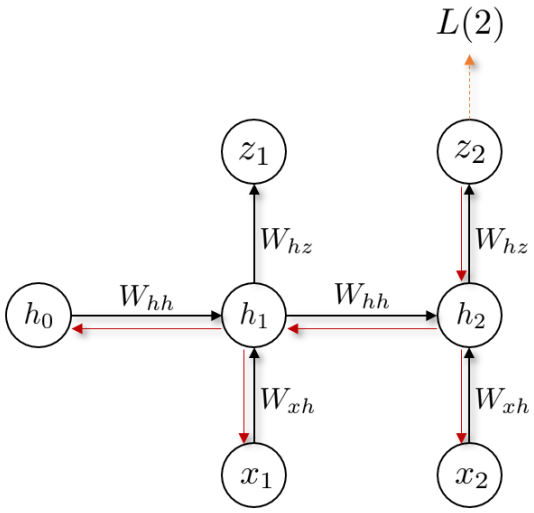


Figure 3: Backpropagation in L(2)

$$\frac{\partial \mathcal{L}(2)}{\partial \alpha_2} = -(y_2 - z_2) \tag{19}$$

$$\begin{aligned} \frac{\partial \mathcal{L}(2)}{\partial W_{hz}} &= \frac{\partial \mathcal{L}(2)}{\partial z_2} \frac{\partial z_2}{\partial \alpha_2} \frac{\partial \alpha_2}{\partial W_{hz}} \\ &= -(y_2 - z_2) h_2 \end{aligned} \tag{20}$$

$$\begin{aligned} \frac{\partial \mathcal{L}(2)}{\partial W_{hh}} &= \frac{\partial \mathcal{L}(2)}{\partial z_2} \frac{\partial z_2}{\partial \alpha_2} \frac{\partial \alpha_2}{\partial h_2} \frac{\partial h_2}{\partial W_{hh}} + \frac{\partial \mathcal{L}(2)}{\partial z_2} \frac{\partial z_2}{\partial \alpha_2} \frac{\partial \alpha_2}{\partial h_2} \frac{\partial h_2}{\partial h_1} \frac{\partial h_1}{\partial W_{hh}} \\ &= -(y_2 - z_2) W_{hz} (1 - {h_2}^2) h_1 - (y_2 - z_2) W_{hz} (1 - {h_2}^2) W_{hh} (1 - {h_1}^2) h_0 \end{aligned} \tag{21}$$

$$\begin{aligned} \frac{\partial \mathcal{L}(2)}{\partial W_{xh}} &= \frac{\partial \mathcal{L}(2)}{\partial z_2} \frac{\partial z_2}{\partial \alpha_2} \frac{\partial \alpha_2}{\partial h_2} \frac{\partial h_2}{\partial W_{xh}} + \frac{\partial \mathcal{L}(2)}{\partial z_2} \frac{\partial z_2}{\partial \alpha_2} \frac{\partial \alpha_2}{\partial h_2} \frac{\partial h_2}{\partial h_1} \frac{\partial h_1}{\partial W_{xh}} \\ &= -(y_2 - z_2) W_{hz} (1 - {h_2}^2) x_2 - (y_2 - z_2) W_{hz} (1 - {h_2}^2) W_{hh} (1 - {h_1}^2) x_1 \end{aligned} \tag{22}$$

**(3) $\mathcal{L}$에 대한 기울기**

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial W_{hz}} &= \sum_{t=1}^{2} \frac{\partial \mathcal{L}}{\partial z_t}\frac{\partial z_t}{\partial \alpha_t}\frac{\partial \alpha_t}{\partial W_{hz}} \\ &= \frac{\partial \mathcal{L}}{\partial z_1}\frac{\partial z_1}{\partial \alpha_1}\frac{\partial \alpha_1}{\partial W_{hz}} + \frac{\partial \mathcal{L}}{\partial z_2}\frac{\partial z_2}{\partial \alpha_2}\frac{\partial \alpha_2}{\partial W_{hz}} \\ &= -(y_1 - z_1)h_1 - (y_2 - z_2)h_2 \\ &= -\sum_{t=1}^{2}(y_t - z_t)h_t\end{aligned} \tag{23}$$

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial b_z} &= \sum_{t=1}^{2} \frac{\partial \mathcal{L}}{\partial z_t}\frac{\partial z_t}{\partial \alpha_t}\frac{\partial \alpha_t}{\partial b_z} \\ &= \frac{\partial \mathcal{L}}{\partial z_1}\frac{\partial z_1}{\partial \alpha_1}\frac{\partial \alpha_1}{\partial b_z} + \frac{\partial \mathcal{L}}{\partial z_2}\frac{\partial z_2}{\partial \alpha_2}\frac{\partial \alpha_2}{\partial b_z} \\ &= -(y_1 - z_1) - (y_2 - z_2) \\ &= -\sum_{t=1}^{2}(y_t - z_t)\end{aligned} \tag{24}$$

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial W_{hh}} &= \sum_{t=1}^{2}\sum_{k=1}^{t} \frac{\partial \mathcal{L}(t)}{\partial z_t}\frac{\partial z_t}{\partial \alpha_t}\frac{\partial \alpha_t}{\partial h_t}\frac{\partial h_t}{\partial h_k}\frac{\partial h_k}{\partial W_{hh}} \\ &= \frac{\partial \mathcal{L}(1)}{\partial z_1}\frac{\partial z_1}{\partial \alpha_1}\frac{\partial \alpha_1}{\partial h_1}\frac{\partial h_1}{\partial W_{hh}} + \frac{\partial \mathcal{L}(2)}{\partial z_2}\frac{\partial z_2}{\partial \alpha_2}\frac{\partial \alpha_2}{\partial h_2}\frac{\partial h_2}{\partial W_{hh}} + \frac{\partial \mathcal{L}(2)}{\partial z_2}\frac{\partial z_2}{\partial \alpha_2}\frac{\partial \alpha_2}{\partial h_2}\frac{\partial h_2}{\partial h_1}\frac{\partial h_1}{\partial W_{hh}} \\ &= -(y_1 - z_1)W_{hz}(1 - {h_1}^2)h_0 \\ &\quad - (y_2 - z_2)W_{hz}(1 - {h_2}^2)h_1 - (y_2 - z_2)W_{hz}(1 - {h_2}^2)W_{hh}(1 - {h_1}^2)h_0\end{aligned} \tag{25}$$

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial W_{xh}} &= \sum_{t=1}^{2}\sum_{k=1}^{t} \frac{\partial \mathcal{L}(t)}{\partial z_t}\frac{\partial z_t}{\partial \alpha_t}\frac{\partial \alpha_t}{\partial h_t}\frac{\partial h_t}{\partial h_k}\frac{\partial h_k}{\partial W_{xh}} \\ &= \frac{\partial \mathcal{L}(1)}{\partial z_1}\frac{\partial z_1}{\partial \alpha_1}\frac{\partial \alpha_1}{\partial h_1}\frac{\partial h_1}{\partial W_{xh}} + \frac{\partial \mathcal{L}(2)}{\partial z_2}\frac{\partial z_2}{\partial \alpha_2}\frac{\partial \alpha_2}{\partial h_2}\frac{\partial h_2}{\partial W_{xh}} + \frac{\partial \mathcal{L}(2)}{\partial z_2}\frac{\partial z_2}{\partial \alpha_2}\frac{\partial \alpha_2}{\partial h_2}\frac{\partial h_2}{\partial h_1}\frac{\partial h_1}{\partial W_{xh}} \\ &= -(y_1 - z_1)W_{hz}(1 - {h_1}^2)x_1 \\ &\quad - (y_2 - z_2)W_{hz}(1 - {h_2}^2)x_2 - (y_2 - z_2)W_{hz}(1 - {h_2}^2)W_{hh}(1 - {h_1}^2)x_1\end{aligned} \tag{26}$$

# 5 Appendix

$y = tanh(x)$의 미분

$$y = tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

$$\left\{ \frac{f(x)}{g(x)} \right\}' = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}$$

$$\frac{\partial e^x}{\partial x} = e^x, \quad \frac{\partial e^{-x}}{\partial x} = -e^{-x}$$

이므로 $y = tanh(x)$를 $x$에 대해 미분하면 다음과 같음을 보일 수 있다.

$$\begin{aligned} \frac{\partial tanh(x)}{\partial x} &= \frac{(e^x + e^{-x})(e^x + e^{-x}) - (e^x - e^{-x})(e^x - e^{-x})}{(e^x + e^{-x})^2} \\ &= 1 - \frac{(e^x - e^{-x})(e^x - e^{-x})}{(e^x + e^{-x})^2} = 1 - \left[ \frac{(e^x - e^{-x})}{(e^x + e^{-x})} \right]^2 \\ &= 1 - tanh(x)^2 = 1 - y^2 \end{aligned}$$