GRU Backpropagation

박주찬

DICE Lab
School of Computer Science and Engineering
KOREATECH
green261535@gmail.com

Introduction

LSTM에서는 출력, 입력, 삭제 게이트라는 3개의 게이트가 존재했었다. 반면, GRU에서는 업데이트 게이트와 리셋 게이트 두 가지 게이트만이 존재한다. GRU는 LSTM보다 게이트 수가 하나 적어 학습 속도면에서는 빠르지만 성능면에서는 LSTM이 GRU보다 성능이 뛰어나다는 것으로 알려져 있다. 이 문서에서는 GRU의 게이트를 살펴보고 GRU의 weigth들이 어떻게 Update되는지 Backpropagation을 통해 살펴보도록 하겠다.

GRU(Gated Recurrent Unit)

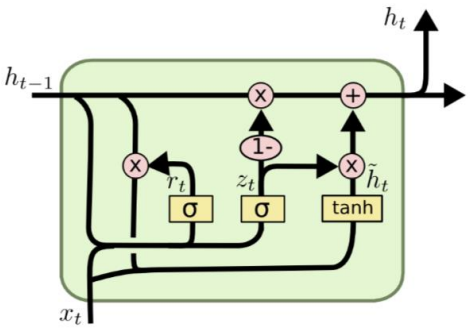

Fig. 1.

GRU의 기본 동작은 **Fig. 1**와 같다.

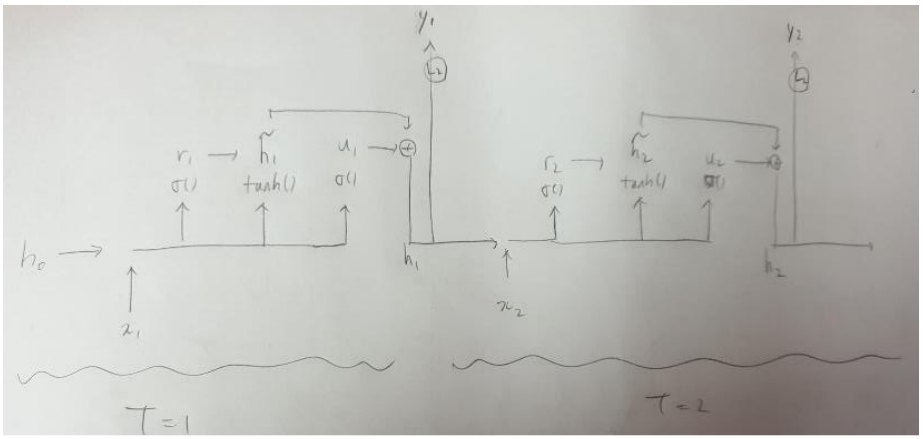

Fig. 2

두개의 GRU가 연속해서 이어져 있을 때를 살펴보면 **Fig. 2**와 같으며, 각 게이트의 식을 time step t로 일반화하면 아래와 같은 식으로 나타낼 수 있다. u는 update 게이트, r은 reset 게이트, h는 hidden State이다. 각 게이트와 State 안에 사용된 대문자 W와 U는 파라미터이다.
그러므로 $\frac{\partial \mathrm{L}}{\partial W_u}, \frac{\partial \mathrm{L}}{\partial U_u}, \frac{\partial \mathrm{L}}{\partial W_r}, \frac{\partial \mathrm{L}}{\partial U_r}, \frac{\partial \mathrm{L}}{\partial W_h}, \frac{\partial \mathrm{L}}{\partial U_h}$를 구하려고 한다.

$$u_t = \sigma(W_u \cdot h_{t-1} + U_u \cdot x_t + b_u)$$

$$r_t = \sigma(W_r \cdot h_{t-1} + U_r \cdot x_t + b_r)$$

$$\begin{aligned}\tilde{h}_t &= tanh(W_h \cdot (h_{t-1} \cdot r_t) + U_h \cdot x_t + b_h) \\ &= tanh(k_t \cdot r_t + U_h \cdot x_t + b_h), \qquad \{k_t = W_h \cdot h_{t-1}\}\end{aligned}$$

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

$$L_2 = \frac{1}{2}(h_t - y_t)^2$$

Fig. 2 그림에서 나와있듯이 T=2인 경우부터 시작해서 T=1인 경우 두가지만을 사용해서 Backpropagation을 해보겠다.

GRU, T=2)

T=2일 때, $\frac{\partial \mathrm{L}_2}{\partial W_u}, \frac{\partial \mathrm{L}_2}{\partial U_u}, \frac{\partial \mathrm{L}_2}{\partial W_r}, \frac{\partial \mathrm{L}_2}{\partial U_r}, \frac{\partial \mathrm{L}_2}{\partial W_h}, \frac{\partial \mathrm{L}_2}{\partial U_h}$를 구해보려고 한다. 체인 룰에 의하면 아래와 같이 식을 정리할 수 있다.

$$\frac{\partial \mathrm{L}_2}{\partial W_u} = \frac{\partial \mathrm{L}_2}{\partial h_2} \cdot \frac{\partial \mathrm{h}_2}{\partial u_2} \cdot \frac{\partial \mathrm{u}_2}{\partial W_u}$$
$$= (h_2 - y_2) \cdot \left(\tilde{h}_2 - h_1\right) \cdot \left(u_2 \cdot (1 - u_2)\right) \cdot h_1$$

$$\frac{\partial \mathrm{L}_2}{\partial U_u} = \frac{\partial \mathrm{L}_2}{\partial h_2} \cdot \frac{\partial \mathrm{h}_2}{\partial u_2} \cdot \frac{\partial \mathrm{u}_2}{\partial U_u}$$
$$= (h_2 - y_2) \cdot \left(\tilde{h}_2 - h_1\right) \cdot \left(u_2 \cdot (1 - u_2)\right) \cdot x_2$$

$$\frac{\partial \mathrm{L}_2}{\partial W_r} = \frac{\partial \mathrm{L}_2}{\partial h_2} \cdot \frac{\partial \mathrm{h}_2}{\partial \tilde{h}_2} \cdot \frac{\partial \tilde{h}_2}{\partial r_2} \cdot \frac{\partial r_2}{\partial W_r}$$
$$= (h_2 - y_2) \cdot u_2 \cdot \left(1 - \left(\tilde{h}_2\right)^2\right) \cdot (W_h \cdot h_1) \cdot \left(r_2 \cdot (1 - r_2)\right) \cdot h_1$$

$$\frac{\partial \mathrm{L}_2}{\partial U_r} = \frac{\partial \mathrm{L}_2}{\partial h_2} \cdot \frac{\partial \mathrm{h}_2}{\partial \tilde{h}_2} \cdot \frac{\partial \tilde{h}_2}{\partial r_2} \cdot \frac{\partial r_2}{\partial U_r}$$
$$= (h_2 - y_2) \cdot u_2 \cdot \left(1 - \left(\tilde{h}_2\right)^2\right) \cdot (W_h \cdot h_1) \cdot \left(r_2 \cdot (1 - r_2)\right) \cdot x_2$$

$$\frac{\partial \mathrm{L}_2}{\partial W_h} = \frac{\partial \mathrm{L}_2}{\partial h_2} \cdot \frac{\partial \mathrm{h}_2}{\partial \tilde{h}_2} \cdot \frac{\partial \tilde{h}_2}{\partial \mathrm{k}_2} \cdot \frac{\partial k_2}{\partial W_h}$$
$$= (h_2 - y_2) \cdot u_2 \cdot \left(1 - \left(\tilde{h}_2\right)^2\right) \cdot r_2 \cdot h_1$$

$$\frac{\partial \mathrm{L}_2}{\partial U_h} = \frac{\partial \mathrm{L}_2}{\partial h_2} \cdot \frac{\partial \mathrm{h}_2}{\partial \tilde{h}_2} \cdot \frac{\partial \tilde{h}_2}{\partial \mathrm{U}_h}$$
$$= (h_2 - y_2) \cdot u_2 \cdot \left(1 - \left(\tilde{h}_2\right)^2\right) \cdot x_2$$

GRU, T=1)

T=1일 때, L은 L_{Total} 이 되며, $\mathrm{L}_{Total} = \mathrm{L}_1 + \mathrm{L}_2$ 를 만족한다. T=2)에서 L_2 를 구했으므로 L_1만 구해서 L_{Total}를 구해주면 된다. 즉,

$\frac{\partial \mathrm{L}_{Total}}{\partial W_u}, \frac{\partial \mathrm{L}_{Total}}{\partial U_u}, \frac{\partial \mathrm{L}_{Total}}{\partial W_r}, \frac{\partial \mathrm{L}_{Total}}{\partial U_r}, \frac{\partial \mathrm{L}_{Total}}{\partial W_h}, \frac{\partial \mathrm{L}_{Total}}{\partial U_h}$ 를 구해보려고 한다. 이 중 $\frac{\partial \mathrm{L}_{Total}}{\partial U_u}$ 만 구하는 것만 정리하고, 나머지는 $\frac{\partial \mathrm{L}_{Total}}{\partial U_u}$를 구한 방식과 유사하므로 생략하도록 하겠다. 체인 룰에 의하면 아래와 같이 식을 정리할 수 있다.

$$\frac{\partial \mathrm{L}_{Total}}{\partial U_u} = \frac{\partial \mathrm{L}_1}{\partial U_u} + \frac{\partial \mathrm{L}_2}{\partial U_u}$$

$$\begin{aligned}\frac{\partial \mathrm{L}_1}{\partial U_u} &= \frac{\partial \mathrm{L}_1}{\partial h_1} \cdot \frac{\partial \mathrm{h}_1}{\partial u_1} \cdot \frac{\partial \mathrm{u}_1}{\partial U_u} \\ &= (h_1 - y_1) \cdot \left(\tilde{h}_2 - h_0\right) \cdot \left(u_1 \cdot (1 - u_1)\right) \cdot x_1\end{aligned}$$

$$\begin{aligned}\frac{\partial \mathrm{L}_2}{\partial U_u} &= \frac{\partial \mathrm{L}_2}{\partial h_2} \cdot \frac{\partial \mathrm{h}_2}{\partial u_2} \cdot \frac{\partial \mathrm{u}_2}{\partial h_1} \cdot \frac{\partial \mathrm{h}_1}{\partial u_1} \cdot \frac{\partial u_1}{\partial U_u} + \frac{\partial \mathrm{L}_2}{\partial h_2} \cdot \frac{\partial \mathrm{h}_2}{\partial h_1} \cdot \frac{\partial \mathrm{h}_1}{\partial u_1} \cdot \frac{\partial u_1}{\partial U_u} \\ &\quad + \frac{\partial \mathrm{L}_2}{\partial h_2} \cdot \frac{\partial \mathrm{h}_2}{\partial \tilde{h}_2} \cdot \frac{\partial \tilde{h}_2}{\partial \mathrm{r}_2} \cdot \frac{\partial \mathrm{h}_1}{\partial h_1} \cdot \frac{\partial \mathrm{h}_1}{\partial u_1} \cdot \frac{\partial u_1}{\partial U_u} + \frac{\partial \mathrm{L}_2}{\partial h_2} \cdot \frac{\partial \mathrm{h}_2}{\partial \tilde{h}_2} \cdot \frac{\partial \tilde{h}_2}{\partial \mathrm{k}_2} \cdot \frac{\partial \mathrm{h}_1}{\partial h_1} \cdot \frac{\partial \mathrm{h}_1}{\partial u_1} \cdot \frac{\partial u_1}{\partial U_u} \\ &= (h_2 - y_2) \cdot \left(\tilde{h}_2 - h_1\right) \cdot \left(u_1 \cdot (1 - u_1)\right) \cdot W_u \cdot \left(\tilde{h}_1 - h_0\right) \cdot \left(u_1 \cdot (1 - u_1)\right) \cdot x_1 \\ &\quad + (h_2 - y_2) \cdot (1 - u_2) \cdot \left(\tilde{h}_1 - h_0\right) \cdot \left(u_1 \cdot (1 - u_1)\right) \cdot x_1 \\ &\quad + (h_2 - y_2) \cdot u_2 \cdot \left(1 - \left(\tilde{h}_2\right)^2\right) \cdot (W_h \cdot h_1) \cdot \left(u_1 \cdot (1 - u_1)\right) \cdot x_1 \\ &\quad + (h_2 - y_2) \cdot u_2 \cdot \left(1 - \left(\tilde{h}_2\right)^2\right) \cdot r_2 \cdot W_h \cdot \left(u_1 \cdot (1 - u_1)\right) \cdot x_1\end{aligned}$$

$$\begin{aligned}\frac{\partial \mathrm{L}_{Total}}{\partial U_u} &= (h_1 - y_1) \cdot \left(\tilde{h}_2 - h_0\right) \cdot \left(u_1 \cdot (1 - u_1)\right) \cdot x_1 \\ &\quad + (h_2 - y_2) \cdot \left(\tilde{h}_2 - h_1\right) \cdot \left(u_1 \cdot (1 - u_1)\right) \cdot W_u \cdot \left(\tilde{h}_1 - h_0\right) \cdot \left(u_1 \cdot (1 - u_1)\right) \cdot x_1 \\ &\quad + (h_2 - y_2) \cdot (1 - u_2) \cdot \left(\tilde{h}_1 - h_0\right) \cdot \left(u_1 \cdot (1 - u_1)\right) \cdot x_1 \\ &\quad + (h_2 - y_2) \cdot u_2 \cdot \left(1 - \left(\tilde{h}_2\right)^2\right) \cdot (W_h \cdot h_1) \cdot \left(u_1 \cdot (1 - u_1)\right) \cdot x_1 \\ &\quad + (h_2 - y_2) \cdot u_2 \cdot \left(1 - \left(\tilde{h}_2\right)^2\right) \cdot r_2 \cdot W_h \cdot \left(u_1 \cdot (1 - u_1)\right) \cdot x_1\end{aligned}$$

Conclusion

GRU에서 각 게이트의 파라미터 들이 어떻게 update되는지 살펴보았다. LSTM에 비해 게이트 수가 하나 적은 만큼 성능이 비교적 떨어지는 것 같고, 일반적인 경우에는 LSTM을 사용한다.

References

1. https://medium.com/