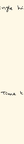
Recurrent Neural Networks (RNNs)

Single hidden layer RNN



Time t:

input: x_t

output: y_t

State or hidden layer: h_t

$$\tilde{h}_t = \phi(Ux_t + Wh_{t-1} + b_h)$$

$$y_t = \phi(Vh_t + b_y)$$

Sequential pattern:

$x = (x_1, x_2, \ldots, x_T)$

hidden state or state: h_t

$$\tilde{h}_t = f(x_t, h_t)$$

$$\therefore \quad \tilde{h}_t = f(x_t, \tilde{h}_t) = f(f(x_t, h_t), h_t)$$

$\Rightarrow$ $\tilde{h}_t$ is function of x_t, x_{t-1}, x_t

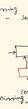
$\therefore$ output $\tilde{h}_t = \phi(V\tilde{h}_t + b_y)$

$= \mathcal{R}(x_t, x_{t-1}, \ldots, x_1)$

Unfolding a RNN:

RNNs for different types of tasks

1. No incoming inputs or outputs:
   - Language modeling
     - Predict w_t given $(w_1, \ldots, w_{t-1})$ or predict w_t or sentence.
2. Incoming outputs:
   - Sequential pattern classification



3. Incoming inputs: x_T
   - Image captioning

4. Incoming outputs followed by incoming inputs:
   - Sequence-to-sequence mapping

Encoder RNN: W_e, U_e, V_e

Decoder RNN: W_d, U_d, V_d

Multiple hidden layer RNN: Deep RNN

Two hidden layer RNN

Training a RNN: Back-propagation through time (BPTT)

Single hidden layer RNN

Sequential pattern classification task

- RNN with incoming outputs

Training data set:

for n sequences and their class labels

$$\mathcal{D} = \{ X_n, \bar{y}_n \}$$

n^{th} sequence $X_n = (x_{n1}, x_{n2}, \ldots, x_{nT})$

Desired output for n^{th} sequence $\bar{y}_n = \begin{cases} 1 & \text{if } n \in \text{class label of } X_n \\ 0 & \text{otherwise} \end{cases}$

Loss fn: Sum-of-squares error

Pattern mode: A training sequence is presented and the parameters are updated.

Hidden layer: h_t	state vector
Hidden layer: $\tilde{h}_t$	λ
Output layer: y_t	μ
Output layer: $\bar{y}_t$	λ

Batch mode: Error:

$$\bar{E} = \frac{1}{2} \sum_{n=1}^{N} (\bar{y}_n - y_n)^2$$

Forward mode:

Compute ΔV_{kh}, ΔW_{jh}, ΔU_{ih}

$$\Delta V_{kh} = \eta \delta_k^{\mu} h_h$$

$$\Delta W_{jh} = \eta \delta_j^{\lambda} h_{T-1}$$

$$\Delta U_{ih} = \eta \delta_i^{\lambda} x_i$$

Backward:

$$\Delta V_{kh} = \eta \delta_k^{\mu} h_{kh}$$

$$\delta_k^{\mu} = -\frac{\partial \bar{E}}{\partial y_k^{\mu}} \frac{\partial y_k^{\mu}}{\partial a_k^{\mu}} = \frac{\partial \bar{E}}{\partial y_k^{\mu}} V_{kh} \delta_h^{\lambda} \cdot \frac{\partial g(a_k^{\mu})}{\partial a_k^{\mu}}$$

$$\Delta W_{jh}^{\lambda} = \eta \delta_j^{\lambda} h_{T-1}$$

$$\delta_j^{\lambda} = \sum_{m=1}^{T} W_{jh} \delta_h^{\lambda} \cdot \frac{\partial g(a_j^{\lambda})}{\partial a_j^{\lambda}}$$

$$\Delta W_{jh}^{\lambda} = \eta \sum_{t=1}^{T} \Delta W_{jh}^{\lambda}$$

$$\Rightarrow \Delta W_{jh}^{\lambda} = \boxed{\eta \sum_{t=1}^{T} \delta_j^{\lambda} h_{T-t}}$$

$$\Delta U_{ih} = \eta \sum_{t=1}^{T} \delta_i^{\lambda} x_{i,t}$$

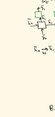
Missing inputs:

Source outputs: $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_T$

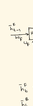
Error at time t: $\bar{E}_t = \frac{1}{2}(\bar{y}_{nt} - y_{nt})^2$

$$\Delta V_{kh} = \eta \sum_{t=1}^{T} \delta_k^{\mu} h_{kh}$$

$$\delta_k^{\mu} = (\bar{y}_{nt} - y_{nt}) \cdot \frac{\partial g(a_k^{\mu})}{\partial a_k^{\mu}}$$



Bidirectional RNN



$$h_t^F = \phi(Ux_t + Wh_{t-1}^F + b_h^F)$$

$$h_t^B = \phi(Ux_t + Wh_{t+1}^B + b_h^B)$$

$$\bar{y}_t = \phi(Vh_t^F + Vh_t^B + b_y)$$

$\bar{y}_t$ depends on the past inputs as well as on the future inputs.

Operation of BRNN:

Present a training sequence $x = (x_1, x_2, \ldots, x_T)$

TO:

$h_1^F, h_2^F, \ldots, h_T^F$ update h_t^F $t = 1, 2, \ldots, T$

$h_T^B, h_{T-1}^B, \ldots, h_1^B$ update h_t^B $t = T, T-1, \ldots, 1$

$\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_T$ update $\bar{y}_t$ $t = 1, 2, \ldots, T$

Perform gradient update

- Compute δ_k^{μ} and $\delta_h^{\lambda}, \delta_h^{\lambda}$
- Compute $\delta_j^{\lambda F}$ $t = T, T-1, \ldots, 1$ and $\delta_j^{\lambda B}$
- Compute $\delta_i^{\lambda F}$ $t = 1, 2, \ldots, T$ and $\delta_i^{\lambda B}$
- Compute ΔV_{kh} and ΔV_{hi}

Vanishing and Exploding Gradients Problem

Unidirectional RNN

Sequential pattern classification task

$$\bar{E} = \frac{1}{2} (\bar{y} - \bar{y})^2 (\bar{y} - \bar{y})$$

$$\Delta W = -\eta \frac{\partial \bar{E}}{\partial W}$$

(Forward pass)

$$\frac{\partial \bar{E}}{\partial W} = \frac{\partial \bar{E}}{\partial \bar{y}} \frac{\partial \bar{y}}{\partial W}$$

(Chain rule)

$$\frac{\partial \bar{y}}{\partial W}$$

$T = 4$

$$\frac{\partial \bar{y}_4}{\partial W} = \underbrace{\frac{\partial \bar{y}_4}{\partial y_4}}_{\text{Repeat}} \cdot \underbrace{\frac{\partial y_4}{\partial h_4}}_{\text{Repeat}} \cdot \frac{\partial h_4}{\partial W}$$

$$\frac{\partial h_4}{\partial W} = \frac{\partial h_4}{\partial h_3} \cdot \frac{\partial h_3}{\partial W}$$

$$\therefore \quad \frac{\partial \bar{y}_4}{\partial W} = \frac{\partial \bar{y}_4}{\partial y_4} \cdot \frac{\partial y_4}{\partial h_4} \cdot \frac{\partial h_4}{\partial h_3} \cdot \frac{\partial h_3}{\partial W}$$

$$\frac{\partial \bar{y}_4}{\partial W} = \frac{\partial \bar{y}_4}{\partial y_4} \cdot \frac{\partial y_4}{\partial h_4} \cdot \frac{\partial h_4}{\partial h_3} \cdot \frac{\partial h_3}{\partial h_2} \cdot \frac{\partial h_2}{\partial h_1} \cdot \frac{\partial h_1}{\partial W}$$

$$= \frac{\partial \bar{y}_4}{\partial y_4} \cdot \frac{\partial y_4}{\partial h_4} \cdot \frac{\partial h_4}{\partial h_3} \cdot \frac{\partial h_3}{\partial h_2} \cdot \frac{\partial h_2}{\partial h_1} \cdot \frac{\partial h_1}{\partial W}$$

$$\frac{\partial \bar{y}_T}{\partial W} = \prod_{t=2}^{T} \frac{\partial h_t}{\partial h_{t-1}} \cdot \frac{\partial h_1}{\partial W}$$

TOM $t = 2$

$$\frac{\partial h_t}{\partial h_{t-1}} = \frac{\partial \phi(a_t)}{\partial h_{t-1}} = \frac{\partial \phi(a_t)}{\partial a_t} \cdot \frac{\partial a_t}{\partial h_{t-1}}$$

$$a_t = \text{Wh}_{t-1} + \text{Ux}_t + b_h$$

$$\partial_t \left[\frac{\partial h_t}{\partial h_{t-1}} \right] = \text{diag}\left[\frac{\partial \phi(a_t)}{\partial a_t} \right] \cdot \frac{\partial a_t}{\partial h_{t-1}}$$

$$= \text{diag}\left[\phi'(a_t) \right] \cdot W$$

$$\left\| \frac{\partial h_t}{\partial h_{t-1}} \right\| = \left\| \text{diag}\left[\phi'(a_t) \right] \cdot W \right\|$$

Cauchy - Schwarz inequality for matrices:

$$\|AB\| \leq \|A\| \cdot \|B\|$$

$$\|\text{diag}[\phi'(a_t)]\| \leq \|\phi'(a_t)\|$$

$$\left\| \frac{\partial h_t}{\partial h_{t-1}} \right\| \leq \left\| \text{diag}\left[\phi'(a_t) \right] \right\| \cdot \|W\|$$

$$\frac{\partial \phi(a_t)}{\partial a_t} = \text{diag}\left[\phi'(a_t) \right]$$

Logistic $\phi(x) = \frac{1}{1+e^{-x}}$

$$\phi'(x) = \phi(x)(1-\phi(x))$$

$$\leq \frac{1}{4}$$

Tanh $\phi(x) = \frac{(1-\phi(x))^2}{4}$

P.S.: γ : Max value of $\phi'(x)$

$$\gamma = \begin{cases} \frac{1}{4} & \text{logistic} \\ 1 & \text{tanh} \end{cases}$$

$$\left\| \text{diag}\left[\phi'(a_t) \right] \right\| \leq \sqrt{\gamma}\, \theta\, \gamma$$

$$\|W\| \leq \theta$$

$$\left\| \frac{\partial h_t}{\partial h_{t-1}} \right\| \leq \sqrt{\gamma}\, \theta\, \gamma\, \lambda$$

$$\left\| \frac{\partial \bar{y}_T}{\partial h_1} \right\| = \left\| \prod_{t=2}^{T} \frac{\partial h_t}{\partial h_{t-1}} \right\|$$

$$\leq \left(\sqrt{\gamma}\, \theta\, \gamma\, \lambda \right)^{T-1}$$

If $\sqrt{\gamma}\, \theta\, \gamma\, \lambda < 1$ $\Rightarrow$ gradient vanishes

If $\sqrt{\gamma}\, \theta\, \gamma\, \lambda > 1$ $\Rightarrow$ gradient explodes