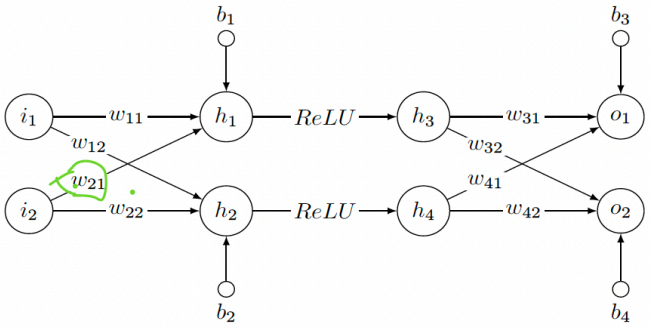
TUM

Part III: Backpropagation (9 points)

1. (9 points) Given the following neural network with fully connection layer and ReLU activations, including two input units (i_1, i_2), four hidden units (h_1, h_2) and (h_3, h_4). The output units are indicated as (o_1, o_2) and their targets are indicated as (t_1, t_2). The weights and bias of fully connected layer are called w and b with specific sub-descriptors.



The values of variables are given in the following table:

Variable	i_1	i_2	w_{11}	w_{12}	w_{21}	w_{22}	w_{31}	w_{32}	w_{41}	w_{42}	b_1	b_2	b_3	b_4	t_1	t_2
Value	2.0	-1.0	1.0	-0.5	0.5	-1.0	0.5	-1.0	-0.5	1.0	0.5	-0.5	-1.0	0.5	1.0	0.5

h_1

* $O_1 = \{(i_1 w_{11} + i_2 w_{21} + b_1)\ Relu \times w_{31}\}$ h_3 $+ \{(i_1 w_{12} + i_2 w_{22} + b_2)\ Relu \times w_{41}\} + b_3$

h_2 h_4

$= \{((2.1 + (-1).0.5 + 0.5)\ Relu \times 0.5) + (2.-0.5 + (-1.0)(-1.0) + (-0.5)\ Relu \times (0.5) + (-1.0)\}$

$= (2 - 0.5 + 0.5)\ Relu \times 0.5 + (-1 + 1 - 0.5)\ Relu \times -0.5 - 1$

$= 1 + 0 - 1 = 0$

$$O_2 = (i_1 w_{12} + i_2 w_{22} + b_2)\ \text{Relu} \times w_{42} + \text{Relu}$$
$$(i_1 w_{11} + i_2 w_{21} + b_1) \times w_{32} + b_4$$
$$= -1.5$$

$$L(O_1, O_2) = \frac{1}{2}(O_1 - t_1)^2 + \frac{1}{2}(O_2 - t_2)^2$$
$$= \frac{1}{2}(0 - 1)^2 + \frac{1}{2}(-1.5 - 0.5)^2$$
$$= \frac{1}{2}(1 + 4) = 2.5$$

* Update the weight w_{21} using gradient descent with learning rate 0.1.

$$\frac{\partial L}{\partial w_{21}} = \frac{\partial L}{\partial O_1} \cdot \frac{\partial O_1}{\partial h_3} \cdot \frac{\partial h_3}{\partial h_1} \cdot \frac{\partial h_1}{\partial w_{21}} + \frac{\partial L}{\partial O_2} \cdot \frac{\partial O_2}{\partial h_3} \cdot \frac{\partial h_3}{\partial h_1} \cdot \frac{\partial h_1}{\partial w_{21}}$$
$$= (O_1 - t_1) \cdot \frac{\partial}{\partial h_3}(h_3 w_{31} + h_4 w_{41} + b_3)$$

$$\cdot \frac{\partial}{\partial h_1}\left(\text{Relu}(h_1)\right) \cdot \frac{\partial}{\partial w_{21}}\left(i_1 w_{11} + i_2 w_{21} - b_1\right)$$

$$+ \quad (o_2 - t_2) \cdot \frac{\partial}{\partial h_3}\left(h_3 w_{32} + h_4 w_{42} + b_4\right)$$

$$\cdot \frac{\partial}{\partial h_1}\left(\text{Relu}(h_1)\right) \cdot \frac{\partial}{\partial w_{21}}\left(i_1 w_{11} + i_2 w_{21} + b_1\right)$$

$$= 0.5 \cdot w_{31} \cdot \text{Relu}'(h_1) \cdot i_2 + \quad (o_2 - t_2)$$

$$\cdot w_{32} \cdot \text{Relu}'(h_1) \cdot i_2$$

$$= -1 \times 0.5 \times 1 \cdot (-1) + (-2) \cdot (-1.0) \cdot 1 \cdot (-1.0)$$

$$= 0.5 - 2$$

$$= -1.5$$

$$w_{21}^{new} = w_{21}^{old} - \eta \nabla \frac{dL}{dw_{21}}$$

$$= 0.5 - 0.1 \cdot (-1.5)$$

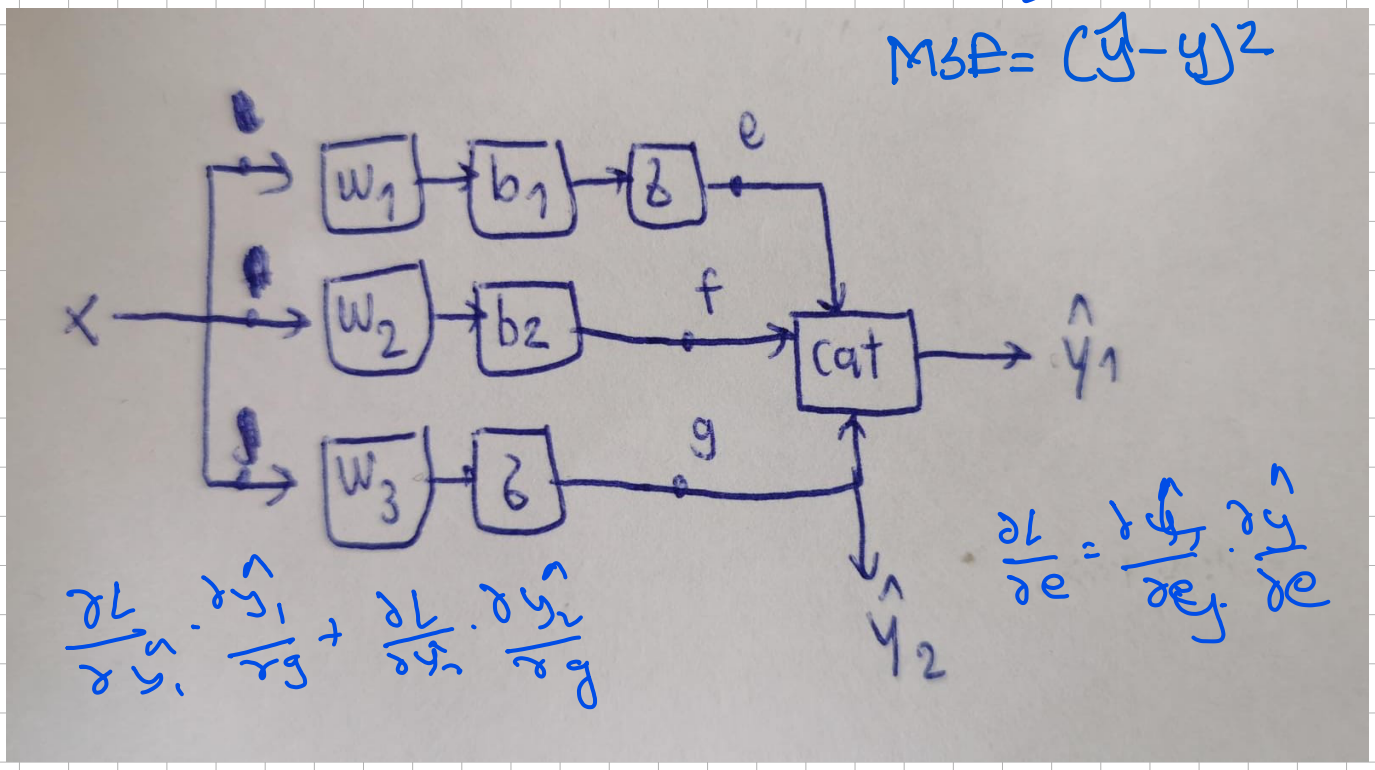
$$= \frac{5}{10} + \frac{15}{100}$$

$$= \frac{50 + 15}{100} = 0.65$$

55-23

$\frac{\partial L}{\partial \hat{y}_1} = \frac{\partial}{\partial \hat{y}_1}(\hat{y}_1 - y)^2 = 2(\hat{y}_1 - y)$

$\frac{\partial L}{\partial \hat{y}_2} = (\hat{y}_2 - y)^2 = 2(\hat{y}_2 - y)$



$e = (xw_1 + b_1)\,6$

$f = xw_2 + b_2$

$g = (xw_3)\,6$

$\hat{y}_1 = e + f + g$

$\hat{y}_2 = g$

$e = 6(w_1 x + b_1)$

$f = xw_2 + b_2$

$\hat{y}_2 = g = 6w_3 x$

$\hat{y}_1 = e + f + g$

$$\frac{\partial L}{\partial \hat{y}_1} = \frac{\partial}{\partial \hat{y}_1}(\hat{y}_1 - y)^2$$

$$= 2(\hat{y}_1 - y)$$

$$\frac{\partial L}{\partial \hat{y}_2} = \frac{\partial}{\partial \hat{y}_2}\left(\hat{y}_2 - y\right)^2$$

$$= 2\left(\hat{y}_2 - y\right)$$

$$\frac{\partial L}{\partial e} = \frac{\partial L}{\partial \hat{y}_1} \cdot \frac{\partial \hat{y}_1}{\partial e}$$

$$= \frac{\partial L}{\partial \hat{y}_1} \cdot \frac{\partial}{\partial e}\left(e + f + g\right)$$

$$= 2\left(\hat{y}_1 - y\right) \cdot 1$$

$$\frac{\partial L}{\partial f} = \frac{\partial L}{\partial \hat{y}_1} \cdot \frac{\partial \hat{y}_1}{\partial f}$$

$$= \frac{\partial L}{\partial \hat{y}_1} \cdot \frac{\partial}{\partial f}\left(e + f + g\right)$$

$$= 2\left(\hat{y}_1 - y\right) \cdot 1$$

$$\frac{\partial L}{\partial g} = \frac{\partial L}{\partial \hat{y}_1} \cdot \frac{\partial \hat{y}_1}{g} + \frac{\partial L}{\partial \hat{y}_2} \cdot \frac{\partial \hat{y}_2}{\partial g}$$

$$= \frac{\partial L}{\partial \hat{y}_1} \cdot \frac{\partial}{\partial g}\left(e + f + g\right) + \frac{\partial L}{\partial \hat{y}_2} \cdot \frac{\partial}{\partial g}\left(g\right)$$

$$= 2(\hat{y}_1 - y) \cdot 1 + 2(\hat{y}_2 - y) \cdot 1$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial c} \cdot \frac{\partial c}{\partial w_1}$$

$$= \frac{\partial L}{\partial c} \cdot \frac{\partial}{\partial w_1} \left(6 (x w_1 + b_1) \right)$$

$$= \frac{\partial L}{\partial c} \cdot 6'(x w_1 + b_1) \cdot x$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial f} \cdot \frac{\partial f}{\partial w_2}$$

$$= \frac{\partial L}{\partial f} \cdot \frac{\partial}{\partial w_2} \left(x w_2 + b_2 \right)$$

$$= \frac{\partial L}{\partial f} \cdot x$$

$$\frac{\partial L}{\partial w_3} = \frac{\partial L}{\partial g} \cdot \frac{\partial g}{\partial w_3}$$

$$= \frac{\partial L}{\partial g} \cdot \frac{\partial}{\partial w_3} \left((x w_3) 6 \right)$$

$$= \frac{\partial L}{\partial g} \cdot 6'(x w_3) \cdot x$$

Mock - 23 July 2021

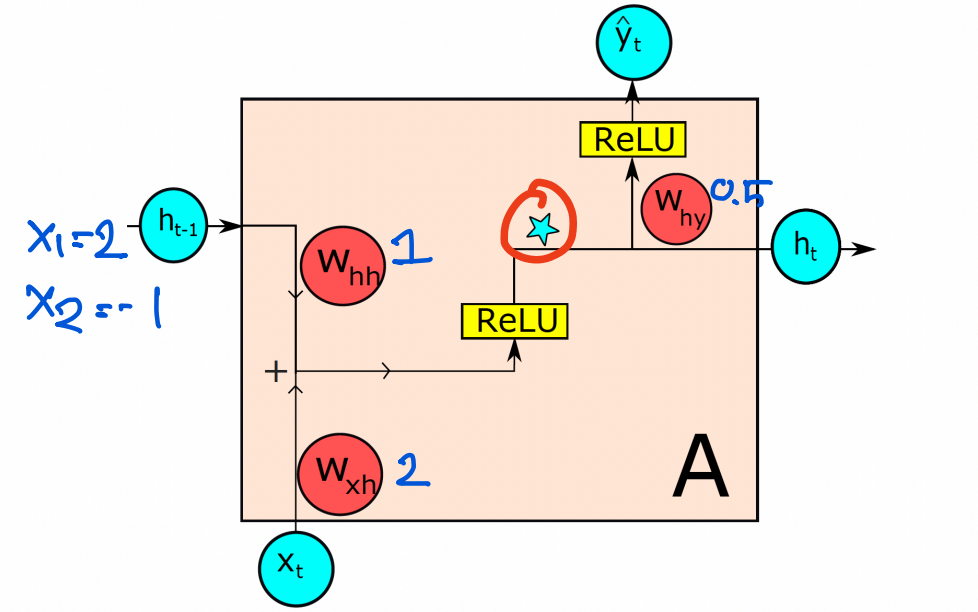


Figure 3: Schematic of a simplified Elman Cell.

* Calculate h_1, h_2, $\hat{y}_1$, $\hat{y}_2$ for $t=1$ and 2.

$$h_1 = \text{Relu}(h_{t-1} w_{hh} + w_{xh} x_t)$$

$t=1$

$$= \text{Relu}(0 \cdot w_{hh} + 2 \cdot 2)$$

$$= 4$$

$$h_2 = \text{Relu}(h_{t-1} w_{hh} + x_t w_{xh}) \quad t=2$$

$$= \text{Relu}(4 \cdot 1 + (-1) \cdot 2)$$

$$= \text{Relu}(2)$$

$$= 2$$

$$\hat{y}_1 = \text{Relu}(h_1 \times w_{hy})$$

$$= \text{Relu}(4 \times 0.5)$$

$$= 2$$

$$\hat{y}_2 = \text{Relu}(h_2 \times w_{hy})$$

$$= \text{Relu}(2 \times 0.5)$$

$$= 1$$

$$\frac{\partial L}{\partial h_t} = \frac{\partial L}{\partial \hat{y}_t} \cdot \frac{\partial \hat{y}_t}{\partial h_t} + \frac{\partial L}{\partial h_{t+1}} \cdot \frac{\partial h_{t+1}}{\partial h_t}$$

$$= 2(\hat{y}_t - y) \cdot \frac{\partial}{\partial h_t} \left(\text{Relu}(h_t \times w_{hy}) \right)$$

$$+ \frac{\partial L}{\partial h_{t+1}} \cdot \frac{\partial h_{t+1}}{\partial h_t}$$

$$= \frac{\partial L}{\partial \hat{y}_t} \cdot \text{Relu}'(w_{hy} \cdot h_t) \cdot w_{hy}$$

$$+ \frac{\partial L}{\partial h_{t+1}} \cdot \frac{\partial}{\partial h_t} \left(\text{Relu}(h_t \cdot w_{hh} + x_{t+1} \cdot w_{xh}) \right)$$

$$= \frac{\partial L}{\partial y_t} \cdot \text{Relu}'(w_{hy} \cdot h_t) \cdot w_{hy}$$

$$+ \frac{\partial L}{\partial h_{t+1}} \cdot \text{Relu}'\underbrace{(h_t \cdot w_{hh} + x_{t+1} \cdot w_{xh})}_{y_{t+1}} \cdot w_{hh}$$

$$\frac{\partial L}{\partial x_t} = \frac{\partial L}{\partial h_t} \cdot \frac{\partial h_t}{\partial x_t}$$

$$= \frac{\partial L}{\partial h_t} \cdot \frac{\partial}{\partial x_t} \left(\text{Relu}(h_{t-1} \cdot w_{hh} + x_t w_{xh}) \right)$$

$$= \frac{\partial L}{\partial h_t} \cdot \text{Relu}'(h_{t-1} \cdot w_{hh} + x_t w_{xh}) \cdot w_{xh}$$

$$\frac{\partial L}{\partial w_{xh_t}} = \frac{\partial L}{\partial h_t} \cdot \frac{\partial h_t}{\partial w_{xh_t}}$$

$$= \frac{\partial L}{\partial h_t} \cdot \frac{\partial}{\partial w_{xh_t}} \left(\text{Relu}(h_{t-1} \cdot w_{hh} + x_t w_{xh}) \right)$$

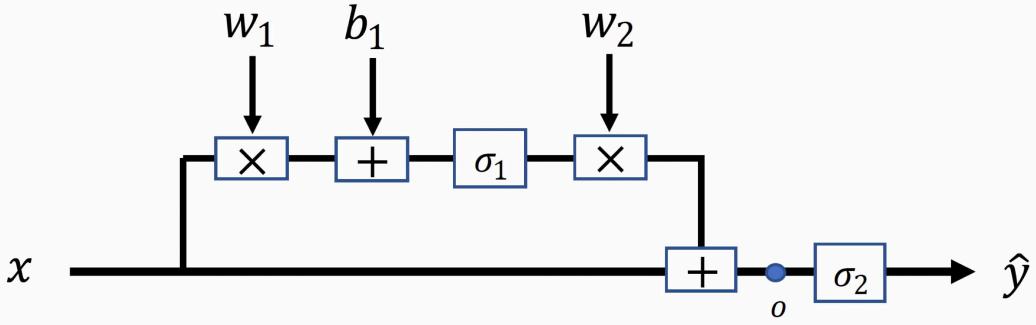
$$= \frac{\partial L}{\partial h_t} \cdot \text{Relu}'(h_{t-1} \cdot w_{hh} + x_t w_{xh}) \cdot x_t$$

$$\frac{\partial L}{\partial w_{hh_t}} = \frac{\partial L}{\partial h_t} \cdot \frac{\partial h_t}{\partial w_{hh_t}} = \frac{\partial L}{\partial h_t} \cdot \frac{\partial}{\partial w_{hh_t}} \left(\text{Relu}(h_{t-1} \cdot w_{hh} + x_t w_{xh}) \right)$$

$$= \frac{\partial L}{\partial h_t} \cdot \text{Relu}'(h_{t-1} \cdot w_{hh} + x_t \cdot w_{xh}) \cdot h_{t-1}$$

26 July 2021

Given is the following network $f(x) = \hat{y}$ receiving $x \in \mathbb{R}$ as input to compute a prediction $\hat{y} \in \mathbb{R}$. It uses two weights $w_1, w_2 \in \mathbb{R}$, one bias $b_1 \in \mathbb{R}$ and two sigmoid activations denoted as functions $\sigma_1(\cdot)$ and $\sigma_2(\cdot)$ shown in the figure below. The network is trained using the L_2 norm, defined as $L(y, \hat{y}) = \|\hat{y} - y\|_2^2$ with labels $y \in \{0, 1\}$. The boxes are mathematical operations, while $\times$ represents the multiplication, $+$ the addition and σ the sigmoid activation. $o \in \mathbb{R}$ marks an intermediate result as indicated in the figure.



$$f(x) = \hat{y} = \left((w_1 x + b_1)\,\sigma_1 \times w_2 + x\right)\sigma_2$$

$$= \left\{\left\{0.5 \cdot 1 + (-0.5)\right\}\sigma_1 \times (-2) + 1\right\}\sigma_2$$

$$= 0.5$$

$$L = (\hat{y} - y)^2$$

$$= (0.5 - 1)^2$$

$$= 0.25$$

$$\frac{\partial L}{\partial \hat{y}} = \frac{\partial}{\partial \hat{y}} (\hat{y} - y)^2$$

$$= 2(\hat{y} - y)$$

$$\frac{\partial L}{\partial o} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial o}$$

$$= \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial}{\partial o} \left(\sigma_2(o) \right)$$

$$= \frac{\partial L}{\partial \hat{y}} \cdot \sigma_2'(o)$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial o} \cdot \frac{\partial o}{\partial w_2}$$

$$= \frac{\partial L}{\partial o} \cdot \frac{\partial}{\partial w_2} \left(x + \{(w_1 x + b_1) \sigma_1 \times w_2\} \right)$$

$$= \frac{\partial L}{\partial o} \cdot \sigma_1 (w_1 x + b_1)$$

$$\frac{\partial L}{\partial b_1} = \frac{\partial L}{\partial o} \cdot \frac{\partial o}{\partial b_1}$$

$$= \frac{\partial L}{\partial o} \cdot \frac{\partial}{\partial b_1} \left(x + \{(w_1 x + b_1) \sigma_1 \times w_2\} \right)$$

$$= \frac{\partial L}{\partial o} \cdot w_2 \cdot \sigma_1' (x w_1 + b_1)$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial o} \cdot \frac{\partial o}{\partial w_1}$$

$$= \frac{\partial L}{\partial o} \cdot \frac{\partial}{\partial w_1} \left(x + f\{(w_1 x + b_1) G_1 \times w_2\} \right)$$

$$= \frac{\partial L}{\partial o} \cdot w_2 \cdot G_1'(w_1 x + b_1) \cdot x$$ ✓

$$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial o} \cdot \frac{\partial o}{\partial x}$$

$$= \frac{\partial L}{\partial o} \cdot \frac{\partial}{\partial x} \left(x + f\{(w_1 x + b_1) G_1 \times w_2\} \right)$$

$$= \frac{\partial L}{\partial o} \cdot \left(1 + w_2 \cdot G_1'(w_1 x + b_1) \cdot w_1 \right)$$ ✓

$$= \frac{\partial L}{\partial o} + \frac{\partial L}{\partial o} \cdot w_2 \cdot G_1'(w_1 x + b_1) w_1$$

24 June 2021

4 Backpropagation & Recurrent Neural Networks (9P)

Backpropagation is an important concept that enables training neural networks and it is also used in recurrent neural networks. Given is the recurrent cell visualized in Figure 2. In the following, the subscript (e.g., x_t) will denote the time step t.

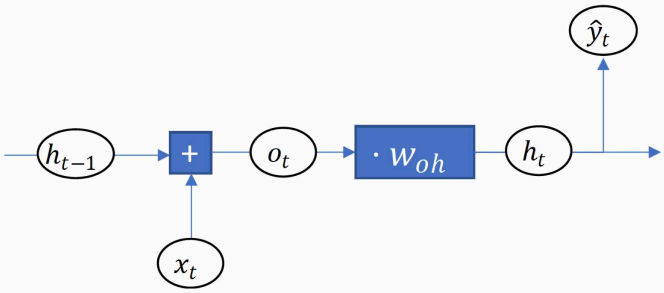


Figure 2: Schematic of a novel recurrent cell.

$$h_1 = (h_{t-1} + x_t)\, w_{oh}$$

$$= (0 + 1)\, 0.5$$

$$= 0.5$$

$$h_2 = (h_{t-1} + x_t)\, w_{oh}$$

$$= (0.5 + 0.5)\, 0.5$$

$$= 0.5$$

$$y_1 = h_t = 0.5$$

$$y_2 = h_t = 0.5$$

$$L = -y_t \ln(\hat{y}_t) - (1-y_t)\ln(1-\hat{y}_t)$$

$$\frac{\partial L}{\partial \hat{y}_t} = -\frac{y}{\hat{y}_t} + \frac{1-y}{1-\hat{y}_t}$$

$$\frac{\partial L}{\partial h_t} = \frac{\partial L}{\partial \hat{y}_t} \cdot \frac{\partial \hat{y}_t}{\partial h_t} + \frac{\partial L}{\partial h_{t+1}} \cdot \frac{\partial h_{t+1}}{\partial h_t}$$

$$= \frac{\partial L}{\partial \hat{y}_t} \cdot \frac{\partial}{\partial h_t}\left(h_t\right) + \frac{\partial L}{\partial h_{t+1}} \cdot \frac{\partial}{\partial h_t}\Big(\qquad (h_t + x_t) w_{oh} \Big)$$

$$= \frac{\partial L}{\partial \hat{y}_t} \cdot 1 + \frac{\partial L}{\partial h_{t+1}} \cdot w_{oh}$$

$$\frac{\partial L}{\partial o_t} = \frac{\partial L}{\partial h_t} \cdot \frac{\partial h_t}{\partial o_t}$$

$$= \frac{\partial L}{\partial h_t} \cdot \frac{\partial}{\partial o_t}\left(o_t \times w_{oh}\right)$$

$$= \frac{\partial L}{\partial h_t} \cdot w_{oh}$$

$$\frac{\partial L}{\partial w_{oh_t}} = \frac{\partial L}{\partial h_t} \cdot \frac{\partial h_t}{w_{oh_t}}$$

$$= \frac{\partial L}{\partial h_t} \cdot \frac{\partial}{\partial w_{oh_t}}\left(o_t \times w_{oh}\right)$$

$$= \frac{\partial L}{\partial h_t} \cdot o_t$$

$$\frac{\partial L}{\partial W_{on}} = \sum_t \frac{\partial L}{\partial W_{ont}}$$

29 March 2023

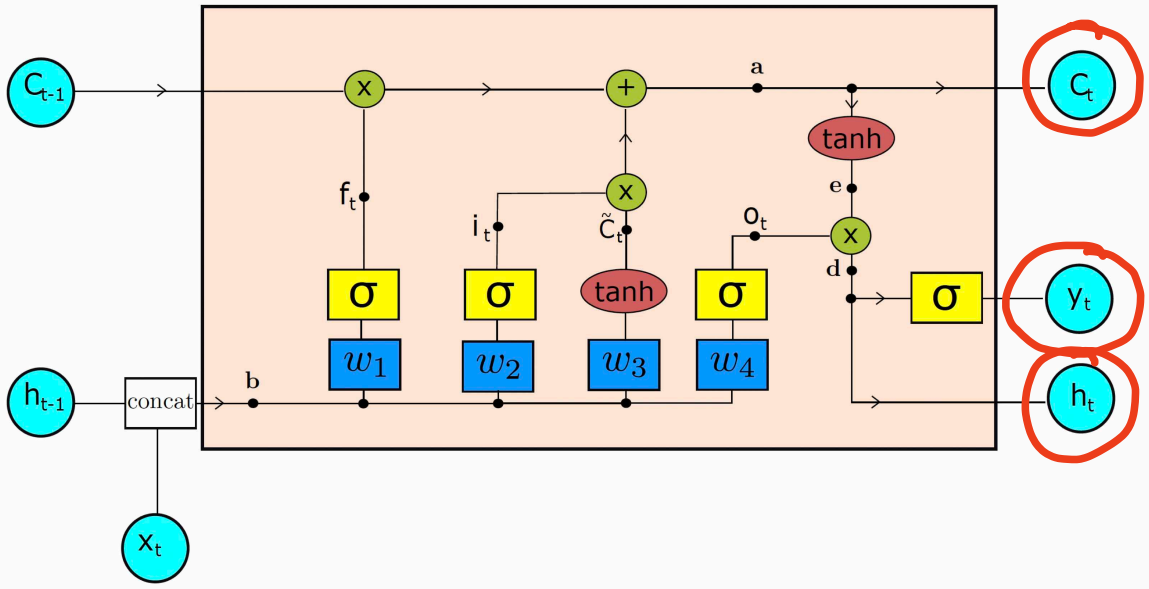


Figure 2: LSTM cell with intermediate steps.

C_t = long term memory

h_t = short term memory

$\hat{y}_t$ = output

$\tanh'(x) = 1 - \tanh^2(x)$

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

$$C_t = (C_{t-1} \cdot f_t) + (i_t \cdot \hat{C}_t)$$

$$Y_t = (\tanh(C_t) \cdot o_t) 6$$

$$h_t = \tanh(C_t) \cdot o_t$$

$$\frac{\partial L}{\partial C_{t-1}} = \frac{\partial L}{\partial C_t} \cdot \frac{\partial C_t}{\partial C_{t-1}}$$

$$= \frac{\partial L}{\partial C_t} \frac{\partial}{\partial C_{t-1}} \left((C_{t-1} \cdot f_t) + (i_t \times \hat{C}_t \right)$$

$$= \frac{\partial L}{\partial C_t} \cdot f_t$$

$$\frac{\partial L}{\partial d} = \frac{\partial L}{\partial y_t} \cdot \frac{\partial y_t}{\partial d} + \frac{\partial L}{\partial h_t} \cdot \frac{\partial h_t}{\partial d}$$

$$= \frac{\partial L}{\partial y_t} \cdot \frac{\partial}{\partial d}\left(\sigma(d)\right) + \frac{\partial L}{\partial h_t} \cdot \frac{\partial}{\partial d}(d)$$

$$= \frac{\partial L}{\partial y_t} \cdot \sigma'(d) + \frac{\partial L}{\partial h_t} \cdot 1$$

$$\frac{\partial L}{\partial e} = \frac{\partial L}{\partial d} \cdot \frac{\partial d}{\partial c}$$

$$= \frac{\partial L}{\partial d} \cdot \frac{\partial}{\partial e}\left(e \cdot o_t\right)$$

$$= \frac{\partial L}{\partial d} \cdot o_t$$

$$\frac{\partial L}{\partial o_t} = \frac{\partial L}{\partial d} \cdot \frac{\partial d}{\partial o_t}$$

$$= \frac{\partial L}{\partial d} \cdot \frac{\partial}{\partial o_t}\left(o_t \cdot e\right)$$

$$= \frac{\partial L}{\partial d} \cdot e$$

$$\frac{\partial L}{\partial a} = \frac{\partial L}{\partial e} \cdot \frac{\partial e}{\partial a} + \frac{\partial L}{\partial c_t}$$

$$= \frac{\partial L}{\partial e} \cdot \frac{\partial}{\partial a}\left(\tanh(a)\right) + \frac{\partial L}{\partial c_t}$$

$$= \frac{\partial L}{\partial e} \cdot \tanh'(a) + \frac{\partial L}{\partial c_t}$$

$$\frac{\partial L}{\partial \hat{C}_t} = \frac{\partial L}{\partial a} \cdot \frac{\partial a}{\partial \hat{C}_t}$$

$$= \frac{\partial L}{\partial a} \cdot \frac{\partial}{\partial \hat{C}_t} \left(C_{t-1} \cdot f_t + i_t \cdot \hat{C}_t \right)$$

$$= \frac{\partial L}{\partial a} \cdot i_t$$

$$\frac{\partial L}{\partial i_t} = \frac{\partial L}{\partial a} \cdot \frac{\partial a}{\partial i_t}$$

$$= \frac{\partial L}{\partial a} \cdot \frac{\partial}{\partial i_t} \left(C_{t-1} \cdot f_t + i_t \cdot \hat{C}_t \right)$$

$$= \frac{\partial L}{\partial a} \cdot \hat{C}_t$$

$$\frac{\partial L}{\partial f_t} = \frac{\partial L}{\partial a} \cdot \frac{\partial a}{\partial f_t} \left(C_{t-1} \cdot f_t + i_t \cdot \hat{C}_t \right)$$

$$= \frac{\partial L}{\partial a} \cdot C_{t-1}$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial f_t} \cdot \frac{\partial f_t}{\partial b} + \frac{\partial L}{\partial i_t} \cdot \frac{\partial i_t}{\partial b} + \frac{\partial L}{\partial \hat{C}_t} \cdot \frac{\partial \hat{C}_t}{\partial b}$$

$$+ \frac{\partial L}{\partial o_t} \cdot \frac{\partial o_t}{\partial b}$$

$$= \frac{\partial L}{\partial f_t} \cdot \frac{\partial}{\partial b}\left(\sigma(bw_1)\right) + \frac{\partial L}{\partial i_t} \cdot \frac{\partial}{\partial b}\left(\sigma(bw_2)\right)$$

$$+ \frac{\partial L}{\partial \tilde{c}_t} \cdot \frac{\partial}{\partial b}\left(\tanh(bw_3)\right) + \frac{\partial L}{\partial o_t} \cdot \frac{\partial}{\partial b}\left(\sigma(bw_4)\right)$$

$$= \frac{\partial L}{\partial f_t} \cdot \sigma'(bw_1) \cdot w_1 + \frac{\partial L}{\partial i_t} \cdot \sigma'(bw_2) \cdot w_2$$

$$+ \frac{\partial L}{\partial \tilde{c}_t} \cdot \tanh'(bw_3) \cdot w_3 + \frac{\partial L}{\partial o_t} \cdot \sigma'(bw_4) \cdot w_4$$

08 August 2022

4 Recurrent Networks and Backpropagation (9P)

Given is the following network which receives an input $x \in \mathbb{R}$ to predict two outputs $f_1(x) = \hat{y}_1$ and $f_2(x) = \hat{y}_2$ where $\hat{y}_{1,2} \in \mathbb{R}$. It only contains one weight $w \in \mathbb{R}$ which is multiplied with its input and a bias $b \in \mathbb{R}$ which is added to its input. At the end a sigmoid function σ is applied. The states k, l and m are highlighted in the figure as intermediate results. The + operation for the skip connection is a simple addition. The network is visualized in the following figure:

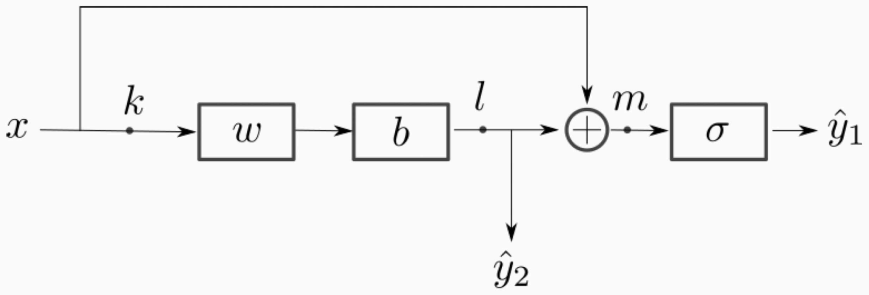


Figure 1: Schematic of a multi-headed residual network.

$l = wx + b$

$m = wx + b + x$

$f_1(x) = \hat{y}_1 = \sigma(wx + b + x)$

$f_2(x) = \hat{y}_2 = xw + b$

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

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

$$\frac{\partial L}{\partial m} = \frac{\partial L}{\partial \hat{y}_1} \cdot \frac{\partial \hat{y}_1}{\partial m}$$

$$= \frac{\partial L}{\partial \hat{y}_1} \cdot \frac{\partial}{\partial m}\left(\sigma(m)\right)$$

$$= \frac{\partial L}{\partial \hat{y}_1} \cdot \sigma'(m)$$

$$\frac{\partial L}{\partial l} = \frac{\partial L}{\partial m} \cdot \frac{\partial m}{\partial l} + \frac{\partial L}{\partial y_2} \cdot \frac{\partial y_2}{\partial l}$$

$$= \frac{\partial L}{\partial m} \cdot \frac{\partial}{\partial l}(l+x) + \frac{\partial L}{\partial y_2} \cdot \frac{\partial}{\partial l}(l)$$

$$= \frac{\partial L}{\partial m} \cdot 1 + \frac{\partial L}{\partial y_2} \cdot 1$$

$$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial l} \cdot \frac{\partial l}{\partial b}$$

$$= \frac{\partial L}{\partial l} \cdot \frac{\partial}{\partial b}(xw+b)$$

$$= \frac{\partial L}{\partial l} \cdot 1$$

$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial l} \cdot \frac{\partial l}{\partial w}$$

$$= \frac{\partial L}{\partial l} \cdot \frac{\partial}{\partial w}(xw+b)$$

$$= \frac{\partial L}{\partial l} \cdot x$$

$$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial l} \cdot \frac{\partial l}{\partial x}$$

$$= \frac{\partial L}{\partial l} \cdot \frac{\partial}{\partial x}(xw+b)$$

$$= \frac{\partial L}{\partial \ell} \cdot \omega$$

$$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial k} \cdot \frac{\partial k}{\partial x} + \frac{\partial L}{\partial m} \cdot \frac{\partial m}{\partial x}$$

[$m = x$ just for short memo ny]

$$= \frac{\partial L}{\partial k} \cdot \frac{\partial}{\partial x}(x) + \frac{\partial L}{\partial m} \cdot \frac{\partial}{\partial x}(x)$$

$$= \frac{\partial L}{\partial k} \cdot 1 + \frac{\partial L}{\partial m} \cdot 1$$

06 April 2022

4 Recurrent Networks and Backpropagation (14 P)

Given is the following modified recurrent network cell $f(x_t, h_{t-1}) = \hat{y}_t$. It receives an input $x_t \in \mathbb{R}$ and a hidden state $h_t \in \mathbb{R}$ to compute a prediction $\hat{y}_t$. It only contains one weight $w \in \mathbb{R}$ which is multiplied with its input. The states s_t and o_t are highlighted in the figure as intermediate results. The + operation is a simple addition. Using the label $y_t \in \{0, 1\}$ and the loss function $L(y_t, \hat{y}_t) = ||\hat{y}_t - y_t||_2^2$ one can compute the loss and backpropagate it through the network for an input sequence of length $T \in \mathbb{N}^+$ with $t \in [1, T]$ being the current time state. The network is visualized in the following figure:

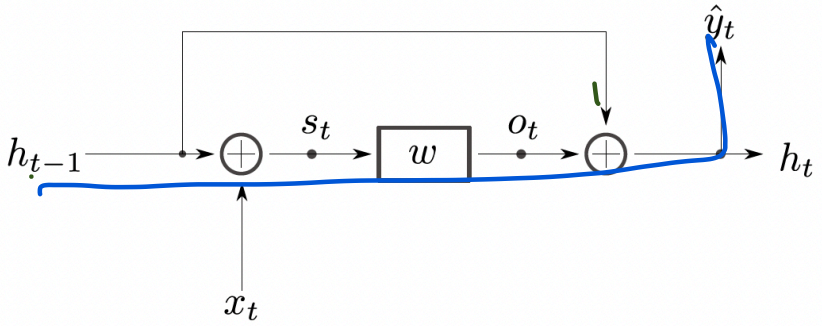


Figure 2: Schematic of a recurrent network cell.

$$s_t = h_{t-1} + x_t$$

$$o_t = (h_{t-1} + x_t)\, w$$

$$f = (h_{t-1} + x_t)\, w + h_{t-1}$$

$$\frac{\partial L}{\partial \hat{y}_t} = \frac{\partial}{\partial \hat{y}_t} (\hat{y}_t - y)^2$$

$$= 2(\hat{y}_t - y)$$

$$\frac{\partial L}{\partial o_t} = \frac{\partial L}{\partial \hat{y}_t} \cdot \frac{\partial \hat{y}_t}{\partial o_t} + \frac{\partial L}{\partial h_t}$$

$$= \frac{\partial L}{\partial \hat{y}_t} \cdot \frac{\partial}{\partial o_t} (o_t + h_{t-1}) + \frac{\partial L}{\partial h_t}$$

$$= \frac{\partial L}{\partial y_t} \cdot 1 + \frac{\partial L}{\partial h_t}$$

$$\frac{\partial L}{\partial w_t} = \frac{\partial L}{\partial o_t} \cdot \frac{\partial o_t}{\partial w_t}$$

$$= \frac{\partial L}{\partial o_t} \cdot \frac{\partial}{\partial w_t} \left(s_t \cdot w \right)$$

$$= \frac{\partial L}{\partial o_t} \cdot s_t$$

$$\frac{\partial L}{\partial w} = \sum \frac{\partial L}{\partial w_t}$$

$$\frac{\partial L}{\partial s_t} = \frac{\partial L}{\partial o_t} \cdot \frac{\partial o_t}{\partial s_t}$$

$$= \frac{\partial L}{\partial o_t} \cdot \frac{\partial}{\partial s_t} \left(s_t \cdot w \right)$$

$$= \frac{\partial L}{\partial o_t} \cdot w$$

$$\frac{\partial L}{\partial x_t} = \frac{\partial L}{\partial s_t} \cdot \frac{\partial s_t}{\partial x_t}$$

$$= \frac{\partial L}{\partial s_t} \cdot \frac{\partial}{\partial x_t} \left(h_{t-1} + x_t \right)$$

$$= \frac{\partial L}{\partial s_t} \cdot 1$$

$$\frac{\partial L}{\partial h_{t-1}} = \frac{\partial L}{\partial s_t} \cdot \frac{\partial s_t}{\partial h_{t-1}} + \frac{\partial L}{\partial h_t} \cdot \frac{\partial h_t}{\partial h_{t-1}} + \frac{\partial L}{\partial \hat{y}_t} \cdot \frac{\partial \hat{y}_t}{\partial h_{t-1}}$$

$$= \frac{\partial L}{\partial s_t} \cdot \frac{\partial}{\partial h_{t-1}} \left(h_{t-1} + x_t \right) + \frac{\partial L}{\partial h_t} \cdot \frac{\partial}{\partial h_{t-1}} \Big(\left(h_{t-1} + o_t \right) + \frac{\partial L}{\partial \hat{y}_t} \cdot \frac{\partial}{\partial h_{t-1}} \left(h_{t-1} + o_t \right)$$

$$= \frac{\partial L}{\partial s_t} \cdot 1 + \underbrace{\frac{\partial L}{\partial h_t} \cdot 1 + \frac{\partial L}{\partial \hat{y}_t} \cdot 1}_{\frac{\partial L}{\partial o_t}}$$

$$= \frac{\partial L}{\partial s_t} + \frac{\partial L}{\partial o_t}$$