TUTOR LESSON 3 - DEEP LEARNING (BPPT + VANISHING GRADIENT + GATED UNITS)

Exercise 1: Given the RNN described by the following equations

$$h^{(0)} = 0$$
$$a_h^{(t)} = Ux^{(t)} + Wh^{(t-1)} + b$$
$$h^{(t)} = \mathrm{ReLU}(a_h^{(t)})$$
$$a_o^{(t)} = Vh^{(t)} + c$$
$$o^{(t)} = \tanh(a_o^{(t)})$$
$$L = \sum_t e^{(t)} = \sum_t (y^{(t)} - o^{(t)})^2$$

INPUT SEQ.

$$\underset{t=1}{\begin{bmatrix}1\\-1\end{bmatrix}} \leftarrow \underset{t=2}{\begin{bmatrix}-1\\1\end{bmatrix}} \leftarrow \underset{t=3}{\begin{bmatrix}0\\1\end{bmatrix}}$$

TARGET SEQ.

$$\underset{t=1}{[1]} \leftarrow \underset{t=2}{[-1]} \leftarrow \underset{t=3}{[-1]}$$

$$U = \begin{bmatrix}1 & 1\\1 & -1\end{bmatrix}, \quad W = \begin{bmatrix}-1 & 1\\2 & 1\end{bmatrix}, \quad V = [1 \ \ -1],$$

$$b = \begin{bmatrix}1\\1\end{bmatrix}, \quad c = 3.$$

Compute $\frac{\partial L}{\partial W}$.

Solution:

$\boxed{\text{FORWARD PASS}}$

$\underline{t=1}$: $a_h^{(1)} = h^{(1)} = \begin{bmatrix}1\\3\end{bmatrix}$; $a_o^{(1)} = 1$; $o^{(1)} = \tanh(1) \simeq 0.76$

$\underline{t=2}$: $a_h^{(2)} = h^{(2)} = \begin{bmatrix}3\\4\end{bmatrix}$; $a_o^{(2)} = 2$; $o^{(2)} = \tanh(2) \simeq 0.96$

$\underline{t=3}$:

$$a_h^{(3)} = \begin{bmatrix}1 & 1\\1 & -1\end{bmatrix}\begin{bmatrix}0\\1\end{bmatrix} + \begin{bmatrix}-1 & 1\\2 & 1\end{bmatrix}\begin{bmatrix}3\\4\end{bmatrix} + \begin{bmatrix}1\\1\end{bmatrix} = \begin{bmatrix}1\\-1\end{bmatrix} + \begin{bmatrix}1\\10\end{bmatrix} + \begin{bmatrix}1\\1\end{bmatrix} = \begin{bmatrix}3\\10\end{bmatrix} = h^{(3)}$$

$$a_o^{(3)} = [1 \ \ -1]\begin{bmatrix}3\\10\end{bmatrix} + 3 = 3 - 10 + 3 = -4.$$

$$o^{(3)} = \tanh(-4) \simeq -0.99$$

$$\longrightarrow L = \sum_{t=1}^{3} e^{(t)} = (1-0.76)^2 + (1-0.96)^2 + (1+0.99)^2 \simeq 4.02$$

$\boxed{\text{BACKWARD PASS}}$

$$\frac{\partial L}{\partial W} = \frac{\partial \sum_t e^{(t)}}{\partial W} = \sum_t \frac{\partial e^{(t)}}{\partial W} = \sum_t \frac{\partial (y(t) - o(t))^2}{\partial W}$$

$$= \sum_t -2(y^{(t)} - o^{(t)}) \cdot \frac{\partial o^{(t)}}{\partial W} = \sum_t -2(y^{(t)} - o^{(t)}) \cdot \frac{\partial o^{(t)}}{\partial a_o(t)} \cdot \frac{\partial a_o^{(t)}}{\partial W}$$

$$= \sum_t \underbrace{-2(y^{(t)} - o^{(t)}) \cdot (1 - (o^{(t)})^2)}_{= c(t)} \cdot \frac{\partial a_o^{(t)}}{\partial h^{(t)}} \cdot \frac{\partial h^{(t)}}{\partial W} = \sum_t c(t) \cdot V \cdot \frac{\partial a_h(t)}{\partial W}$$

CRITICAL POINT

$\boxed{t=1}$

$$\frac{\partial a_h^{(1)}}{\partial W} = \frac{\partial (Ux^{(1)} + b)}{\partial W} = \underline{0}$$

$$\mathbb{R}^{2\times(2\times 2)}$$

$\boxed{t=2}$

$$\frac{\partial a_h^{(2)}}{\partial W} = \frac{\partial (Ux^{(2)} + Wh^{(1)} + b)}{\partial W} \in \mathbb{R}^{2\times(2\times 2)}$$

Let's break down the computation.

$$(1)\ \frac{\partial a_{h,1}^{(2)}}{\partial W} = \begin{bmatrix} h_1^{(1)} & h_2^{(1)} \\ 0 & 0 \end{bmatrix} ; \quad (2)\ \frac{\partial a_{h,2}^{(2)}}{\partial W} = \begin{bmatrix} 0 & 0 \\ h_1^{(1)} & h_2^{(1)} \end{bmatrix}$$

This can be easily seen from the fact that:

$$\begin{bmatrix} a_{h,1}^{(2)} \\ a_{h,2}^{(2)} \end{bmatrix} = \begin{bmatrix} u_{11} x_1^{(2)} + u_{12} x_2^{(2)} + W_{11} h_1^{(1)} + W_{12} h_2^{(1)} + b_1 \\ u_{21} x_1^{(2)} + u_{22} x_2^{(2)} + W_{21} h_1^{(1)} + W_{22} h_2^{(1)} + b_2 \end{bmatrix}$$

We multiply this tensor with V:

$$V \cdot \frac{\partial a_h^{(2)}}{\partial W} = \begin{bmatrix} v_1 & v_2 \end{bmatrix} \begin{bmatrix} \begin{bmatrix} h_1^{(1)} & h_2^{(1)} \\ 0 & 0 \end{bmatrix} \\ \begin{bmatrix} 0 & 0 \\ h_1^{(1)} & h_2^{(1)} \end{bmatrix} \end{bmatrix} = \begin{bmatrix} v_1 h_1^{(1)} & v_1 h_2^{(1)} \\ v_2 h_1^{(1)} & v_2 h_2^{(1)} \end{bmatrix}$$

$$= v_1 \odot \begin{bmatrix} h_1^{(1)} & h_2^{(1)} \\ 0 & 0 \end{bmatrix} + v_2 \odot \begin{bmatrix} 0 & 0 \\ h_1^{(1)} & h_2^{(1)} \end{bmatrix} \in \mathbb{R}^{2\times 2}$$

In our case: $V \cdot \frac{\partial a_h^{(2)}}{\partial W} = \begin{bmatrix} 1 & 3 \\ -1 & -3 \end{bmatrix}$

$\boxed{t=3}$

$h^{(3)} = Ux^{(3)} + Wh^{(2)} + b$ *Note: remember that $h^{(t)} \equiv a_h^{(t)} \quad \forall t$

$h^{(2)} = Ux^{(2)} + Wh^{(1)} + b$

$h^{(1)} = Ux^{(1)} + b$

$$\underset{\mathbb{R}^{2\times(2\times2)}}{\frac{\partial h^{(3)}}{\partial W}} = \underset{\mathbb{R}^{2\times2}}{\frac{\partial h^{(3)}}{\partial h^{(2)}}} \cdot \underset{\mathbb{R}^{2\times(2\times2)}}{\frac{\partial h^{(2)}}{\partial W}}$$

- Now: $\frac{\partial h^{(3)}}{\partial h^{(2)}} = W$

- $\frac{\partial h^{(2)}}{\partial W} \longrightarrow$ Already computed! (Same as $\frac{\partial a_h^{(2)}}{\partial W}$)

In the end:

$$\underset{\mathbb{R}^{2\times(2\times2)}}{\frac{\partial h^{(3)}}{\partial W}} = \underset{\mathbb{R}^{2\times2}}{\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}} \cdot \underset{\mathbb{R}^{2\times(2\times2)}}{\begin{bmatrix} \begin{bmatrix} h_1^{(1)} & h_2^{(1)} \\ 0 & 0 \end{bmatrix} \\ \begin{bmatrix} 0 & 0 \\ h_1^{(1)} & h_2^{(1)} \end{bmatrix} \end{bmatrix}} = \begin{bmatrix} \begin{bmatrix} W_{11} h_1^{(1)} & W_{11} h_2^{(1)} \\ W_{21} h_1^{(1)} & W_{21} h_2^{(1)} \end{bmatrix} \\ \begin{bmatrix} W_{12} h_1^{(1)} & W_{12} h_2^{(1)} \\ W_{22} h_1^{(1)} & W_{22} h_2^{(1)} \end{bmatrix} \end{bmatrix}$$

Also in this case we have the multiplication: $\underset{\mathbb{R}^{1\times2}}{V} \cdot \underset{\mathbb{R}^{2\times(2\times2)}}{\frac{\partial a_h^{(3)}}{\partial W}}$

$$= \begin{bmatrix} v_1 W_{11} h_1^{(1)} + v_2 W_{21} h_1^{(1)} & v_1 W_{11} h_2^{(1)} + v_2 W_{21} h_2^{(1)} \\ v_1 W_{12} h_1^{(1)} + v_2 W_{22} h_1^{(1)} & v_1 W_{12} h_2^{(1)} + v_2 W_{22} h_2^{(1)} \end{bmatrix} \in \mathbb{R}^{2\times2}$$

Which we can rewrite as:

$$(v_1 W_{11} + v_2 W_{21}) \odot \underset{=A}{\begin{bmatrix} h_1^{(1)} & h_2^{(1)} \\ 0 & 0 \end{bmatrix}} + (v_1 W_{12} + v_2 W_{22}) \odot \underset{=B}{\begin{bmatrix} 0 & 0 \\ h_1^{(1)} & h_2^{(1)} \end{bmatrix}}$$

In the end:

$$\underset{\mathbb{R}^{2\times2}}{\frac{\partial L}{\partial W}} = C(2) \cdot [v_1 \odot A + v_2 \odot B] + C(3) \cdot \{(v_1 W_{11} + v_2 W_{21}) \odot A + (v_1 W_{12} + v_2 W_{22}) \odot B\}$$

$$= -2 (y^{(2)} - o^{(2)}) (1 - (o^{(2)})^2) \cdot \begin{bmatrix} 1 & 3 \\ -1 & -3 \end{bmatrix} + (-2) \cdot (y^{(3)} - o^{(3)}) \cdot (1 - (o^{(3)})^2) \cdot \begin{bmatrix} -3 & -9 \\ 0 & 0 \end{bmatrix}$$

LEAKY INTEGRATOR UNIT

Integrates the input and has a leakage component:

$$h^{(t)} = (1-\alpha)\, h^{(t-1)} + \sigma\left(Ux^{(t)} + Wh^{(t-1)}\right)$$

where α is the LEAKY DECAY RATE.

These UNITS are used in RESERVOIR COMPUTING.

Reservoir is a network that maps the input signals into higher dimensional spaces through the dynamics of a fixed, non-linear dynamical system called RESERVOIR.

<u>Exercise</u>: A leaky integrator unit with RELU activation function and $\alpha = 0.5$ can be implemented by a GRU with...

LEAKY: $h^{(t)} = (1-\alpha)\, h^{(t-1)} + \mathrm{ReLU}\left(Ux^{(t)} + Wh^{(t-1)}\right)$

GRU: $h^{(t)} = (1-z^{(t)}) \odot h^{(t-1)} + z^{(t)} \odot f\left(Ux^{(t)} + W(r^{(t)} \odot h^{(t-1)})\right)$

ACTIVATION FUNCTION

In order for a GRU to implement a Leaky Integrator Unit,

- $f = \mathrm{ReLU}(\cdot)$
- $r^{(t)} = 1 \implies W(r^{(t)} \odot h^{(t-1)}) = Wh^{(t-1)}$
- $z^{(t)} = 0.5$, because $1-\alpha = 1 - z^{(t)}$.

With these values we have:

$$h^{(t)} = 0.5 \odot h^{(t-1)} + 0.5 \odot \mathrm{ReLU}\left(Ux^{(t)} + Wh^{(t-1)}\right)$$

this value should be $=1$ (like in leaky units)

$\implies$ To obtain 1, the input and hidden weights, U and W, should be MULTIPLIED by 2; The 2 is taken outside the ReLU function and when multiplied by 0.5 gives 1.