ECE/CS/ME 539 – Fall 2024 — Activity Solution 33

Solution

1. Forward Pass

Given:

$$x_1 = 0, \quad x_2 = 1.0, \quad y = 2, \quad h_0 = 0,$$

and parameters:

$$\begin{aligned} W_{\text{xz}} = 0.5, \quad W_{\text{hz}} = 0.1, \quad b_z = 0.0, \\ W_{\text{xr}} = 0.4, \quad W_{\text{hr}} = 0.2, \quad b_r = 0.0, \\ W_{\text{xh}} = 0.3, \quad W_{\text{hh}} = 0.3, \quad b_h = 0.0, \end{aligned}$$

Compute for $t = 1$

$$h_1 = 0$$

Compute for $t = 2$

$$h_2 = 0.11$$

Compute $\frac{\partial L}{\partial h_T}$:

$$\frac{\partial L}{\partial h_T} = h_T - y = 0.11 - 2 = -1.89.$$

Compute $\frac{\partial L}{\partial W_{\text{xz}}}$:

Given:

$$h_2 = z_2 h_1 + (1 - z_2)\tilde{h}_2$$

Since $h_1 = 0$, this simplifies to:

$$h_2 = (1 - z_2)\tilde{h}_2.$$

To compute $\frac{\partial L}{\partial W_{\text{xz}}}$, we use the chain rule:

$$\begin{aligned} \frac{\partial L}{\partial W_{\text{xz}}} &= \frac{\partial L}{\partial h_2} \cdot \frac{\partial h_2}{\partial W_{\text{xz}}} \\ &= -1.89 \cdot \frac{\partial h_2}{\partial W_{\text{xz}}}. \\ &= -1.89 \cdot \left(\frac{\partial (1 - z_2)}{\partial W_{\text{xz}}} \tilde{h}_2 + \frac{\partial \tilde{h}_2}{\partial W_{\text{xz}}} (1 - z_2) \right). \\ &= 1.89 \cdot \frac{\partial z_2}{\partial W_{\text{xz}}} \tilde{h}_2 \end{aligned}$$

$\frac{\partial \tilde{h}_2}{\partial W_{\text{xz}}} = 0$ and the details are provided in the last page.

Now, compute $\frac{\partial z_2}{\partial W_{\text{xz}}}$:

$$\frac{\partial z_2}{\partial W_{\text{xz}}} = z_2(1 - z_2)x_2,$$

where

$$z_2 = \sigma(W_{\text{xz}} x_2) = \sigma(0.5 \cdot 1) = \sigma(0.5) = 0.6225.$$

Thus:

$$\frac{\partial z_2}{\partial W_{\text{xz}}} = 0.6225 \cdot (1 - 0.6225) \cdot 1 = 0.235.$$

Next, compute $\tilde{h}_2$:

$$\tilde{h}_2 = \tanh(W_{\text{xh}} x_2 + W_{\text{hh}}(r_2 \odot h_1) + b_h).$$

Given $h_1 = 0$ and $r_2 = \sigma(W_{\text{xr}} x_2)$:

$$\tilde{h}_2 = \tanh(0.3 \cdot 1 + 0.3 \cdot 0 + 0) = \tanh(0.3) = 0.2913.$$

Finally, compute $\frac{\partial L}{\partial W_{\text{xz}}}$:

$$\frac{\partial L}{\partial W_{\text{xz}}} = -1.89 \cdot \big(- 0.235 \cdot 0.2913\big) = 0.1294.$$

Compute $\frac{\partial L}{\partial W_{\mathbf{xr}}}$

$$\frac{\partial L}{\partial W_{\text{xr}}} = \frac{\partial L}{\partial h_2} \frac{\partial h_2}{\partial \tilde{h}_2} \frac{\partial \tilde{h}_2}{\partial W_{\text{xr}}}$$

TheW_{xr} influences $\tilde{h}_2$ through h_1, but when $h_1 = 0$, this dependency is severed. As a result, the connection between W_{xr} and $\tilde{h}_2$ no longer exists, leading to

$$\frac{\partial \tilde{h}_2}{\partial W_{\text{xr}}} = 0.$$

So,

$$\frac{\partial L}{\partial W_{\text{xr}}} = 0.$$

Compute $\frac{\partial L}{\partial W_{\mathbf{xh}}}$

Given:

$$h_2 = (1 - z_2)\tilde{h}_2.$$

To compute $\frac{\partial L}{\partial W_{\text{xh}}}$, we use the chain rule:

$$\frac{\partial L}{\partial W_{\text{xh}}} = \frac{\partial L}{\partial h_2} \cdot \frac{\partial h_2}{\partial W_{\text{xh}}}.$$

Expanding $\frac{\partial h_2}{\partial W_{\text{xh}}}$:

$$\frac{\partial h_2}{\partial W_{\text{xh}}} = \frac{\partial}{\partial W_{\text{xh}}}\Big((1 - z_2)\tilde{h}_2\Big) = (1 - z_2) \cdot \frac{\partial \tilde{h}_2}{\partial W_{\text{xh}}} + \tilde{h}_2 \cdot \frac{\partial (1 - z_2)}{\partial W_{\text{xh}}}.$$

However, since z_2 does not depend on W_{xh} in this scenario (as $h_1 = 0$):

$$\frac{\partial (1 - z_2)}{\partial W_{\text{xh}}} = 0.$$

Thus:

$$\frac{\partial h_2}{\partial W_{\text{xh}}} = (1 - z_2) \cdot \frac{\partial \tilde{h}_2}{\partial W_{\text{xh}}}.$$

Now, compute $\frac{\partial \tilde{h}_2}{\partial W_{\text{xh}}}$:

$$\tilde{h}_2 = \tanh(W_{\text{xh}}x_2 + W_{\text{hh}}(r_2 \cdot h_1) + b_h) = \tanh(W_{\text{xh}}x_2 + b_h).$$

$$\frac{\partial \tilde{h}_2}{\partial W_{\text{xh}}} = (1 - \tilde{h}_2^2) \cdot x_2 = (1 - (0.2913)^2) \times 1 = 0.9151.$$

Compute $1 - z_2$:

$$z_2 = \sigma(W_{\text{xz}}x_2 + W_{\text{hz}}h_1 + b_z) = \sigma(0.5 \times 1) = \sigma(0.5) = 0.6225,$$

$$1 - z_2 = 0.3775.$$

Finally, compute $\frac{\partial h_2}{\partial W_{\text{xh}}}$:

$$\frac{\partial h_2}{\partial W_{\text{xh}}} = 0.3775 \times 0.9151 = 0.3454.$$

Now, apply the chain rule:

$$\frac{\partial L}{\partial W_{\text{xh}}} = \frac{\partial L}{\partial h_2} \cdot \frac{\partial h_2}{\partial W_{\text{xh}}} = -1.8901 \times 0.3454 = -0.6528.$$

3. Gradient Stability:

In simple RNNs, the hidden state is updated as:

$$h_t = \tanh(W_h h_{t-1} + W_x x_t + b),$$

which involves repeated multiplications of the hidden state by weights (W_h) across many timesteps. This can lead to vanishing or exploding gradients.

In GRUs, the hidden state update involves gated mechanisms:

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t,$$

where

$$\tilde{h}_t = \tanh(W_r(r_t \odot h_{t-1}) + W_x x_t).$$

The interpolation between the previous hidden state (h_{t-1}) and the candidate hidden state ($\tilde{h}_t$) reduces the risk of vanishing or exploding gradients.

depends on h_1

$$h_2 = z_2 \cdot h_1 + (1-z_2)\cdot \tilde{h}_2$$

depends on W_{xz} depends on z_1

depends on W_{xz}

$$\frac{\partial h_2}{\partial W_{xz}} = \frac{\partial z_2}{\partial W_{xz}} \cdot \underbrace{h_1}_{=0} + \underbrace{\frac{\partial h_1}{\partial W_{xz}}}_{=0\ (*)} \cdot z_2 + \underbrace{\frac{\partial \tilde{h}_2}{\partial W_{xz}}}_{=0\ (*)} (1-z_2) + \frac{\partial (1-z_2)}{\partial W_{xz}} \cdot \tilde{h}_2$$

$$\left(\begin{aligned} h_1 &= z_1 \cdot h_0 + (1-z_1)\cdot \tilde{h}_1 \\ &= (1-z_1)\,\tilde{h}_1 \\ &= (1-z_1)\tanh(W_{xh} x_1) \\ \Rightarrow \frac{\partial h_1}{\partial W_{xz}} &= \frac{\partial (1-z_1)}{\partial W_{xz}} \cdot \underbrace{\tanh(W_{xh}\cdot x_1)}_{0} \end{aligned}\right) \quad (*)$$

$$\left(\begin{aligned} \tilde{h}_2 &= \tanh(W_{xh}x_2 + W_{hh}\cdot r_2 \cdot h_1 + b_h) \\ \frac{\partial \tilde{h}_2}{\partial W_{xz}} &= \tanh'(\sim)\cdot W_{hh} \cdot \left(\frac{\partial r_2}{\partial W_{xz}} \cdot \underbrace{h_1}_{0} + \underbrace{\frac{\partial h_1}{\partial W_{xz}}}_{=0} \cdot r_2 \right) \\ &= 0 \end{aligned}\right) \quad (*)$$

Thus, $\dfrac{\partial h_2}{\partial W_{xz}} = -\dfrac{\partial z_2}{\partial W_{xz}} \cdot \tilde{h}_2$