Recurrent Neural Network (RNN)

Forward Propagation

Single Example

$$z^{(t)} = u x^{(t)} + w h^{(t-1)} + b^{(1)}$$
$$h^{(t)} = \phi(z^{(t)})$$
$$r^{(t)} = v h^{(t)} + b^{(2)}$$
$$y^{(t)} = \phi(r^{(t)})$$

Batch

$$\mathcal{M}^{(t)} = \mathcal{X}^{(t)} \mathcal{U}^{\tau} + \mathcal{H}^{(t-1)} \mathcal{W}^{\tau} + b^{(1)}$$
$$\mathcal{H}^{(t)} = \phi(\mathcal{M}^{(t)})$$
$$\mathcal{O}^{(t)} = \mathcal{H}^{(t)} \mathcal{V}^{\tau} + b^{(2)}$$
$$\mathcal{L} = \frac{1}{\tau} \sum_{t=1}^{\tau} l(\mathcal{O}_t, \mathcal{Y}_t)$$

Backpropagation Through Time (BPTT)

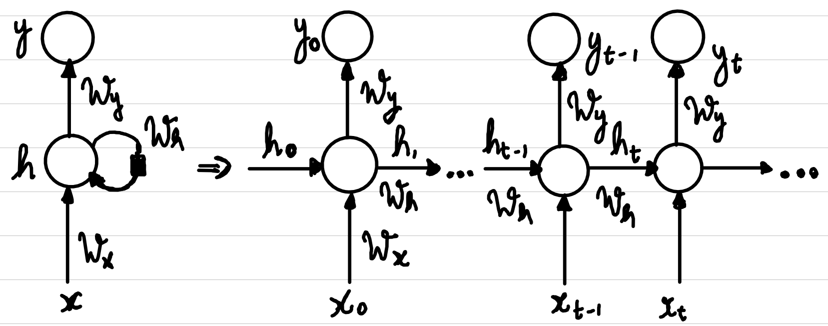
$$\overline{\mathcal{L}} = 1$$
$$\overline{y^{(t)}} = \overline{\mathcal{L}} \cdot \frac{\partial y}{\partial y^{(t)}}$$
$$\overline{r^{(t)}} = \overline{y^{(t)}} \phi'(r^{(t)})$$
$$\overline{h^{(t)}} = \overline{r^{(t)}} v + \overline{z^{(t+1)}} w$$
$$\overline{z^{(t)}} = \overline{h^{(t)}} \phi'(z^{(t)})$$
$$\overline{x^{(t)}} = \overline{z^{(t)}} u$$

$$\overline{v} = \sum_t \overline{r^{(t)}} h^{(t)}$$
$$\overline{u} = \sum_t \overline{z^{(t)}} x^{(t)}$$
$$\overline{w} = \sum_t \overline{z^{(t+1)}} h^{(t)}$$

$$\overline{\mathcal{L}} = 1$$
$$\overline{\mathcal{O}^{(t)}} = \frac{1}{\tau}(\mathcal{Y}^{(t)} - \mathcal{Y}^{(t)})$$
$$\overline{\mathcal{H}^{(t)}} = \overline{\mathcal{O}^{(t)}} \mathcal{V} + \overline{\mathcal{M}^{(t+1)}} \mathcal{W}$$
$$\overline{b^{(2)}} = (\overline{\mathcal{O}^{(t)}})^{\tau} \mathbb{1}$$
$$\overline{\mathcal{M}^{(t)}} = \overline{\mathcal{H}^{(t)}} \circ \phi'(\mathcal{M}^{(t)})$$
$$\overline{\mathcal{X}^{(t)}} = \overline{\mathcal{M}^{(t)}} \mathcal{U}$$
$$\overline{b^{(1)}} = (\overline{\mathcal{M}^{(t)}})^{\tau} \mathbb{1}$$
$$\overline{\mathcal{V}} = \sum^{\tau} (\overline{\mathcal{O}^{(t)}})^{\tau} \mathcal{H}^{(t)}$$
$$\overline{\mathcal{U}} = \sum^{\tau} (\overline{\mathcal{M}^{(t)}})^{\tau} \mathcal{X}^{(t)}$$
$$\overline{\mathcal{W}} = \sum^{\tau-1} (\overline{\mathcal{M}^{(t)}})^{\tau} \mathcal{H}^{(t+1)}$$

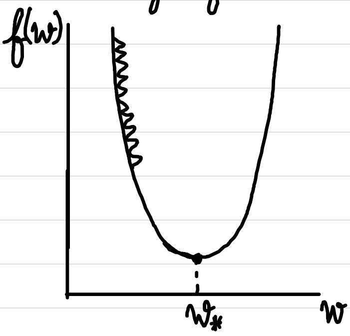
$\mathcal{X}$ - (batch_size, sequence_length, number_of_features)

Diagrams:



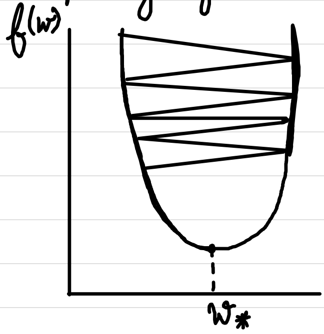
$$h_{t+1} = f(W_x x_t + W_h h_t + b_h)$$

Vanishing Gradients



- fails to capture long-terms information and dependencies

Exploding Gradients



- falls wide off optimal solution