RNNs (HW9) Tutorial

08.01.25 14:18

Exercises for the course
Deep Learning 1
Winter Semester 2024/25

Machine Learning Group
Faculty IV – Electrical Engineering and Computer Science
Technische Universität Berlin

Exercise Sheet 9

Exercise 1: Computing Gradients in RNNs ($5 \times 10 + 5 \times 10 = 100$ P)

We consider the task of binary classifying univariate time series (only two time steps for the purpose of the exercise) using a recurrent neural network. Let (x_1, x_2) be the time series given as input. The recurrent neural network is given by the equations:

$$h_1 = w \cdot x_1 + \tanh(h_0)$$
$$h_2 = w \cdot x_2 + \tanh(h_1)$$
$$y = h_1 + h_2,$$

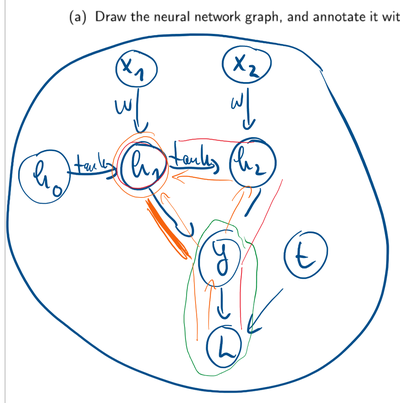
and we assume that the neural network has initial state $h_0 = 0$. The variable y is the neural network output and w is the model parameter. We further assume that the univariate time series (x_1, x_2) comes with a binary target label $t \in \{-1, 1\}$ and the prediction error for this data point is modeled via the log-loss function

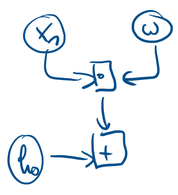
$$\mathcal{L}(y, t) = \log(1 + \exp(-yt)).$$

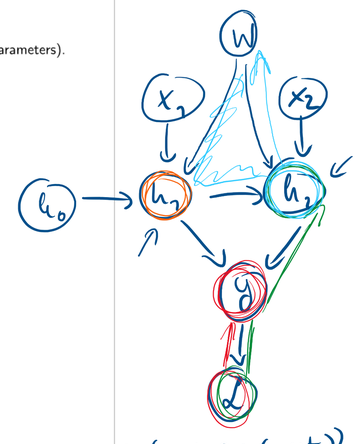
We would like to extract the gradient of the objective w.r.t. the parameter w.

(a) Draw the neural network graph, and annotate it with relevant variables (inputs, activations, and parameters).

$$h_2 = w \cdot x_2 + \tanh(h_1)$$
$$= w \cdot x_2 + \tanh(w \cdot x_1 + \tanh(h_0))$$







$$\frac{\partial \mathcal{L}}{\partial y} \quad \frac{\partial^+ y}{\partial^+ h_1}$$
$$\frac{\partial \mathcal{L}}{\partial h_2} \quad \frac{\partial^+ h_2}{\partial^+ h_1}$$
$$+ \rightarrow \frac{\partial \mathcal{L}}{\partial h_1}$$

(b) *Compute* $\partial \mathcal{L} / \partial y$.

$$\frac{\partial \mathcal{L}}{\partial y} = \frac{\partial \log(1 + \exp(-yt))}{\partial y} = \frac{\partial \log(1+\exp(-yt))}{\partial (1+\exp(-yt))} \cdot \frac{\partial (1+\exp(-yt))}{\partial(-yt)} \cdot \frac{\partial -yt}{\partial y}$$

$$= \frac{1}{1+\exp(-yt)} \cdot \exp(-yt) \cdot (-t) = \underbrace{\frac{-t \cdot \exp(-yt)}{1+\exp(-yt)}}_{=g} = -t \cdot \mathrm{sigm}(-yt)$$

(c) Assuming the last computation was stored in g, *compute* $\partial \mathcal{L} / \partial h_2$ as a function of g.

$$\frac{\partial \mathcal{L}}{\partial h_2} = \frac{\partial \mathcal{L}}{\partial y} \cdot \frac{\partial y}{\partial h_2} = g \cdot \underbrace{\frac{\partial h_1 + h_2}{\partial h_2}}_{=1} = g = \delta_2$$

(d) Assuming the last computation was stored in δ_2, *compute* $\partial \mathcal{L} / \partial h_1$ as a function of g and δ_2.

$$\frac{\partial \mathcal{L}}{\partial h_1} = \frac{\partial \mathcal{L}}{\partial h_2} \cdot \frac{\partial^+ h_2}{\partial^+ h_1} + \frac{\partial \mathcal{L}}{\partial y} \cdot \frac{\partial^+ y}{\partial^+ h_1}$$

$$\frac{\partial (h_1 + h_2)}{\partial h_1} = \underbrace{\frac{\partial h_1}{\partial h_1}}_{=1} + \frac{\partial h_2}{\partial h_1}$$

$$= \delta_2 \cdot \frac{\partial^+ (\cancel{x_2} \cdot w + \tanh(h_1))}{\partial^+ h_1} + g \cdot \underbrace{\frac{\partial^+ (h_1 + h_2)}{\partial^+ h_1}}_{=1}$$

$$= \underbrace{\delta_2}_{=g} \cdot \tanh'(h_1) + g = g(1 + \tanh'(h_1)) = \delta_1$$

$$\frac{\partial \mathcal{L}}{\partial h_1} = \frac{\partial \mathcal{L}}{\partial y} \cdot \frac{\partial y}{\partial h_1} = \frac{\partial \mathcal{L}}{\partial y} \cdot \frac{\partial (h_1 + h_2)}{\partial h_1} = \frac{\partial \mathcal{L}}{\partial y} \cdot \left(\frac{\partial h_1}{\partial h_1} + \frac{\partial h_2}{\partial h_1} \right)$$
$$= \frac{\partial \mathcal{L}}{\partial y} \cdot \frac{\partial h_1}{\partial h_1} + \frac{\partial \mathcal{L}}{\partial y} \cdot \frac{\partial h_2}{\partial h_1}$$

(e) Assuming the last computation was stored in δ_1, *compute* $\partial \mathcal{L} / \partial w$ as a function of g, δ_2 and δ_1.

$$\frac{\partial \mathcal{L}}{\partial w} = \frac{\partial \mathcal{L}}{\partial h_2} \cdot \frac{\partial^+ h_2}{\partial^+ w} + \frac{\partial \mathcal{L}}{\partial h_1} \cdot \frac{\partial^+ h_1}{\partial^+ w}$$

$$= \delta_2 \cdot x_2 + \delta_1 \cdot x_1$$

$$\frac{\partial \mathcal{L}}{\partial w} = \frac{\partial \mathcal{L}}{\partial y} \cdot \frac{\partial y}{\partial w} = \frac{\partial \mathcal{L}}{\partial y} \cdot \frac{\partial (h_1 + h_2)}{\partial w}$$
$$= \frac{\partial \mathcal{L}}{\partial y} \cdot \left(\frac{\partial h_1}{\partial w} + \frac{\partial h_2}{\partial w} \right)$$
$$= \frac{\partial \mathcal{L}}{\partial y} \cdot \frac{\partial h_1}{\partial w} + \frac{\partial \mathcal{L}}{\partial y} \cdot \frac{\partial h_2}{\partial w}$$
$$= g \cdot x_1 + g \cdot \frac{\partial (w x_2 + \tanh(h_1))}{\partial w} \rightarrow \frac{\partial w x_2}{\partial w} + \frac{\partial \tanh(h_1)}{\partial w}$$
$$= g \cdot x_1 + g \cdot (x_2 + \tanh'(h_1) \cdot x_1)$$
$$= g \cdot x_1 + g \cdot x_2 + g \cdot \tanh'(h_1) \cdot x_1$$
$$= g \cdot x_2 + \underbrace{(g + g \cdot \tanh'(h_1))}_{= g(1 + \tanh'(h_1)) = \delta_1} \cdot x_1$$

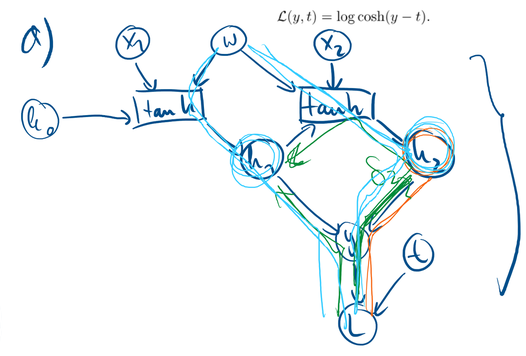
(f) Repeat the steps above (a–e) for the case where the recurrent neural network is given by the equations:

$$h_1 = \tanh(x_1 + w + h_0)$$
$$h_2 = \tanh(x_2 + w + h_1)$$
$$y = h_1 + h_2,$$

where the initial state is set to $h_0 = 0$, the target is real-valued ($t \in \mathbb{R}$), and the error function is given by

$$\mathcal{L}(y, t) = \log \cosh(y - t).$$

a)



b)

$$\frac{\partial \mathcal{L}}{\partial y} = \frac{\partial \log(\cosh(y-t))}{\partial y} = \frac{\partial \log(\cosh(y-t))}{\partial \cosh(\ldots)} \cdot \frac{\partial \cosh(y-t)}{\partial y}$$

$$= \frac{1}{\cosh(y-t)} \cdot \sinh(y-t) \cdot \underbrace{\frac{\partial y - t}{\partial y}}_{= \underbrace{\frac{\partial y}{\partial y}}_{=1} + \underbrace{\frac{\partial t}{\partial y}}_{=0}}$$

$$= \frac{\sinh(y-t)}{\cosh(y-t)} = \tanh(y-t)$$

c)

$$\frac{\partial \mathcal{L}}{\partial h_2} = \frac{\partial \mathcal{L}}{\partial y} \cdot \frac{\partial^+ y}{\partial^+ h_2} = g \cdot \underbrace{\frac{\partial^+ (h_1 + h_2)}{\partial^+ h_2}}_{=1} = \delta_2$$

d)

$$\frac{\partial \mathcal{L}}{\partial h_1} = \frac{\partial \mathcal{L}}{\partial y} \cdot \underbrace{\frac{\partial^+ y}{\partial^+ h_1}}_{=1} + \frac{\partial \mathcal{L}}{\partial h_2} \cdot \frac{\partial^+ h_2}{\partial^+ h_1} \rightarrow \frac{\partial^+ \tanh(x_2 + w + h_1)}{\partial^+ h_1}$$
$$= \tanh'(x_2 + w + h_1) \cdot 1$$

$$= g + g \cdot \tanh'(x_2 + w + h_1)$$
$$= g(1 + \tanh'(x_2 + w + h_1)) = \delta_1$$

e)

$$\frac{\partial \mathcal{L}}{\partial w} = \frac{\partial \mathcal{L}}{\partial h_1} \cdot \frac{\partial^+ h_1}{\partial^+ w} + \frac{\partial \mathcal{L}}{\partial h_2} \cdot \frac{\partial^+ h_2}{\partial^+ w}$$

$$= \delta_1 \cdot \frac{\partial^+ \tanh(h_0 + w + x_1)}{\partial^+ w} + \delta_2 \cdot \frac{\partial^+ \tanh(x_2 + w + h_1)}{\partial^+ w}$$

$$= \delta_1 \cdot \tanh'(h_0 + w + x_1) + \delta_2 \tanh'(x_2 + w + h_1)$$