IN5400
Theoretical Exercises

Mohamed Ismail

April 3, 2020

These exercises can give you a hint about how exercises for the written exam can be.

1 Week 10 - Recurrent neural networks

1.1 Task 1

You are given a simple RNN network as illustrated in the computational graph. Assume that we use identity functions as activation functions.

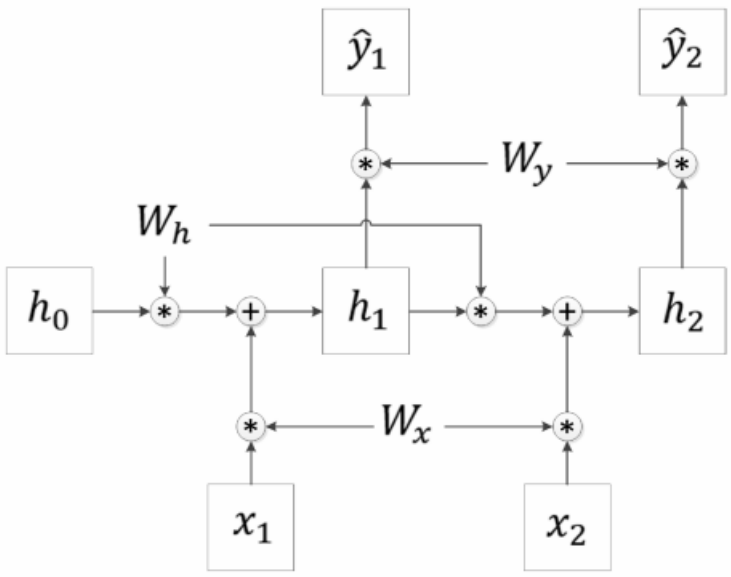

Figure 1: The structure of a very simple RNN network

Figure 1: Caption

Assume that the input at a given time, the hidden state, and the output at a given time are

scalar.
Let $h_0 = 1$, $x_1 = x_2 = 10$ and $y_1 = y_2 = 5$.
We assume the initial values of the weights are: $W_h = 1$, $W_x = 0.1$ and $W_y = 2$.

1.1.1 1.a

Compute the predicted value $\hat{y}_2$.

Answer

$$h_1 = x_1 \times W_x + h_0 \times W_h = 2 \tag{1}$$
$$h_2 = h_1 \times W_h + x_2 \times W_x = 3 \tag{2}$$
$$\hat{y}_2 = h_2 \times W_y = 6 \tag{3}$$

1.1.2 1.b

If we use the quadratic loss, the loss at a given time t is $L_t = (\hat{y}_t - y_t)^2$. Compute the total loss given the values for the weights and inputs given above.

Answer

$$h_1 = x_1 \times W_x + h_0 \times W_h = 2 \tag{4}$$
$$\hat{y}_1 = h_1 \times W_y = 4 \tag{5}$$
$$h_2 = h_1 \times W_h + x_2 \times W_x = 3 \tag{6}$$
$$\hat{y}_2 = h_2 \times W_y = 6 \tag{7}$$
$$L_1 = (\hat{y}_1 - y_1)^2 = 1 \tag{8}$$
$$L_2 = (\hat{y}_2 - y_2)^2 = 1 \tag{9}$$
$$L = L_1 + L_2 = 2 \tag{10}$$

1.1.3 1.c

Compute the derivative of the total loss with respect to h_1, $\frac{\partial L}{\partial h_1}$.

Answer

$$\hat{y}_1 = h_1 \times W_y \tag{11}$$
$$\hat{y}_2 = h_2 \times W_y = \big(h_1 \times W_h + x_2 \times W_x\big) \times W_y \tag{12}$$
$$L = L_1 + L_2 = (\hat{y}_1 - y_1)^2 + (\hat{y}_2 - y_2)^2 \tag{13}$$
$$\frac{\partial L}{\partial h_1} = \frac{\partial L_1}{\partial h_1} + \frac{\partial L_2}{\partial h_1} \tag{14}$$
$$\frac{\partial L_1}{\partial h_1} = \frac{\partial L_1}{\partial \hat{y}_1} \times \frac{\partial \hat{y}_1}{\partial h_1} = 2 \times (\hat{y}_1 - y_1) \times W_y = -4 \tag{15}$$
$$\frac{\partial L_2}{\partial h_1} = \frac{\partial L_2}{\partial \hat{y}_2} \times \frac{\partial \hat{y}_2}{\partial h_1} = 2 \times (\hat{y}_2 - y_2) \times W_h W_y = 4 \tag{16}$$
$$\frac{\partial L}{\partial h_1} = -4 + 4 = 0 \tag{17}$$

1.1.4 1.d

Compute the derivative of the total loss with respect to W_h, $\frac{\partial L}{\partial W_h}$

Answer

$$h_1 = x_1 \times W_x + h_0 \times W_h \tag{18}$$
$$\hat{y}_1 = h_1 \times W_y \tag{19}$$
$$\hat{y}_2 = h_2 \times W_y = \big(h_1 \times W_h + x_2 \times W_x\big) \times W_y \tag{20}$$
$$L = L_1 + L_2 = (\hat{y}_1 - y_1)^2 + (\hat{y}_2 - y_2)^2 \tag{21}$$
$$\frac{\partial L}{\partial W_h} = \frac{\partial L_1}{\partial W_h} + \frac{\partial L_2}{\partial W_h} \tag{22}$$
$$\frac{\partial L_1}{\partial W_h} = \frac{\partial L_1}{\partial \hat{y}_1} \times \frac{\partial \hat{y}_1}{\partial W_h} = 2 \times (\hat{y}_1 - y_1) \times h_0 W_y = -4 \tag{23}$$
$$\frac{\partial L_2}{\partial W_h} = \frac{\partial L_2}{\partial \hat{y}_2} \times \frac{\partial \hat{y}_2}{\partial W_h} = 2 \times (\hat{y}_2 - y_2) \times \big(h_0 W_h W_y + h_1 W_y\big) = 12 \tag{24}$$
$$\frac{\partial L}{\partial W_h} = -4 + 12 = 8 \tag{25}$$

1.2 Task 2

Recurrent neural networks are powerful models for processing sequential data.

1.2.1 2.a

Show and describe the most general recurrence formula for a recurrent neural network.

Answer

$$h_t = f_W(h_{t-1}, x_t) \tag{26}$$

Where

- h_t is the new state
- f_W is some function with parameters W
- h_{t-1} is the old state
- x_t is the input vector at time step t.

Note: We use the same function and parameters for every "time" step.

1.2.2 2.b

Why are long range dependencies difficult to learn in a recurrent neural network?

Answer
The challenge in training a recurrent neural network is to preserve long range dependencies.

The activation functions

- tanh can easily cause vanishing gradients.
  - tanh solves the exploding value problem
  - However, it does not solve the exploding gradient problem
  - Think of a scalar input and a scalar hidden state

$$h_t = \tanh W_{hh} h_{t-1} + W_{hx} x_t + b \tag{27}$$

$$\frac{\partial h_t}{\partial h_{t-1}} = \Big[1 - \tanh^2 W_{hh} h_{t-1} + W_{hx} x_t + b\Big] \times W_{hh} \tag{28}$$

  - The gradients can explode/vanish exponentially in time (steps)
  - If $|W_{hh}| < 1$, vanishing gradients
  - If $|W_{hh}| > 1$, exploding gradients
- *relu* can easily cause exploding values and/or gradients.

1.2.3 2.c

Why is Gated Recurrent Units (GRU) more efficient in preserving long range dependencies than vanilla RNNS?

Answer
GRU is a more advanced recurrent unit. It uses gates to control information flow. It is used to improve long term dependencies, because it has the ability to add and to remove from the state, not "transforming the state" only.

1.2.4 2.d

What is the advantage and disadvantage with using Truncated Backpropagation Through Time (TBTT)?

Answer

Advantages

- It reduces the memory requirements.
- It can update parameters faster.

Disadvantages

- It is not able to capture longer dependencies then the truncated length.