Deep Learning

Theoretical Exercises – Week 7 – Chapter 6

Exercises on the book "Deep Learning" written by Ian Goodfellow,
Yoshua Bengio, and Aaron Courville.
Exercises and solutions by T. Méndez and G. Schuster

FS 2024

1 Exercises on Deep Feedforward Networks

1. What is the advantage of several hidden layers over one very large hidden layer?

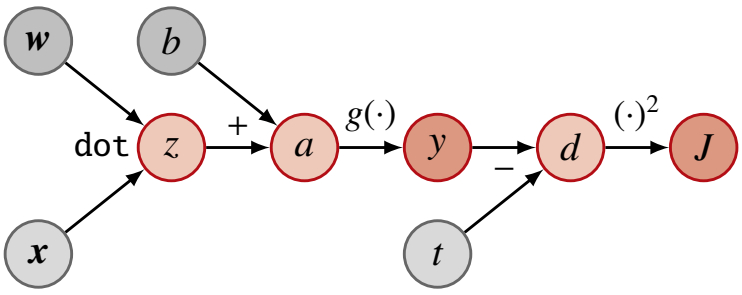
   Solution:
   In theory, a feedforward neural network with a single hidden layer is sufficient to represent any function, but the hidden layer may be unfeasibly large and may fail to learn and generalize properly.

   In many cases, using deeper models can significantly reduce the number of neurons required to represent the desired function and can reduce the amount of generalization error.

   Deep models also encode a very general belief that the function we want to learn should involve composition of several simpler functions. Since many real-world problems have this hierarchical structure, deep models work better. This is also demonstrated by many recent successes in practice, which rely on deep networks. Greater depth seems to result in better generalization for a wide variety of tasks.

2. Given is the following computational graph, where $\boldsymbol{x} = [x_1\ x_2]^T$, $\boldsymbol{w} = [w_1\ w_2]^T$ and

$$g(a) = \text{softplus(a)} = \log(1 + \exp(a)).$$



Calculate the following partial derivatives or gradients, respectively,

$$\frac{\partial J}{\partial b}, \qquad \nabla_w J, \qquad \nabla_x J, \qquad \text{and} \qquad \frac{\partial J}{\partial t}$$

by using the chain rule and the computational graph.

Solution:

The two partial derivatives can be calculated with the scalar chain rule as

$$\frac{\partial J}{\partial b} = \frac{\partial J}{\partial d} \cdot \frac{\partial d}{\partial y} \cdot \frac{\partial y}{\partial a} \cdot \frac{\partial a}{\partial b},$$
$$\frac{\partial J}{\partial t} = \frac{\partial J}{\partial d} \cdot \frac{\partial d}{\partial t},$$

where

$$\frac{\partial J}{\partial d} = 2 \cdot d = 2 \cdot (y - t),$$
$$\frac{\partial d}{\partial y} = 1,$$
$$\frac{\partial d}{\partial t} = -1,$$
$$\frac{\partial y}{\partial a} = \frac{\partial}{\partial a}\text{softplus(a)} = \frac{\exp(a)}{1 + \exp(a)},$$
$$\frac{\partial a}{\partial b} = 1.$$

Putting everything together results in

$$\frac{\partial J}{\partial b} = 2 \cdot (y - t) \cdot \frac{\exp(a)}{1 + \exp(a)},$$
$$\frac{\partial J}{\partial t} = -2 \cdot (y - t).$$

For the gradients it is a bit more complicated. They can be calculated with the vector chain rule as

$$\nabla_w J = \left(\frac{\partial z}{\partial \boldsymbol{w}}\right)^T \cdot \frac{\partial J}{\partial z} = \begin{bmatrix} \frac{\partial z}{\partial w_1} & \frac{\partial z}{\partial w_2} \end{bmatrix}^T \cdot \frac{\partial J}{\partial z}$$

and

$$\nabla_x J = \left(\frac{\partial z}{\partial \boldsymbol{x}}\right)^T \cdot \frac{\partial J}{\partial z} = \begin{bmatrix} \frac{\partial z}{\partial x_1} & \frac{\partial z}{\partial x_2} \end{bmatrix}^T \cdot \frac{\partial J}{\partial z},$$

where

$$\frac{\partial J}{\partial z} = \frac{\partial J}{\partial d} \cdot \frac{\partial d}{\partial y} \cdot \frac{\partial y}{\partial a} \cdot \frac{\partial a}{\partial z} = 2 \cdot (y - t) \cdot \frac{\exp(a)}{1 + \exp(a)}$$

and

$$\frac{\partial a}{\partial z} = 1.$$

This results in

$$\nabla_{\boldsymbol{w}} J = \begin{bmatrix} \dfrac{\partial z}{\partial w_1} \\ \dfrac{\partial z}{\partial w_2} \end{bmatrix} \cdot 2 \cdot (y - t) \cdot \frac{\exp(a)}{1 + \exp(a)} = \begin{bmatrix} 2 \cdot (y - t) \cdot \frac{\exp(a)}{1+\exp(a)} \cdot x_1 \\ 2 \cdot (y - t) \cdot \frac{\exp(a)}{1+\exp(a)} \cdot x_2 \end{bmatrix},$$

$$\nabla_{\boldsymbol{x}} J = \begin{bmatrix} \dfrac{\partial z}{\partial x_1} \\ \dfrac{\partial z}{\partial x_2} \end{bmatrix} \cdot 2 \cdot (y - t) \cdot \frac{\exp(a)}{1 + \exp(a)} = \begin{bmatrix} 2 \cdot (y - t) \cdot \frac{\exp(a)}{1+\exp(a)} \cdot w_1 \\ 2 \cdot (y - t) \cdot \frac{\exp(a)}{1+\exp(a)} \cdot w_2 \end{bmatrix},$$

since

$$\frac{\partial z}{\partial w_i} = \frac{\partial (x_1 \cdot w_1 + x_2 \cdot w_2)}{\partial w_i} = x_i,$$

$$\frac{\partial z}{\partial x_i} = \frac{\partial (x_1 \cdot w_1 + x_2 \cdot w_2)}{\partial x_i} = w_i.$$