CS5787: Exercises 1

FULL_CODE_FOLDER_URL

Mitchell Krieger
mak483

Meitong (Estelle) He
mh2585

1 Theory: Question 1 [12.5 pts]

a. What is the input X?
There are m samples in the batch with 10 features each so the input shape is $(m, 10)$

b. What is the shape of the hidden layer's weight vector W_h, and the shape of its bias vector b_h?
To transform an $(m, 10)$ input into a hidden layer of 50 neurons we need a shape W_h to have a shape of $(10, 50)$ so that we get a $(m, 50)$ result. After the weight matrix is applied, we then add the bias b_h to each row of W_h, which means that the result needs to be $(1, 50)$

c. What is the shape of the output layer's weight vector W_o, and its bias vector b_o?
$W_h : (50, 3), b_h : (1, 3)$

d. What is the shape of the network's output matrix Y? $(m, 3)$

e. Write an equation that computes the network's output matrix Y as a function of X, W_h, b_h, W_o, and b_o?
$Y = W_o a(W_h X + b_h) + b_o$
Where a is the ReLU activation funciton, $a(x) = \max(0, x)$.

2 Theory: Question 2 [12.5 pts]

For the first layer there are 3x3x3 parameters in each of the 100 kernels, and then we need bias parameters for each kernel as well. This totals 3x3x3x100 + 100 = 2800 parameters. In the second layer there are 3x3x100 parameters in each of the 200 kernels, and then we need bias parameters for each kernel as well. This totals 3x3x100x200 + 200 = 180200. In the final layer there are 3x3x200 parameters in each of the 200 kernels, and then we need bias parameters for each kernel as well. This totals 3x3x200x400 + 400 = 720400. Adding all the layers together we get 2800 + 180200 + 720400 = 903400 total parameters.

3 Theory: Question 3 [25 pts]

a. $\frac{\partial f}{\partial \gamma} = \frac{\partial f}{\partial y}\frac{\partial y}{\partial \gamma} = \frac{\partial f}{\partial y}\widehat{x}_i$

b. $\frac{\partial f}{\partial \beta} = \frac{\partial f}{\partial y}\frac{\partial y}{\partial \beta} = \frac{\partial f}{\partial y}$

c. $\frac{\partial f}{\partial \widehat{x}_i} = \frac{\partial f}{\partial y}\frac{\partial y}{\partial \widehat{x}_i} = \frac{\partial f}{\partial y}\gamma$

d. $\frac{\partial f}{\partial \sigma^2} = \frac{\partial f}{\partial y}\frac{\partial y}{\partial \widehat{x}_i}\frac{\partial \widehat{x}_i}{\partial \sigma^2} = \frac{\partial f}{\partial y}\gamma\frac{-1}{2}\frac{(x_i-\mu)}{\sqrt{(\sigma^2+\epsilon)^3}}$

e. $\frac{\partial f}{\partial \mu} = \frac{\partial f}{\partial y}\frac{\partial y}{\partial \widehat{x}_i}\frac{\partial \widehat{x}_i}{\partial \mu} + \frac{\partial f}{\partial y}\frac{\partial y}{\partial \widehat{x}_i}\frac{\partial \widehat{x}_i}{\partial \sigma^2}\frac{\partial \sigma^2}{\partial \mu} = \frac{\partial f}{\partial y}\gamma\frac{-1}{\sqrt{(\sigma^2+\epsilon)^3}} + \frac{\partial f}{\partial y}\gamma\frac{-1}{2}\frac{(x_i-\mu)}{\sqrt{(\sigma^2+\epsilon)^3}}\frac{-2}{m}\sum_{i=1}^{m}(x_i - \mu)$

f.

$$
\begin{aligned}
\frac{\partial f}{\partial x_i} &= \frac{\partial f}{\partial \widehat{x}_i}\frac{\partial \widehat{x}_i}{\partial x_i} + \frac{\partial f}{\partial \sigma^2}\frac{\partial \sigma^2}{\partial x_i} + \frac{\partial f}{\partial \mu}\frac{\partial \mu}{\partial x_i} \\
\frac{\partial f}{\partial x_i} &= \frac{\partial f}{\partial y}\gamma\frac{1}{\sqrt{(\sigma^2+\epsilon)^3}} \\
&\quad + \frac{\partial f}{\partial y}\gamma\frac{-1}{2}\frac{(x_i-\mu)}{\sqrt{(\sigma^2+\epsilon)^3}}\frac{2}{m}(x_i-\mu) \\
&\quad + \frac{\partial f}{\partial y}\gamma\frac{-1}{\sqrt{(\sigma^2+\epsilon)^3}} \\
&\quad + \frac{\partial f}{\partial y}\gamma\frac{-1}{2}\frac{(x_i-\mu)}{\sqrt{(\sigma^2+\epsilon)^3}}\frac{-2}{m}\sum_{i=1}^{m}(x_i-\mu)\frac{1}{m}
\end{aligned}
$$

4 Practical [50 pts]

TODO: Provide a report detailing your experiments, results and discussion in this section.