Deniz Orkun Eren
905624625

ECE C247 HW #2

2) Softmax Classifier Gradient

Let $w_i^T x^{(j)} + b_i = o_i$

$$\text{Likelihood} = \prod_{j=1}^{m} \prod_{n=1}^{c} Pr(y^{(j)} = n \mid x^{(j)}, \theta)^{\delta_n(y^{(j)})}$$

where $\delta_n(y^{(j)}) = \begin{cases} 1 & \text{if } y^{(j)} = n \\ 0 & \text{otw.} \end{cases}$

We are given that $Pr(y^{(j)} = n \mid x^{(j)}, \theta) = \text{softmax}_n(x^{(j)}) = \dfrac{e^{o_n}}{\sum_{k=0}^{c} e^{o_k}}$

So, $$\text{likelihood} = \prod_{j=1}^{m} \prod_{n=1}^{c} \text{softmax}_n(x^{(j)})^{\delta_n(y^{(j)})}$$

$$\mathcal{L} = \text{Log-likelihood} = \sum_{j=1}^{m} \sum_{n=1}^{c} \delta_n(y^{(j)}) \log(\text{softmax}_n(x^{(j)}))$$

$$\frac{\partial \mathcal{L}}{\partial w_i} = \frac{\partial \mathcal{L}}{\partial \text{softmax}_n} \cdot \frac{\partial \text{softmax}_n}{\partial o_i} \cdot \frac{\partial o_i}{\partial w_i}$$

$$\frac{dL}{do_i} = \sum_{j=1}^{m} \delta_i(y^{(j)}) \cdot \frac{1}{\text{Softmax}_i(x^{(j)})} \cdot \frac{d\,\text{Softmax}_i(x^{(j)})}{do_i} + \sum_{n \neq i} \delta_n(y^{(j)}) \cdot \frac{1}{\text{Softmax}_n(x^{(j)})} \cdot \frac{d\,\text{Softmax}_n(x^{(j)})}{do_i}$$

$$\frac{d\,\text{Softmax}_i(x^{(j)})}{do_i} = \frac{d\,\frac{e^{o_i}}{\sum_{k=1}^{c} e^{o_k}}}{do_i} = \frac{e^{o_i} \cdot \sum_{k=1}^{c} e^{o_k} - e^{o_i} \cdot e^{o_i}}{\left(\sum_{k=1}^{c} e^{o_k}\right)^2}$$

$$= \frac{e^{o_i}}{\sum_k e^{o_k}} - \left(\frac{e^{o_i}}{\sum_k e^{o_k}}\right)^2$$

$$= \text{Softmax}_i(x^{(j)}) \left[1 - \text{Softmax}_i(x^{(j)})\right]$$

$$\frac{d\,\text{Softmax}_n(x^{(j)})}{do_i} \quad \text{when } n \neq i = \frac{d\,\frac{e^{o_n}}{\sum_k e^{o_k}}}{do_i}$$

$$= \frac{0 \cdot \sum_k e^{o_k} - e^{o_n} \cdot e^{o_i}}{\left(\sum_k e^{o_k}\right)^2}$$

$$= -\text{Softmax}_n(x^{(j)}) \cdot \text{Softmax}_i(x^{(j)})$$

Plugging in to $\frac{dL}{do_i}$ gives:

$$\frac{dL}{do_i} = \sum_{j=1}^{m} \delta_i(y^{(j)}) \cdot \frac{softmax_i(x^{(j)}) \cdot [1 - softmax_i(x^{(j)})]}{softmax_i(x^{(j)})} + \sum_{n \neq i} \delta_n(y^{(j)}) \cdot \frac{-softmax_n(x^{(j)}) \cdot softmax_i(x^{(j)})}{softmax_n(x^{(j)})}$$

$$= \sum_{j=1}^{m} \delta_i(y^{(j)}) \cdot [1 - softmax_i(x^{(j)})] + \sum_{n \neq i} \delta_n(y^{(j)}) \cdot -softmax_i(x^{(j)})$$

$$= \sum_{j=1}^{m} \delta_i(y^{(j)}) - softmax_i(x^{(j)}) \cdot \sum_{n=1}^{c} \delta_n(y^{(j)})$$

We know that $\sum_{n=1}^{c} \delta_n(y^{(j)}) = 1$ since $\delta_n(y^{(j)})$ is 1 when $y^{(j)} = n$, 0 otw.

So, $$\frac{dL}{do_i} = \sum_{j=1}^{m} \delta_i(y^{(j)}) - softmax_i(x^{(j)})$$

$$\frac{do_i}{dw_i} = \frac{d\, w_i^T x^{(j)} + b_i}{dw_i} = x^{(j)}$$

$$\frac{do_i}{db_i} = \frac{d\, w_i^T x^{(j)} + b_i}{db_i} = 1$$

So, $$\frac{dL}{dw_i} = \frac{dL}{do_i} \cdot \frac{do_i}{dw_i} = \sum_{j=1}^{m} [\delta_i(y^{(j)}) - softmax_i(x^{(j)})] \cdot x^{(j)}$$

$$\frac{dL}{db_i} = \frac{dL}{do_i} \cdot \frac{do_i}{db_i} = \sum_{j=1}^{m} \delta_i(y^{(j)}) - softmax_i(x^{(j)})$$