Deep Learning Paper Assignment 1

Frederik Harder

November 9, 2016

Exercise 1

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial W_{out}} =& \frac{\partial \mathcal{L}}{\partial z_{out}} \frac{\partial z_{out}}{\partial s_{out}} \frac{\partial s_{out}}{\partial W_{out}} \\
=& (z_{out} - y_{gt}) f_3'(s_{out}) z_2 \\
\frac{\partial \mathcal{L}}{\partial W_2} =& \frac{\partial \mathcal{L}}{\partial z_{out}} \frac{\partial z_{out}}{\partial s_{out}} \frac{\partial s_{out}}{\partial z_2} \frac{\partial z_2}{\partial s_2} \frac{\partial s_2}{\partial W_2} \\
=& (z_{out} - y_{gt}) f_3'(s_{out}) W_{out} f_2'(s_2) z_1 \\
\frac{\partial \mathcal{L}}{\partial W_1} =& \frac{\partial \mathcal{L}}{\partial z_{out}} \frac{\partial z_{out}}{\partial s_{out}} \frac{\partial s_{out}}{\partial z_2} \frac{\partial z_2}{\partial s_2} \frac{\partial s_2}{\partial z_1} \frac{\partial z_1}{\partial s_1} \frac{\partial s_1}{\partial W_1} \\
=& (z_{out} - y_{gt}) f_3'(s_{out}) W_{out} f_2'(s_2) W_2 f_1'(s_1) x_{in}
\end{aligned}$$

Exercise 2

$$\Delta W_k = \frac{\partial \mathcal{L}}{\partial W_k} = \frac{\partial \mathcal{L}}{\partial s_k} \frac{\partial s_k}{\partial W_k} = \delta_k z_{k-1}$$

With $z_0 = x_{in}$

In the following $\circ$ denotes the Hadamard product, whereas $\cdot$ is the regular matrix product. α is set to 0.1.

$$s_{hid} = X \cdot W = \begin{bmatrix} 0.458 & 0.869 & 0.704 \\ 0.1205 & 0.1615 & 0.044 \\ -0.442 & -0.181 & 0.704 \\ 0.1195 & 0.1185 & -0.044 \end{bmatrix}$$

$$z_{hid} = ReLU(s_{hid}) = \begin{bmatrix} 0.458 & 0.869 & 0.704 \\ 0.1205 & 0.1615 & 0.044 \\ 0. & 0. & 0.704 \\ 0.1195 & 0.1185 & 0. \end{bmatrix}$$

$$s_{out} = z_{hid} \cdot w = \begin{bmatrix} 0.09859 \\ 0.011215 \\ 0.06336 \\ 0.005945 \end{bmatrix}$$

$$z_{out} = \tanh(s_{out}) = \begin{bmatrix} 0.09827181 \\ 0.01121453 \\ 0.06327535 \\ 0.00594493 \end{bmatrix}$$

$$\mathcal{L} = \frac{1}{2}(z_{out} - y)^2 = \begin{bmatrix} 0.40655687 \\ 0.48884835 \\ 0.56527723 \\ 0.5059626 \end{bmatrix}$$

$$\delta_{out} = (z_{out} - y) \circ (1 - z_{out} \circ z_{out}) = \begin{bmatrix} -0.89301989 \\ -0.98866111 \\ 1.05901824 \\ 1.00590938 \end{bmatrix}$$

$$\delta_{hid} = \delta_{out} \cdot w^T \circ \mathbb{1}[z_{hid} > 0] = \begin{bmatrix} -0.0178604 & -0.0267906 & -0.08037179 \\ -0.01977322 & -0.02965983 & -0.0889795 \\ 0. & 0. & 0.09531164 \\ 0.02011819 & 0.03017728 & 0. \end{bmatrix}$$

$$\Delta w = z_{hid}^T \delta_{out} = \begin{bmatrix} -0.4079306 \\ -0.81650279 \\ 0.07336175 \end{bmatrix}$$

$$\Delta W = x^T \cdot \delta_{hid} = \begin{bmatrix} -0.01332631 & -0.01998946 & -0.14955847 \\ -0.01628289 & -0.02442433 & 0.00750291 \end{bmatrix}$$

$$w' = w - \alpha \Delta w = \begin{bmatrix} 0.06079306 \\ 0.11165028 \\ 0.08266383 \end{bmatrix}$$

$$W' = W - \alpha \Delta W = \begin{bmatrix} 0.60133263 & 0.70199895 & 0.01495585 \\ 0.01162829 & 0.43244243 & 0.87924971 \end{bmatrix}$$

Values for the two repetitions are listed in the following

$$s'_{hid} = \begin{bmatrix} 0.4603021 & 0.87245316 & 0.71461665 \\ 0.12084794 & 0.16202191 & 0.04695365 \\ -0.44169684 & -0.18054526 & 0.69218288 \\ 0.11968511 & 0.11877767 & -0.04097132 \end{bmatrix}$$

$$z'_{hid} = \begin{bmatrix} 0.4603021 & 0.87245316 & 0.71461665 \\ 0.12084794 & 0.16202191 & 0.04695365 \\ 0. & 0. & 0.69218288 \\ 0.11968511 & 0.11877767 & 0. \end{bmatrix}$$

$$s'_{out} = \begin{bmatrix} 0.18446576 \\ 0.02931788 \\ 0.05721848 \\ 0.02053758 \end{bmatrix}, z'_{out} = \begin{bmatrix} 0.18240154 \\ 0.02930948 \\ 0.05715612 \\ 0.0205347 \end{bmatrix}, \mathcal{L}' = \begin{bmatrix} 0.33423362 \\ 0.47112004 \\ 0.55878953 \\ 0.52074553 \end{bmatrix}$$

$$\delta'_{out} = \begin{bmatrix} -0.7903967 \\ -0.96985665 \\ 1.05370258 \\ 1.02010436 \end{bmatrix}, \delta'_{hid} = \begin{bmatrix} -0.04805063 & -0.08824801 & -0.06533721 \\ -0.05896055 & -0.10828477 & -0.08017206 \\ 0. & 0. & 0.08710309 \\ 0.06201527 & 0.11389494 & 0. \end{bmatrix}$$

$$\Delta w' = \begin{bmatrix} -0.35893514 \\ -0.7255565 \\ 0.11898593 \end{bmatrix}, \Delta W' = \begin{bmatrix} -0.03542703 & -0.06506397 & -0.13036464 \\ -0.0444893 & -0.08170739 & 0.01340409 \end{bmatrix}$$

$$w'' = \begin{bmatrix} 0.09668657 \\ 0.18420593 \\ 0.07076523 \end{bmatrix}, W'' = \begin{bmatrix} 0.60487533 & 0.70850534 & 0.02799231 \\ 0.01607722 & 0.44061317 & 0.8779093 \end{bmatrix}$$

$$s''_{hid} = \begin{bmatrix} 0.46651828 & 0.88386955 & 0.72332167 \\ 0.12177893 & 0.16373173 & 0.04949393 \\ -0.44079473 & -0.17888847 & 0.68133321 \\ 0.12017121 & 0.11967041 & -0.038297 \end{bmatrix}$$

$$z''_{hid} = \begin{bmatrix} 0.46651828 & 0.88386955 & 0.72332167 \\ 0.12177893 & 0.16373173 & 0.04242393 \\ 0. & 0. & 0.68133321 \\ 0.12017121 & 0.11967041 & 0. \end{bmatrix}$$

$$s''_{out} = \begin{bmatrix} 0.25910609 \\ 0.04543719 \\ 0.0482147 \\ 0.03366294 \end{bmatrix}, z''_{out} = \begin{bmatrix} 0.25345924 \\ 0.04540595 \\ 0.04817738 \\ 0.03365023 \end{bmatrix}, \mathcal{L}'' = \begin{bmatrix} 0.27866155 \\ 0.4556249 \\ 0.54933791 \\ 0.5342164 \end{bmatrix}$$

$$\delta''_{out} = \begin{bmatrix} -0.6985818 \\ -0.95262596 \\ 1.04574449 \\ 1.03247979 \end{bmatrix}, \delta''_{hid} = \begin{bmatrix} -0.06754348 & -0.12868291 & -0.0494353 \\ -0.09210614 & -0.17547935 & -0.0674128 \\ 0. & 0. & 0.07400235 \\ 0.09982693 & 0.1901889 & 0. \end{bmatrix}$$

$$\Delta w'' = \begin{bmatrix} -0.3178366 \\ -0.64987299 \\ 0.16005189 \end{bmatrix}, \Delta W'' = \begin{bmatrix} -0.04911345 & -0.09357027 & -0.1060608 \\ -0.06363144 & -0.12122974 & 0.016283 \end{bmatrix}$$

$$w''' = \begin{bmatrix} 0.12847023 \\ 0.24919323 \\ 0.05476004 \end{bmatrix}, W''' = \begin{bmatrix} 0.60978668 & 0.71786237 & 0.03859839 \\ 0.02244036 & 0.45273615 & 0.876281 \end{bmatrix}$$

Exercise 3

(i)

Hinge loss penalizes all activations p_j which are not smaller than activation p_{y_i} for the target label y_i by a given *margin.* activations smaller than $(p_{y_i} - margin)$ all produce a loss of 0.

(ii)

Here $\mathbb{1}$ denotes the indicator function. First we derive the derivative of the softmax function.

$$\begin{aligned} \frac{\partial p_j}{\partial o_k} =& \frac{\partial}{\partial o_k} \frac{\exp(o_j)}{\sum_n \exp(o_n)} = \frac{\frac{\partial}{\partial o_k} \exp(o_j) \sum_n \exp(o_n) - \exp(o_j)\exp(o_k)}{\left(\sum_n \exp(o_n)\right)^2} \\ =& \frac{\exp(o_j)}{\sum_n \exp(o_n)} \left(\mathbb{1}[j = k] - \frac{exp(o_k)}{\sum_n \exp(o_n)} \right) = p_j(\mathbb{1}[j = k] - p_k) \end{aligned}$$

With this result we can find an expression for the derivatives of the hinge loss.

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial o_k} =& \frac{\partial}{\partial o_k} \sum_{j \neq y_i} \max(0, p_j - p_{y_i} - margin) \\ =& \sum_{j \neq y_i} \mathbb{1}[p_j > p_{y_i} - margin](p_j(\mathbb{1}[j = k] - p_k) - p_{y_i}(\mathbb{1}[y_i = k] - p_k))\end{aligned}$$

Assume first the case $k \neq y_i$:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial o_k} =& \mathbb{1}[p_k > p_{y_i} - margin](p_k(1 - p_k) + p_{y_i} p_k) + \sum_{j \notin \{k, y_i\}} \mathbb{1}[p_j > p_{y_i} - margin](-p_j p_k) + p_{y_i} p_k)) \\ =& \mathbb{1}[p_k > p_{y_i} - margin](p_k(1 + p_{y_i} - p_k)) + \sum_{j \notin \{k, y_i\}} \mathbb{1}[p_j > p_{y_i} - margin](p_k(p_{y_i} - p_j)) \\ =& \mathbb{1}[p_k > p_{y_i} - margin](p_k(1 + p_{y_i} - p_k)) + \sum_{j \notin \{k, y_i\}} \mathbb{1}[p_j > p_{y_i} - margin](p_k(p_{y_i} - p_j)) \\ =& \sum_{j \neq y_i} (\mathbb{1}[p_j > p_{y_i} - margin] p_k(p_{y_i} - p_j)) + \mathbb{1}[p_k > p_{y_i} - margin] p_k\end{aligned}$$

Now for the case of $k = y_i$:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial o_k} =& \sum_{j \neq y_i} \mathbb{1}[p_j > p_{y_i} - margin](p_j(-p_k) - p_{y_i}(1 - p_k)) \\ =& \sum_{j \neq y_i} \mathbb{1}[p_j > p_{y_i} - margin](p_{y_i}(p_k - 1) - p_j p_k) \\ =& \sum_{j \neq y_i} \mathbb{1}[p_j > p_{y_i} - margin](p_k(p_{y_i} - p_j) - p_{y_i}) \\ =& \sum_{j \neq y_i} (\mathbb{1}[p_j > p_{y_i} - margin] p_k(p_{y_i} - p_j)) - \sum_{j \neq y_i} \mathbb{1}[p_j > p_{y_i} - margin] p_{y_i}\end{aligned}$$

There are probably more elegant derivations, but the results shown here should still be correct and easy enough to implement, if a bit odd in notation.