Deep Learning paper assignment

Martijn Wardenaar

November 2016

1

$L = 0.5(y_{out} - y_{gt})^2$

Derivative to W_{out}

$\frac{\partial L}{\partial W_{out}} = \frac{\partial}{\partial W_{out}} 0.5(y_{out} - y_{gt})^2$

$= \frac{\partial}{\partial W_{out}} 0.5(f_3(W_{out} \cdot z_2) - y_{gt})^2$

$= (f_3(W_{out} \cdot z_2) - y_{gt}) \cdot f_3^{'}(W_{out} \cdot z_2) \cdot z_2$

Derivative to W_2

$\frac{\partial L}{\partial W_2} = \frac{\partial}{\partial W_2} 0.5(y_{out} - y_{gt})^2$

$= \frac{\partial}{\partial W_{out}} 0.5(f_3(W_{out} \cdot f_2(W_2 \cdot z_1)) - y_{gt})^2$

$= (f_3(W_{out} \cdot f_2(W_2 \cdot z_1)) - y_{gt}) \cdot f_3^{'}(W_{out} \cdot f_2(W_2 \cdot z_1)) \cdot (W_{out} \cdot f_2^{'}(W_2 \cdot z_1)) \cdot z_1$

$= (f_3(W_{out} \cdot f_2(s_2)) - y_{gt}) \cdot f_3^{'}(W_{out} \cdot f_2(s_2)) \cdot (W_{out} \cdot f_2^{'}(s_2)) \cdot z_1$

Derivative to W_1

$\frac{\partial L}{\partial W_1} = \frac{\partial}{\partial W_1} 0.5(y_{out} - y_{gt})^2$

$= \frac{\partial}{\partial W_{out}} 0.5(f_3(W_{out} \cdot f_2(W_2 \cdot f_1(W_1 \cdot x_{in})) - y_{gt})^2$

$= (f_3(W_{out} \cdot f_2(W_2 \cdot f_1(W_1 \cdot x_{in}))) - y_{gt}) \cdot f_3^{'}(W_{out} \cdot f_2(W_2 \cdot f_1(W_1 \cdot x_{in}))) \cdot (W_{out} \cdot f_2^{'}(W_2 \cdot f_1(W_1 \cdot x_{in}))) \cdot (W_2 \cdot f_1^{'}(W_1 \cdot x_{in})) \cdot x_{in}$

$= (f_3(W_{out} \cdot f_2(W_2 \cdot f_1(s_1))) - y_{gt}) \cdot f_3^{'}(W_{out} \cdot f_2(W_2 \cdot f_1(s_1))) \cdot (W_{out} \cdot f_2^{'}(s_2)) \cdot (W_2 \cdot f_1^{'}(s_{in})) \cdot x_1$

2 Prelude

The ΔW_k is the change in weights between neurons, so it is the error (δ_k) times the activation of the first node of the two nodes that are connected by k.

$$\Delta W_k = -\eta(\delta_k \cdot \frac{\partial f(x)}{\partial x})$$

where f(x) is the activation of the first layer of the connection of W_k.

3

The derivative of the ReLU is

$$\frac{\delta ReLU(x)}{\delta x} = \begin{cases} 1 & x > 0 \\ 0 & x \leq 0 \end{cases}$$

3.1 First iteration

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

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

$$s_{out} = s_{ReLU} \cdot W_{out} = \begin{bmatrix} 0.09859 \\ 0.011215 \\ 0.06336 \\ 0.005945 \end{bmatrix}$$

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

$$Loss = 0.5(y_{out} - Y)^2 = \begin{bmatrix} 0.40655687 \\ 0.48884835 \\ 0.56527723 \\ 0.5059626 \end{bmatrix}$$

$$\delta_{s_{out}} = tanh^{'}(s_{out}) \times (y_o ut - Y) = (1 - tanh^2(s_{out})) \times (y_{out} - Y) = \begin{bmatrix} -0.89301989 \\ -0.98866111 \\ 1.05901824 \\ 1.00590938 \end{bmatrix}$$

$$\delta_s = ReLU^{'}(s) \times (\delta_{s_{out}} \cdot W_{out}^T) = \begin{bmatrix} -0.0178604 & -0.0267906 & -0.08037179 \\ -0.01977322 & -0.02965983 & -0.0889795 \\ 0. & 0. & 0.09531164 \\ 0.02011819 & 0.03017728 & 0. \end{bmatrix}$$

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

$$\Delta W_{out} = s_{ReLU}^T \cdot \delta_{s_{out}} = \begin{bmatrix} -0.407930603934 \\ -0.816502794351 \\ 0.0733617484642 \end{bmatrix}$$

$$W = W - 0.5\Delta W = \begin{bmatrix} 0.60666315 & 0.70999473 & 0.07477924 \\ 0.01814144 & 0.44221217 & 0.87624855 \end{bmatrix}$$

$$W_{out} = W_{out} - 0.5\Delta W_{out} = \begin{bmatrix} 0.223965301967 \\ 0.438251397176 \\ 0.0533191257679 \end{bmatrix}$$

3.2 Second iteration

$$s = X \cdot W = \begin{bmatrix} 0.469510519996 & 0.886265779994 & 0.757083265289 \\ 0.122239702751 & 0.164109554126 & 0.0587682747412 \\ -0.440484208993 & -0.17872631349 & 0.644914409837 \\ 0.120425558313 & 0.11988833747 & -0.0288565799541 \end{bmatrix}$$

$$s_{ReLU} = ReLU(s) = \begin{bmatrix} 0.469510519996 & 0.886265779994 & 0.757083265289 \\ 0.122239702751 & 0.164109554126 & 0.0587682747412 \\ 0.0 & 0.0 & 0.644914409837 \\ 0.120425558313 & 0.11988833747 & 0.0 \end{bmatrix}$$

$$s_{out} = s_{ReLU} \cdot W_{out} = \begin{bmatrix} 0.533928299578 \\ 0.102432166357 \\ 0.0343862725276 \\ 0.0795123779332 \end{bmatrix}$$

$$y_{out} = tanh(s_{out}) = \begin{bmatrix} 0.488378173197 \\ 0.102075412221 \\ 0.0343727259781 \\ 0.0793452357355 \end{bmatrix}$$

$$Loss = 0.5(y_{out} - Y)^2 = \begin{bmatrix} 0.130878446831 \\ 0.403134282669 \\ 0.534963468124 \\ 0.582493068952 \end{bmatrix}$$

$$\delta_{s_{out}} = tanh^{'}(s_{out}) \times (y_o ut - Y) = (1 - tanh^2(s_{out})) \times (y_{out} - Y) = \begin{bmatrix} -0.38959324721 \\ -0.888568761505 \\ 1.03315063085 \\ 1.07255003816 \end{bmatrix}$$

$$\delta_s = ReLU^{'}(s) \times (\delta_{s_{out}} \cdot W_{out}^T) = \begin{bmatrix} -0.0872553692557 & -0.17073978492 & -0.0207727713463 \\ -0.199008570989 & -0.389416501216 & -0.0473777095481 \\ 0.0 & 0.0 & 0.0550866884235 \\ 0.240213993172 & 0.470046552766 & 0.0 \end{bmatrix}$$

$$\Delta W = X^T \cdot \delta_s = \begin{bmatrix} -0.0572004425051 & -0.11192882838 & -0.066370136737 \\ -0.0917654236126 & -0.179564980635 & 0.0250822481844 \end{bmatrix}$$

$$\Delta W_{out} = s_{ReLU}^T \cdot \delta_{s_{out}} = \begin{bmatrix} -0.1623740722 \\ -0.362519545451 \\ 0.319119548533 \end{bmatrix}$$

$$W = W - 0.5\Delta W = \begin{bmatrix} 0.635263373912 & 0.765959143179 & 0.107964305336 \\ 0.0640241561831 & 0.531994656883 & 0.863707422861 \end{bmatrix}$$

$$W_{out} = W_{out} - 0.5\Delta W_{out} = \begin{bmatrix} 0.305152338067 \\ 0.619511169901 \\ -0.106240648499 \end{bmatrix}$$

3.3 Third iteration

$$s = X \cdot W = \begin{bmatrix} 0.52766685538 & 1.00006508289 & 0.771939167291 \\ 0.130253882592 & 0.17979156148 & 0.0647782322103 \\ -0.425228205487 & -0.148873631878 & 0.609992709287 \\ 0.123851466973 & 0.126592095792 & -0.0215925100758 \end{bmatrix}$$

$$s_{ReLU} = ReLU(s) = \begin{bmatrix} 0.52766685538 & 1.00006508289 & 0.771939167291 \\ 0.130253882592 & 0.17979156148 & 0.0647782322103 \\ 0.0 & 0.0 & 0.609992709287 \\ 0.123851466973 & 0.126592095792 & 0.0 \end{bmatrix}$$

$$s_{out} = s_{ReLU} \cdot W_{out} = \begin{bmatrix} 0.698558946384 \\ 0.144248076007 \\ -0.0648060210142 \\ 0.116218782084 \end{bmatrix}$$

$$y_{out} = tanh(s_{out}) = \begin{bmatrix} 0.603452286645 \\ 0.143255852498 \\ -0.0647154486174 \\ 0.115698345523 \end{bmatrix}$$

$$Loss = 0.5(y_{out} - Y)^2 = \begin{bmatrix} 0.0786250444837 \\ 0.36700526714 \\ 0.437378596027 \\ 0.622391399102 \end{bmatrix}$$

$$\delta_{s_{out}} = tanh^{'}(s_{out}) \times (y_o ut - Y) = (1 - tanh^2(s_{out})) \times (y_{out} - Y) = \begin{bmatrix} -0.25214301473 \\ -0.83916183911 \\ 0.93136749617 \\ 1.10076348792 \end{bmatrix}$$

$$\delta_s = ReLU^{'}(s) \times (\delta_{s_{out}} \cdot W_{out}^T) = \begin{bmatrix} -0.076942030472 & -0.156205414038 & 0.0267878373993 \\ -0.256072197221 & -0.519870132683 & 0.0891530979825 \\ 0.0 & 0.0 & -0.0989490867838 \\ 0.335900551996 & 0.681935276183 & -0.0 \end{bmatrix}$$

$$\Delta W = X^T \cdot \delta_s = \begin{bmatrix} -0.041740851899 & -0.0847410318283 & 0.112133312734 \\ -0.0911522618385 & -0.185054601673 & -0.0532713446084 \end{bmatrix}$$

$$\Delta W_{out} = s_{ReLU}^T \cdot \delta_{s_{out}} = \begin{bmatrix} -0.106020426586 \\ -0.263685685408 \\ 0.319128893026 \end{bmatrix}$$

$$W = W - 0.5\Delta W = \begin{bmatrix} 0.656133799861 & 0.808329659093 & 0.0518976489694 \\ 0.109600287102 & 0.624521957719 & 0.890343095166 \end{bmatrix}$$

$$W_{out} = W_{out} - 0.5\Delta W_{out} = \begin{bmatrix} 0.35816255136 \\ 0.751354012605 \\ -0.265805095012 \end{bmatrix}$$

4

4.1

The hinge loss is designed to penalize for classes that are bigger or within a certain margin from the ground truth class. The vector P has a probability of the example belonging to each of the classes. If the probability of a class is higher than the probability of the ground truth class plus some margin, the difference between the probability of that class minus the probability of the ground truth class plus the margin is added to the loss.

4.2

The derivative of the softmax is

$$\frac{\partial p_j}{\partial o_i} = \begin{cases} p_i(1 - p_i) & j = i \\ -p_j p_i & j \neq i \end{cases}$$

The derivative of the hinge loss can be found by taking the derivative of the part that is larger than zero. So the general form is

$$\frac{\partial \mathcal{L}}{\partial oj} = \begin{cases} 0 & p_j - p_{y_i} + margin \leq 0 \\ \frac{\partial}{\partial o_j}(p_j - p_{y_i} + margin) & p_j - p_{y_i} + margin > 0 \end{cases}$$

and if we fill in the values from the derivative of the softmax we get:

$$\frac{\partial \mathcal{L}}{\partial oj} = \begin{cases} 0 & p_j - p_{y_i} + margin \leq 0 \\ -p_i p_j - p_i(1 - p_i) & p_j - p_{y_i} + margin > 0 \end{cases}$$