Deep Learning BackPropagation — Gradient of Sigmoid Cross Entropy

Zhe Li

May 20, 2018

Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be n training examples, and let $\sigma(z) = \frac{1}{1+\exp(-z)}$, for logistic regression, we would like to minimize the following L2 squared loss:

$$\mathcal{L} = \frac{1}{2}\sum_{i=1}^{n} ||\hat{y}_i - y_i||_2^2 \tag{1}$$

where $\hat{y}_i$ is the prediction for $\mathbf{x}_i$, computed by $\hat{y}_i = \sigma(z_i) = \sigma(W^T\mathbf{x}_i + b)$. In order to solve the minimization problem, we need to compute the gradient of $\mathcal{L}$ w.r.t W

$$\frac{\partial \mathcal{L}}{\partial W} = \sum_{i=1}^{n}(\hat{y}_i - y_i)\hat{y}_i(1-\hat{y}_i)\mathbf{x}_i \tag{2}$$

The gradient of $\mathcal{L}$ w.r.t

1 Cross-entropy loss function

If we take the cross entropy is loss function given as:

$$\mathcal{L} = \sum_{i=1}^{n} -y_i \log(\sigma(z_i)) - (1-y_i)\log(1-\sigma(z_i)) \tag{3}$$

Compute gradient of $\mathcal{L}$ w.r.t W:

$$\frac{\partial L}{\partial W} = \sum_{i=1}^{n} -y_i \frac{1}{\sigma(z_i)}\frac{\partial \sigma(z_i)}{\partial W} - (1-y_i)\frac{1}{1-\sigma(z_i)}\frac{\partial(1-\sigma(z_i))}{\partial W} \tag{4}$$

Using the property:

$$\frac{\partial \sigma(z_i)}{\partial W} = \sigma(z_i)(1-\sigma(z_i))\mathbf{x}_i \tag{5}$$

Plugging in the above, we can obtain

$$\frac{\partial L}{\partial W} = \sum_{i=1}^{n} -y_i \frac{1}{\sigma(z_i)} \frac{\partial \sigma(z_i)}{\partial W} - (1-y_i) \frac{1}{1-\sigma(z_i)} \frac{\partial (1-\sigma(z_i))}{\partial W} \tag{6}$$

$$= \sum_{i=1}^{n} -y_i (1-\sigma(z_i)) \mathbf{x}_i + (1-y_i)\sigma(z_i)\mathbf{x}_i \tag{7}$$

$$= \sum_{i=1}^{n} (\sigma(z_i) - y_i)\mathbf{x}_i = \sum_{i=1}^{n} (\hat{y}_i - y_i)\mathbf{x}_i \tag{8}$$

2 Two hidden layers

$$\begin{aligned}
t_1 &= XW^1 + b^1 \\
z_1 &= \sigma(t_1) \\
\hat{y} &= z_1 W^2 + b^2 \\
\mathcal{L} &= \sum_{i=1}^{k} -y_i \log(\hat{y}_i) - (1-y_i)\log(1-\hat{y}_i)
\end{aligned}$$

where k is the number of class. Here let's only consider at each iteration we use one example to update weights W^1, b^1, W^2, b^2. First compute gradient in terms of W^2, b^2, it is not that clear to directly compute gradient in terms of whole matrix W^2. So let's first consider the first column of $W^2 \in \mathbb{R}^{512\times 10}$

$$W^2 = \begin{bmatrix} w^2_{0,0} & w^2_{0,1} & \cdots & w^2_{0,9} \\ w^2_{1,0} & w^2_{1,1} & \cdots & w^2_{1,9} \\ \vdots & \vdots & \ddots & \vdots \\ w^2_{511,0} & w^2_{511,1} & \cdots & w^2_{511,9} \end{bmatrix}$$

$$y = \begin{bmatrix} y_0 & y_1 & \cdots & y_9 \end{bmatrix}$$

$$\hat{y}_i = \begin{bmatrix} \hat{y}_0 & \hat{y}_1 & \cdots & \hat{y}_9 \end{bmatrix}$$

$$z_1 = \begin{bmatrix} z_{1,0} & z_{1,1} & \cdots & z_{1,511} \end{bmatrix}$$

Let's look the i^{th} element $\hat{y}_i$, which is computed by

$$\hat{y}_i = \sigma(z_{1,0} * w^2_{0,i} + z_{1,1} * w^2_{1,i} + \cdots + z_{1,511} * w^2_{511,i})$$

Denote dW^2 as gradient of $\mathcal{L}$ in terms of W^2, so the i^th column of dW^2 is given as

$$(y_i - \hat{y}_i) * z_1^T$$

where $(y_i - \hat{y}_i)$ is the scalar and $z_1^T \in \mathbb{R}^{512}$, so the entire dW^2 matrix will be

$$dW^2 = \begin{bmatrix} (y_0 - \hat{y}_0) * z_1^T & (y_1 - \hat{y}_1) * z_1^T & \cdots & (y_9 - \hat{y}_9) * z_1^T \end{bmatrix}$$

As we seen $dW^2 \in \mathbb{R}^{512 \times 10}$. It is easy to see that

$$db^2 = \begin{bmatrix} (y_0 - \hat{y}_0) & (y_1 - \hat{y}_1) & \cdots & (y_9 - \hat{y}_9) \end{bmatrix}$$

2.1 Consider Mini Batch

if we consider mini batch (batch size = 128), we change our notation a little bit:

$$Y = \begin{bmatrix} y_{0,0} & y_{0,1} & \cdots & y_{0,9} \\ y_{1,0} & y_{1,1} & \cdots & y_{1,9} \\ \vdots & \vdots & \vdots & \vdots \\ y_{127,0} & y_{127,1} & \cdots & y_{127,9} \end{bmatrix}$$

$$\hat{Y} = \begin{bmatrix} \hat{y}_{0,0} & \hat{y}_{0,1} & \cdots & \hat{y}_{0,9} \\ \hat{y}_{1,0} & \hat{y}_{1,1} & \cdots & \hat{y}_{1,9} \\ \vdots & \vdots & \vdots & \vdots \\ \hat{y}_{127,0} & \hat{y}_{127,1} & \cdots & \hat{y}_{127,9} \end{bmatrix}$$

$$Z = \begin{bmatrix} \cdots z_0 \cdots \\ \cdots z_1 \cdots \\ \vdots \\ \cdots z_{127} \cdots \end{bmatrix} = \begin{bmatrix} z_{0,0} & z_{0,1} & \cdots & z_{0,512} \\ z_{1,0} & z_{1,1} & \cdots & z_{1,512} \\ \vdots & \vdots & \vdots & \vdots \\ z_{127,0} & z_{127,1} & \cdots & z_{127,512} \end{bmatrix}$$

Thus, the gradient dW^2 is

$$dW^2 = \begin{bmatrix} (y_{0,0} - \hat{y}_{0,0}) * z_0^T & (y_{0,1} - \hat{y}_{0,1}) * z_0^T & \cdots & (y_{0,9} - \hat{y}_{0,9}) * z_0^T \\ + & + & \cdots & + \\ (y_{1,0} - \hat{y}_{1,0}) * z_1^T & (y_{1,1} - \hat{y}_{1,1}) * z_1^T & \cdots & (y_{1,9} - \hat{y}_{1,9}) * z_0^T \\ + & + & \cdots & + \\ (y_{127,0} - \hat{y}_{127,0}) * z_{127}^T & (y_{127,1} - \hat{y}_{127,1}) * z_{127}^T & \cdots & (y_{127,9} - \hat{y}_{127,9}) * z_{127}^T \end{bmatrix}$$

We know that

$$Y - \hat{Y} = \begin{bmatrix} y_{0,0} - \hat{y}_{0,0} & y_{0,1} - \hat{y}_{0,1} & \cdots & y_{0,9} - \hat{y}_{0,9} \\ y_{1,0} - \hat{y}_{1,0} & y_{1,1} - \hat{y}_{1,1} & \cdots & y_{1,9} - \hat{y}_{1,9} \\ \vdots & \vdots & \cdots & \vdots \\ y_{127,0} - \hat{y}_{127,0} & y_{127,1} - \hat{y}_{127,1} & \cdots & y_{127,9} - \hat{y}_{127,9} \end{bmatrix}$$

$Y - \hat{Y} \in \mathbb{R}^{128 \times 10}$, from the above we could see that

$$dW^2 = Z^T * (Y - \hat{Y}) \tag{9}$$

3 gradient of $\mathcal{L}$ in terms of W^1

First, $W^1 \in \mathbb{R}^{3072\times 512}$. We focus on the first column W_0^1, since the first column W_1^1 only has relation with the first neuron in the first hidden layer.

$$t_{1,0} = xW_0^1 + b_0, \tag{10}$$

For simplicity we can drop b_0 at this moment. In order to compute the gradient only in terms of W_0^1, we check how this term connect to loss $\mathcal{L}$

$$\begin{aligned}\mathcal{L} =&\{-y_0\log(\hat{y}_0) - (1-y_0)\log(1-\hat{y}_0)\} + \{-y_1\log(\hat{y}_1) - (1-y_1)\log(1-\hat{y}_1)\}\\ &+\cdots+\{-y_9\log(\hat{y}_9) - (1-y_9)\log(1-\hat{y}_9)\}\end{aligned}$$

Let's check the 0^th term

$$\mathcal{L}_0 = -y_i\log(\hat{y}_0) - (1-y_0)\log(1-\hat{y}_0)$$

Compute gradient of $\mathcal{L}_0$ in terms of the first column W_0^1

$$\frac{\partial\mathcal{L}_0}{\partial W_0^1} = (y_0-\hat{y}_0)\frac{\partial t_{2,0}}{\partial z_0}\frac{\partial z_0}{\partial t_{1,0}}\frac{\partial t_{1,0}}{\partial W_0^1}$$

In the above equation, note that $\frac{\partial t_{1,0}}{\partial W_0^1} = x \in \mathbb{R}^{3072}$, $\frac{\partial t_{2,0}}{\partial z_0} = W_{0,0}^2$ and $\frac{\partial z_0}{\partial t_{1,0}} = \sigma(t_{1,0})(1-\sigma(t_{1,0}))$, which is denoted as δ_0 Thus, the gradient of $\mathcal{L}$

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial W_0^1} &= \frac{\partial\mathcal{L}_0}{\partial W_0^1} + \frac{\partial\mathcal{L}_1}{\partial W_0^1} + \cdots + \frac{\partial\mathcal{L}_9}{\partial W_0^1}\\ &= (y_0-\hat{y}_0)W_{0,0}^2\delta_0 x + (y_1-\hat{y}_1)W_{0,1}^2\delta_0 x + \cdots + (y_9-\hat{y}_9)W_{0,9}^2\delta_0 x\\ &= [(y_0-\hat{y}_0)(y_1-\hat{y}_1)\cdots(y_9-\hat{y}_9)][W_{0,0}^2W_{0,1}^2\cdots W_{0,9}^2]^T\circ\delta_0\end{aligned}$$

From the above we could obtain the gradient of $\mathcal{L}$ in terms of entire matrix W^1

$$\frac{\partial\mathcal{L}}{\partial W^1} = x\begin{bmatrix}(y_0-\hat{y}_0) & (y_1-\hat{y}_1) & \cdots & (y_9-\hat{y}_9)\end{bmatrix} * \begin{bmatrix} w_{0,0}^2 & w_{1,0}^2 & \cdots & w_{511,0}^2\\ w_{0,1}^2 & w_{1,1}^2 & \cdots & w_{511,1}^2\\ \vdots & \vdots & \cdots & \vdots\\ w_{0,9}^2 & w_{1,0}^2 & \cdots & w_{511,9}^2\end{bmatrix}\circ\begin{bmatrix}\delta_0 & \delta_1 & \cdots & \delta_{511}\end{bmatrix}$$

Let check if the dimensions of matrix are matchable or not, $x \in \mathcal{R}^{3072}$, $y-\hat{y} \in \mathbb{R}^{1\times 10}$, $W^2 \in \mathbb{R}^{10\times 512}$, $\delta \in \mathcal{R}^{512}$, so the $\frac{\partial\mathcal{L}}{\partial W^1} \in \mathbb{R}^{3072\times 512}$