1 Structure of dense layers

$$a^{[0]}_\mu \equiv \delta_{\mu 0} + \delta_{\mu i} x_i \tag{1}$$

$$a^{[l]}_\mu \equiv \delta_{\mu 0} + \delta^{[l]}_{\mu i}\; g^{[l]}\left(a^{[l-1]}_\nu \theta^{[l]}_{\nu i}\right) \tag{2}$$

$$z^{[l]}_i \equiv a^{[l-1]}_\nu \theta^{[l]}_{\nu i} \tag{3}$$

It should be understood that in $a^{[l]}_i$ one has $i \in \{1, ..., s_l\}$, and that in $a^{[l]}_\mu$ one has $\mu \in \{0, ..., s_l\}$. The binary cross entropy cross function is

$$cost = \frac{1}{m}\sum_n^m J\Big|_{x=x_n, y=y_n} \tag{4}$$

where

$$J \equiv -\sum_k^{s_L}\left(y_k \log\left(a^{[L]}_k\right) + (1-y_k)\log\left(1-a^{[L]}_k\right)\right) \tag{5}$$

2 Backward propagation. Recursion relation

The following recursion relations do the job

$$\underbrace{\frac{\partial J}{\partial \theta^{[l]}_{\nu_{l-1} j_l}}}_{\mathrm{d}\theta^{[l]}_{j_l j_{l-1}}} = \underbrace{\frac{\partial J}{\partial z^{[l]}_{j_l}}}_{\mathrm{d}z^{[l]}_{j_l}} \underbrace{\frac{\partial z^{[l]}_{j_l}}{\partial \theta^{[l]}_{\nu_{l-1} j_l}}}_{a^{[l-1]}_{\nu_{l-1}}} \tag{6}$$

$$\underbrace{\frac{\partial J}{\partial z^{[l]}_{j_l}}}_{\mathrm{d}z^{[l]}_{j_l}} = \underbrace{\frac{\partial J}{\partial z^{[l+1]}_{j_{l+1}}} \frac{\partial z^{[l+1]}_{j_{l+1}}}{\partial a^{[l]}_{j_l}}}_{\mathrm{d}a^{[l]}_{j_l}} \frac{\partial a^{[l]}_{j_l}}{\partial z^{[l]}_{j_l}} = \underbrace{\frac{\partial J}{\partial z^{[l+1]}_{j_{l+1}}} \theta^{[l+1]}_{j_l j_{l+1}}}_{\mathrm{d}a^{[l]}_{j_l} = \mathrm{d}z^{[l+1]}_{j_{l+1}} (\theta^{[l+1]T})_{j_{l+1} j_l}} \quad g^{[l]\prime}\left(z^{[l]}_{j_l}\right) \tag{7}$$

Hence, in an abbreviated notation one has:

$$\mathrm{d}\theta^{[l]}_{\nu_{l-1} j_l} = a^{[l-1]}_{\nu_{l-1}}\, \mathrm{d}z^{[l]}_{j_l} \tag{8}$$

$$\mathrm{d}a^{[l-1]}_{j_{l-1}} = \mathrm{d}z^{[l]}_{j_l} (\theta^{[l]T})_{j_l \nu_{l-1}} \tag{9}$$

$$\mathrm{d}z^{[l]}_{j_l} = \mathrm{d}a^{[l]}_{j_l}\; g^{[l]\prime}\left(z^{[l]}_{j_l}\right) \tag{10}$$

Hence, by providing $\{\mathrm{d}a^{[l]}, \theta^{[l]}\}$ one gets $\{\mathrm{d}\theta^{[l]}_{j_l \mu}, \mathrm{d}a^{[l-1]}_{j_{l-1}}\}$ and this process can be iterated. The initial condition is

$$\mathrm{d}\theta^{[L]}_{j_L\mu} = \frac{\partial J}{\partial \theta^{[L]}_{j_L\mu}} = \underbrace{\frac{(a^{[L]}_{j_L} - y_{j_L})}{a^{[L]}_{j_L}(1 - a^{[L-1]}_{j_L})}}_{\mathrm{d}a^{[L]}_{j_L}} \; g^{[L]\prime}\left(z^{[L]}_{j_L}\right) \, a^{[L-1]}_{\mu} \tag{11}$$

$$\mathrm{d}z^{[L]}_{j_L} = \mathrm{d}a^{[L]}_{j_L} \, g^{[L]\prime}\left(z^{[L]}_{j_L}\right) \tag{12}$$

$$\mathrm{d}\theta^{[L]}_{j_L\mu} = \mathrm{d}z^{[L]}_{j_L} \, a^{[L-1]}_{\mu} \tag{13}$$

This layer actually simplifies because one chooses $g^{[L]} = \sigma_{\text{sigmoid}}$, which means that $\sigma'_{\text{sigmoid}} = \sigma_{\text{sigmoid}}(1 - \sigma_{\text{sigmoid}})$.

3 Forward propagation convolutional neural networks

No padding case.

$$a^{[0]}_{ij} \equiv x_{ij} \tag{14}$$

$$z^{[l]}_{ij} = \theta^{[l]}_{rs} a^{[l-1]}_{(iS^{[l]}+r)(jS^{[l]}+s)} + b^{[l]}\delta_{ij} \tag{15}$$

$$a^{[l]}_{ij} \equiv g^{[l]}\left(z^{[l]}_{ij}\right) \tag{16}$$

It should be understood that incides of the activation of the l layer, $a^{[l]}_{ij}$, run over $i \in \{0, ..., n^{[l]}_H - 1\}$ and $j \in \{0, ..., n^{[l]}_W - 1\}$, where $n^{[l]}_X = (n^{[l-1]}_X - f^{[l]})/S^{[l]} + 1$. Finally, the indices of the weights of the l layer, $\theta^{[l]}_{rs}$, runs over $r, s \in \{0, ..., f^{[l]} - 1\}$.

4 Back prop

The following recursion relations do the job

$$\underbrace{\frac{\partial J}{\partial \theta^{[l]}_{rs}}}_{\mathrm{d}\theta^{[l]}_{rs}} = \underbrace{\frac{\partial J}{\partial z^{[l]}_{ij}}}_{\mathrm{d}z^{[l]}_{ij}} \times \underbrace{\frac{\partial z^{[l]}_{ij}}{\partial \theta^{[l]}_{rs}}}_{a^{[l]}_{(iS^{[l]}+r)(jS^{[l]}+s)}} \tag{17}$$

$$\underbrace{\frac{\partial J}{\partial z_{ij}^{[l]}}}_{\mathrm{d}z_{ij}^{[l]}} = \underbrace{\frac{\partial J}{\partial z_{uv}^{[l+1]}} \frac{\partial z_{uv}^{[l+1]}}{\partial a_{(uS^{[l]}+r)(vS^{[l]}+s)}^{[l]}} \delta_{i(uS^{[l]}+r)} \delta_{j(vS^{[l]}+s)}}_{\mathrm{d}a_{ij}^{[l]}} \frac{\partial a_{ij}^{[l]}}{\partial z_{ij}^{[l]}} \tag{18}$$

$$= \underbrace{\frac{\partial J}{\partial z_{uv}^{[l+1]}} \theta_{rs}^{[l+1]} \delta_{i(uS^{[l]}+r)} \delta_{j(vS^{[l]}+s)}}_{\mathrm{d}a_{ij}^{[l]} = \theta_{rs}^{[l+1]} \mathrm{d}z_{uv}^{[l+1]} \delta_{i(uS^{[l]}+r)} \delta_{j(vS^{[l]}+s)}} g^{[l]\prime}\left(z_{ij}\right) \tag{19}$$

Hence, in an abbreviated notation one has:

$$\mathrm{d}\theta_{j_l \mu}^{[l]} = \mathrm{d}z_{j_l}^{[l]} a_{j_{l-1}}^{[l-1]} \tag{20}$$

$$\mathrm{d}a_{j_{l-1}}^{[l-1]} = (\theta^{[l]T})_{j_{l-1} j_l} \mathrm{d}z_{j_l}^{[l]} \tag{21}$$

$$\mathrm{d}z_{j_l}^{[l]} = \mathrm{d}a_{j_l}^{[l]} \; g^{[l]\prime}\left(z_{j_l}^{[l]}\right) \tag{22}$$

5 Propagation with softmax

$$a_\mu^{[0]} \equiv \delta_{\mu 0} + \delta_{\mu i} x_i \tag{23}$$

$$a_\mu^{[l]} \equiv \delta_{\mu 0} + \delta_{\mu i}^{[l]} \; g^{[l]}\left(a_\nu^{[l-1]} \theta_{\nu i}^{[l]}\right), \quad l < L \tag{24}$$

$$z_i^{[l]} \equiv a_\nu^{[l-1]} \theta_{\nu i}^{[l]}, \quad l \leq L \tag{25}$$

$$\hat{y}_i \equiv \frac{e^{z_i^{[L]}}}{\sum_c e^{z_c^{[L]}}} \tag{26}$$

It should be understood that in $a_i^{[l]}$ one has $i \in \{1, ..., s_l\}$, and that in $a_\mu^{[l]}$ one has $\mu \in \{0, ..., s_l\}$. The loss function is

$$J \equiv -\sum_i y_i \log\left(\hat{y}_i\right) \tag{27}$$

6 Back propagation

The initial condition in this case changes to

$$\mathrm{d}\theta_{\nu_{L-1} j_L}^{[L]} = \frac{\partial J}{\partial \theta_{\nu_{L-1} j_L}^{[L]}} = \left(\frac{\partial J}{\partial \hat{y}_i}\right) \left(\frac{\partial \hat{y}_i}{\partial z_{j_L}^{[L]}}\right) \left(\frac{\partial z_{j_L}^{[L]}}{\partial \theta_{\nu_{L-1} j_L}^{[L]}}\right) \tag{28}$$

$$\frac{\partial J}{\partial \hat{y}_i} = \left(-\frac{y_i}{\hat{y}_i}\right) \tag{29}$$

$$\frac{\partial \hat{y}_i}{\partial z^{[L]}_{j_L}} = \hat{y}_i(\delta_{ij_L} - \hat{y}_{j_L}) \tag{30}$$

$$\frac{\partial z^{[L]}_{j_L}}{\partial \theta^{[L]}_{\nu_{L-1}j_L}} = a^{L-1}_{\nu_{L-1}} \tag{31}$$

Note that after summing over i, the first two terms in (28) simplify as:

$$\left(-\frac{y_i}{\hat{y}_i}\right)\left(\frac{\partial \hat{y}_i}{\partial z^{[L]}_{j_L}}\right) = -y_i + \hat{y}_i \tag{32}$$

****

7 Old, don't look!

8 Machine Learning coursera Andrew NG

In the lectures notes, one has $\theta^{[l]}_{\text{here}} = \theta^{[l-1]}_{\text{there}}$. We do this change to make the connexion with the deep learning specialization lectures.

$$a^{[0]}_{\mu} \equiv \delta_{\mu 0} + \delta_{\mu i} x_i \tag{33}$$

$$a^{[l]}_{\mu} \equiv \delta_{\mu 0} + \delta^{[l]}_{\mu i}\; g^{[l]}\left(\theta^{[l]}_{i\nu} a^{[l-1]}_{\nu}\right) \tag{34}$$

$$z^{[l]}_i \equiv \theta^{[l]}_{i\nu} a^{[l]}_{\nu} \tag{35}$$

It should be understood that in $a^{[l]}_i$ one has $i \in \{1, ..., s_l\}$, and that in $a^{[l]}_{\mu}$ one has $\mu \in \{0, ..., s_l\}$.

The cost fucntion has the structure

$$cost = \frac{1}{m}\sum_{n}^{m} J\Big|_{x=x_n, y=y_n} \tag{36}$$

where

$$J \equiv -\sum_{k}^{s_L}\left(y_k \log\left(a^{[L]}_k\right) + (1-y_k)\log\left(1 - a^{[L]}_k\right)\right) \tag{37}$$

To implement gradient descent we need the next partial derivatives ($l \in \{1, ..., L\}$)

$$\frac{\partial J}{\partial \theta^{[l]}_{j_l \mu}} = \frac{\partial J}{\partial a^{[L]}_{j_L}} \left(\frac{\partial a^{[L]}_{j_L}}{\partial a^{[L-1]}_{j_{L-1}}} \cdots \frac{\partial a^{[l+1]}_{j_{l+1}}}{\partial a^{[l]}_{j_l}} \right) \frac{\partial a^{[l]}_{j_l}}{\partial z^{[l]}_{j_l}} \frac{\partial z^{[l]}_{j_l}}{\partial \theta^{[l]}_{j_l \mu}} \tag{38}$$

through simple partial derivation one can show that

$$\frac{\partial J}{\partial a^{[L]}_{j_L}} = \frac{(a^{[L]}_{j_L} - y_{j_L})}{a^{[L]}_{j_L}(1 - a^{[L]}_{j_L})} \tag{39}$$

$$\frac{\partial z^{[l]}_{j_l}}{\partial \theta^{[l]}_{j_l \mu}} = a^{[l-1]}_{\mu} \tag{40}$$

$$\frac{\partial a^{[l]}_{j_l}}{\partial a^{[l-1]}_{j_{l-1}}} = \left(\frac{\partial a^{[l]}_{j_l}}{\partial z^{[l]}_{j_l}} \right) \left(\frac{\partial z^{[l]}_{j_l}}{\partial a^{[l-1]}_{j_{l-1}}} \right) = \left(a^{[l]\prime}_{j_l} \right) \left(\theta^{[l]}_{j_l j_{l-1}} \right) \tag{41}$$

$$\frac{\partial a^{[l]}_{j_l}}{\partial z^{[l]}_{j_l}} = a^{[l]\prime}_{j_l} \tag{42}$$

Then, we can write

$$\frac{\partial J}{\partial \theta^{[l]}_{j_l \mu}} = \frac{(a^{[L]}_{j_L} - y_{j_L})}{a^{[L]}_{j_L}(1 - a^{[L-1]}_{j_L})} \left(\frac{\partial a^{[L]}_{j_L}}{\partial a^{[L-1]}_{j_{L-1}}} \cdots \frac{\partial a^{[l+1]}_{j_{l+1}}}{\partial a^{[l]}_{j_l}} \right) a^{[l]\prime}_{j_l} \, a^{[l-1]}_{\mu} \tag{43}$$

9 Recursion relation

Let us write

$$\frac{\partial J}{\partial \theta^{[l]}_{j_l \mu}} = \frac{\partial J}{\partial a^{[l]}_{j_l}} \frac{\partial a^{[l]}_{j_l}}{\partial z^{[l]}_{j_l}} \frac{\partial z^{[l]}_{j_l}}{\partial \theta^{[l]}_{j_l \mu}} \tag{44}$$

since

$$\frac{\partial J}{\partial a^{[l]}_{j_l}} = \frac{\partial J}{\partial a^{[l+1]}_{j_{l+1}}} \frac{\partial a^{[l+1]}_{j_{l+1}}}{\partial a^{[l]}_{j_l}} = \frac{\partial J}{\partial a^{[l+1]}_{j_{l+1}}} \left(\frac{\partial a^{[l+1]}_{j_{l+1}}}{\partial z^{[l+1]}_{j_{l+1}}} \right) \left(\frac{\partial z^{[l+1]}_{j_{l+1}}}{\partial a^{[l]}_{j_l}} \right) = \frac{\partial J}{\partial a^{[l+1]}_{j_{l+1}}} a^{[l+1]\prime}_{j_{l+1}} \theta^{[l+1]}_{j_{l+1} j_l} \tag{45}$$

In Andrew's program they use the following notation:

$$\underbrace{\frac{\partial J}{\partial a^{[l]}_{j_l}}}_{\mathrm{d}z^{[l]}_{j_l}} = \underbrace{\theta^{[l+1]}_{j_{l+1} j_l} \frac{\partial J}{\partial a^{[l+1]}_{j_{l+1}}}}_{\mathrm{d}a^{[l]}_{j_l} = (\theta^{[l+1]T})_{j_l j_{l+1}} \mathrm{d}z^{[l+1]}_{j_{l+1}}} a^{[l+1]\prime}_{j_{l+1}} \tag{46}$$