3.10.17.

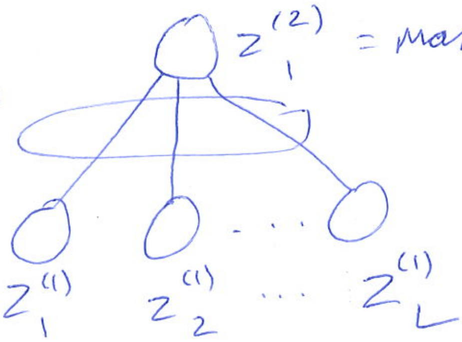
Sumeet Agarwal

ELL784: Notes on Backprop with MaxPooling

Consider a MaxPooling unit $z_1^{(2)}$, which is pooling over a bunch of feature map units denoted $z_1^{(1)}, z_2^{(1)}, \ldots z_L^{(1)}$

$$z_1^{(2)} = \max(z_1^{(1)}, z_2^{(1)}, \ldots z_L^{(1)})$$

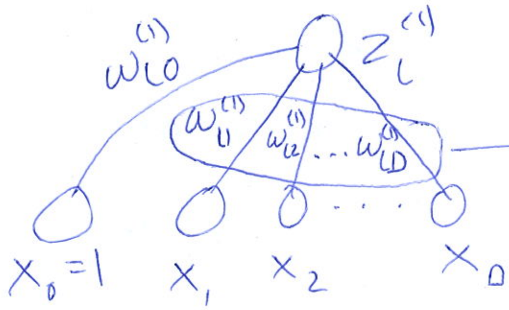
NB: These links don't have 'weights'



(receptive field)

Note that each feature map unit will in turn look like the below:

$l \in \{1, 2, \ldots L\}$



Most of these wts. will be 0 due to sparse connectivity

(In this whole derivation $\sigma()$ could be replaced by any activation fn. $h()$)

Formally, we have $z_l^{(1)} = \sigma(a_l^{(1)})$

where $$a_l^{(1)} = \sum_{i=0}^{D} w_{li}^{(1)} x_i = \sum_{i \in R_l} w_{li}^{(1)} x_i$$

Here R_l denotes the <u>receptive field</u> for $z_l^{(1)}$, i.e. the set of input units it is connected to

(Assume all $z_l^{(1)}$ connect to $x_0 = 1$)

Now, suppose we wish to compute a partial derivative like

$$\frac{\partial E(\underline{w})}{\partial w_{li}^{(1)}} \quad \left[\text{Assume } i \in R_l \text{, otherwise } w_{li}^{(1)} = 0 \text{ by defn.}\right]$$

The usual chain rule process gives us:

$$\frac{\partial E(\underline{w})}{\partial w_{li}^{(1)}} = \underbrace{\frac{\partial E(\underline{w})}{\partial a_l^{(1)}}}_{\triangleq \delta_l^{(1)}} \cdot \underbrace{\frac{\partial a_l^{(1)}}{\partial w_{li}^{(1)}}}_{= x_i}$$

So we need to get $\delta_l^{(1)}$ to compute this. This needs to come via backpropagation, as for all 'hidden' units where there's no direct way to define 'error'.

$$\delta_l^{(1)} = \frac{\partial E(\underline{w})}{\partial a_l^{(1)}} = \underbrace{\frac{\partial E(\underline{w})}{\partial z_1^{(2)}}}_{\triangleq \delta_1^{(2)}} \cdot \frac{\partial z_1^{(2)}}{\partial z_l^{(1)}} \cdot \underbrace{\frac{\partial z_l^{(1)}}{\partial a_l^{(1)}}}_{= \sigma'(a_l^{(1)})}$$

(Here I am proceeding as if $z_1^{(2)}$ is the only pooling unit to which $z_l^{(1)}$ connects. If there are others, we would get a summation as before. But note that pooling IS normally applied to non-overlapping regions of a feature map.)

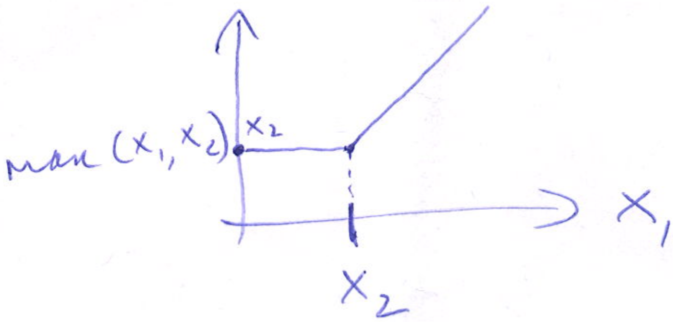
So we have

$$\delta_l^{(1)} = \delta_1^{(2)} \cdot \frac{\partial z_1^{(2)}}{\partial z_l^{(1)}} \sigma'(a_l^{(1)}) \qquad = ?$$

The first term is from backprop. above $z_1^{(2)}$. So only the middle term needs to be obtained. This is:

$$\frac{\partial}{\partial z_l^{(1)}} \max\left(z_1^{(1)}, z_2^{(1)}, \ldots z_L^{(1)}\right)$$

In general the max() fn. is not differentiable everywhere. But it is actually differentiable at most points. Consider the following:



(Here x_2 is taken fixed.)

For $x_1 < x_2$, $\dfrac{\partial \max(x_1, x_2)}{\partial x_1} = 0$

For $x_1 > x_2$, $\dfrac{\partial \max(x_1, x_2)}{\partial x_1} = 1$

The problem arises <u>only</u> at $x_1 = x_2$

So, coming back to $\max(z_1^{(1)}, z_2^{(1)}, \dots z_L^{(1)})$, we can actually differentiate it, as long as we assume that no two (or more) of its arguments have the same value which is also the maximum value. In this case, we have a unique maximum:

$$l_{max} \triangleq \underset{l}{\operatorname{argmax}} \; z_l^{(1)}, \quad l \in \{1, \dots L\}$$

And we thus have

$$\frac{\partial}{\partial z_l^{(1)}} \max(z_1^{(1)}, z_2^{(1)}, \dots z_L^{(1)})$$

$$= \begin{cases} 1 \,; & \text{if } l = l_{max} \\ 0 \,; & \text{otherwise.} \end{cases}$$

This in turn gives us:

$$\delta_l^{(1)} = \begin{cases} \delta_1^{(2)} \cdot \sigma'(a_l^{(1)}) \,; & l = l_{max} \\ 0 \,; & \text{otherwise} \end{cases}$$

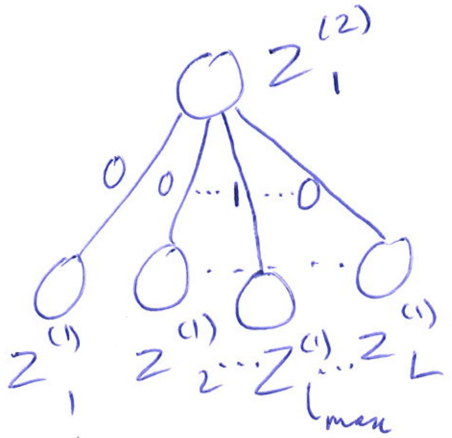
So we have all we need to compute any

$$\frac{\partial E(\underline{w})}{\partial w_{li}^{(1)}} .$$

Notice that the entire error at $z_1^{(2)}$ is being propagated back to <u>just a single unit</u> in the layer below : $z_{l_{max}}^{(1)}$.

This is intuitively reasonable : For the current setting of the weights, $z_{l_{max}}^{(1)}$ is the controlling unit for $z_1^{(2)}$, and no other unit has an effect on it.

There is another way to think of this. Although we said that a MaxPooling unit has no 'weights', supposing we were to write it out as if it did. We can do that as follows:



i.e., all wts. are 0 except for $z_{l_{max}}^{(1)}$

If we imagine these 'weights' and apply the normal backprop principle (propagate the error back along each link, multiplying by its weight), we get the same thing as above!