THE INCREDIBLE SHRINKING NEURAL NETWORK: PRUNING TO OPERATE IN CONSTRAINED MEMORY ENVIRONMENTS

Nikolas Wolfe, Aditya Sharma, Lukas Drude & Bhiksha Raj
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{nwolfe, adityasharma, bhiksha}@cmu.edu
{drude@nt.upb.de}

ABSTRACT

We propose and evaluate a method for pruning neural networks to operate in constrained memory environments such as mobile or embedded devices. We evaluate a simple pruning technique using first-order derivative approximations of the gradient of each neuron in an optimally trained network, and turning off those neurons which contribute least to the output of the network. We then show the limitations of this type of approximation by comparing against the ground truth value for the change in error resulting from the removal of a given neuron. We attempt to improve on this using a second-order derivative approximation. We also explore the correlation between neurons in a trained network and attempt to improve our choice of candidate neurons for removal to account for faults that can occur from the removal of a single neuron at a time. We argue that this method of pruning allows for the optimal tradeoff in network size versus accuracy in order to operate within the memory constraints of a particular device or application environment.

1 TAYLOR SERIES REPRESENTATION OF ERROR

Let us denote the total error from the optimally trained neural network for any given validation dataset with N instances as E_{total}. Then,

$$E_{\text{total}} = \sum_n E_n, \tag{1}$$

where E_n is the error from the network over one validation instance. E_n can be seen as a function O, where O is the output of any general neuron in the network (In reality this error depends on each neuron's output, but for the sake of simplicity we use O to represent that). This error can be approximated at a particular neuron's output (say O_k) by using the 2nd order Taylor Series as,

$$\hat{E}_n(O) \approx E_n(O_k) + (O - O_k) \cdot E_n(O_k) + \left.\frac{\partial E_n}{\partial O}\right|_{O_k} + 0.5 \cdot (O - O_k)^2 \cdot \left.\frac{\partial^2 E_n}{\partial O^2}\right|_{O_k}, \tag{2}$$

where $\hat{E}_n(O_k)$ represents the contribution of a neuron k to the total error E_n of the network for any given validation instance n. When this neuron is pruned, its output O_k becomes 0. From equation 2, the contribution $E_n(0)$ of this neuron, then becomes:

$$\hat{E}_n(0) \approx E_n(O_k) - O_k \cdot \left.\frac{\partial E_n}{\partial O}\right|_{O_k} + 0.5 \cdot O_k^2 \cdot \left.\frac{\partial^2 E_n}{\partial O^2}\right|_{O_k} \tag{3}$$

Replacing O by O_k in equation 2 shows us that the error is approximated perfectly by equation 2 at O_k. Using this and equation 3 we get:

$$\Delta E_{n,k} = \hat{E}_n(0) - \hat{E}_n(O_k) = -O_k \cdot \left.\frac{\partial E_n}{\partial O}\right|_{O_k} + 0.5 \cdot O_k^2 \cdot \left.\frac{\partial^2 E_n}{\partial O^2}\right|_{O_k}, \tag{4}$$

where $\Delta E_{n,k}$ is the change in the total error of the network given a validation instance n, when exactly one neuron (k) is turned off.

2 Second Derivative Gradient Terms

Given a neuron n with output a_i, and outgoing weights $[w_{i,1}, w_{i,2}, \ldots w_{i,j}]$, the input x to each of its j forward-connected neurons is given by:

$$x_j = \sum_i (w_{i,j} \cdot a_i) \tag{5}$$

For simplicity's sake, we will drop the index variable i and examine only the connection between the output a of neuron n and each of its j connections to the next forward layer. So the contribution c_j from neuron n to the input of each forward-connected neuron is given by:

$$c_j = w_j \cdot a \tag{6}$$

Where w_j is the weight connecting the output a from n to the input of the jth neuron. We will denote the error function of an optimally trained network as E. The second-derivative of E with respect to the output of neuron n is given by:

$$\frac{d^2E}{da^2} = \frac{d}{da}\frac{dE}{da} = \frac{d}{da}\sum_j \left(\frac{dE}{dc_j} \cdot \frac{dc_j}{da}\right) = \sum_j \left(\frac{d}{da}\frac{dE}{dc_j} \cdot \frac{dc_j}{da}\right) \tag{7}$$

Which states that the 2nd derivative of E with respect to a is the sum of the 2nd derivative terms of all outgoing connections.

3 SECOND DERIVATIVE BACK-PROPAGATION

Name and network definitions:

$$E = \frac{1}{2}\sum_i (o_i^{(0)} - t_i)^2 \quad o_i^{(m)} = \sigma(x_i^{(m)}) \quad x_i^{(m)} = \sum_j w_{ji}^{(m)} o_j^{(m+1)} \quad c_{ji}^{(m)} = w_{ji}^{(m)} o_j^{(m+1)} \tag{8}$$

Superscripts represent the index of the layer of the network in question, with 0 representing the output layer. E is the squared-error network cost function. $o_i^{(m)}$ is the ith output in layer m generated by the activation function σ, which in this paper is is the standard logistic sigmoid. $x_i^{(m)}$ is the weighted sum of inputs to the ith neuron in the mth layer, and $c_{ji}^{(m)}$ is the contribution of the jth neuron in the $m+1$ layer to the input of the ith neuron in the mth layer.

3.1 FIRST AND SECOND DERIVATIVES

The first and second derivatives of the cost function with respect to the outputs:

$$\frac{\partial E}{\partial o_i^{(0)}} = o_i^{(0)} - t_i \tag{9}$$

$$\frac{\partial^2 E}{\partial {o_i^{(0)}}^2} = 1 \tag{10}$$

The first and second derivatives of the sigmoid function in forms depending only on the output:

$$\sigma'(x) = \sigma(x)\left(1 - \sigma(x)\right) \tag{11}$$
$$\sigma''(x) = \sigma'(x)\left(1 - 2\sigma(x)\right) \tag{12}$$

The second derivative of the sigmoid is easily derived from the first derivative:

$$\sigma'(x) = \sigma(x)\left(1 - \sigma(x)\right) \tag{13}$$
$$\sigma''(x) = \frac{\mathrm{d}}{\mathrm{d}x} \underbrace{\sigma(x)}_{f(x)} \underbrace{\left(1 - \sigma(x)\right)}_{g(x)} \tag{14}$$
$$\sigma''(x) = f'(x)g(x) + f(x)g'(x) \tag{15}$$
$$\sigma''(x) = \sigma'(x)(1 - \sigma(x)) - \sigma(x)\sigma'(x) \tag{16}$$
$$\sigma''(x) = \sigma'(x) - 2\sigma(x)\sigma'(x) \tag{17}$$
$$\sigma''(x) = \sigma'(x)(1 - 2\sigma(x)) \tag{18}$$

And for future convenience:

$$\frac{\mathrm{d}o_i^{(m)}}{\mathrm{d}x_i^{(m)}} = \frac{\mathrm{d}}{\mathrm{d}x_i^{(m)}}\left(o_i^{(m)} = \sigma(x_i^{(m)})\right) \tag{19}$$
$$= \left(o_i^{(m)}\right)\left(1 - o_i^{(m)}\right) \tag{20}$$
$$= \sigma'\left(x_i^{(m)}\right) \tag{21}$$
$$\frac{\mathrm{d}^2 o_i^{(m)}}{\mathrm{d}{x_i^{(m)}}^2} = \frac{\mathrm{d}}{\mathrm{d}x_i^{(m)}}\left(\frac{\mathrm{d}o_i^{(m)}}{\mathrm{d}x_i^{(m)}} = \left(o_i^{(m)}\right)\left(1 - o_i^{(m)}\right)\right) \tag{22}$$
$$= \left(o_i^{(m)}\left(1 - o_i^{(m)}\right)\right)\left(1 - 2o_i^{(m)}\right) \tag{23}$$
$$= \sigma''\left(x_i^{(m)}\right) \tag{24}$$

Derivative of the error with respect to the ith neuron's input $x_i^{(0)}$ in the output layer:

$$\frac{\partial E}{\partial x_i^{(0)}} = \frac{\partial E}{\partial o_i^{(0)}} \frac{\partial o_i^{(0)}}{\partial x_i^{(0)}} \tag{25}$$

$$= \underbrace{\left(o_i^{(0)} - t_i\right)}_{\text{from (9)}} \underbrace{\sigma\left(x_i^{(0)}\right)\left(1 - \sigma\left(x_i^{(0)}\right)\right)}_{\text{from (11)}} \tag{26}$$

$$= \left(o_i^{(0)} - t_i\right)\left(o_i^{(0)}\left(1 - o_i^{(0)}\right)\right) \tag{27}$$

$$= \left(o_i^{(0)} - t_i\right)\sigma'\left(x_i^{(0)}\right) \tag{28}$$

Second derivative of the error with respect to the ith neuron's input $x_i^{(0)}$ in the output layer:

$$\frac{\partial^2 E}{\partial {x_i^{(0)}}^2} = \frac{\partial}{\partial x_i^{(0)}}\left(\frac{\partial E}{\partial o_i^{(0)}} \frac{\partial o_i^{(0)}}{\partial x_i^{(0)}}\right) \tag{29}$$

$$= \frac{\partial^2 E}{\partial x_i^{(0)} o_i^{(0)}} \frac{\partial o_i^{(0)}}{\partial x_i^{(0)}} + \frac{\partial E}{\partial o_i^{(0)}} \frac{\partial^2 o_i^{(0)}}{\partial {x_i^{(0)}}^2} \tag{30}$$

$$= \frac{\partial^2 E}{\partial x_i^{(0)} o_i^{(0)}} \underbrace{\left(o_i^{(0)}\left(1 - o_i^{(0)}\right)\right)}_{\text{from (11)}} + \underbrace{\left(o_i^{(0)} - t_i\right)}_{\text{from (9)}} \underbrace{\left(o_i^{(0)}\left(1 - o_i^{(0)}\right)\right)\left(1 - 2o_i^{(0)}\right)}_{\text{from (12)}} \tag{31}$$

$$\left(\frac{\partial^2 E}{\partial x_i^{(0)} o_i^{(0)}}\right) = \frac{\partial}{\partial x_i^{(0)}} \frac{\partial E}{\partial o_i^{(0)}} = \frac{\partial}{\partial x_i^{(0)}} \underbrace{\left(o_i^{(0)} - t_i\right)}_{\text{from (9)}} = \frac{\partial o_i^{(0)}}{\partial x_i^{(0)}} = \underbrace{\left(o_i^{(0)}\left(1 - o_i^{(0)}\right)\right)}_{\text{from (11)}} \tag{32}$$

$$\frac{\partial^2 E}{\partial {x_i^{(0)}}^2} = \left(o_i^{(0)}\left(1 - o_i^{(0)}\right)\right)^2 + \left(o_i^{(0)} - t_i\right)\left(o_i^{(0)}\left(1 - o_i^{(0)}\right)\right)\left(1 - 2o_i^{(0)}\right) \tag{33}$$

$$= \left(\sigma'\left(x_i^{(0)}\right)\right)^2 + \left(o_i^{(0)} - t_i\right)\sigma''\left(x_i^{(0)}\right) \tag{34}$$

First derivative of the error with respect to a single input contribution $c_{ji}^{(0)}$ from neuron j to neuron i with weight $w_{ji}^{(0)}$ in the output layer:

$$\frac{\partial E}{\partial c_{ji}^{(0)}} = \frac{\partial E}{\partial o_i^{(0)}} \frac{\partial o_i^{(0)}}{\partial x_i^{(0)}} \frac{\partial x_i^{(0)}}{\partial c_{ji}^{(0)}} \tag{35}$$

$$= \underbrace{\left(o_i^{(0)} - t_i\right)}_{\text{from (9)}} \underbrace{\left(o_i^{(0)}\left(1 - o_i^{(0)}\right)\right)}_{\text{from (11)}} \frac{\partial x_i^{(0)}}{\partial c_{ji}^{(0)}} \tag{36}$$

$$\left(\frac{\partial x_i^{(m)}}{\partial c_{ji}^{(m)}}\right) = \frac{\partial}{\partial c_{ji}^{(m)}}\left(x_i^{(m)} = \sum_j w_{ji}^{(m)} o_j^{(m+1)}\right) = \frac{\partial}{\partial c_{ji}^{(m)}}\left(c_{ji}^{(m)} + k\right) = 1 \tag{37}$$

$$\frac{\partial E}{\partial c_{ji}^{(0)}} = \left(o_i^{(0)} - t_i\right)\left(o_i^{(0)}\left(1 - o_i^{(0)}\right)\right) \tag{38}$$

$$= \underbrace{\left(o_i^{(0)} - t_i\right)\sigma'\left(x_i^{(0)}\right)}_{\text{from (28)}} \tag{39}$$

$$\frac{\partial E}{\partial c_{ji}^{(0)}} = \frac{\partial E}{\partial x_i^{(0)}} \tag{40}$$

Second derivative of the error with respect to a single input contribution $c_{ji}^{(0)}$:

$$\frac{\partial^2 E}{\partial {c_{ji}^{(0)}}^2} = \frac{\partial}{\partial c_{ji}^{(0)}} \left(\underbrace{\frac{\partial E}{\partial c_{ji}^{(0)}} = \left(o_i^{(0)} - t_i\right) \sigma'\left(x_i^{(0)}\right)}_{\text{from (39)}} \right) \tag{41}$$

$$= \frac{\partial}{\partial c_{ji}^{(0)}} \left(\sigma\left(x_i^{(0)}\right) - t_i\right) \sigma'\left(x_i^{(0)}\right) \tag{42}$$

$$= \frac{\partial}{\partial c_{ji}^{(0)}} \left(\sigma\left(\sum_j w_{ji}^{(m)} o_j^{(m+1)}\right) - t_i\right) \sigma'\left(\sum_j w_{ji}^{(m)} o_j^{(m+1)}\right) \tag{43}$$

$$= \frac{\partial}{\partial c_{ji}^{(0)}} \left(\sigma\left(\sum_j c_{ji}^{(0)}\right) - t_i\right) \sigma'\left(\sum_j c_{ji}^{(0)}\right) \tag{44}$$

$$= \frac{\partial}{\partial c_{ji}^{(0)}} \underbrace{\left(\sigma\left(c_{ji}^{(0)} + k\right) - t_i\right)}_{f\left(c_{ji}^{(0)}\right)} \underbrace{\sigma'\left(c_{ji}^{(0)} + k\right)}_{g\left(c_{ji}^{(0)}\right)} \tag{45}$$

$$= f'\left(c_{ji}^{(0)}\right) g\left(c_{ji}^{(0)}\right) + f\left(c_{ji}^{(0)}\right) g'\left(c_{ji}^{(0)}\right) \tag{46}$$

$$= \sigma'\left(c_{ji}^{(0)} + k\right) \sigma'\left(c_{ji}^{(0)} + k\right) + \left(\sigma\left(c_{ji}^{(0)} + k\right) - t_i\right) \sigma''\left(c_{ji}^{(0)} + k\right) \tag{47}$$

$$= \sigma'\left(c_{ji}^{(0)} + k\right)^2 + \left(o_i^{(0)} - t_i\right) \sigma''\left(c_{ji}^{(0)} + k\right) \tag{48}$$

$$\left(c_{ji}^{(0)} + k = \sum_j c_{ji}^{(0)} = \sum_j w_{ji}^{(m)} o_j^{(m+1)} = x_i^{(0)}\right) \tag{49}$$

$$\frac{\partial^2 E}{\partial {c_{ji}^{(0)}}^2} = \underbrace{\left(\sigma'\left(x_i^{(0)}\right)\right)^2 + \left(o_i^{(0)} - t_i\right) \sigma''\left(x_i^{(0)}\right)}_{\text{from (34)}} \tag{50}$$

$$\frac{\partial^2 E}{\partial {c_{ji}^{(0)}}^2} = \frac{\partial^2 E}{\partial {x_i^{(0)}}^2} \tag{51}$$

3.1.1 Summary Of Output Layer Derivatives

$$\frac{\partial E}{\partial o_i^{(0)}} = o_i^{(0)} - t_i \qquad \frac{\partial^2 E}{\partial {o_i^{(0)}}^2} = 1 \tag{52}$$

$$\frac{\partial E}{\partial x_i^{(0)}} = \left(o_i^{(0)} - t_i\right) \sigma'\left(x_i^{(0)}\right) \qquad \frac{\partial^2 E}{\partial {x_i^{(0)}}^2} = \left(\sigma'\left(x_i^{(0)}\right)\right)^2 + \left(o_i^{(0)} - t_i\right) \sigma''\left(x_i^{(0)}\right) \tag{53}$$

$$\frac{\partial E}{\partial c_{ji}^{(0)}} = \frac{\partial E}{\partial x_i^{(0)}} \qquad \frac{\partial^2 E}{\partial {c_{ji}^{(0)}}^2} = \frac{\partial^2 E}{\partial {x_i^{(0)}}^2} \tag{54}$$

3.1.2 HIDDEN LAYER DERIVATIVES

The first derivative of the error with respect to a neuron with output $o_j^{(1)}$ in the first hidden layer, summing over all partial derivative contributions from the output layer:

$$\frac{\partial E}{\partial o_j^{(1)}} = \sum_i \frac{\partial E}{\partial o_i^{(0)}} \frac{\partial o_i^{(0)}}{\partial x_i^{(0)}} \frac{\partial x_i^{(0)}}{\partial c_{ji}^{(0)}} \frac{\partial c_{ji}^{(0)}}{\partial o_j^{(1)}} = \sum_i \underbrace{\left(o_i^{(0)} - t_i\right) \sigma'\left(x_i^{(0)}\right)}_{\text{from (28)}} w_{ji}^{(0)} \tag{55}$$

$$\frac{\partial c_{ji}^{(m)}}{\partial o_j^{(m+1)}} = \frac{\partial}{\partial o_j^{(m+1)}} \left(c_{ji}^{(m)} = w_{ji}^{(m)} o_j^{(m+1)}\right) = w_{ji}^{(m)} \tag{56}$$

$$\frac{\partial E}{\partial o_j^{(1)}} = \sum_i \frac{\partial E}{\partial x_i^{(0)}} w_{ji}^{(0)} \tag{57}$$

The second derivative of the error with respect to a neuron with output $o_j^{(1)}$ in the first hidden layer:

$$\frac{\partial^2 E}{\partial {o_j^{(1)}}^2} = \frac{\partial}{\partial o_j^{(1)}} \frac{\partial E}{\partial o_j^{(1)}} \tag{58}$$

$$= \frac{\partial}{\partial o_j^{(1)}} \sum_i \frac{\partial E}{\partial x_i^{(0)}} w_{ji}^{(0)} \tag{59}$$

$$= \frac{\partial}{\partial o_j^{(1)}} \sum_i \left(o_i^{(0)} - t_i\right) \sigma'\left(x_i^{(0)}\right) w_{ji}^{(0)} \tag{60}$$

$$o_i^{(0)} = \sigma\left(x_i^{(0)}\right) = \sigma\left(\sum_j \left(w_{ji}^{(0)} o_j^{(1)}\right)\right) \tag{61}$$

$$= \frac{\partial}{\partial o_j^{(1)}} \sum_i \underbrace{\left(\sigma\left(\sum_j w_{ji}^{(0)} o_j^{(1)}\right) - t_i\right)}_{f\left(o_j^{(1)}\right)} \underbrace{\sigma'\left(\sum_j w_{ji}^{(0)} o_j^{(1)}\right)}_{g\left(o_j^{(1)}\right)} w_{ji}^{(0)} \tag{62}$$

$$= \sum_i \left(f'\left(o_j^{(1)}\right) g\left(o_j^{(1)}\right) + f\left(o_j^{(1)}\right) g'\left(o_j^{(1)}\right)\right) \tag{63}$$

$$= \sum_i \sigma'\left(\sum_j w_{ji}^{(0)} o_j^{(1)}\right) w_{ji}^{(0)} \; \sigma'\left(\sum_j w_{ji}^{(0)} o_j^{(1)}\right) w_{ji}^{(0)} + \ldots \tag{64}$$

$$\sum_i \left(\sigma\left(\sum_j w_{ji}^{(0)} o_j^{(1)}\right) - t_i\right) \sigma''\left(\sum_j w_{ji}^{(0)} o_j^{(1)}\right) \left(w_{ji}^{(0)}\right)^2 \tag{65}$$

$$= \sum_i \left(\left(\sigma'\left(x_i^{(0)}\right)\right)^2 \left(w_{ji}^{(0)}\right)^2 + \left(o_i^{(0)} - t_i\right) \sigma''\left(x_i^{(0)}\right) \left(w_{ji}^{(0)}\right)^2\right) \tag{66}$$

$$= \sum_i \underbrace{\left(\left(\sigma'\left(x_i^{(0)}\right)\right)^2 + \left(o_i^{(0)} - t_i\right) \sigma''\left(x_i^{(0)}\right)\right)}_{\text{from (34)}} \left(w_{ji}^{(0)}\right)^2 \tag{67}$$

$$\frac{\partial^2 E}{\partial {o_j^{(1)}}^2} = \sum_i \frac{\partial^2 E}{\partial {x_i^{(0)}}^2} \left(w_{ji}^{(0)}\right)^2 \tag{68}$$

Note that the equation in (68) does not depend on the form of $\frac{\partial^2 E}{\partial x_x^{(0)^2}}$, which means we can replace the specific indexes with general ones:

$$\frac{\partial^2 E}{\partial o_j^{(m+1)^2}} = \sum_i \frac{\partial^2 E}{\partial x_i^{(m)^2}} \left(w_{ji}^{(m)} \right)^2 \tag{69}$$

At this point we are beginning to see the recursion in the form of the 2nd derivative terms which can be thought of analogously to the first derivative recursion which is central to the back-propagation algorithm. The formulation above which makes specific reference to layer indexes also works in the general case.

Consider the ith neuron in any layer m with output $o_i^{(m)}$ and input $x_i^{(m)}$. The first and second derivatives of the error E with respect to this neuron's *input* are:

$$\frac{\partial E}{\partial x_i^{(m)}} = \frac{\partial E}{\partial o_i^{(m)}} \frac{\partial o_i^{(m)}}{\partial x_i^{(m)}} \tag{70}$$

$$\frac{\partial^2 E}{\partial x_i^{(m)^2}} = \frac{\partial}{\partial x_i^{(m)}} \frac{\partial E}{\partial x_i^{(m)}} \tag{71}$$

$$= \frac{\partial}{\partial x_i^{(m)}} \left(\frac{\partial E}{\partial o_i^{(m)}} \frac{\partial o_i^{(m)}}{\partial x_i^{(m)}} \right) \tag{72}$$

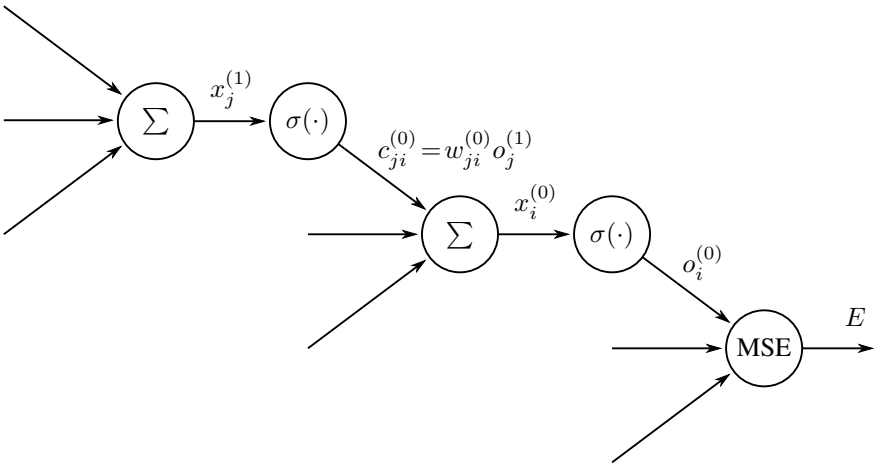
$$= \frac{\partial E}{\partial x_i^{(m)} o_i^{(m)}} \frac{\partial o_i^{(m)}}{\partial x_i^{(m)}} + \frac{\partial E}{\partial o_i^{(m)}} \frac{\partial^2 o_i^{(m)}}{\partial x_i^{(m)^2}} \tag{73}$$

$$= \frac{\partial}{\partial x_i^{(m)}} \frac{\partial E}{\partial o_i^{(m)}} \frac{\partial o_i^{(m)}}{\partial x_i^{(m)}} + \frac{\partial E}{\partial o_i^{(m)}} \sigma'' \left(x_i^{(m)} \right) \tag{74}$$

$$= \frac{\partial}{\partial x_i^{(m)}} \frac{\partial E}{\partial o_i^{(m)}} \frac{\partial o_i^{(m)}}{\partial x_i^{(m)}} + \frac{\partial E}{\partial o_i^{(m)}} \sigma'' \left(x_i^{(m)} \right) \tag{75}$$

$$????? = \frac{\partial^2 E}{\partial o_i^{(m)^2}} \left(\sigma' \left(x_i^{(m)} \right) \right)^2 + \frac{\partial E}{\partial o_i^{(m)}} \sigma'' \left(x_i^{(m)} \right) \tag{76}$$

Both of these terms are easily computable and can be stored as we propagate back from the output of the network to the input. And the derivative of the error E with respect to the output $o_j^{(m+1)}$ in the $m+1$ layer is:
??????????


$\sum$
$x_j^{(1)}$
$\sigma(\cdot)$
$c_{ji}^{(0)} = w_{ji}^{(0)} o_j^{(1)}$
$\sum$
$x_i^{(0)}$
$\sigma(\cdot)$
$o_i^{(0)}$
MSE
E