neural network theory

Charles Snider
(Dated: September 12, 2022)

Define the cost function J relative to some loss function L and the outputs $a_{ij}^{(n)}$ of the neural network as

$$J \equiv \sum_{i,j} L(a_{ij}^{(n)})$$

where the index i refers to a given training example, and the index j refers to the jth element of the prediction of the neural network, or the jth neuron's activation in the output. The superscript (n) refers to the layer of the network in question– the nth layer refers to the output, i.e. the network has $n+1$ layers counting the input, with the input layer being layer 0. We are interested in the derivative of the cost function J with respect to a weight $W_{\mu\nu}^{(k)}$. The weights for each layer are stored as matrices such that

$$a^{(k)} = g(W^{(k)}a^{(k-1)} + B^{(k)})$$

for some activation function g, and for a given training example. The μth row of $W_{\mu\nu}^{(k)}$ is used to calculate the contributions of $a^{(k-1)}$ to $a_{\mu}^{(k)}$– the $\mu\nu$th element $W_{\mu\nu}^{(k)}$ indicates the degree to which $a_{\nu}^{(k-1)}$ contributes to $a_{\mu}^{(k)}$. To perform gradient descent, we are interested in quantities $\partial J/\partial W_{\mu\nu}^{(k)}$. We can calculate these iteratively. We first change notation $W_{\mu\nu}^{(k)} \to W_{\mu\nu}^{(n-k)}$ where k now indicates how many layers *backward* we are from the output, with k ranging from 0 to $n-1$ and $k=0$ referring to the output layer. We begin with $k=0$:

$$\frac{\partial J}{\partial W_{\mu\nu}^{(n)}} = \sum_{ij} \frac{\partial L(a_{ij}^{(n)})}{\partial W_{\mu\nu}^{(n)}} = \sum_{ij} L'(a_{ij}^{(n)}) \frac{\partial a_{ij}^{(n)}}{\partial W_{\mu\nu}^{(n)}}$$

The main thrust of the derivation therefore revolves around calculating $\partial a_{ij}^{(n)}/\partial W_{\mu\nu}^{(n)}$.

I. SIMPLE CASE: $a_{ij}^{(n)} = h(z_{ij}^{(n)})$

For an output activation function h, we have

$$a_{ij}^{(n)} = h\left(\sum_k W_{jk}^{(n)} a_{ik}^{(n-1)} + B_j^{(n)}\right) = h(z_{ij}^{(n)})$$

We have defined $z^{(k)} \equiv W^{(k)}a^{(k-1)} + B^{(k)}$ for convenience. Note that the weights lack a "training" index i because the weights are the same for all training examples. Until later in the derivation, the index i can generally be ignored. The derivative with respect to the weight becomes

$$\begin{aligned}\frac{\partial a_{ij}^{(n)}}{\partial W_{\mu\nu}^{(n)}} &= \frac{\partial}{\partial W_{\mu\nu}^{(n)}} h(z_{ij}^{(n)}) = h'(z_{ij}^{(n)}) \frac{\partial z_{ij}^{(n)}}{\partial W_{\mu\nu}^{(n)}} = h'(z_{ij}^{(n)}) \frac{\partial}{\partial W_{\mu\nu}^{(n)}} \left[\sum_k W_{jk}^{(n)} a_{ik}^{(n-1)} + B_j^{(n)}\right] \\ &= h'(z_{ij}^{(n)}) a_{ik}^{(n-1)} \delta_{j\mu}\delta_{k\nu} = \boxed{h'(z_{ij}^{(n)}) a_{i\nu}^{(n-1)} \delta_{j\mu}}\end{aligned}$$

With this result in hand, we can now look at the derivative of an output $a_{ij}^{(n)}$ relative to a weight of the $(n-1)$th layer:

$$\begin{aligned}\frac{\partial a_{ij}^{(n)}}{\partial W_{\mu\nu}^{(n-1)}} &= \frac{\partial}{\partial W_{\mu\nu}^{(n-1)}} h(z_{ij}^{(n)}) = h'(z_{ij}^{(n)}) \frac{\partial z_{ij}^{(n)}}{\partial W_{\mu\nu}^{(n-1)}} = h'(z_{ij}^{(n)}) \frac{\partial}{\partial W_{\mu\nu}^{(n-1)}} \left[\sum_k W_{jk}^{(n)} a_{ik}^{(n-1)} + B_j^{(n)}\right] \\ &= h'(z_{ij}^{(n)}) \left[\sum_k W_{jk}^{(n)} \frac{\partial a_{ik}^{(n-1)}}{\partial W_{\mu\nu}^{(n-1)}}\right]\end{aligned}$$

The derivative relative to $W_{\mu\nu}^{(n-1)}$ passes into the sum and onto $a_{ik}^{(n-1)}$ since the weights are assumed independent. However, the term $\partial a_{ik}^{(n-1)}/\partial W_{\mu\nu}^{(n-1)}$ is the same as the derivative of the output relative to the nth layer weights, with $n \to n-1$, $j \to k$, and $h \to g$ (replacing the output activation with the hidden layer activation). We can substitute in our first layer result with these changes to find

$$\frac{\partial a_{ij}^{(n)}}{\partial W_{\mu\nu}^{(n-1)}} = h'(z_{ij}^{(n)}) \left[\sum_k W_{jk}^{(n)} g'(z_{ik}^{(n-1)}) a_{i\nu}^{(n-2)} \delta_{k\mu} \right] = \boxed{h'(z_{ij}^{(n)}) \left[W_{j\mu}^{(n)} g'(z_{i\mu}^{(n-1)}) \right] a_{i\nu}^{(n-2)}}$$

We can continue this recursive process. Consider the next layer's derivative:

$$\begin{aligned} \frac{\partial a_{ij}^{(n)}}{\partial W_{\mu\nu}^{(n-2)}} &= \frac{\partial}{\partial W_{\mu\nu}^{(n-2)}} h(z_{ij}^{(n)}) = h'(z_{ij}^{(n)}) \frac{\partial z_{ij}^{(n)}}{\partial W_{\mu\nu}^{(n-2)}} = h'(z_{ij}^{(n)}) \frac{\partial}{\partial W_{\mu\nu}^{(n-1)}} \left[\sum_k W_{jk}^{(n)} a_{ik}^{(n-1)} + B_j^{(n)} \right] \\ &= h'(z_{ij}^{(n)}) \left[\sum_k W_{jk}^{(n)} \frac{\partial a_{ik}^{(n-1)}}{\partial W_{\mu\nu}^{(n-2)}} \right] \end{aligned}$$

The derivative within the sum is the same as the derivative of the output with respect to one layer back, again with $n \to n-1$, $j \to k$, and $h \to g$. Making this substitution, we have

$$\frac{\partial a_{ij}^{(n)}}{\partial W_{\mu\nu}^{(n-2)}} = \boxed{h'(z_{ij}^{(n)}) \left[\sum_k W_{jk}^{(n)} g'(z_{ik}^{(n-1)}) W_{k\mu}^{(n-1)} g'(z_{i\mu}^{(n-2)}) \right] a_{i\nu}^{(n-3)}}$$

By the same process, we can find the next layer's derivatives:

$$\frac{\partial a_{ij}^{(n)}}{\partial W_{\mu\nu}^{(n-3)}} = \boxed{h'(z_{ij}^{(n)}) \left[\sum_{k,\ell} W_{jk}^{(n)} g'(z_{ik}^{(n-1)}) W_{k\ell}^{(n-1)} g'(z_{i\ell}^{(n-2)}) W_{\ell\mu}^{(n-2)} g'(z_{i\mu}^{(n-3)}) \right] a_{i\nu}^{(n-4)}}$$

By this recursive process, we can arrive at a general answer for the derivative with respect to $W_{\mu\nu}^{(n-k)}$

$$\frac{\partial a_{ij}^{(n)}}{\partial W_{\mu\nu}^{(n-k)}} = \boxed{h'(z_{ij}^{(n)}) \left[\sum_{\gamma_1,\ldots,\gamma_{k-1}} W_{j\gamma_1}^{(n)} g'(z_{i\gamma_1}^{(n-1)}) \ldots W_{\gamma_\beta\gamma_{\beta+1}}^{(n-\beta+1)} g'(z_{i\gamma_{\beta+1}}^{(n-\beta)}) \ldots W_{\gamma_{k-1}\mu}^{(n-k+1)} g'(z_{i\mu}^{(n-k)}) \right] a_{i\nu}^{(n-k-1)}}$$

We can make this equation more manageable by recognizing that the sum in brackets represents a matrix multiplication. Define matrices α as follows:

$$\boxed{\alpha_{ijk}^{(\ell)} \equiv W_{jk}^{(\ell)} g'(z_{ik}^{(\ell-1)})}$$

The derivative then becomes

$$\frac{\partial a_{ij}^{(n)}}{\partial W_{\mu\nu}^{(n-k)}} = h'(z_{ij}^{(n)}) \left[\sum_{\gamma_1,\ldots,\gamma_{k-1}} \alpha_{ij\gamma_1}^{(n)} \ldots \alpha_{i\gamma_\beta\gamma_{\beta+1}}^{(n-\beta+1)} \ldots \alpha_{i\gamma_{k-1}\mu}^{(n-k+1)} \right] a_{i\nu}^{(n-k-1)}$$

The sum is simply the $j\mu$th element of the matrix multiplication of the αs, ignoring the training index i (i.e. perform this multiplication for each training example independently). We can once again define a matrix Λ such that

$$\Lambda_{ijk}^{(n-\ell)} \equiv \left(\prod_{\gamma=0}^{k-1} \alpha^{(n-\gamma)} \right)_{ijk}$$

with the understanding that the product occurs over the slices of α and in a manner such that if $p > q$, $\alpha^{(p)}$ occurs to the left of $\alpha^{(q)}$ in the product, i.e. the product is ordered. In "MATLAB notation", where $\alpha(i,:,:)$ denotes the ith slice of α, we can write this

$$\boxed{\Lambda^{(n-\ell)}(i,:,:) \equiv \prod_{\gamma=0}^{k-1} \left[\alpha^{(n-\gamma)}(i,:,:) \right]}$$

With this definition, we can write the derivative as

$$\frac{\partial a_{ij}^{(n)}}{\partial W_{\mu\nu}^{(n-k)}} = h'(z_{ij}^{(n)})\Lambda_{ij\mu}^{(n-k)} a_{i\nu}^{(n-k-1)}$$

With this shorthand established, we can substitute this expression into the original cost function derivative:

$$\frac{\partial J}{\partial W_{\mu\nu}^{(n-k)}} = \sum_{ij} L'(a_{ij}^{(n)})h'(z_{ij}^{(n)})\Lambda_{ij\mu}^{(n-k)} a_{i\nu}^{(n-k-1)}$$

Splitting up the sums and regrouping can give some additional insight:

$$\frac{\partial J}{\partial W_{\mu\nu}^{(n-k)}} = \sum_{i} \left[\sum_{j} L'(a_{ij}^{(n)})h'(z_{ij}^{(n)})\Lambda_{ij\mu}^{(n-k)} \right] a_{i\nu}^{(n-k-1)}$$

For clarity, define a matrix η such that

$$\boxed{\eta_{ij} \equiv L'(a_{ij}^{(n)})h'(z_{ij}^{(n)})}$$

With this substitution, we have

$$\frac{\partial J}{\partial W_{\mu\nu}^{(n-k)}} = \sum_{i} \left[\sum_{j} \eta_{ij}\Lambda_{ij\mu}^{(n-k)} \right] a_{i\nu}^{(n-k-1)}$$

The term in brackets is now clearly a matrix-vector product, when ignoring the training index i. Defining a matrix $\Lambda^{T,(\ell)}$ such that $\Lambda_{ijk}^{(\ell)} = \Lambda_{ikj}^{T,(\ell)}$, we have

$$\frac{\partial J}{\partial W_{\mu\nu}^{(n-k)}} = \sum_{i} \left[\sum_{j} \eta_{ij}\Lambda_{i\mu j}^{T,(n-k)} \right] a_{i\nu}^{(n-k-1)} \equiv \sum_{i} \Delta_{i\mu}^{(n-k)} a_{i\nu}^{(n-k-1)}$$

where $\Delta^{(n-k)}$, in "MATLAB notation", is defined by

$$\boxed{\Delta^{(n-k)}(i,:) \equiv \Lambda^{T,(n-k)}(i,:,:)\eta(i,:)}$$

Taking a transpose of Δ, we can write the derivative of J as one final matrix multiplication:

$$\boxed{\frac{\partial J}{\partial W_{\mu\nu}^{(n-k)}} = \left[\left(\Delta^{(n-k)}\right)^T a^{(n-k-1)} \right]_{\mu\nu}}$$

for $a^{(n-k-1)}$ the $m \times d$ matrix containing the d activations of the $(n-k-1)$th layer for each of the m training examples.

II. SPECIAL CASE: SOFTMAX OUTPUT ACTIVATION

When the output activation function is the softmax function, a given output activation $a_{ij}^{(n)}$ is now dependent on all values of $z_{i\mu}^{(n)}$ and not just $j = \mu$. The derivative of the output activation with regards to a given weight $W_{\mu\nu}^{(n)}$ no longer picks up the δ function and must be reconsidered. The softmax function is

$$a_{ij}^{(n)} = h(z_i^{(n)}) = \frac{e^{z_{ij}^{(n)}}}{\sum_k e^{z_{ik}^{(n)}}}$$

with $z_i^{(n)}$ denoting the vector whose jth element is $z_{ij}^{(n)}$. The derivative of this function with regards to a given weight $W_{\mu\nu}^{(n-k)}$ is

$$\frac{\partial a_{ij}^{(n)}}{\partial W_{\mu\nu}^{(n-k)}} = \boxed{a_{ij}^{(n)} \frac{\partial z_{ij}^{(n)}}{\partial W_{\mu\nu}^{(n-k)}} - a_{ij}^{(n)} \sum_{\ell} \left[a_{i\ell}^{(n)} \frac{\partial z_{i\ell}^{(n)}}{\partial W_{\mu\nu}^{(n-k)}} \right]}$$

With this adjustment, we now have to re-consider the derivation from the previous section. Beginning with $k=0$, we have

$$\frac{\partial a_{ij}^{(n)}}{\partial W_{\mu\nu}^{(n)}} = a_{ij}^{(n)} a_{i\nu}^{(n-1)} \delta_{j\mu} - a_{ij}^{(n)} \sum_{\ell} \left[a_{i\ell}^{(n)} a_{i\nu}^{(n-1)} \delta_{\ell\mu} \right] = \boxed{a_{ij}^{(n)} \left[\delta_{j\mu} - a_{i\mu}^{(n)} \right] a_{i\nu}^{(n-1)}}$$

We now proceed as before, recursively building up each derivative. Looking at the derivative of a term one layer back, we have

$$\frac{\partial a_{ij}^{(n)}}{\partial W_{\mu\nu}^{(n-1)}} = a_{ij}^{(n)} \frac{\partial z_{ij}^{(n)}}{\partial W_{\mu\nu}^{(n-1)}} - a_{ij}^{(n)} \sum_{\ell} \left[a_{i\ell}^{(n)} \frac{\partial z_{i\ell}^{(n)}}{\partial W_{\mu\nu}^{(n-1)}} \right]$$

Plugging in $z_{ij}^{(n)} = \sum_k W_{jk}^{(n)} a_{ik}^{(n-1)} + B_j^{(n)}$, we arrive at

$$\frac{\partial a_{ij}^{(n)}}{\partial W_{\mu\nu}^{(n-1)}} = a_{ij}^{(n)} \left[\sum_k W_{jk}^{(n)} \frac{\partial a_{ik}^{(n-1)}}{\partial W_{\mu\nu}^{(n-1)}} \right] - a_{ij}^{(n)} \sum_{\ell} \left[a_{i\ell}^{(n)} \left[\sum_k W_{\ell k}^{(n)} \frac{\partial a_{ik}^{(n-1)}}{\partial W_{\mu\nu}^{(n-1)}} \right] \right]$$

Here, we make the assumption that only the output layer uses the softmax activation function, and that the interior layers use some activation function (e.g. reLU) such that $a_{ij}^{(n)} = h(z_{ij}^{(n)})$ rather than $h(z_i^{(n)})$. Therefore, for the derivatives of $a_{ij}^{(n-1)}$ relative to a weight $W_{\mu\nu}^{(n-1)}$ we can use the result from the previous section. Plugging this in, we get

$$\frac{\partial a_{ij}^{(n)}}{\partial W_{\mu\nu}^{(n-1)}} = \boxed{a_{ij}^{(n)} \left[W_{j\mu}^{(n)} g'(z_{i\mu}^{(n-1)}) - \sum_{\ell} a_{i\ell}^{(n)} \left(W_{\ell\mu}^{(n)} g'(z_{i\mu}^{(n-1)}) \right) \right] a_{i\nu}^{(n-2)}}$$

We do the same for two layers back, which gives us

$$\begin{aligned} \frac{\partial a_{ij}^{(n)}}{\partial W_{\mu\nu}^{(n-2)}} &= a_{ij}^{(n)} \left[\sum_k W_{jk}^{(n)} \frac{\partial a_{ik}^{(n-1)}}{\partial W_{\mu\nu}^{(n-2)}} \right] - a_{ij}^{(n)} \sum_{\ell} \left[a_{i\ell}^{(n)} \left[\sum_k W_{\ell k}^{(n)} \frac{\partial a_{ik}^{(n-1)}}{\partial W_{\mu\nu}^{(n-2)}} \right] \right] \\ &= a_{ij}^{(n)} \left[\sum_k \left[W_{jk}^{(n)} g'(z_{ik}^{(n-1)}) W_{k\mu}^{(n-1)} g'(z_{i\mu}^{(n-2)}) \right] - \sum_{\ell} a_{i\ell}^{(n)} \left(\sum_k \left[W_{\ell k}^{(n)} g'(z_{ik}^{(n-1)}) W_{k\mu}^{(n-1)} g'(z_{i\mu}^{(n-2)}) \right] \right) \right] a_{i\nu}^{(n-3)} \end{aligned}$$

Here, we begin to see a pattern. Using our definition of the matrix Λ from the previous section, we can infer that

$$\frac{\partial a_{ij}^{(n)}}{\partial W_{\mu\nu}^{(n-k)}} = \boxed{a_{ij}^{(n)} \left[\Lambda_{ij\mu}^{(n-k)} - \sum_{\ell} a_{i\ell}^{(n)} \Lambda_{i\ell\mu}^{(n-k)} \right] a_{i\nu}^{(n-k-1)}}$$

If we replace the exponential term with a δ function $\delta_{j\ell}$, this expression reduces to the expression from the previous section. We can therefore use our previous results, and approach, with an adjusted matrix

$$\Lambda_{ij\mu}^{(n-k)} \quad \longrightarrow \quad \Gamma_{ij\mu}^{(n-k)} = \Lambda_{ij\mu}^{(n-k)} - \sum_{\ell} a_{i\ell}^{(n)} \Lambda_{i\ell\mu}^{(n-k)} \qquad \text{and} \qquad h'(z_{ij}^{(n)}) \quad \longrightarrow \quad a_{ij}^{(n)}$$