3Blue1Brown AI Notes

Wednesday, November 6, 2024 10:06 PM

BACKPROPOGATION

Simple neural network example

$$\left(a^{(L-3)}\right)\xrightarrow[b^{(L-2)}]{w^{(L-2)}}\left(a^{(L-2)}\right)\xrightarrow[b^{(L-1)}]{w^{(L-1)}}\left(a^{(L-1)}\right)\xrightarrow[b^{(L)}]{w^{(L)}}\left(a^{(L)}\right)\qquad (y)$$

desired output

$$Cost = C_0 = (a^{(L)} - y)^2 \qquad z^{(L)} = w^{(L)}a^{(L-1)} + b^{(L)}, \quad a^{(L)} = \sigma(z^{(L)})$$

$$\Delta w^{(L)} \to \Delta z^{(L)} \to \Delta a^{(L)} \to \Delta C_0$$

$$\frac{\partial z^{(L)}}{\partial w^{(L)}} \times \frac{\partial a^{(L)}}{\partial z^{(L)}} \times \frac{\partial C_0}{\partial a^{(L)}} = \frac{\partial C_0}{\partial w^{(L)}} \quad \text{(chain rule)}$$

$$\frac{\partial z^{(L)}}{\partial w^{(L)}} = \frac{\partial}{\partial w^{(L)}}\left(w^{(L)}a^{(L-1)} + b^{(L)}\right) = a^{(L-1)}$$

$$\frac{\partial a^{(L)}}{\partial z^{(L)}} = \sigma'(z^{(L)})$$

$$\frac{\partial C_0}{\partial a^{(L)}} = \frac{\partial}{\partial a^{(L)}}(a^{(L)} - y)^2 = 2(a^{(L)} - y)$$

$$\frac{\partial C_0}{\partial w^{(L)}} = 2a^{(L-1)}\sigma'(z^{(L)})(a^{(L)} - y)$$

$$\frac{\partial C_0}{\partial b^{(L)}} = 2\sigma'(z^{(L)})(a^{(L)} - y) \qquad \frac{\partial C_0}{\partial a^{(L-1)}} = 2w^{(L)}\sigma'(z^{(L)})(a^{(L)} - y)$$

- We need to average all $\frac{\partial C_k}{\partial w^{(L)}}$ and $\frac{\partial C_k}{\partial b^{(L)}}$ from each training example to obtain $\frac{\partial C}{\partial w^{(L)}}$ and $\frac{\partial C}{\partial b^{(L)}}$, the partial derivatives of the full cost function:

$$\frac{\partial C}{\partial w^{(L)}} = \frac{1}{n}\sum_{n=0}^{n-1}\frac{\partial C_n}{\partial w^{(L)}} \qquad \frac{\partial C}{\partial b^{(L)}} = \frac{1}{n}\sum_{n=0}^{n-1}\frac{\partial C_n}{\partial b^{(L)}}$$

$$\nabla C = \begin{bmatrix} \partial C/\partial w^{(1)} \\ \partial C/\partial b^{(1)} \\ \vdots \\ \partial C/\partial w^{(L)} \\ \partial C/\partial b^{(L)} \end{bmatrix} \longleftarrow \text{gradient}$$

Repeat process recursively.

Multineuron layer example

$$\left(a_0^{(L-1)}\right),\ \left(a_1^{(L-1)}\right),\ \left(a_2^{(L-1)}\right) \xrightarrow{w_{00}^{(L)}} \left(a_0^{(L)}\right),\ \left(a_1^{(L)}\right)$$

Let each neuron in a given layer be denoted by a 0-based index.

Index layer L neurons with j and L-1 with k.

$$C_0 = \sum_{j=0}^{n_L-1}(a_j^{(L)} - y_j)^2 \qquad z_j^{(L)} = \sum_{k=0}^{n_{(L-1)}-1} w_{jk}^{(L)} a_k^{(L-1)} + b_j^{(L)}$$

- Chain rule mostly stays the same
- Only difference is:

$$\frac{\partial C_0}{\partial a_k^{(L-1)}} = \sum_{j=0}^{n_L-1}\frac{\partial C_0}{\partial a_j^{(L)}}\cdot\frac{\partial a_j^{(L)}}{\partial z_j^{(L)}}\cdot\frac{\partial z_j^{(L)}}{\partial a_k^{(L-1)}}$$

since a single neuron from the previous layer influences multiple neurons in L.

$$\frac{\partial C_0}{\partial w_{jk}^{(L)}} = \frac{\partial C_0}{\partial a_j^{(L)}} \times \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \times \frac{\partial z_j^{(L)}}{\partial w_{jk}^{(L)}} = 2(a_j^{(L)} - y_j)\cdot\sigma'(z_j^{(L)})\cdot a_k^{(L-1)}$$

$$\frac{\partial C_0}{\partial b_j^{(L)}} = \frac{\partial C_0}{\partial a_j^{(L)}} \times \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \times \frac{\partial z_j^{(L)}}{\partial b_j^{(L)}} = 2(a_j^{(L)} - y_j)\cdot\sigma'(z_j^{(L)})$$

$$\frac{\partial C_0}{\partial a_k^{(L-1)}} = \sum_{j=0}^{n_L-1}\frac{\partial C_0}{\partial a_j^{(L)}} \times \frac{\partial a_j^{(L)}}{\partial z_j^{(L)}} \times \frac{\partial z_j^{(L)}}{\partial a_k^{(L-1)}} = \sum_{j=0}^{n_L-1} 2(a_j^{(L)} - y_j)\cdot\sigma'(z_j^{(L)})\cdot w_{jk}^{(L)}$$

- If $\sigma(x) := ReLU(x)$,

$$\sigma'(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x \le 0 \end{cases}$$

- If $\sigma(x) :=$ sigmoid(x),

$$\sigma'(x) = \frac{d}{dx}\left(\frac{1}{1+e^{-x}}\right) = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1}{1+e^{-x}}\cdot\frac{e^{-x}}{1+e^{-x}}$$

$$= \sigma(x)\cdot\frac{1-1+e^{-x}}{1+e^{-x}} = \sigma(x)\cdot\left(\frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}}\right)$$

$$= \sigma(x)\cdot(1-\sigma(x)) = a(1-a)$$

- If $\sigma(\vec{x}) :=$ softmax$(\vec{x})$ where $\vec{x}$ is a vector,

$$\sigma_i = \frac{e^{x_i}}{\sum e^{x_n}} \to \frac{\partial\sigma_i}{\partial x_j} = \frac{\frac{\partial}{\partial x_j}(e^{x_i})\cdot\sum e^{x_n} - e^{x_i}\cdot\frac{\partial}{\partial x_j}(\sum e^{x_n})}{(\sum e^{x_n})^2}$$

Consider $i \ne j$:

$$\frac{\partial\sigma_i}{\partial x_j} = \frac{\cancel{0\cdot\sum e^{x_n}} - e^{x_i}e^{x_j}}{(\sum e^{x_n})^2} = -\left(\frac{e^{x_i}}{\sum e^{x_n}}\right)\left(\frac{e^{x_j}}{\sum e^{x_n}}\right) = \boxed{-\sigma_i\sigma_j}$$

Consider $i = j$:

$$\frac{\partial\sigma_i}{\partial x_j} = \frac{e^{x_i}\sum e^{x_n} - e^{x_i}e^{x_j}}{(\sum e^{x_n})^2} = \frac{e^{x_i}(\sum e^{x_n} - e^{x_j})}{(\sum e^{x_n})^2}$$

$$= \left(\frac{e^{x_i}}{\sum e^{x_n}}\right)\left(\frac{\sum e^{x_n} - e^{x_j}}{\sum e^{x_n}}\right) = \left(\frac{e^{x_i}}{\sum e^{x_n}}\right)\left(1 - \frac{e^{x_j}}{\sum e^{x_n}}\right)$$

$$= \boxed{\sigma_i(1-\sigma_j)}$$

In general,

$$\frac{\partial\sigma_i}{\partial x_j} = \sigma_i(\delta_{ij} - \sigma_j), \text{ where } \delta_{ij} := \begin{cases} 0 & \text{if } i \ne j \\ 1 & \text{if } i = j \end{cases}$$

- The Jacobian of softmax$(\vec{x})$ is

$$J_s = \begin{bmatrix} \partial\sigma_1/\partial x_1 & \cdots & \partial\sigma_1/\partial x_n \\ \partial\sigma_2/\partial x_1 & \ddots & \partial\sigma_2/\partial x_n \\ \vdots & & \vdots \\ \partial\sigma_n/\partial x_1 & \cdots & \partial\sigma_n/\partial x_n \end{bmatrix}$$

where n is the number of components in $\vec{x}$.

$$J_{s_{ij}} = \frac{\partial\sigma_i}{\partial x_j} = \sigma_i(\delta_{ij} - \sigma_j) \longrightarrow \begin{bmatrix} \sigma_1(1-\sigma_1) & \sigma_1\sigma_2 & \\ \sigma_2\sigma_1 & \sigma_2(1-\sigma_2) & \\ & & \ddots \\ & & & \sigma_n(1-\sigma_n) \end{bmatrix}$$

Let $a(w) = \sigma(z(w))$ represent the activation of the output layer w.r.t its weights.

$$\frac{\partial a}{\partial w} = \frac{\partial a}{\partial z}\cdot\frac{\partial z}{\partial w} = J_s\cdot\frac{\partial z}{\partial w}$$

Recall: $z_j^{(L)} = \sum_k w_{jk}a_k^{(L-1)} + b_j^{(L)} \to \frac{\partial z_j^{(L)}}{\partial w_{jk}} = a_k^{(L-1)}$

$$\to \frac{\partial z^{(L)}}{\partial[j\text{th row of } w]} = a^{(L-1)}$$

Considering the entire matrix w,

$$\frac{\partial z_j^{(L)}}{\partial w_{ik}} = \begin{cases} a_k^{(L-1)} & \text{if } i = j \\ 0 & \text{if } i \ne j \end{cases}$$

$$\frac{\partial a_t^{(L)}}{\partial w_{ij}} = \sum_k\frac{\partial a_t^{(L)}}{\partial z_k^{(L)}}\cdot\frac{\partial z_k^{(L)}}{\partial w_{ij}} = \frac{\partial a_t^{(L)}}{\partial z_i^{(L)}}\cdot a_j^{(L-1)} = \sigma_t(\delta_{ti} - \sigma_i)\cdot a_j^{(L-1)}$$

(only nonzero for $i = k$)

$$\frac{\partial C_0}{\partial w_{ij}} = \frac{\partial C_0}{\partial a_t^{(L)}}\cdot\frac{\partial a_t^{(L)}}{\partial w_{ij}}$$

$C_0 = -\sum_k y_k \log a_k^{(L)}$; we use one-hot encoding so y_k is only $=1$ for one index t

$$C_0 = -\log a_t^{(L)}$$

$$\frac{\partial C_0}{\partial a_t^{(L)}} = -\frac{1}{a_t^{(L)}} \to \frac{\partial C_0}{\partial a_j^{(L)}} = -\frac{\delta_{jt}}{a_j^{(L)}}$$

$$\to \frac{\partial C_0}{\partial w_{ij}} = -\frac{1}{a_t^{(L)}}\cdot\sigma_t(\delta_{ti} - \sigma_i)\,a_j^{(L-1)}$$

$$= -\frac{1}{\cancel{a_t}}\cdot\cancel{\sigma_t}(\delta_{ti} - \sigma_i)\,a_j^{(L-1)}$$

$$= \boxed{(\sigma_i - \delta_{ti})\,a_j^{(L-1)}}$$

$$= (a_i^{(L)} - \delta_{ti})\,a_j^{(L-1)}$$

$$\frac{\partial z_j^{(L)}}{\partial b_i^{(L)}} = \delta_{ij}$$

$$\frac{\partial a_j^{(L)}}{\partial b_i^{(L)}} = \sum_k\frac{\partial a_j^{(L)}}{\partial z_k^{(L)}}\cdot\frac{\partial z_k^{(L)}}{\partial b_i^{(L)}} = \frac{\partial a_j^{(L)}}{\partial z_i^{(L)}} = a_j^{(L)}(\delta_{ij} - a_i^{(L)})$$

$$\frac{\partial C_0}{\partial b_i^{(L)}} = \frac{\partial C_0}{\partial a_j^{(L)}}\cdot\frac{\partial a_j^{(L)}}{\partial b_i^{(L)}} = -\frac{1}{\cancel{a_j^{(L)}}}\cdot\cancel{a_j^{(L)}}(\delta_{ij} - a_i^{(L)}) = \boxed{a_i^{(L)} - \delta_{ij}}$$

$$\frac{\partial C_0}{\partial z_j^{(L)}} = \sum_k\frac{\partial C_0}{\partial a_k^{(L)}}\cdot\frac{\partial a_k^{(L)}}{\partial z_j^{(L)}} = \frac{\partial C_0}{\partial a_t^{(L)}}\cdot\frac{\partial a_t^{(L)}}{\partial z_j^{(L)}}$$

$$= -\frac{1}{\cancel{a_t^{(L)}}}\cdot\cancel{a_t^{(L)}}(\delta_{tj} - a_j^{(L)}) = a_j^{(L)} - \delta_{tj} = a_j^{(L)} - y_j$$

$$\frac{\partial C_0}{\partial z} = a^{(L)} - y = \nabla_z C$$

$$\begin{bmatrix} \frac{\partial z_1^{(L)}}{\partial w_{11}}, & \frac{\partial z_1^{(L)}}{\partial w_{12}}, & \cdots & \frac{\partial z_1^{(L)}}{\partial w_{1k}} \\ \frac{\partial z_2^{(L)}}{\partial w_{21}}, & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \frac{\partial z_j^{(L)}}{\partial w_{j1}}, & \cdots & \cdots & \frac{\partial z_j^{(L)}}{\partial w_{jk}} \end{bmatrix}\begin{matrix} \cdot\frac{\partial C_0}{\partial z_1} \\ \cdot\frac{\partial C_0}{\partial z_2} \\ \vdots \\ \cdot\frac{\partial C_0}{\partial z_j} \end{matrix} \longrightarrow \frac{\partial C_0}{\partial w} = \begin{bmatrix} \frac{\partial C_0}{\partial z_1} \\ \frac{\partial C_0}{\partial z_2} \\ \vdots \\ \frac{\partial C_0}{\partial z_j} \end{bmatrix}\cdot\left[a_1^{(L-1)}, a_2^{(L-1)}, \ldots a_k^{(L-1)}\right] = \nabla_z C_0\cdot(a^{(L-1)})^T$$

$$\begin{matrix} \| & \| & & \| \\ a_1^{(L-1)} & a_2^{(L-1)} & \cdots & a_k^{(L-1)} \end{matrix}$$

$$\begin{bmatrix} \frac{\partial z_1^{(L)}}{\partial b_1^{(L)}} \\ \frac{\partial z_2^{(L)}}{\partial b_2^{(L)}} \\ \vdots \\ \frac{\partial z_j^{(L)}}{\partial b_j^{(L)}} \end{bmatrix}\begin{matrix} \cdot\frac{\partial C_0}{\partial z_1} \\ \cdot\frac{\partial C_0}{\partial z_2} \\ \vdots \\ \cdot\frac{\partial C_0}{\partial z_j} \end{matrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}\begin{matrix} \cdot\frac{\partial C_0}{\partial z_1} \\ \cdot\frac{\partial C_0}{\partial z_2} \\ \vdots \\ \cdot\frac{\partial C_0}{\partial z_j} \end{matrix} = \begin{pmatrix} \frac{\partial C_0}{\partial z_1} \\ \frac{\partial C_0}{\partial z_2} \\ \vdots \\ \frac{\partial C_0}{\partial z_j} \end{pmatrix} \longrightarrow \frac{\partial C_0}{\partial b} = \frac{\partial C_0}{\partial z}$$

$$\frac{\partial C_0}{\partial a^{(L-1)}} = \frac{\partial C_0}{\partial z^{(L)}}\cdot\frac{\partial z^{(L)}}{\partial a^{(L-1)}} = W^T\cdot\nabla_z C_0$$

$$\begin{bmatrix} \frac{\partial C_0}{\partial a_1^{(L-1)}} \\ \frac{\partial C_0}{\partial a_2^{(L-1)}} \\ \vdots \\ \frac{\partial C_0}{\partial a_k^{(L-1)}} \end{bmatrix} \qquad \begin{bmatrix} \frac{\partial z_1^{(L)}}{\partial a_1^{(L-1)}} & \frac{\partial z_1^{(L)}}{\partial a_2^{(L-1)}} & \cdots & \frac{\partial z_1^{(L)}}{\partial a_k^{(L-1)}} \\ \frac{\partial z_2^{(L)}}{\partial a_1^{(L-1)}} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \frac{\partial z_j^{(L)}}{\partial a_1^{(L-1)}} & \cdots & \cdots & \frac{\partial z_j^{(L)}}{\partial a_k^{(L-1)}} \end{bmatrix} = \begin{bmatrix} w_{11}^{(L)}, & w_{12}^{(L)}, & \cdots & w_{1k}^{(L)} \\ w_{21}^{(L)}, & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ w_{j1}^{(L)}, & \cdots & \cdots & w_{jk}^{(L)} \end{bmatrix} = W$$

$$\frac{\partial C_0}{\partial w^{(L-1)}} = \frac{\partial C_0}{\partial a^{(L-1)}}\cdot\frac{\partial a^{(L-1)}}{\partial z^{(L-1)}}\cdot\frac{\partial z^{(L-1)}}{\partial w^{(L-1)}} = \frac{\partial C_0}{\partial a^{(L-1)}}\odot\frac{\partial a^{(L-1)}}{\partial z^{(L-1)}}\cdot(a^{(L-2)})^T$$

$$\begin{bmatrix} \frac{\partial C_0}{\partial a_1^{(L-1)}} \\ \frac{\partial C_0}{\partial a_2^{(L-1)}} \\ \vdots \\ \frac{\partial C_0}{\partial a_k^{(L-1)}} \end{bmatrix} \odot \begin{bmatrix} \frac{\partial a_1^{(L-1)}}{\partial z_1^{(L-1)}} \\ \frac{\partial a_2^{(L-1)}}{\partial z_2^{(L-1)}} \\ \vdots \\ \frac{\partial a_k^{(L-1)}}{\partial a_k^{(L-1)}} \end{bmatrix} \qquad \begin{bmatrix} \frac{\partial z_1^{(L)}}{\partial w_{11}}, & \frac{\partial z_1^{(L)}}{\partial w_{12}}, & \cdots & \frac{\partial z_1^{(L)}}{\partial w_{1k}} \\ \frac{\partial z_2^{(L)}}{\partial w_{21}}, & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \frac{\partial z_j^{(L)}}{\partial w_{j1}}, & \cdots & \cdots & \frac{\partial z_j^{(L)}}{\partial w_{jk}} \end{bmatrix}$$

$$\begin{matrix} \| & \| & & \| \\ a_1^{(L-2)} & a_2^{(L-2)} & \cdots & a_k^{(L-2)} \end{matrix}$$

$$\frac{\partial C_0}{\partial a^{(L-2)}} = \frac{\partial C_0}{\partial z^{(L-1)}}\cdot\frac{\partial z^{(L-1)}}{\partial a^{(L-2)}} = W^T\cdot\nabla_{z^{(L-1)}} C_0$$

$$\begin{bmatrix} \frac{\partial z_1}{\partial a_1} & \frac{\partial z_1}{\partial a_2} & \cdots & \frac{\partial z_1}{\partial a_k} \\ \frac{\partial z_2}{\partial a_1} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \frac{\partial z_j}{\partial a_1} & \cdots & \cdots & \frac{\partial z_j}{\partial a_k} \end{bmatrix} = \begin{bmatrix} W_{11}, & W_{12}, & \cdots & W_{1k} \\ W_{21} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ W_{j1} & \cdots & \cdots & W_{jk} \end{bmatrix} = W^{(L-1)}$$