Basic Knowledge

$$\underline{C} = \underline{A}\,\underline{B} \longrightarrow \underline{C}^T = \underline{B}^T \underline{A}^T, \quad \underline{C}^{-1} = \underline{B}^{-1} \underline{A}^{-1}$$

$$y \in \mathbb{R}^m, x \in \mathbb{R}^n \quad \frac{\partial \underline{y}}{\partial \underline{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots \\ \vdots & \ddots & \\ & & \frac{\partial y_m}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

$$\underline{y} = \underline{A}\,\underline{x} \quad \frac{\partial \underline{y}}{\partial \underline{x}} = \underline{A} \quad \frac{\partial \underline{y}}{\partial \underline{z}} = \underline{A} \frac{\partial \underline{x}}{\partial \underline{z}}$$

Key Idea:

General tensor $\underline{X}$, elements X_i with i: tuple of indexes

$$\underline{Y} = \underline{g}(\underline{X}), \quad z = f(\underline{Y}) \Rightarrow \nabla_{\underline{X}} z = \sum_j (\nabla_{\underline{X}} Y_j) \frac{\partial z}{\partial Y_j}$$

↳ again, tuple of indeces

Backprop:

$$\underline{G} = \underbrace{\begin{bmatrix} & & \\ & & \\ & & \end{bmatrix}}_{n^o \text{ wide},\ m \text{ tall}} \quad \begin{matrix} \frac{\partial L}{\partial \underline{o}_1^l} \\ \vdots \\ \frac{\partial L}{\partial \underline{o}_m^l} \end{matrix}$$

n^o: number of outputs (units),

m: batch size

Every row j is gradient with respect to the j-th sample.

First step: undo activation.

$$\underset{(m \times n^o)}{\underline{G}} \longleftarrow \underline{G} \odot \begin{bmatrix} \underline{f}'(\underline{z}_1^l) \\ \vdots \\ \underline{f}'(\underline{z}_m^l) \end{bmatrix}$$

where $\underline{o}_j^l = \underline{f}(\underline{z}_j^l)$

↓

activation function element-wise

Next step: $\nabla_{\underline{W}} J$ and $\nabla_{\underline{b}} J$.

→ sample: $\underline{z}_j^l = \underline{W}^T \underline{o}_j^{l-1} + \underline{b}^l$ (vectors are rows)

$\underline{z}_j^l = \underline{W}\,\underline{o}_j^{l-1} + \underline{b}^l$ (vectors are columns)

$(\underline{O}^{l-1})$ → whole batch

→ Whole batch:

$$\underline{Z}^l = \underset{(m \times n^o)}{\begin{bmatrix} \underline{z}_1^l \\ \vdots \\ \underline{z}_m^l \end{bmatrix}} = \underset{(m \times n^i)}{\begin{bmatrix} \underline{o}_1^{l-1} \\ \vdots \\ \underline{o}_m^{l-1} \end{bmatrix}} \underset{(n^i \times n^o)}{\begin{bmatrix} \underline{W}^l \end{bmatrix}} + \underset{(1 \times n^o)}{\begin{bmatrix} \underline{b}^l \end{bmatrix}}$$

broadcasting

n^i: number of inputs, W_{ij}^l = weight for input i of unit j at layer l

So we have,

$$\underline{G} = \nabla_{\underline{Z}^\ell} \mathcal{L} \quad (m \times n^o) \qquad G_{ij} = \frac{\partial \mathcal{L}}{\partial Z^\ell_{ij}} \rightarrow$$ component j of sample i of layer ℓ

At this point we apply the key idea:

$$\nabla_{\underline{W}^\ell} \mathcal{L} = \sum_{i,j} \frac{\partial Z^\ell_{ij}}{\partial \underline{W}^\ell} \frac{\partial \mathcal{L}}{\partial Z^\ell_{ij}} \Rightarrow (\nabla_{\underline{W}^\ell} \mathcal{L})_{hk} = \sum_{i,j} \frac{\partial Z^\ell_{ij}}{\partial W^\ell_{hk}} G_{ij}$$

Considering that $Z^\ell_{ij} = O^{\ell-1}_{i1} W^\ell_{1j} + O^{\ell-1}_{i2} W^\ell_{2j} + \dots + O^{\ell-1}_{in^i} W^\ell_{n^i j} + b^\ell_i$,

$$\frac{\partial Z^\ell_{ij}}{\partial W^\ell_{hk}} = \begin{cases} \varnothing & \text{if } j \neq k \\ O^{\ell-1}_{ih} & \text{if } j = k \end{cases}$$

So for every W_{hk}, there are m non-zero values,

$$\frac{\partial Z_{1k}}{\partial W_{hk}}, \frac{\partial Z_{2k}}{\partial W_{hk}}, \dots, \frac{\partial Z_{mk}}{\partial W_{hk}} \quad [i = 1, \dots, m \,;\, j = k]$$

Which is correct: one for every sample!

$$(\nabla_{\underline{W}^\ell} \mathcal{L})_{hk} = O^{\ell-1}_{1h} G_{1k} + O^{\ell-1}_{2h} G_{2k} + \dots + O^{\ell-1}_{mh} G_{mk}$$

↳ Column h of $\underline{O}^{\ell-1}$ and k of $\underline{G}$

So we obtain, equivalently:

$$\underset{(n^i \times n^o)}{\nabla_{\underline{W}^\ell} \mathcal{L}} = \underset{(n^i \times m)}{\underline{O}^{\ell-1\,T}} \; \underset{(m \times n^o)}{\underline{G}}$$

Adding regularization, $\underline{\theta}$ = all parameters

$$\Rightarrow A) \quad \nabla_{\underline{W}^\ell} J = \underline{O}^{\ell-1\,T} \underline{G} + \lambda \nabla_{\underline{W}^\ell} \Omega(\underline{\theta})$$

For bias, first notice that the correct way to write the equation for $\underline{Z}^\ell$ is:

$$\underline{Z}^\ell = \underline{O}^{\ell-1}\,\underline{W}^\ell + \underbrace{\left.\begin{bmatrix} \underline{b}^\ell \\ \underline{b}^\ell \\ \vdots \\ \underline{b}^\ell \end{bmatrix}\right\} m \text{ times}}_{(m \times n^\ell)} \quad \text{, without broadcasting}$$

So $$\underline{Z}^\ell = \underline{O}^{\ell-1}\,\underline{W}^\ell + \underbrace{\begin{bmatrix} 1 \end{bmatrix}}_{(m\times 1)} \underbrace{[\,b_1^\ell \; b_2^\ell \; \dots \; b_{n^\ell}^\ell\,]}_{(1 \times n^\ell)}$$

Apply key idea

$$\nabla_{\underline{b}^\ell} \mathcal{L} = \sum_{i,j} \frac{\partial Z_{ij}^\ell}{\partial \underline{b}^\ell} \frac{\partial \mathcal{L}}{\partial Z_{ij}^\ell} = \sum_{i,j} \frac{\partial Z_{ij}^\ell}{\partial \underline{b}^\ell} G_{ij}$$

$$(\nabla_{\underline{b}^\ell} \mathcal{L})_k = \sum_{i,j} \frac{\partial Z_{ij}^\ell}{\partial b_k^\ell} G_{ij}$$

Considering again that $Z_{ij}^\ell = \dots + b_j^\ell$,

$$\frac{\partial Z_{ij}^\ell}{\partial b_k^\ell} = \begin{cases} \emptyset & \text{if } j \neq k \\ 1 & \text{if } j = k \end{cases}$$

So, again, m non-zero values, all ones.

$$(\nabla_{\underline{b}^\ell} \mathcal{L})_k = \sum_i^m G_{ik} \Rightarrow \text{sum of } \underline{G} \text{ } k\text{-th column}$$

$$\Rightarrow \text{B)} \quad \underbrace{\nabla_{\underline{b}^\ell} \mathcal{L}}_{(1 \times n^\ell)} = \underbrace{[\,1\,]}_{(1 \times m)} \underbrace{\underline{G}}_{(m \times n^\ell)}$$

At this point we need to calculate $\underline{G}$ needed for the next step of the backprop.

We need $\nabla_{\underline{O}^{\ell-1}} \mathcal{L}$.

Applying key idea: $$\nabla_{\underline{O}^{\ell-1}} \mathcal{L} = \sum_{i,j} \frac{\partial Z_{ij}^\ell}{\partial \underline{O}^{\ell-1}} \frac{\partial \mathcal{L}}{\partial Z_{ij}^\ell}$$

$$(\nabla_{\underline{O}^{\ell-1}} \mathcal{L})_{hk} = \sum_{i,j} \frac{\partial Z_{ij}^\ell}{\partial O_{hk}^{\ell-1}} \frac{\partial \mathcal{L}}{\partial Z_{ij}^\ell}$$

Considering again that $z_{ij} = o_{i1} w_{1j} + \dots + o_{in^i} w_{n^i j} + b_j$

$$\frac{\partial z_{ij}^{l}}{\partial o_{hk}^{l-1}} = \begin{cases} \emptyset & \text{if } i \neq h \\ W_{kj}^{l} & \text{if } i = h \end{cases}$$

So for every o_{hk}^{l-1} there are n^o non-zero values

$$W_{k1}^{l}, W_{k2}^{l}, \dots, W_{kn^o}^{l} \quad [j = 1, \dots, n^o ; \; i = h]$$

$$(\nabla_{\underline{o}^{l-1}} \mathcal{L})_{hk} = W_{k1}^{l} G_{h1} + \dots + W_{kn^o}^{l} G_{hn^o}$$

$\hookrightarrow$ Row k of $\underline{W}^l$ and h of $\underline{G}$

$$\Rightarrow c) \quad \underset{(m \times n^i)}{\nabla_{\underline{o}^{l-1}} \mathcal{L}} = \underset{(m \times n^o)}{\underline{G}} \; \underset{(n^o \times n^i)}{\underline{W}^{l^T}}$$