Backpropagation

Hung-yi Lee

Background

- Cost Function $C(\theta)$
  - Given training examples: $\{(x^1, \hat{y}^1), \cdots, (x^r, \hat{y}^r), \cdots, (x^R, \hat{y}^R)\}$
  - Find a set of parameters θ* minimizing C(θ)
    - $C(\theta) = \frac{1}{R}\sum_r C^r(\theta), C^r(\theta) = \|f(x^r; \theta) - \hat{y}^r\|$
- Gradient Descent
  - $\nabla C(\theta) = \frac{1}{R}\sum_r \nabla C^r(\theta)$
  - Given w_{ij}^l and b_i^l, we have to compute $\partial C^r / \partial w_{ij}^l$ and $\partial C^r / \partial b_i^l$
- There is an efficient way to compute the gradients of the network parameters – ***backpropagation***.

Chain Rule

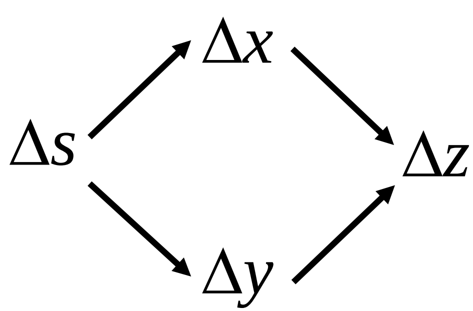
Case 1

$$y = g(x) \quad z = h(y)$$

$$\Delta x \to \Delta y \to \Delta z \qquad \frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx}$$

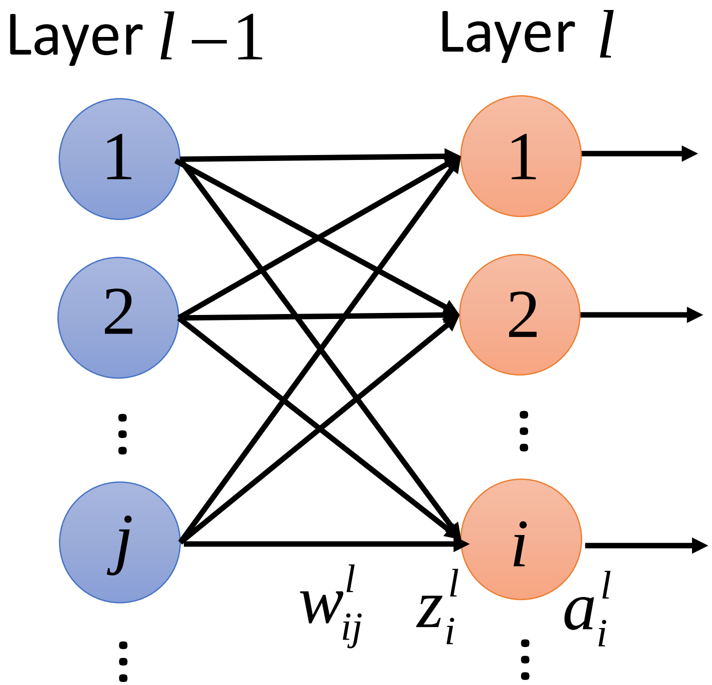
Case 2

$$x = g(s) \qquad y = h(s) \qquad z = k(x, y)$$



$$\frac{\partial z}{\partial s} = \frac{\partial z}{\partial x}\frac{\partial x}{\partial s} + \frac{\partial z}{\partial y}\frac{\partial y}{\partial s}$$

$\partial C^r / \partial w_{ij}^l$

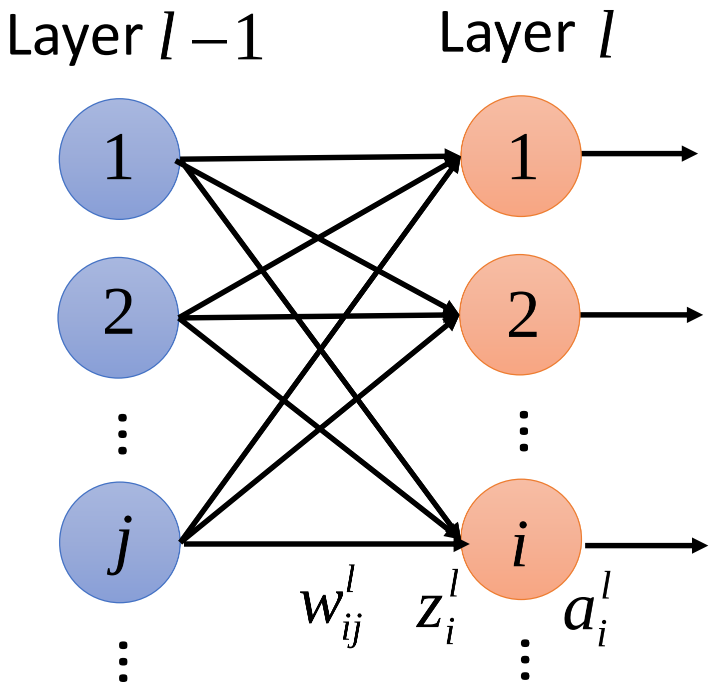


$$\Delta w_{ij}^l \rightarrow \Delta z_i^l \cdots\cdots \rightarrow \Delta C^r$$

$$\frac{\partial C^r}{\partial w_{ij}^l} = \frac{\partial z_i^l}{\partial w_{ij}^l} \frac{\partial C^r}{\partial z_i^l}$$

- $\frac{\partial C^r}{\partial w_{ij}^l}$ is the multiplication of two terms

$\partial C^r / \partial w_{ij}^l$ - First Term



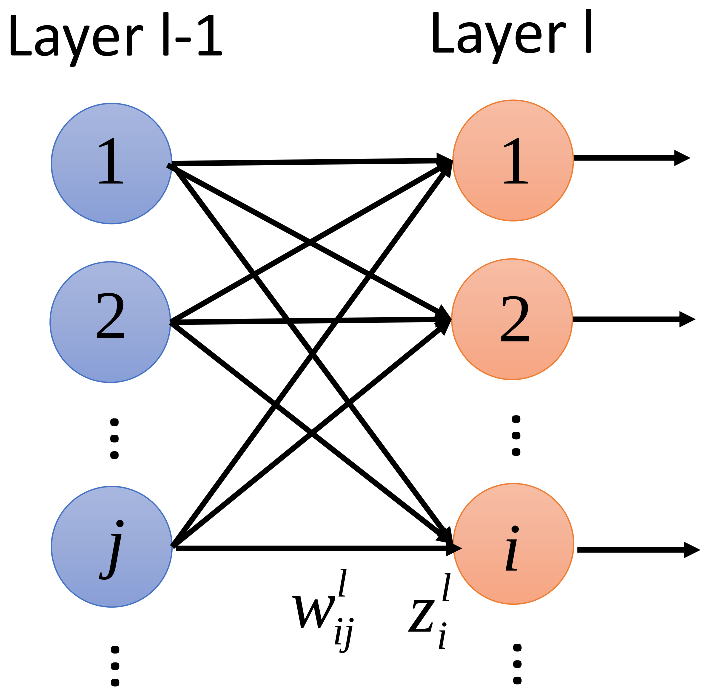
$$\Delta w_{ij}^l \rightarrow \Delta z_i^l \cdots\cdots \rightarrow \Delta C^r$$

$$\frac{\partial C^r}{\partial w_{ij}^l} = \boxed{\frac{\partial z_i^l}{\partial w_{ij}^l}} \frac{\partial C^r}{\partial z_i^l}$$

- $\dfrac{\partial C^r}{\partial w_{ij}^l}$ is the multiplication of two terms

$\partial C^r / \partial w_{ij}^l$ - First Term

$$\frac{\partial C^r}{\partial w_{ij}^l} = \boxed{\frac{\partial z_i^l}{\partial w_{ij}^l}} \frac{\partial C^r}{\partial z_i^l}$$

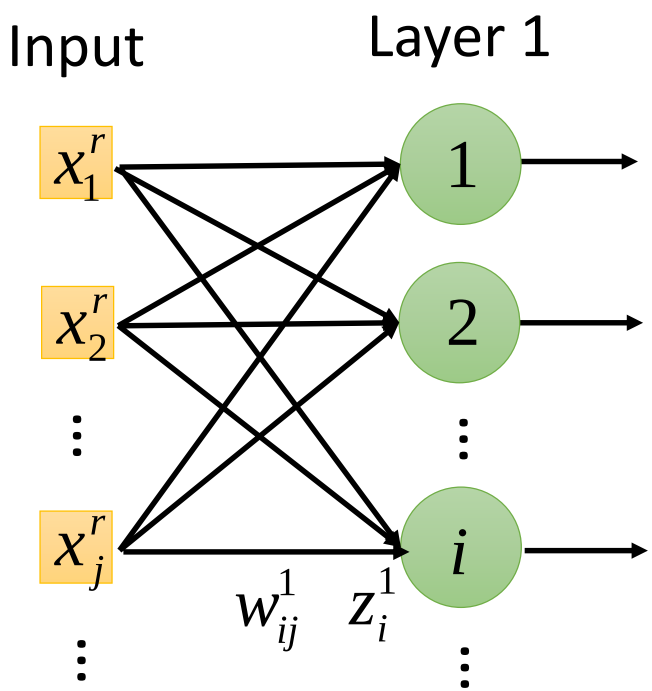


If l > 1

$$z_i^l = \sum_j w_{ij}^l a_j^{l-1} + b_i^l \qquad \frac{\partial z_i^l}{\partial w_{ij}^l} = a_j^{l-1}$$

$\partial C^r / \partial w_{ij}^l$ - First Term

$$\frac{\partial C^r}{\partial w_{ij}^l} = \boxed{\frac{\partial z_i^l}{\partial w_{ij}^l}} \frac{\partial C^r}{\partial z_i^l}$$



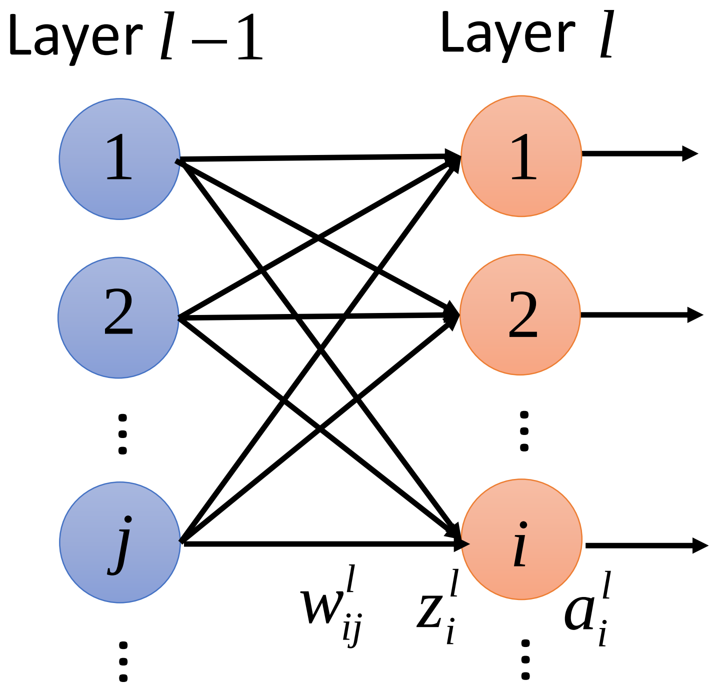
If l > 1

$$z_i^l = \sum_j w_{ij}^l a_j^{l-1} + b_i^l \qquad \frac{\partial z_i^l}{\partial w_{ij}^l} = a_j^{l-1}$$

If l = 1

$$z_i^1 = \sum_j w_{ij}^1 x_j^r + b_i^1 \qquad \frac{\partial z_i^1}{\partial w_{ij}^1} = x_j^r$$

$\partial C^r / \partial w_{ij}^l$ - Second Term



$$\Delta w_{ij}^l \rightarrow \Delta z_i^l \cdots\cdots \rightarrow \Delta C^r$$

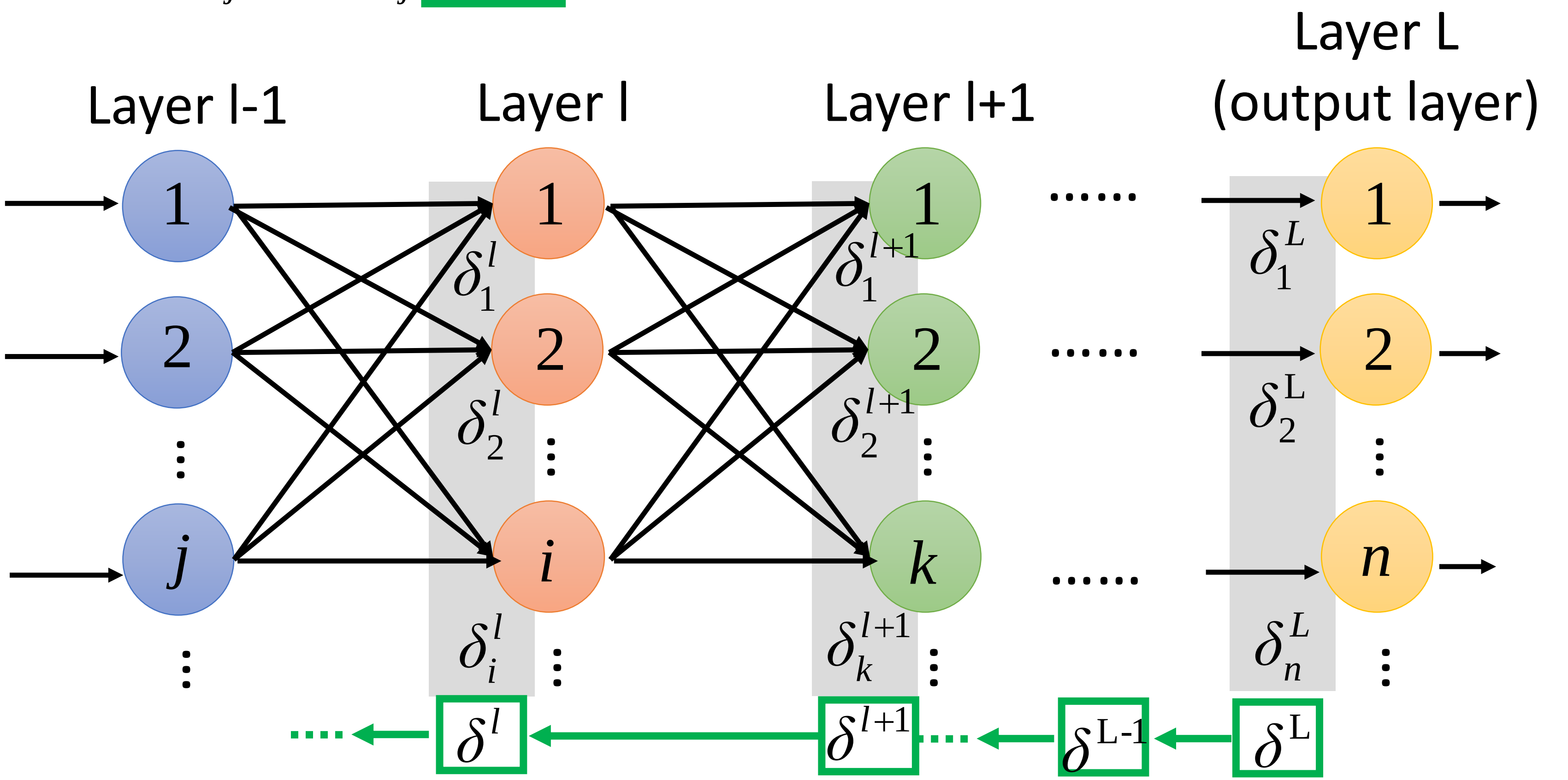
$$\frac{\partial C^r}{\partial w_{ij}^l} = \frac{\partial z_i^l}{\partial w_{ij}^l} \boxed{\frac{\partial C^r}{\partial z_i^l}} \rightarrow \delta_i^l$$

- $\dfrac{\partial C^r}{\partial w_{ij}^l}$ is the multiplication of two terms

$\partial C^r / \partial w_{ij}^l$ - Second Term

$$\frac{\partial C^r}{\partial w_{ij}^l} = \frac{\partial z_i^l}{\partial w_{ij}^l} \boxed{\frac{\partial C^r}{\partial z_i^l}} \rightarrow \delta_i^l$$

1. How to compute δ^L
2. The relation of δ^l and δ^{l+1}



$\partial C^r / \partial w_{ij}^l$ - Second Term

$$\frac{\partial C^r}{\partial w_{ij}^l} = \frac{\partial z_i^l}{\partial w_{ij}^l} \frac{\partial C^r}{\partial z_i^l} \rightarrow \delta_i^l$$

1. How to compute δ^L
2. The relation of δ^l and δ^{l+1}

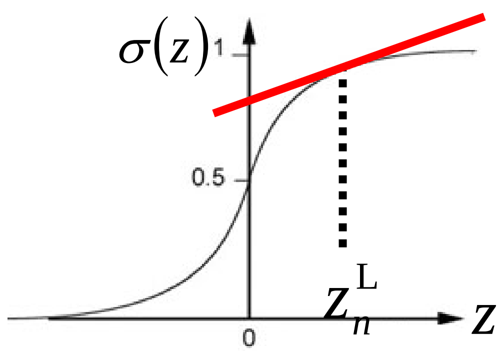
$$\delta_n^L = \frac{\partial C^r}{\partial z_n^L}$$

$$\Delta z_n^L \rightarrow \Delta a_n^L = \Delta y_n^r \rightarrow \Delta C^r$$

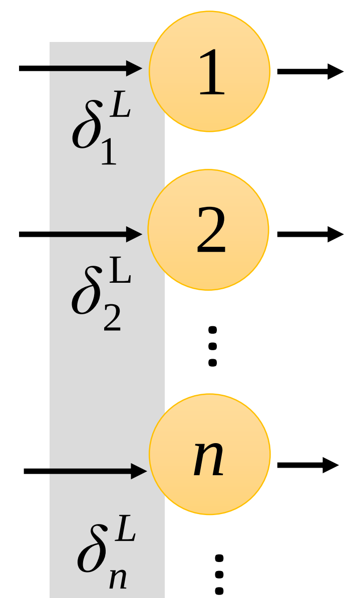
$$= \frac{\partial y_n^r}{\partial z_n^L} \frac{\partial C^r}{\partial y_n^r}$$

$\frac{\partial y_n^r}{\partial z_n^L} \rightarrow \sigma'(z_n^L)$

Depending on the definition of cost function



Layer L
(output layer)



$\partial C^r / \partial w_{ij}^l$ - Second Term

$$\frac{\partial C^r}{\partial w_{ij}^l} = \frac{\partial z_i^l}{\partial w_{ij}^l} \frac{\partial C^r}{\partial z_i^l} \rightarrow \delta_i^l$$

1. How to compute δ^L

2. The relation of δ^l and δ^{l+1}

$$\delta_n^L = \frac{\partial C^r}{\partial z_n^L}$$

$$= \frac{\partial y_n^r}{\partial z_n^L} \frac{\partial C^r}{\partial y_n^r}$$

$$= \sigma'(z_n^L) \frac{\partial C^r}{\partial y_n^r}$$

δ^L?

$$\sigma'(z^L) = \begin{bmatrix} \sigma'(z_1^L) \\ \sigma'(z_2^L) \\ \vdots \\ \sigma'(z_n^L) \\ \vdots \end{bmatrix} \qquad \nabla C^r(y^r) = \begin{bmatrix} \partial C^r / \partial y_1^r \\ \partial C^r / \partial y_2^r \\ \vdots \\ \partial C^r / \partial y_n^r \\ \vdots \end{bmatrix}$$

$$\delta^L = \sigma'(z^l) \bullet \nabla C^r(y^r)$$

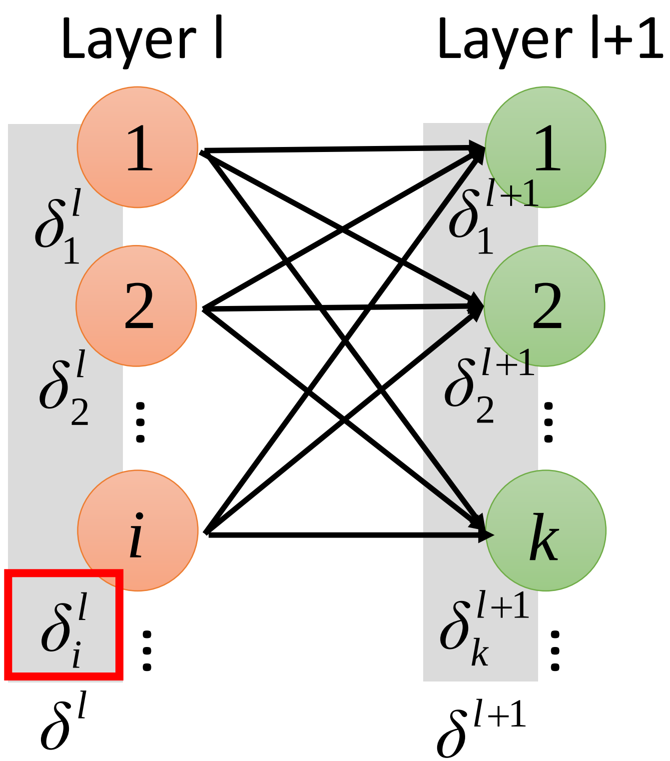
element-wise multiplication

$\partial C^r / \partial w_{ij}^l$ - Second Term

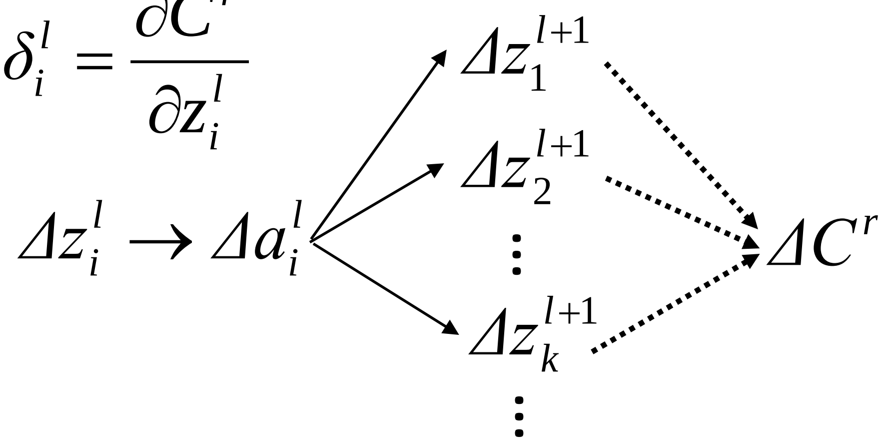
$$\frac{\partial C^r}{\partial w_{ij}^l} = \frac{\partial z_i^l}{\partial w_{ij}^l} \boxed{\frac{\partial C^r}{\partial z_i^l}} \rightarrow \delta_i^l$$

1. How to compute δ^L

2. The relation of δ^l and δ^{l+1}



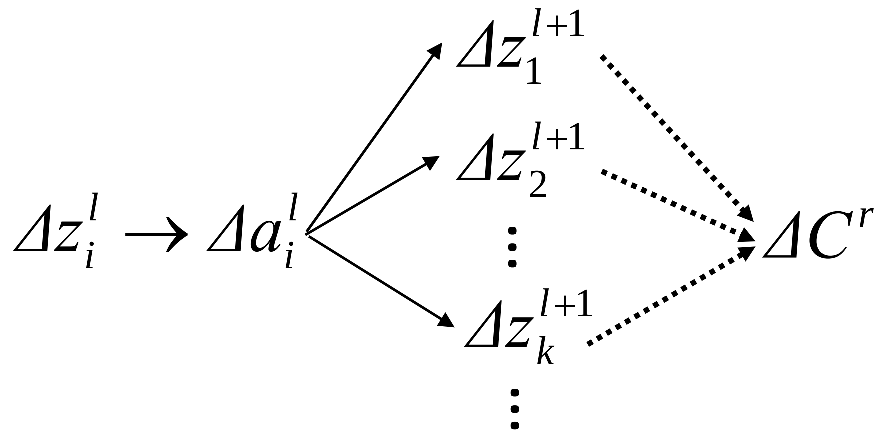
$$\delta_i^l = \frac{\partial C^r}{\partial z_i^l}$$

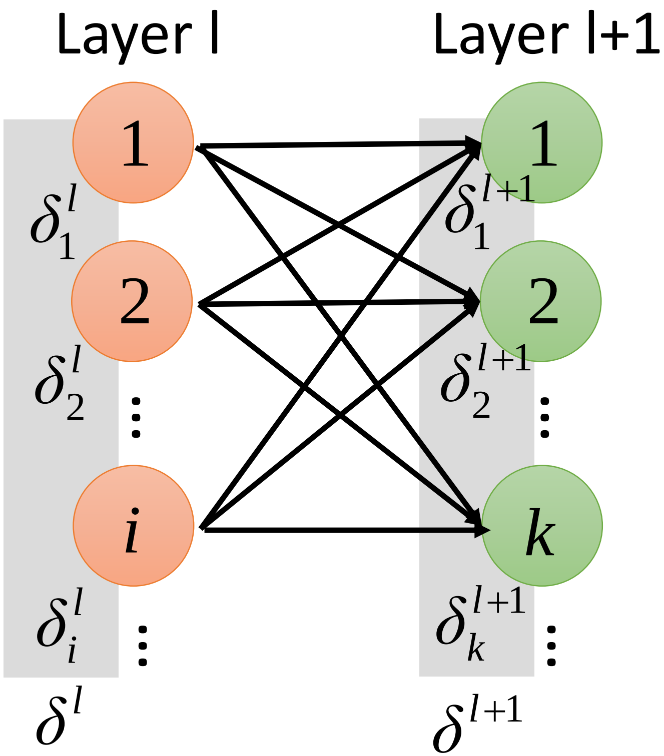


$$\delta_i^l = \frac{\partial C^r}{\partial z_i^l} = \frac{\partial a_i^l}{\partial z_i^l} \sum_k \frac{\partial z_k^{l+1}}{\partial a_i^l} \boxed{\frac{\partial C^r}{\partial z_k^{l+1}}} \rightarrow \delta_k^{l+1}$$

$\partial C^r / \partial w_{ij}^l$ - Second Term

$$\frac{\partial C^r}{\partial w_{ij}^l} = \frac{\partial z_i^l}{\partial w_{ij}^l} \boxed{\frac{\partial C^r}{\partial z_i^l}} \rightarrow \delta_i^l$$





$$\delta_i^l = \frac{\partial a_i^l}{\partial z_i^l} \sum_k \frac{\partial z_k^{l+1}}{\partial a_i^l} \delta_k^{l+1}$$

$$\sigma'\left(z_i^l\right)$$

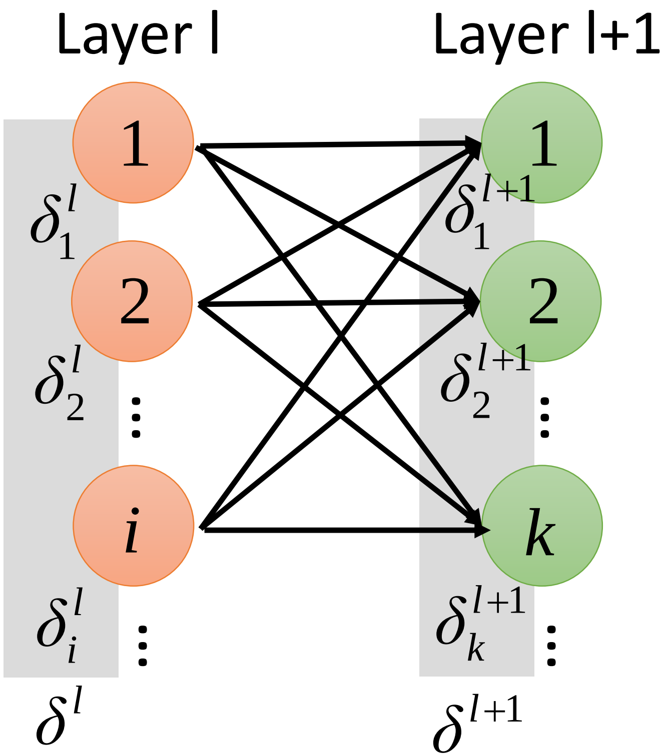
$$z_k^{l+1} = \sum_i w_{ki}^{l+1} a_i^l + b_k^{l+1}$$

$$\delta_i^l = \sigma'\left(z_i^l\right) \sum_k w_{ki}^{l+1} \delta_k^{l+1}$$

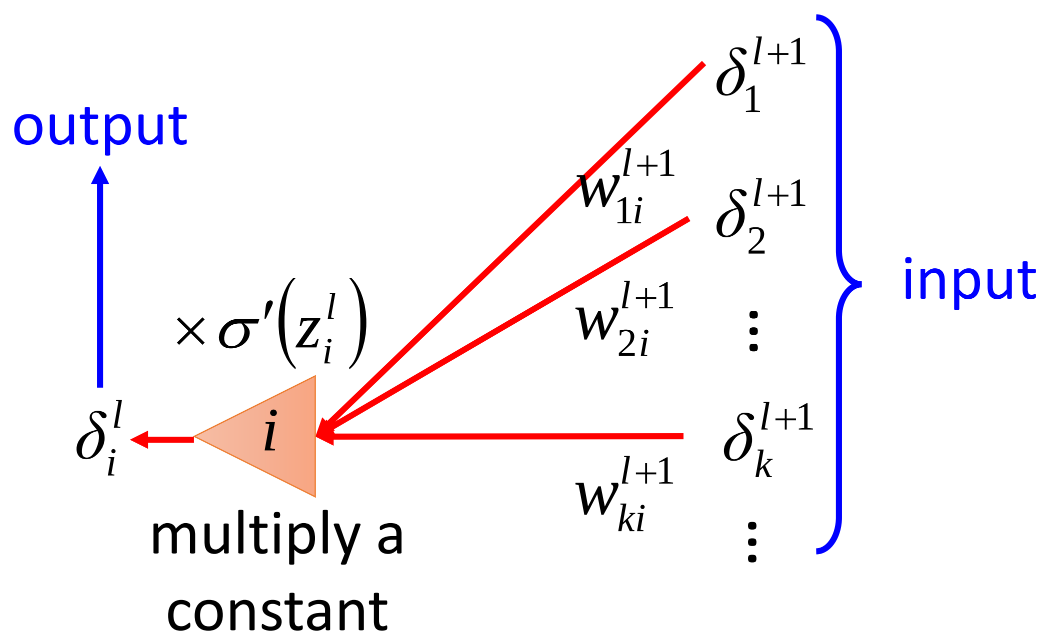
$\partial C^r / \partial w_{ij}^l$ - Second Term

$$\frac{\partial C^r}{\partial w_{ij}^l} = \frac{\partial z_i^l}{\partial w_{ij}^l}\boxed{\frac{\partial C^r}{\partial z_i^l}} \rightarrow \delta_i^l$$

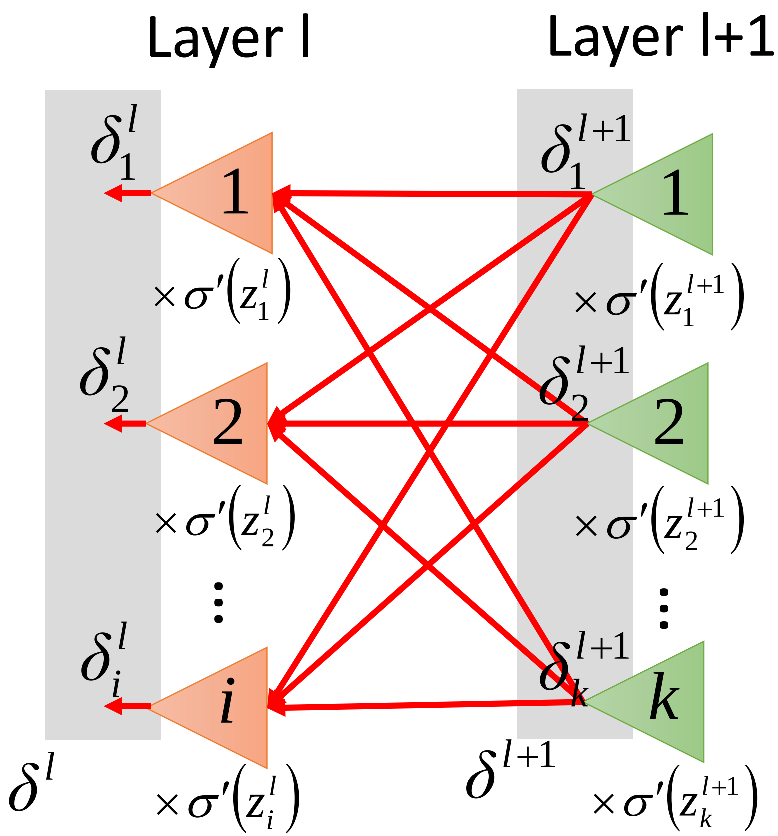
$$\delta_i^l = \sigma'\left(z_i^l\right)\sum_k w_{ki}^{l+1}\delta_k^{l+1}$$



new type of neuron



$\partial C^r / \partial w_{ij}^l$ - Second Term

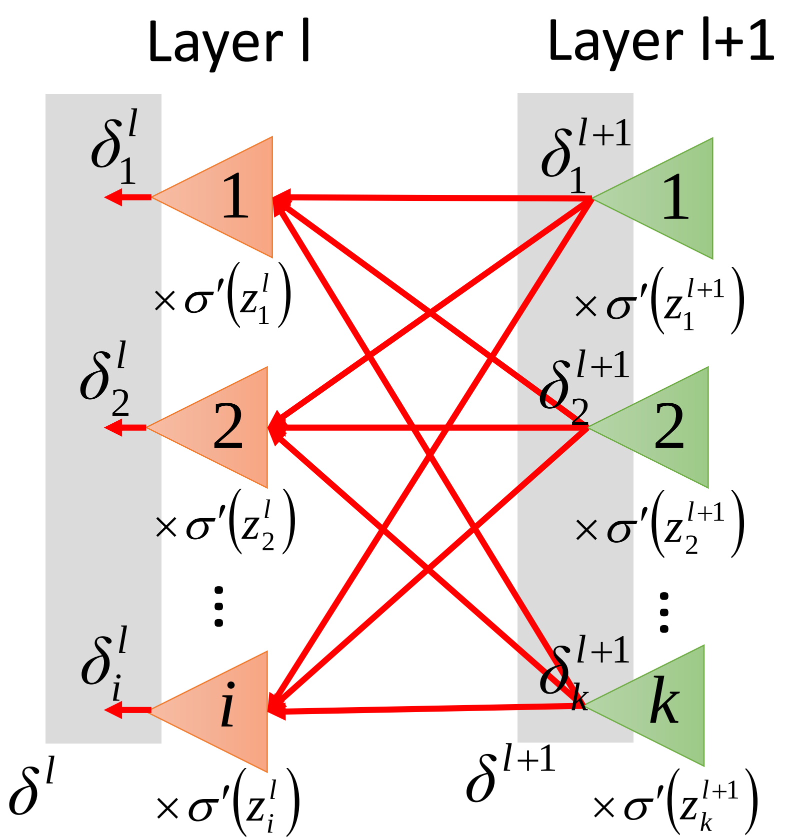


$$\delta_i^l = \sigma'(z_i^l)\sum_k w_{ki}^{l+1}\delta_k^{l+1}$$

$$\sigma'(z^l) = \begin{bmatrix} \sigma'(z_1^l) \\ \sigma'(z_2^l) \\ \vdots \\ \sigma'(z_i^l) \\ \vdots \end{bmatrix}$$

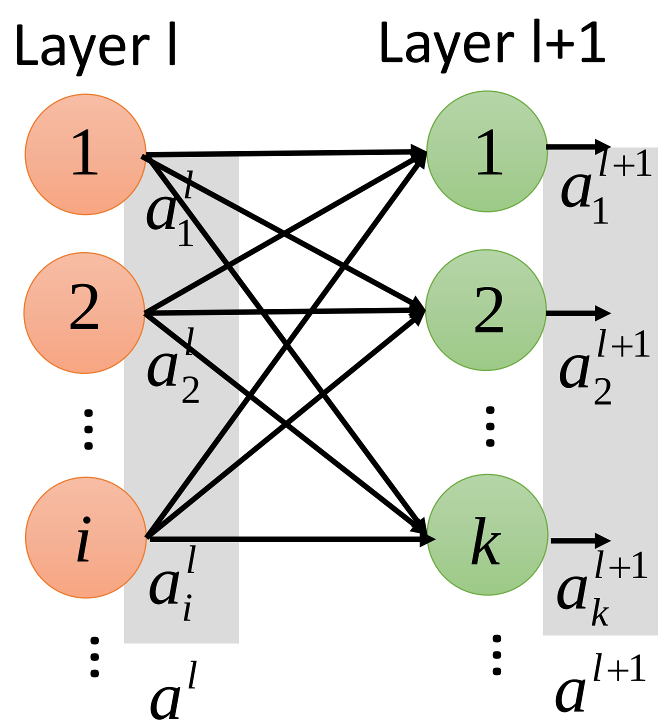
$$\delta^l = \sigma'(z^l) \bullet (W^{l+1})^T \delta^{l+1}$$

$\partial C^r / \partial w_{ij}^l$ - Second Term



$$\delta^l = \sigma'(z^l) \bullet (W^{l+1})^T \delta^{l+1}$$

Compare



$$a^{l+1} = \sigma(W^{l+1} a^l + b^{l+1})$$

$$\frac{\partial C^r}{\partial w_{ij}^l} = \frac{\partial z_i^l}{\partial w_{ij}^l} \boxed{\frac{\partial C^r}{\partial z_i^l}}$$

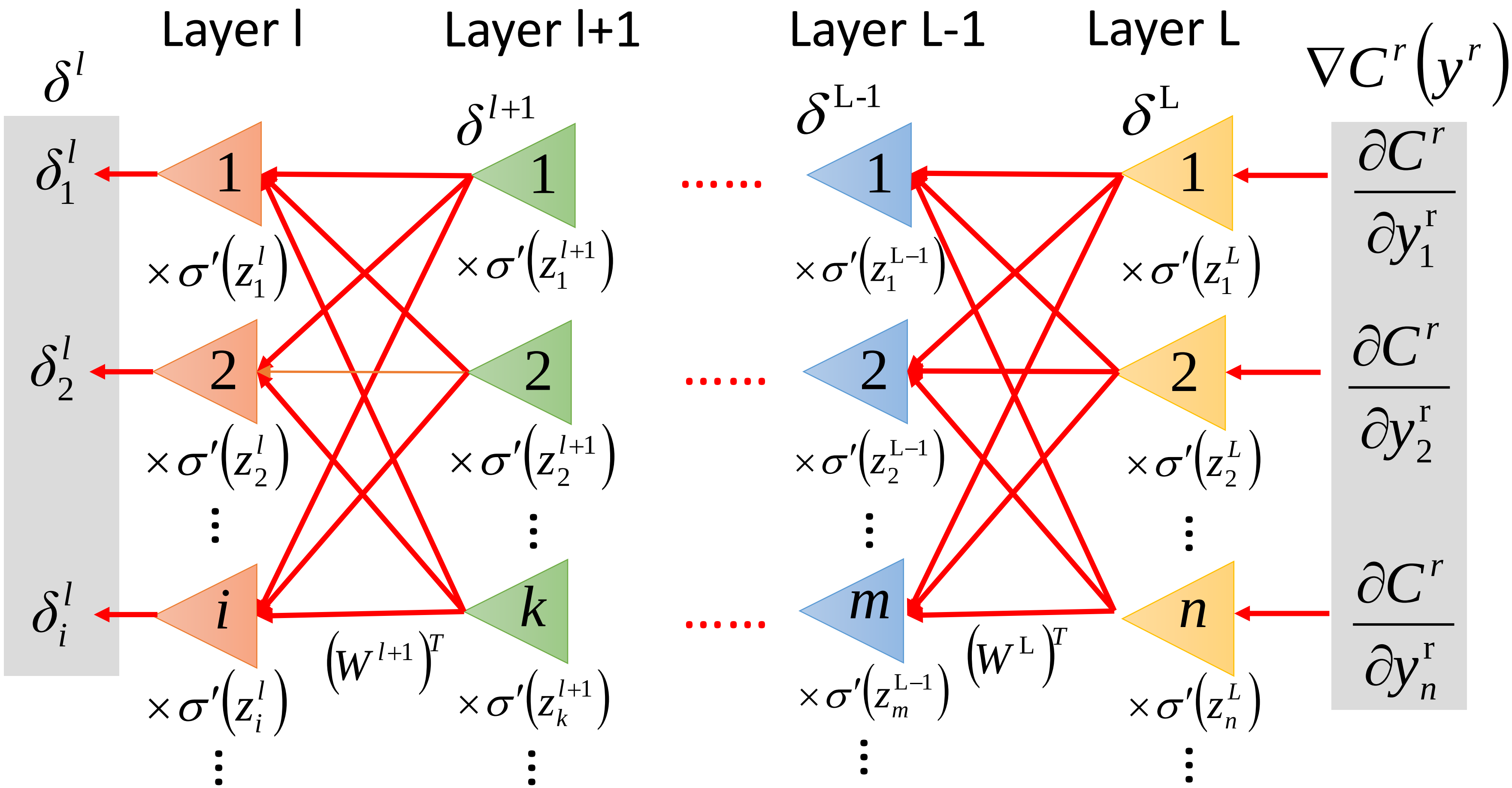
δ_i^l

1. How to compute δ^L

$$\Rightarrow \delta^{\mathrm{L}} = \sigma'(z^{\mathrm{L}}) \bullet \nabla C^r(y^r)$$

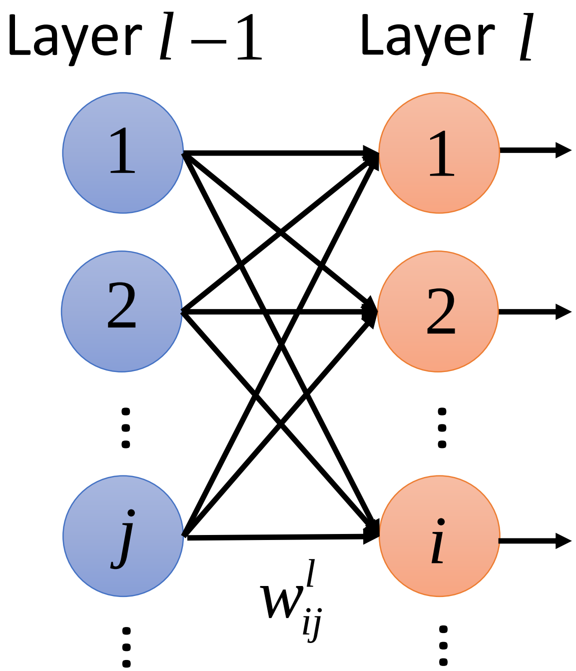
2. The relation of δ^l and δ^{l+1}

$$\Rightarrow \delta^l = \sigma'(z^l) \bullet (W^{l+1})^T \delta^{l+1}$$



Concluding Remarks

$$\frac{\partial C^r}{\partial w_{ij}^l} = \frac{\partial z_i^l}{\partial w_{ij}^l} \frac{\partial C^r}{\partial z_i^l}$$



$$\begin{cases} a_j^{l-1} & l > 1 \\ x_j^r & l = 1 \end{cases}$$

$$\delta_i^l$$

Forward Pass

$$z^1 = W^1 x^r + b^1$$

$$a^1 = \sigma\left(z^1\right)$$

$$\cdots\cdots$$

$$z^{l-1} = W^{l-1} a^{l-2} + b^{l-1}$$

$$a^{l-1} = \sigma\left(z^{l-1}\right)$$

Backward Pass

$$\delta^L = \sigma'\left(z^L\right) \bullet \nabla C^r\left(y^r\right)$$

$$\delta^{L-1} = \sigma'\left(z^{L-1}\right) \bullet \left(W^L\right)^T \delta^L$$

$$\cdots\cdots$$

$$\delta^l = \sigma'\left(z^l\right) \bullet \left(W^{l+1}\right)^T \delta^{l+1}$$

$$\cdots\cdots$$

Acknowledgement

- 感謝 Ryan Sun 來信指出投影片上的錯誤