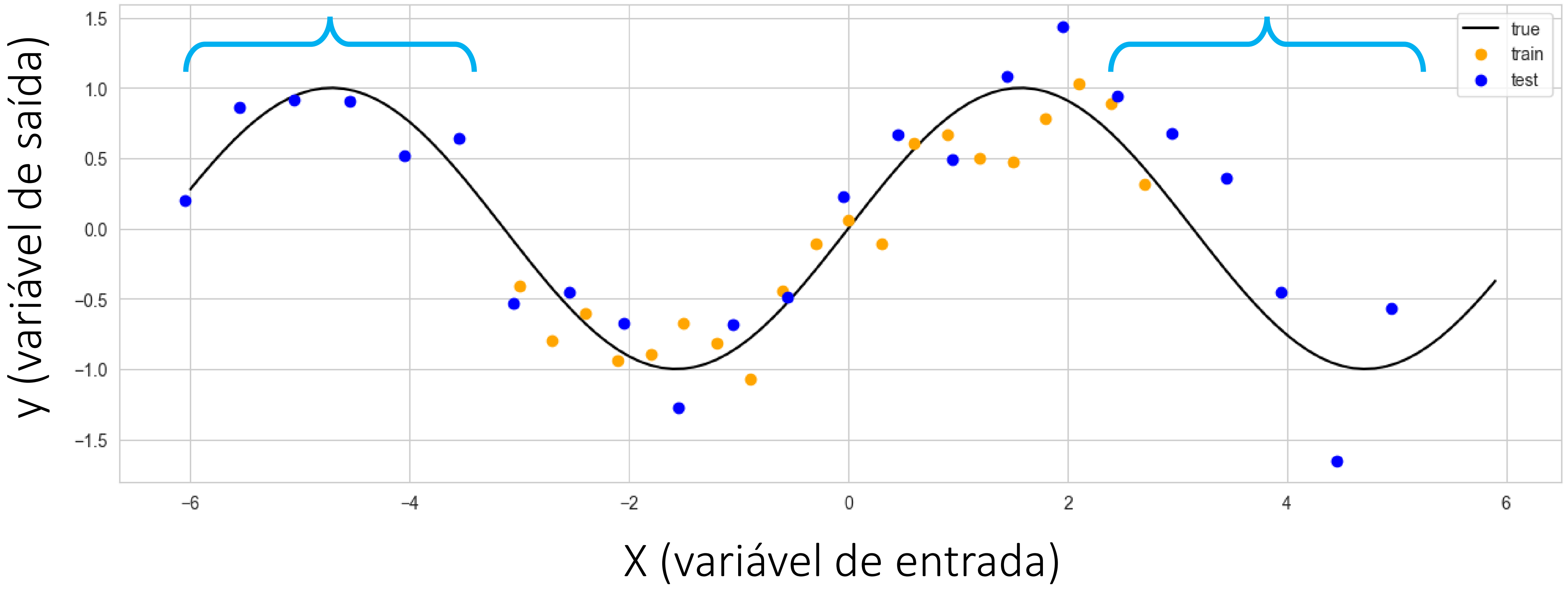
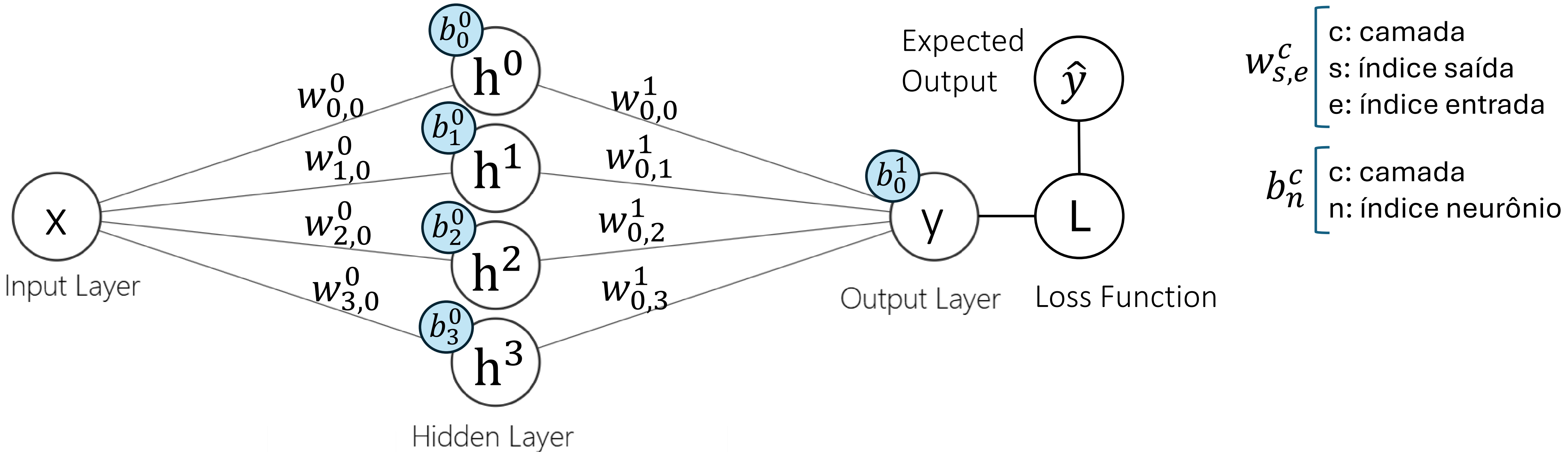
Derivação do *Backpropagation* para uma MLP com uma camada oculta na tarefa de regressão univariada

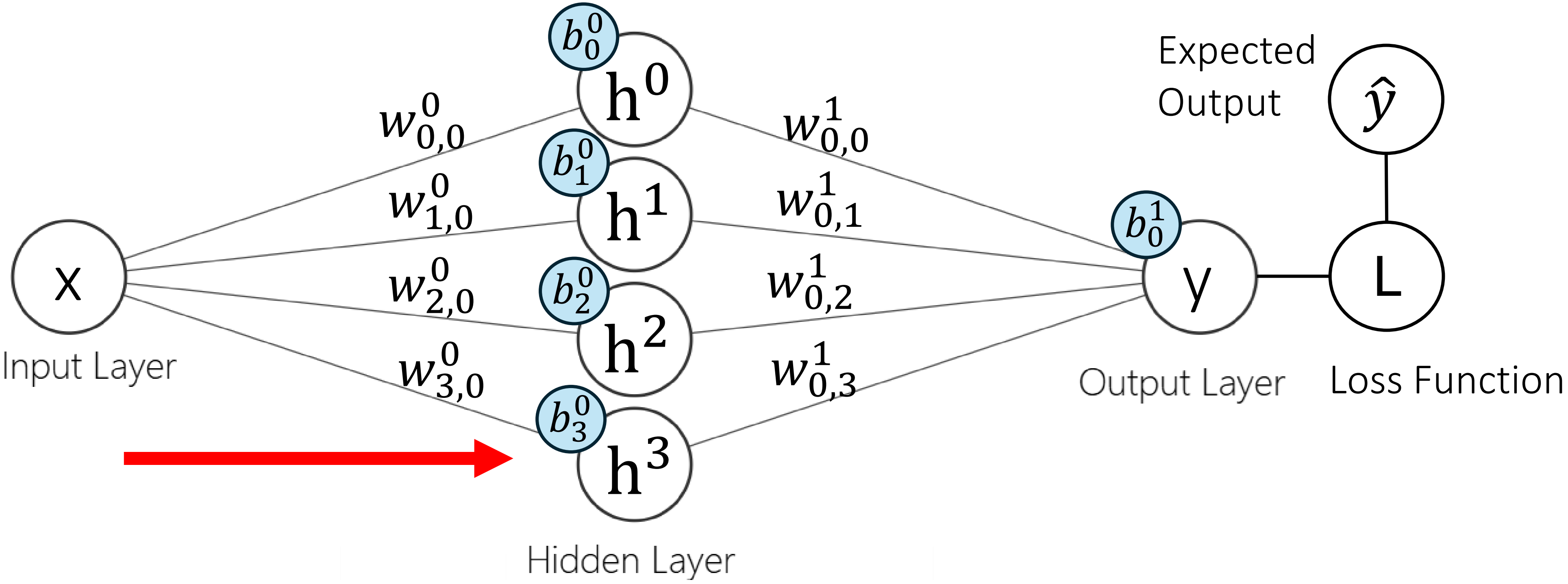
Tarefa



y (variável de saída)
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−6
−4
−2
0
2
4
6
true
train
test
X (variável de entrada)



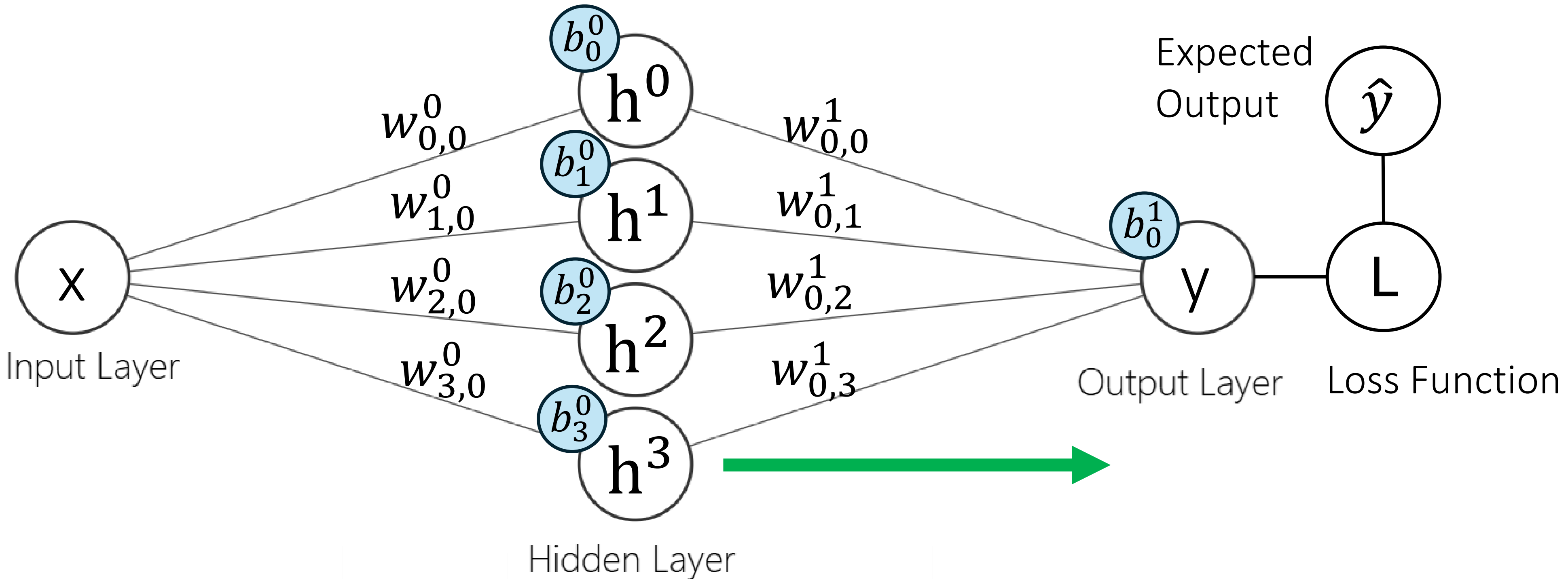
x
Input Layer
$w_{0,0}^0$
$w_{1,0}^0$
$w_{2,0}^0$
$w_{3,0}^0$
b_0^0 h^0
b_1^0 h^1
b_2^0 h^2
b_3^0 h^3
Hidden Layer
$w_{0,0}^1$
$w_{0,1}^1$
$w_{0,2}^1$
$w_{0,3}^1$
b_0^1 y
Output Layer
Expected Output
$\hat{y}$
L
Loss Function
$w_{s,e}^c$
c: camada
s: índice saída
e: índice entrada
b_n^c
c: camada
n: índice neurônio



$w^c_{s,e}$
- c: camada
- s: índice saída
- e: índice entrada

b^c_n
- c: camada
- n: índice neurônio

$$\sigma\left(\begin{bmatrix} w^0_{0,0} \\ w^0_{1,0} \\ w^0_{2,0} \\ w^0_{3,0} \end{bmatrix} \begin{bmatrix} x_i \end{bmatrix} + \begin{bmatrix} b^0_0 \\ b^0_1 \\ b^0_2 \\ b^0_3 \end{bmatrix}\right) = \begin{bmatrix} h^0 \\ h^1 \\ h^2 \\ h^3 \end{bmatrix}$$

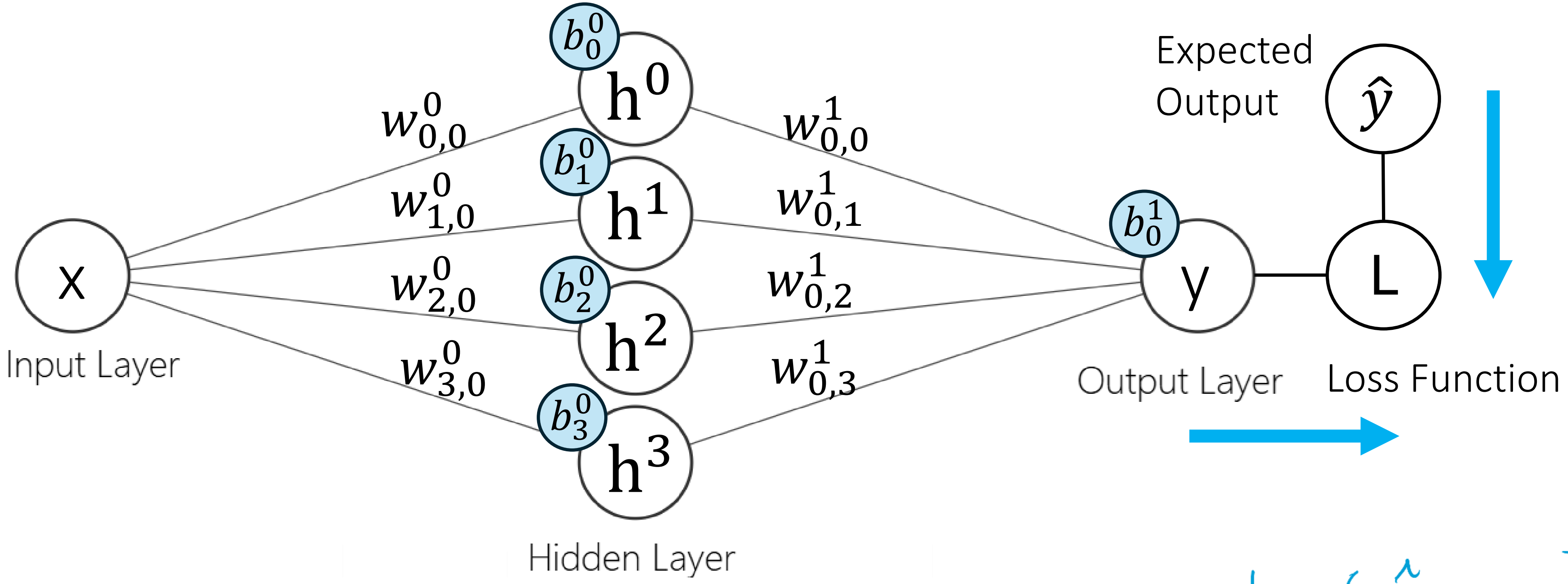


$w^c_{s,e}$ [c: camada
s: índice saída
e: índice entrada

b^c_n [c: camada
n: índice neurônio

como a tarefa é de regressão, a função de ativação da última camada será a identidade

$$\sigma\left(\begin{bmatrix} w^1_{0,0} & w^1_{0,1} & w^1_{0,2} & w^1_{0,3} \end{bmatrix}\begin{bmatrix} h^0 \\ h^1 \\ h^2 \\ h^3 \end{bmatrix} + \begin{bmatrix} b^1_0 \end{bmatrix}\right) = \begin{bmatrix} y_i \end{bmatrix}$$

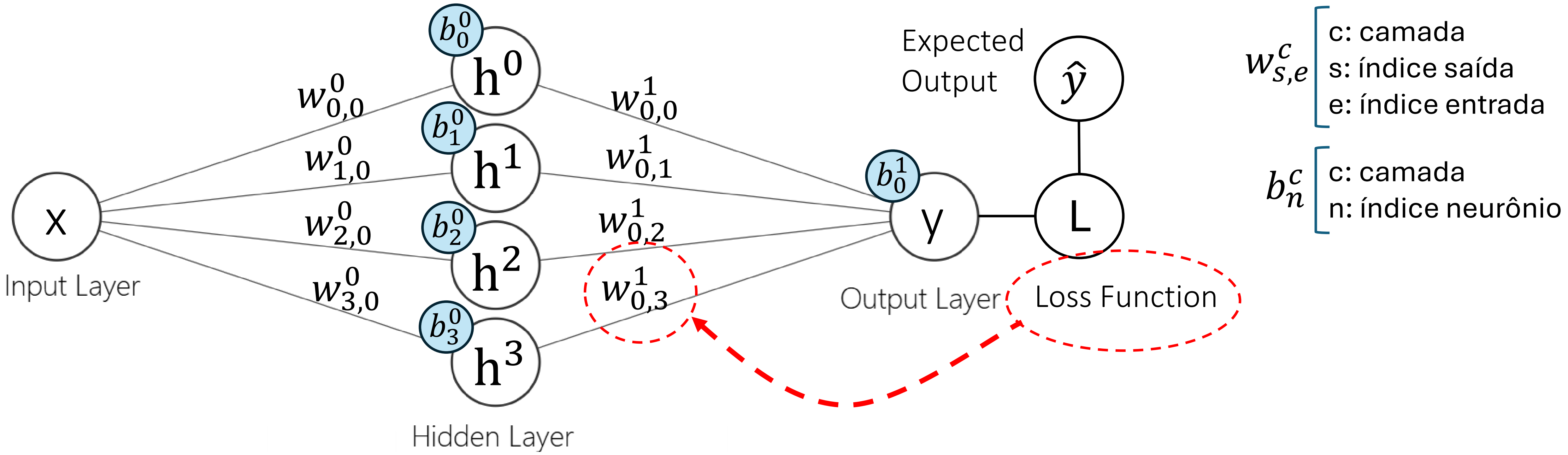


$w_{s,e}^c$
- c: camada
- s: índice saída
- e: índice entrada

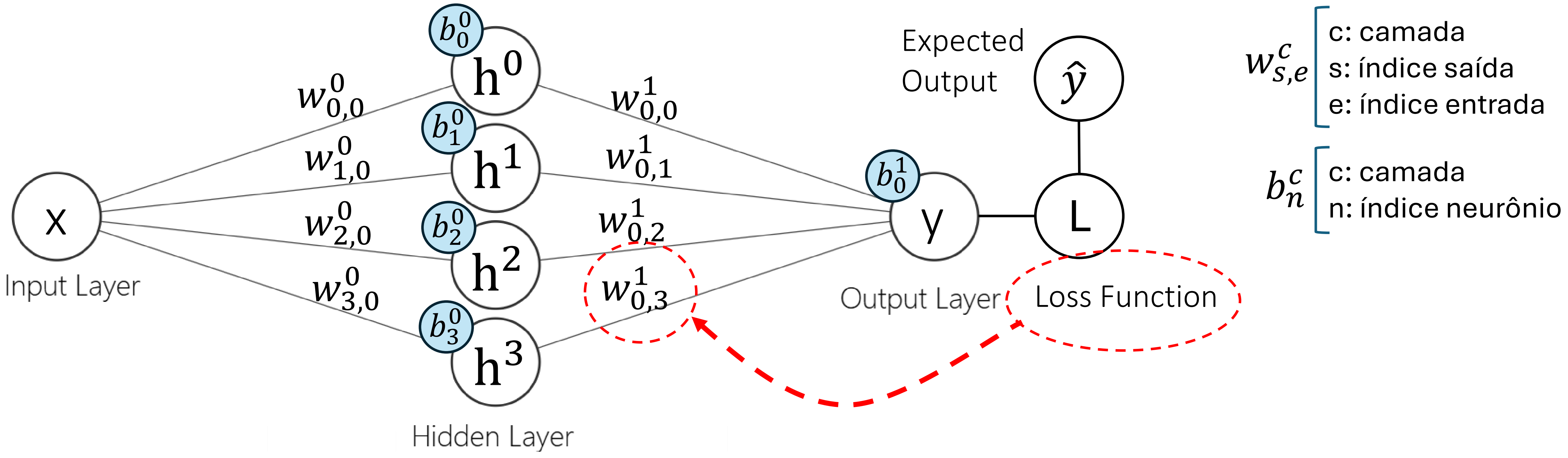
b_n^c
- c: camada
- n: índice neurônio

para amostra i: $L_i = (\hat{y}_i - y_i)^2$

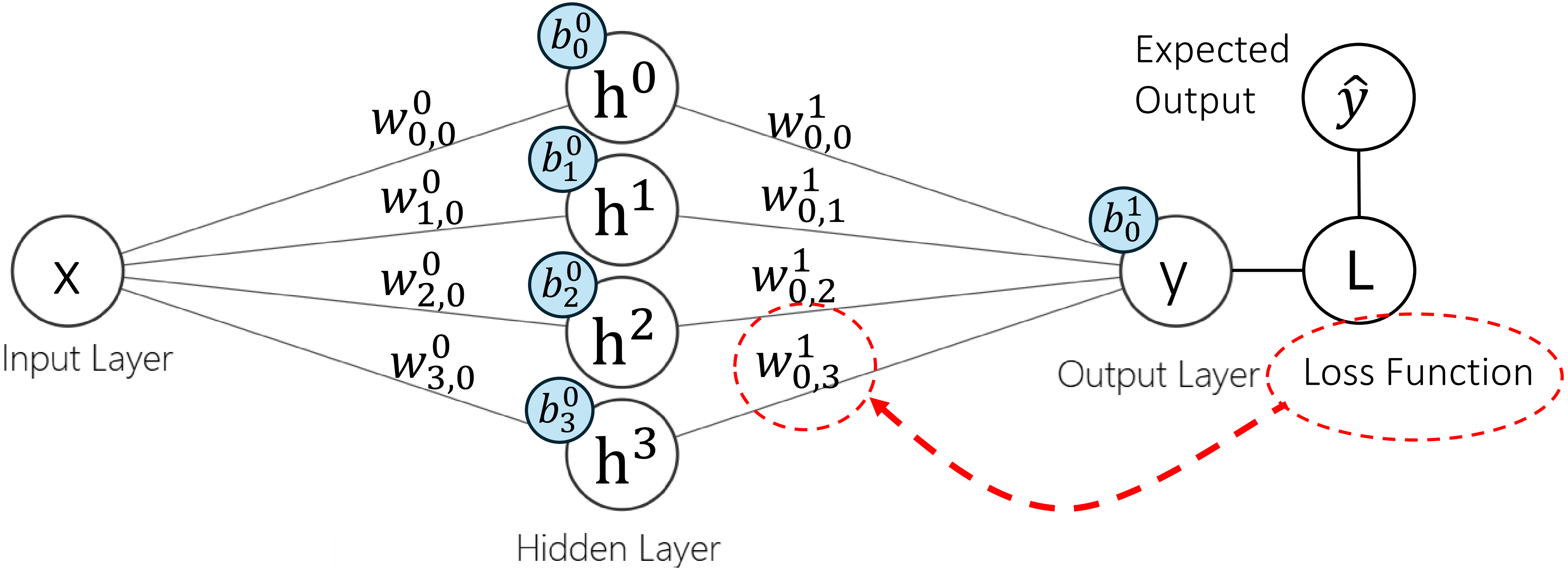
MSE para dataset: $L = \frac{1}{N} \sum_{i=0}^{N} (\hat{y}_i - y_i)^2$



$$\frac{\partial L}{\partial w_{0,3}^1}$$



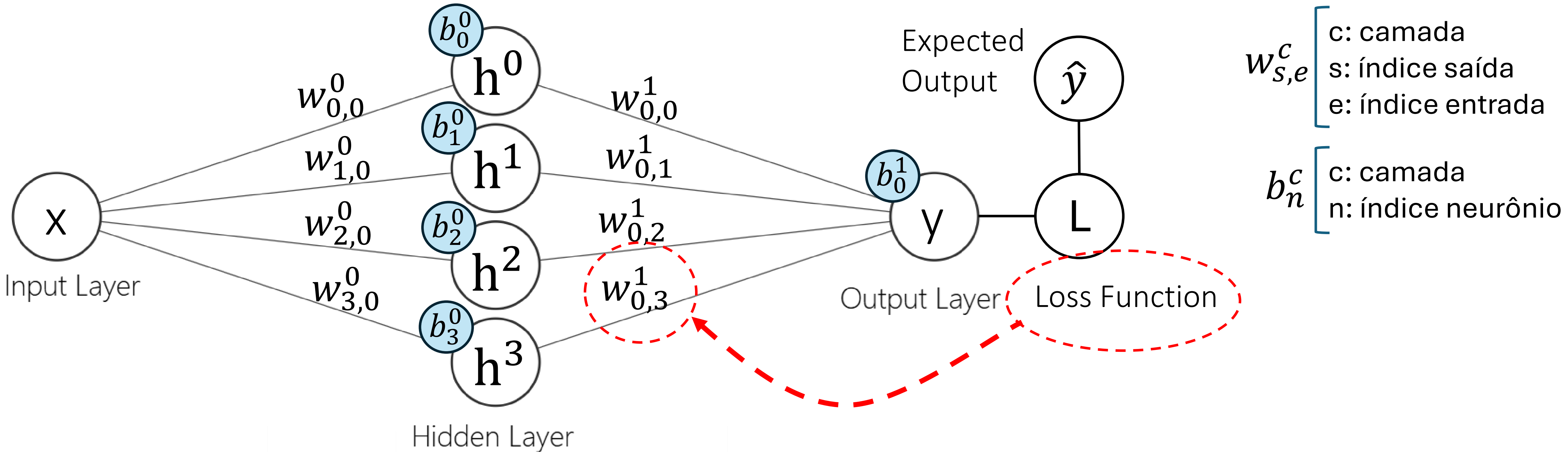
$$\frac{\partial L}{\partial w^1_{0,3}} = \frac{\partial}{\partial w^1_{0,3}} \frac{1}{N} \sum_{i=0}^{N} (y_i - \hat{y}_i)^2$$



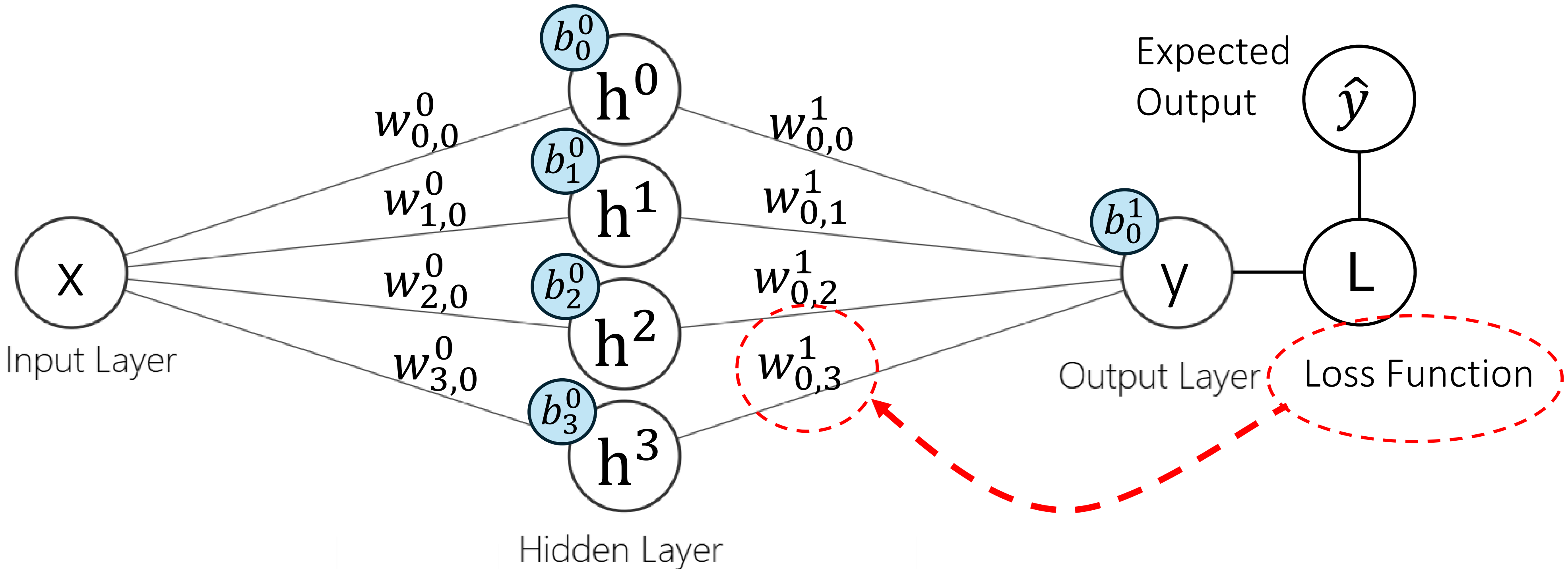
$w_{s,e}^{c}$ c: camada
s: índice saída
e: índice entrada

b_n^c c: camada
n: índice neurônio

$$\frac{\partial L}{\partial w_{0,3}^1} = \frac{\partial}{\partial w_{0,3}^1} \frac{1}{N} \sum_{i=0}^{N} (y_i - \hat{y}_i)^2 = \frac{1}{N} \sum_{i=0}^{N} \frac{\partial}{\partial w_{0,3}^1} (y_i - \hat{y}_i)^2$$



$$\frac{\partial L}{\partial w^1_{0,3}} = \frac{\partial}{\partial w^1_{0,3}} \frac{1}{N} \sum_{i=0}^{N} (y_i - \hat{y_i})^2 = \frac{1}{N} \sum_{i=0}^{N} \frac{\partial}{\partial w^1_{0,3}} (y_i - \hat{y_i})^2 = \frac{1}{N} \sum_{i=0}^{N} 2(y_i - \hat{y_i}) \frac{\partial}{\partial w^1_{0,3}} y_i$$



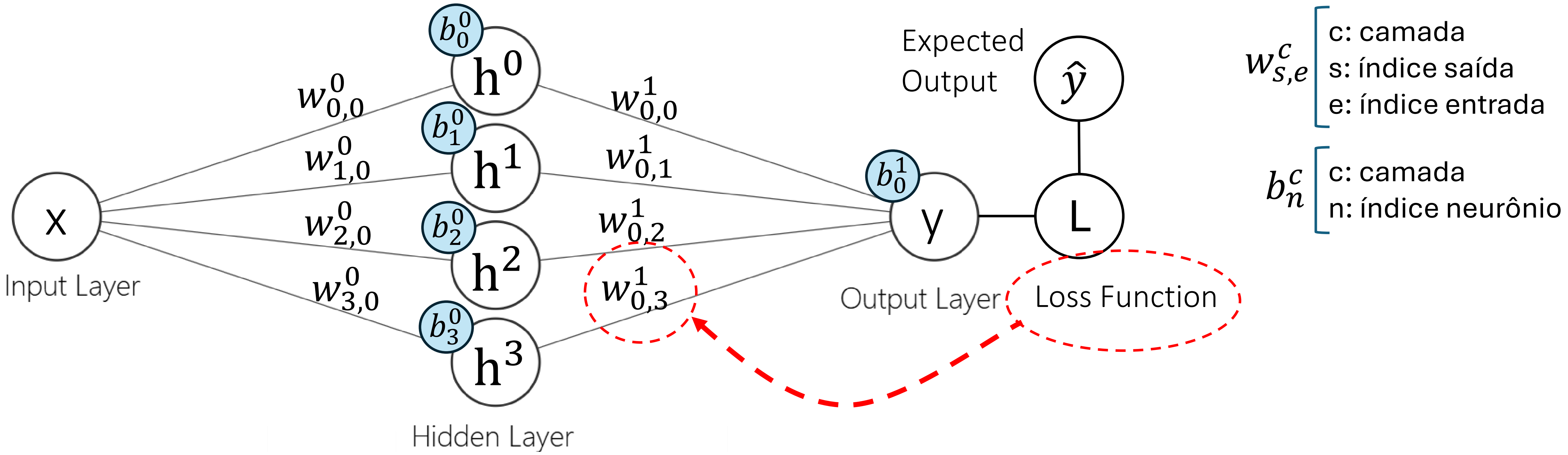
$w_{s,e}^c$ c: camada
s: índice saída
e: índice entrada

b_n^c c: camada
n: índice neurônio

$$\frac{\partial L}{\partial w_{0,3}^1} = \frac{\partial}{\partial w_{0,3}^1}\frac{1}{N}\sum_{i=0}^{N}(y_i - \hat{y_i})^2 = \frac{1}{N}\sum_{i=0}^{N}\frac{\partial}{\partial w_{0,3}^1}(y_i - \hat{y_i})^2 = \frac{1}{N}\sum_{i=0}^{N}2(y_i - \hat{y_i})\frac{\partial}{\partial w_{0,3}^1}y_i$$

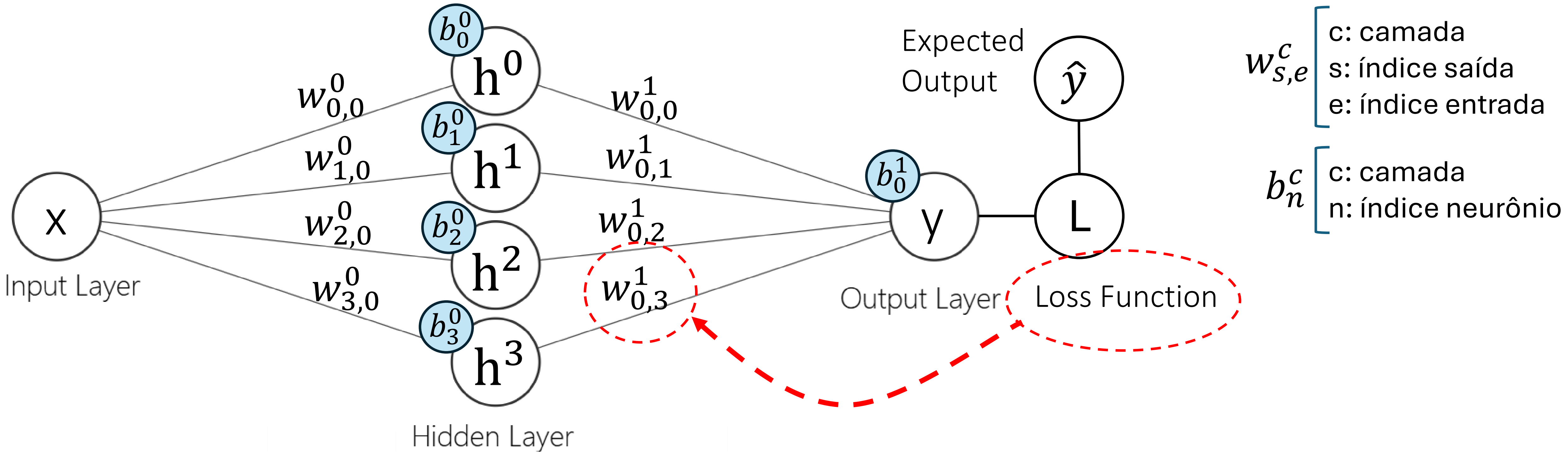
$$\frac{\partial}{\partial w_{0,3}^1}y_i = \frac{\partial}{\partial w_{0,3}^1}\left[h^0 w_{0,0}^1 + h^1 w_{0,1}^1 + h^2 w_{0,2}^1 + h^3 w_{0,3}^1 + b_0^1\right]$$

lembrando que a função de ativação é identidade na última camada



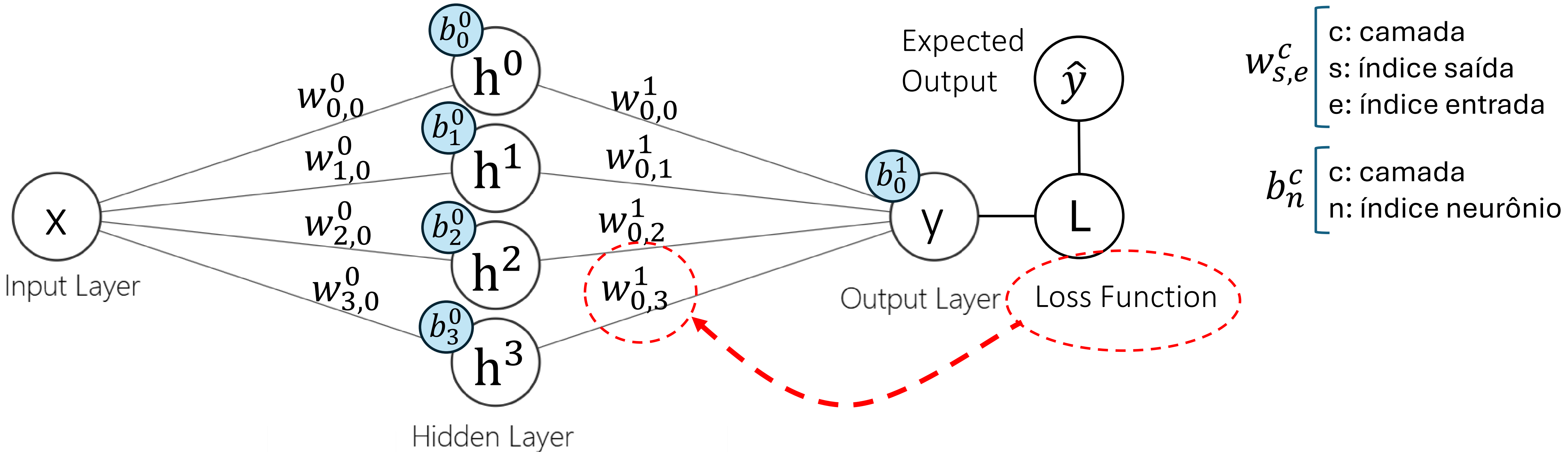
$$\frac{\partial L}{\partial w_{0,3}^1} = \frac{\partial}{\partial w_{0,3}^1}\frac{1}{N}\sum_{i=0}^{N}(y_i - \widehat{y_i})^2 = \frac{1}{N}\sum_{i=0}^{N}\frac{\partial}{\partial w_{0,3}^1}(y_i - \widehat{y_i})^2 = \frac{1}{N}\sum_{i=0}^{N}2(y_i - \widehat{y_i})\frac{\partial}{\partial w_{0,3}^1}y_i$$

$$\frac{\partial}{\partial w_{0,3}^1}y_i = \frac{\partial}{\partial w_{0,3}^1}\left[h^0 w_{0,0}^1 + h^1 w_{0,1}^1 + h^2 w_{0,2}^1 \boxed{+ h^3 w_{0,3}^1 +} b_0^1\right]$$

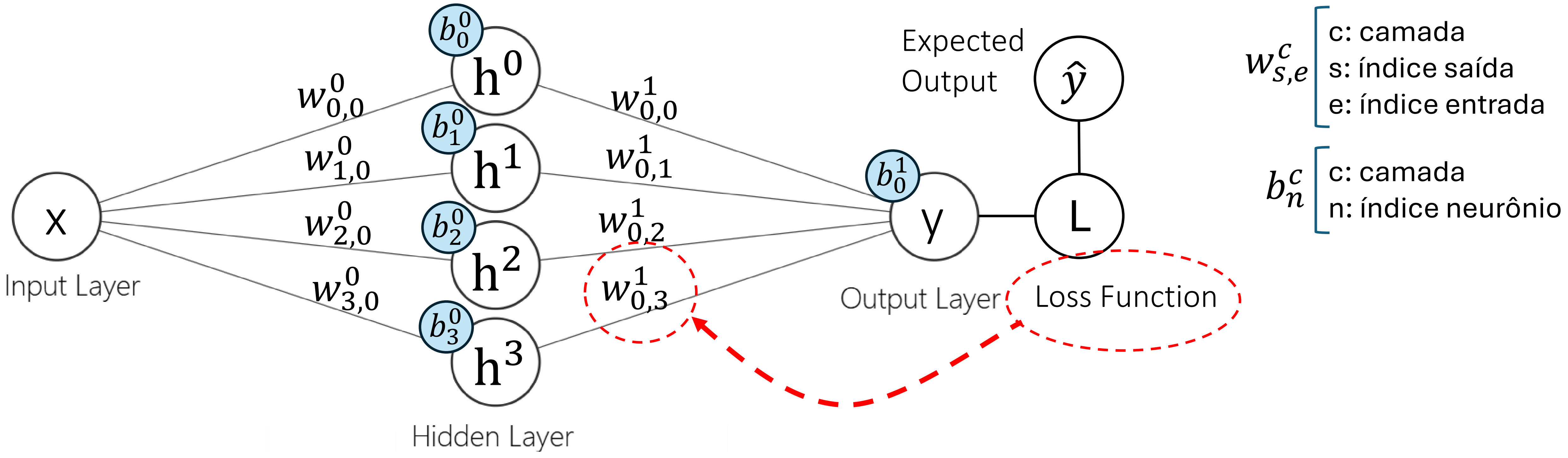


$$\frac{\partial L}{\partial w_{0,3}^{1}} = \frac{\partial}{\partial w_{0,3}^{1}} \frac{1}{N} \sum_{i=0}^{N} (y_i - \hat{y_i})^2 = \frac{1}{N} \sum_{i=0}^{N} \frac{\partial}{\partial w_{0,3}^{1}} (y_i - \hat{y_i})^2 = \frac{1}{N} \sum_{i=0}^{N} 2(y_i - \hat{y_i}) \frac{\partial}{\partial w_{0,3}^{1}} y_i$$

$$\frac{\partial}{\partial w_{0,3}^{1}} y_i = h^3$$

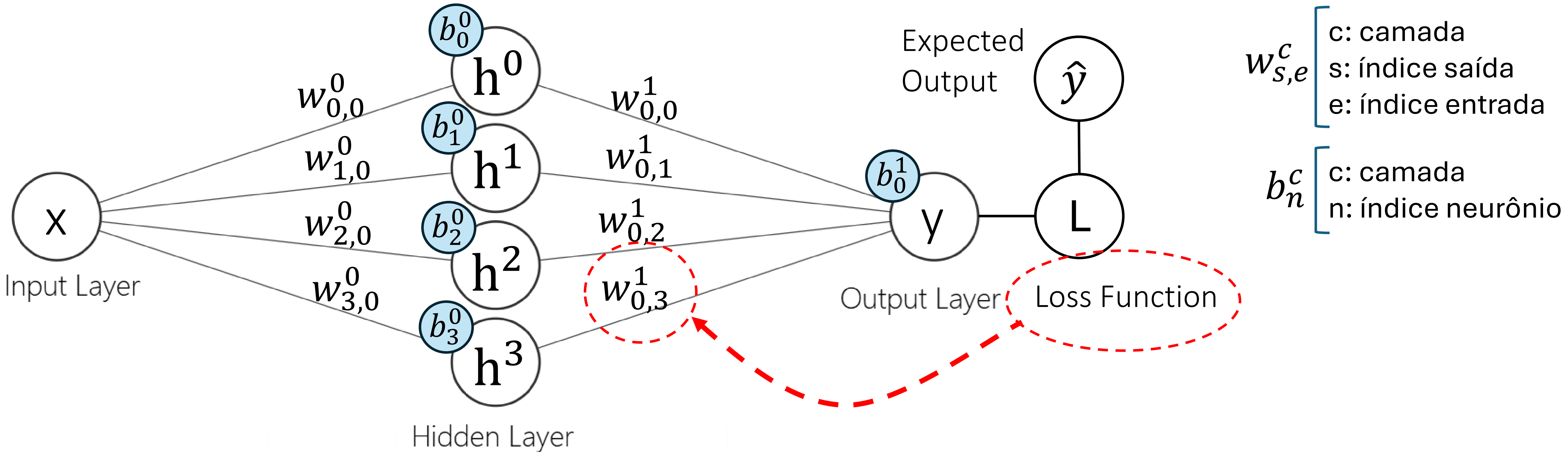


$$\frac{\partial L}{\partial w_{0,3}^1} = \frac{\partial}{\partial w_{0,3}^1}\frac{1}{N}\sum_{i=0}^{N}(y_i - \hat{y}_i)^2 = \frac{1}{N}\sum_{i=0}^{N}\frac{\partial}{\partial w_{0,3}^1}(y_i - \hat{y}_i)^2 = \frac{1}{N}\sum_{i=0}^{N}2(y_i - \hat{y}_i)\,h^3$$



$$\frac{\partial L}{\partial w^1_{0,3}} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y_i})\, h^3$$

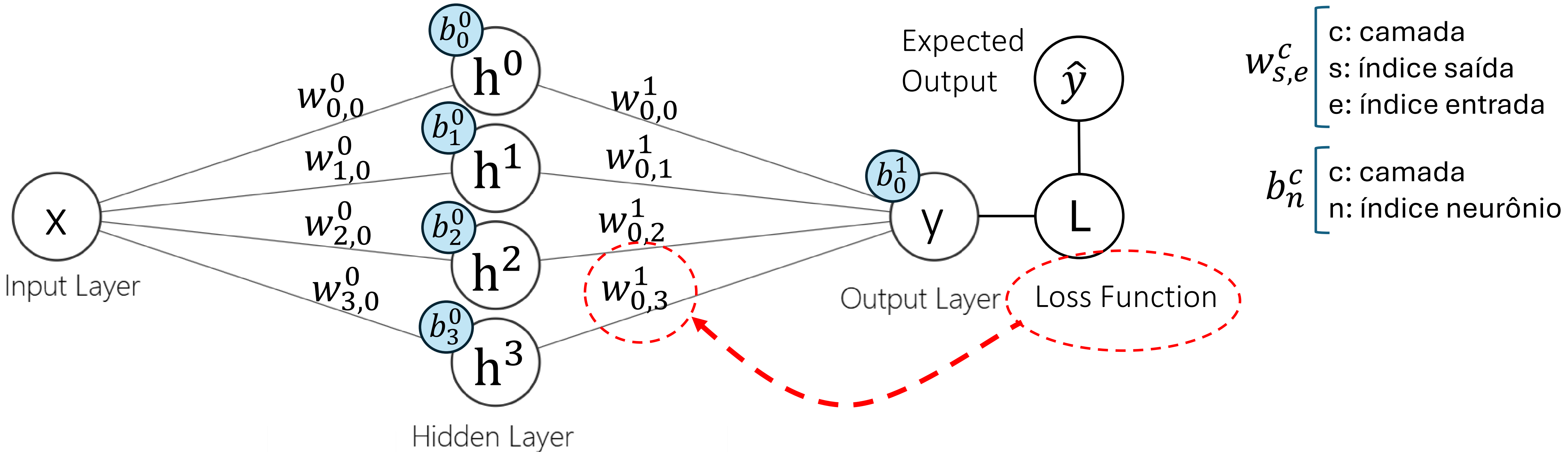
Tire um momento para refletir sobre a intuição da equação



$$\frac{\partial L}{\partial w_{0,3}^1} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y_i})\, h^3$$

- $w_{0,3}^1$ influencia L via y_i

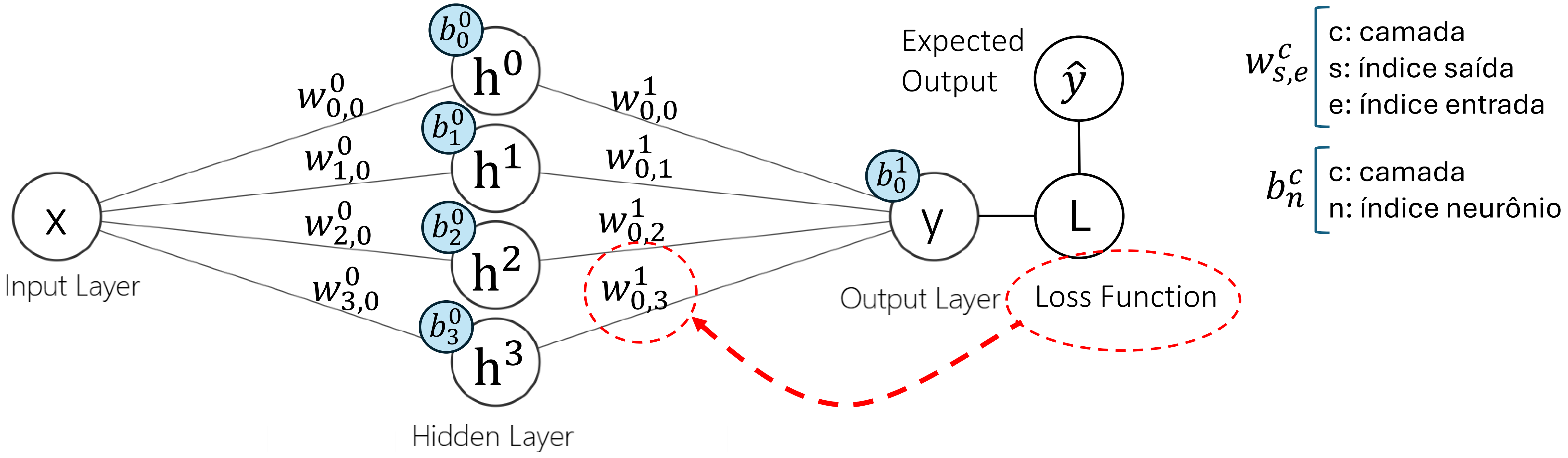
Tire um momento para refletir sobre a intuição da equação



$$\frac{\partial L}{\partial w_{0,3}^1} = \frac{1}{N}\sum_{i=0}^{N} 2\overbrace{(y_i - \hat{y}_i)}\, h^3$$

- $w_{0,3}^1$ influencia L via y_i
- mas o impacto na *loss* também depende de $\hat{y}_i$

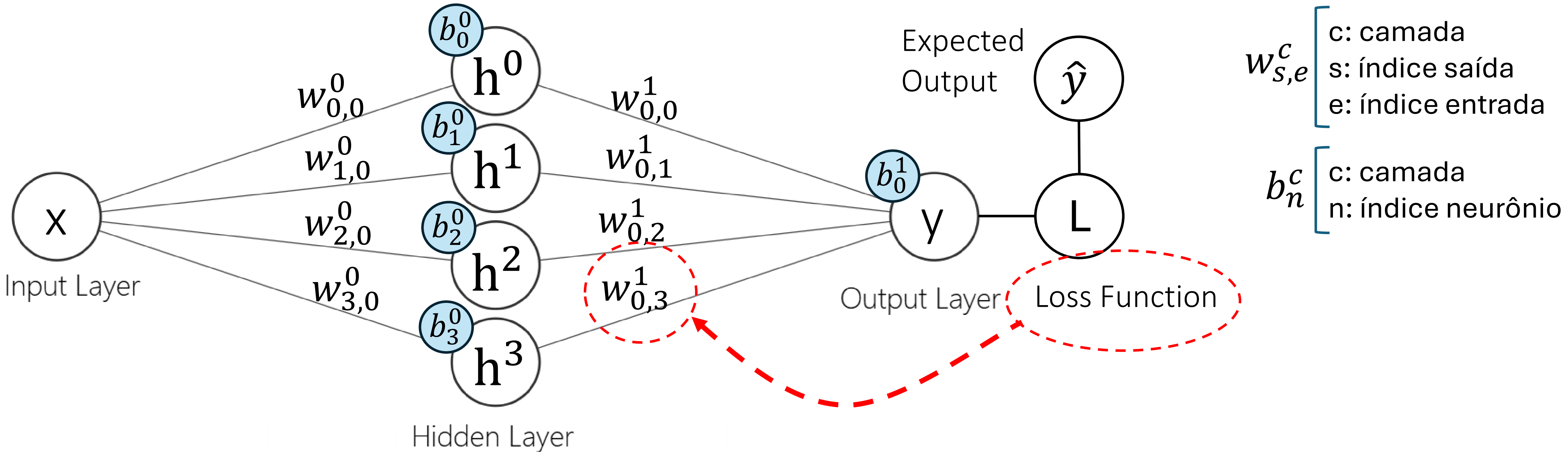
Tire um momento para refletir sobre a intuição da equação



$$\frac{\partial L}{\partial w_{0,3}^1} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y}_i)\, h^3$$

- $w_{0,3}^1$ influencia L via y_i
- mas o impacto na *loss* também depende de $\hat{y}_i$
- h^3 define o quanto $w_{0,3}^1$ vai mudar y_i

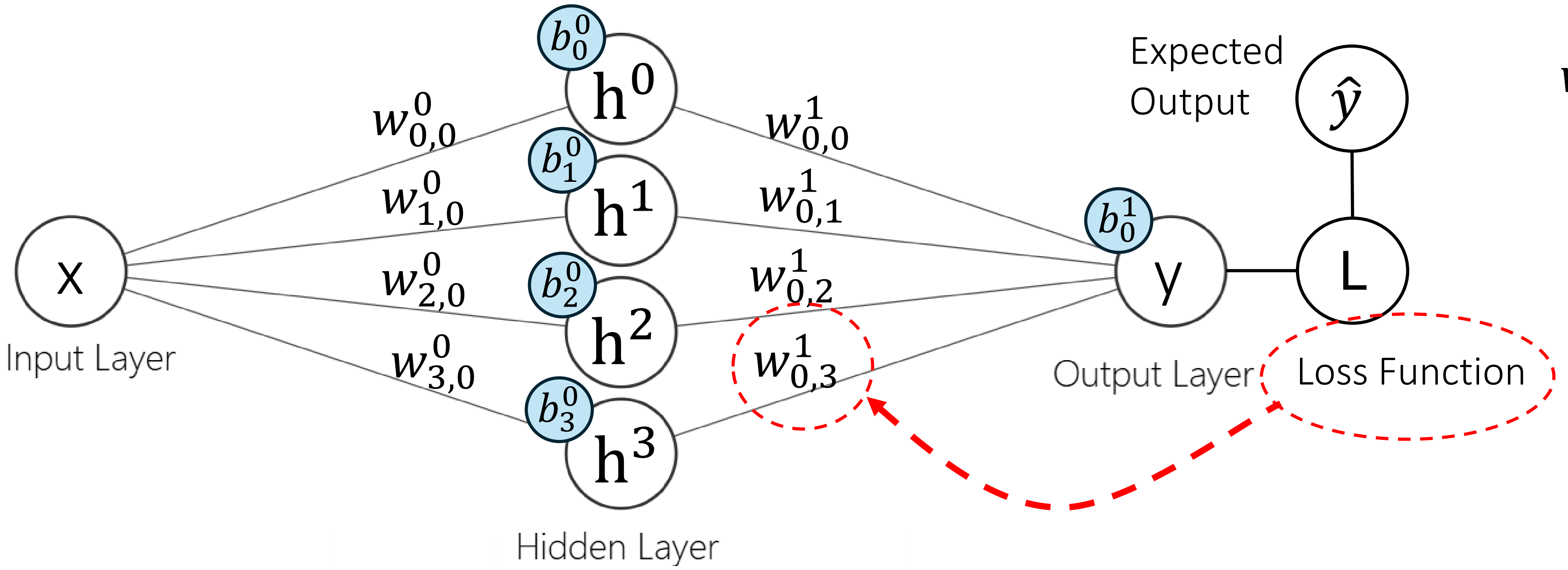
Tire um momento para refletir sobre a intuição da equação



$$\frac{\partial L}{\partial w_{0,3}^1} = \overbrace{\frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y}_i)\, h^3}$$

Tire um momento para refletir sobre a intuição da equação

- $w_{0,3}^1$ influencia L via y_i
- mas o impacto na *loss* também depende de $\hat{y}_i$
- h^3 define o quanto $w_{0,3}^1$ vai mudar y_i
- a modificação em $w_{0,3}^1$ será a média das modificações por amostra de treino.



$w_{s,e}^c$ — c: camada; s: índice saída; e: índice entrada

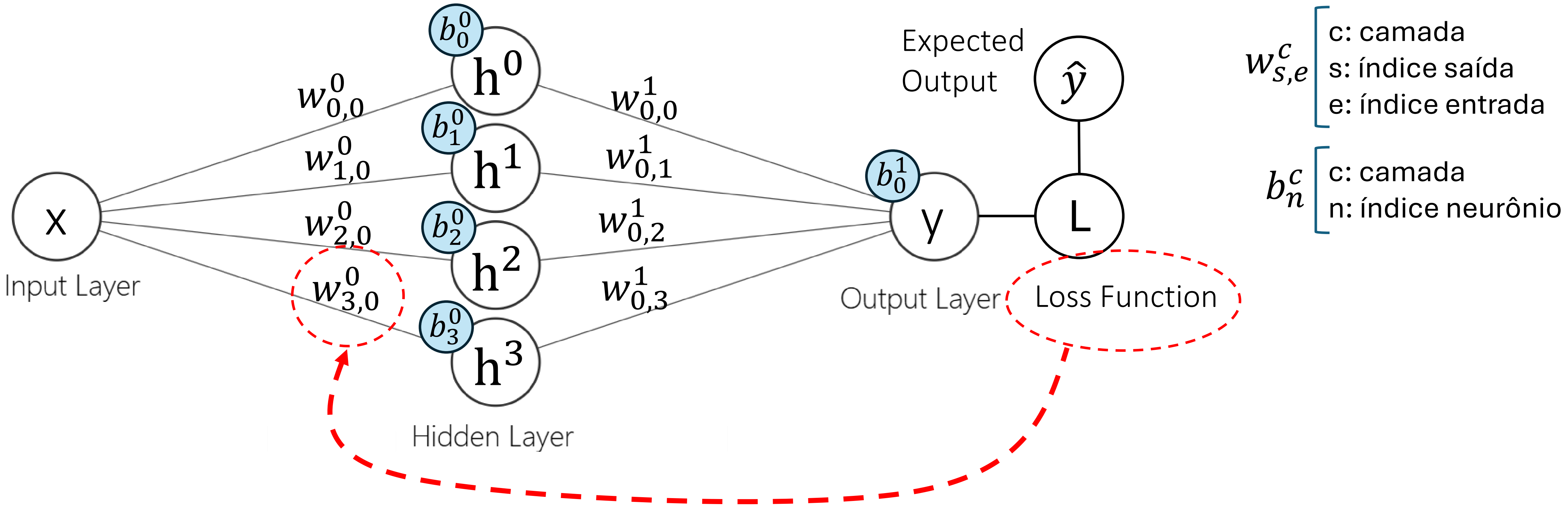
b_n^c — c: camada; n: índice neurônio

$$\frac{\partial L}{\partial w_{0,k}^1} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y}_i)\, h^k \qquad \Bigg| \qquad \frac{\partial L}{\partial b_0^1} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y}_i)$$

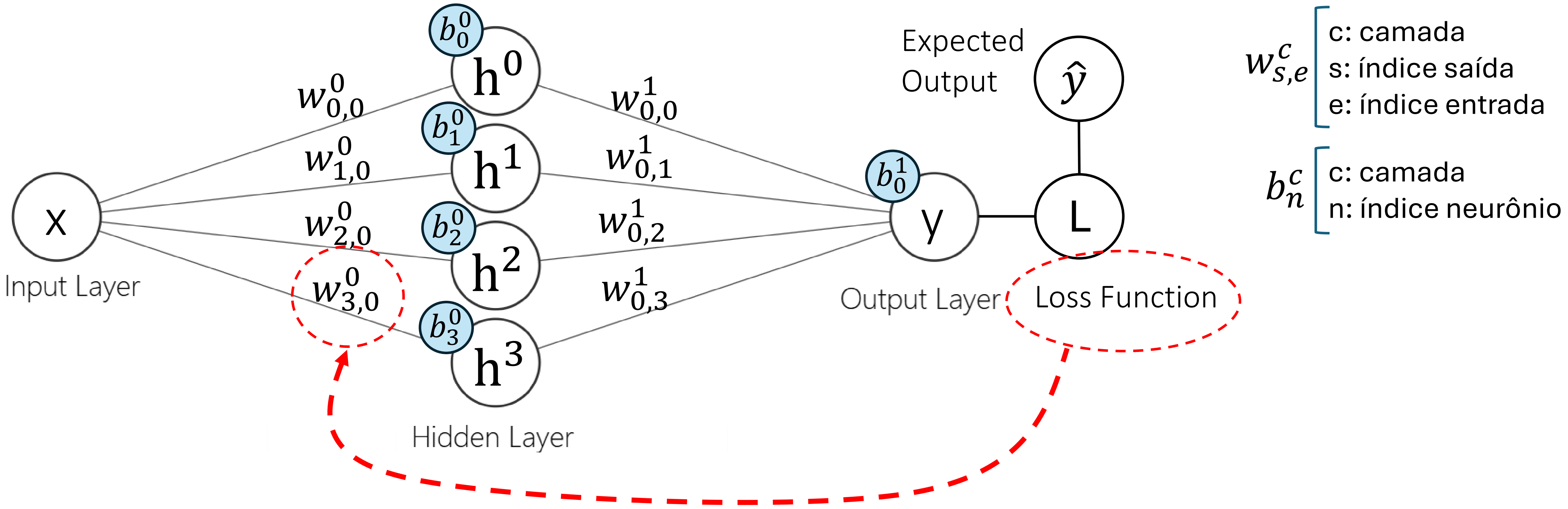
A derivação é análoga para os demais pesos e para o bias, mesmo com camadas intermediárias com mais neurônios.

$$\frac{\partial}{\partial w_{0,k}^1}\left[h^0 w_{0,0}^1 + h^1 w_{0,1}^1 + \cdots + h^k w_{0,k}^1 + \cdots + h^m w_{0,m}^1 + b_0^1\right]$$

(apenas para lembrar como era a expressão que levou às derivadas)

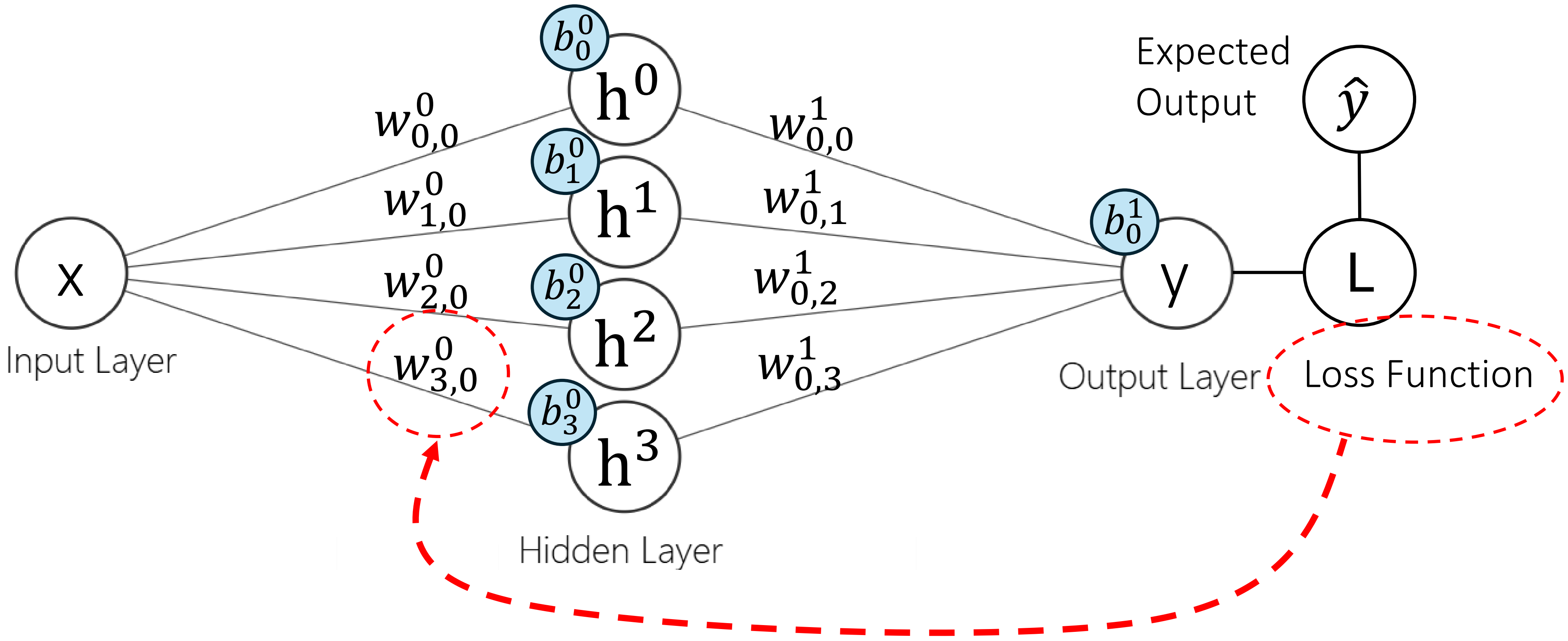


$$\frac{\partial L}{\partial w_{3,0}^0}$$



$$\frac{\partial L}{\partial w^0_{3,0}} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y_i})\frac{\partial}{\partial w^0_{3,0}} y_i$$

$$\frac{\partial}{\partial w^0_{3,0}} y_i = \frac{\partial}{\partial w^0_{3,0}}\left[h^0 w^1_{0,0} + h^1 w^1_{0,1} + h^2 w^1_{0,2} + h^3 w^1_{0,3} + b^1_0\right]$$

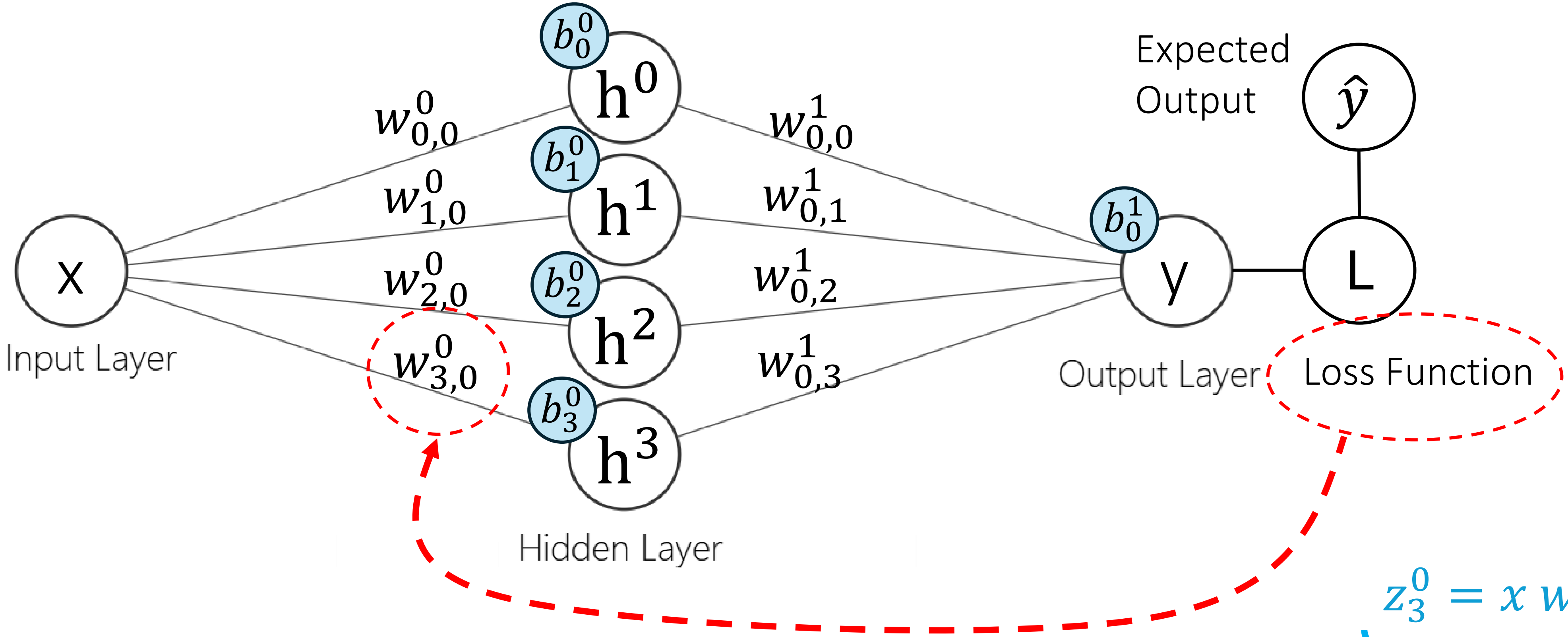


$w^c_{s,e}$
- c: camada
- s: índice saída
- e: índice entrada

b^c_n
- c: camada
- n: índice neurônio

$$\frac{\partial L}{\partial w^0_{3,0}} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y_i})\frac{\partial}{\partial w^0_{3,0}} y_i$$

$$\frac{\partial}{\partial w^0_{3,0}} y_i = \frac{\partial}{\partial w^0_{3,0}}\left[h^0 w^1_{0,0} + h^1 w^1_{0,1} + h^2 w^1_{0,2} + h^3 w^1_{0,3} + b^1_0\right] = w^1_{0,3}\frac{\partial}{\partial w^0_{3,0}} h^3$$



$w^c_{s,e}$
- c: camada
- s: índice saída
- e: índice entrada

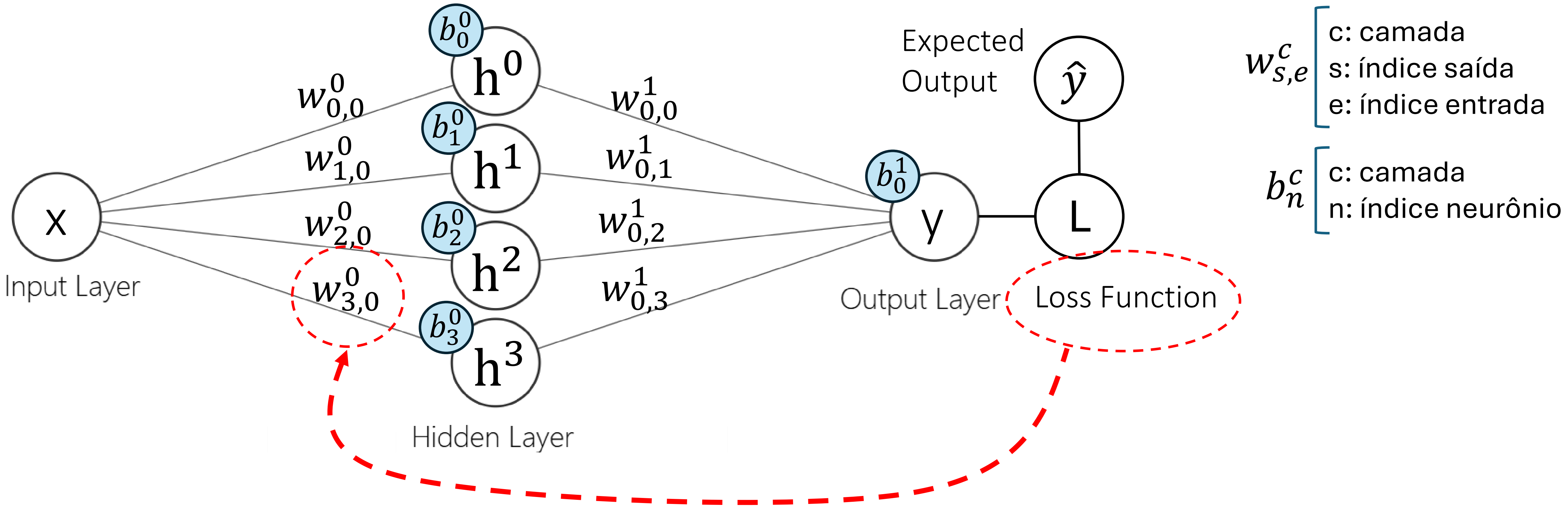
b^c_n
- c: camada
- n: índice neurônio

$$\underbrace{z^0_3 = x\, w^0_{3,0} + b^0_3}_{\text{valor pré-ativação do neurônio 3 da camada 0}}$$

$$\frac{\partial L}{\partial w^0_{3,0}} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y}_i)\frac{\partial}{\partial w^0_{3,0}} y_i$$

$$\frac{\partial}{\partial w^0_{3,0}} h^3 = \frac{\partial}{\partial w^0_{3,0}} \sigma(z^0_3)$$

$$\frac{\partial}{\partial w^0_{3,0}} y_i = w^1_{0,3}\frac{\partial}{\partial w^0_{3,0}} h^3$$

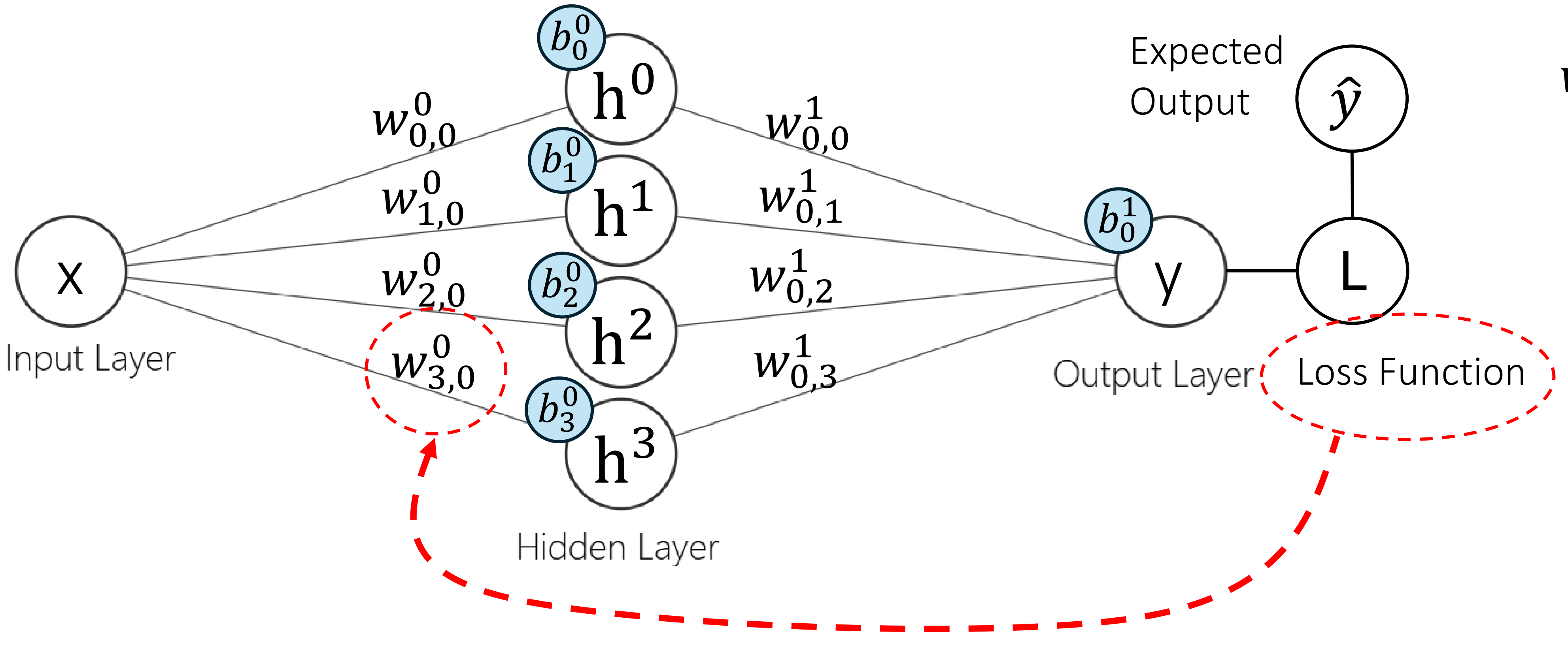


$$\frac{\partial L}{\partial w^0_{3,0}} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y}_i)\frac{\partial}{\partial w^0_{3,0}} y_i$$

$$\frac{\partial}{\partial w^0_{3,0}} y_i = w^1_{0,3}\frac{\partial}{\partial w^0_{3,0}} h^3$$

$$\frac{\partial}{\partial w^0_{3,0}} h^3 = \frac{\partial}{\partial w^0_{3,0}}\sigma(z^0_3)$$

$$= \sigma'(z^0_3)\frac{\partial}{\partial w^0_{3,0}} z^0_3$$



$w_{s,e}^{c}$
- c: camada
- s: índice saída
- e: índice entrada

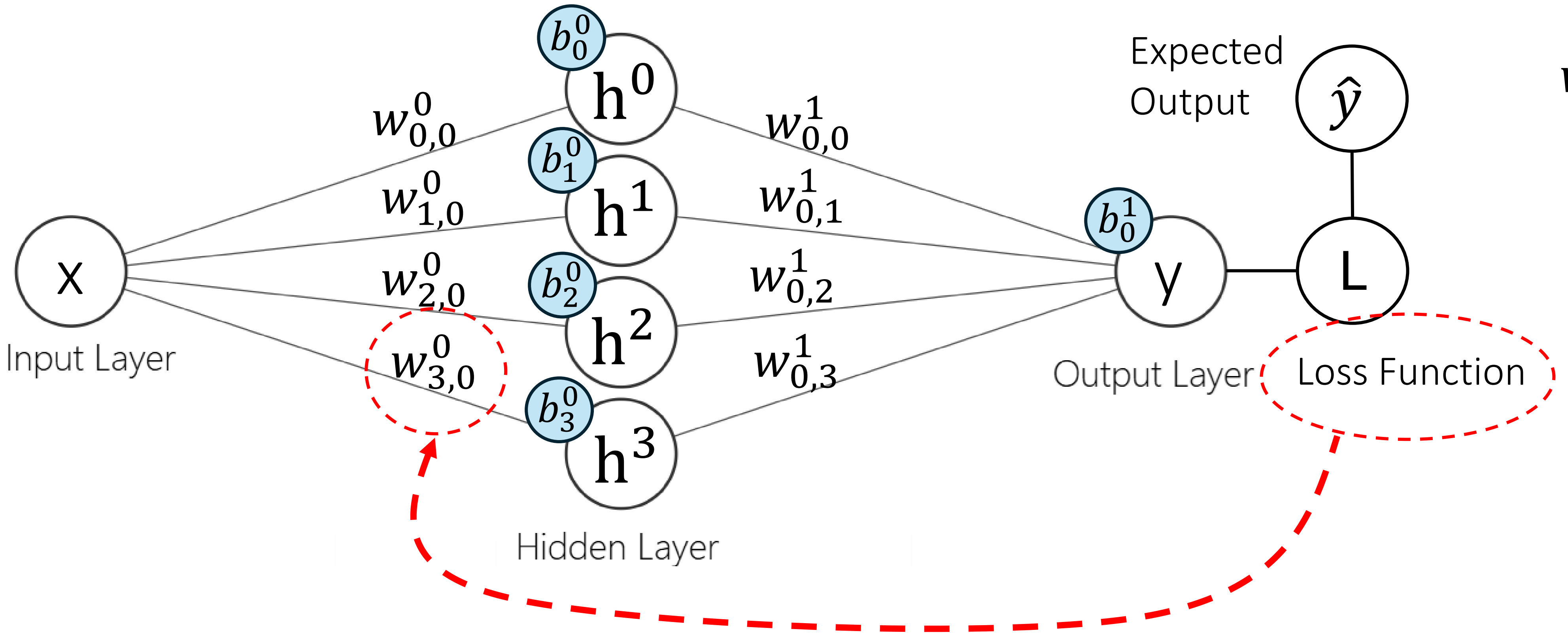
b_n^c
- c: camada
- n: índice neurônio

$$\frac{\partial L}{\partial w_{3,0}^{0}} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y}_i)\frac{\partial}{\partial w_{3,0}^{0}} y_i$$

$$\frac{\partial}{\partial w_{3,0}^{0}} y_i = w_{0,3}^{1}\frac{\partial}{\partial w_{3,0}^{0}} h^3$$

$$\frac{\partial}{\partial w_{3,0}^{0}} h^3 = \frac{\partial}{\partial w_{3,0}^{0}} \sigma(z_3^0)$$

$$= \sigma'(z_3^0)\frac{\partial}{\partial w_{3,0}^{0}}\left[x\, w_{3,0}^{0} + b_3^0\right]$$



$w_{s,e}^c$ c: camada
s: índice saída
e: índice entrada

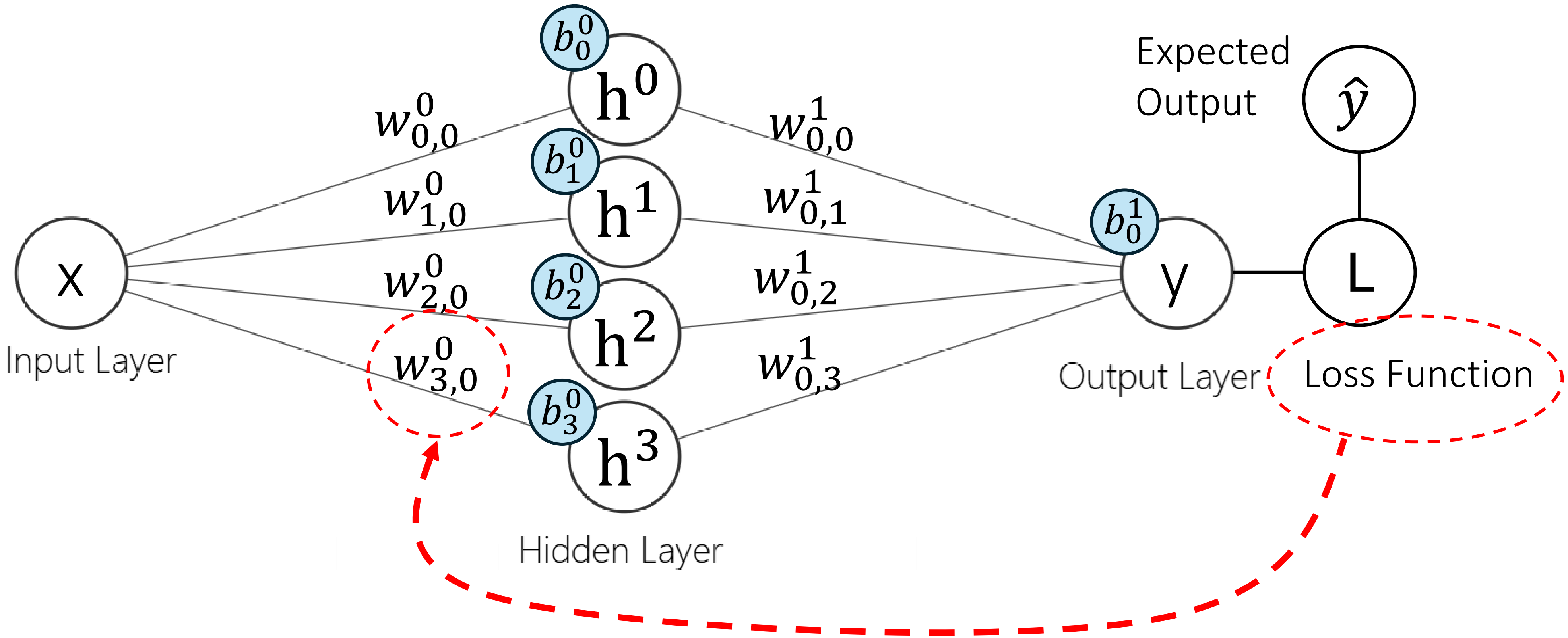
b_n^c c: camada
n: índice neurônio

$$\frac{\partial L}{\partial w_{3,0}^0} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y}_i)\frac{\partial}{\partial w_{3,0}^0} y_i$$

$$\frac{\partial}{\partial w_{3,0}^0} y_i = w_{0,3}^1 \frac{\partial}{\partial w_{3,0}^0} h^3$$

$$\frac{\partial}{\partial w_{3,0}^0} h^3 = \frac{\partial}{\partial w_{3,0}^0} \sigma(z_3^0)$$

$$= \sigma'(z_3^0)\, x$$



$w^c_{s,e}$
- c: camada
- s: índice saída
- e: índice entrada

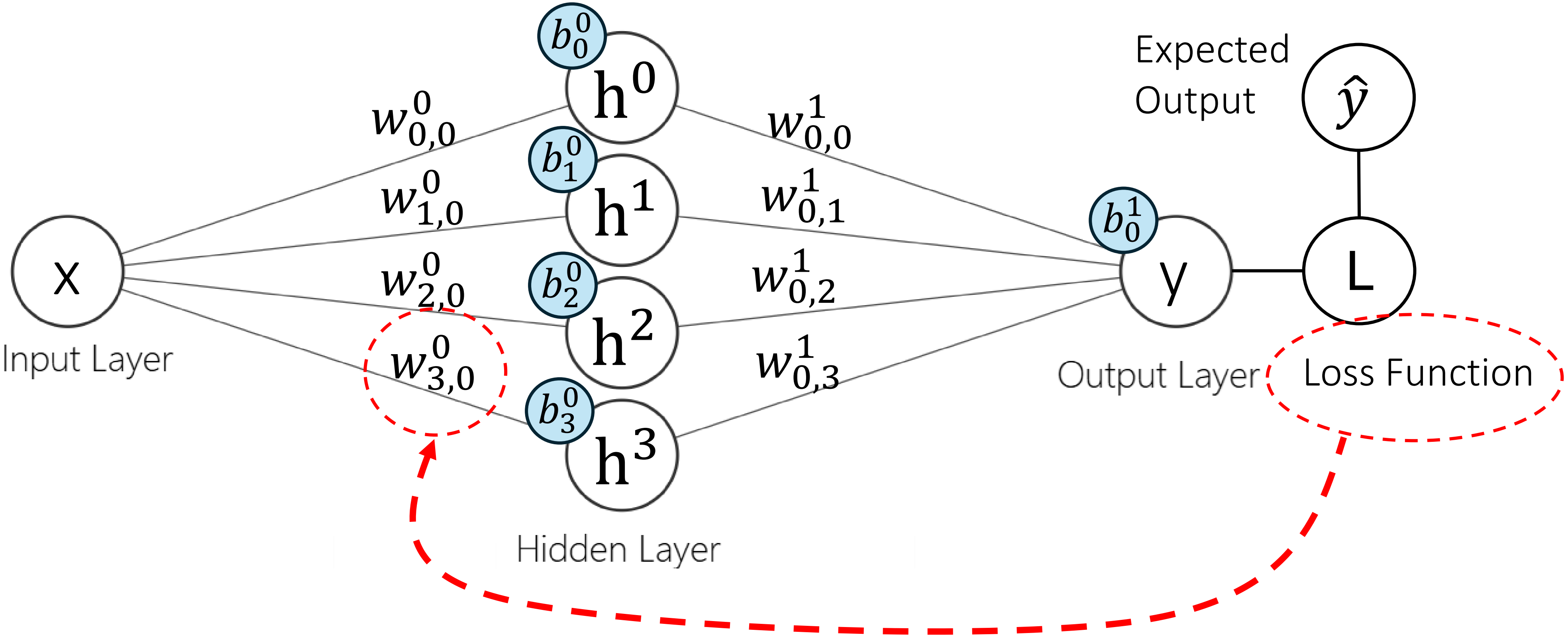
b^c_n
- c: camada
- n: índice neurônio

$$\frac{\partial L}{\partial w^0_{3,0}} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y_i})\frac{\partial}{\partial w^0_{3,0}} y_i$$

$$\frac{\partial}{\partial w^0_{3,0}} h^3 = \frac{\partial}{\partial w^0_{3,0}} \sigma(z^0_3)$$

$$= \sigma'(z^0_3)\, x$$

$$\frac{\partial}{\partial w^0_{3,0}} y_i = w^1_{0,3}\frac{\partial}{\partial w^0_{3,0}} h^3$$



$w_{s,e}^{c}$ — c: camada; s: índice saída; e: índice entrada

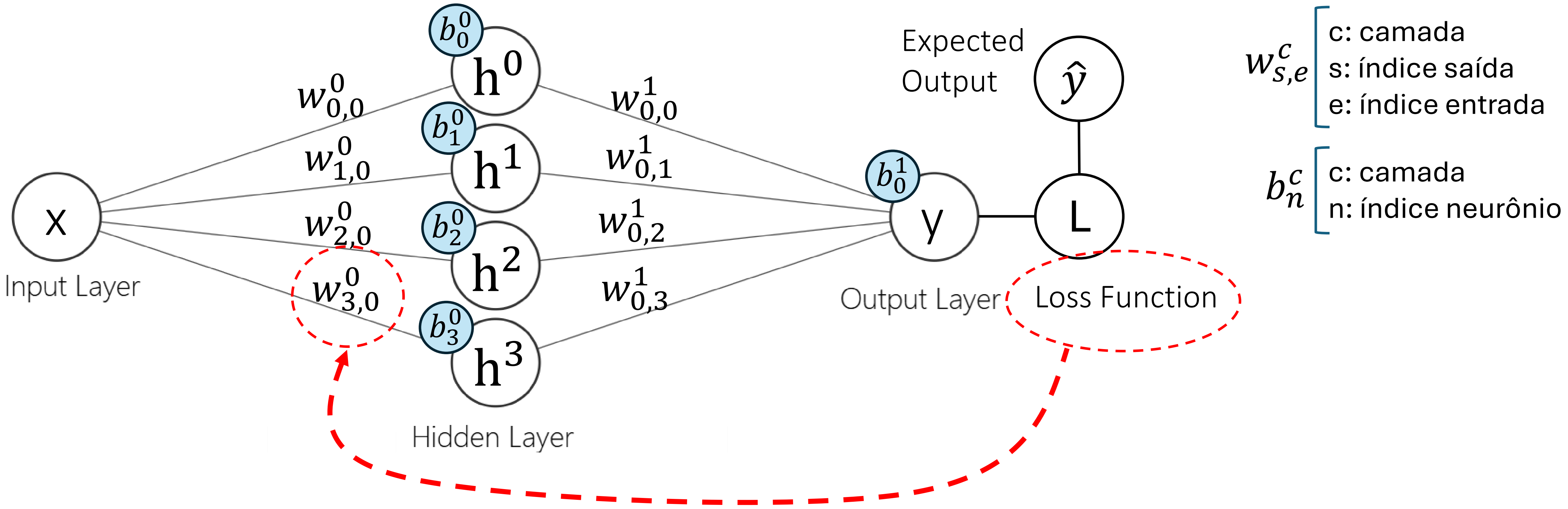
b_{n}^{c} — c: camada; n: índice neurônio

$$\frac{\partial L}{\partial w_{3,0}^{0}} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y}_i)\frac{\partial}{\partial w_{3,0}^{0}} y_i$$

$$\frac{\partial}{\partial w_{3,0}^{0}} h^3 = \frac{\partial}{\partial w_{3,0}^{0}} \sigma(z_3^0)$$

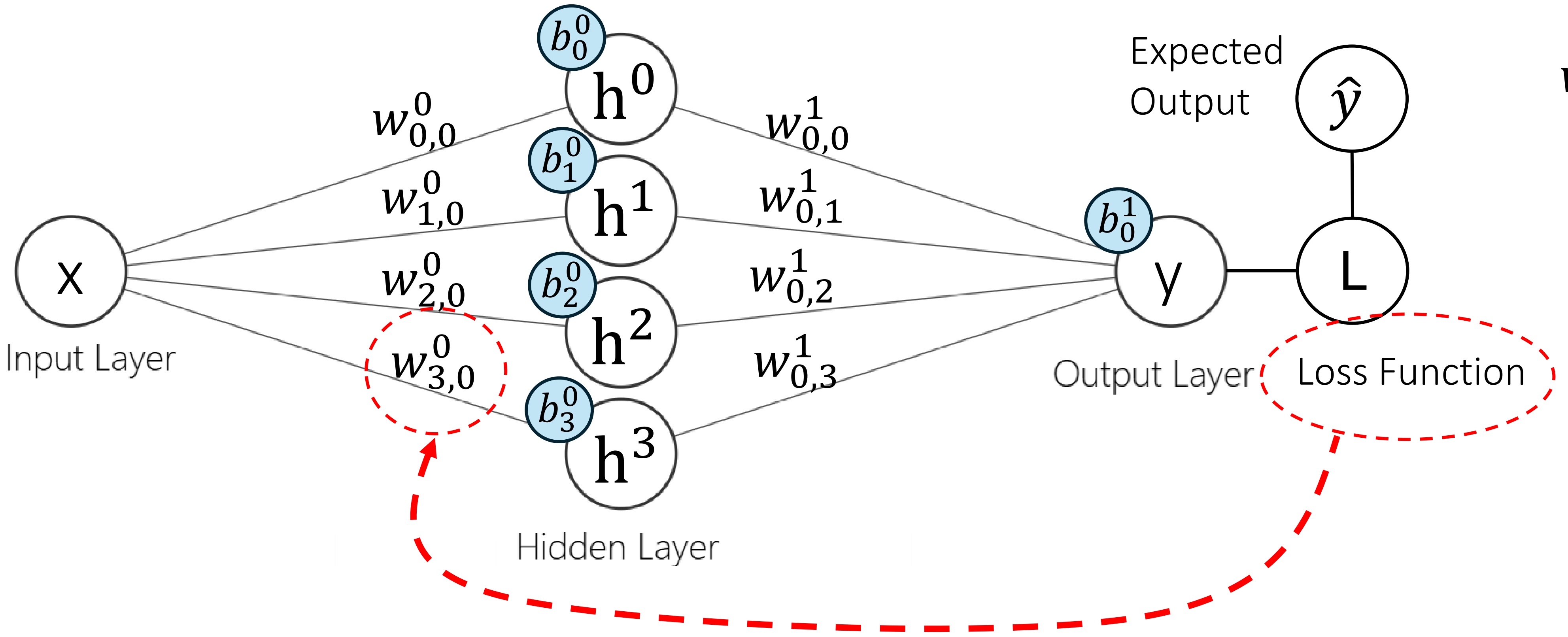
$$= \sigma'(z_3^0)\, x$$

$$\frac{\partial}{\partial w_{3,0}^{0}} y_i = w_{0,3}^{1}\, \sigma'(z_3^0)\, x$$



$$\frac{\partial L}{\partial w_{3,0}^0} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y}_i)\frac{\partial}{\partial w_{3,0}^0} y_i$$

$$\frac{\partial}{\partial w_{3,0}^0} y_i = w_{0,3}^1\, \sigma'(z_3^0)\, x$$

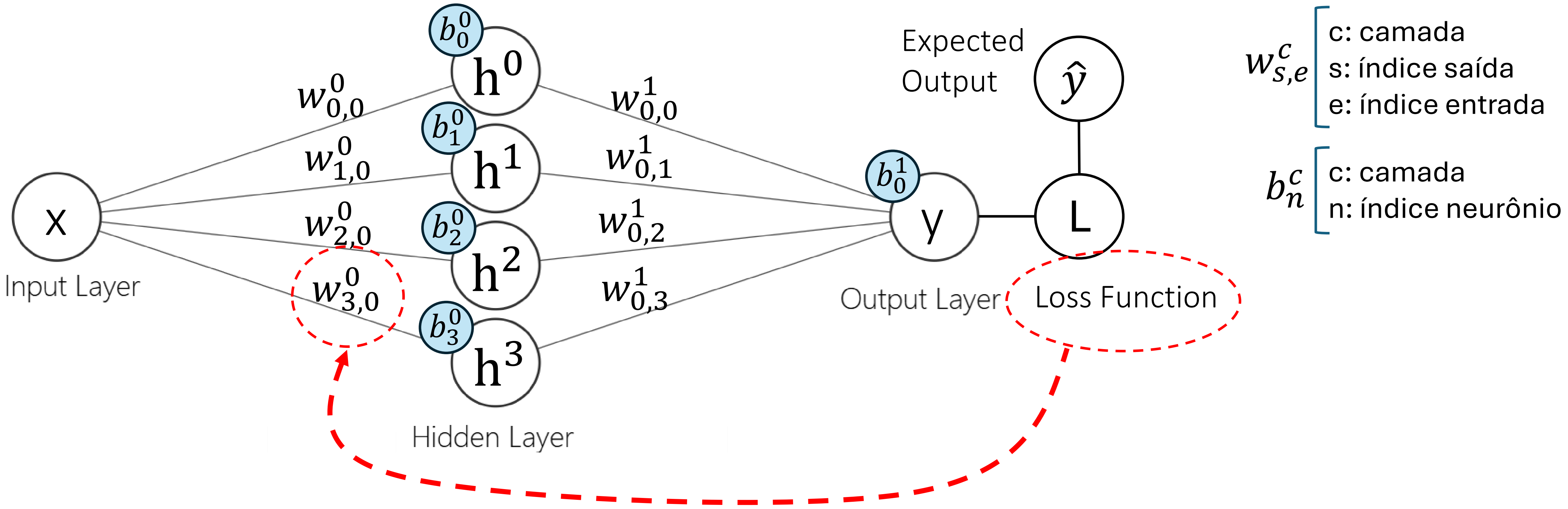


$w_{s,e}^c$
- c: camada
- s: índice saída
- e: índice entrada

b_n^c
- c: camada
- n: índice neurônio

$$\frac{\partial L}{\partial w_{3,0}^0} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y_i})\, w_{0,3}^1\, \sigma'(z_3^0)\, x$$

$$\frac{\partial}{\partial w_{3,0}^0} y_i = w_{0,3}^1\, \sigma'(z_3^0)\, x$$

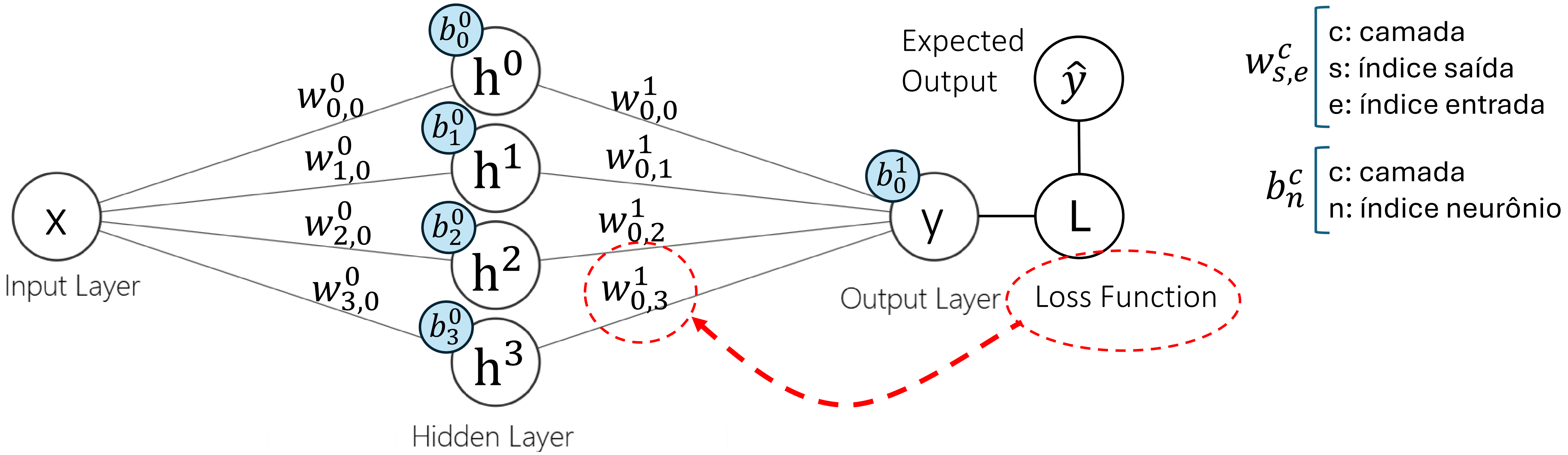


$$\left.\begin{aligned} \frac{\partial L}{\partial w^0_{k,0}} &= \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y}_i)\, w^1_{0,k}\, \sigma'\left(z^0_k\right) x \\ \frac{\partial L}{\partial b^0_k} &= \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y}_i)\, w^1_{0,k}\, \sigma'\left(z^0_k\right) \end{aligned}\right\}$$

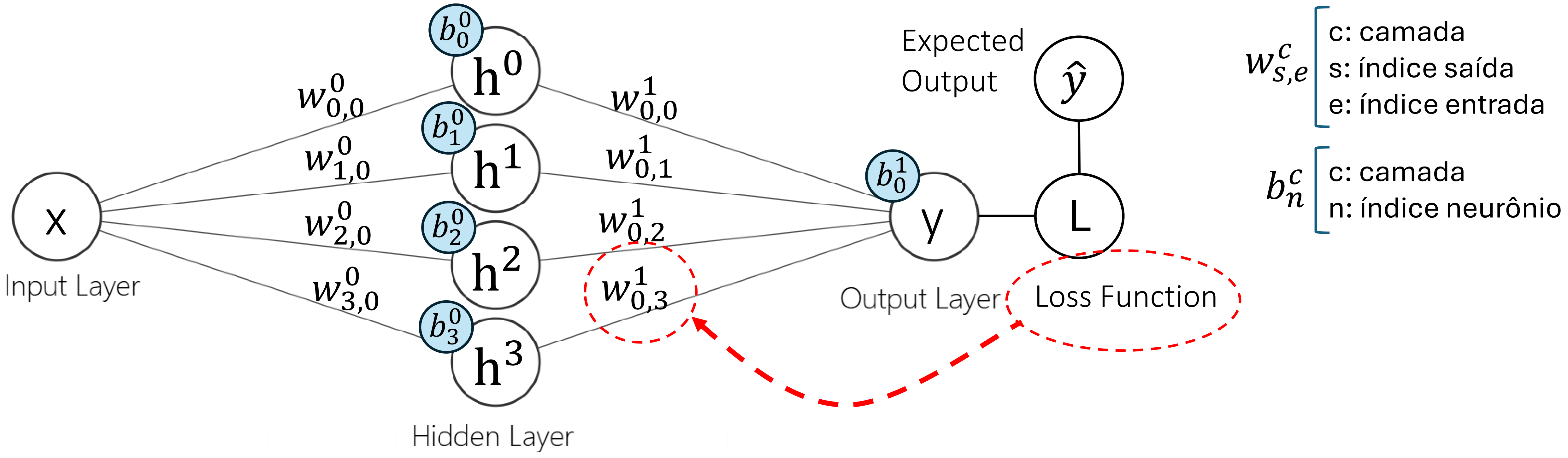
Generalizando para um peso e bias qualquer da camada 0

A próxima seção refaz a derivação anterior resolvendo as derivadas parciais que compõe a regra da cadeia separadamente ao invés de tentar resolver a derivada parcial da função de perda em relação aos pesos diretamente. Este caminho pode ser mais fácil de digerir para algumas pessoas.

Olhando com outra perspectiva: Tornando a regra de cadeia explícita e resolvendo os fatores independentemente



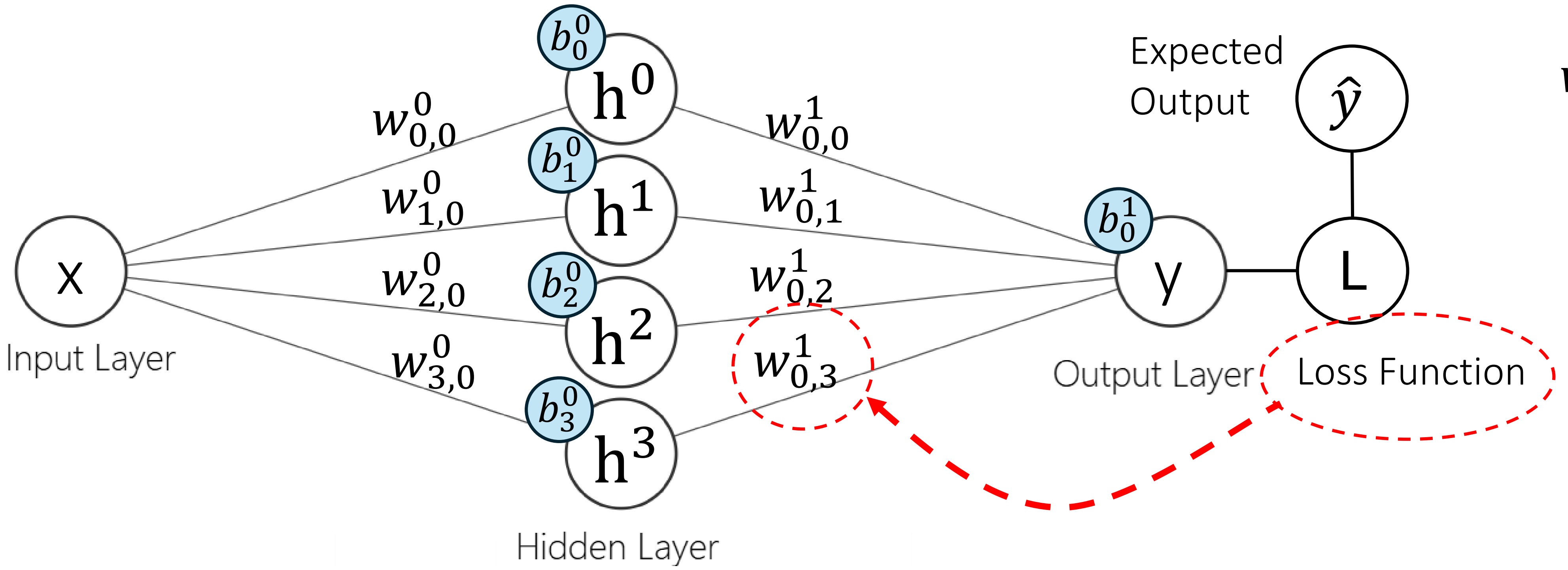
$$\frac{\partial L}{\partial w_{0,3}^1} = \frac{\partial L}{\partial y_i} \frac{\partial y_i}{\partial w_{0,3}^1}$$



$$\frac{\partial L}{\partial w_{0,3}^1} = \frac{\partial L}{\partial y_i}\frac{\partial y_i}{\partial w_{0,3}^1}$$

$$\frac{\partial L}{\partial y_i} = \frac{\partial}{\partial y_i}\frac{1}{N}\sum_{i=0}^{N}(y_i - \hat{y_i})^2$$

Resolvemos a derivada parcial assumindo que y_i é nossa variável de interesse (sem considerar que na verdade queremos a derivada parcial em relação a $w_{0,3}^1$ e que y_i é uma função de $w_{0,3}^1$.

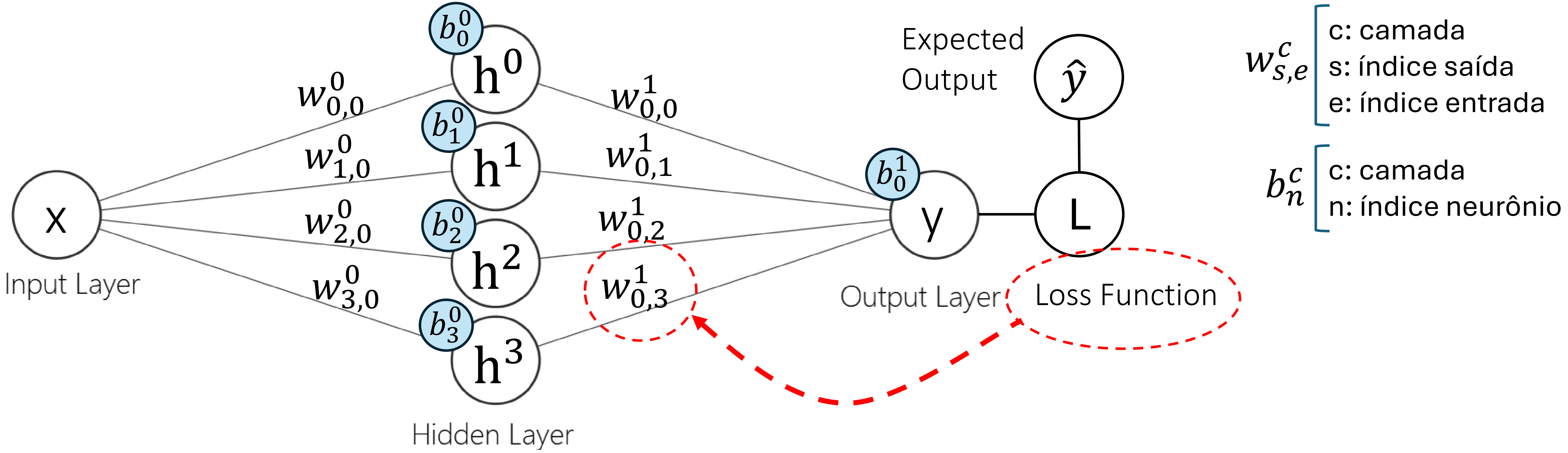


$w_{s,e}^c$ c: camada
s: índice saída
e: índice entrada

b_n^c c: camada
n: índice neurônio

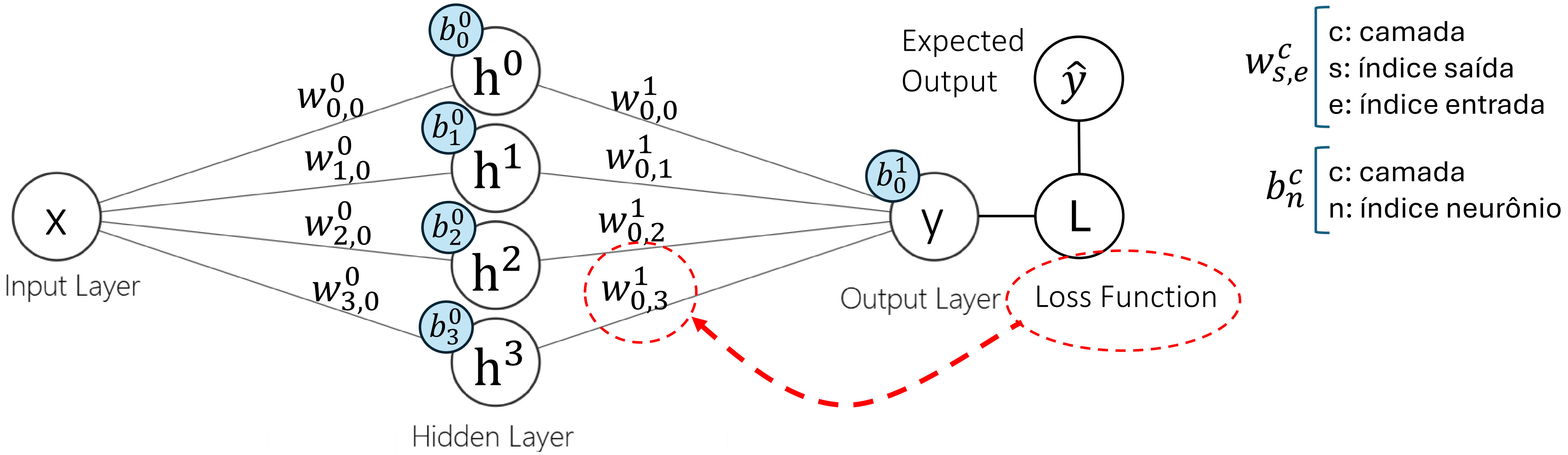
$$\frac{\partial L}{\partial w_{0,3}^1} = \frac{\partial L}{\partial y_i} \frac{\partial y_i}{\partial w_{0,3}^1}$$

$$\frac{\partial L}{\partial y_i} = \frac{\partial}{\partial y_i} \frac{1}{N} \sum_{i=0}^{N} (y_i - \hat{y_i})^2 = \frac{1}{N} \sum_{i=0}^{N} \frac{\partial}{\partial y_i} (y_i - \hat{y_i})^2$$



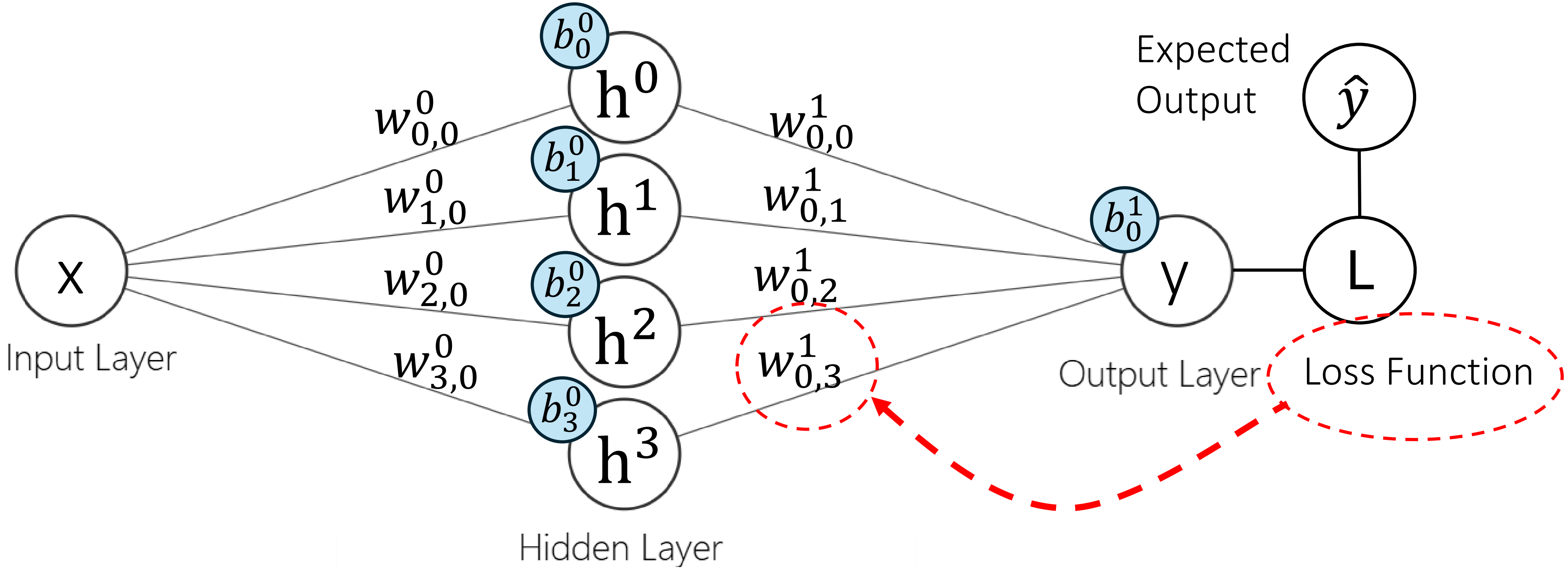
$$\frac{\partial L}{\partial w_{0,3}^1} = \frac{\partial L}{\partial y_i}\frac{\partial y_i}{\partial w_{0,3}^1}$$

$$\frac{\partial L}{\partial y_i} = \frac{\partial}{\partial y_i}\frac{1}{N}\sum_{i=0}^{N}(y_i - \hat{y_i})^2 = \frac{1}{N}\sum_{i=0}^{N}\frac{\partial}{\partial y_i}(y_i - \hat{y_i})^2 = \frac{1}{N}\sum_{i=0}^{N}2(y_i - \hat{y_i})$$



$$\frac{\partial L}{\partial w_{0,3}^1} = \frac{\partial L}{\partial y_i}\frac{\partial y_i}{\partial w_{0,3}^1}$$

$$\frac{\partial L}{\partial y_i} = \frac{\partial}{\partial y_i}\frac{1}{N}\sum_{i=0}^{N}(y_i - \hat{y_i})^2 = \frac{1}{N}\sum_{i=0}^{N}\frac{\partial}{\partial y_i}(y_i - \hat{y_i})^2 = \frac{1}{N}\sum_{i=0}^{N}2(y_i - \hat{y_i})$$

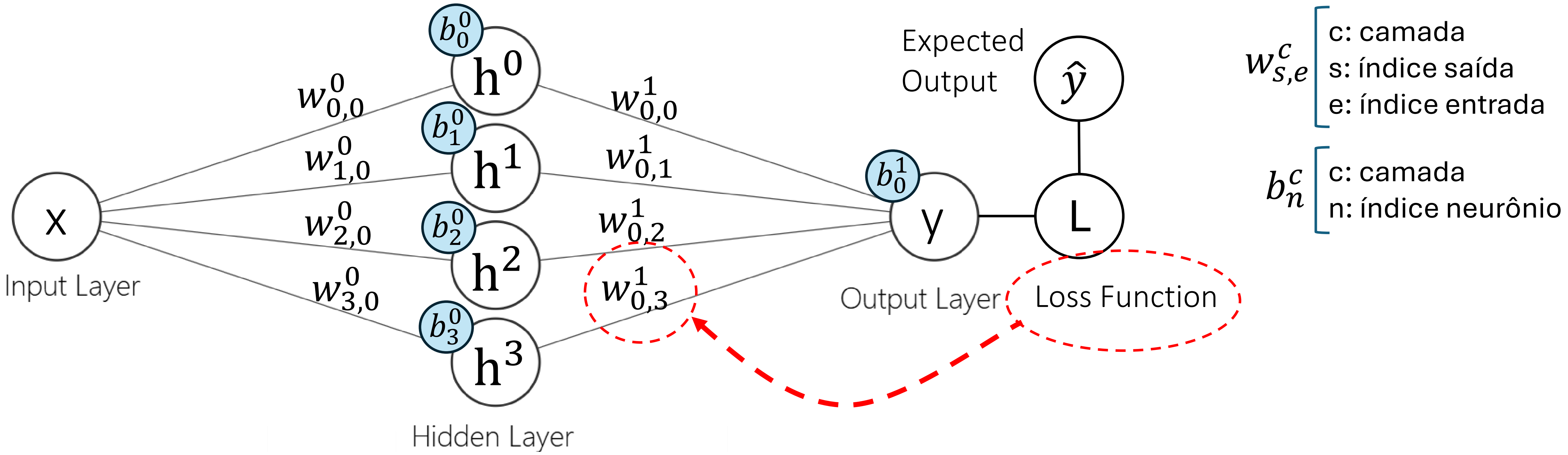


$w^c_{s,e}$
- c: camada
- s: índice saída
- e: índice entrada

b^c_n
- c: camada
- n: índice neurônio

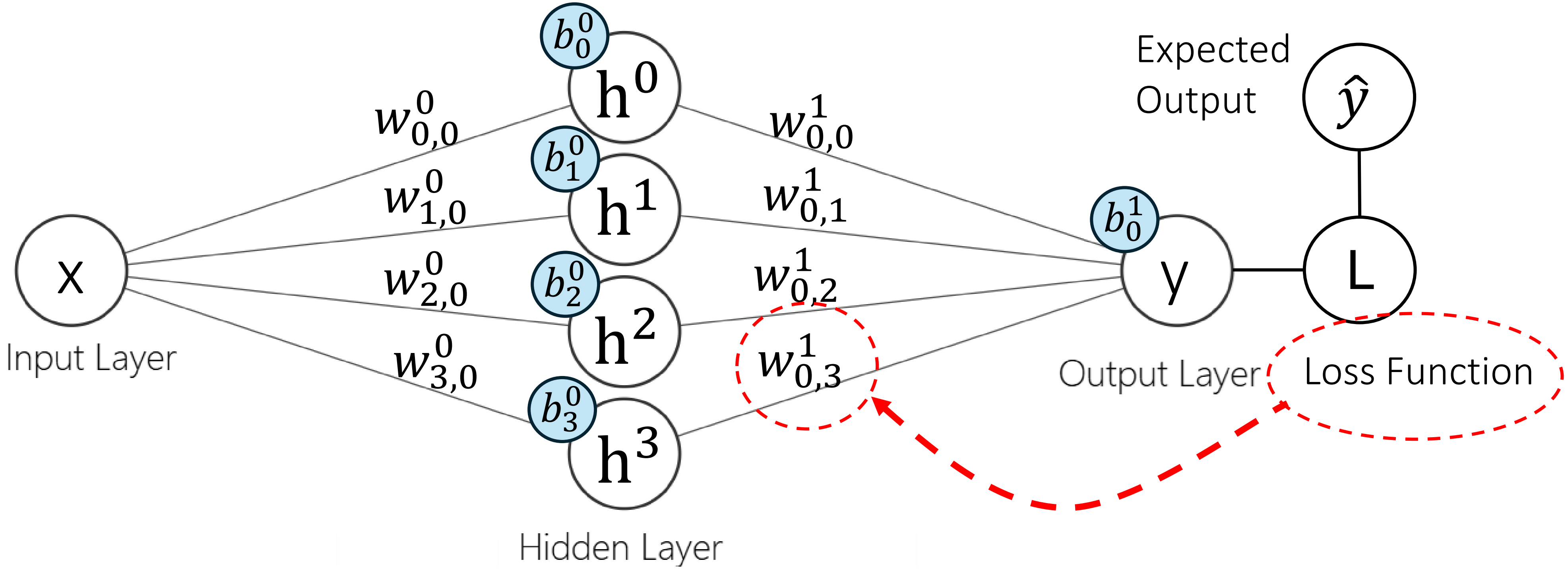
$$\frac{\partial L}{\partial w^1_{0,3}} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \widehat{y_i})\,\frac{\partial y_i}{\partial w^1_{0,3}}$$

$$\frac{\partial L}{\partial y_i} = \frac{\partial}{\partial y_i}\frac{1}{N}\sum_{i=0}^{N}(y_i - \widehat{y_i})^2 = \frac{1}{N}\sum_{i=0}^{N}\frac{\partial}{\partial y_i}(y_i - \widehat{y_i})^2 = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \widehat{y_i})$$



$$\frac{\partial L}{\partial w_{0,3}^1} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y}_i)\,\frac{\partial y_i}{\partial w_{0,3}^1}$$

$$\frac{\partial y_i}{\partial w_{0,3}^1} = \frac{\partial}{\partial w_{0,3}^1}\left[h^0 w_{0,0}^1 + h^1 w_{0,1}^1 + h^2 w_{0,2}^1 + h^3 w_{0,3}^1 + b_0^1\right]$$

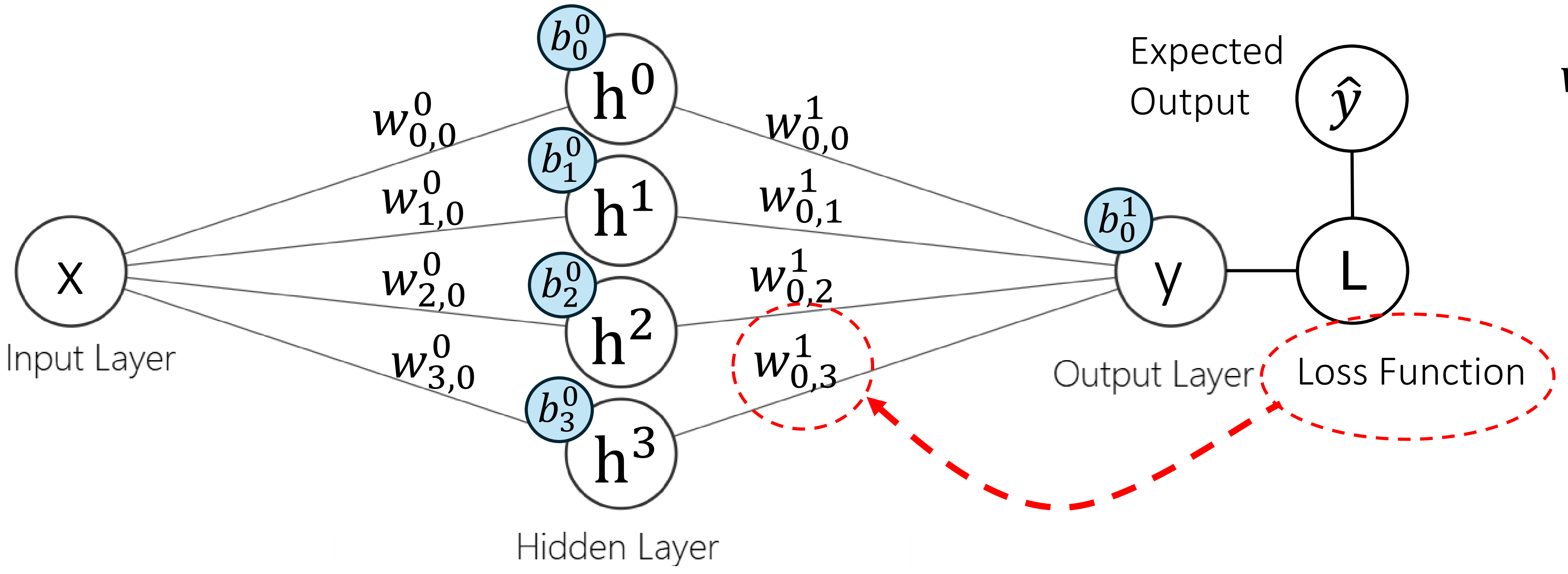


$w^c_{s,e}$ c: camada; s: índice saída; e: índice entrada

b^c_n c: camada; n: índice neurônio

$$\frac{\partial L}{\partial w^1_{0,3}} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y}_i)\frac{\partial y_i}{\partial w^1_{0,3}}$$

$$\frac{\partial y_i}{\partial w^1_{0,3}} = \frac{\partial}{\partial w^1_{0,3}}\left[h^0 w^1_{0,0} + h^1 w^1_{0,1} + h^2 w^1_{0,2} + h^3 w^1_{0,3} + b^1_0\right] = h^3$$

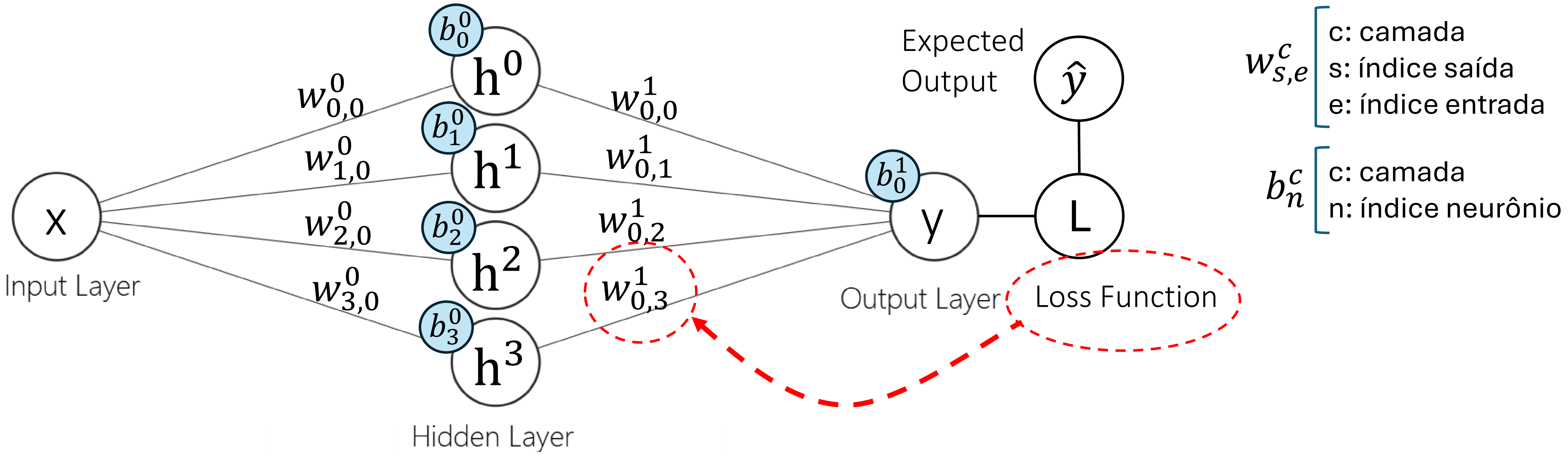


$w^c_{s,e}$ c: camada
s: índice saída
e: índice entrada

b^c_n c: camada
n: índice neurônio

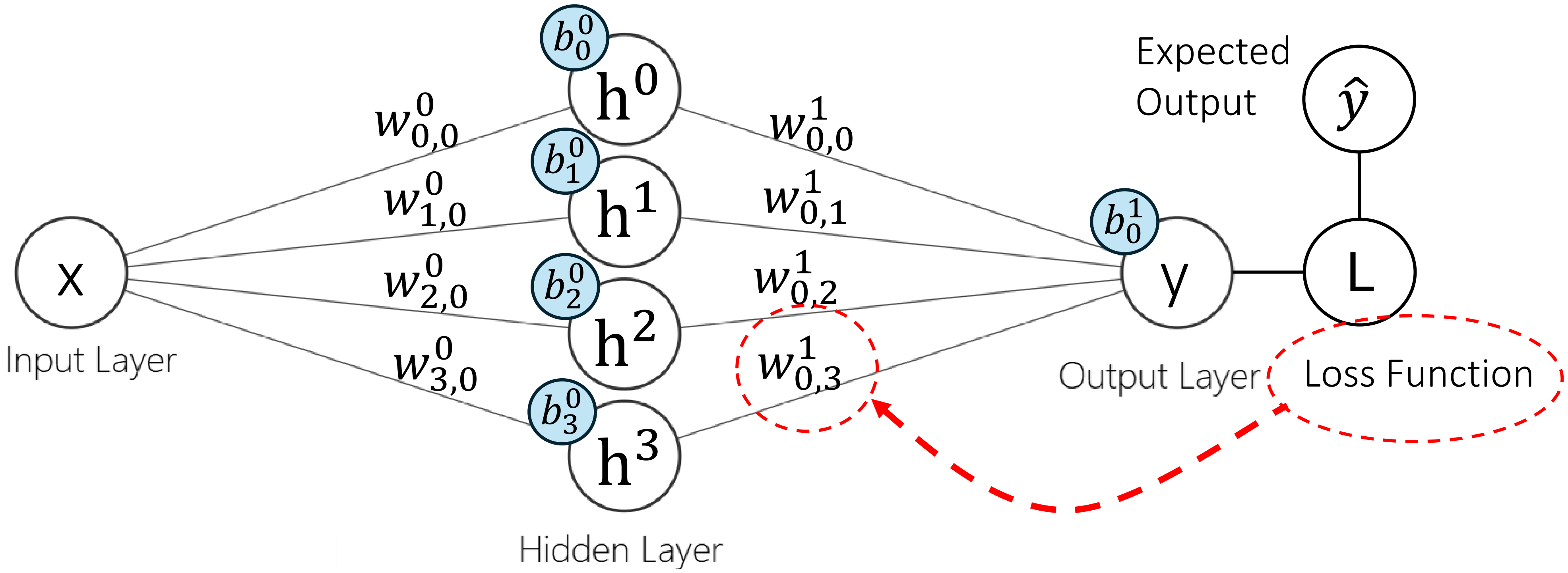
$$\frac{\partial L}{\partial w^1_{0,3}} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y}_i)\frac{\partial y_i}{\partial w^1_{0,3}}$$

$$\frac{\partial y_i}{\partial w^1_{0,3}} = \frac{\partial}{\partial w^1_{0,3}}\left[h^0 w^1_{0,0} + h^1 w^1_{0,1} + h^2 w^1_{0,2} + h^3 w^1_{0,3} + b^1_0\right] = h^3$$



$$\frac{\partial L}{\partial w_{0,3}^1} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y}_i)\, h^3$$

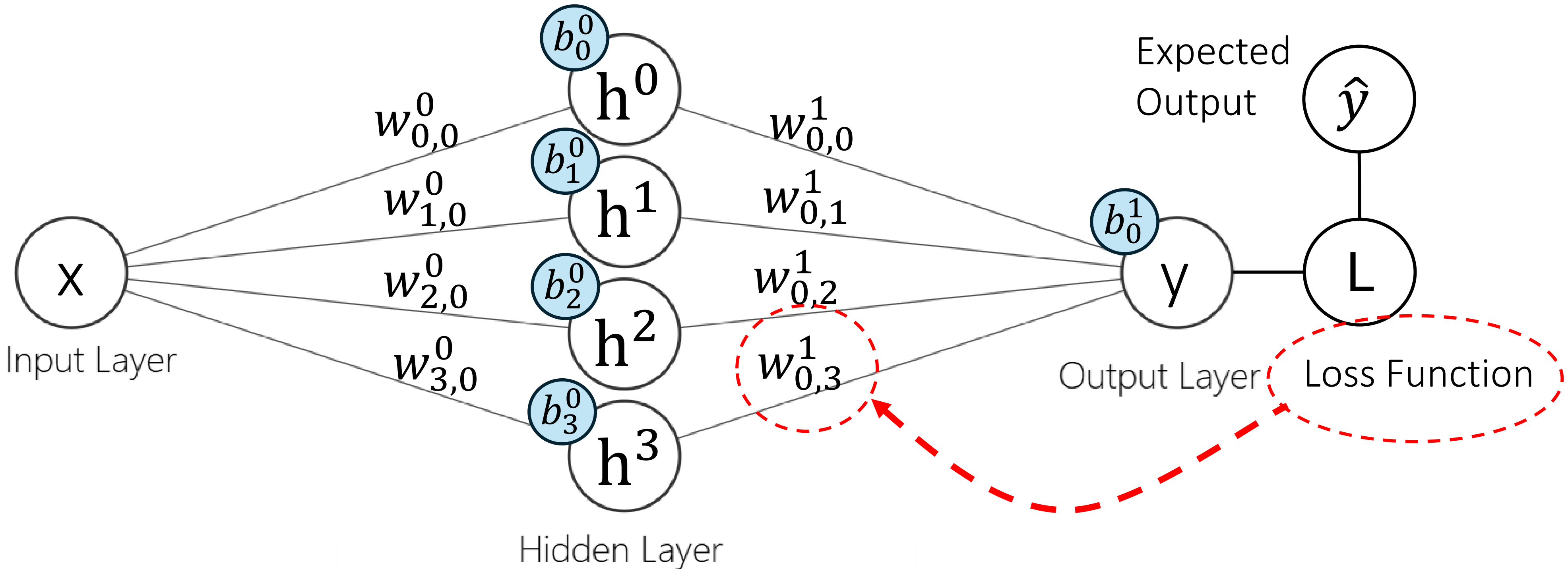
$$\frac{\partial y_i}{\partial w_{0,3}^1} = \frac{\partial}{\partial w_{0,3}^1}\left[h^0 w_{0,0}^1 + h^1 w_{0,1}^1 + h^2 w_{0,2}^1 + h^3 w_{0,3}^1 + b_0^1\right] = h^3$$



$w_{s,e}^{c}$ c: camada
s: índice saída
e: índice entrada

b_n^c c: camada
n: índice neurônio

$$\frac{\partial L}{\partial w_{0,3}^1} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y_i})\, h^3$$

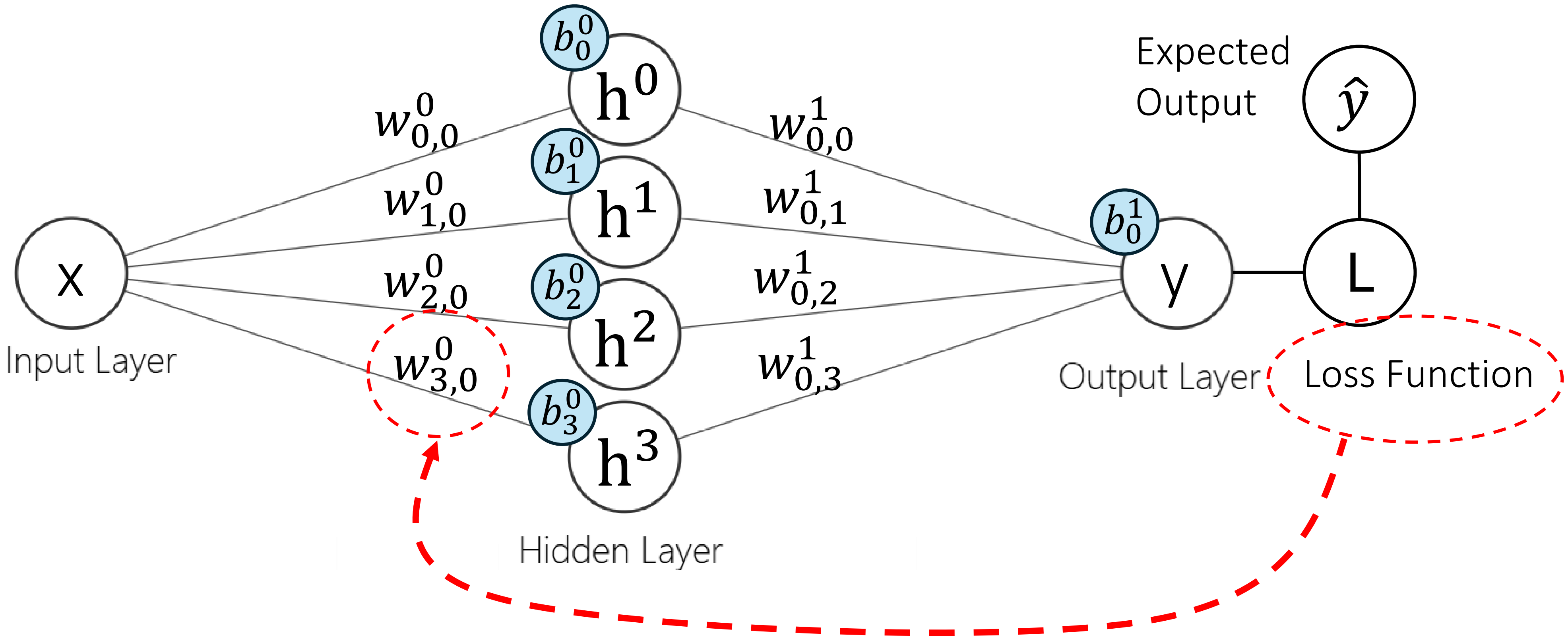


$w_{s,e}^c$ c: camada
s: índice saída
e: índice entrada

b_n^c c: camada
n: índice neurônio

Generalizando para o k-ésimo peso e para o bias da última camada:

$$\frac{\partial L}{\partial w_{0,k}^1} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \widehat{y_i})\, h^k \qquad \frac{\partial L}{\partial b_0^1} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \widehat{y_i})$$



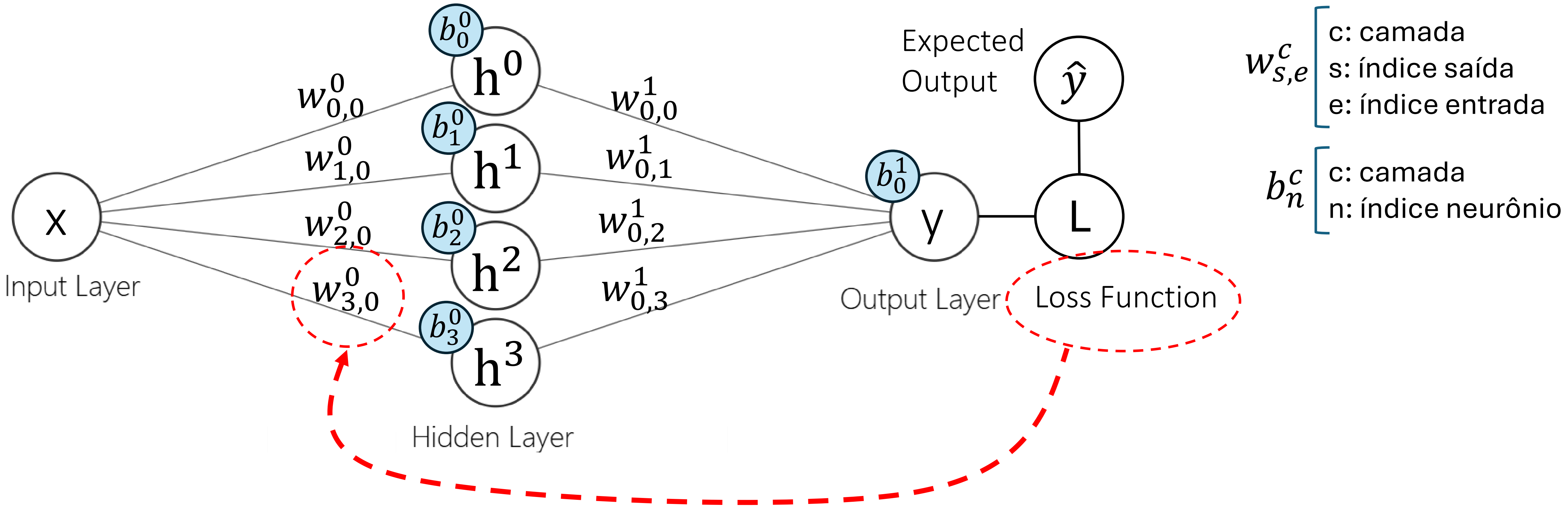
$w_{s,e}^c$ c: camada
s: índice saída
e: índice entrada

b_n^c c: camada
n: índice neurônio

$$\frac{\partial L}{\partial w_{3,0}^0} = \frac{\partial L}{\partial y} \frac{\partial y}{\partial h^3} \frac{\partial h^3}{\partial z^3} \frac{\partial z^3}{\partial w_{3,0}^0}$$

$$\frac{\partial L}{\partial y_i} = \frac{1}{N} \sum_{i=0}^{N} 2(y_i - \hat{y_i})$$

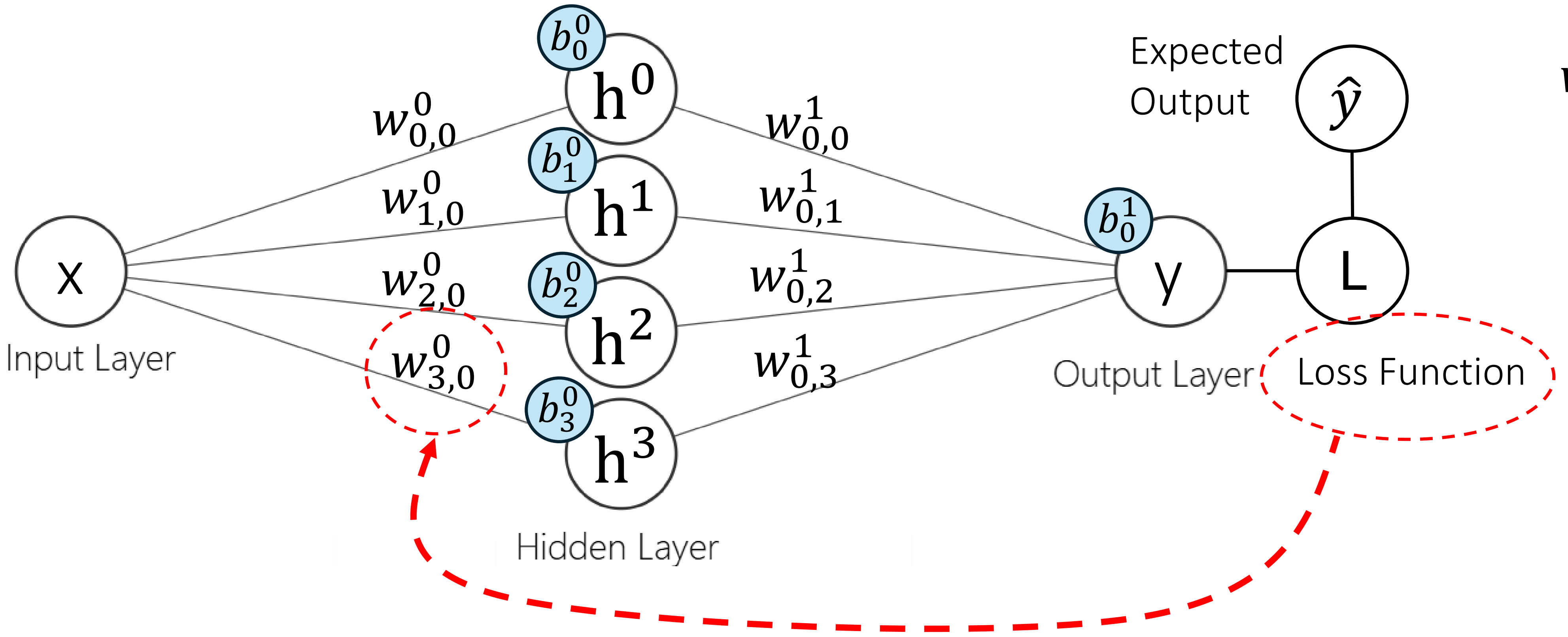
da derivação da última camada



$$\frac{\partial L}{\partial w^0_{3,0}} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y_i}) \frac{\partial y}{\partial h^3}\frac{\partial h^3}{\partial z^3}\frac{\partial z^3}{\partial w^0_{3,0}}$$

$$\frac{\partial L}{\partial y_i} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y_i})$$

da derivação da última camada



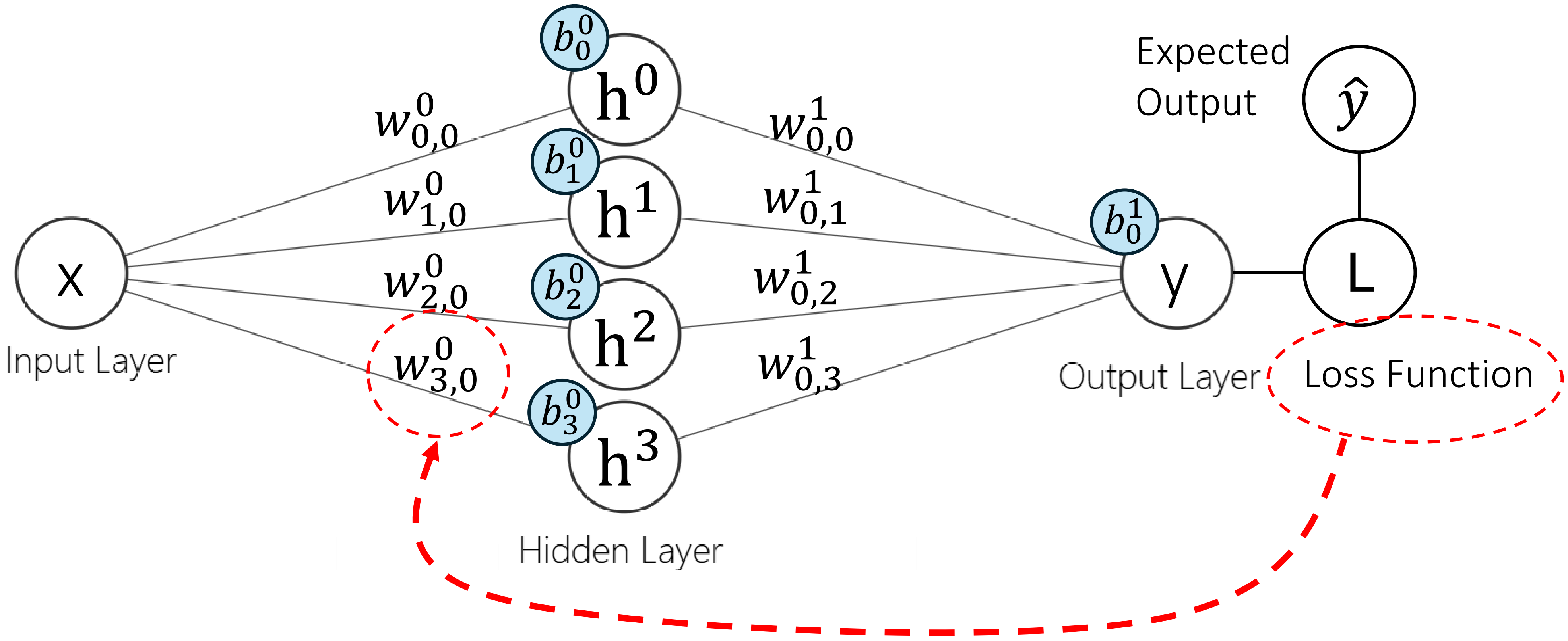
$w_{s,e}^c$ — c: camada; s: índice saída; e: índice entrada

b_n^c — c: camada; n: índice neurônio

$$\frac{\partial L}{\partial w_{3,0}^0} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y_i})\frac{\partial y}{\partial h^3}\frac{\partial h^3}{\partial z^3}\frac{\partial z^3}{\partial w_{3,0}^0}$$

$$\frac{\partial y}{\partial h^3} = \frac{\partial}{\partial h^3}\left[h^0 w_{0,0}^1 + h^1 w_{0,1}^1 + h^2 w_{0,2}^1 + h^3 w_{0,3}^1 + b_0^1\right]$$

Novamente, resolvemos a derivada parcial como se h^3 fosse nossa variável de interesse

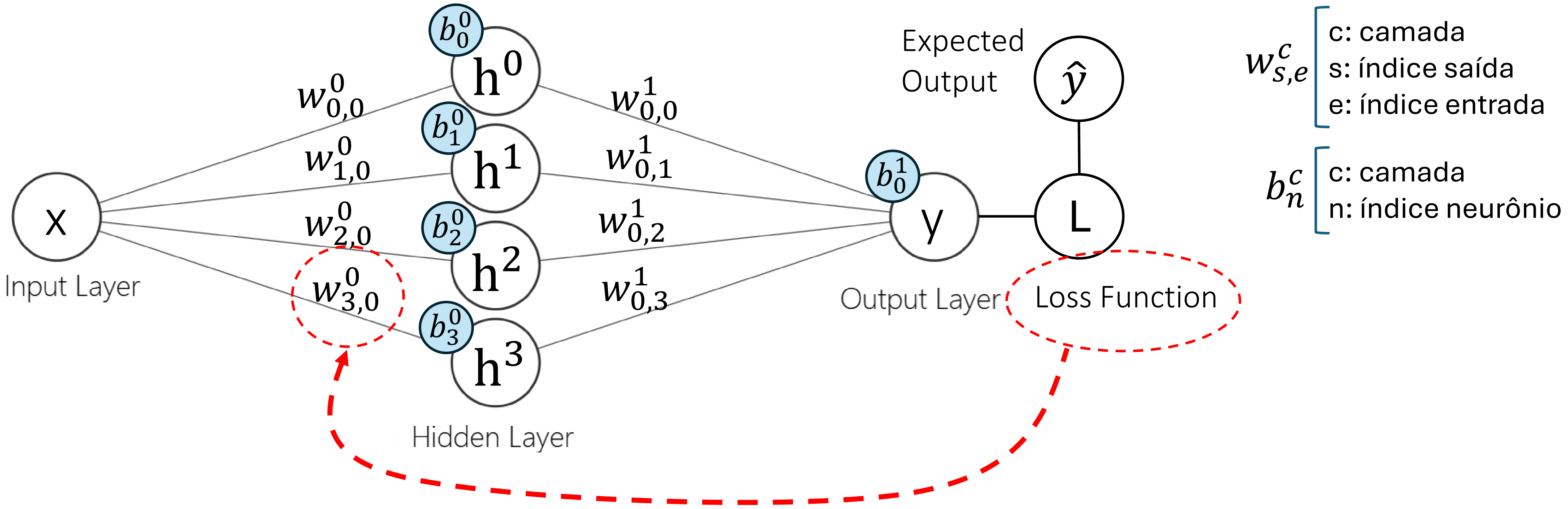


$w_{s,e}^c$
- c: camada
- s: índice saída
- e: índice entrada

b_n^c
- c: camada
- n: índice neurônio

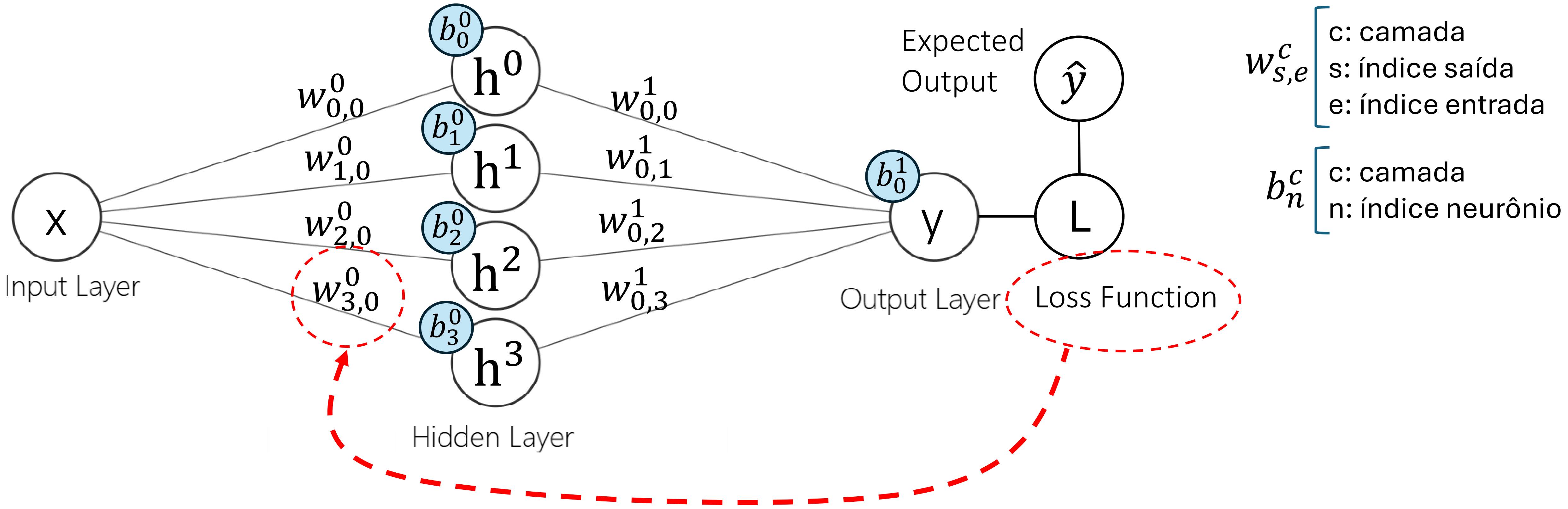
$$\frac{\partial L}{\partial w_{3,0}^0} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y_i})\, \frac{\partial y}{\partial h^3}\frac{\partial h^3}{\partial z^3}\frac{\partial z^3}{\partial w_{3,0}^0}$$

$$\frac{\partial y}{\partial h^3} = \frac{\partial}{\partial h^3}\left[h^0 w_{0,0}^1 + h^1 w_{0,1}^1 + h^2 w_{0,2}^1 + h^3 w_{0,3}^1 + b_0^1\right] = w_{0,3}^1$$



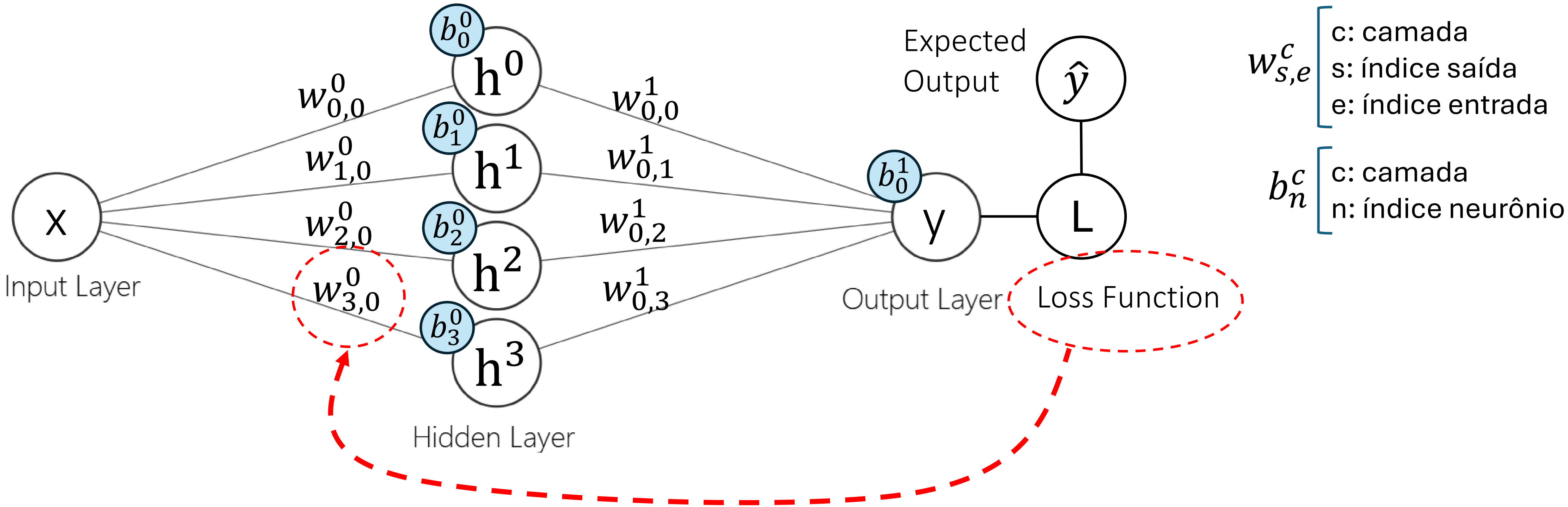
$$\frac{\partial L}{\partial w_{3,0}^{0}} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y_i})\, w_{0,3}^{1}\, \frac{\partial h^3}{\partial z^3}\frac{\partial z^3}{\partial w_{3,0}^{0}}$$

$$\frac{\partial y}{\partial h^3} = \frac{\partial}{\partial h^3}\left[h^0 w_{0,0}^1 + h^1 w_{0,1}^1 + h^2 w_{0,2}^1 + h^3 w_{0,3}^1 + b_0^1\right] = w_{0,3}^1$$



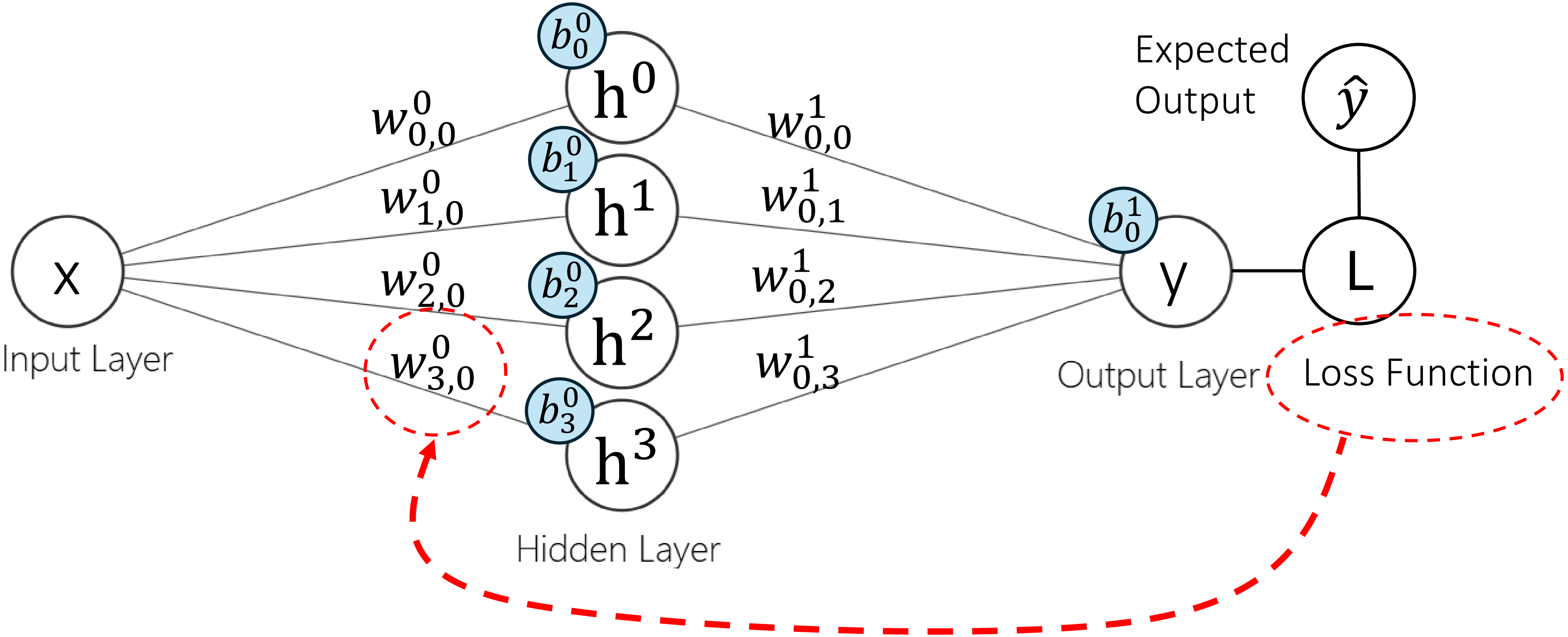
$$\frac{\partial L}{\partial w^0_{3,0}} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y}_i)\, w^1_{0,3}\, \frac{\partial h^3}{\partial z^3} \frac{\partial z^3}{\partial w^0_{3,0}}$$

$$\frac{\partial h^3}{\partial z^3} = \frac{\partial}{\partial z^3}\sigma(z^0_3)$$



$$\frac{\partial L}{\partial w_{3,0}^{0}} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y}_i)\, w_{0,3}^{1} \frac{\partial h^3}{\partial z^3} \frac{\partial z^3}{\partial w_{3,0}^{0}}$$

$$\frac{\partial h^3}{\partial z^3} = \frac{\partial}{\partial z^3}\sigma(z_3^0) = \sigma'(z_3^0)$$

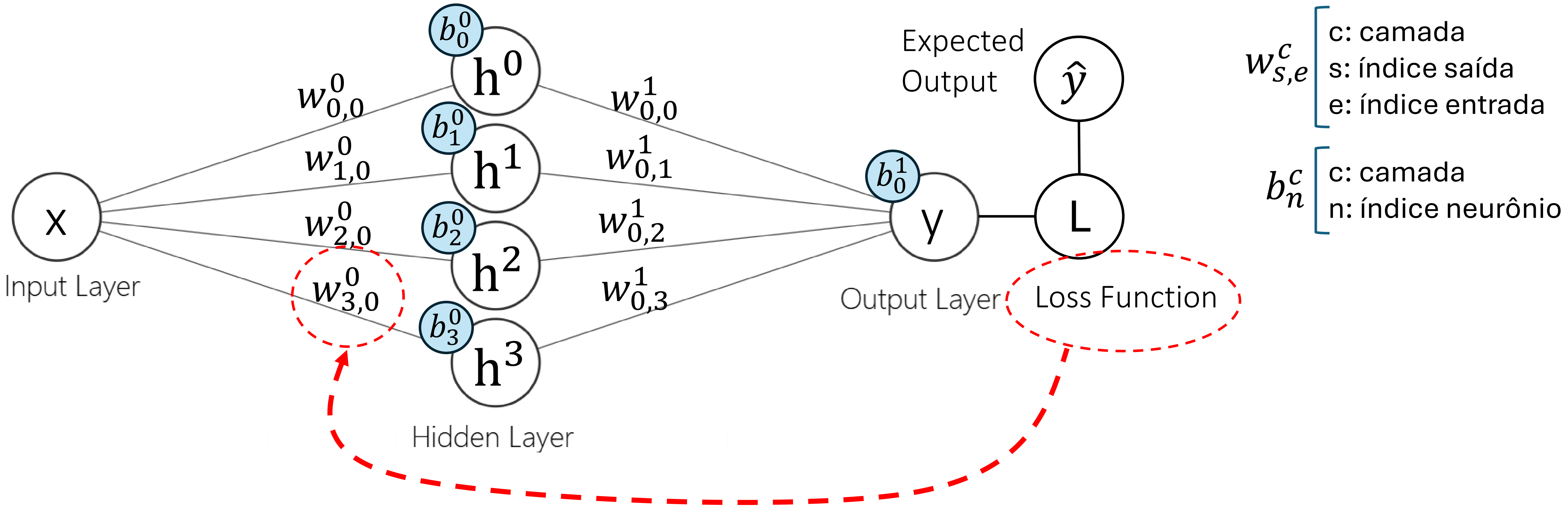


$w^c_{s,e}$
- c: camada
- s: índice saída
- e: índice entrada

b^c_n
- c: camada
- n: índice neurônio

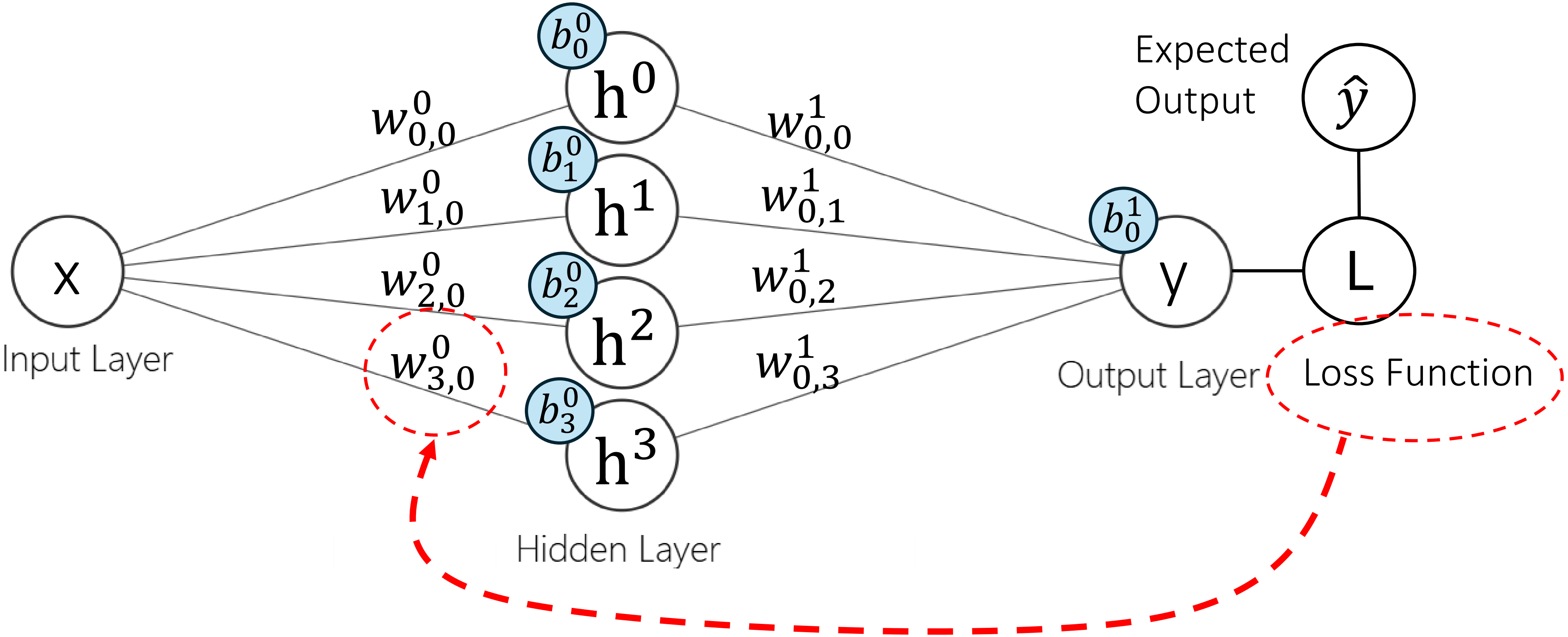
$$\frac{\partial L}{\partial w^0_{3,0}} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y_i})\, w^1_{0,3}\, \sigma'(z^0_3)\, \frac{\partial z^3}{\partial w^0_{3,0}}$$

$$\frac{\partial h^3}{\partial z^3} = \frac{\partial}{\partial z^3}\sigma(z^0_3) = \sigma'(z^0_3)$$



$$\frac{\partial L}{\partial w^0_{3,0}} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \widehat{y_i})\, w^1_{0,3}\ \sigma'(z^0_3)\,\frac{\partial z^3}{\partial w^0_{3,0}}$$

$$\frac{\partial z^3}{\partial w^0_{3,0}} = \frac{\partial}{\partial w^0_{3,0}}\left[x\, w^0_{3,0} + b^0_3\right]$$

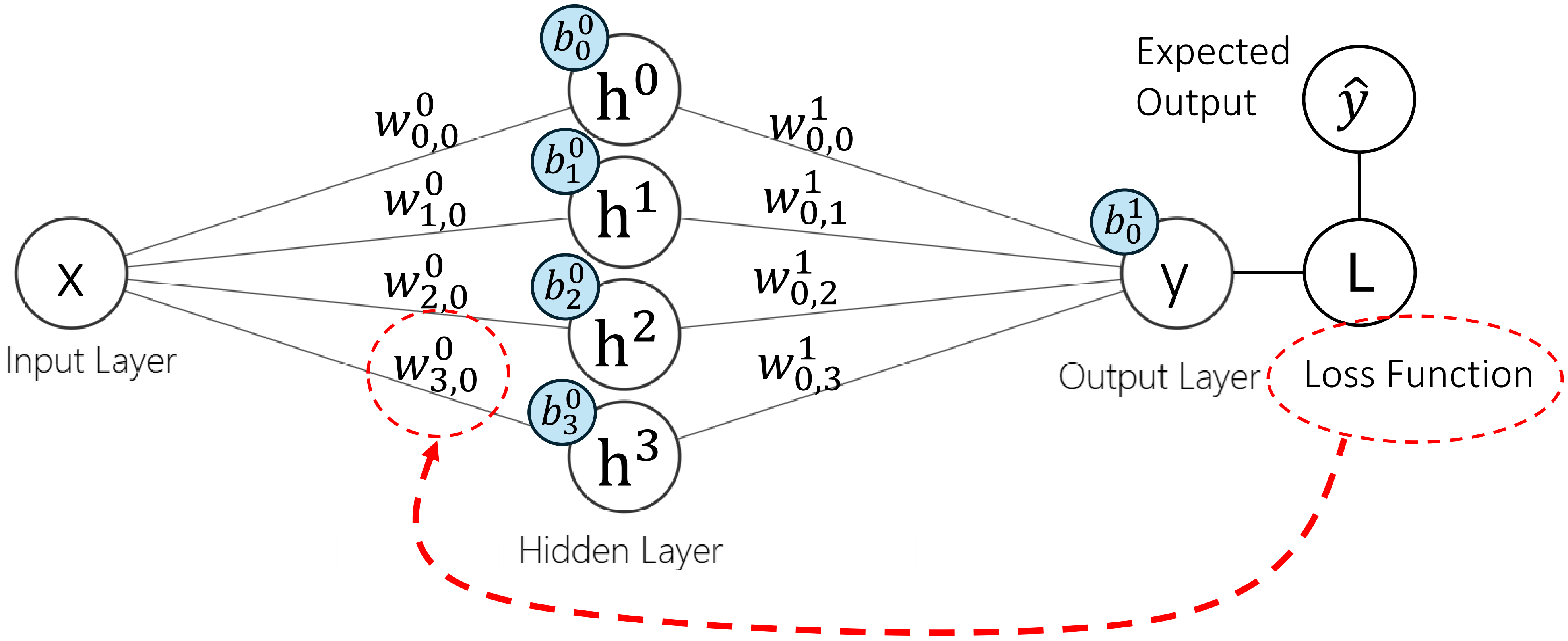


$w_{s,e}^c$
- c: camada
- s: índice saída
- e: índice entrada

b_n^c
- c: camada
- n: índice neurônio

$$\frac{\partial L}{\partial w_{3,0}^0} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y_i})\, w_{0,3}^1\, \sigma'(z_3^0)\, \frac{\partial z^3}{\partial w_{3,0}^0}$$

$$\frac{\partial z^3}{\partial w_{3,0}^0} = \frac{\partial}{\partial w_{3,0}^0}\left[x\, w_{3,0}^0 + b_3^0\right] = x$$

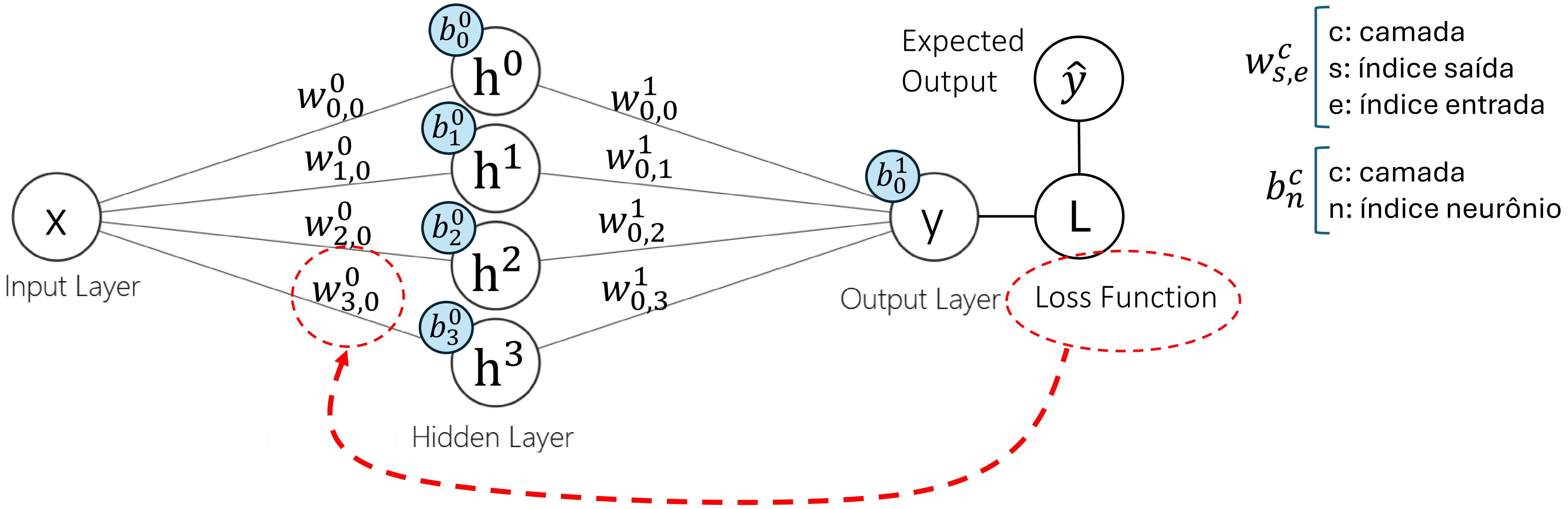


$w_{s,e}^c$ c: camada
s: índice saída
e: índice entrada

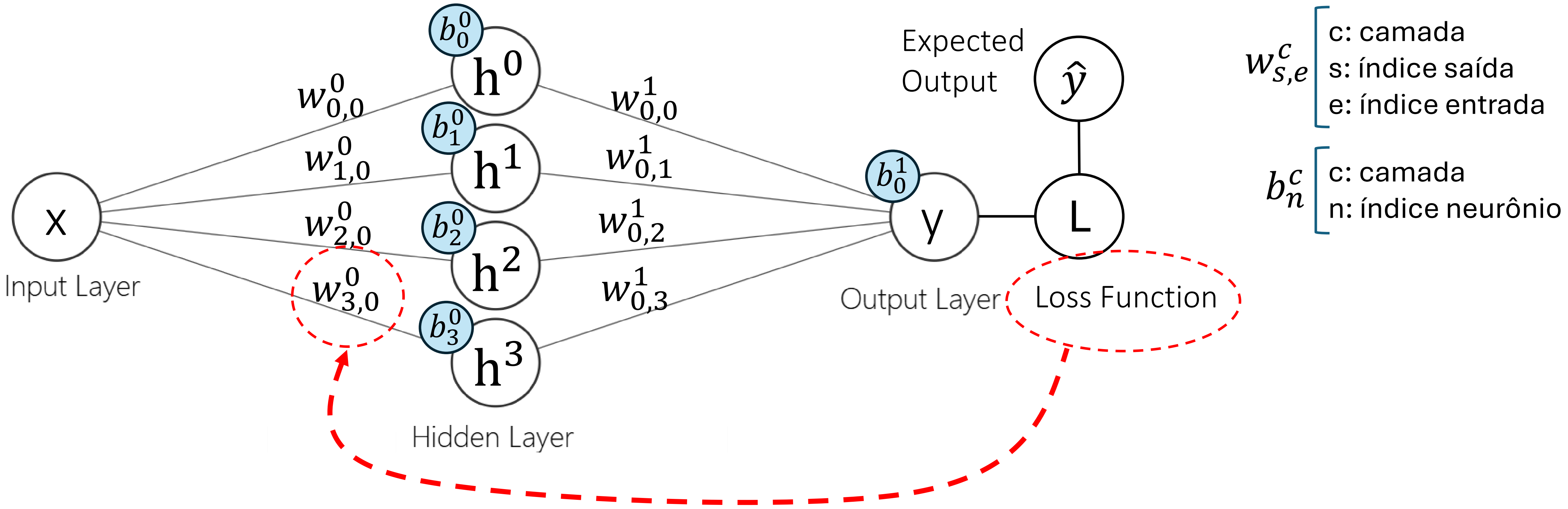
b_n^c c: camada
n: índice neurônio

$$\frac{\partial L}{\partial w_{3,0}^0} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y_i})\, w_{0,3}^1\, \sigma'(z_3^0)x$$

$$\frac{\partial z^3}{\partial w_{3,0}^0} = \frac{\partial}{\partial w_{3,0}^0}\left[x\, w_{3,0}^0 + b_3^0\right] = x$$



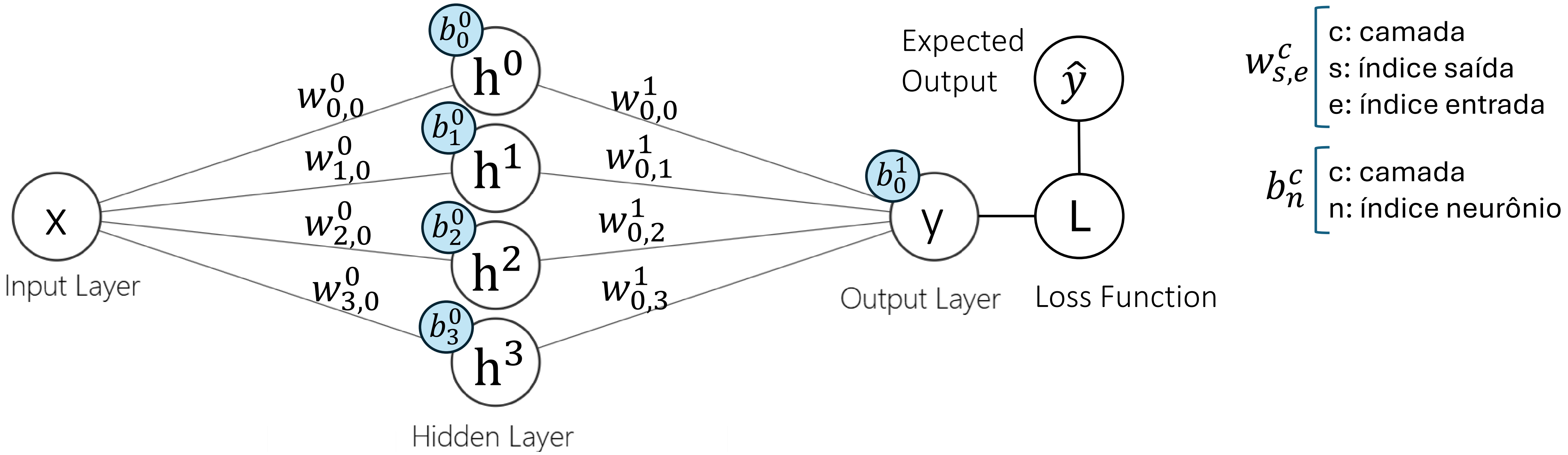
$$\frac{\partial L}{\partial w_{3,0}^0} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y_i})\, w_{0,3}^1\, \sigma'(z_3^0)x$$



Generalizando para o k-ésimo peso e para o bias da camada oculta:

$$\frac{\partial L}{\partial w^0_{k,0}} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y_i})\, w^1_{0,k}\, \sigma'\left(z^0_k\right)x \qquad \frac{\partial L}{\partial b^0_k} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y_i})\, w^1_{0,k}\, \sigma'\left(z^0_k\right)$$

Resumo do Resultado Final



Última camada:

$$\frac{\partial L}{\partial w_{0,k}^1} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y}_i)\, h^k$$

$$\frac{\partial L}{\partial b_0^1} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y}_i)$$

Camada oculta:

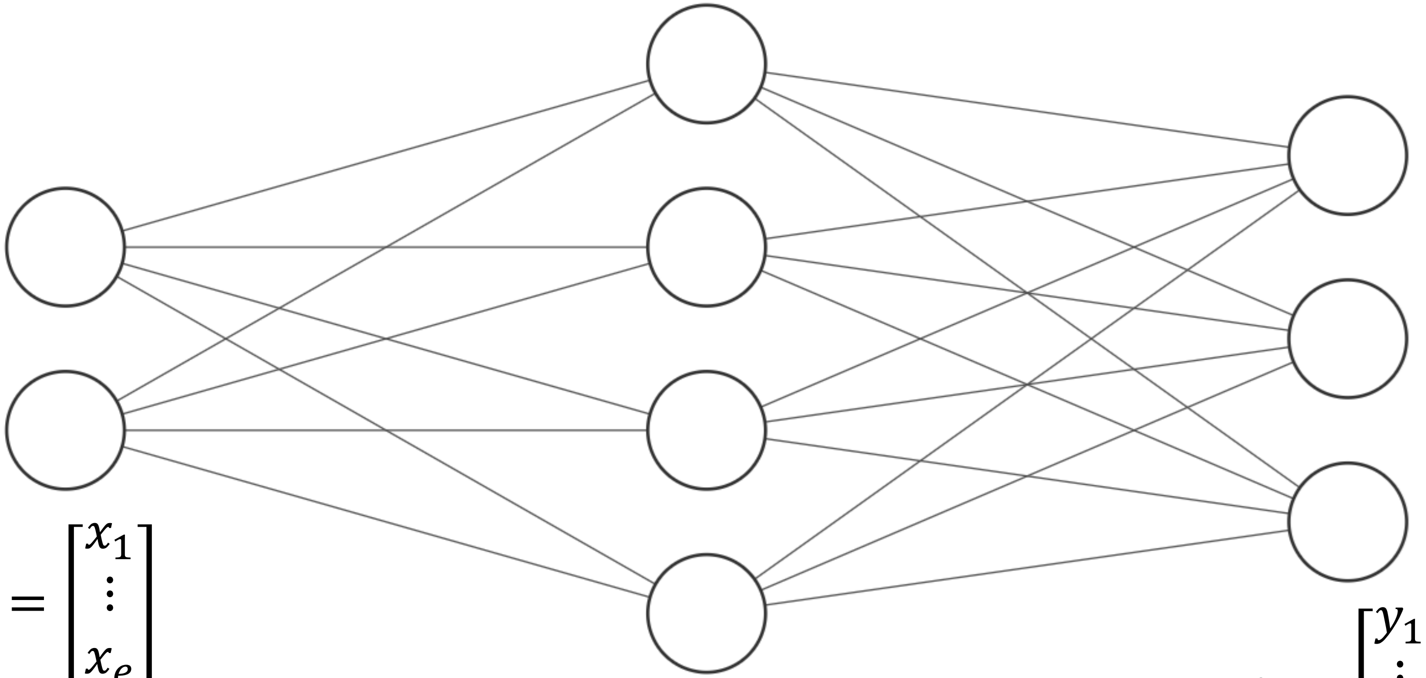
$$\frac{\partial L}{\partial w_{k,0}^0} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y}_i)\, w_{0,k}^1\, \sigma'\left(z_k^0\right) x$$

$$\frac{\partial L}{\partial b_k^0} = \frac{1}{N}\sum_{i=0}^{N} 2(y_i - \hat{y}_i)\, w_{0,k}^1\, \sigma'\left(z_k^0\right)$$

Veja a implementação do método no Jupyter Notebook

Os objetivos são possibilitar uma implementação mais computacionalmente eficiente usando numpy e generalizar a solução para entradas e saídas com quaisquer tamanhos.

Implementando *Backpropagation* Matricialmente

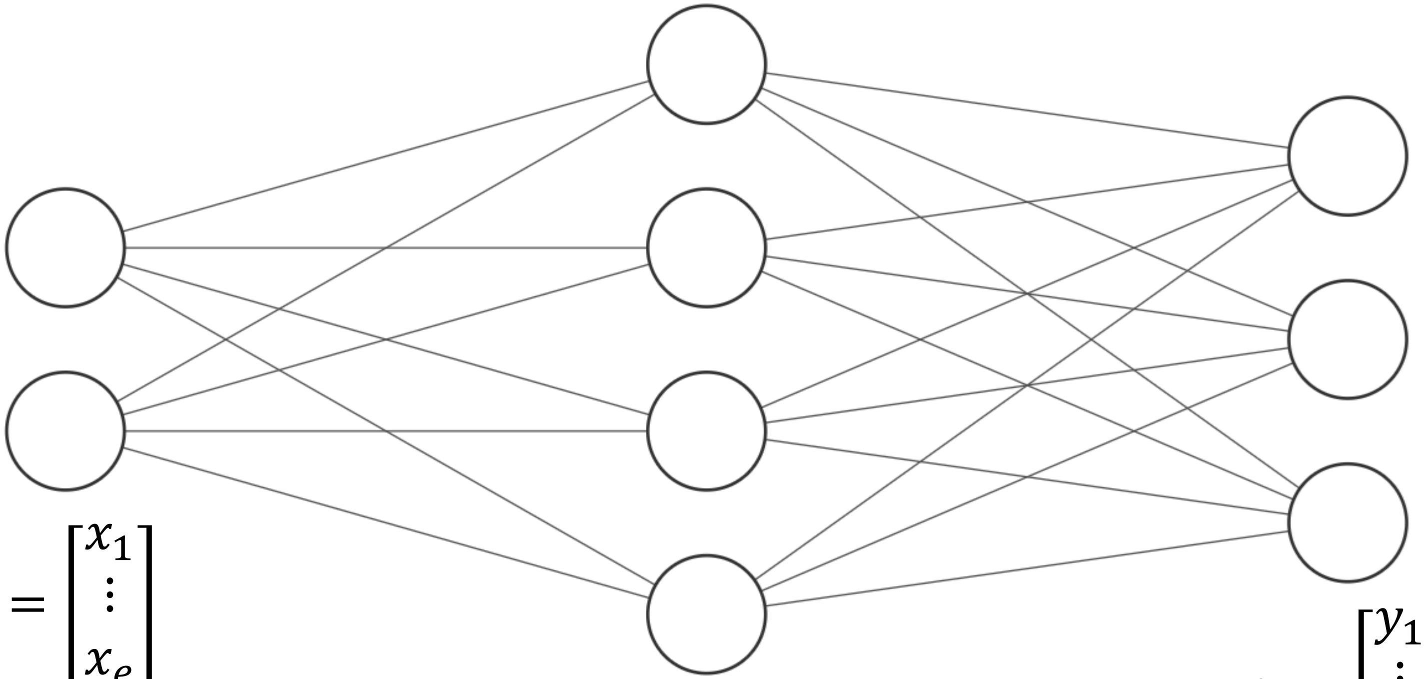


$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_e \end{bmatrix}$$

$$h = \begin{bmatrix} h_1 \\ \vdots \\ h_m \end{bmatrix}$$

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_s \end{bmatrix}$$

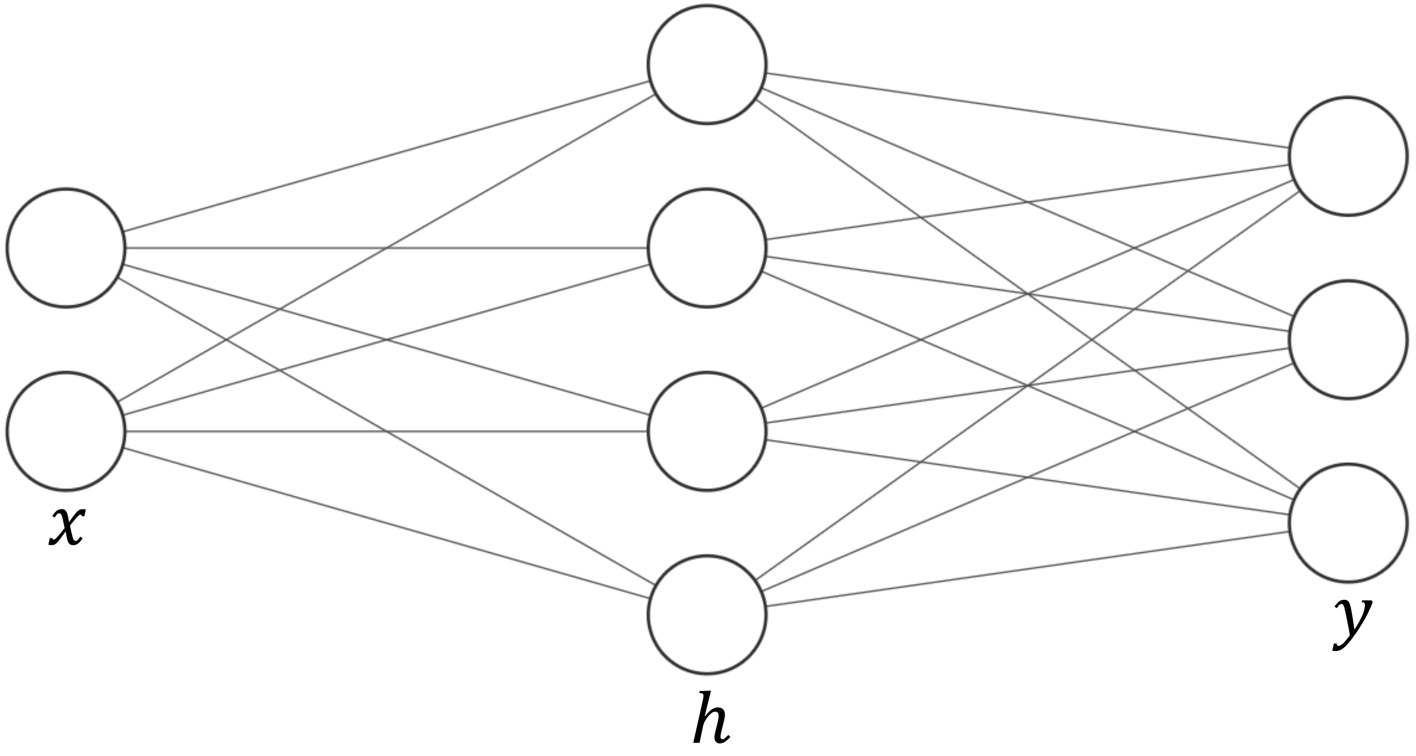
Forward Pass



$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_e \end{bmatrix} \qquad h = \begin{bmatrix} h_1 \\ \vdots \\ h_m \end{bmatrix} \qquad y = \begin{bmatrix} y_1 \\ \vdots \\ y_s \end{bmatrix}$$

$$\underbrace{h}_{m \times 1} = \sigma\Big(\underbrace{W^h}_{m \times e} \underbrace{x}_{\mathrm{e} \times 1} + \underbrace{b^h}_{m \times 1}\Big)$$

$$\underbrace{y}_{s \times 1} = \underbrace{W^y}_{s \times m} \underbrace{h}_{m \times 1} + \underbrace{b^y}_{s \times 1}$$



$$\underset{m\times 1}{h} = \sigma\left(\underset{m\times e}{W^h}\,\underset{\mathrm{e}\times 1}{x} + \underset{m\times 1}{b^h}\right)$$

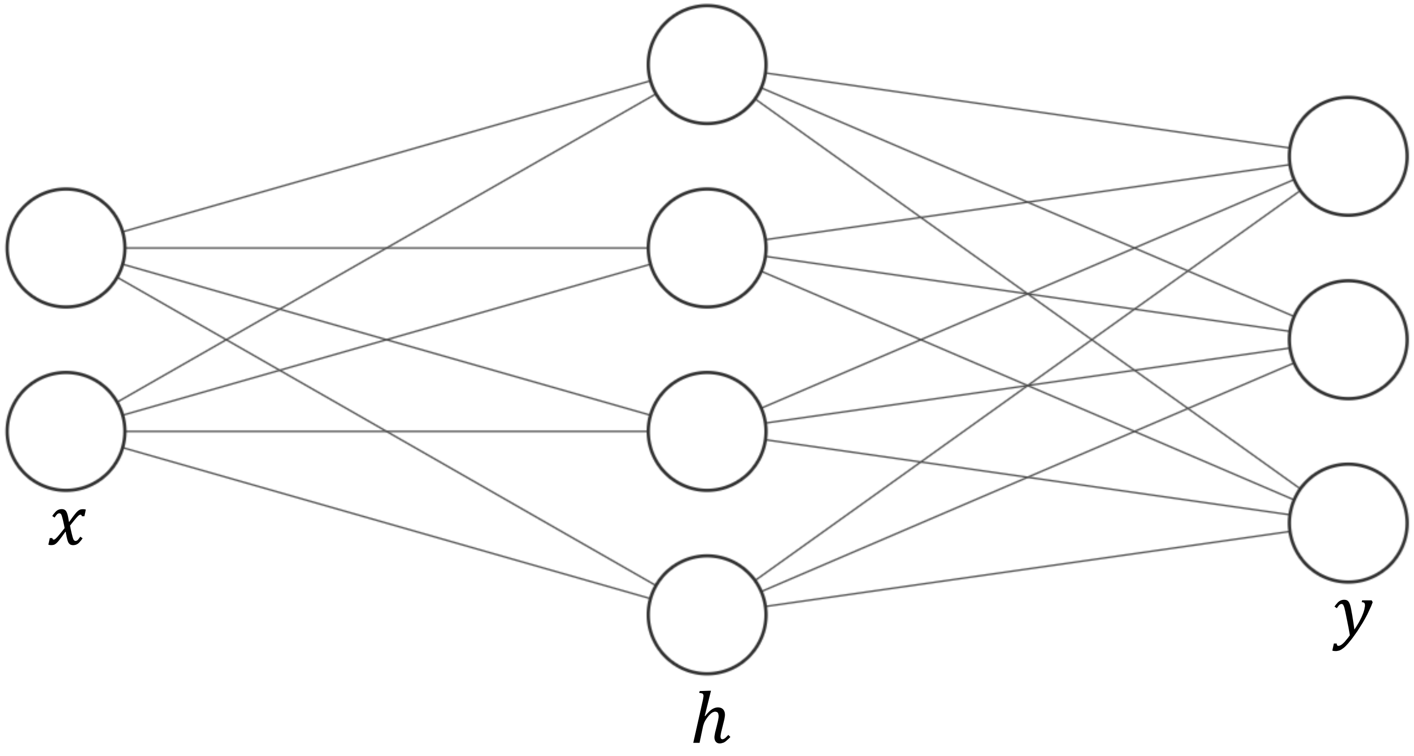
$$\underset{s\times 1}{y} = \underset{s\times m}{W^y}\,\underset{m\times 1}{h} + \underset{s\times 1}{b^y}$$

$$\frac{\partial L}{\partial W^y} = \sum_{i=0}^{N}\left[\frac{2}{N}\overset{s\times 1}{(y_i - \hat{y_i})}\,\overset{1\times m}{h^T}\right]$$

$$\frac{\partial L}{\partial b^y} = \sum_{i=0}^{N}\left[\frac{2}{N}\overset{s\times 1}{(y_i - \hat{y_i})}\right]$$

$$\frac{\partial L}{\partial y}\quad\frac{\partial y}{\partial W^y}$$

Gradiente do erro em relação aos parâmetros da última camada

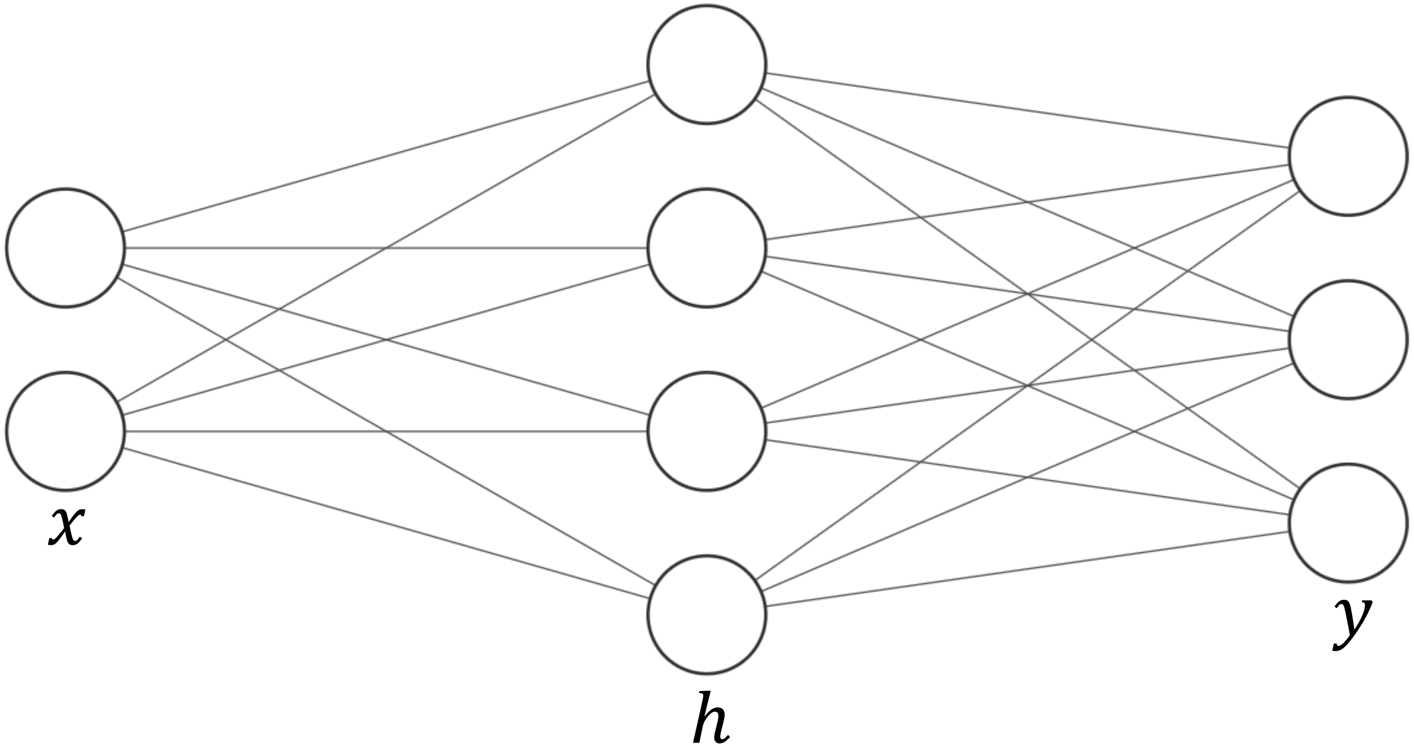


$$\underbrace{h}_{m \times 1} = \sigma\left(\underbrace{W^h}_{m \times e} \underbrace{x}_{e \times 1} + \underbrace{b^h}_{m \times 1}\right)$$

$$\underbrace{y}_{s \times 1} = \underbrace{W^y}_{s \times m} \underbrace{h}_{m \times 1} + \underbrace{b^y}_{s \times 1}$$

$$\frac{\partial L}{\partial W^h} = \sum_{i=0}^{N} \left[\sigma'(z^h) \odot \left(W^{y^T} \frac{2}{N} (y_i - \hat{y}_i)\right)\right] x^T$$

Gradiente do erro em relação aos parâmetros da camada oculta

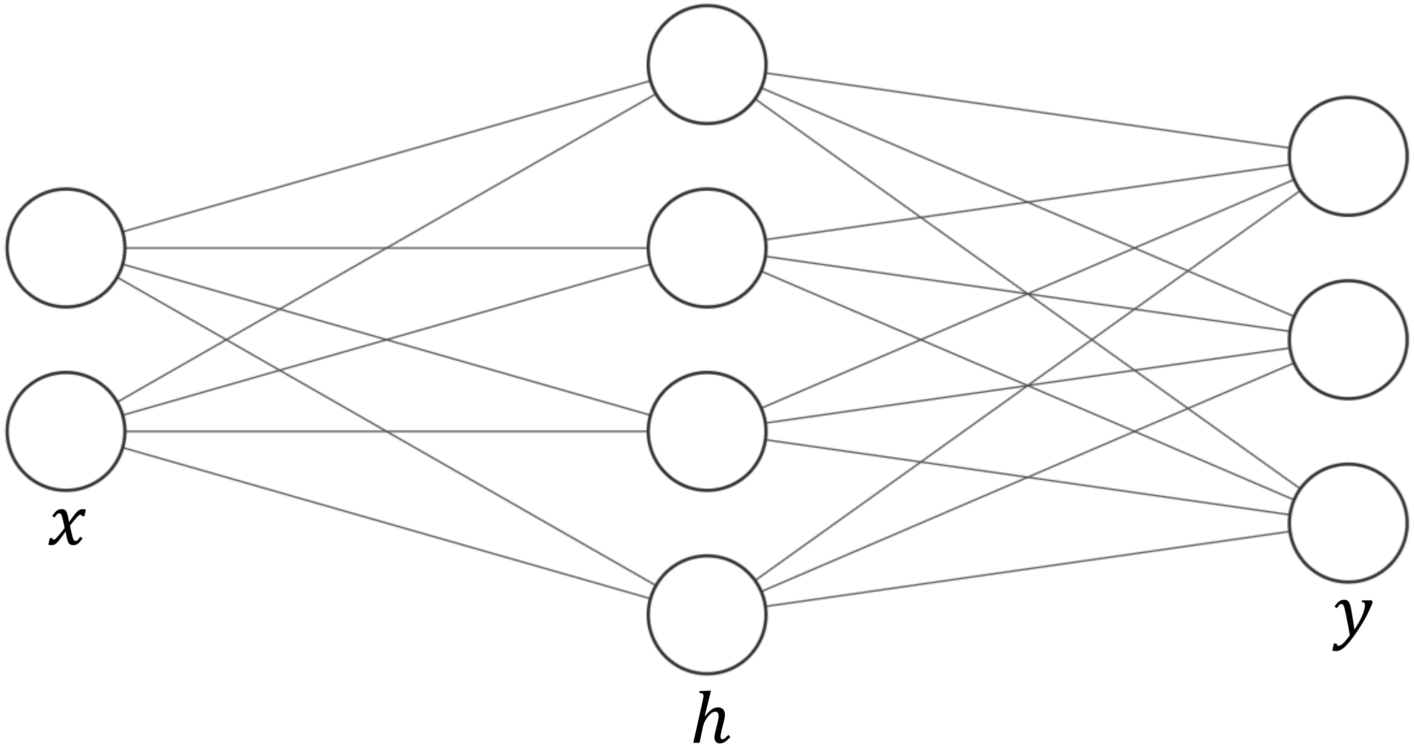


$$\underset{m \times 1}{h} = \sigma\left(\underset{m \times e}{W^h}\,\underset{\mathrm{e} \times 1}{x} + \underset{m \times 1}{b^h}\right)$$

$$\underset{s \times 1}{y} = \underset{s \times m}{W^y}\,\underset{m \times 1}{h} + \underset{s \times 1}{b^y}$$

$$\frac{\partial L}{\partial W^h} = \sum_{i=0}^{N} \left[\sigma'(z^h) \odot \underbrace{\left(\overset{m \times s}{W^{y^T}} \frac{2}{N} \overset{s \times 1}{(y_i - \hat{y}_i)} \right)}_{\underset{m \times 1}{\frac{\partial L}{\partial h}}} \right] x^T$$

Gradiente do erro em relação aos parâmetros da camada oculta

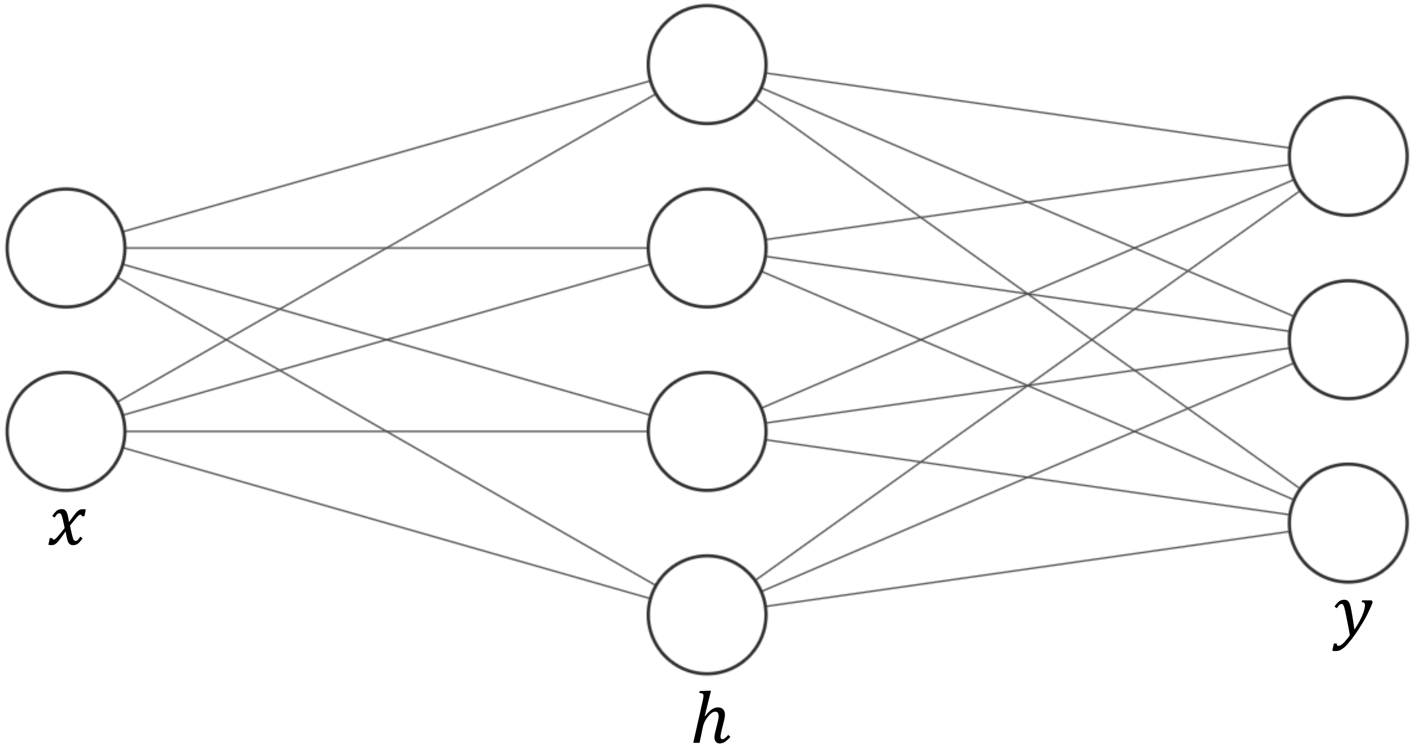


$$\underset{m\times 1}{h} = \sigma\Big(\underset{m\times e}{W^h}\,\underset{e\times 1}{x} + \underset{m\times 1}{b^h}\Big)$$

$$\underset{s\times 1}{y} = \underset{s\times m}{W^y}\,\underset{m\times 1}{h} + \underset{s\times 1}{b^y}$$

$$\frac{\partial L}{\partial W^h} = \sum_{i=0}^{N} \underbrace{\left[\sigma'(z^h) \overset{\text{produto ponto a ponto}}{\odot} \left(\overset{m\times s}{W^{y^T}}\frac{2}{N}\overset{s\times 1}{(y_i - \widehat{y}_i)}\right)\right]}_{\underset{m\times 1}{\frac{\partial L}{\partial z^h}}} x^T$$

Gradiente do erro em relação aos parâmetros da camada oculta



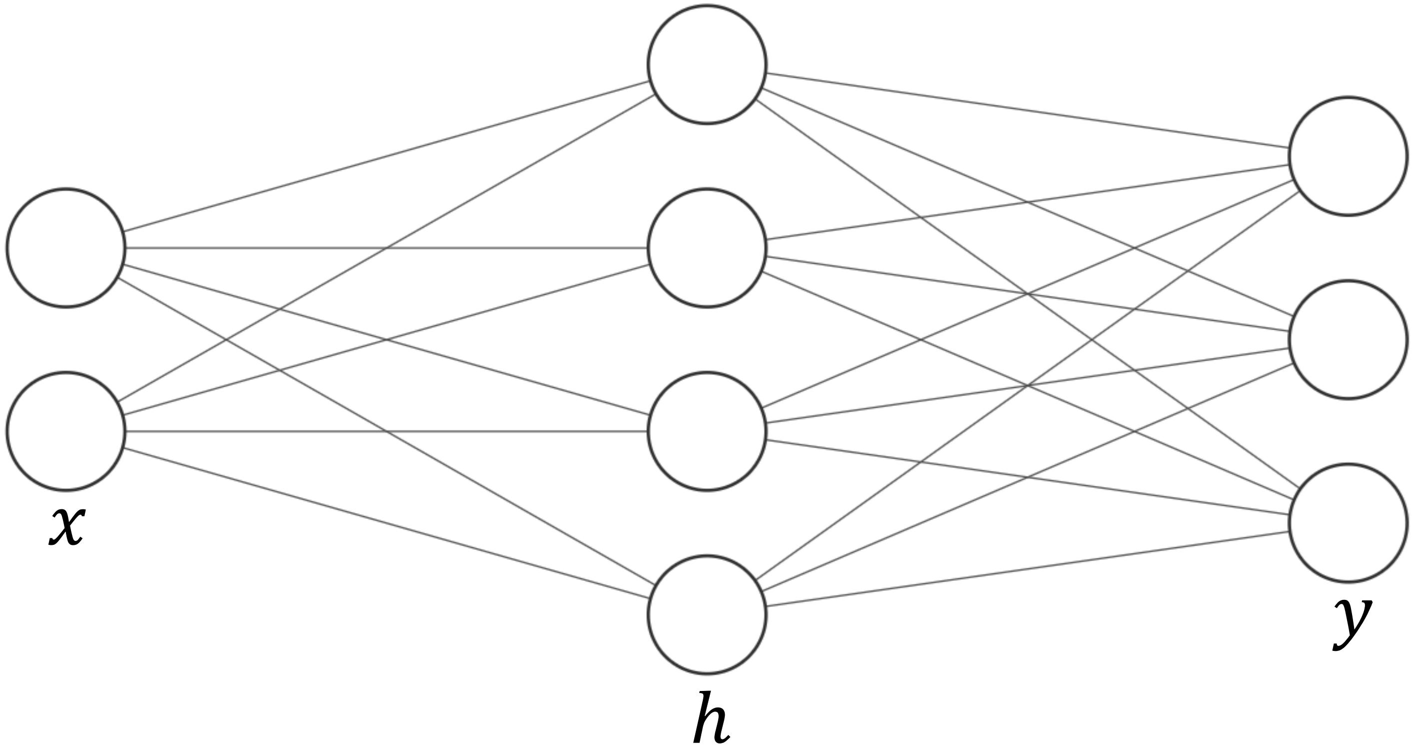
$$\underbrace{h}_{m \times 1} = \sigma\big(\underbrace{W^h}_{m \times e} \underbrace{x}_{e \times 1} + \underbrace{b^h}_{m \times 1}\big)$$

$$\underbrace{y}_{s \times 1} = \underbrace{W^y}_{s \times m} \underbrace{h}_{m \times 1} + \underbrace{b^y}_{s \times 1}$$

produto ponto a ponto

$$\underbrace{\frac{\partial L}{\partial W^h}}_{m \times e} = \sum_{i=0}^{N} \underbrace{\left[\sigma'(z^h) \odot \left(\overbrace{W^{y^T}}^{m \times s} \frac{2}{N} \overbrace{(y_i - \widehat{y}_i)}^{s \times 1}\right)\right]}_{\underbrace{\frac{\partial L}{\partial z^h}}_{m \times 1}} \overbrace{x^T}^{1 \times e}$$

Gradiente do erro em relação aos parâmetros da camada oculta



$$\underset{m\times 1}{h} = \sigma\big(\underset{m\times e}{W^h}\,\underset{e\times 1}{x} + \underset{m\times 1}{b^h}\big)$$

$$\underset{s\times 1}{y} = \underset{s\times m}{W^y}\,\underset{m\times 1}{h} + \underset{s\times 1}{b^y}$$

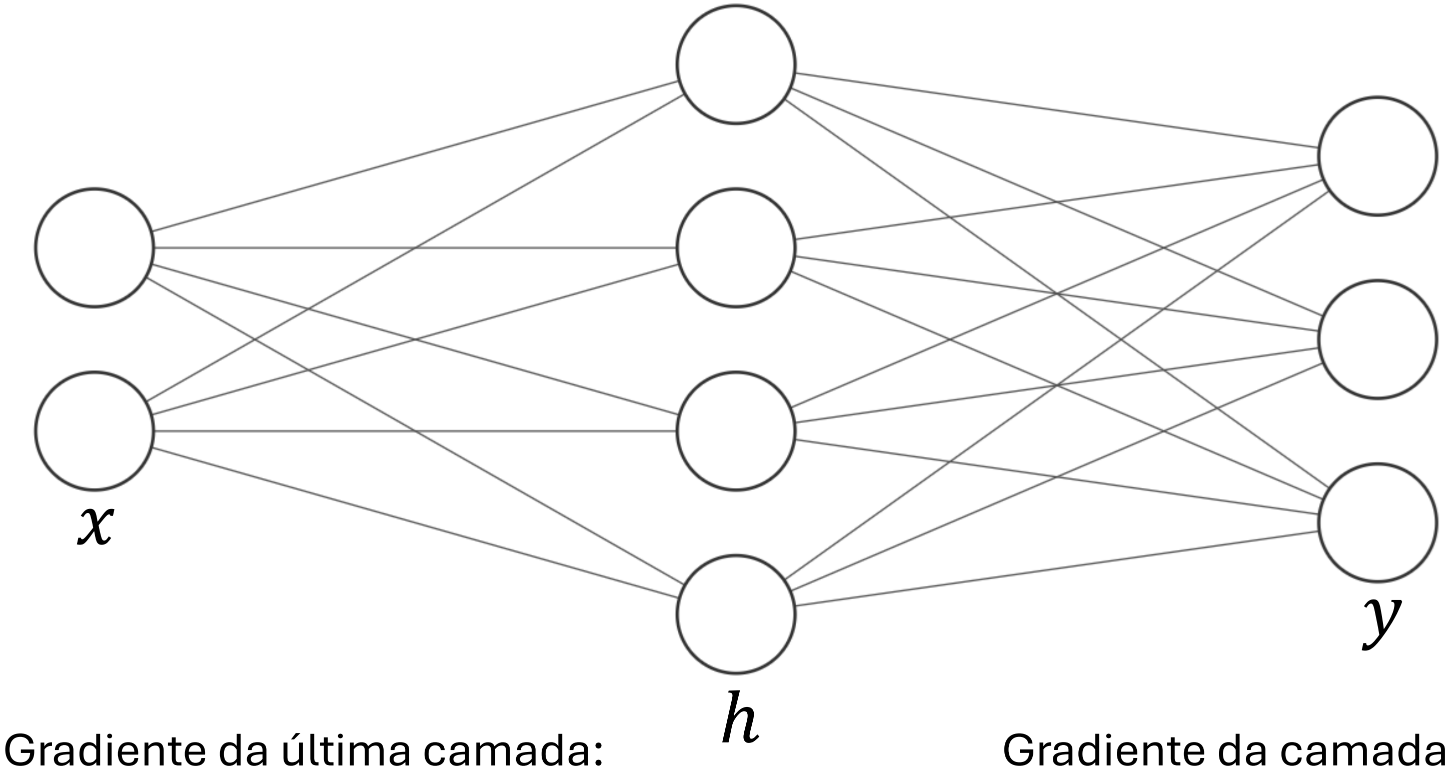
analogamente...

produto ponto a ponto

$$\underset{m\times 1}{\frac{\partial L}{\partial b^h}} = \sum_{i=0}^{N} \underbrace{\left[\sigma'(z^h)\odot\left(\overset{m\times s}{W^{y^T}}\,\frac{2}{N}\,\overset{s\times 1}{(y_i - \hat{y}_i)}\right)\right]}_{m\times 1}$$

Gradiente do erro em relação aos parâmetros da camada oculta

Resumo do Resultado Final



$$h = \sigma(W^h x + b^h)$$

h: $m \times 1$, W^h: $m \times e$, x: $e \times 1$, b^h: $m \times 1$

$$y = W^y h + b^y$$

y: $s \times 1$, W^y: $s \times m$, h: $m \times 1$, b^y: $s \times 1$

Gradiente da última camada:

$$\frac{\partial L}{\partial W^y} = \sum_{i=0}^{N} \left[\frac{2}{N}(y_i - \hat{y}_i)\, h^T\right]$$

$$\frac{\partial L}{\partial b^y} = \sum_{i=0}^{N} \left[\frac{2}{N}(y_i - \hat{y}_i)\right]$$

Gradiente da camada oculta:

$$\frac{\partial L}{\partial W^h} = \sum_{i=0}^{N} \left[\sigma'(z^h) \odot \left(W^{y^T} \frac{2}{N}(y_i - \hat{y}_i)\right)\right] x^T$$

$$\frac{\partial L}{\partial b^h} = \sum_{i=0}^{N} \left[\sigma'(z^h) \odot \left(W^{y^T} \frac{2}{N}(y_i - \hat{y}_i)\right)\right]$$

Veja a implementação do método no Jupyter Notebook