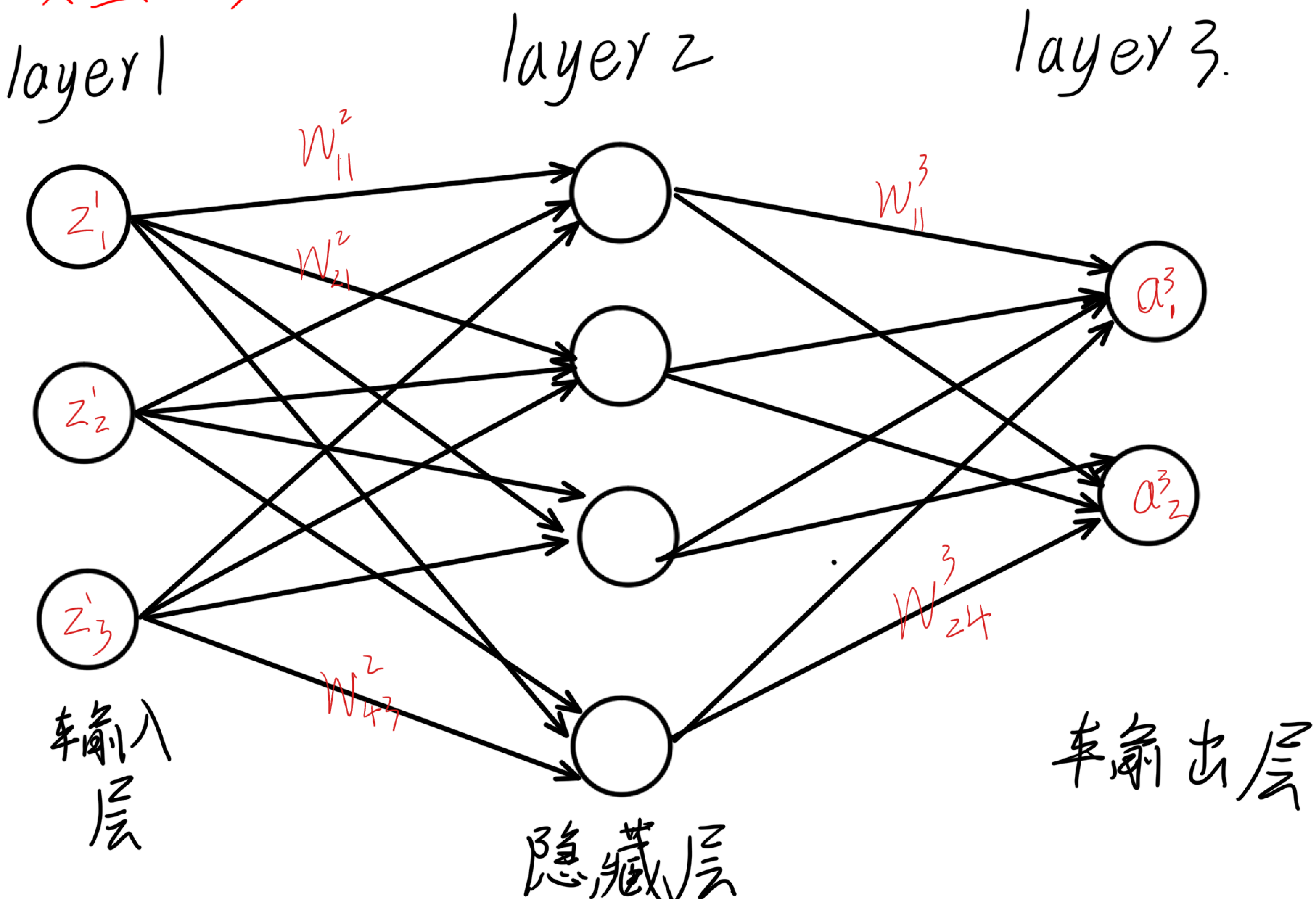
反向传播推导

1. 变量定义



上图是一个三层神经网络，对变量进行定义：

w_{jk}^{l} 表示第 $(l-1)$ 层的第 k 个神经元连接到第 l 层第 j 个神经元的权重。

b_j^l 表示第 l 层第 j 个神经元的偏置。

z_j^l 表示第 l 层第 j 个神经元的输入，即：$z_j^l = \sum_k w_{jk}^l a_k^{l-1} + b_j^l$

a_j^l 表示第 l 层第 j 个神经元的输入，即：$a_j^l = \sigma(\sum_k w_{jk}^l a_k^{l-1} + b_j^l)$

σ 表示激活函数，常见的例如 sigmoid 函数、Relu 函数等。 $= \sigma(z_j^l)$

2. 公式推导

首先，将第 l 层第 j 个神经元中产生的错误（即实际值与预测值之间的误差）定义为：

$$\delta_j^l = \frac{\partial C}{\partial z_j^l}$$

以一个输入样本为例进行说明，此时损失函数与代价函数的均方误差表示形式均为：

$$C=\frac{1}{2}\sum_j(y_j-a_j^L)^2$$ y_j表示真实值，a_j^L表示预测的输出值

公式1（计算最后一层神经网络产生的误差）

$$\delta^L=\nabla_a C\odot\sigma'(z^L)=\frac{\partial L}{\partial z^L}\odot\frac{\partial a^L}{\partial z^L}$$

其中，$\odot$表示Hadamard乘积，用于矩阵或向量之间点对点的乘法用算，∇表示梯度运算符。⇒ 例 $f=[f_1,f_2,\cdots,f_m]$ $x=[x_1,x_2,\cdots x_n]$

$$\nabla_x f=\begin{bmatrix}\frac{\partial f_1}{\partial x_1}, & \frac{\partial f_1}{\partial x_2}, & \cdots & \frac{\partial f_1}{\partial x_n}\\ \vdots & \vdots & \ddots & \vdots\\ \frac{\partial f_m}{\partial x_1}, & \frac{\partial f_m}{\partial x_2}, & \cdots & \frac{\partial f_m}{\partial x_n}\end{bmatrix}$$

公式一推导过程：

$$\because\ \delta_j^L=\frac{\partial C}{\partial z_j^L}=\frac{\partial C}{\partial a_j^L}\cdot\frac{\partial a_j^L}{\partial z_j^L}$$

$$\therefore \delta^L = \frac{\partial L}{\partial z^L} \odot \frac{\partial a^L}{\partial z^L} = \nabla_a C \odot \sigma'(z^L)$$

公式2(由后向前，计算每一层神经网络产生的误差)：

☆ $\delta^l = ((w^{l+1})^T \delta^{l+1}) \odot \sigma'(z^l)$ (T表示矩阵转置运算)

推导过程：

$$\because \delta_j^l = \frac{\partial C}{\partial z_j^l} = \sum_k \frac{\partial C}{\partial z_k^{l+1}} \cdot \frac{\partial z_k^{l+1}}{\partial a_j^l} \cdot \frac{\partial a_j^l}{\partial z_j^l}$$

$$= \sum_k \delta_k^{l+1} \cdot \frac{\partial (w_{kj}^{l+1} a_j^l + b_k^{l+1})}{\partial a_j^l} \cdot \sigma'(z_j^l)$$

$$= \sum_k \delta_k^{l+1} \cdot w_{kj}^{l+1} \cdot \sigma'(z_j^l)$$

$$\therefore \delta^l = ((w^{l+1})^T \delta^{l+1}) \odot \sigma'(z^l)$$

公式3（计算权重梯度）：

$$\frac{\partial C}{\partial w_{jk}^{l}} = a_k^{l-1}\delta_j^{l}$$

推导过程：

$$\frac{\partial C}{\partial w_{jk}^{l}} = \frac{\partial C}{\partial z_j^{l}} \cdot \frac{\partial z_j^{l}}{\partial w_{jk}^{l}} = \delta_j^{l} \cdot \frac{\partial (w_{jk}^{l} a_k^{l-1} + b_j^{l})}{\partial w_{jk}^{l}} = a_k^{l-1}\delta_j^{l}$$

公式4（计算偏置的梯度）：

$$\frac{\partial C}{\partial b_j^{l}} = \delta_j^{l}$$

推导过程：

$$\frac{\partial C}{\partial b_j^{l}} = \frac{\partial C}{\partial z_j^{l}} \cdot \frac{\partial z_j^{l}}{\partial b_j^{l}} = \delta_j^{l} \cdot \frac{\partial (w_{jk}^{l} a_k^{l-1} + b_j^{l})}{\partial b_j^{l}} = \delta_j^{l}$$

更新权重和偏置

设学习率为 η

$$(w_{jk}^{l})_{新} = w_{j}^{l} - \eta \cdot \frac{\partial C}{\partial w_{jk}^{l}} = w_{jk}^{l} - \eta a_{k}^{l-1} \delta_{j}^{l}$$

$$(b_{j}^{l})_{新} = b_{j}^{l} - \eta \cdot \frac{\partial C}{\partial b_{j}^{l}} = b_{j}^{l} - \eta \delta_{j}^{l}$$

总结：四个公式：

① $\delta^{L} = \nabla_{a} C \odot \sigma'(z^{L}) = \frac{\partial L}{\partial z^{l}} \odot \frac{\partial a^{l}}{\partial z^{l}}$.

② $\delta^{l} = ((w^{l+1})^{T} \delta^{l+1}) \odot \sigma'(z^{l})$

③ $\frac{\partial C}{\partial w_{jk}^{l}} = a_{k}^{l-1} \delta_{j}^{l}$

4 $\dfrac{\partial C}{\partial b_j^l} = \delta_j^l.$