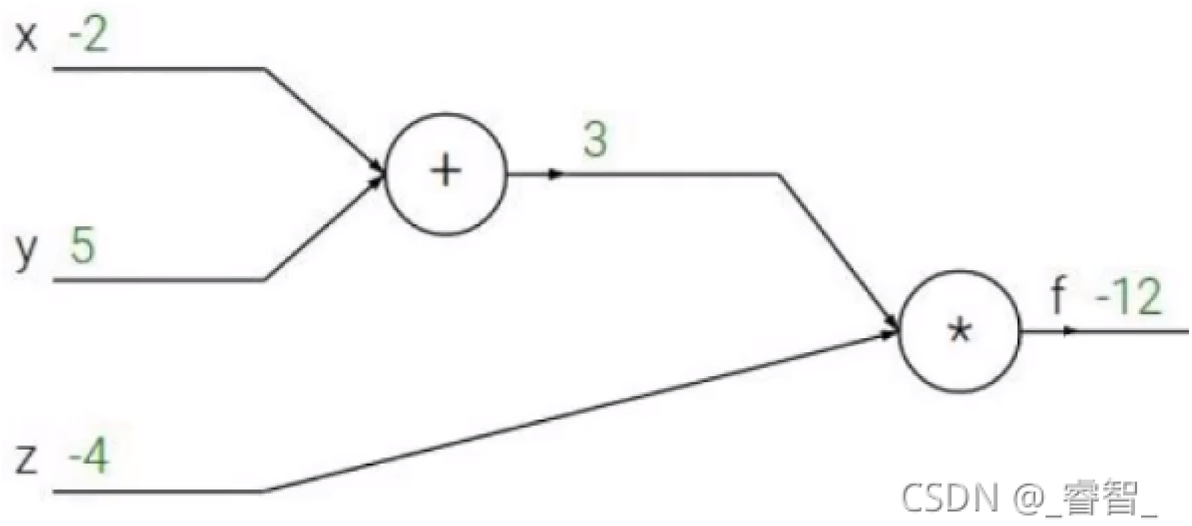
二、反向传播（梯度下降）

1. 意义

从后往前，逐层求导，可以得到前一项的权重

作用：更新模型，达到学习的效果

从后往前传递梯度，得到新计算出来的权重



$$\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$$

三个求偏导意义分别是：x、y、z对 f(结果)的影响

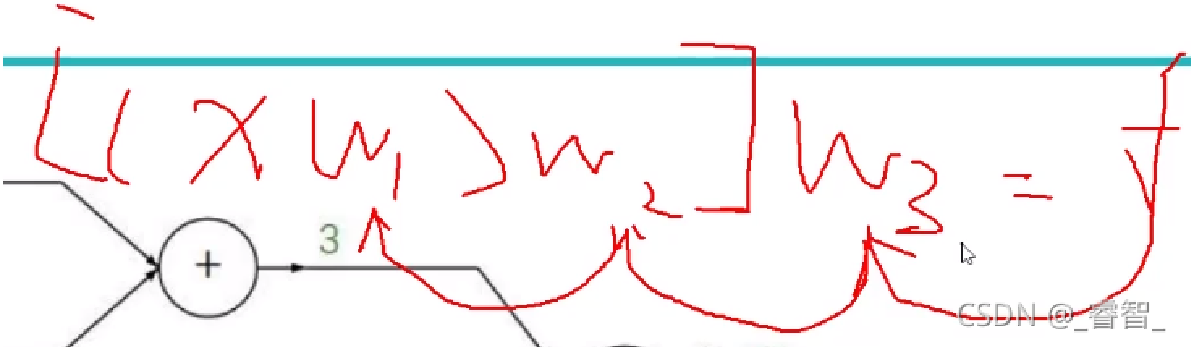
2. 目的

反向传播目的：优化权重参数（根据前向传播得到的损失值）

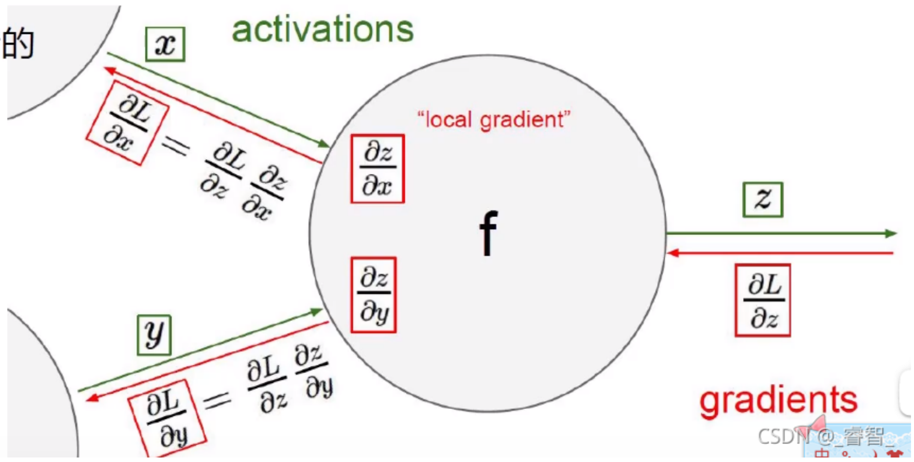
3.过程

反向传播：从后往前，逐层计算。

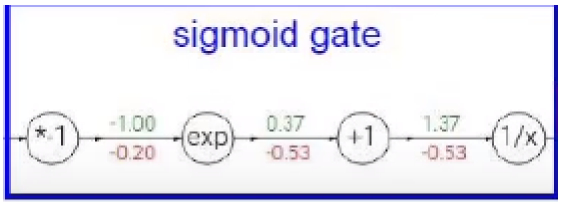
目的：为了让损失函数的值最小。



梯度是一步一步传的

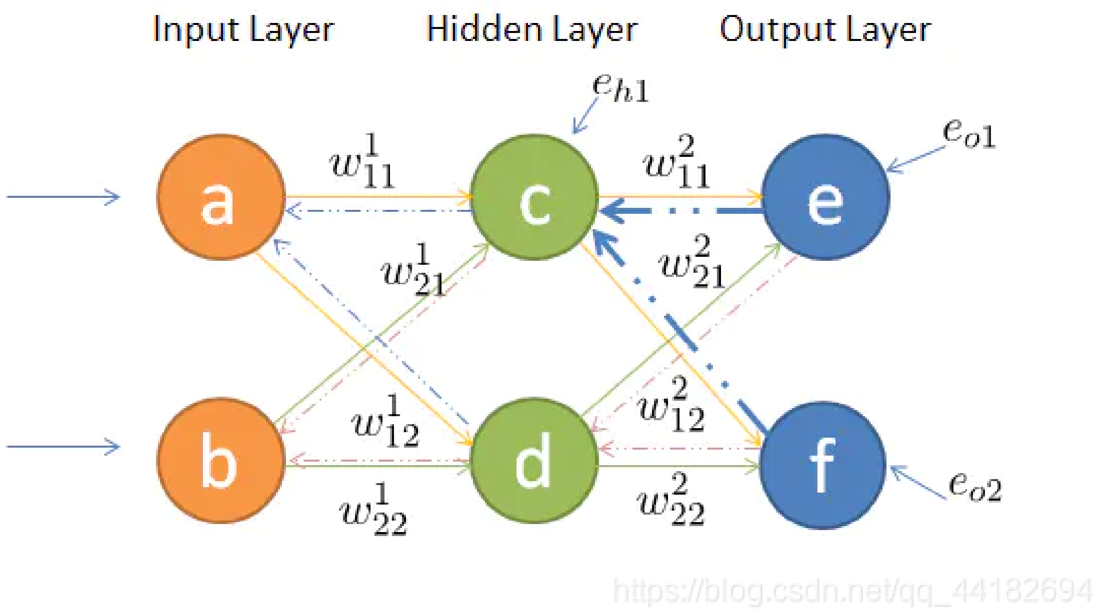


反向偏导也可以一大块一大块地求。



卷积神经网络反向传播

卷积神经网络反向传播无非就是误差溯源和调整超参数的过程。神经网络在训练过程中，通过前向传播得到的输出结果有时候与真实值不同，因为目前没有哪个神经网络结构能够做到100%的准确率。既然存在误差，那就要将误差反向传播并逐层返回，通过链式法则调整训练参数以减小误差。这就涉及到卷积神经网络的反向传播过程。前向传递输入数据到输出产生误差，反向传播则误差信息更新权重参数矩阵。



由上图可知，e01为e节点的误差值，e02为f节点的误差值，要向隐藏层节点传递误差值。以c节点为例，从图中看出，e,f节点都指向了c节点，每个箭头上都有相应的权值，因此c节点的输入误差由e,f两项构成，相应比例应该用权值比表示，用数学公式可以表示为：

$$e_{h1} = \frac{w_{11}^2}{w_{11}^2 + w_{21}^2} \cdot e_{o1} + \frac{w_{12}^2}{w_{12}^2 + w_{22}^2} \cdot e_{o2}$$

同理，d节点的输入可以表示为：

$$e_{h2} = \frac{w_{21}^2}{w_{11}^2 + w_{21}^2} \cdot e_{o1} + \frac{w_{22}^2}{w_{12}^2 + w_{22}^2} \cdot e_{o2}$$

如果以矩阵的形式表达，可以为：

$$\begin{pmatrix} e_{h1} \\ e_{h2} \end{pmatrix} = \begin{pmatrix} \frac{w_{11}^2}{w_{11}^2+w_{21}^2} & \frac{w_{12}^2}{w_{12}^2+w_{22}^2} \\ \frac{w_{21}^2}{w_{11}^2+w_{21}^2} & \frac{w_{22}^2}{w_{12}^2+w_{22}^2} \end{pmatrix} \bullet \begin{pmatrix} e_{o1} \\ e_{o2} \end{pmatrix}$$

但是，我们发现中间矩阵表示太麻烦，因为我们要的是它们之间的比例关系，因此可以把中间矩阵的分母全部去掉：

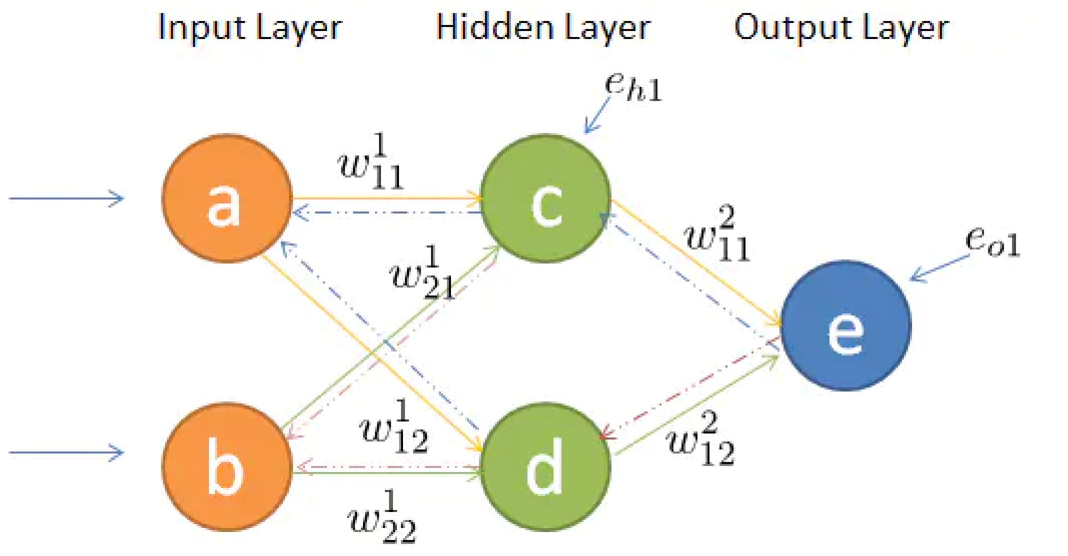
$$\begin{pmatrix} e_{h1} \\ e_{h2} \end{pmatrix} = \begin{pmatrix} w_{11}^2 & w_{12}^2 \\ w_{21}^2 & w_{22}^2 \end{pmatrix} \bullet \begin{pmatrix} e_{o1} \\ e_{o2} \end{pmatrix}$$

还可以表示为：

$$\mathbf{E}_h = \mathbf{W}^T \cdot \mathbf{E}_o$$

我们可以从上式发现，反向传播时的权重矩阵其实就是前向传播时的权重矩阵的转置。至此，我们完成了输出层向隐藏层的传播过程，隐藏层向输入层传播过程同理，这里不再赘述。

链式法则
我们通过神经网络误差反向传播机制，将误差反馈到了各层，那么接下来就是利用误差值去更新训练参数，这就用到了链式法则。
同样以上文的神经网络模型为例：



我们首先更新隐藏层的权重参数w11，从后往前推导，首先我们可得：

$$e_{o1} = \frac{1}{2}(a_1^3 - y_1)^2$$

$$a_1^3 = sigmoid(z_1^3)$$

$$z_1^3 = (w_{11}^2 \cdot a_1^2 + w_{12}^2 \cdot a_2^2 + b_1^3)$$

神经网络参数更新是利用梯度下降法进行更新，以梯度的反方向前进，梯度下降一般公式：

$$w \leftarrow w + \Delta w$$

$$\Delta w = -\alpha \frac{\partial Loss}{\partial w}$$

由此公式可知，我们要求出误差对参数的偏导数。结合本文，误差对w11的偏导为：

$$\frac{\eth e_{o1}}{\eth w_{11}^2} = \frac{\eth e_{o1}}{\eth a_1^3} \cdot \frac{\eth a_1^3}{\eth z_1^3} \cdot \frac{\eth z_1^3}{\eth w_{11}^2}$$

代入后得：

$$\frac{\eth e_{o1}}{\eth w_{11}^2} = (a_1^3 - y_1) \cdot sigmoid(z_1^3) \cdot (1 - sigmoid(z_1^3)) \cdot a_1^2$$

同理，误差对w12的偏导为：

$$\frac{\eth e_{o1}}{\eth w_{12}^2} = \frac{\eth e_{o1}}{\eth a_1^3} \cdot \frac{\eth a_1^3}{\eth z_1^3} \cdot \frac{\eth z_1^3}{\eth w_{12}^2}$$

代入后得：

$$\frac{\eth e_{o1}}{\eth w_{12}^2} = (a_1^3 - y_1) \cdot sigmoid(z_1^3) \cdot (1 - sigmoid(z_1^3)) \cdot a_2^2$$

同理，误差对偏置向量b的求导为：

$$\frac{\eth e_{o1}}{\eth b_1^3} = \frac{\eth e_{o1}}{\eth a_1^3} \cdot \frac{\eth a_1^3}{\eth z_1^3} \cdot \frac{\eth z_1^3}{\eth b_1^3}$$

代入后得：

$$\frac{\eth e_{o1}}{\eth b_1^3} = (a_1^3 - y_1) \cdot sigmoid(z_1^3) \cdot (1 - sigmoid(z_1^3))$$

至此，我们可以利用梯度下降公式更新隐藏层的权重参数，接着我们可以对输入层的权重参数w11进行更新：

$$e_{o1} = \frac{1}{2}(a_1^3 - y_1)^2$$
$$a_1^3 = sigmoid(z_1^3)$$
$$z_1^3 = (w_{11}^2 \cdot a_1^2 + w_{12}^2 \cdot a_2^2 + b_1^3)$$
$$a_1^2 = sigmoid(z_1^2)$$
$$z_1^2 = (w_{11}^1 \cdot a_1^1 + w_{12}^1 \cdot a_2^1 + b_1^2)$$

误差对输入层的w11的偏导为：

$$\frac{\partial e_{o1}}{\partial w_{11}^1} = \frac{\partial e_{o1}}{\partial a_1^3} \cdot \frac{\partial a_1^3}{\partial z_1^3} \cdot \frac{\partial z_1^3}{\partial a_1^2} \cdot \frac{\partial a_1^2}{\partial z_1^2} \cdot \frac{\partial z_1^2}{\partial w_{11}^1}$$

代入后，具体可得：

$$\frac{\partial e_{o1}}{\partial w_{11}^1} = (a_1^3 - y_1) \cdot sigmoid(z_1^3) \cdot (1 - sigmoid(z_1^3)) \cdot w_{11}^2 \cdot sigmoid(z_1^2) \cdot (1 - sigmoid(z_1^2)) \cdot a_1^1$$

同理，输入层的其余参数也是用同方法求得，这里不再赘述。最后我们可以对误差对w11的偏导代入梯度下降公式：

$$w_{11} = w_{11} - \eta \frac{\partial e_{o1}}{\partial w_{11}}$$
$$b_1 = b_1 - \eta \frac{\partial e_{o1}}{\partial b_1}$$

这样我们便完成了参数更新的全过程。