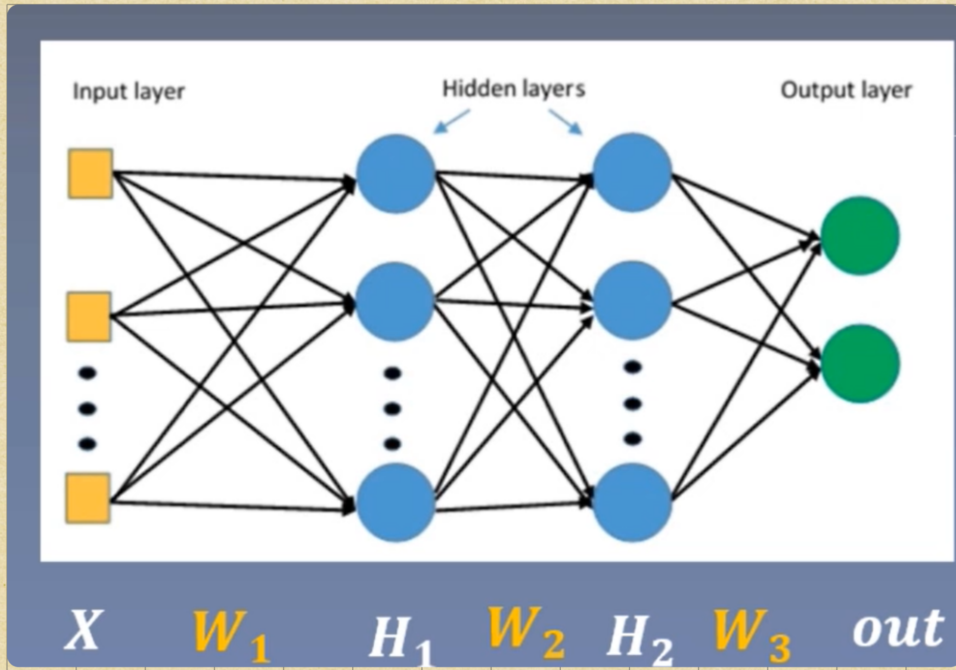

$H_2 = H_1 \times W_2$

$\Delta W_2 = \frac{\partial Loss}{\partial W_2} = \frac{\partial Loss}{\partial out} \cdot \frac{\partial out}{\partial H_2} \cdot \frac{\partial H_2}{\partial W_2} = \frac{\partial Loss}{\partial out} \cdot \frac{\partial out}{\partial H_2} \cdot H_1$

梯度消失：$H_1 \to 0$　$\Delta W_2 \to 0$

梯度爆炸：$H_1 \to \infty$　$\Delta W_2 \to \infty$

无激活函数时

$D(W) = \frac{1}{n} \Rightarrow std(W) = \sqrt{\frac{1}{n}}$ 时能保证方差始终为1

Xavier 初始化 (sigmoid, tanh)

$D(W) = \frac{2}{n_i + n_{i+1}} = \frac{a^2}{3}$　$[-a, a]$

$\Rightarrow W \sim U\left[-\frac{\sqrt{6}}{\sqrt{n_i + n_{i+1}}}, \frac{\sqrt{6}}{\sqrt{n_i + n_{i+1}}}\right]$

※需要传入激活函数的增益(gain)

Kaiming 初始化(ReLU及变种)

$$D(W) = \frac{2}{n_i} = \frac{2}{(1+a^2)\cdot n_i}$$

$$std(W) = \sqrt{\frac{2}{(1+a^2)\cdot n_i}} = \sqrt{\frac{2}{n_i}}$$

pytorch 十种初始化方法:

Xavier 均匀分布
Xavier 正态分布
kaiming 均匀分布
kaiming 正态分布
均匀分布
正态分布
常数分布
正交矩阵初始化
单位矩阵初始化
稀疏矩阵初始化

函数名 参数

nn.init.calculate_gain(nonlinearity, param).

计算激活函数方差尺度变化

$$\frac{std_{in}}{std_{out}}$$