人工智能基础 第 6 次作业

2、试设计一个前馈神经网络来解决 XOR（异或）问题，要求该前馈神经网络具有两个隐藏神经元和一个输出神经元，要求激活函数为 ReLU。

（a）. 请给出你设计的网络的具体参数，并验证其能够满足上述要求。

（b）. 请证明：若网络的输入限制为两输入 (x_1, x_2)，且激活函数为线性函数，则无法解决上述异或问题。

【解】:

（a）. 网络以及参数如下图所示。

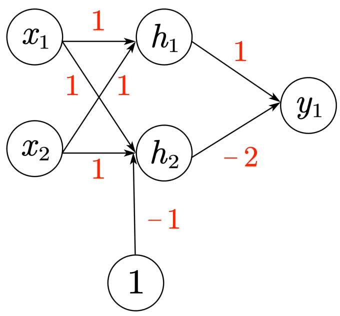

图 1: 第 2 题 (a)

隐藏层神经元的输出经过 ReLU 函数激活。则 h_1, h_2 的输出以及最终的输出 y 为：

$$\begin{bmatrix} h_1 \\ h_2 \end{bmatrix} = \mathrm{ReLU}\left(\begin{bmatrix} x_1 + x_2 \\ x_1 + x_2 - 1 \end{bmatrix}\right),$$

$$y = h_1 - 2h_2$$

- 当 $x_1 = 1, x_2 = 1$ 时，经计算可知 $h_1 = 2, h_2 = 1$，从而 $y = 0$。
- 当 $x_1 = 1, x_2 = 0$ 时，经计算可知 $h_1 = 1, h_2 = 0$，从而 $y = 1$。
- 当 $x_1 = 0, x_2 = 1$ 时，经计算可知 $h_1 = 1, h_2 = 0$，从而 $y = 1$。
- 当 $x_1 = 0, x_2 = 0$ 时，经计算可知 $h_1 = 0, h_2 = 0$，从而 $y = 0$。

综上所述，网络完成了 XOR（异或）的功能。

（b）. 激活函数为线性时，则无论中间过程，输出 y 必然为输入 x_1, x_2 的线性函数。

$$y = \omega_1 x_1 + \omega_2 x_2 + b$$

对于异或问题来说，输入输出对应关系如下：红色“+”表示输出为 1，蓝色“O”表示输出为 0。

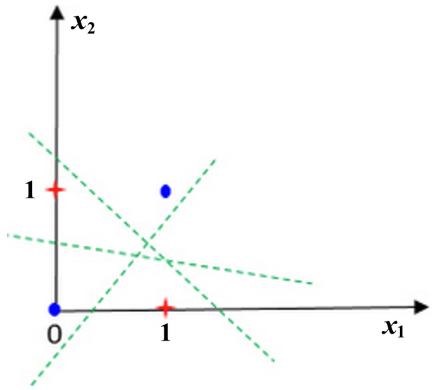

图 2: 第 2 题 (b)

不存在一条直线能将红色的“+”与蓝色的“O”分开，因此若 y 是 x_1, x_2 的线性函数，则无法解决异或问题。

4、现有两输入的前馈神经网络及其参数如下所示，且隐藏层和输出层神经元均使用了 Sigmoid 激活函数，请回答如下问题：

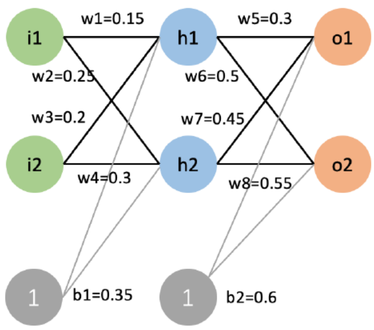

图 3: 第 4 题

(a)．当输入为 $(i_1, i_2) = (0.05, 0.10)$ 时，计算该神经网络的输出分别为多少。请写明必要的计算过程。

(b)．在（a）的基础上，若 $(o_1, o_2) = (0.12, 0.95)$，采用最小化均方误差作为优化准则，请根据反向传播算法计算参数 w_5 和 w_6 的梯度。

(c)．在（b）的基础上，若采用梯度下降更新参数，且学习率设为 0.1，写出更新后的参数 w_5 和 w_6。

【解】：

(a)． 首先计算隐藏层单元的输出：

$$h_1 = \text{sigmoid}\,(0.15 i_1 + 0.2 i_2 + 0.35) = \text{sigmoid}(0.3775) = 0.5932$$
$$h_2 = \text{sigmoid}\,(0.25 i_1 + 0.3 i_2 + 0.35) = \text{sigmoid}(0.3925) = 0.5969$$

从而输出为：

$$o_1 = \text{sigmoid}(0.3 h_1 + 0.45 h_2 + 0.6) = \text{sigmoid}(1.0466) = 0.7401$$
$$o_2 = \text{sigmoid}(0.5 h_1 + 0.55 h_2 + 0.6) = \text{sigmoid}(1.2249) = 0.7729$$

(b)． 最小化均方误差。设真实的输出为 (y_1, y_2)，则：

$$e = (\hat{o}_1 - y_1)^2 + (\hat{o}_2 - y_1)^2 = (0.7401 - 0.12)^2 + (0.7729 - 0.95)^2 = 0.3532$$

可知：

$$\frac{\partial e}{\partial o_1} = 2(o_1 - y_1) = 2 \times (0.7401 - 0.12) = 1.2402$$
$$\frac{\partial e}{\partial o_2} = 2(o_2 - y_2) = 2 \times (0.7729 - 0.95) = -0.3542$$

设 $z_1 = w_5 h_1 + w_7 h_2 + b_2$，$z_2 = w_6 h_1 + w_8 h_2 + b_2$，则 $o_1 = \text{sigmoid}(z_1)$，$o_2 = \text{sigmoid}(z_2)$。由此可得：

$$\frac{\partial o_1}{\partial z_1} = \frac{e^{-z_1}}{(1+e^{-z_1})^2} = o_1(1 - o_1) = 0.7401 \times (1 - 0.7401) = 0.1924$$
$$\frac{\partial o_2}{\partial z_2} = \frac{e^{-z_2}}{(1+e^{-z_2})^2} = o_2(1 - o_2) = 0.7729 \times (1 - 0.7729) = 0.1755$$
$$\frac{\partial z_1}{\partial w_5} = h_1 = 0.5932$$
$$\frac{\partial z_2}{\partial w_6} = h_1 = 0.5932$$

因此，根据链式法则，可以得到：

$$\frac{\partial e}{\partial w_5} = \frac{\partial e}{\partial o_1} \times \frac{\partial o_1}{\partial z_1} \times \frac{\partial z_1}{\partial w_5} = 1.2402 \times 0.1924 \times 0.5932 = 0.141525$$
$$\frac{\partial e}{\partial w_6} = \frac{\partial e}{\partial o_2} \times \frac{\partial o_2}{\partial z_2} \times \frac{\partial z_2}{\partial w_6} = -0.3542 \times 0.1755 \times 0.5932 = -0.036875$$

(c)． 按梯度下降更新参数，学习率 $\eta = 0.1$，则新的 w_5' 和 w_6' 为：

$$w_5' = w_5 - \eta \times \frac{\partial e}{\partial w_5} = 0.3 - 0.1 \times 0.141525 = 0.2858$$
$$w_6' = w_6 - \eta \times \frac{\partial e}{\partial w_6} = 0.5 - 0.1 \times (-0.036875) = 0.5037$$

5、关于激活函数，请回答如下的问题：

(a)．若在多层的神经网络中使用 Tanh（双曲正切单元）作为神经元的激活函数，请根据 BP 算法推导出参数梯度的表示形式，说明其可能导致梯度消失问题的原因，并提出一种解决方案。

(b)．"死亡 ReLU" 问题是指当使用 ReLU 作为激活函数时，大梯度流经某个神经元后，容易导致神经元输出始终为零，请举例说明该问题，并给出两种解决的思路。

(c)．对于下述的两种激活函数：

$$\mathrm{swish}(x) = x \cdot \mathrm{sigmoid}(\beta x)$$
$$\mathrm{GELU}(x) = xP(X \le x)$$

其中 β 为常数，X 服从标准高斯分布，其概率密度函数为 $f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$。

请求出这两种激活函数的导数，试比较 GELU 与 ReLU 激活函数，并说明 GELU 激活函数的优势。

【解】：

(a)． 设损失函数为 e，对于某一层的神经元而言：

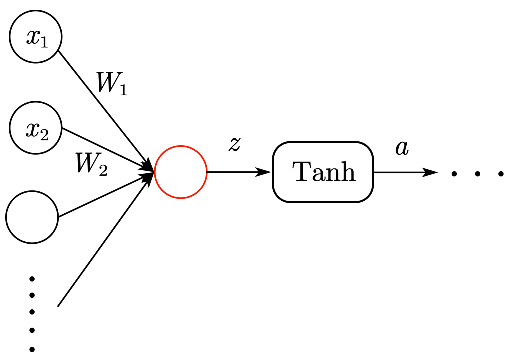

图 4: 第 5 题 (a)

对于参数 W_1 的梯度为:

$$
\begin{aligned}
\frac{\partial e}{\partial W_1} &= \frac{\partial e}{\partial a} \times \frac{\partial a}{\partial z} \times \frac{\partial z}{\partial W_1} \\
&= \frac{\partial e}{\partial a} \times \frac{\mathrm{d}}{\mathrm{d}z}\left(\frac{e^z - e^{-z}}{e^z + e^{-z}}\right) \times x_1 \\
&= \frac{\partial e}{\partial a} \times (1+a)(1-a) \times x_1
\end{aligned}
$$

对其它参数可同理得到。

当 z 较大或较小时，导致 a 接近于 1 或 -1，从而使得 $\frac{\partial a}{\partial z} = (1+a)(1-a) \approx 0$。在链式法则中，多个导数相乘，则有可能使得参数的梯度接近 0，从而参数难以更新。

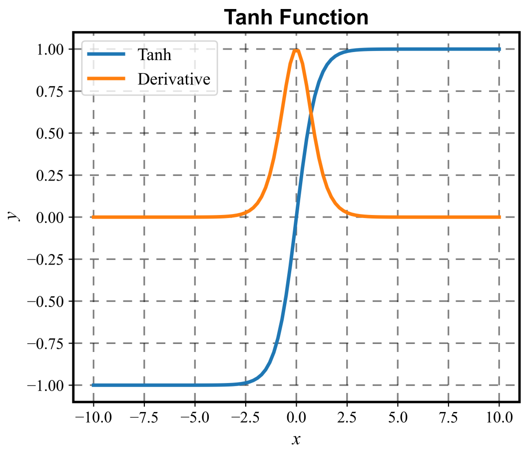

图 5: Tanh 函数及其导数

解决方案：改用其它的激活函数，或是采用 Batch Normalization，通过对每一层输出进行规范，削弱 W 带来的放大缩小的影响，使 Tanh 的输入不会过大过小而导致饱和。

(b) . 大梯度流经某个神经元，若此时学习率恰好又较大，会使得其权重更新量很多。有可能导致对于正常的训练集样本输入，输出 $z = W_1x_1 + W_2x_2 + W_3x_3 + W_4x_4$ 都是负值。而激活函数为 ReLU。

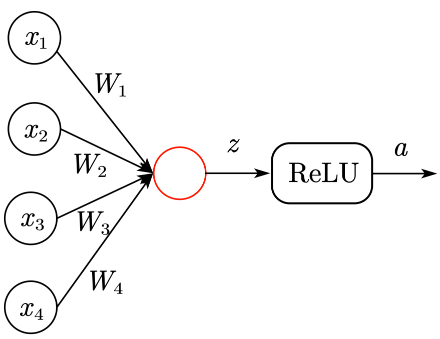

图 6: 第 5 题 (b)

设损失函数为 e，则它对于参数 W_1 的导数为：

$$\begin{aligned}\frac{\partial e}{\partial W_1} &= \frac{\partial e}{\partial a} \times \frac{\partial a}{\partial z} \times \frac{\partial z}{\partial W_1} \\ &= \frac{\partial e}{\partial a} \times 0 \times x_1 \\ &= 0\end{aligned}$$

对 W_2, W_3, W_4 同理。则无论学习率为何值，$W_1 \sim W_4$ 都无法得到更新，将一直保持这个非常差的状态，导致神经元的输出始终为 0。

解决思路：1、采用 Leaky ReLU 等激活函数，使得输入小于零时也有非零的输出和梯度。2、采用较小的学习率或采用学习率调整策略。

(c)． 计算导数：

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}(\mathrm{Swish}(x)) &= \mathrm{sigmoid}(\beta x) + x \times \frac{\mathrm{d}}{\mathrm{d}x}(\mathrm{sigmoid}(\beta x)) \\ &= \frac{1}{1+e^{-\beta x}} + x \cdot \frac{\beta e^{-\beta x}}{\left(1+e^{-\beta x}\right)^2}\end{aligned}$$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}(\mathrm{GELU}(x)) &= P(X \le x) + x \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \\ &= \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t + x \cdot \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)\end{aligned}$$

绘制这几种激活函数的图像：

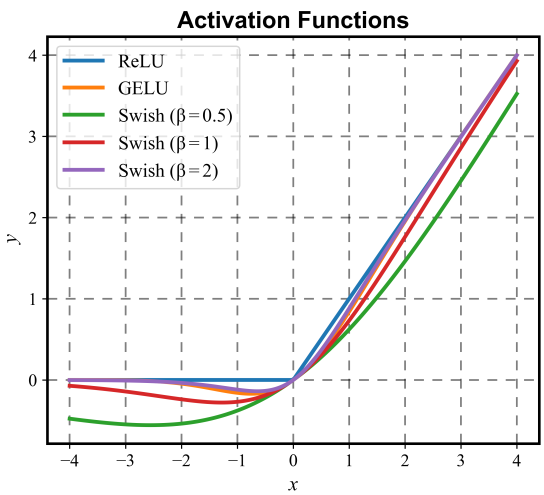

图 7: 激活函数的图像

它们的一阶导数的函数图像：

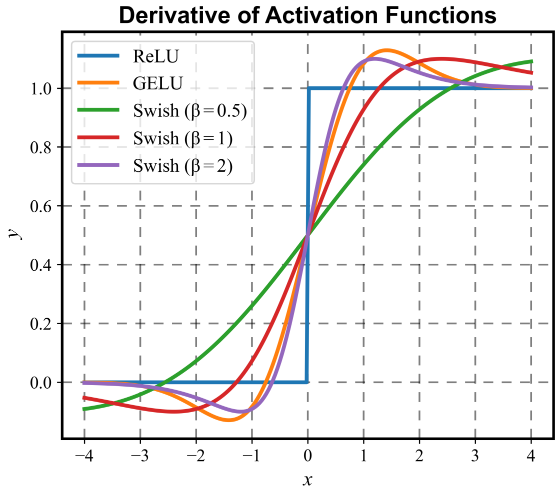

图 8: 激活函数导数的图像

比起 ReLU，GELU 函数更加光滑；GELU 在输入小于零时也可以有非零的输出和梯度，在一定程度上避免了神经元“死亡”的问题。