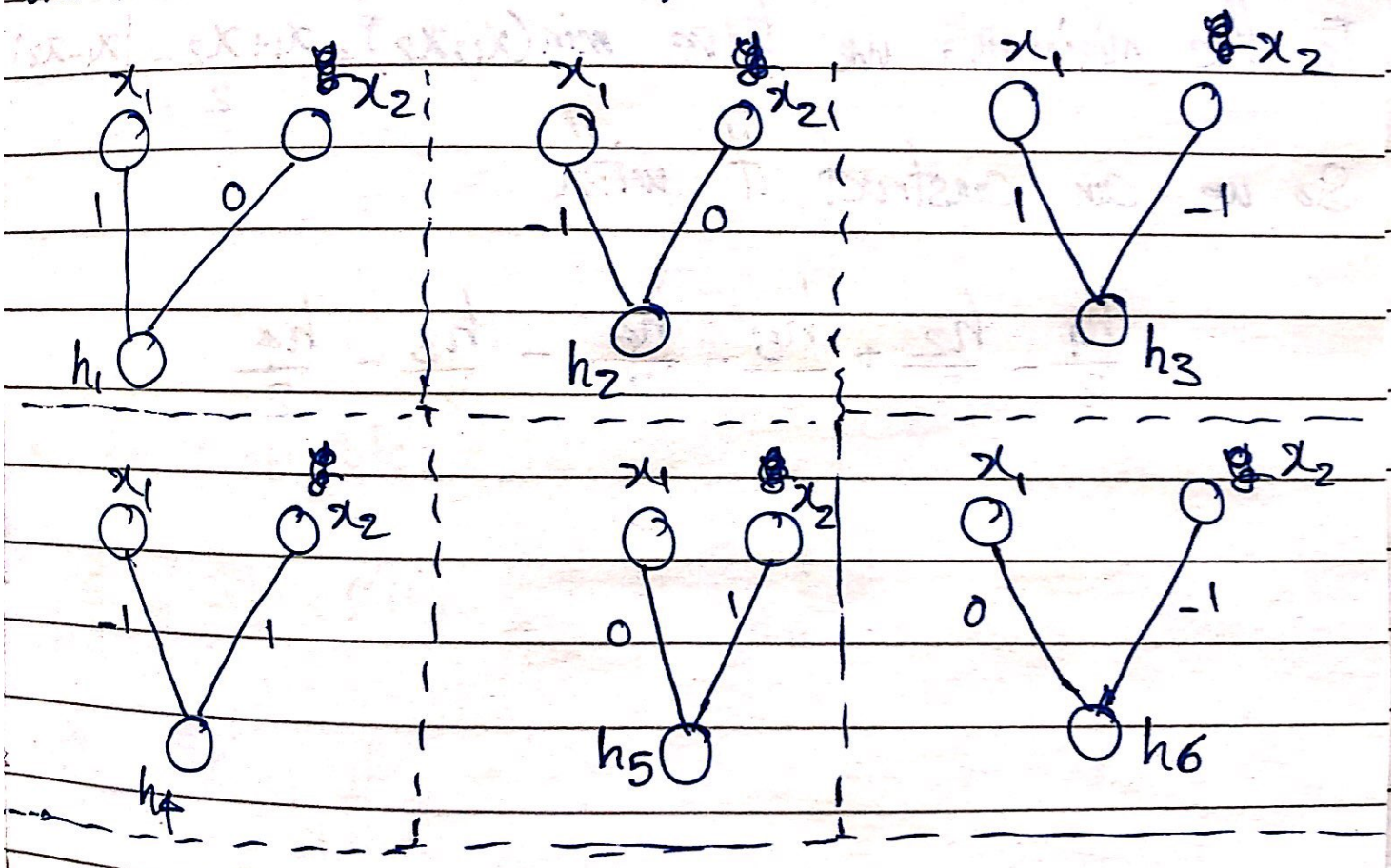
1. Note that $\max(x_1, x_2) = \frac{x_1 + x_2 + |x_1 - x_2|}{2}$.

We use this fact and construct the following network (To make things more clear, I have drawn hidden units seperately):



Note that $h_1 = \text{Relu}(x_1 + 0 * x_2)$ and so on for the other hidden units. It's easy to check that $h_1 - h_2 = x_1$, $h_5 - h_6 = x_2$ and $h_3 + h_4 = |x_1 - x_2|$. Therefore with the output unit as

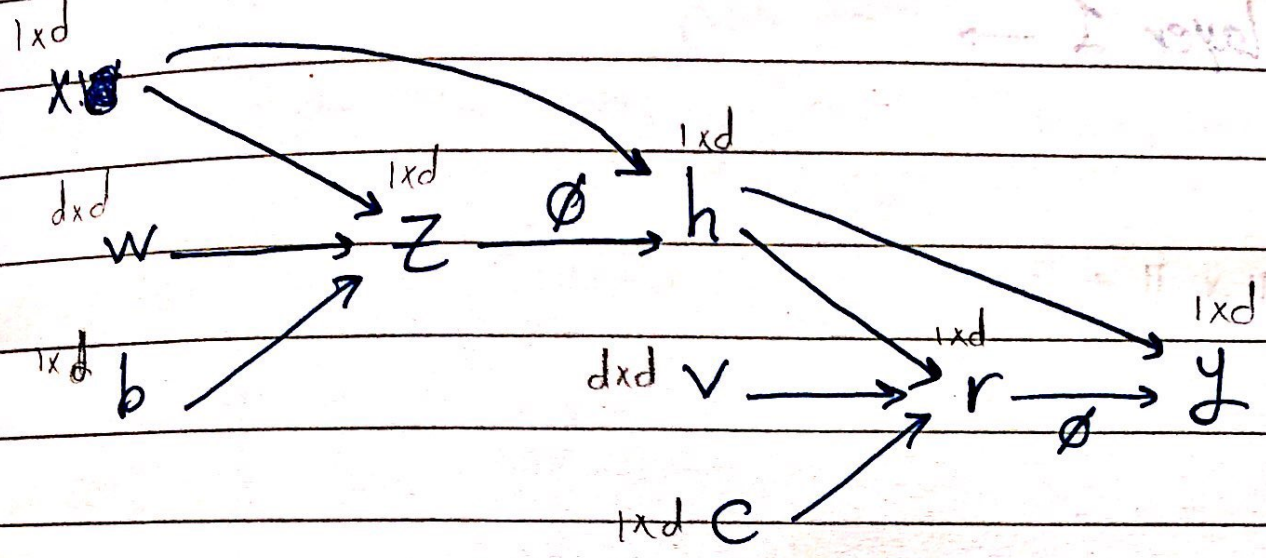
$O = \frac{(h_1 - h_2)}{2} + \frac{(h_3 + h_4)}{2} + \frac{(h_5 - h_6)}{2}$ [which is a linear layer with $+\frac{1}{2}$ and $-\frac{1}{2}$ weights] we can output $\max(x_1, x_2)$.

For the minimum, we have $\min(x_1, x_2) = \frac{x_1 + x_2 - |x_1 - x_2|}{2}$

So we can construct it with

$$\frac{h_1}{2} - \frac{h_2}{2} + \frac{h_5}{2} - \frac{h_6}{2} - \frac{h_3}{2} - \frac{h_4}{2}$$

2. Computation Graph:



b) backprop rules:

$$\frac{\partial y}{\partial c} = \frac{\partial y}{\partial r} = \phi(r) \qquad \frac{\partial y}{\partial V} = h^T \phi(r)$$

$$\frac{\partial y}{\partial h} = \phi(r)\, r + \boxed{1} \longrightarrow \text{vector of ones}$$

$$\frac{\partial y}{\partial z} = \frac{\partial y}{\partial h} \cdot \phi'(z) \qquad \frac{\partial y}{\partial b} = \frac{\partial y}{\partial z}$$

$$\frac{\partial y}{\partial W} = X^T \frac{\partial y}{\partial z} \qquad \frac{\partial y}{\partial X} = \frac{\partial y}{\partial z} W + \boxed{1}$$

vector of ones

3. a)

Conv. layer 1:

#units: 55 * 55 * 48 * 2 = 290,400

#weights: 96 * 11 * 11 * 3 = 34,848

#connections: #units * 11 * 11 * 3 = 10,542,972

Conv. layer 2:

#units: 186,624

#weights: 256 * 5 * 5 * 48 = 307,200

#connections: #units * 5 * 5 * 48 = 223,948,800

Conv. layer 3: #units: 64,896

weights: 384 * 3 * 3 * 256 = 884,736

Connections: #units * 256 * 3 * 3 = 149,520,384

Conv. layer 4: #units: 64,896

weights: 384 * 3 * 3 * 192 = 663,552

Connections: #units * 192 * 3 * 3 = 112,140,288

Conv. layer 5: #units: 43,264

weights: 256 * 3 * 3 * 192 = 442,368

Connections: #units * 192 * 3 * 3 = 74,760,192

F.C. layer 1 : # units: 4096

weights: 9216 * 4096 = 37,748, 736

connections: 37,748, 736

F.C. layer 2: # units: 4096

weights: 4096 * 4096 = 16,777, 216

connections: 16,777, 216

Output layer: # units: 1000

weights: 4096 * 1000 = 4,096,000

connections: 4,096,000

b)

i. I would reduce the number of F.C. layers or reduce the nomber of units in them as most of the parameters are in these layers.

ii. I would reduce the number of units of the convolutional layers. as most of the connection are in these layers. This can be done by reducing the depth of the convolutional part of the model, or having a bigger stride for kernels, or having more max-pooling layers (or max-pooling layers with bigger sizes), or having smaller number of kernels for each layer.