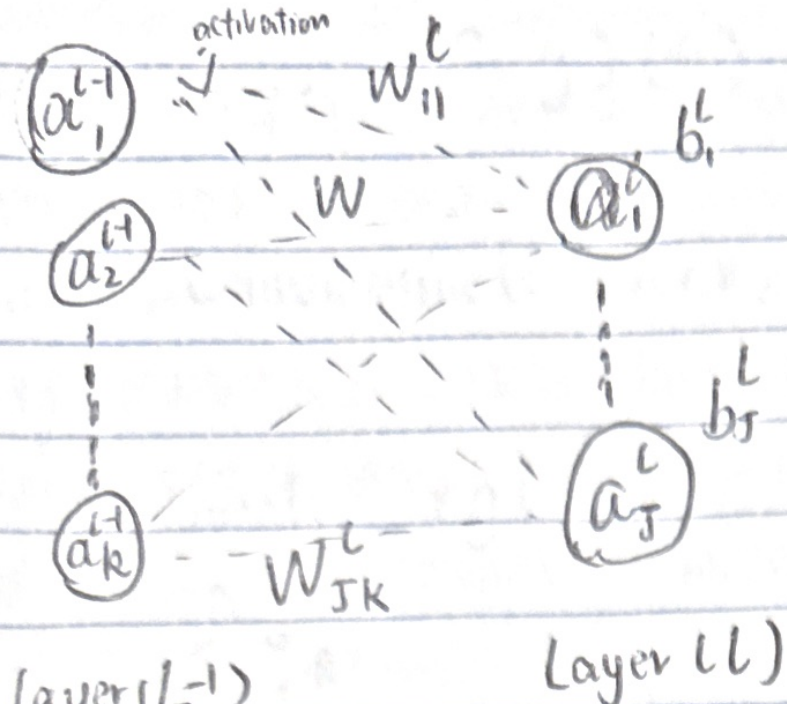

w (weights) is for different edges (each pair)

b (bias)

$$a_j^l = \sigma\left(\sum_k w_{jk}^l a_k^{l-1} + b_j^l\right)$$

σ : activation function

$$C_1 = \frac{1}{2}\sum_j (y_j - a_j^l)^2$$

$$C = \frac{1}{n}\sum_n \frac{1}{2}\sum_j (y_j - a_j^l)^2 = \frac{1}{n}\sum_n C_1 = \frac{1}{2n}\sum_j (y_j - a_j^l)^2$$

n : # of sample for training

Gradient Descent

$$w = w - \eta \frac{\partial C}{\partial w}$$

η : learning rate

too high: fail to converge

too low: inefficient, slow to converge