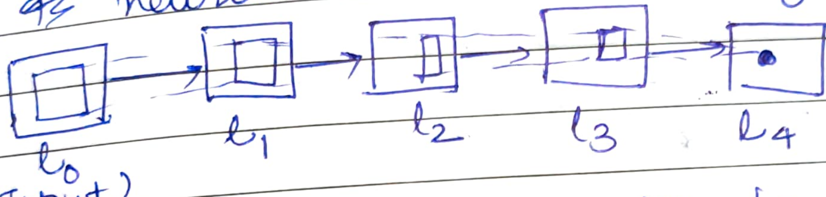
Assignment-3

Puneet Mangla CS17BTECH11029

Q.2 When a 3x3 filter with stride 1 is applied on an $n \times n$ input, the dimensions get reduced to $(n-2)\times(n-2)$

For ~~4th~~ neuron in 4^{th} non-image layer



the support at l_3 will be 3×3
" " at l_2 " " 5×5
" " at l_1 " " 7×7
" " at l_0 " " 9×9

Thus, the support ~~is~~ at l_0 or input layer is 81

Q.3 Adding an extra hidden layer will decrease the bias and increase the variance.
Adding a hidden layer will increase the representation power and thus will lead to lower bias on given run.

However, Adding layer will increase the complexity of ANN, making loss surfaces complex and difficult to converge at a same point, leading to ~~diff~~ variance in performance.

Q.5 $E(w) = E(w^*) + \frac{1}{2}(w-w^*)^T H (w-w^*)$

and H has eigenvalues λ_i corresponding to eigenvector u_i

By linear independence of eigenvectors u_i

$$(w-w^*) = \sum_{i} \alpha_i u_i = U\alpha$$

where $U = [u_1, u_2\ u_3 \dots u_n]$ and $\alpha = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{bmatrix}$

H has eigenvalue decomposition ie

$H = U \Lambda U^T$ where $\Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \lambda_2 & \\ 0 & & \ddots \ \lambda_n \end{bmatrix}_{n \times n}$

and $UU^T = I$

now $E(w) = E(w^*) + \frac{1}{2}(U\alpha)^T U \Lambda U^T (U\alpha)$

$= E(w^*) + \frac{1}{2}\alpha^T U^T U \Lambda U^T U \alpha$ $\{ U^T U = I$

$= E(w^*) + \frac{1}{2}\alpha^T \Lambda \alpha$

$= E(w^*) + \frac{1}{2}[\alpha_1^2 \lambda_1 + \lambda_2 \alpha_2^2 \dots \lambda_n \alpha_n^2]$

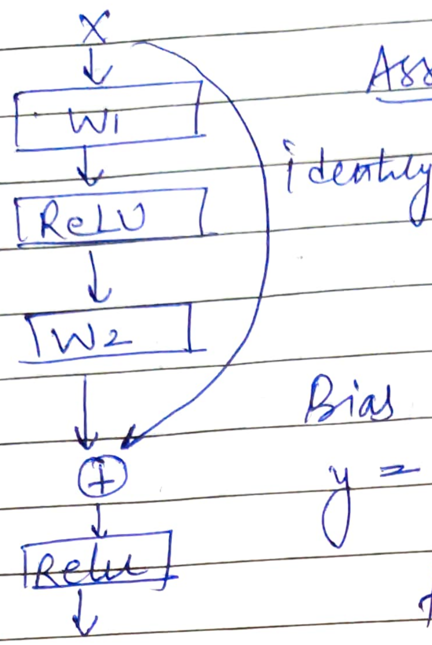
$= E(w^*) + \frac{1}{2}\sum_i \lambda_i \alpha_i^2$

$\left\{ \frac{x_1^2}{a_1^2} + \frac{x_2^2}{a_2^2} \dots = 1 \right\}$

This represents an elipse. Since we represented vectors as linear combination of U_i's, the axis will be U_i's

now the length $a_i = \frac{\sqrt{2(E(w) - E(w^*))}}{\sqrt{\lambda_i}}$ Hence $a_i \propto \frac{1}{\sqrt{\lambda_i}}$

Q.1



Assumption: we assume a simple 2 layer fully connected residual block with Relu activation. Let $\sigma = \text{Relu}()$

Bias is false.

$$y = \cancel{\sigma \&}\ \sigma(W_2 \cdot \sigma(W_1 x) + x)$$

↓ feed forward equation

$$\cancel{\frac{\partial Y}{\partial W_2} = (W_2 \sigma(W_1 x) + x)\ \sigma(W}$$

$$\frac{\partial Y}{\partial W_2^{ij}} = \sigma'(W_2\, \sigma(W_1 x) + x) \cdot (\sigma(W_1 x)_i)$$

Let $Z = W_1 x$

$$\frac{\partial Y}{\partial Z_i} = \sigma'(W_2\, \sigma(W_1 x) + x) \cdot \sum_{\forall j} W_2^{ji} \cdot \sigma'(Z_i)$$

$$\frac{\partial Y}{\partial W_1^{ij}} = \frac{\partial Y}{\partial Z_j} \cdot \frac{\partial Z_j}{\partial W_2^{ij}} = \frac{\partial Y}{\partial Z_j} \cdot x_i$$

$$\frac{\partial Y}{\partial x_j} = \sigma'(W_2\, \sigma(W_1 x) + x) + \left[\sum_{\forall i} \frac{\partial Y}{\partial Z_i} \cdot \frac{\partial Z_i}{\partial x_j} \right]$$

$$= \sigma'(W_2\, \sigma(W_1 x) + x) + \left[\sum_{\forall i} \frac{\partial Y}{\partial Z_i} \cdot \sigma'(x_j \cdot W_{ji}') \right]$$

Q.4 $\sigma(a) = \frac{1}{1+e^{-a}}$ also $\sigma(-a) = 1 - \sigma(a)$

$$\tanh(a) = \frac{e^{a} - e^{-a}}{e^{a}+e^{-a}} = \frac{1-e^{-2a}}{1+e^{-2a}} = \frac{1}{1+e^{-2a}} - \frac{1}{1+e^{2a}}$$

$$= \sigma(2a) - \sigma(-2a) \quad \{\because \sigma(-a) = 1-\sigma(a)\}$$

$$= 2\cdot\sigma(2a) - 1 \quad \text{or} \quad \sigma(a) = \frac{\tanh(a/2)+1}{2}$$

$$y_k(x,w) = \sum_{j=1}^{M} w_{kj}^{(2)} \sigma\left(\sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)}\right) + w_{k0}^{(2)} \quad -(1)$$

$$y_k(x,\tilde{w}) = \sum_{j=1}^{M} \tilde{w}_{kj}^{(2)} \tanh\left(\sum_{i=1}^{D} \tilde{w}_{ji}^{(1)} x_i + \tilde{w}_{j0}^{(1)}\right) + \tilde{w}_{k0}^{(2)} \quad -(2)$$

$$y_k(x,w) = \sum_{j=1}^{M} w_{kj}^{(2)} \left[\frac{\tanh\left(\frac{\cancel{a}\sum_{i=1}^{D} w_{ji} x_i + w_{j0}^{(1)}}{2}\right) + 1}{2}\right] + w_{k0}^{(2)}$$

$$y_k(x,w) = \sum_{j=1}^{M} w_{kj}^{(2)} \frac{1}{2} \tanh\left(\frac{\sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)}}{2}\right) + \sum_{j=1}^{M} w_{kj}^{(2)} \frac{1}{2} + w_{k0}^{(2)} \quad -(3)$$

Comparing (2) and (3)

$$\tilde{w}_{kj}^{(2)} = \frac{w_{kj}^{(2)}}{2}, \quad \tilde{w}_{ji}^{(1)} = \frac{w_{ji}^{(1)}}{2}, \quad \tilde{w}_{j0}^{(1)} = \frac{w_{j0}^{(1)}}{2}$$

$$\tilde{w}_{k0}^{(2)} = w_{k0}^{(2)} + \sum_{j=1}^{M} w_{kj}^{(2)} \frac{1}{2}$$

Q.6 Since the custom data set is small and both target and source datasets are similar.

Q.4

- We will use transfer learning, where ~~specti~~ specialized features and generic layers both will be fixed.

- Parameters of final classification will be initialized again and trained.