Name- Pratham Singhal 2021082

__/__/__

ASSIGNMENT – I

Answer 1) (a) MSE does not take account of the probabilistic nature of the dependent and independent variable because it assumes the relationships between two ~~to be~~ as linear. Also in classification based learning, it leads ~~to~~ to slow convergence.

(b) $BCE = -[(+y \log \hat{y}) + ((1-y)\log(1-\hat{y}))]$

(c) $BCE = -[y \log \hat{y} + (1-y)\log(1-\hat{y})]$

$\Rightarrow y = 0 \quad \hat{y} = 0.9$

$BCE = -1 \log(0.1)$

$\boxed{BCE = 2.303}$

(d) $Avg\ Loss = \frac{-1}{N}\sum_{1}^{N}\left(y \log \hat{y} + (1-y)\log(1-\hat{y})\right)$

$= -\frac{1}{3}\left[1\log(0.1) + 1(\log(0.8) + 1\log(0.3)\right.$

$= \frac{-1}{3}\left[-(\log 2.303 + 0.223 + 1.204)\right.$

Avg Loss = 1.243

(e) with L2 regularization.

$$L_{BCE} = \frac{1}{m}\sum -y\log\hat{y} + -(1-y)\log(1-\hat{y}) + \frac{\lambda}{2m}\sum_{i=1}^{n} w_i^2$$

Regularization Term

$$L_{BCE} = 1/m \sum -y \log \sigma(z) - (1-y)\log(1-\sigma(z)) + \frac{\lambda}{2m}\sum w_j^2$$

$$\partial L/\partial w = \frac{-1}{m}\left(\sum y \cdot \frac{1}{\sigma(z)} \cdot \sigma(z)[1-\sigma(z)] \cdot \frac{\partial w^T x}{\partial w} + (1-y)\frac{\partial w x}{\partial w} \cdot \frac{1}{1-\sigma(z)}[-\sigma(z)(1-\sigma(z))]\right) + \frac{\lambda}{2m} \cdot 2\sum_{i=1}^{N} w_j$$

$$\frac{\partial L}{\partial w} = \frac{1}{m}\left(-\left(\sum (y-\hat{y}) x_i\right) + \lambda \sum_{i}^{N} w_j\right.$$

$$W_{New} = W_{old} - \alpha\left[\frac{1}{m}\left[\left[-\left(\sum_{j=1}^{i=m}(y-\hat{y})x_j\right) + \lambda\Sigma w_j\right]\right]\right]$$

f) KL divergence is a measure of the average distribution diverges from a second, expected probability distribution.

Cross entropy $H(P,Q)$ measure of average no. of bits needed to identify an event from a set of possibilities if a coding scheme is used based on a given prob of distribution Q, rather than the "true" distribution P.

KL - divergence

$$D_{KL}(P\|Q) = \sum P(x) * \log\left(\frac{P(x)}{Q(x)}\right)$$

i.e.

$$D_{KL}(P\|Q) = \sum P(x)(\log P(x) - \log Q(x))$$

$$= \sum P(x)\log P(x) - \sum P(x)\log Q(x)$$

$$-\sum P(x)\log P(x) + D_{KL}(P\|Q) = -\sum P(x)\log Q(x)$$

i.e $E_{x\sim p}[\log P] + D_{KL}(P\|Q) = E_{x\sim p}[\log Q]$

$H P + D_{KL}(P\|Q) = H(P,Q)$

Ans 2) a) $W_2 = K \times D_a$

$b_2 = K \times 1$

$(D_a \times m)$

b) $\frac{\partial \hat{y}_k}{\partial z_k^{[2]}} = \hat{y}_k(1-\hat{y}_k)$

c) $\frac{\partial \hat{y}_k}{\partial z_i^{[2]}} = -\hat{y}_k \hat{y}_p$

d) $L = \sum -y_i \log \hat{y}_i$ where, $\hat{y}_i = \text{softmax}(z_i^{[2]})$

$\frac{\partial L}{\partial z_i} = \frac{\partial \sum -y_i \log \hat{y}_i}{\partial z_i}$ $\hat{y}_k = \text{softmax}(z_k)$

$= \frac{\partial}{\partial z_i} \sum -y_i \log(\text{softmax}(z_i^{[2]}))$

$= -y_j \sum \frac{\partial}{\partial z_i} \log(\text{softmax}(z_i^{[2]}))$

$$-y_i \frac{1}{\text{softmax}(z_i)} \cdot \frac{\partial}{\partial z_i} \text{softmax}(z_i)$$

$$= \frac{-y_i}{\text{softmax}(z_i)} \cdot \frac{\partial}{\partial z_i} \frac{e^{z_i}}{\sum e^{z_i}}$$

Case 1

$$= -y_i \cdot \hat{y}_i (1 - \hat{y}_i) \quad (i = k)$$

$$\frac{\partial L}{\partial z_i} = -y_i (1 - \hat{y}_i)$$

Case 2 $\quad i \neq k$

$$\frac{\partial L}{\partial z_i} = \frac{-y_i}{\hat{y}_i} \cdot -\hat{y}_k \hat{y}_i \quad \left[\frac{\partial}{\partial z} \text{softmax}\, i_k = -\hat{y}_k \hat{y}_i \right]$$

$$\frac{\partial L}{\partial z_i} = y_i \hat{y}_k$$

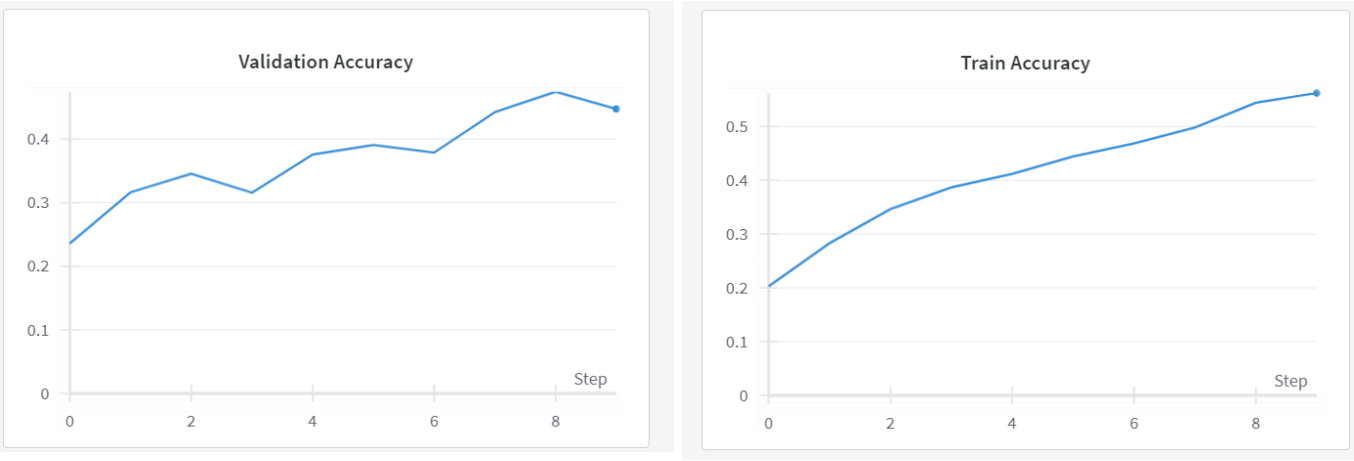
(c) Num stability in softmax is due to computation of exponents of large +ve / -ve values causing an overflow / underflow issue.

Overflow can lead to ∞ & underflow can lead to 0.

This can issue a provides meaningful gradients to the function

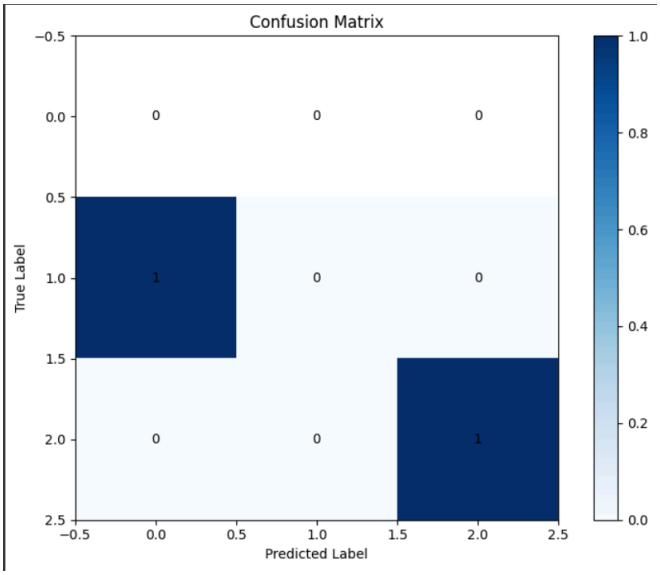
$$\hat{y}_k = \frac{e^{z_k^{[2]} - \max(z^{[2]})}}{\sum_{j=1}^{K} e^{z_j^{[2]} - \max(z^{[2]})}}$$

Answer2>
Ans 2.2.c>

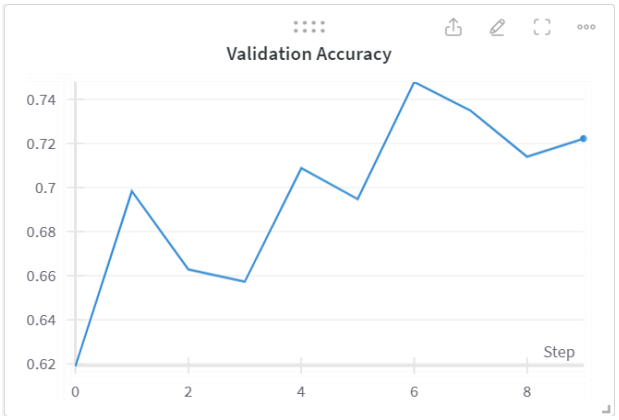


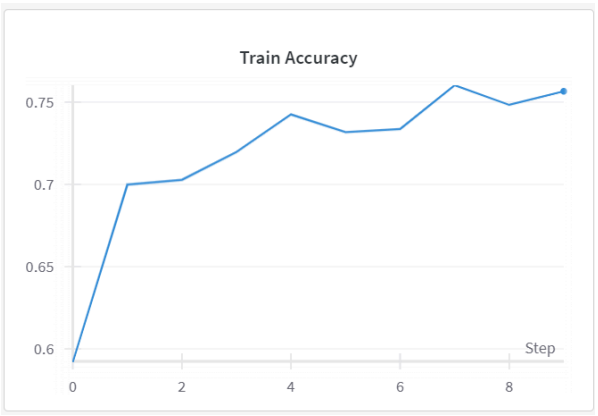
The trained CNN model is overfitting to a slight extent, since the training accuracy is 56% and validation accuracy is 46%.

ans2.2.d>
Test accuracy is 58% and F1 score = 0.5



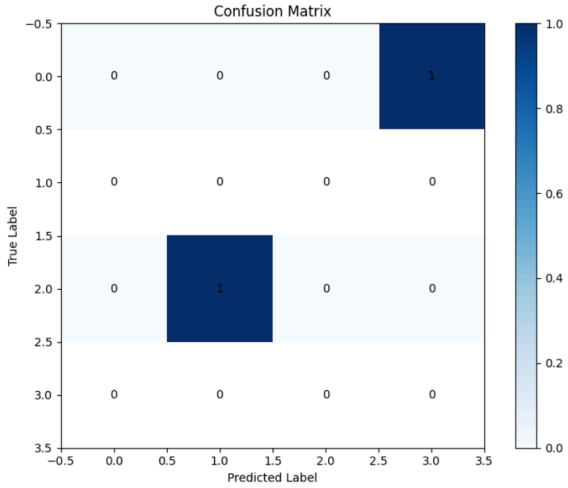
Ans 2.3.b>



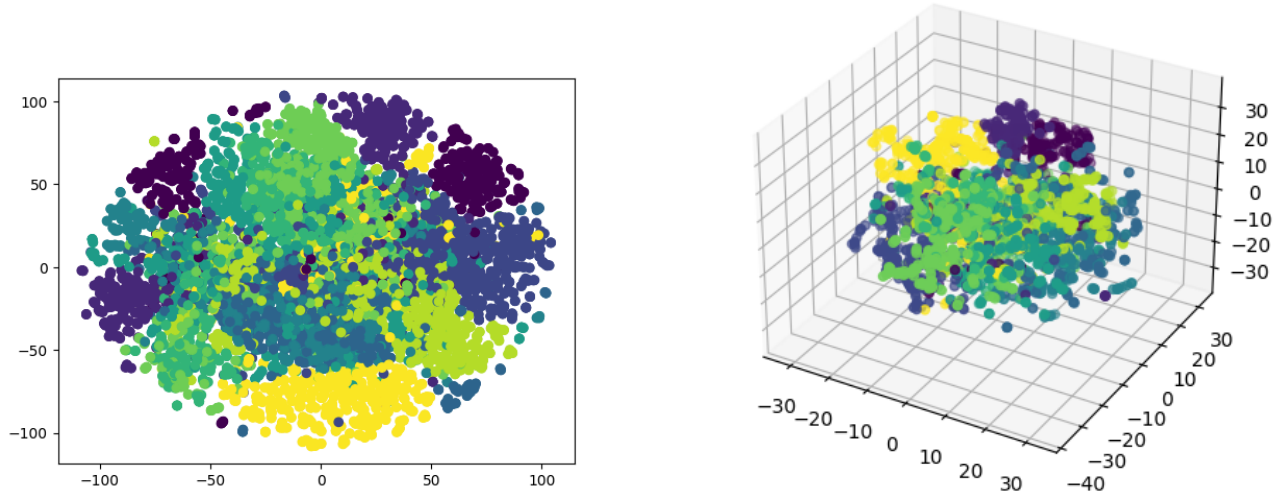


The model is not overfitting, since the training accuracy is 75% and validation accuracy is 72%. The gap is low between them.

Ans2.3.c> Test Accuracy = 71% and f score = 0.68



Ans 2.3.d>



Ans2.4.c>

The test accuracy = 71 % and F1 score = .69

Ans2.6> the cnn model performed a little less accurately compared to all the other models.It was overfitting to a bit. The resnet model without data augmentation performed better than the other all models. This was the best accurate model, but was overfitting a little. The resnet model with data augmentation was not overfitting but performed a little less accurately. This was due to the different types of datasets.