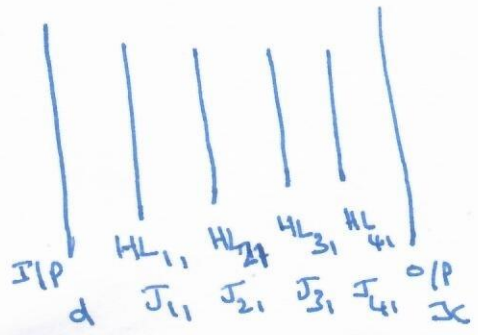
1. (a) Adam is expected to give the best convergence mainly because it combines the momentum based method and RMS Prop. Numerator corresponds to the exponentially weighted average of the first moment of gradient. Denominator term corresponds to the exponentially weighted average of the second moment of gradient.

(b) Internal covariant shift: change in the distribution of output of a hidden node due to different mini-batches. Mean and variance used in normalization are computed using the output of a node for examples in a mini-batch. The scale and shift parameters are learnt during training.

2. (b) The $AANN_1$ is trained using $\mathcal{D}_1 = \{\bar{x}_n\}_{n=1}^{N}$.
After $AANN_1$ is trained, every $\bar{x}_n$ is passed through it to get $\bar{z}_{1n}$.
The $AANN_2$ is trained using $\mathcal{D}_2 = \{\bar{z}_{1n}\}_{n=1}^{N}$.
After $AANN_2$ is trained, every $\bar{z}_{1n}$ is passed through it to get $\bar{z}_{2n}$.
This process is continued to train $AANN_3$ and $AANN_4$.

(c)




$HL_{i1}$ is the first hidden layer of $AANN_i$, with $J_{i1}$ nodes. $i = 1, 2, 3, 4$

3. Grey level image.

Input: $W \times H \times 1$

CL : F×F kernel — Tanh

D feature maps

$W_1 \times H_1 \times D$

$$W_1 = \left\lfloor \frac{W - 2F + 2P}{S} \right\rfloor + 1$$

$$W_1 = W - 2F + 1$$

$$H_1 = H - 2F + 1$$

PL : $W_2 = \frac{W_1}{2}$

$H_2 = \frac{H_1}{2}$

$W_2 \times H_2 \times D$

FC : L nodes — Tanh

O/P Layer : K nodes — Logistic

(a) No. of Connections:

CL: No. of neurons in a feature map: $W_1 \times H_1$

No. of connections to a neuron: $F \times F$.

Total no. of connections: $W_1 \times H_1 \times F \times F \times D$

FC: No. of inputs to a neuron in FC layer: $W_2 \times H_2 \times$

No. of connections to FC layer: $W_2 \times H_2 \times D \times L$

O/P Layer: No. of connections: $K \times L$

Total no. of connections with weights: $W_1 \times H_1 \times F \times F \times D +$
$W_2 \times H_2 \times D \times L + K$

No. of weight parameters in CL to be learnt: $F \times F \times D$

Total no. of weight parameters: $F \times F \times D + W_2 \times H_2 \times D \times L$
$+ K * L$

(b) $\tilde{E} = \frac{1}{2} \sum_{k=1}^{K} (t_k - y_k)^2$

$$\Delta w_{\lambda k} = -\eta \frac{\partial \tilde{E}}{\partial w_{\lambda k}}$$

$$= \eta (t_k - y_k) \cdot \frac{d \phi_o(a_k)}{d a_k} \cdot A_\lambda^{FC}$$

$$\Delta w_{\lambda k} = \eta (t_k - y_k) \beta_o \cdot \phi_o(a_k) (1 - \phi_o(a_k)) \cdot A_\lambda^{FC}$$

$$\delta_k = (t_k - y_k) \cdot \beta_o \phi_o(a_k) (1 - \phi_o(a_k))$$

(c) $\Delta w_{mijk} = \eta \delta_\lambda A_{mij} \cdot r_{mij}$

$$\delta_\lambda = \left( \sum_{k=1}^{K} w_{\lambda k} \delta_k \right) \cdot \frac{d \phi_{FC}(a_\lambda)}{d a_\lambda}$$

$$\delta_\lambda = \left( \sum_{k=1}^{K} w_{\lambda k} \delta_k \right) \cdot \beta_{FC} \left(1 - \phi_{FC}^2(a_\lambda)\right)$$

$A_{mij}$ : output of $(i,j)^{th}$ neuron in the $m^{th}$ FM

$r_{mij}$ : 1, if $A_{mij}$ is maximum of all the inputs to the neuron in PL to which the $(i,j)^{th}$ neuron is connected

= 0, otherwise.

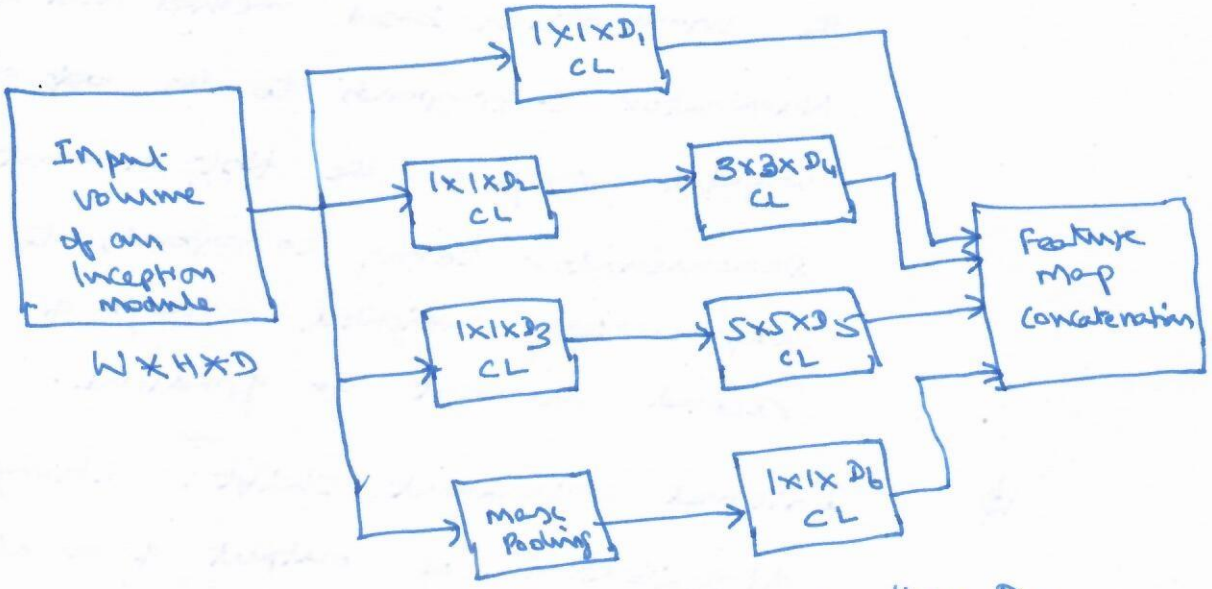
(d) $\Delta w_{pql}^m = \frac{1}{W_1 \times H_1} \sum_{i=1}^{W_1} \sum_{j=1}^{H_1} \Delta w_{pqij}^m$

$$\Delta w_{pqij}^m = \eta \delta_{ij}^m x_{pq}$$

$$\delta_{ij}^m = \left( \sum_{\lambda=1}^{L} \delta_\lambda \cdot w_{mij\lambda} \cdot r_{mij} \right) \cdot \frac{d \phi_{CL}(a_{ij}^m)}{d a_{ij}^m}$$

$$= \left( \sum_{\lambda=1}^{L} \delta_\lambda \cdot w_{mij\lambda} \cdot r_{mij} \right) \cdot \beta_{CL} \left(1 - \phi_{CL}^2(a_{ij}^m)\right)$$

4. Typical structure of inception module:




$D_1$, $D_2$, $D_3$, $D_6$ are smaller than D.

Main purpose: Extract features with different resolution, using kernels of different sizes.

1x1x $D_i$ CLs are used to reduce the depth of volume on which kernels of different sizes are applied.