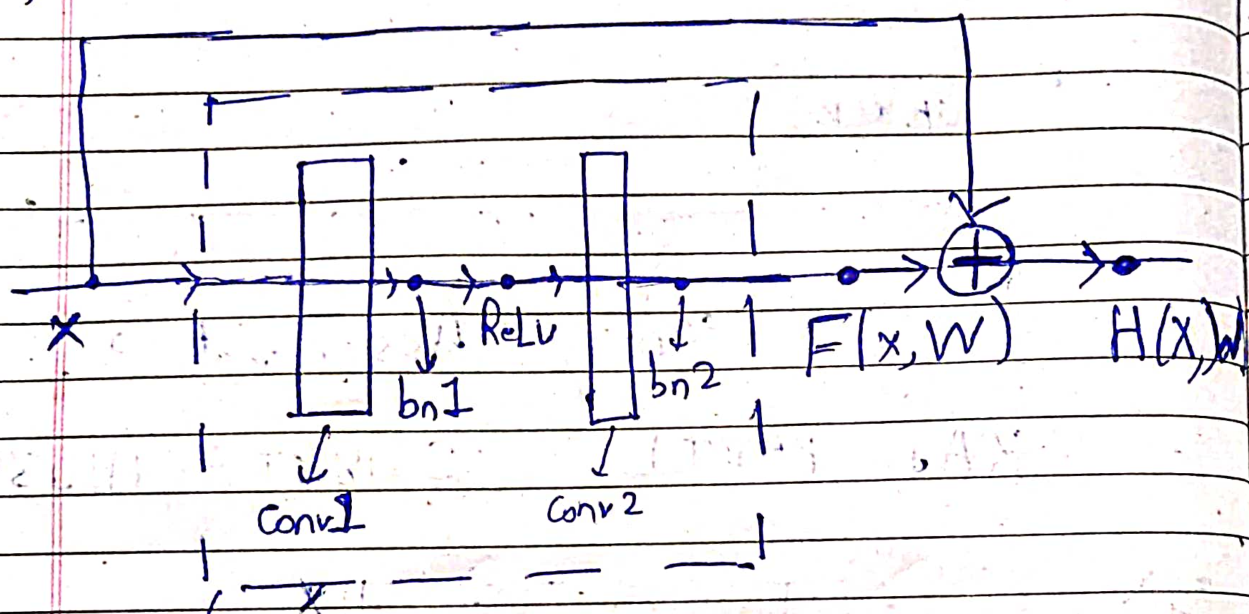
Deep Learning : Assignment 3

RAJ PATIL : CS18BTECH11039

1.) Basic Block :-



Representing all weights of the basic Block by "W" & input as "X"

Then activation of residual Block is :- $F(X,W)$

Now, during Back Prop, $\dfrac{\partial L}{\partial H(X,W)}$ is available

need 2 things to update weights & allow backprop for previous Layers :-

i) $\dfrac{\partial L}{\partial W}$ ii) $\dfrac{\partial L}{\partial X}$

Note that $H(x) = F(x,w) + x$

now

i) $$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial H} \cdot \frac{\partial H}{\partial F} \cdot \frac{\partial F}{\partial W}$$

available

$$\cancel{\frac{\partial L}{\partial H}} ($$

ii) $$\frac{\partial L}{\partial X} = \frac{\partial L}{\partial H} \frac{\partial H}{\partial X}$$

available

$$= \frac{\partial L}{\partial H} \cdot \left(\frac{\partial X}{\partial X} + \frac{\partial H(x,w)}{\partial F(x,w)} \frac{\partial F(x,w)}{\partial X} \right)$$

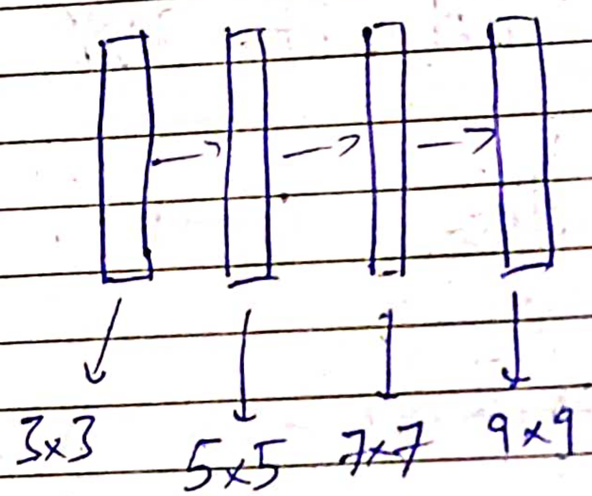
Shape according to Tensor calculus

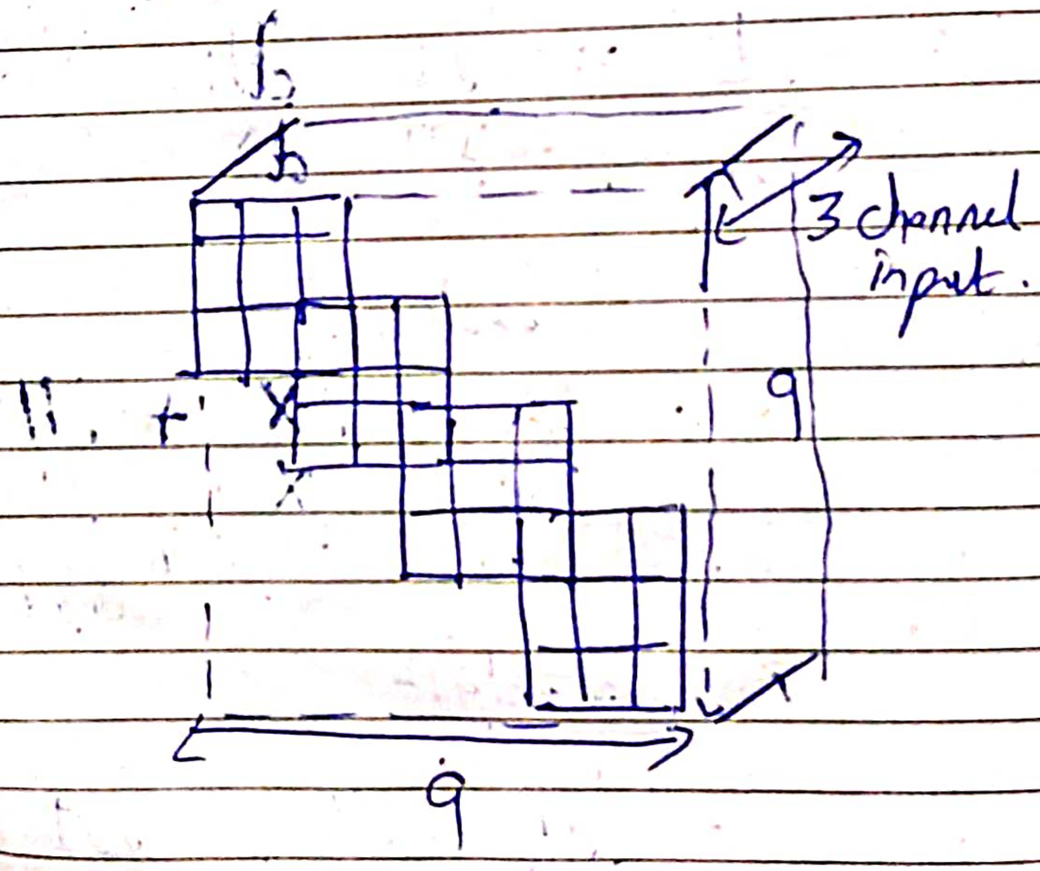
more precisely:-

$$= \frac{\partial L}{\partial H} \left(\frac{\partial X}{\partial X} + \frac{\partial H}{\partial F} \frac{\partial F}{\partial X} \right)$$

now we can update W & backprop to previous layers as well.

2) The receptive field of a neuron in the 4rth non-input layer .....





Page No.:
Date. / /

even though 3 channel input is considered, one pixel is a set of 3 values

$\therefore$ the support for such a neuron is

243 values or $\boxed{81 \text{ pixels}}$

3.) Increasing the # of hidden units will increase the capacity of the network (estimator)

(assuming non-linearities are present)

This increases chances of overfitting to a sample from the data distribution.

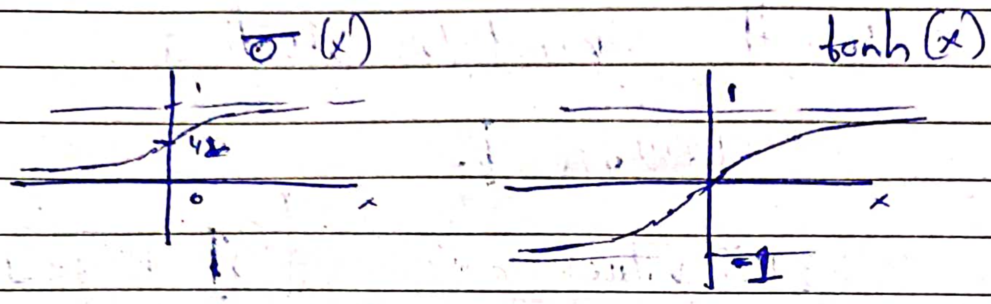
This will lead to greater variations in the estimator for same numbers of different samples from the data distribution.

i.e. The variance of the estimator increases.

& correspondingly, The bias will drop

Page No:
Date: / /

4) Yes she is correct; to direct ~~a~~ a linear map between a sigmoid & a tanh function ...



→ (for 0 mean)

we can shift the σ by $1/2$ downward & scale it by two → (for range -1 to 1)

corresponds to

i.e. $2\left(\sigma(x) - \frac{1}{2}\right) \Longleftrightarrow \tanh(y)$

some $y = P(x)$

verifying

$$2\left(\frac{1}{1+e^{-x}} - \frac{1}{2}\right) = \frac{2 - 1 - e^{-x}}{1+e^{-x}}$$

$$= \frac{1 - e^{-x}}{1+e^{-x}}$$

$$= \frac{\cancel{\tan}\ e^{-x/2}}{e^{-x/2}}\left(\frac{e^{x/2} - e^{-x/2}}{e^{x/2} + e^{-x/2}}\right)$$

$$= \left(\frac{e^{x/2} - e^{-x/2}}{e^{x/2} + e^{-x/2}}\right) = \tanh(x/2)$$

$$\therefore \boxed{\tanh(x) = 2\sigma(2x) - 1}$$

i.e., with a linear transformation of the parameters, one can emulate sigmoid by tanh & hence estimate a similar class of functions using the network.

5.) A constant error corresponds to a ~~certain surface on a~~ contour

for that $E(w) = \text{constant} \triangleq c$

$\therefore$ rearranging given error definition

$$\frac{1}{2}(w - w^*)^T H (w - w^*) \cong E(w) - E(w^*)$$
$$\cong c$$

as w^* is a fixed point.

This is the equation of the contour.

AKA

(representing)

now ; nudging w from w^* (center of contour surface)

by some linear combination of eigen vectors of H

we have

$$W = w^* \sum \alpha_i U_i$$

i.e. $w - w^* = \sum \alpha_i U_i$

substituting in original equation

$$\left(\sum_i \alpha_i v_i\right)^T H \left(\sum_j \alpha_j U_j\right) \simeq \text{constant}(c)$$

$$= \sum \alpha_i^2 U_i^T H v_i$$

$$+ \sum_{i \neq j} \alpha_i \alpha_j U_i^T H U_j \simeq \text{constant.}$$

$$= \sum \alpha_i^2 \lambda_i U_i^T U_i$$

$$+ \sum_{i \neq j} \alpha_i \alpha_j \lambda_j U_i^T U_j = \text{constant}$$

Page No.:
Date: / /

$$\cancel{= \sum \frac{v_i^T v_i}{\cancel{\quad}}}$$

$$= \sum \frac{\sum v_i^T v_i}{\left(\frac{1}{\alpha_i \sqrt{\lambda_i}}\right)^2} + \sum \alpha_i \alpha_j \lambda_j v_i^T v_j = \text{constant}$$

This is the equation of an n-dimensional ellipse with length of axis along the eigen vector v_i as $\frac{1}{\alpha_i \sqrt{\lambda_i}}$

i.e. it is inversely proportional to the square root of corresponding eigen value

The second term can be ignored in the analysis as it only shifts the center of the ellipse and doesn't affect its shape.

Page No.:
Date. / /

to observe the 2-d ellipses:

setting, without loss of generality, all except 2 eigenvalue α_i's to zero say that of p & q, we have

$$\frac{U_p^T U_p}{\left(\frac{1}{\sqrt{\alpha_p^2 \lambda_p}}\right)^2} + \frac{U_q^T U_q}{\left(\frac{1}{\sqrt{\alpha_q^2 \lambda_q}}\right)^2} + (\quad) = \text{const.}$$

center shifting terms

This the vector equation of the two dimensional ellipse

with major axis & minor axes as U_p & U_q depending on the relation b/w the magnitude of their eigen values

:- Shown.

6.) Suggested steps :-

1) retrieve pretained model from the Park in Washington.

2) Augment their data as it is less in amount such that the semantics of the photo - animal pair are retained; ~~as a~~ This aids in regularization as well as increases data available for training.

3) ~~If domain distribution between photos taken~~

3) ~~Freeze the initial layers of the model,~~ remove the final classification layer; append a custom fully connected classification layer with 200 ~~custom~~ classes specific to this Park.

Page No.:
Date / /

4.) Freeze initial layers
& Fine Tune latter layers :

initial v/s ~~final~~ latter ratio is a hyperparameter.

~~Alternativ~~

→ Use low learning rate
& no weight decay