NPTEL Week 11 Live Sessions

on Deep Learning (noc24_ee04)

A course offered by: Prof. Prabir Kumar Biswas, IIT Kharagpur

- **Quiz 10 Solution**
- **Practice Problems for week 11**

Week 8 Q9 → Softmax Leaky;

Triplet loss
one shot learning
Segmentation.

By

Arka Roy

NPTEL PMRF TA

Prime Minister's Research Fellow
Department of Electrical Engineering, IIT Patna
Web: https://sites.google.com/view/arka-roy/home

Powered by:

FREE ONLINE EDUCATION swayam शिक्षित भारत, उन्नत भारत

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR भारतीय प्रौद्योगिकी संस्थान खड़गपुर 1951 योगः कर्मसु कौशलम्

PMRF
Prime Minister's Research Fellows
Ministry of Education
Government of India
सत्यमेव जयते

भारतीय प्रौद्योगिकी संस्थान पटना INDIAN INSTITUTE OF TECHNOLOGY PATNA विद्यार्थी लभते विद्यान

1) Which of the following is False regarding batch-normalization??

a. Batch normalization prevents the problem of exploding — True

b. For an input **x**, Batch Normalization produces same output regardless of in training mode or inference/test mode. → False.

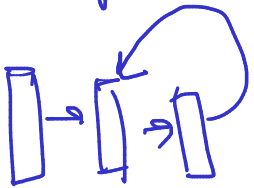
c. Batch normalization may prevent covariate shift

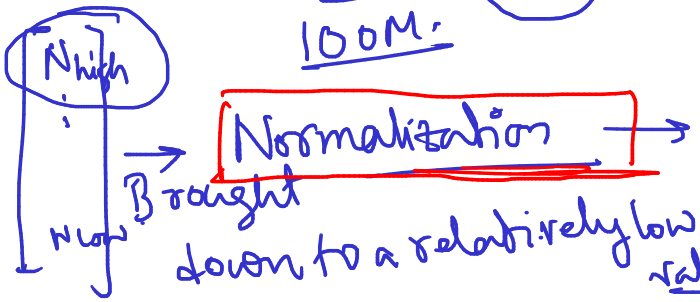
d. All of them

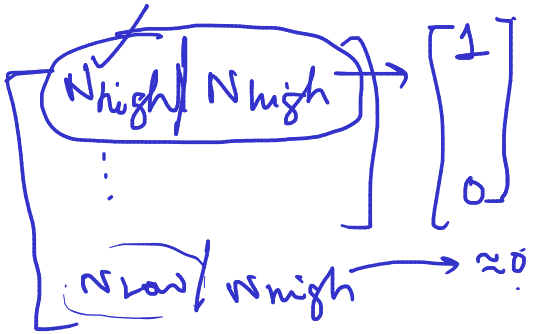
i) Vanishing Gradient ($\nabla_W L(W) \to 0$): $W_{n+1} \leftarrow W_n - \eta \nabla_W L(W)$ → 0

- Regularization.
- Drop out.
- Early stopping.

ii) Exploding gradient $\nabla_W L(W) \to$ high → inf / NaN







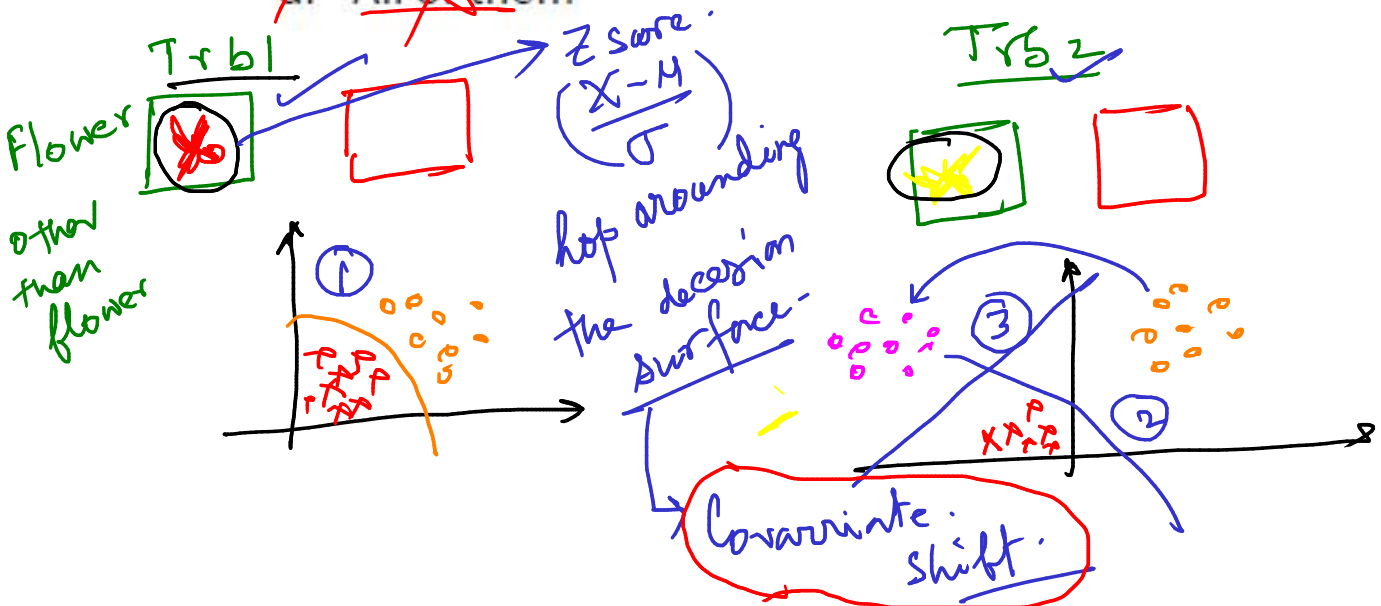
Which of the following is False regarding batch-normalization??

a. Batch normalization prevents the problem of exploding — True.

b. For an input **x**, Batch Normalization produces same output regardless of in training mode or inference/test mode.

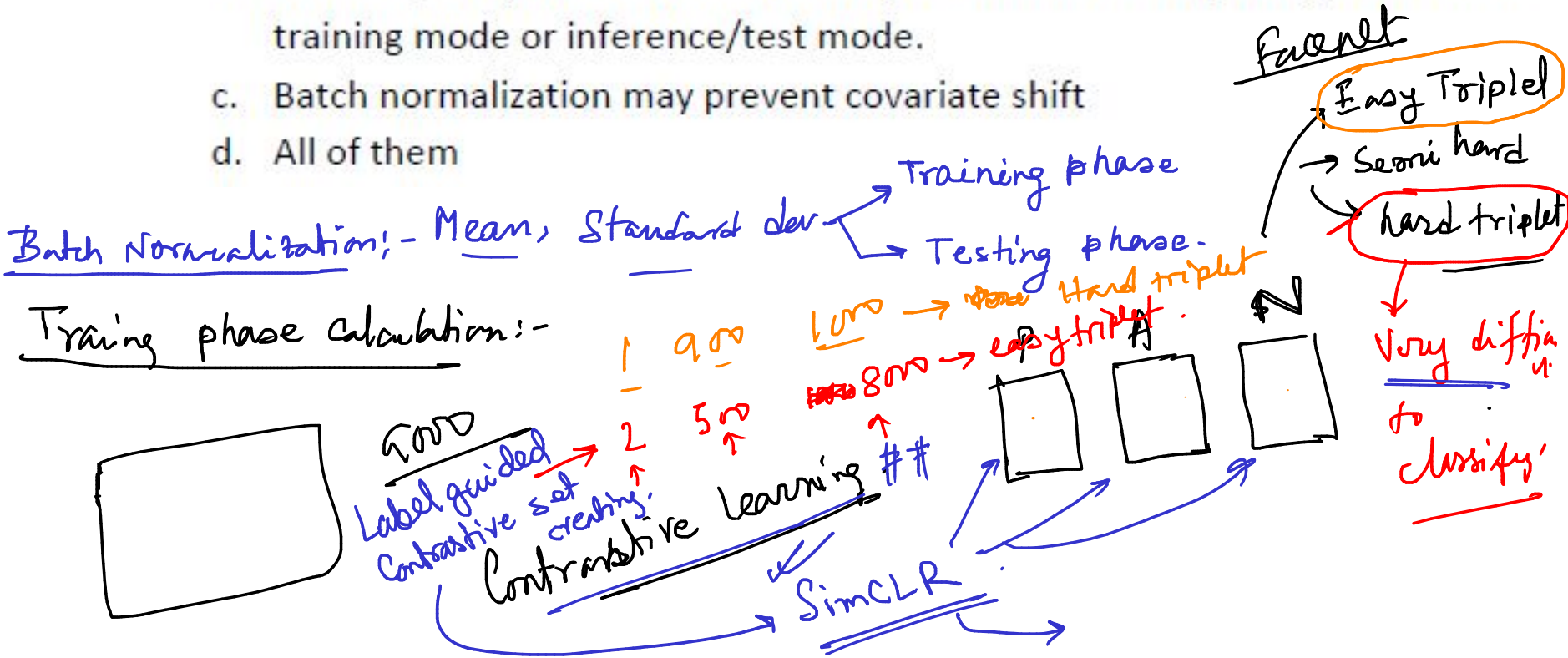
c. Batch normalization may prevent covariate shift — True

d. All of them



Which of the following is False regarding batch-normalization??

a. Batch normalization prevents the problem of exploding
b. For an input **x**, Batch Normalization produces same output regardless of in training mode or inference/test mode.
c. Batch normalization may prevent covariate shift
d. All of them



(7) A neural network has 3 neurons in a hidden layer. Activations of the neurons for three batches are $\begin{bmatrix}4\\4\\3\end{bmatrix}, \begin{bmatrix}0\\2\\5\end{bmatrix}, \begin{bmatrix}8\\9\\1\end{bmatrix}$ respectively. What will be the value of mean if we use batch normalization in this layer?

a. $\begin{bmatrix}4\\5\\3\end{bmatrix}$

b. $\begin{bmatrix}2\\2\\5\end{bmatrix}$

c. $\begin{bmatrix}1\\1\\1\end{bmatrix}$

d. $\begin{bmatrix}0\\0\\0\end{bmatrix}$

$\begin{bmatrix}4+0+8/3\\4+2+9/3\\3+5+1/3\end{bmatrix} = \begin{bmatrix}4\\5\\3\end{bmatrix}$.

$A = \begin{bmatrix}4\\4\\3\end{bmatrix}$ $\quad A^T = \begin{bmatrix}4 & 4 & 3\end{bmatrix}$. $\quad 1\times 3$.

Numpy :- $W^T X$.

$y = x^T . w$.

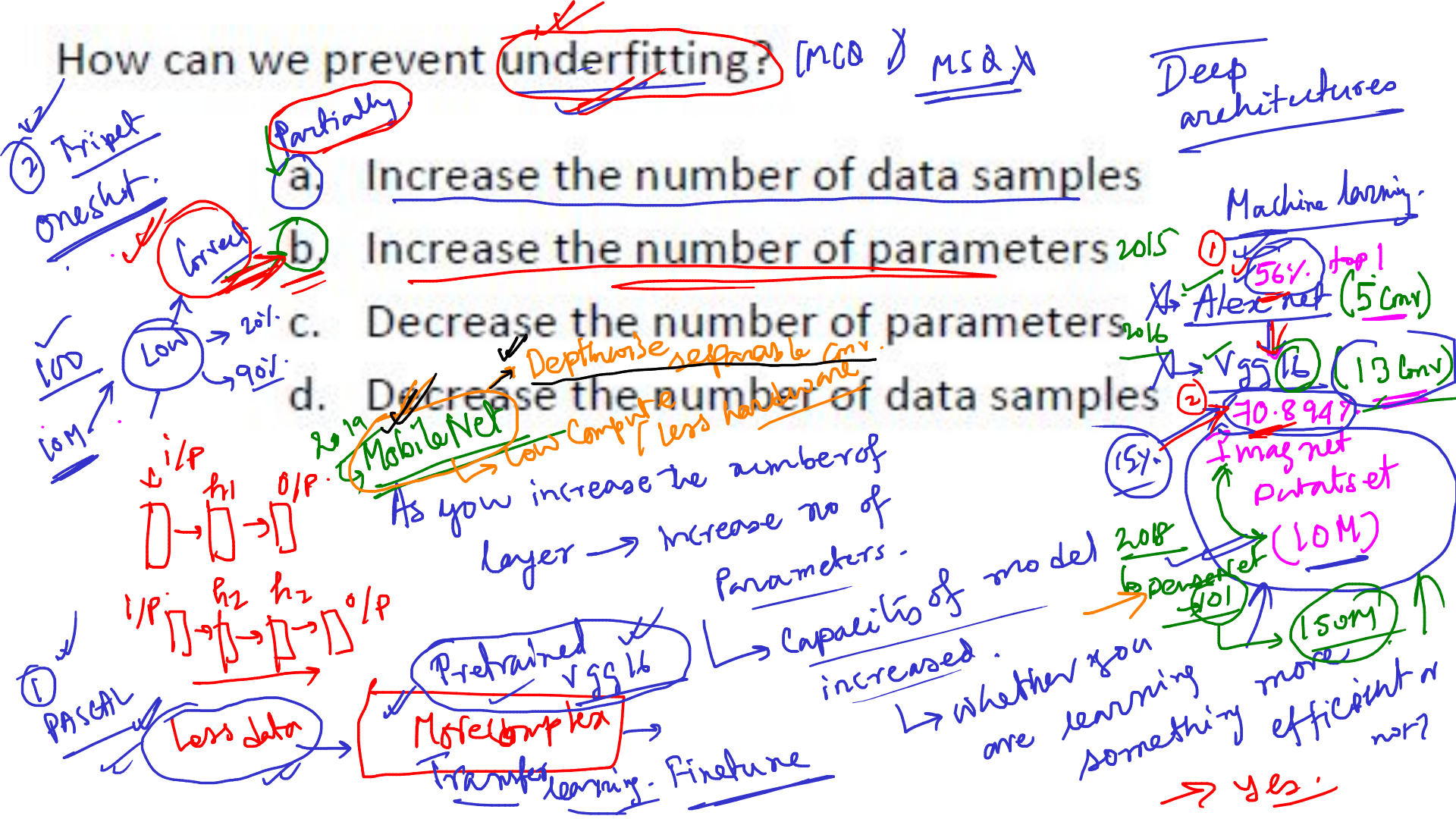
How can we prevent underfitting? [MCQ / MSQ]

a. Increase the number of data samples

b. Increase the number of parameters

c. Decrease the number of parameters

d. Decrease the number of data samples

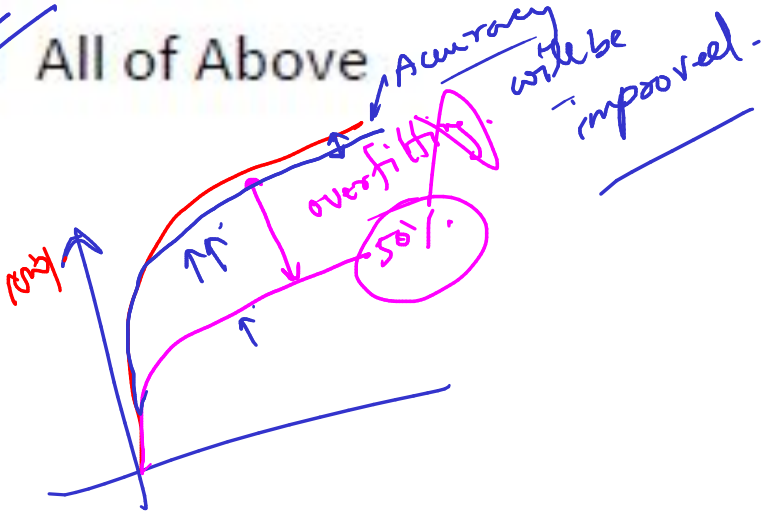


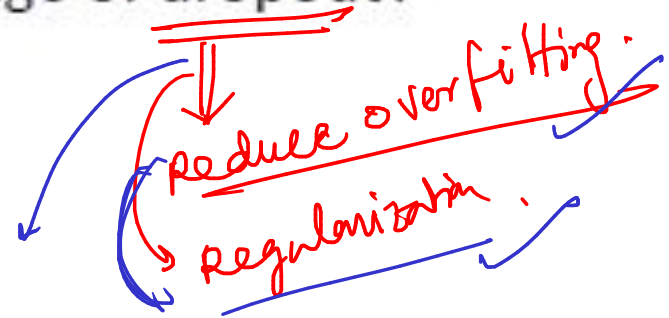
How do we generally calculate mean and variance during testing?

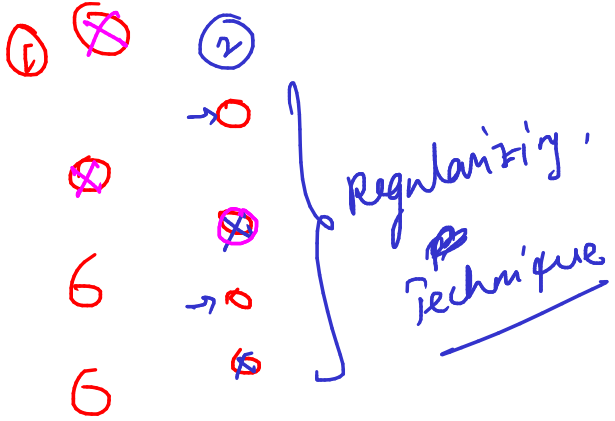
a. Batch normalization is not required during testing
b. Mean and variance based on test image
c. Estimated mean and variance statistics during training
d. None of the above

Which one of the following is an advantage of dropout?

a. Regularization
b. Prevent Overfitting
c. Improve Accuracy
d. All of Above







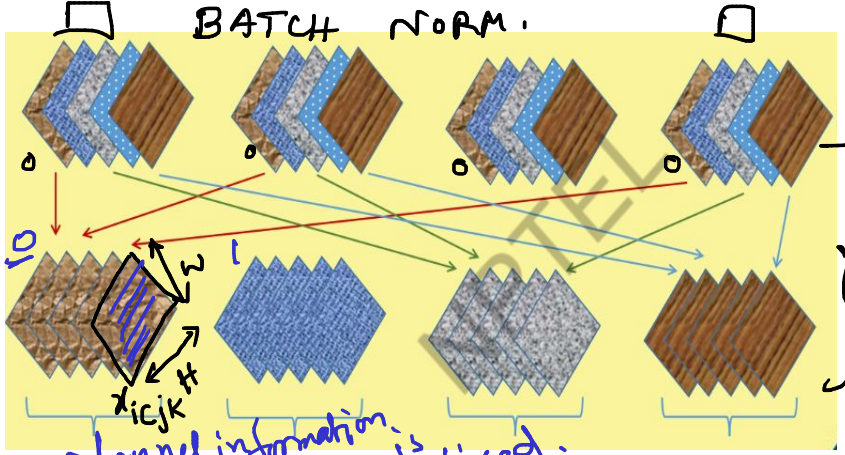
Which of the following is True regarding layer normalization and batch normalization?

a. Layer normalization normalizes over **spatial** and **channel** dimension whereas Batch normalization normalizes over **spatial** and **batch** dimensions

b. Layer normalization normalizes over **spatial** and **channel** dimension whereas Batch normalization normalizes over **batch**, **spatial** and **channel** dimension

c. Batch normalization normalizes over **spatial** and **channel** dimension whereas Layer normalization normalizes over **batch**, **spatial** and **channel** dimension

d. None of these



Spatial dimension (height width of Tensors)

Batch information.

Channel information is fixed.

$$x \in \mathbb{R}^{N \times C \times W \times H}$$

$$\mu_C = \frac{1}{NWH} \sum_{i=1}^{N} \sum_{j=1}^{W} \sum_{k=1}^{H} x_{iCjk}$$

Mean

no of batches

$$\sigma_C^2 = \frac{1}{NWH} \sum_{i=1}^{N} \sum_{j=1}^{W} \sum_{k=1}^{H} (x_{iCjk} - \mu_C)^2$$

$$\hat{x} = \frac{x - \mu_C}{\sqrt{\sigma_C^2 + \epsilon}}$$

Training $\neq$ Testing.

Training Mean and Varriance extraction process

Testing :→ Avg of all the μ and σ evaluated in the training phase

Exponentially decaying average based m

** Estimated statistics

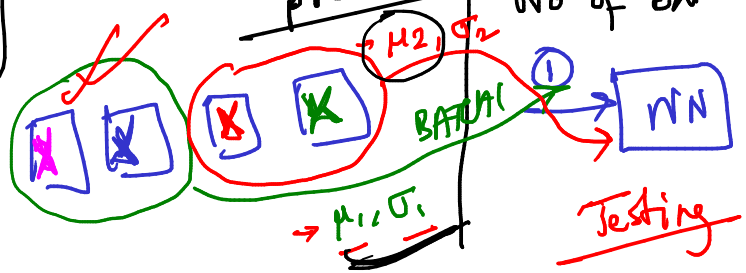
DATASET :- 10,000 sample.

9000 samples → Training

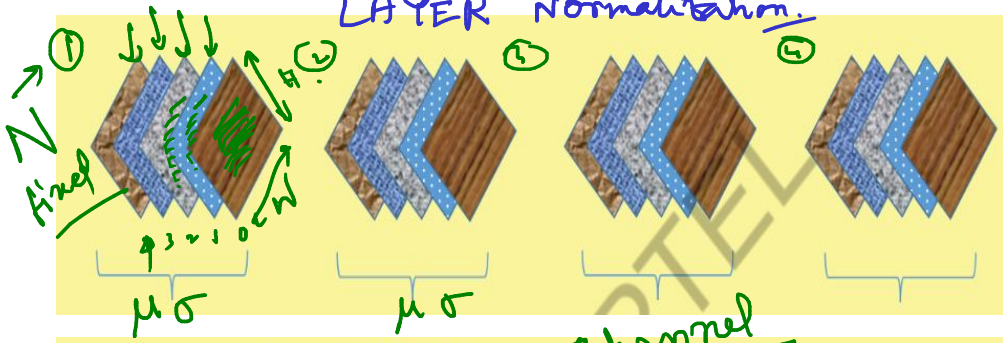
1000 " → Testing.

Batch size = 10.

No of batches = 900



LAYER Normalization.



$$x \in \mathbb{R}^{N \times C \times W \times H}$$

Channel dimension.

$$\mu_N = \frac{1}{CWH} \sum_{i=1}^{C} \sum_{j=1}^{W} \sum_{k=1}^{H} x_{Nijk}$$

spatial width / height.

$$\sigma_N^2 = \frac{1}{CWH} \sum_{i=1}^{C} \sum_{j=1}^{W} \sum_{k=1}^{H} (x_{Nijk} - \mu_N)^2$$

$$\hat{x} = \frac{x - \mu_N}{\sqrt{\sigma_N^2 + \epsilon}}$$

. Which one of the following regularization methods induces sparsity among the trained weights?

a. L_1 regularizer
b. L_2 regularizer
c. Both L_1 & L_2
d. None of the above

Lasso. Absolute value.
Ridge

$\nabla_w L(w) = \frac{\partial L(w)}{\partial w}$

Loss(w) = Data Loss + Regularizer.

$\Rightarrow$ L(w) = Data Loss + L_1 regularizer of weight
= g(w) + L_1 regularizer.

$w(n+1) \leftarrow w(n) - \eta \nabla_w L(w)$.

$|-2| = 2$

$|2| = 2$.

$|-1| = 1$.

$|1| = 1$

$|w| = \begin{cases} \\ \\ \end{cases}$

$L(w) = g(w) + |w|$

$L(w) = g(w) + \lambda \sum_{i=1}^{n} |w_i|$

strength of regularizer

$\nabla_w L(w) = g'(w) + \lambda \, sgn(w)$

Data loss differentiation with respect to w.

$\frac{d|w|}{dw}$

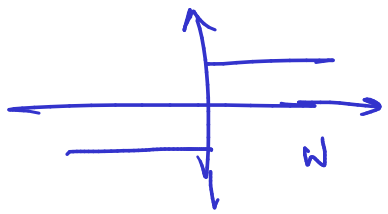
sgn(x)

$\frac{dy}{dx} = \frac{y_2 - y_1}{x_2 - x_1} = \frac{1-0}{-1-0} = -1$

$\frac{dy}{dx} = 1 \cdot \frac{y_2 - y_1}{x_2 - x_1} = \frac{1-0}{1-0} = 1$

$w_{n+1} \leftarrow w_n - \nabla_w L(w)$

$w_{n+1} \leftarrow w_n - [g'(w) + \lambda \operatorname{sgn}(w)]$.



$\|w\|_2^2$ → $\sigma \rightarrow \sqrt{a^2}$

$w > 0$,

$w_{n+1} \leftarrow w_n - g'(w) - \lambda$ → Reducing / shrinking → iteration 1, 2, 3 ...

$w_1^2 + w_2^2 + \cdots$

$L(w) = g(w) + L^2 \text{reg}(w)$ → w_i^2

$L(w) = g(w) + \lambda \sum_{i=1}^{N} \|w_i\|^2$ → $\lambda \cdot w_i^2$

$\frac{\partial L(w_i)}{\partial w} = g'(w_i) + 2\lambda w_i$

$L^2 = \sqrt{a^2}$ → $\sqrt{w_1^2} \rightarrow w_i$

Norm

w_1, w_2, w_3

$(-2)^2 = 4$ $\quad |-2|^2 = 4$.

L^1 reg. → 0 → sparsity

$\sqrt{w_1^2 + w_2^2 + \cdots}$

$w_{n+1} \leftarrow w_n - \nabla_w L(A)$

$w_{n+1} \leftarrow w_n - g'(w) - 2\lambda w_i$ (- Vt)

→ either they will be very small in nature

→ or they will be neglected.

Reduce the value of weight.

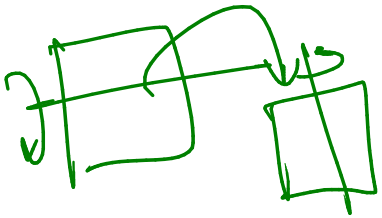
Which among the following is NOT a data augmentation technique?

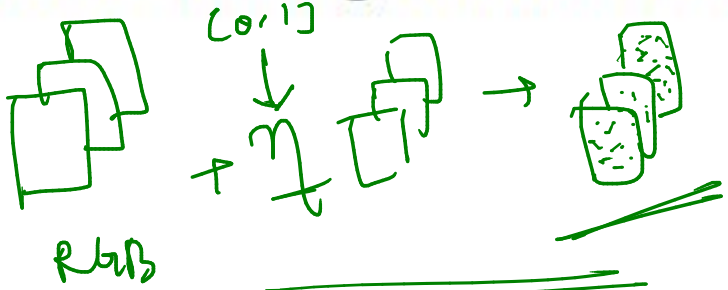
a. Random horizontal and vertical flip of image

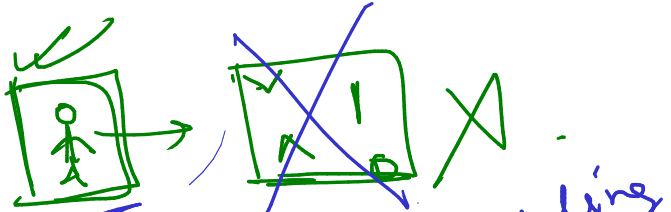
b. Random shuffle all the pixels of an image

c. Random color jittering

d. All the above are data augmentation techniques







output is totally tampered version of input

Which of the following is true about model capacity (where model capacity means the ability of neural network to approximate complex functions)?

a. As number of hidden layers increase, model capacity increases
b. As dropout ratio increases, model capacity increases
c. As learning rate increases, model capacity increases
d. None of these

Two variant training schedule samples its minibatches in the following manner

Training Schedule 1

Mini batch 1=[Image1, Image2, Image3]
Mini batch 2=[Image4, Image5, Image6]

Training Schedule 2

Mini batch 1=[Image1, Image4, Image3]
Mini batch 2=[Image2, Image 5, Image6]

The output activations of each corresponding image is compared across Training schedule 1 and Training schedule 2 for a CNN with batch norm layers. Choose the correct statement

a. Activation outputs of corresponding image will be same across Training schedule 1 and Training schedule 2
b. Activation outputs of corresponding image will be different across Training schedule 1 and Training schedule 2
c. Some activations outputs of corresponding images will be same but some will be different
d. None of these.

Suppose you have a $1D$ signal $x = [5,4,3,2,1]$ and a filter $f = [1,2,3,4]$, and you perform stride 2 transpose convolution on the signal x by the filter f to get the signal y. What will be the signal y if we don't perform cropping?

a. $y = [1,2,5,8,9,14,13,20,19,26,3,4]$

b. $y = [5,10,19,28,15,22,11,16,7,10,3,4]$

c. $y = [1,2,5,8,9,14,13,20,17,26,15,20]$

d. $y = [0,0,5,8,9,14,13,20,19,26,0,0]$

even no of kernel / odd no

$[5 \quad 4 \quad 3 \quad 2 \quad 1]$

y_0 1 → 5

y_1 2 → 10

y_2 3 → 15+4=19

y_3 4

y_4

y_5

y_6

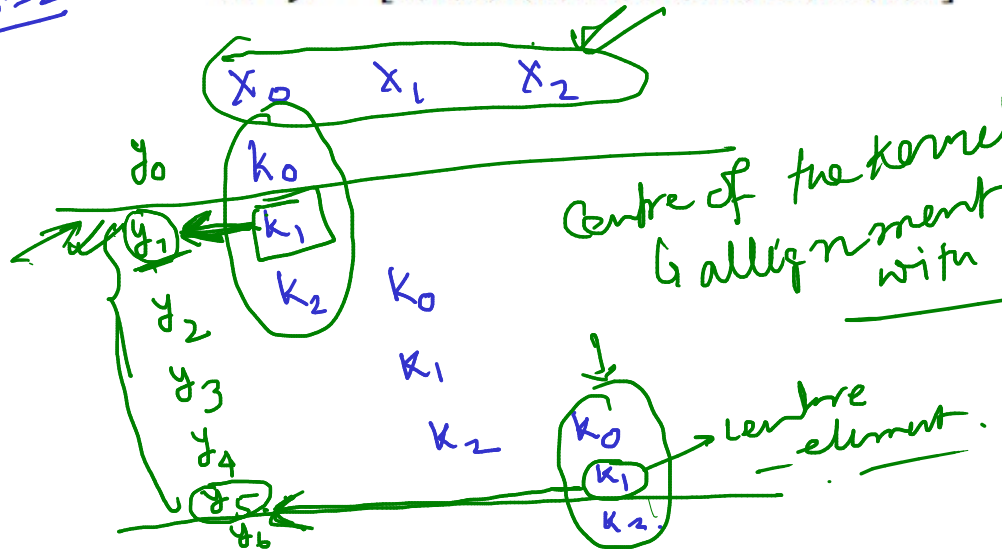
y_7

y_8

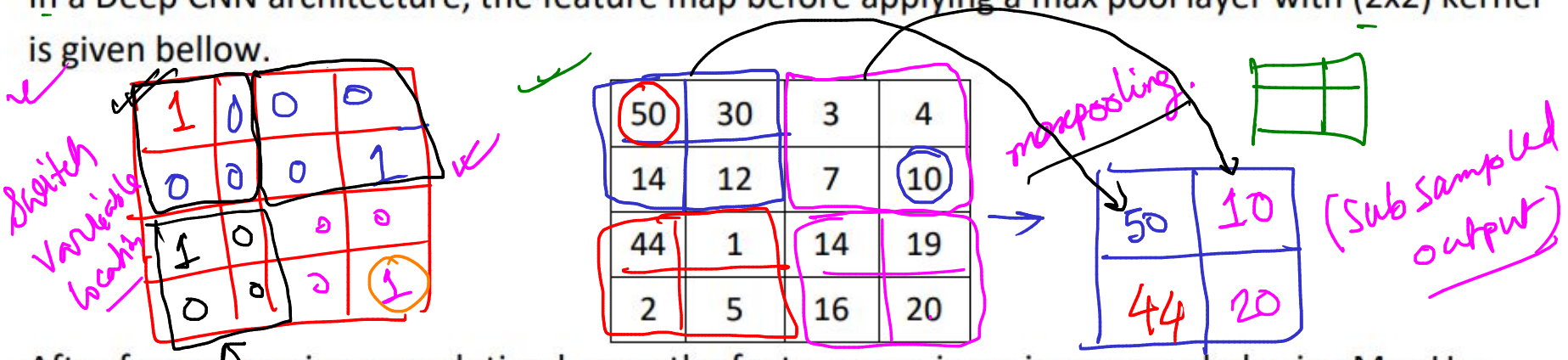
y_9

y_{10} y_{11}

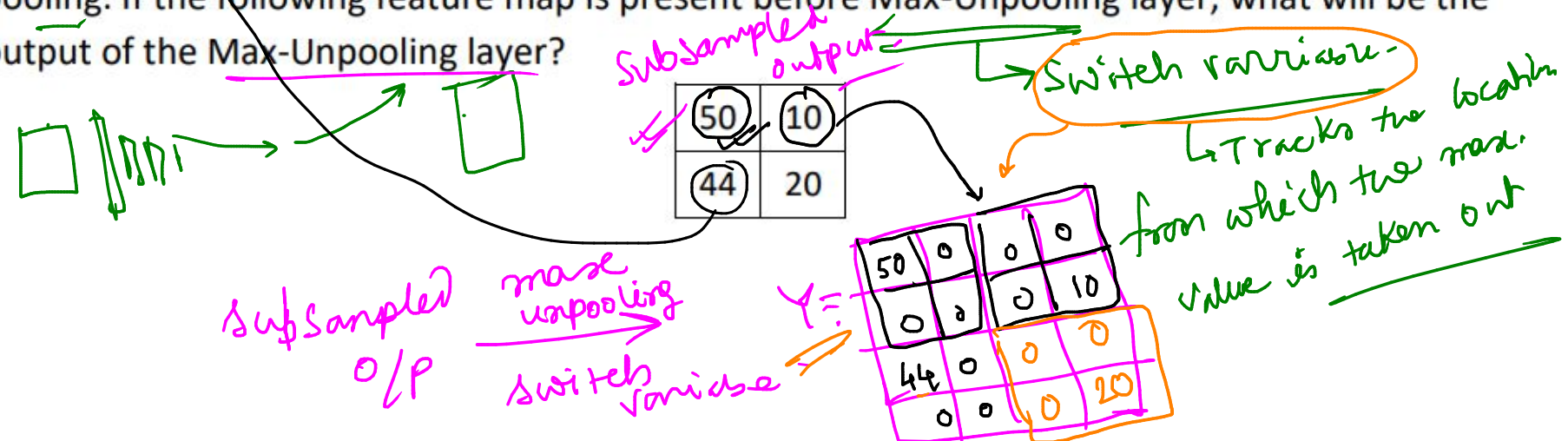
S=2



In a Deep CNN architecture, the feature map before applying a max pool layer with (2x2) kernel is given bellow.



After few successive convolution layers, the feature map is again up-sampled using Max Un-pooling. If the following feature map is present before Max-Unpooling layer, what will be the output of the Max-Unpooling layer?



a.

50	0	0	0
0	0	0	10
44	0	0	0
0	0	0	20

b.

50	0	10	0
0	0	0	0
44	0	20	0
0	0	0	0

c.

50	0	0	10
0	0	0	0
44	0	0	20
0	0	0	0

d.

0	0	0	0
0	50	0	10
0	0	0	0
0	44	0	20

0.10	0.47	0.88	0.25
0.65	0.11	0.66	1
0.91	0.00	0.99	0.29
0.78	1	0.40	0.99

X

0.01	0.44	0.48	0.18
0.60	0.09	0.97	0.99
0.87	0.02	0.98	0.20
0.77	0.99	0.40	0.99

A
Ground truth mask

B
Prediction mask

Dice Coefficient (A, B) =

$$\frac{\sum_{i=1}^{16} A \times B}{\sum_{i=1}^{8} A + \sum_{i=1}^{8} B}$$