# NPTEL Week 10 Live Sessions

## on Deep Learning (noc24_ee04)

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

Powered by:

FREE ONLINE EDUCATION swayam

शिक्षित भारत, उन्नत भारत

- **Quiz 9 Solution**

→ optimizers → Adam
↳ Adgrad
↳ RMS Prop.
↳ Momentum
↳ NAG.

Week 10 → Min-Max
→ Z-score.
→ Batch Norm.
↳ layer, Instance, group.

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](https://sites.google.com/view/arka-roy/home)**


PMRF

Prime Minister's Research Fellows

Ministry of Education

Government of India

1. Comment on the learning rate of Adagrad. Choose the correct option.

Adaptive gradient optimizer.

$grad_v \gg grad_h$.

a. Learning rate is adaptive

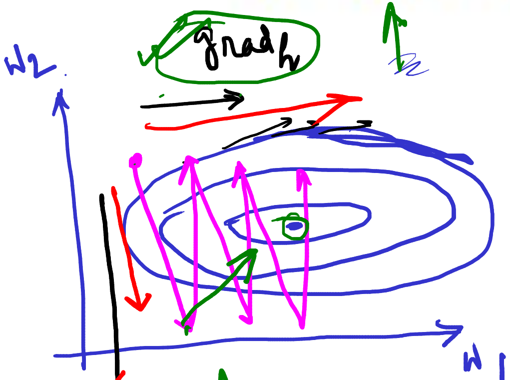
b. Learning rate increases for each time step

c. Learning rate remains the same for each update

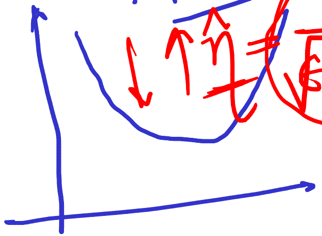
d. None of the above

$L(w)$

$L(w_1, w_2)$








$\downarrow r_t \uparrow$

$\uparrow \frac{1}{r_t} \downarrow$

$\uparrow \eta = \frac{\eta}{\sqrt{\epsilon + r_t I}}$

Multiply element wise

$g_t \downarrow \quad r_t \downarrow \quad \frac{1}{r_t}\uparrow$

$w_{n+1} = w_n - \eta \nabla_w L(w) + \gamma v_{t-1}$

Momentum optimizer.

$g_t = \frac{1}{N} \sum_{x \in N_{minibatch}} \nabla L(w)$.

Avg.

$r_t = \sum g_t \circ g_t$

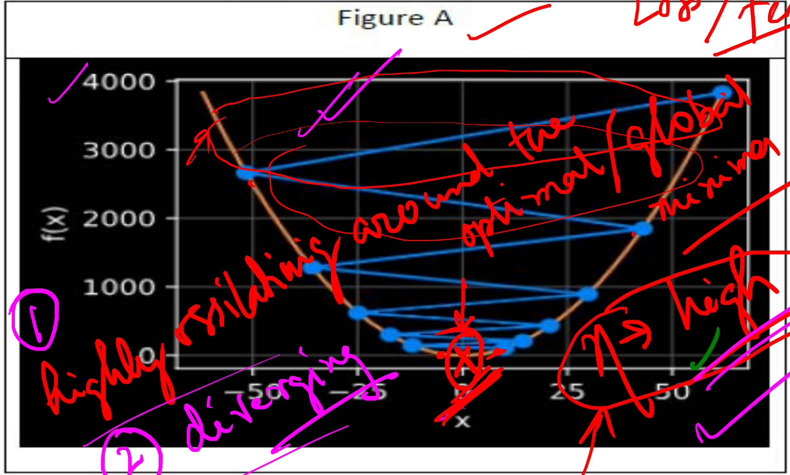
$r_t = g_t^2$

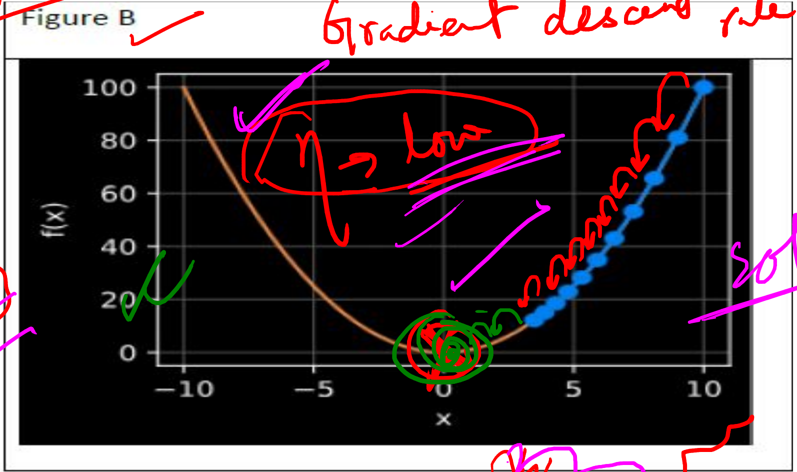
$w_{n+1} = w_n - \frac{\eta}{\sqrt{\epsilon + r_t I}} g_t$

For the following figure A and figure B of loss landscape, choose correct statement

Loss/feature

Gradient descent rate/optimization Technique




softmax

a. Figure A has very small learning rate, Figure B has optimal learning rate

b. Figure A has optimal learning rate, Figure B has very small learning rate

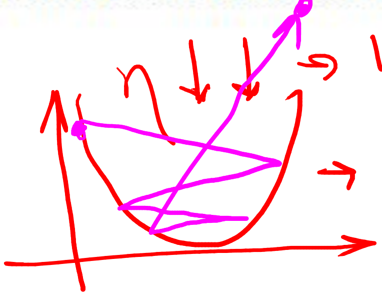
high

c. Figure A and Figure B have different Loss function

d. None of Above

$W_{n+1} = W_n - \eta \nabla_w L(w)$

→ learning rate



→ learning process is very slow.

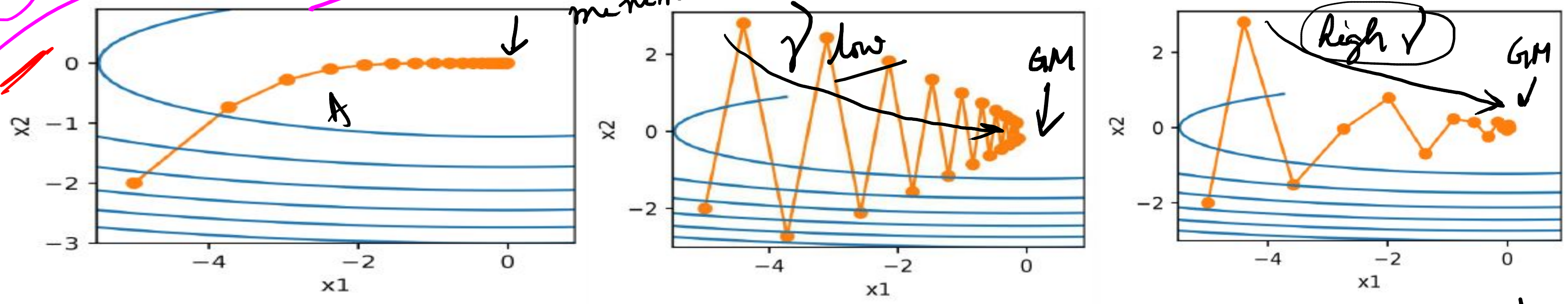
→ optimization process will take huge time to converge.

3 Error Contour plot

Global minima

Momentum optimize.

Momentum opt




a. Figure A is SGD momentum optimizer with high momentum, Figure B is RMSProp or AdaGrad and Figure C is SGD momentum optimizer with low Momentum.

b. Figure A is RMSProp or AdaGrad, Figure B is SGD Momentum with low Momentum and Figure C is SGD momentum with high momentum.

c. Figure A is SGD Momentum optimizer with low momentum, Figure B is RMSProp or AdaGrad and Figure C is SGD Momentum optimizer with high Momentum.

d. None of the above

$\eta \uparrow\uparrow$ → ① Adaptive nature of learning rate

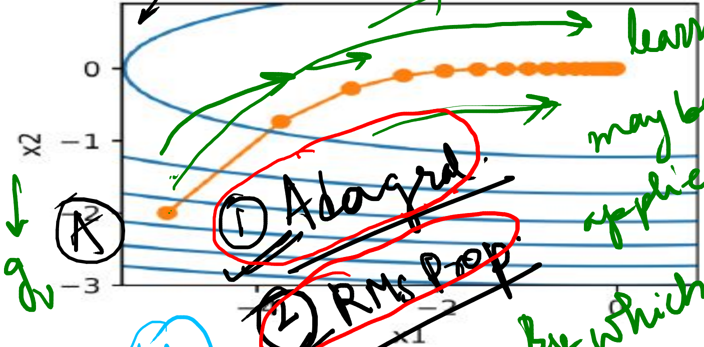
may be applied

① Adagrad

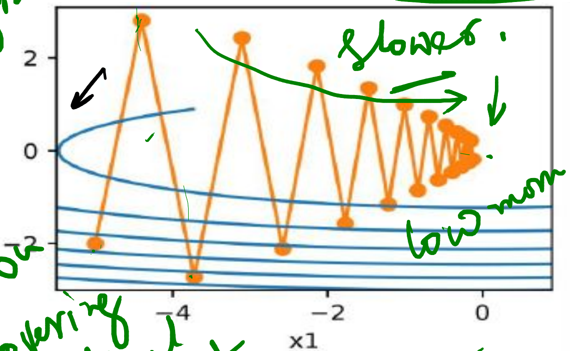
② RMSProp

A

$g_v$

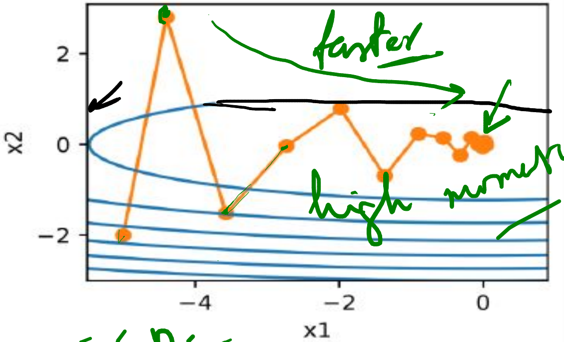



② Momentum optimizer.




slower.

low mom




faster

high moment

By which you are not hovering in the vertical direct unlike the

SGD

$W_1$

momentum SGD

$w_2$

Momentum.

$g_v$

$\eta, v$

MSGD → Vanilla

Basic Version of any algorithm

→ Momentum opt or SGD:-

$$w_{n+1} = w_n - \eta \nabla_w L(w) + \gamma V_{n-1}$$

changes $\eta$ based on the grad. Scales the gradient

$S_5$ Constant

$S_4 \to S_3$ $S_4$ $S_2$ $S_1$

$g_v >> g_h$

Exponential gradient Avg decay

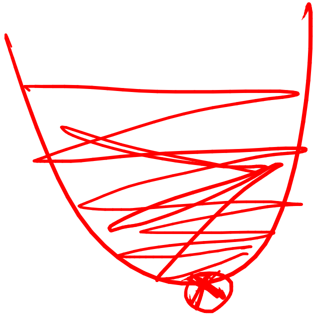
$$r_t = \beta S_t + (1-\beta) S_{t-1} \quad g_t, r_t$$

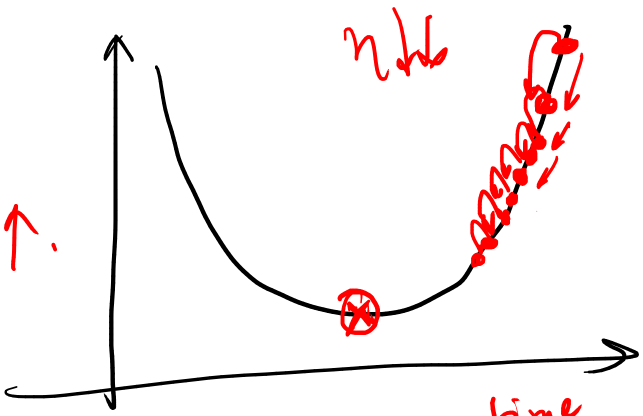
$\beta = 0.9$, $0.1$

$$W_{n+1} = W_n - \frac{\eta}{\sqrt{\epsilon + r_t}} g_t$$

What can be a possible consequence of choosing a very small learning rate? Choose the correct option.

a. Slow convergence
b. Overshooting minima → $\eta \uparrow\uparrow$.
c. Oscillations around the minima → $\eta \uparrow\uparrow$.
d. All of the above






① High settling time
② Slower convergence.

Two version of SGD are implemented as follows:

SGD1: SGD1 samples data points in same order for every epoch while constructing minibatch

SGD2: SGD2 samples data samples in random order for every epoch to construct minibatch

stochastic → Random in nature

Select the correct statement

a. SGD1 is faster than SGD2 and robust to local minima entrapment

b. SGD2 is faster than SGD1 and robust to local minima entrapment

c. SGD1 and SGD2 have same convergence characteristics

d. None of above

model.compile(

Shuffle = True

Memorize the features.

Generalize the understanding.

RMSProp resolves the limitation of which optimizer?

a. Adagrad (exponential decaying Avg. grad)

b. Momentum

c. Solves problem of option b but not a

d. Neither a nor b

$$V_t = \beta S_t + (1-\beta) S_{t-1}$$

Which of the following is a possible advantage of momentum optimizer over of mini-batch gradient descent?

Vanilla

a. Mini-batch gradient descent performs better than momentum optimizer when the surface of the loss function has a much more elongated curvature along X-axis than along Y-axis

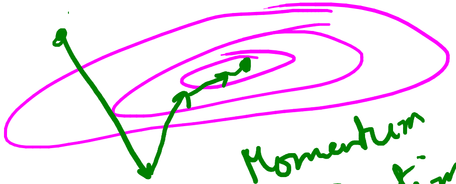
b. Mini-batch gradient descent always performs better than momentum optimizer

c. Mini-batch gradient descent will always overshoot the optimum point even with a lower learning rate value → only for high $\eta$ (overshoot will be there)

$V_{t-1}$ and grad term

d. Mini-batch gradient might oscillate in its path towards convergence which can reduced by momentum optimizer

minibatch grad



Momentum optimizer

The following is the equation of update vector for momentum optimizer. Which of the following is true for $\gamma$?

Momentum factor / term

$$V_t = \gamma V_{t-1} + \eta \nabla_\theta J(\theta)$$

gradient associated with learning rate $\eta$.

a. $\gamma$ is the momentum term which indicates how much acceleration you want

b. $\gamma$ is the step size

c. $\gamma$ is the first order moment

d. $\gamma$ is the second order moment

Velocity by which you have landed to $w_t$ from $w_{t-1}$

Why it is at all required to choose different learning rates for different weights?

$\eta \downarrow\downarrow$

a. To avoid the problem of diminishing learning rate (To generalise

b. To avoid overshooting the optimum point

c. To reduce vertical oscillations while navigating the optimum poin

d. This would aid to reach the optimum point faster

It ensures the convergence of the ~~optimal~~ problem in a faster manner.

$W_{n+1} = W_n - \eta \nabla L$

stopping of the gradient rule

Let $J(\theta)$ be the cost function. Let the gradient descent update rule for $\theta_i$ be,

$$\theta_{i+1} = \theta_i + \nabla\theta_i$$

What is the correct expression of $\nabla\theta_i$, $\alpha$ is the learning rate.

a. $-\alpha \frac{dJ(\theta_i)}{d\,\theta_i}$

b. $\alpha \frac{dJ(\theta_i)}{d\,\theta_i}$

c. $-\frac{dJ(\theta_i)}{d\theta_{i+1}}$

d. $\frac{dJ(\theta_i)}{d\theta_i}$

$$\theta_{i+1} = \theta_i - \underbrace{\alpha \cdot \frac{\partial J(\theta_i)}{\partial \theta_i}}$$

A neural network has 3 neurons in a hidden layer. Activations of the neurons for three batches are $\begin{bmatrix}1\\2\\3\end{bmatrix}, \begin{bmatrix}0\\2\\5\end{bmatrix}, \begin{bmatrix}6\\9\\2\end{bmatrix}$ respectively. What will be the value of mean if we use batch normalization in this layer?

MLP

a. $\begin{bmatrix}2.33\\4.33\\3.33\end{bmatrix}$ $a \to a.$

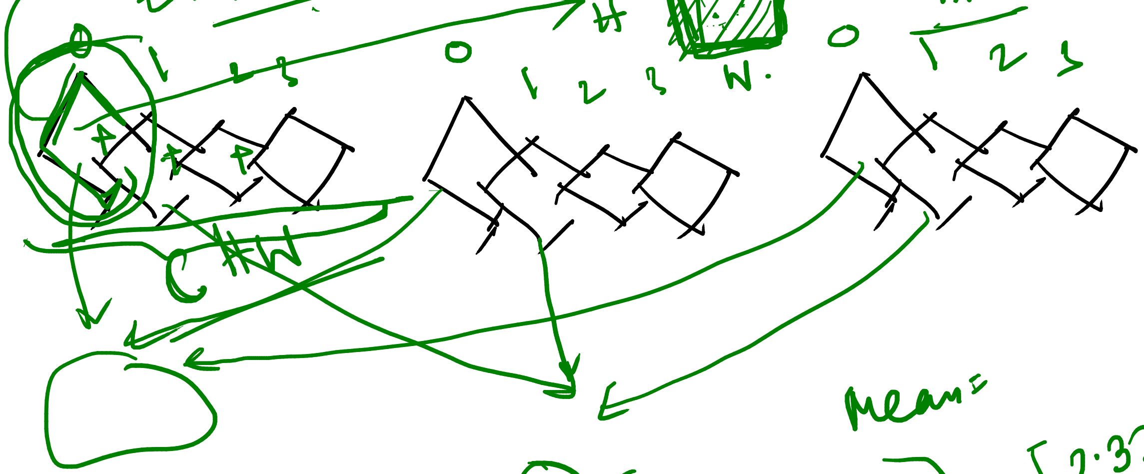
b. $\begin{bmatrix}2.00\\2.33\\5.66\end{bmatrix}$

c. $\begin{bmatrix}1.00\\1.00\\1.00\end{bmatrix}$

d. $\begin{bmatrix}0.00\\0.00\\0.00\end{bmatrix}$

Instance :-> $\mu, \sigma$

$\mu = \frac{\Sigma \text{All value}}{HW}$



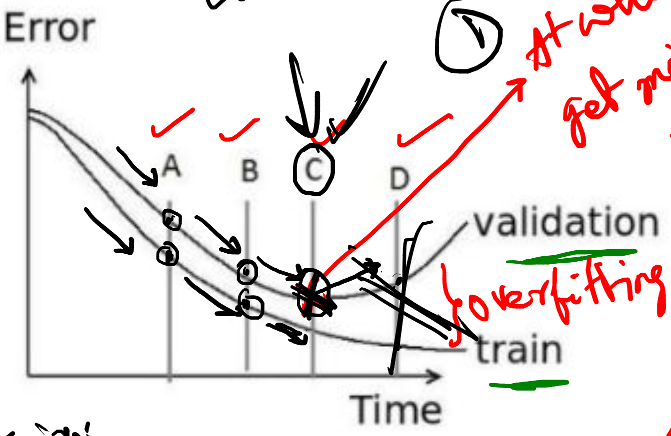

$\mu, \sigma$.

Mean =

$$\begin{bmatrix}\mu_1\\\mu_2\\\mu_3\end{bmatrix} = \begin{bmatrix}\frac{1+0+6}{3}\\\frac{2+2+9}{3}\\\frac{3+5+2}{3}\end{bmatrix} = \begin{bmatrix}2.33\\4.33\\3.33\end{bmatrix}$$

2) While training a neural network for image recognition task, we plot the graph of training error and validation error. Which is the best for early stopping?

Early stopping
↳ Premature stopping of updation of weights to alliviate the issue of overfitting.

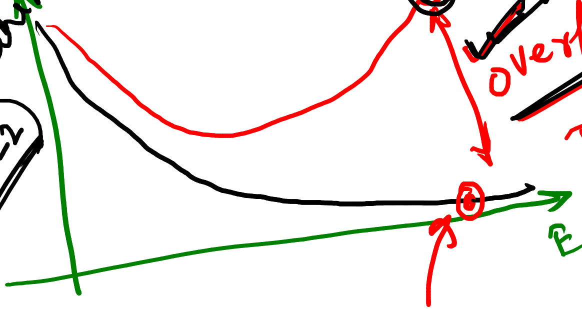



① At which you get minimum validation error
② Beyond this you get overfitting

overfitting has started

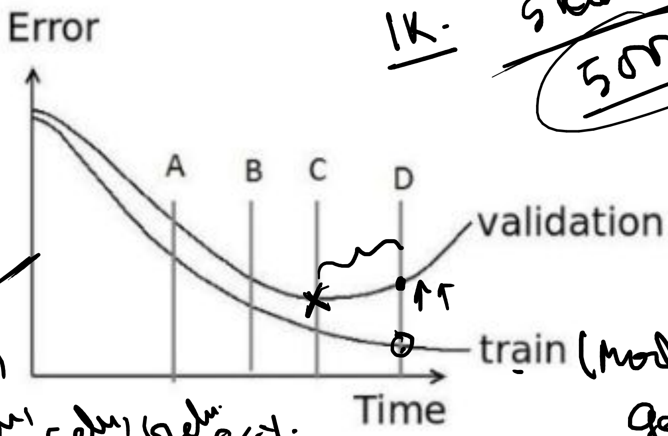
a. A
b. B
c. C
d. D

Overfitting
↳ Early stopping
↳ Dropout
↳ Regularization Technique (L1, L2)




Validation ↑↑.
overfitting → Memorization of DL.
Training loss ↓↓

While training a neural network for image recognition task, we plot the graph of training error and validation error. Which is the best for early stopping?




a. A

b. B

c. C

d. D

Data

Reg.

DROP

Batch

90%.

Activation

Softmax, Relu, elu, selu, gelu

95%.

(Test error) ↑

(Model has lost generalization capability).

Imagenet challenge

Alex, Vgg, Resnet

Dense

Data loss + Regularization loss

CVPR, ICLR.

ConvNet

1K.

5 class

500

VGG16

VGG19

5%. NLP

Transformer

chatGPT

Which among the following is NOT a data augmentation technique?

Image

d

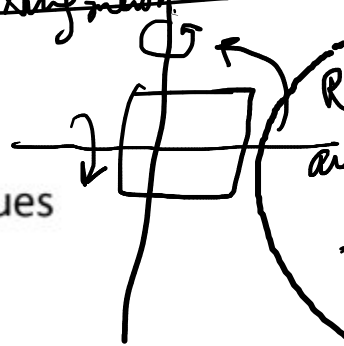
a. Random horizontal and vertical flip of image

b. Random shuffle all the pixels of an image

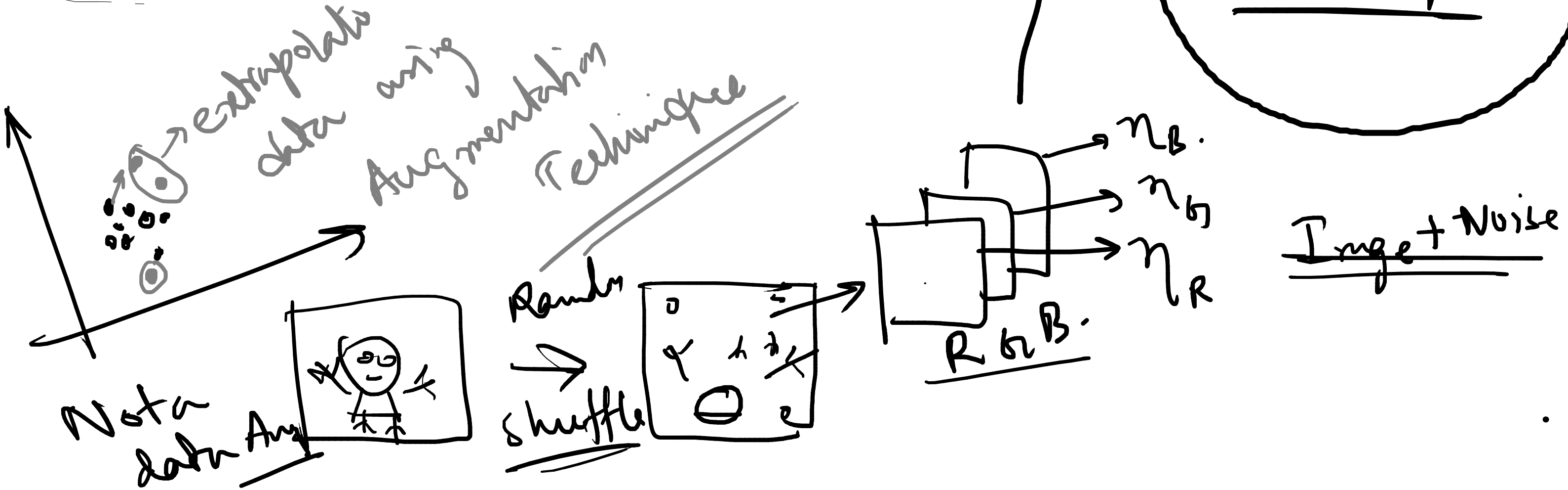
c. Random color jittering

d. All the above are data augmentation techniques

Augmentation

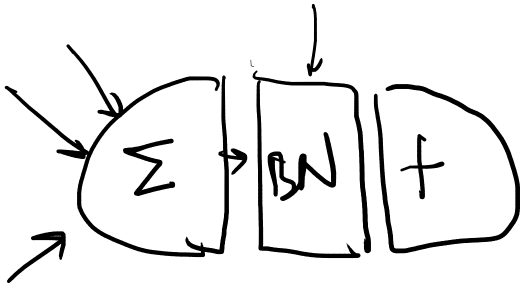


Rotation is an augmentation Technique



Batch Normalization is helpful because

a. It normalizes all the input before sending it to the next layer
b. ~~It returns back the normalized mean and standard deviation of weights~~
c. ~~It is a very efficient back-propagation technique~~
d. None of these



A Batch Norm layer accepts batch of 128D vector. How many parameters of Batch norm get trained via backpropagation during the course of training

a. 256
b. 512
c. 128
d. 1024

BN

BN $\gamma$, $\beta$.

$1 \rightarrow 2$

$128D \rightarrow 2 \times 128$

$= 256$

Which of the following is a regularization method?

a. Data augmentation
b. Dropout
c. Weight decay
d. All of the above

Two variant training schedulesamples its minibatches in the following manner

Training Schedule 1

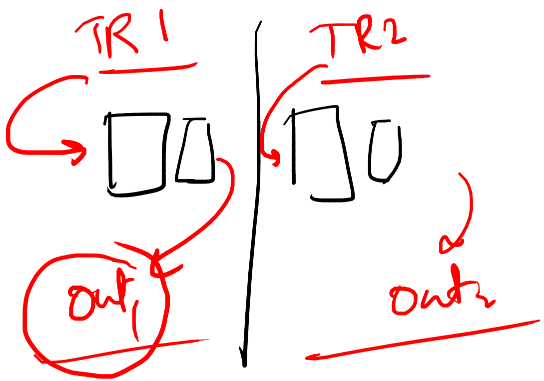
Mini batch 1=[Image1, Image2, Image3]

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.

Thank You