$(3 \times 3 \times 3 \times 16) + 16 = 448$

$\underbrace{\gamma + \beta + EMA_{\mu} + EMA_{\sigma}}_{4 \times 16 \text{ channels}} \rightarrow 64$

$64 \times 64 \times 16$

$\boxed{3 \times 3 \times 16} \times 32 = 4608 + \text{bias}$

in - $3 \times 3 \times 3$ RGB

CNN-Reduce Overfitting

$\underline{IP}$ $128 \times 128 \times 3$

$\downarrow \underline{3 \times 3 \times 3} \times 16$

$128 \times 128 \times 16$

$\downarrow$

$IP_{L1} \rightarrow 64 \times 64 \times 16 \Leftarrow FM$

$\downarrow \underbrace{3 \times 3 \times 16}_{} \times 32$

$4608 + b = 4640$

# of channels

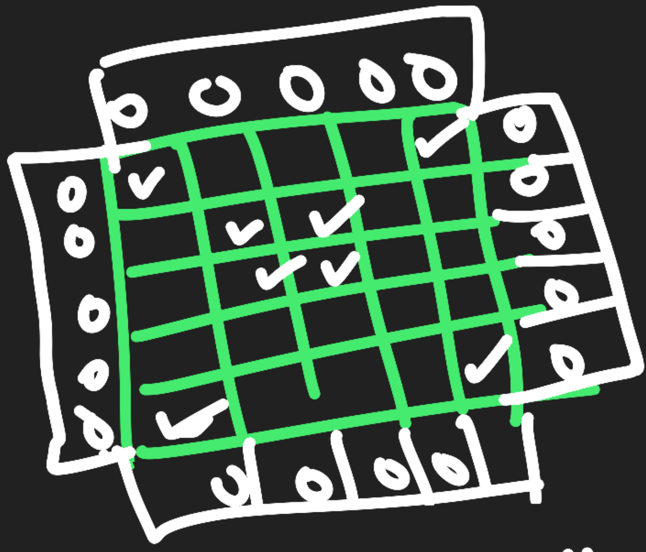
32×4

γ β EMA_{μ} EMA_{σ}

Revision

"Same Padding".

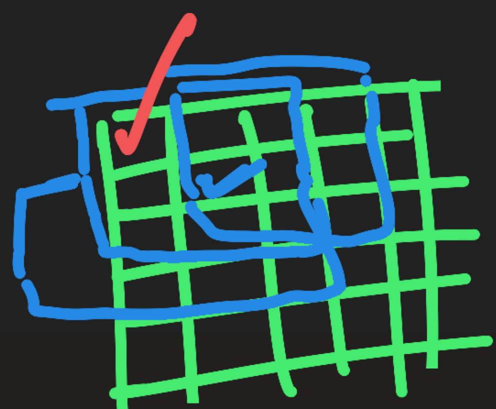
5×5



→ FM 5X5

"valid padding"

no padding



FM 2X3X1

→ Convolution ✓

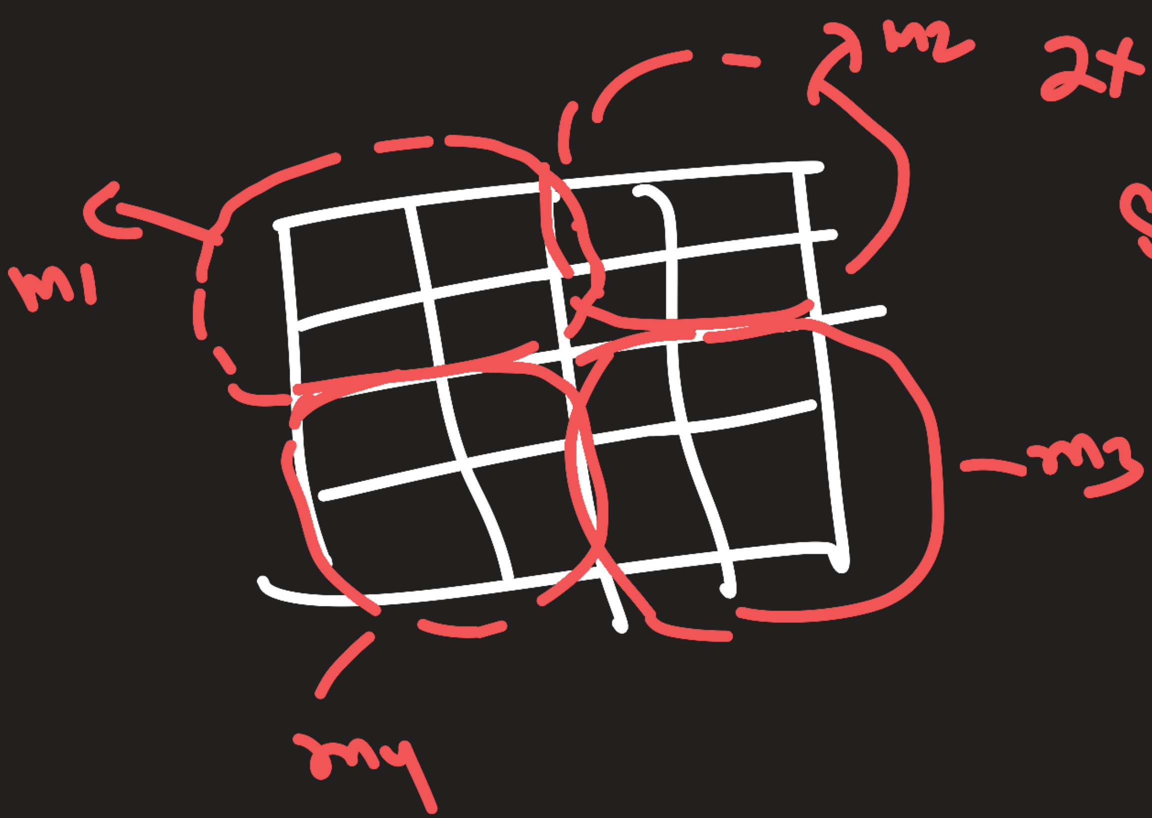
→ Padding ✓

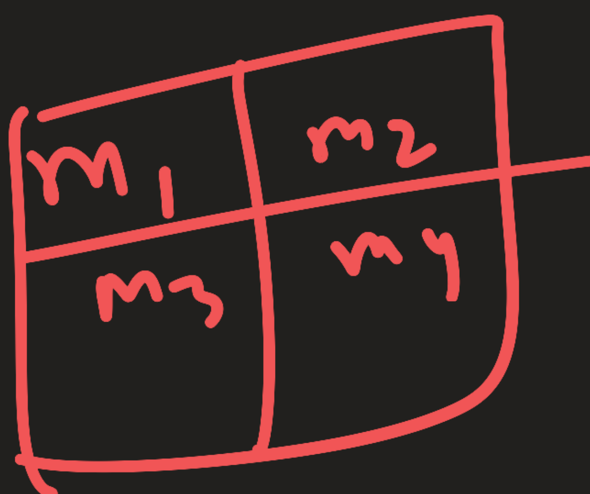
→ Pooling → Max Pooling (2×2) 2

→ Stride → ✓

2×2 maxpool

Stride = 2





Revision

Inner layers of network

Conv Block

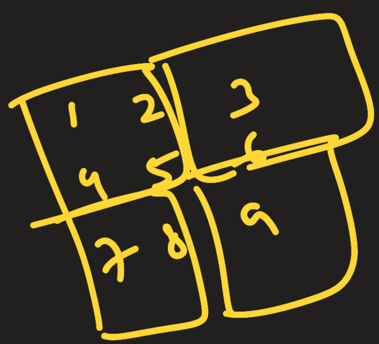
[Conv2D ReLU Maxpool] → [Conv2D ReLU Maxpool]

B1

Conv
Relu
Maxpool
Conv

} ← feature extracting

Dense(softmax) } CCE



max(1,2,3,4)

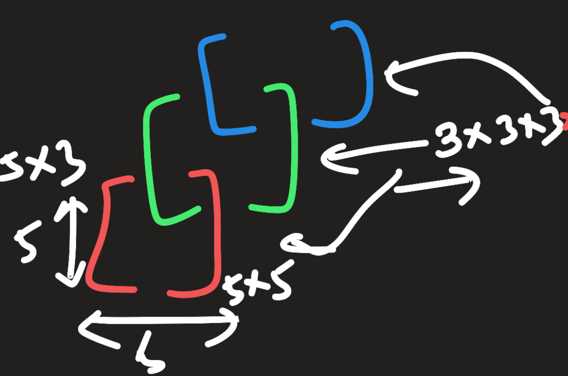
mean(1,2)

1k

5×5×3 $\xrightarrow{3\times3\times3\times5}$ 3×3×5

flatten

3×3×5 → [1 45] 45



5×5×3

5

5

5×5

3×3×3×10

3×3×10

1 [ ] 3×3

2 [ ] 3×3

10 [ ] 3×10

↓ flatten

[1 2 90] 90

↓

L2

L1 — [1 2 10] 1×10

(softmax)

Input

5×5×3

×10

[ ] 3×3×3

Feature Map

[ ]

3×3×10

Agenda

→ Global Average Pooling
→ Batchnorm, L1/L2, Dropout
→ Data Augmentation

→ Rebuilding the model pipeline
→ Rebuilt the Data Pipeline

Reduce Overfitting?

Data

Data Augmentation

300 → 900

Model

→ GAP
→ Batchnorm
→ Dropout

Optimizer

→ Learning Rate Scheduler
→ Early Stopping
→ Model checkpoint

128x128x3 → conv → Maxpool → Dense(ReLU) → Dense(Softmax)

→ Rootcause } Dense Layer

Where we left off? A typical CNN pipeline

Titanic → M/F Age

Image

conv
ReLU
Maxpool

→ feature correlation

→ invariant

← non linearity

Dog

Features →

feature

Softmax

$[\ ]_{H\times W\times C} \rightarrow$ flatten $\rightarrow$ D (ReLU) $\rightarrow$ D(softmax)

Global Average Pooling

I/P $\rightarrow$ [Conv, Relu, Maxpool] $\rightarrow$ [Conv, Relu, Maxpool] $\rightarrow$ $H \times W \times C$

$5 \times 5 \times 3$

Mo

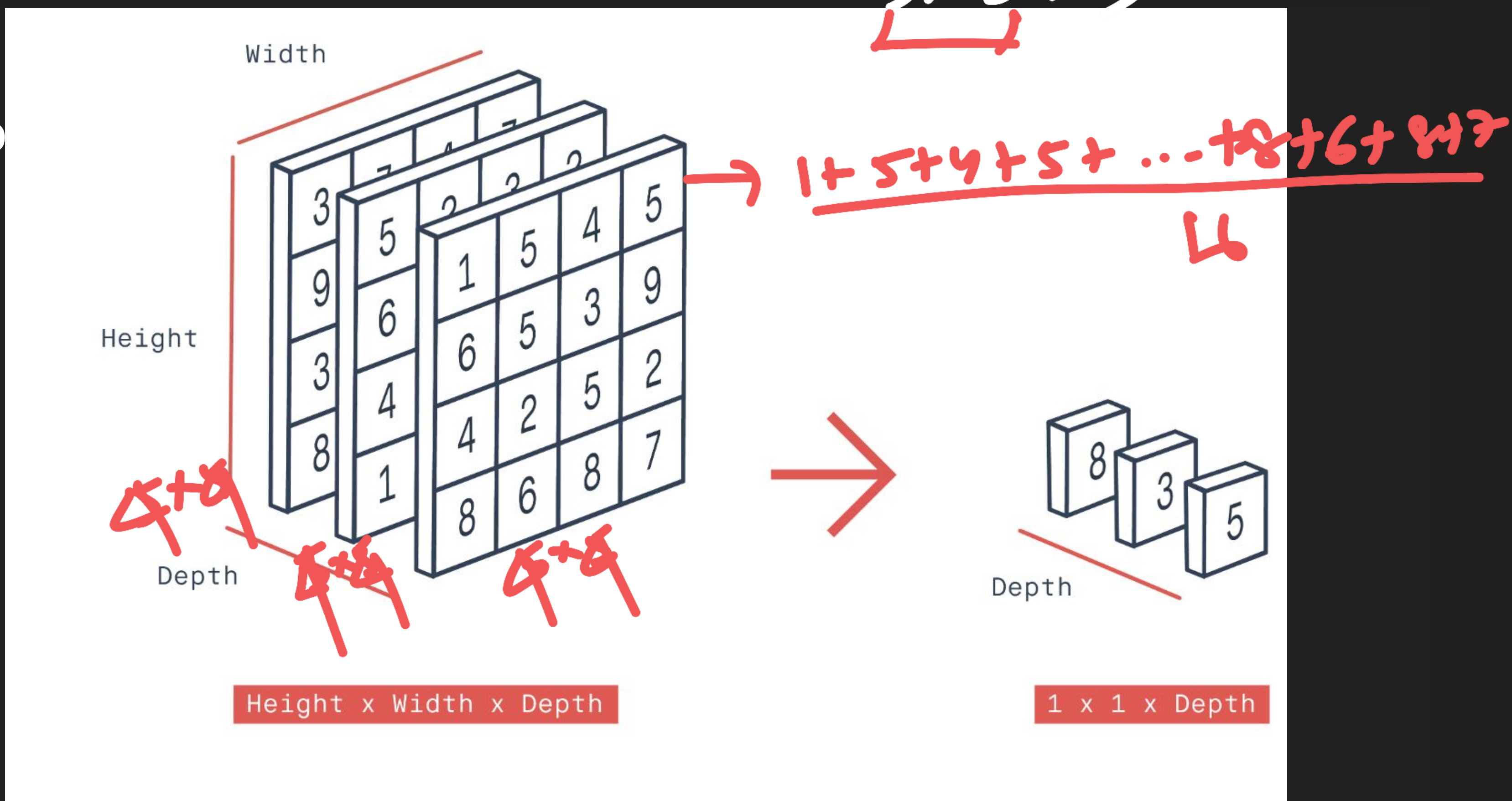


Figure 1. 2D Global average pooling operation. Here 'Depth' = 'Filters'

$4 \times 4 \times 3 \xrightarrow{GAP} 1 \times 3$

C1
C2
C3

Dense(softmax)

10 Classes

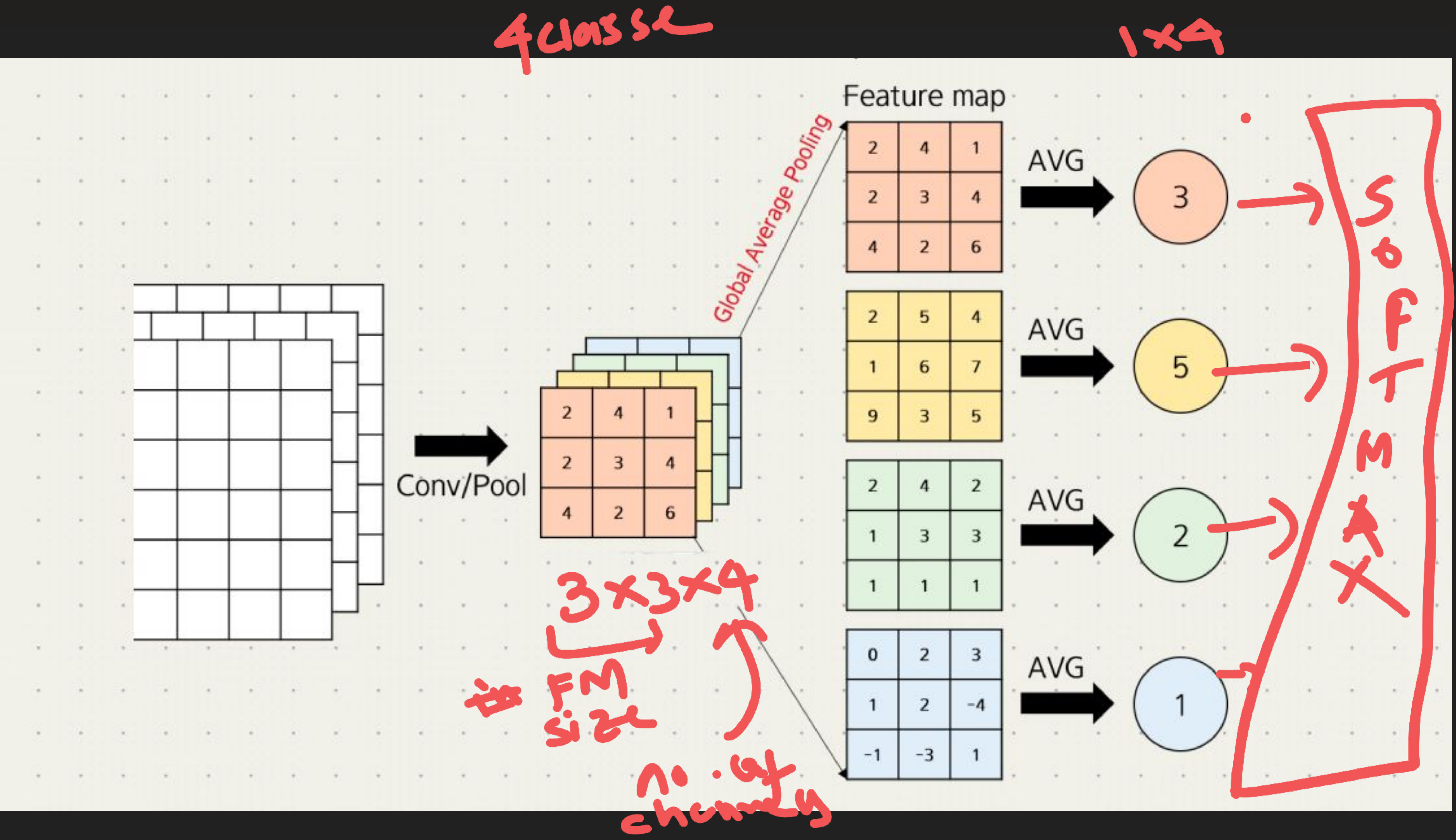
I/P
$128 \times 128 \times 3$ $\rightarrow$ CBlock1 $\rightarrow$ ... $\rightarrow$ CBlockN

$1 \times 10 \xleftarrow{GAP} 10 \times 10 \times 10$

1 ○
2 ○
⋮
10 ○

} softmax $\rightarrow$ CCE

fitting



1 2 . . . 10

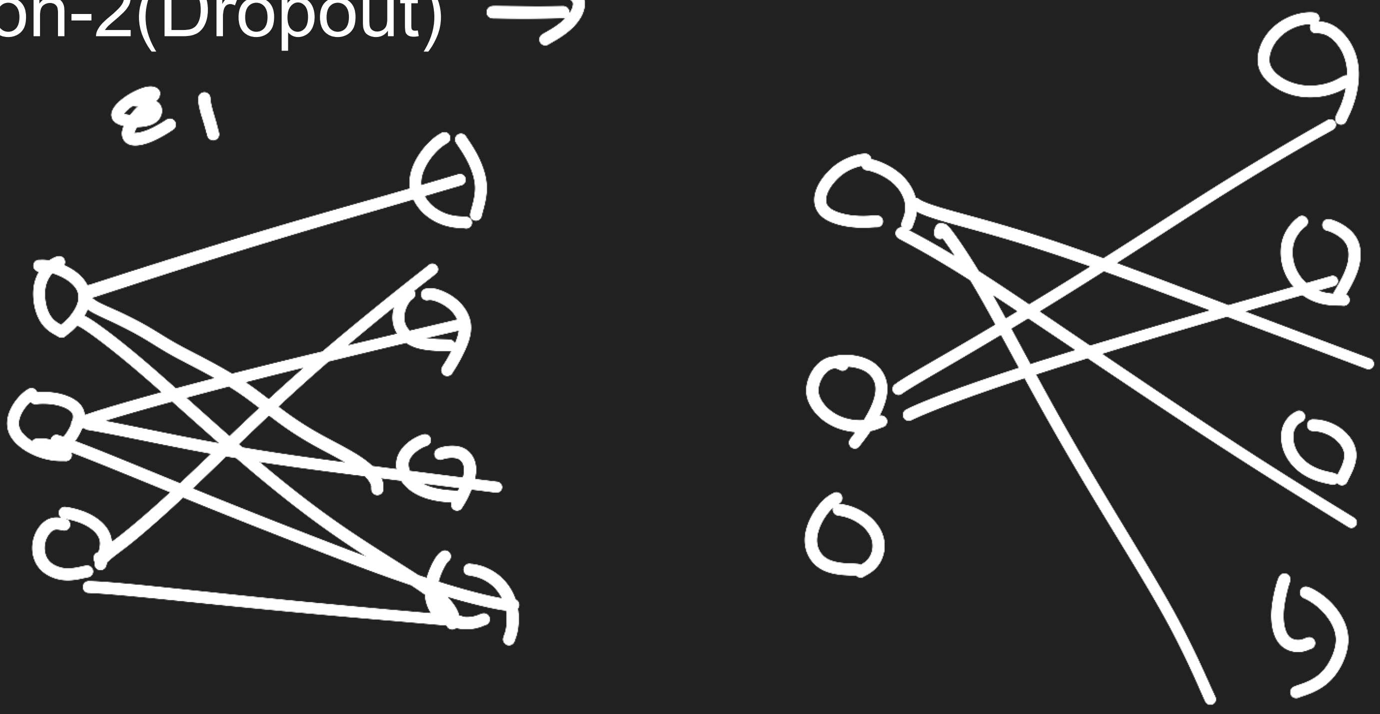
argmax(   )

flatten

3x3x10 → [   ] 90

GAP → 1x10

Modification-2(Dropout) →



I/P

↓

[conv
R
MP]

↓

[conv
R
MP]

↓

D(Relu)

↓

D(softmax)

← Dropout

Modification-3(Batch Norm)

$[\ \ ]_{5\times5\times3}$

$BS = 8/16/32$

$3 \times \underbrace{2 \times 2}_{h\ w} \times 4$

Usually $BS = 3$

Layer

K1 $\quad$ K2

$\cancel{3\times3\times3\times4}$ 4 kernels

	C1	C2	C3	C4
I_1	$[\ \ ]_{2\times2}$	$[\ \ ]_{2\times2}$	-	-
I_2	$[\ \ ]_{2\times2}$	$[\ \ ]_{2\times2}$	-	-
I_3	$[\ \ ]_{2\times2}$	$[\ \ ]_{2\times2}$	-	-
Batch	μ_1, σ_1	μ_2, σ_2	μ_3, σ_3	μ_4, σ_4

γ, β 2 parameters

$$\hat{x} = \frac{x-\mu}{\sigma}$$

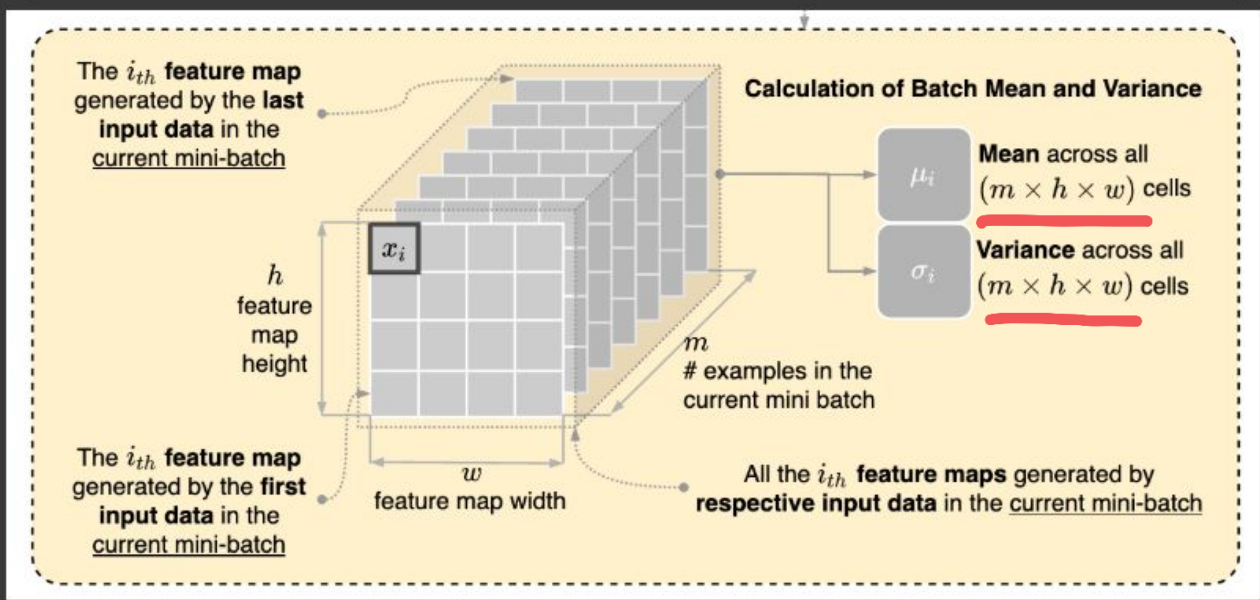
$\gamma = \sigma$
$\beta = \mu$

$\gamma = 1$
$\beta = 0$

$$\gamma \hat{x} + \beta = \cancel{\sigma} \cdot \frac{x-\mu}{\cancel{\sigma}} + \cancel{\mu}$$

no need of normaliz

1. Computing the mean of your minibatch: $\mu_B^{(k)} \leftarrow \frac{1}{m} \sum_{i=1}^{m} x_{B_i}^{(k)}$.
2. Computing the variance of your minibatch: $\sigma_B^{2(k)} \leftarrow \frac{1}{m} \sum_{i=1}^{m} (x_{B_i}^{(k)} - \mu_B^{(k)})^2$
3. Normalizing the value: $\hat{x}_B^{(k)} \leftarrow \frac{x_B^{(k)} - \mu_B^{(k)}}{\sqrt{\sigma_B^{2(k)} + \epsilon}}$
4. Scaling and shifting: $y_i \leftarrow \gamma \hat{x}_B^{(k)} + \beta$.



# of DP in Batch
h
w

$m \times h \times w \times c$ ← # of filters

# of DP in Batch

image h image w

γ
β
EMA μ
EMA σ
→ 4 params of BN

inference

~~X_P~~

1 DP

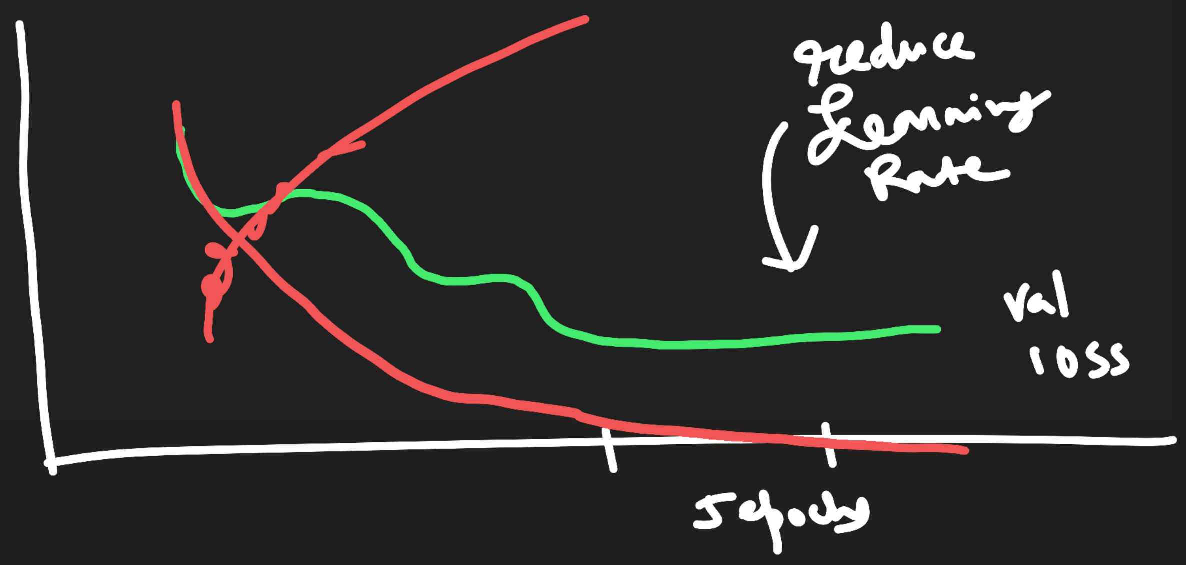
μ, σ

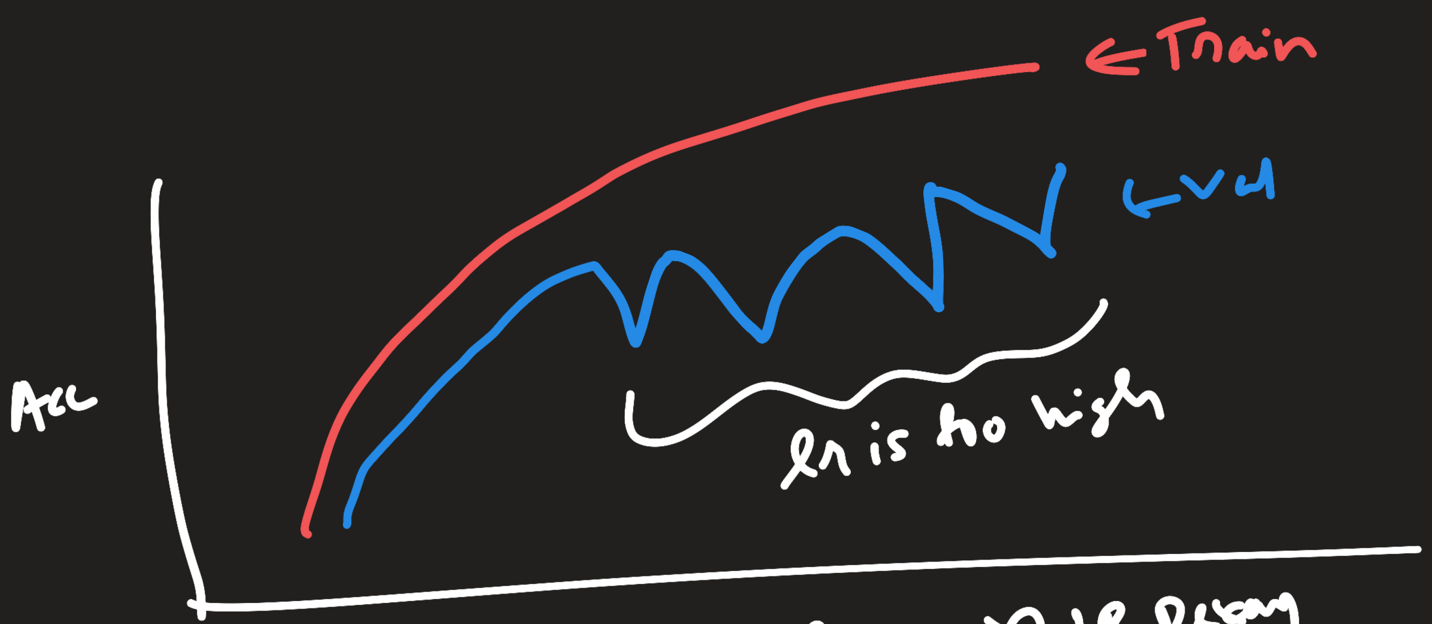
D1 D2 D100

EMA of μ, σ

$B1 = 0.9 * B0 + 0.1 * \mu$
$B2 = 0.9 * B1 + 0.1 * \mu$

Modification 4(ReducaLROnPlateu)





Reduce LR on Plateau → LR Decay

Only 1

→ GAP } Model

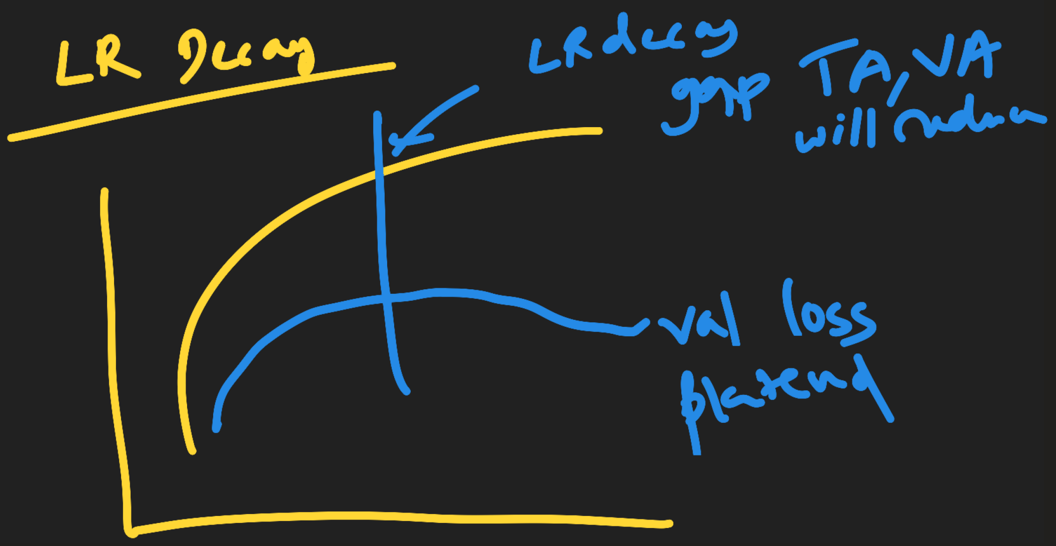
→ Batchnorm } Model

→ L1/L2 regularization } Model

→ Reduce LR on plateau
Early stopping
Model checkpt } optimizing

→ Data } DATA

Modification 5-L1/L2



T

EMA_μ
EMA_σ

$\mu_1, \sigma_1 \leftarrow B1$
$\mu_2, \sigma_2 \leftarrow B2$

$\mu_n, \sigma_n \leftarrow BN$

I

1 D?

EMA_μ

$$\frac{x-\mu}{\sigma}$$

EMA_σ

Inference $\mathrm{EMA}\ \mu$

Train Batch

$$\hat{x} = \frac{x-\mu}{\sigma}$$

$\mathrm{EMA}_\mu, \mathrm{EMA}_\sigma$

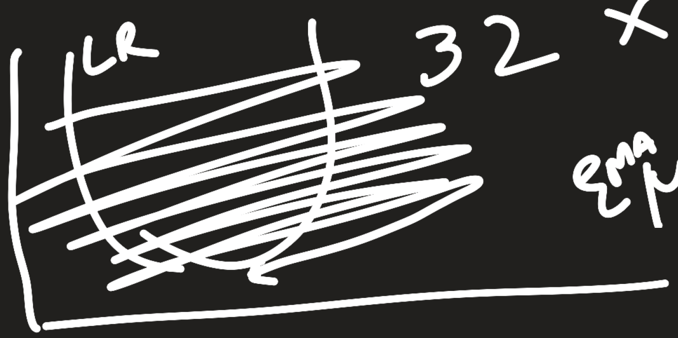
actual op $\gamma \hat{x} + \beta$

2

BN → γ fn

Overfitting

TA
VA



32 × γ

EMA μ, EMA σ, γ, β

99
95
90
85

Data Augmentation