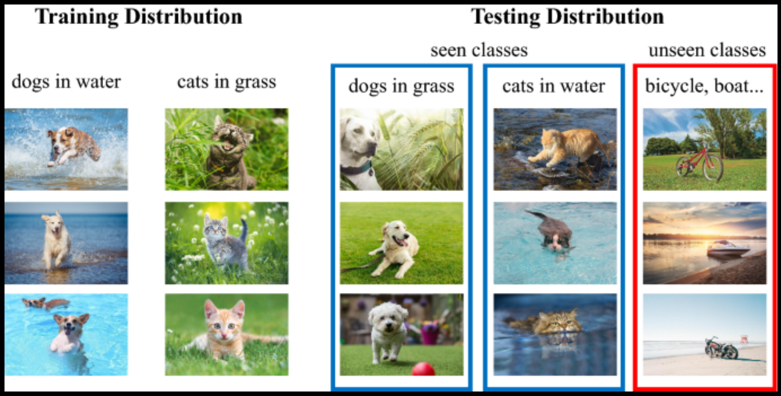
Today's Agenda

1) Backprop Recap
2) Batch Norm
3) Optimizers

Batch Normalization

1) Internal covariate shift



train

test

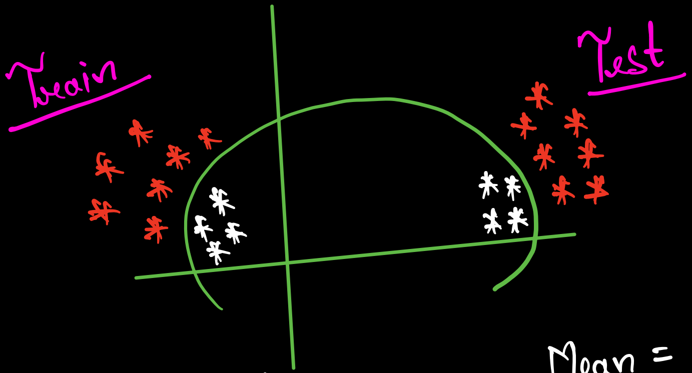
↳ unseen

overfitting

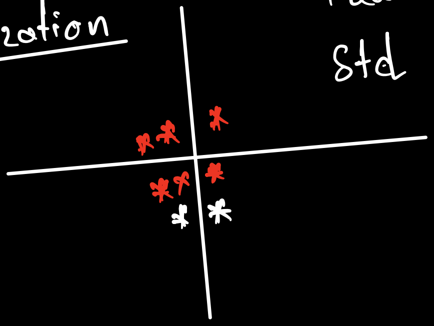
{Object}- Dog

Background — Water

Retrain the network



Normalization

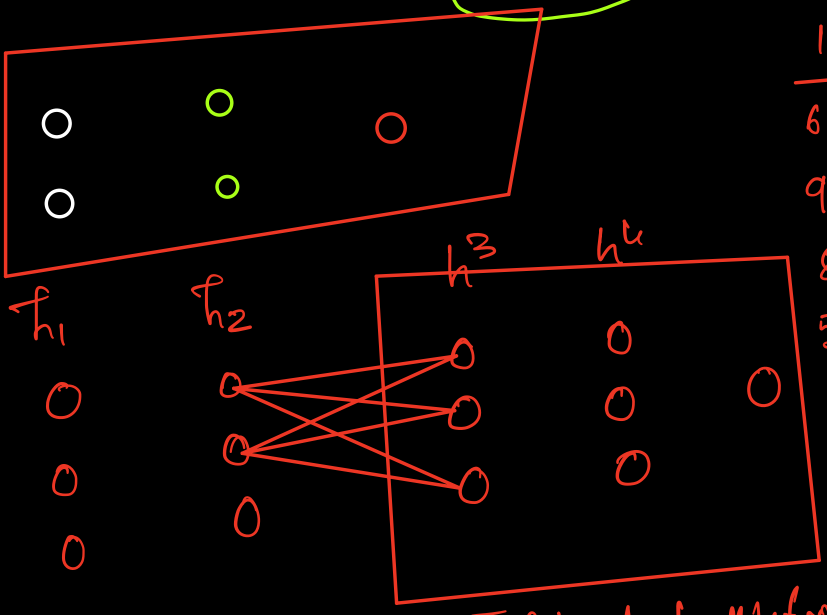


Mean = 0
Std = 1

$$X_i = \frac{X_i - Mean_i}{Std\ Dev_i}$$

Mini batch = hyperparameter → Batch Size

10	12	Placed
6	8	0
9	9	1
8	8	1
5	6	1
⋮	⋮	⋮



Independent Network

1. Training Time

↳ low LR

$z_{11} = wx + b$

$= w_1 10 + w_2 12 + b$

z_{12}
13
14

$g(z) = a_{11}$

✓ $z_{11} \rightarrow z_{11}^N \rightarrow g(z_{11}^N) \rightarrow a_{11}$

OR

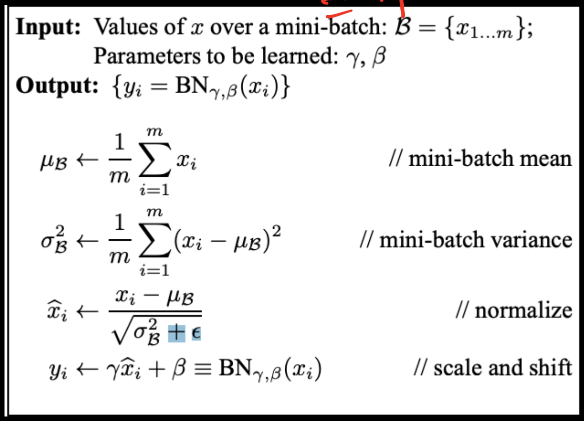
* $z_{11} \rightarrow g(z_{11}) \rightarrow a_{11} \rightarrow a_{11}^N$ *

$$\frac{z_{11} - \mu}{\sigma}$$

Mini Batch
||
Batch Size
= 4

Mean μ $= \frac{1}{4} \cdot ?$

$$\underline{\underline{\mu}} = \frac{1}{m} \sum z_{11}^i$$



```
Input:  Values of $x$ over a mini-batch: $\mathcal{B} = \{x_{1...m}\}$;
        Parameters to be learned: $\gamma, \beta$
Output: $\{y_i = \text{BN}_{\gamma,\beta}(x_i)\}$

  $\mu_\mathcal{B} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i$                              // mini-batch mean
  $\sigma_\mathcal{B}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_\mathcal{B})^2$           // mini-batch variance
  $\widehat{x}_i \leftarrow \frac{x_i - \mu_\mathcal{B}}{\sqrt{\sigma_\mathcal{B}^2 + \epsilon}}$                        // normalize
  $y_i \leftarrow \gamma\widehat{x}_i + \beta \equiv \text{BN}_{\gamma,\beta}(x_i)$              // scale and shift
```

$$\sigma = \sqrt{\frac{1}{m}\sum_{i=0}^{m}(z_{11}^{i} - \mu)}$$

Normalize

$$z_{11}^{N} = \frac{z_{11} - \mu}{\sqrt{\sigma^{2} + \epsilon}}$$

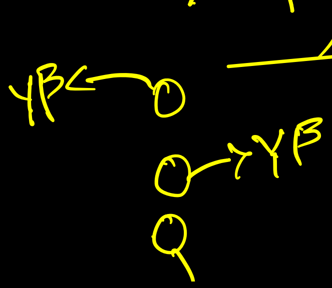
$$= \frac{z_{11} - \mu}{\sigma + \sqrt{\epsilon}}$$ → Error Term

Scale & Shift

$$z_{11}^{BN} = \gamma z_{11}^{N} + \beta$$

Batch Normalization Parameters

$\gamma \; \beta$ → Learnable Parameters



$\gamma \overset{\checkmark}{\beta}$

Test / Inference

Moving Average

Mean

μ

Std. dev

σ

During Treaing

data points = 1000

Batch Size = 100

Iterations = $\underline{\underline{10}}$

$$\frac{\mu_1 + \mu_2 + \mu_3 + \mu_4 + \mu_5 \ldots \underline{\underline{\mu_{10}}}}{10}$$

$= 2$

Non learnable Paramaters

1) Mov. Avg Mean *

2) Mov. Avg Std dev