Subject :

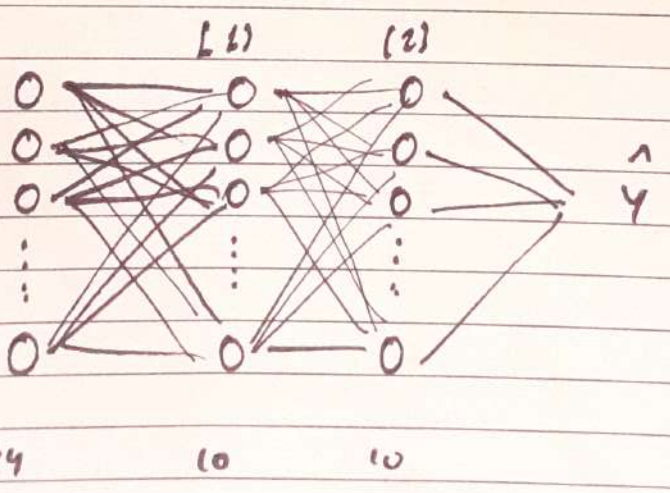
Date :

Neural Net !!

$$\underset{28 * 28}{\boxed{784}} \Rightarrow \underbrace{0, 1, 2, \ldots, 9}_{10 \text{ classes}}$$

$$\left.\begin{bmatrix} [1, 2, \ldots\ldots & 28] \\ [&] \\ [&] \\ \vdots & \\ [&] \end{bmatrix}\right\} 28$$

$$[1, 2, \ldots\ldots\ldots\ldots\ldots, 784]$$



Subject :

Date :

Forward Pass :)

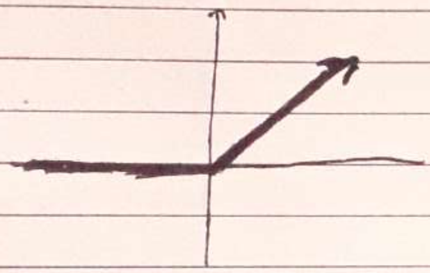
$$A^{(0)} = X \quad (784 \times m)$$

$$Z^{(1)} = W^{(1)} A^{(0)} + b^{(1)}$$

$10 \times m \qquad 10 \times 784 \qquad 784 \times m \qquad 10 \times 1 \rightarrow 10 \times m$

$$A^{(1)} = \sigma(Z^{(1)}) = ReLu(Z^{(1)})$$

Rectified Linear Unit



$$ReLu(x) \begin{cases} x & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}$$

$$Z^{(2)} = W^{(2)} A^{(1)} + b^{(2)}$$

$10 \times m \qquad 10 \times 10 \qquad 10 \times m \qquad 10 \times 1 \rightarrow 10 \times m$

$$A^{(2)} = Softmax(Z^{(2)})$$

Softmax

$$S(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$$

Subject :

Date :

Back Prop :)

$$dZ^{(2)} = A^{(2)} - Y \text{ (one hot encoding)}$$

10×m 10×m 10×m

$$dw^{(2)} = \frac{1}{m} dZ^{(2)} A^{(1)T}$$

10×10 10×m m×10

$$db^{(2)} = \frac{1}{m} \sum dz^{(2)}$$

10×1 10×1

$$dz^{(1)} = W^{(2)T} dz^{(1)} \cdot \sigma'(z^{(1)})$$

10×m 10×10 10×m 10×m

$$dw^{(1)} = \frac{1}{m} dz^{(1)} X^{T}$$

10×784 10×m m×784

$$db^{(2)} = \frac{1}{m} \sum dz^{(1)}$$

10×1 10×1

31 lines, 7mm

Subject :

Date :

UPDATE !!

$$w^{(1)} := w^{(1)} - \alpha \, dw^{(1)}$$

$$b^{(1)} := b^{(1)} - \alpha \, db^{(1)}$$

$$w^{(2)} := w^{(2)} - \alpha \, dW^{(2)}$$

$$b^{(2)} := b^{(2)} - \alpha \, db^{(2)}$$

α = learning rate

By HAFIDZ

Subject :

Date :

Notation

· Scalars $\Rightarrow$ 1 number (3, -4, 2.000.000.000)

$u \in \mathbb{R} / \mathbb{N} / \mathbb{Z}$

· Vectors $\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \approx>$ 3 dimensional Vector

$\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} = u \approx u \in \mathbb{R}^n$, n = amount of elements

$\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \approx u \in \mathbb{R}^2$

· Matrices $\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$

$A \in \mathbb{R}^{m \cdot n}$, m: row and n = columns

Subject :

Date :

· Tensors

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \quad 3$$

$A \in \mathbb{R}^{m \cdot n \cdot j}$; m = row , n : columns ,
j : layer / channel

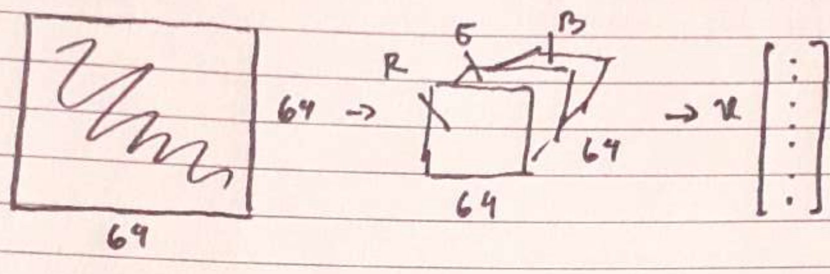
$$\begin{bmatrix} (v,v,v) & (v,v,v) & (v,v,v) \\ (v,v,v) & (v,v,v) & (v,v,v) \\ (v,v,v) & (v,v,v) & (v,v,v) \end{bmatrix}$$

1st layer 2nd layer 3rd layer

Subject :

Date :

Notation



$= 64 \cdot 64 \cdot 3$
$= 12288$

data for 1 image $= (x, Y)$; (input, label)

$x \in \mathbb{R}^{n_x}$, $Y \in$ (one_hot_encoding)

m training exam : $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots (x^{(m)}, y^{(m)})\}$

$$X_{input} = \begin{bmatrix} \vdots & \vdots & & \vdots \\ x^{(1)} & x^{(2)} & \cdots & x^{(m)} \\ \vdots & \vdots & & \vdots \end{bmatrix} \updownarrow n_x \ (12288)$$

$\longleftarrow m \longrightarrow$

amount of images

$X_{input} \in \mathbb{R}^{n_x \cdot m}$ with (n_x, m) dimensional matrix

31 lines, 7mm

Jolo

Subject :

Date :

MATRIX

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \text{row} \Rightarrow (2,2) \text{ or } (\text{row}, \text{colum})$$

colum

$$B = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix} \Rightarrow (3,2) \text{ dimension / shape}$$

Transpose

$$A' = A^T = \begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}$$

switch row into colum, and colum into row

Determinant

$a \cdot d - b \cdot c$

$$Det(A) = |A| = \begin{pmatrix} 2 & 4 \\ 6 & 8 \end{pmatrix} = \begin{pmatrix} 2=a & 4=b \\ 6=c & 8=d \end{pmatrix} \Rightarrow 2 \cdot 8 - 4 \cdot 6 \Rightarrow 16 - 24 \Rightarrow -8$$

31 lines, 7mm

Subject :

Date :

Invers Matrix

$$A^{-1} = \frac{1}{|A|}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \quad \left\{ \begin{pmatrix} a & c \\ b & d \end{pmatrix} \right.$$

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}^{-1} = ?$$

find the determinant

$$|A| = a \cdot d - b \cdot c$$
$$= 1 \cdot 4 - 2 \cdot 3$$
$$= 4 - 6$$
$$= -2$$

$$A^{-1} = \frac{1}{-2} \cdot \begin{pmatrix} 4 & -3 \\ -2 & 1 \end{pmatrix} = \begin{pmatrix} 4/-2 & -3/-2 \\ -2/-2 & 1/-2 \end{pmatrix}$$

Addition

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} + \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} = \begin{pmatrix} 6 & 8 \\ 10 & 12 \end{pmatrix}$$

Subtraction

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} - \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix} = \begin{pmatrix} -4 & -4 \\ -4 & -4 \end{pmatrix}$$

Subject :

Date :

Multiplication

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix} \cdot \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} 1\cdot1+2\cdot3+3\cdot5 & 1\cdot2+2\cdot4+3\cdot6 \\ 4\cdot1+5\cdot3+6\cdot5 & 4\cdot2+5\cdot4+6\cdot6 \end{pmatrix} = \begin{pmatrix} 1+6+15 & 2+8+18 \\ 4+15+30 & 8+20+36 \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} 22 & 28 \\ 49 & 64 \end{pmatrix}$$

$$r \cdot c = r \cdot c \begin{cases} (2 \times 3) \cdot (3 \times 2) & \checkmark \\ (2 \times 3) \cdot (2 \times 3) & \times \end{cases}$$

Division

There's no such thing as matrix division

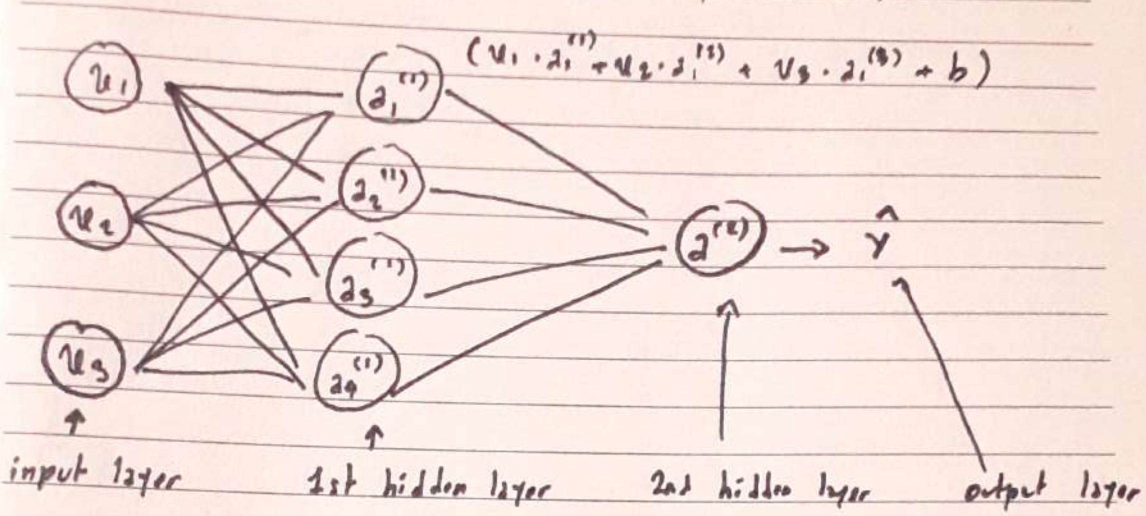
31 lines, 7mm

Hapidel

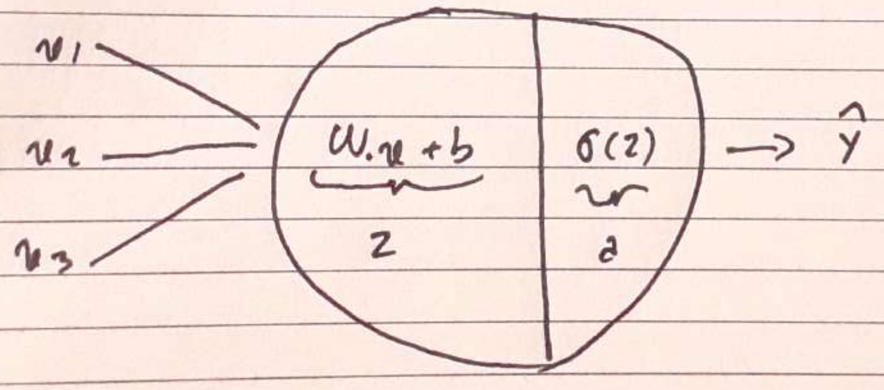
Subject :

Date :

RECAP (Neural Net Representation)



One Neuron Representation



Subject :

Date :

$$\underbrace{\begin{array}{ccc} \begin{bmatrix} x_1 & x_1 & x_1 \\ x_2 & x_2 & x_2 \\ x_3 & x_3 & x_3 \\ \vdots & \vdots & \vdots \\ x_n & x_n & x_n \end{bmatrix} \\ \downarrow \quad\quad \downarrow \quad\quad \downarrow \\ \text{img } 1 \quad \text{img } 2 \quad \text{img } 3 \end{array}}_{\text{images data}} \cdot \underbrace{\begin{bmatrix} w_1^1 & w_2^1 & w_3^1 & \dots & w_n^1 \\ w_1^2 & w_2^2 & w_3^2 & \dots & w_n^2 \\ w_1^3 & w_2^3 & w_3^3 & \dots & w_n^3 \\ \vdots & \vdots & \vdots & & \vdots \\ w_1^5 & w_2^5 & w_3^5 & \dots & w_n^5 \end{bmatrix}}_{\text{weight values}}$$

$$\begin{bmatrix} S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix} \Rightarrow \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \\ z_4 \end{bmatrix}$$

$$S_5 \rightarrow \sum_{i=1}^{n} (w_i \cdot x_i)$$

31 lines, 7mm

Subject :

Date :

$$z_1^{(1)} = w_1^{(1)} x + b_1^{(1)}, \quad a_1^{(1)} = \sigma(z_1^{(1)})$$
$$z_2^{(1)} = w_2^{(1)} x + b_2^{(1)}, \quad a_2^{(1)} = \sigma(z_2^{(1)})$$
$$z_3^{(1)} = w_3^{(1)} x + b_3^{(1)}, \quad a_3^{(1)} = \sigma(z_3^{(1)})$$
$$z_4^{(1)} = w_4^{(1)} x + b_4^{(1)}, \quad a_4^{(1)} = \sigma(z_4^{(1)})$$

$$\underbrace{\begin{bmatrix} \cdots w_1^{(1)} \cdots \\ \cdots w_2^{(1)} \cdots \\ \cdots w_3^{(1)} \cdots \\ \cdots w_4^{(1)} \cdots \end{bmatrix}}_{w^{(1)}\ (4,3)} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \underbrace{\begin{bmatrix} b_1^{(1)} \\ b_2^{(1)} \\ b_3^{(1)} \\ b_4^{(1)} \end{bmatrix}}_{b^{(1)}\ (4,1)}$$

$$\Rightarrow \begin{bmatrix} w_1^{(1)} \cdot x + b_1^{(1)} \\ w_2^{(1)} \cdot x + b_2^{(1)} \\ w_3^{(1)} \cdot x + b_3^{(1)} \\ w_4^{(1)} \cdot x + b_4^{(1)} \end{bmatrix} = \begin{bmatrix} z_1^{(1)} \\ z_2^{(1)} \\ z_3^{(1)} \\ z_4^{(1)} \end{bmatrix} = z^{[1]}$$

$$a^{[1]} = \begin{bmatrix} a_1^{(1)} \\ a_2^{(1)} \\ a_3^{(1)} \\ a_4^{(1)} \end{bmatrix} = \sigma(z^{[1]})$$

input x :

$$\rightarrow z^{(1)} = w^{[1]} x + b^{(1)}$$
$$\rightarrow a^{(1)} = \sigma(z^{(1)})$$
$$\rightarrow z^{(2)} = w^{(2)} a^{(1)} + b^{(2)}$$
$$\rightarrow a^{(2)} = \sigma(z^{(2)})$$

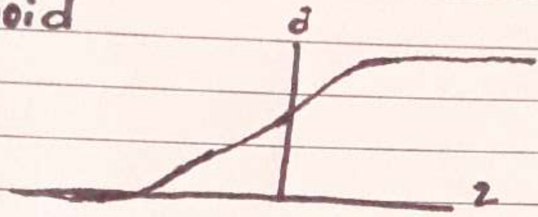
31 lines, 7mm

Subject :

Date :

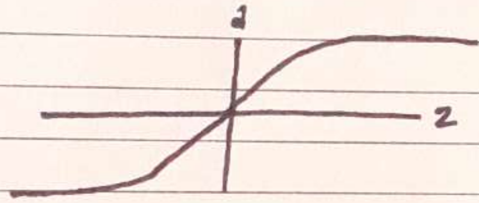
ACTIVATIONS

Sigmoid



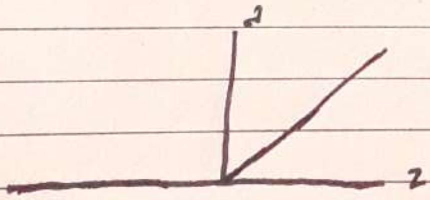
$$a = \frac{1}{1+e^{-z}}$$

tanh



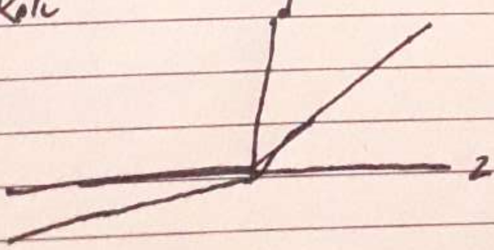
$$a = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$$

Relu



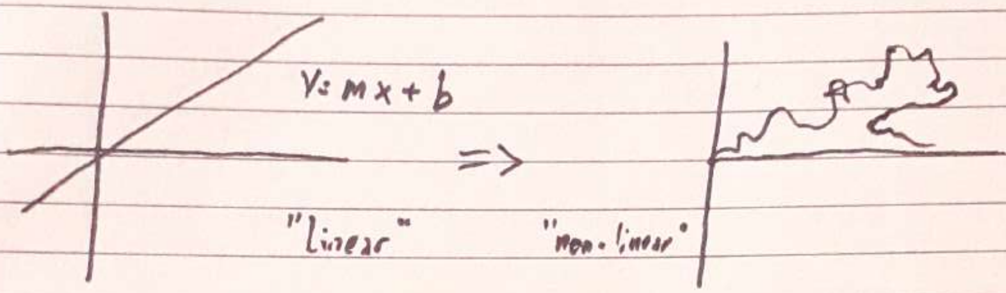
$$a = \max(0, z)$$

Leaky Relu

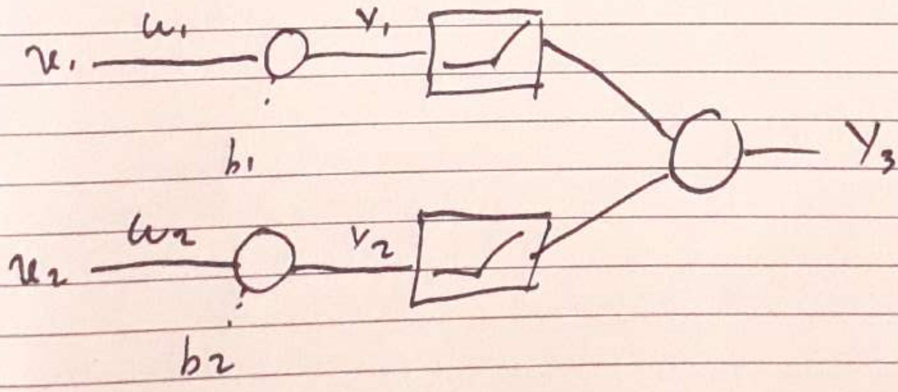


$$a = \max(0.01z, z)$$

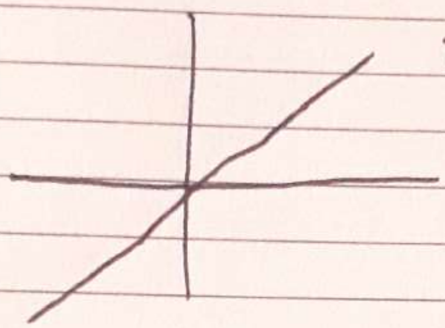
Universal Approximator



- Relu
- Tanh
- Sigmoid
- etc

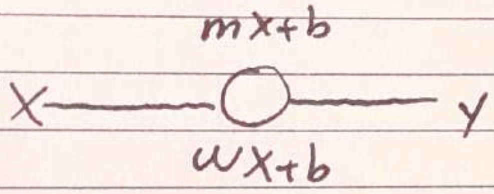


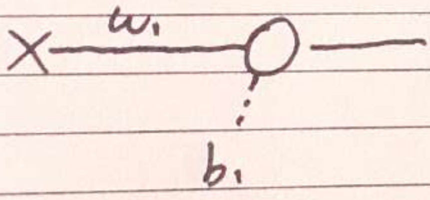
Why non-linear?



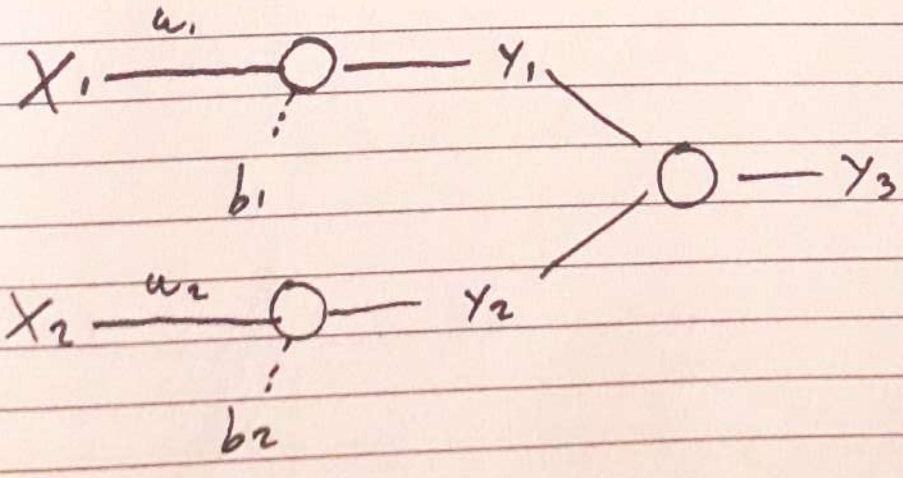
$y = mx + b$

"Linear function"





$\boxed{y_1 = w_1 x + b_1}$

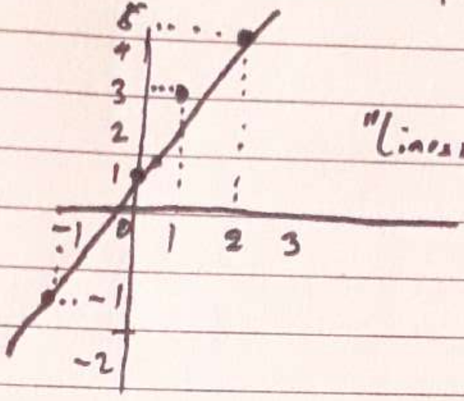


Subject :

Date :

$$Y_1 = W_1 X_1 + b_1$$
$$= 2X_1 + 1$$

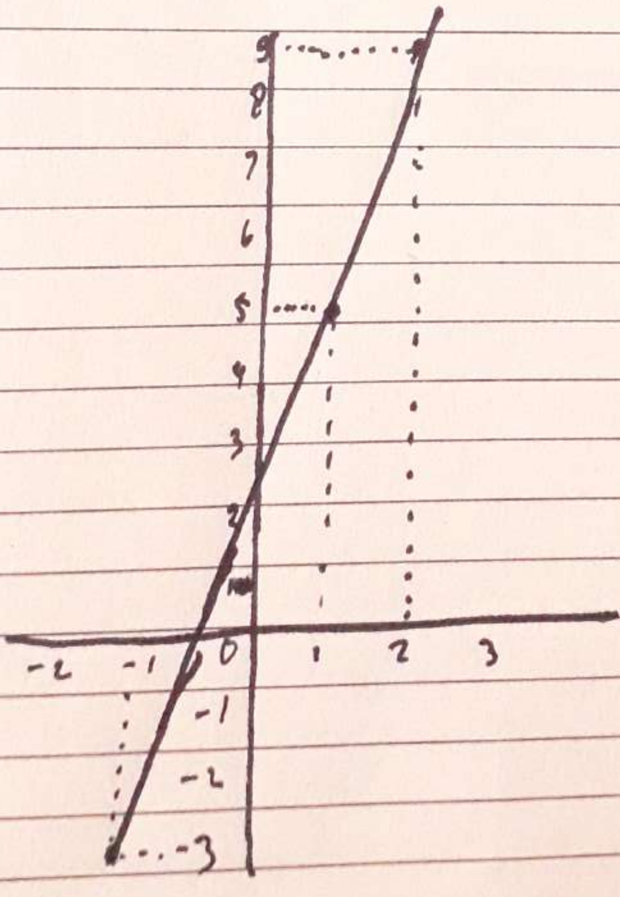
x	-1	0	1	2
y	-1	1	3	5



"Linear line"

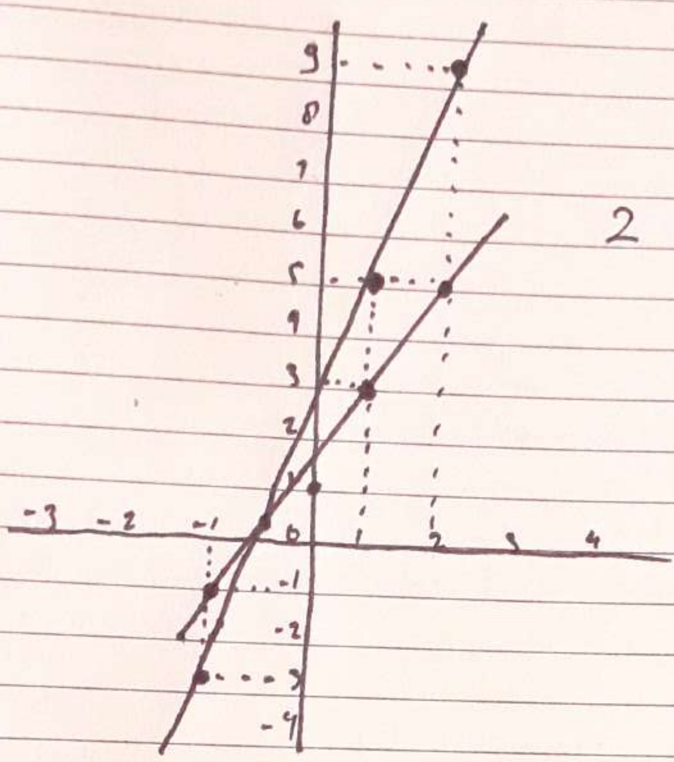
$$Y_2 = W_2 X_2 + b_2$$
$$= 4 X_2 + 1$$

x	-1	0	1	2
y	-3	1	5	9



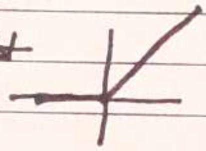
Subject :

Date :



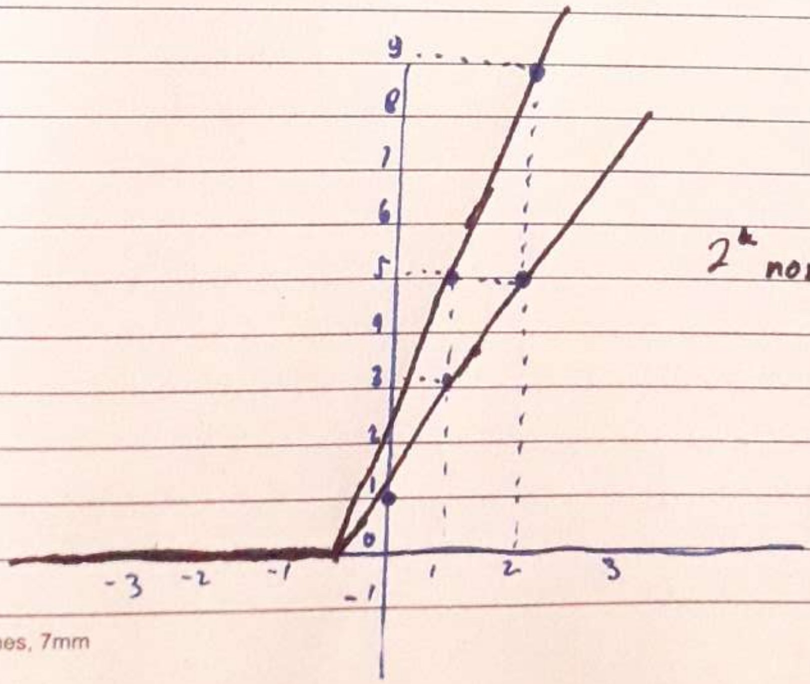
2 linear line

⇓ ReLU



x if $x > 0$

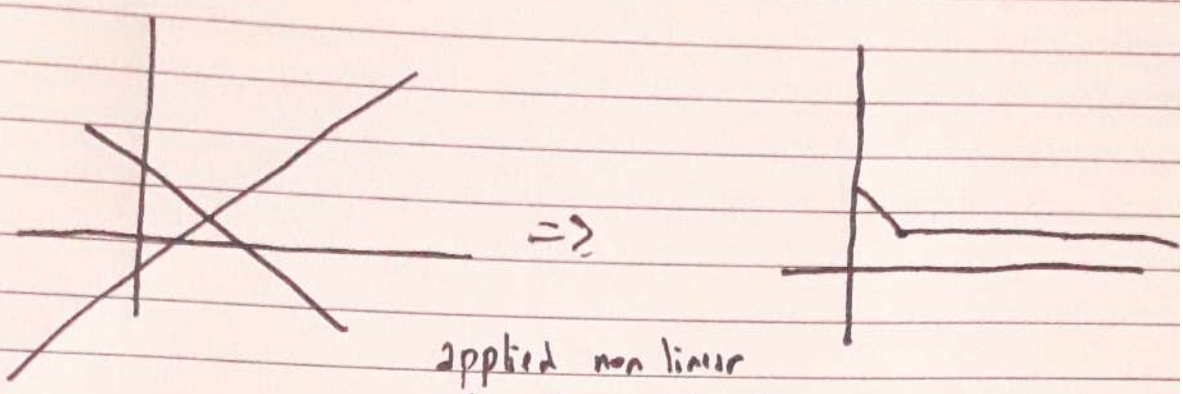
0 if $x \leq 0$



2 non-linear line

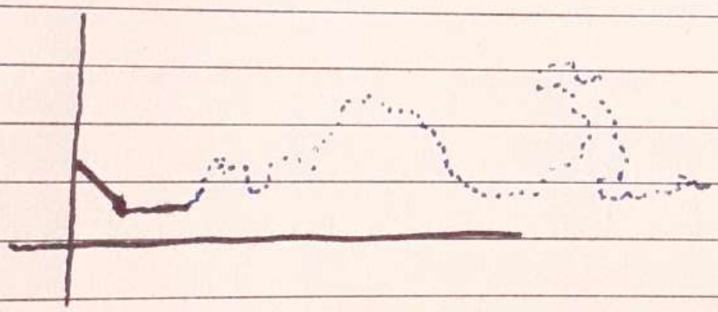
Subject :

Date :



applied non linear
function & combine it

use & combine many non-linear func



31 lines, 7mm

Subject :

Date :

Softmax

$$a^{o\_i} = \text{softmax (one\_image)} = \begin{bmatrix} 5 \\ 2 \\ -1 \\ 3 \end{bmatrix}$$

$$\text{Softmax}(x) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$$

$$a^{o\_i} = \begin{bmatrix} e^{5} / (e^{5}+e^{2}+e^{-1}+e^{3}) \\ e^{2} / (e^{5}+e^{2}+e^{-1}+e^{3}) \\ e^{-1} / (e^{5}+e^{2}+e^{-1}+e^{3}) \\ e^{3} / (e^{5}+e^{2}+e^{-1}+e^{3}) \end{bmatrix} = \begin{bmatrix} 0.842 \\ 0.041 \\ 0.002 \\ 0.113 \end{bmatrix}$$

$$\sum_{j=1}^{n} a^{o\_i\,j} = 1.$$

31 lines, 7mm

Hafidz L

Subject :

Date :

Loss Function

- MSE (Mean Squared Error)

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

n = amount of data
y_i = true label
$\hat{y}_i$ = prediction

- Negative Log-Likelihood (NLL)

$$\log P(\text{labels in training set}) = \log \prod_{i=1}^{m} P(y^{(i)} \mid x^{(i)})$$

$$\log p(\ldots) = \sum_{i=1}^{m} \underbrace{\log p(y^{(i)} \mid x^{(i)})}_{-\ell(\hat{y}^{(i)}, y^{(i)})}$$

$$= -\sum_{i=1}^{m} \ell(\hat{y}^{(i)}, y^{(i)})$$

31 lines, 7mm

Hafidz L

Subject :

Date :

Cost Function

- MSE (Mean Squared Error)

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

n = amount of data
y_i = true label
$\hat{y}_i$ = prediction label

- Negative Log-Likelihood (NLL)

$$J(w,b) = \frac{1}{m} \sum_{i=1}^{m} \ell(a^{(i)}, y)$$

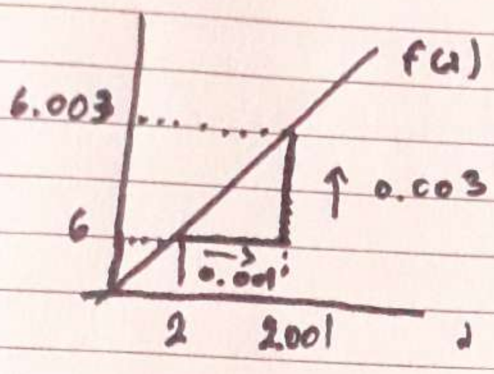
$$a^{(i)} = \hat{Y}^{(i)} = \text{softmax}(z^{(i)})$$

31 lines, 7mm

Subject :

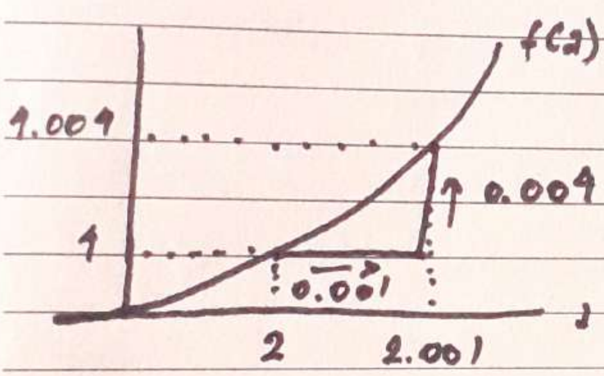
Date :

Derivative



if $a = 2$ then $f(a) = 6$
$a = 2.001$ then $f(a) = 6.003$

Slope / derivative of $f(a)$ at $a = 2$ is 3, because $\frac{\text{height}}{\text{width}} = \frac{0.003}{0.001} = 3$



if $a = 2$ then $f(a) = 4$
$a = 2.001$ then $f(a) \approx 4.004$

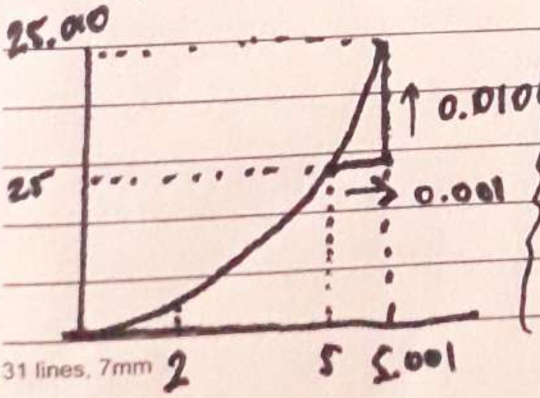
Slope of $f(a)$ at $a = 2$ is 4

$f(a) = a^2$, if $a = 5$ then $f(a) = 25$
if $a = 5.001$ then $f(a) = 25.010$

Slope of $f(a)$ at $a = 5$ is 10

High School Math

derivative of $a^2 = 2a$ using $da^n = na^{n-1}$, $x^3 = 3x^2$, $x^4 = 4x^3$



Hafidal

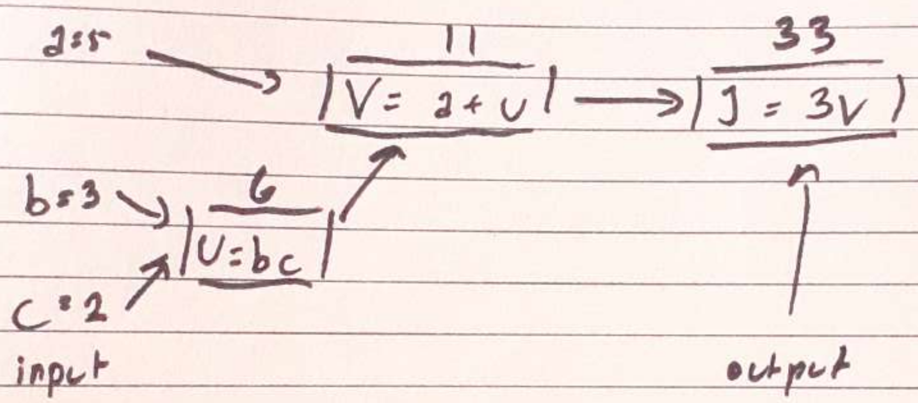
Subject :

Date :

Computation Graph

$$J(a,b,c) = 3(\underbrace{a+\underbrace{bc}_{u}}_{V})$$

$$\begin{cases} u = bc \\ V = a + u \\ J = 3V \end{cases}$$



$$\frac{dJ}{dV} = ?$$

$J = 33$, if $V = 11$ then $J = 3 \cdot 11 = 33$

$V = 11.001$ then $J = 3 \cdot 11.001 = 33.003$

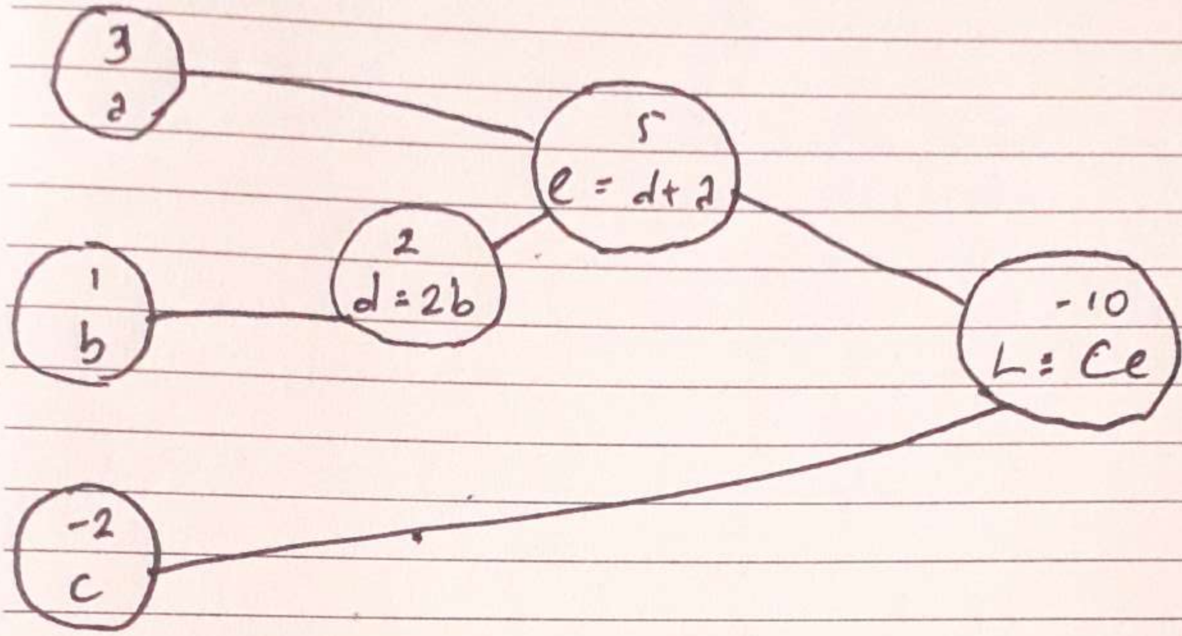
Slope of J when $V=11$ is 3, so $\frac{dJ}{dV} = 3$

$$\frac{dJ}{da} = \frac{dJ}{dV} \cdot \frac{dV}{da} \quad \Big\{ \underbrace{a \rightsquigarrow V \rightsquigarrow J}_{\frac{dJ}{da}}$$

Subject :

Date :

Derivative of Computational Graph



$$L(a, b, c) = c(a + 2b)$$

$$\frac{\partial L}{\partial a} = ?$$

$$\frac{\partial L}{\partial b} = ?$$

$$\frac{\partial L}{\partial c} = ?$$

Chain Rule

$$f(x) = u(v(x)) \Rightarrow \frac{df}{dx} = \frac{du}{dv} \cdot \frac{dv}{dx}$$

$$f(x) = u(v(w(x))) \Rightarrow \frac{df}{dx} = \frac{du}{dv} \cdot \frac{dv}{dw} \cdot \frac{dw}{dx}$$

Subject :

Date :

$d = 2 \cdot b$

$e = a + d$

$L = c \cdot e$

$\frac{\partial L}{\partial c} = e$

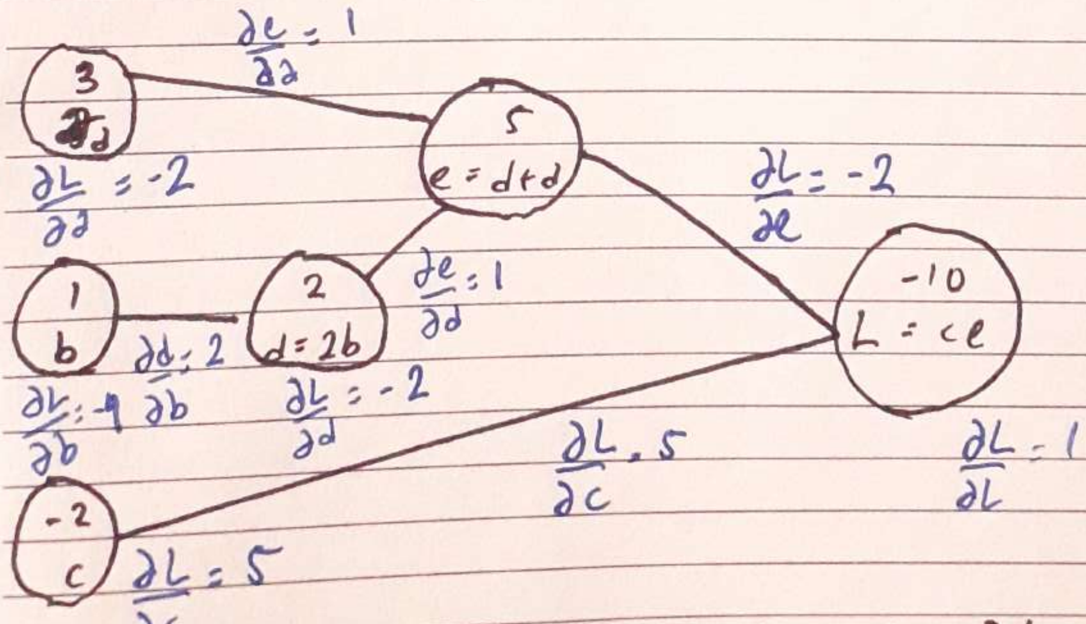
$\frac{\partial L}{\partial a} = \frac{\partial L}{\partial e} \cdot \frac{\partial e}{\partial a}$

$\frac{\partial L}{\partial b} = \frac{\partial L}{\partial e} \cdot \frac{\partial e}{\partial d} \cdot \frac{\partial d}{\partial b}$

$L = c \cdot e \; : \; \frac{\partial L}{\partial e} = c \; , \; \frac{\partial L}{\partial c} = e$

$e = a + d \; : \; \frac{\partial e}{\partial a} = 1 \; , \; \frac{\partial e}{\partial d} = 1$

$d = 2b \; : \; \frac{\partial d}{\partial b} = 2$



$\Longrightarrow \frac{\partial L}{\partial a} = -2 \; , \; \frac{\partial L}{\partial b} = -4 \; , \; \frac{\partial L}{\partial c} = 5$

31 lines, 7mm

Hafidz L

Subject :

Date :

Derivatives with Computation Graph

$z = w^T x + b$

$\hat{y} = a = \sigma(z)$

$L(a,y) = -(y\log(a) + (1-y)\log(1-a))$

$$x_1, w_1, x_2, w_2, b \rightarrow \boxed{z = w_1x_1 + w_2x_2 + b} \xleftarrow{\partial L/\partial z} \boxed{\hat{y} = a = \sigma(z)} \xleftarrow{\frac{\partial L}{\partial \hat{y}/a}} \boxed{L(a,y)}$$

$\frac{\partial L}{\partial w_1} = ?$, $\frac{\partial L}{\partial w_2} = ?$, $\frac{\partial L}{\partial b} = ?$

"da" $= \frac{\partial L}{\partial a} = \frac{\partial L(a,y)}{\partial a}$, $\frac{\partial L}{\partial a} = -(y\log(a) + (1-y)\log(1-a))$

$\frac{\partial}{\partial a}\log(a) = \frac{1}{a}$

$\frac{\partial}{\partial a}((1-y)\log(1-a)) = (1-y)\cdot\left(-\frac{1}{1-a}\right) = -\frac{1-y}{1-a}$

$\frac{\partial L}{\partial a} = -\left(\frac{y}{a} - \frac{1-y}{1-a}\right)$

"da" $= \frac{\partial L}{\partial a} = -\frac{y}{a} + \frac{1-y}{1-a}$

Subject :

Date :

$$"dz" = \frac{\partial L}{\partial z} = \frac{\partial L(a,y)}{\partial z} = \frac{\partial L(a,y)}{\partial a} \cdot \frac{\partial a}{\partial z}$$

$$= \frac{-y}{a} + \frac{1-y}{1-a} \cdot 1-a$$

$$"dz" = a - y$$

$$"dw_i" = \frac{\partial L}{\partial w_i} = \frac{\partial L(a,y)}{\partial w_i} = \frac{\partial L}{\partial a} \cdot \frac{\partial a}{\partial z} \cdot \frac{\partial z}{\partial w_i}$$

$$= \frac{\partial L}{\partial z} \cdot \frac{\partial z}{\partial w_i}$$

$$= a - y \cdot x_i$$

$$dw_i = x_i \cdot "dz"$$

$$"db" = "dz"$$

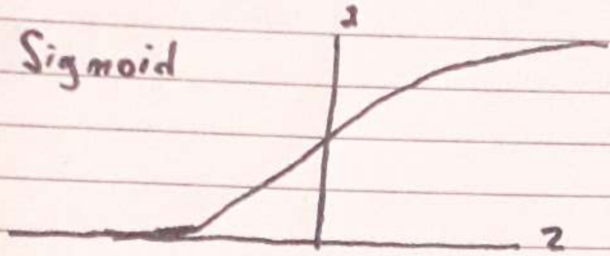
31 lines, 7mm

Subject :

Date :

Derivatives of Activations func

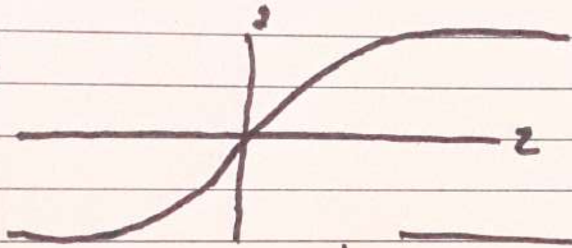
Sigmoid



$$g(z) = \frac{1}{1+e^{-z}}$$

$$\frac{d}{dz} g(z) = \frac{1}{1+e^{-z}}\left(1 - \frac{1}{1+e^{-z}}\right) = \boxed{g(z)(1-g(z))}$$

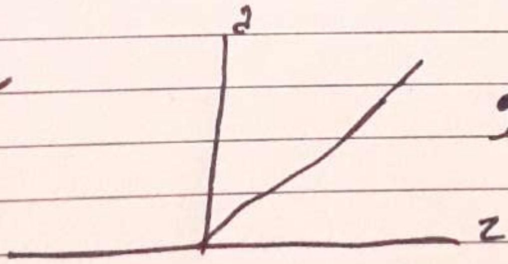
tanh



$$g(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$$

$$g'(z) = \frac{d}{dz} g(z) = \boxed{1 - (\tan(z))^2}$$

ReLU



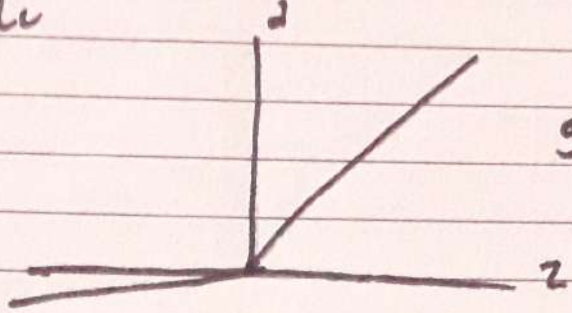
$$g(z) = \max(0, z)$$

$$g'(z) = \begin{cases} 0 & \text{if } z < 0 \\ 1 & \text{if } z \geq 0 \end{cases}$$

Subject :

Date :

Leaky Relu



$g(z) = \max(0.01z, z)$

$$g'(z) = \frac{d}{dz} g(z) = \begin{cases} 0.01 & \text{if } z < 0 \\ 1 & \text{if } z \geq 0 \end{cases}$$

Subject :

Date :

why derive activation ?

forward pass (single layer)

$Z = W \cdot x + b$

$A = f(Z)$; Relu , Sigmoid , tanh , etc

we need $\frac{\partial L}{\partial W}$

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial A} \cdot \frac{\partial A}{\partial Z} \cdot \frac{\partial Z}{\partial W}$$

it tells us how sensitive the activation A is to changes in Z

example

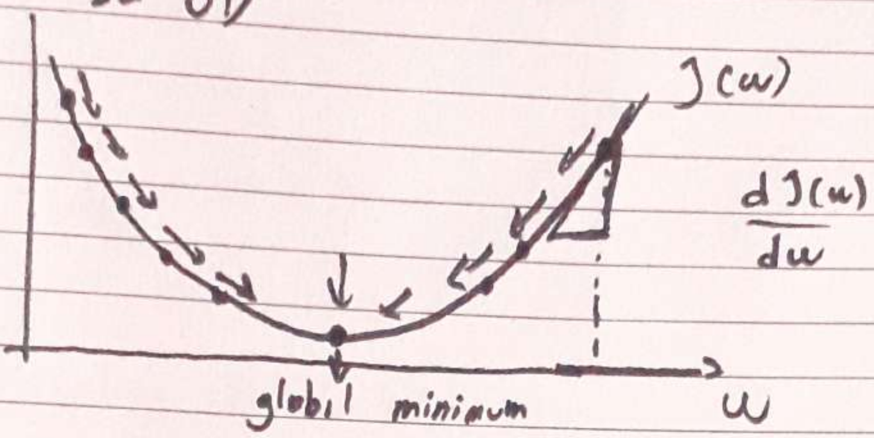
$$\text{Relu}(A) = \max(0, Z) , \quad \frac{\partial A}{\partial Z} \text{ is } \begin{cases} 1 \text{ if } Z > 0 \\ 0 \text{ if } Z \leq 0 \end{cases}$$

Subject :

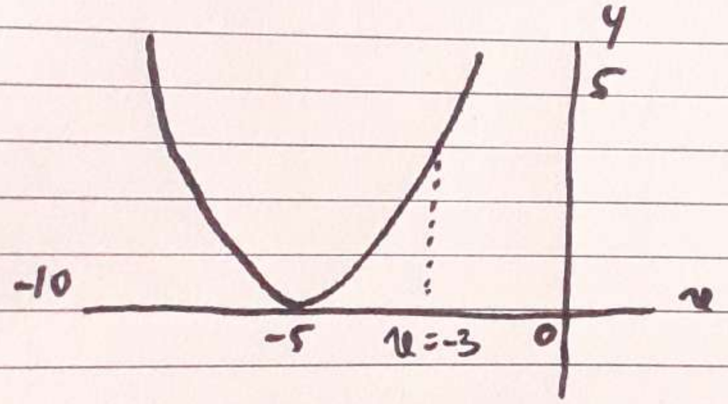
Date :

Gradient Descent :)

1d GD



in this case $w := w - \alpha \frac{dJ(w)}{dw}$



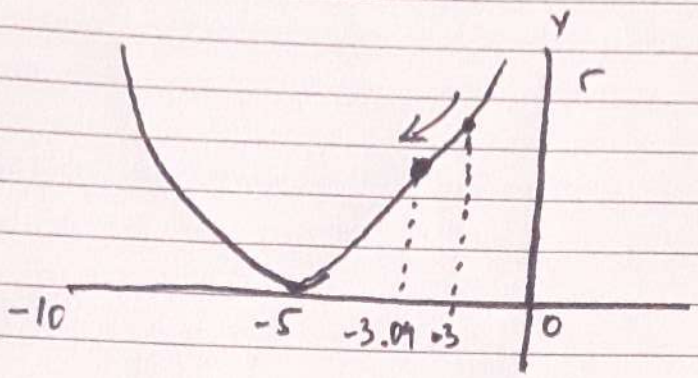
$y = (x+5)^2$

$\frac{dy}{dx} = 2 \cdot (x+5)$, if learning rate $(\alpha) = 0.01$

iteration 1

$$X_1 = X_0 - \alpha \cdot \frac{dy}{dx}$$

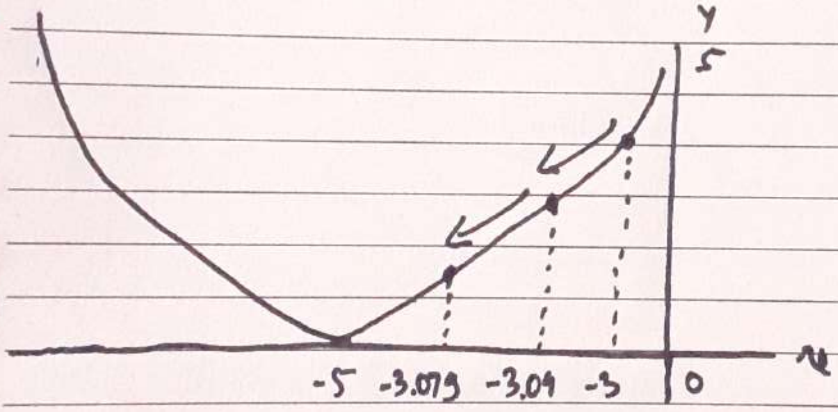
$$= (-3) - (0.01) \cdot (2 \cdot (-3+5)) = -3.04$$



iteration 2

$$X_2 = X_1 - \alpha \cdot \frac{dY}{dx}$$

$$X_2 = (-3.04) - (0.01) \cdot (2 \cdot (-3.04 + 5)) = 3.0792$$



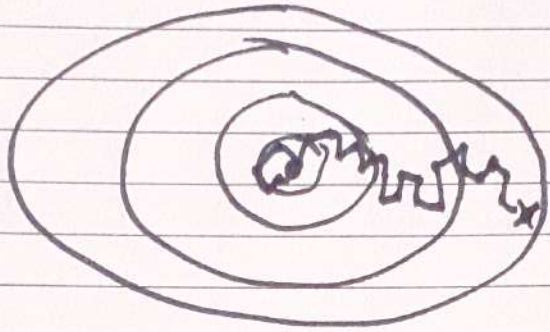
Hafidel

Subject :

Date :

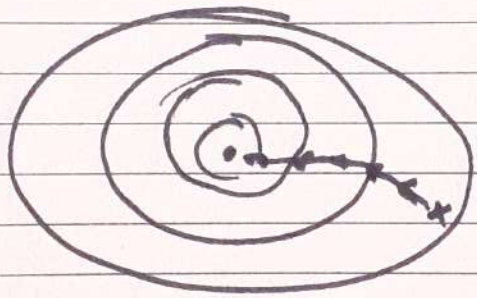
Kinds of Gradient Descent

Stochastic Gradient Descent



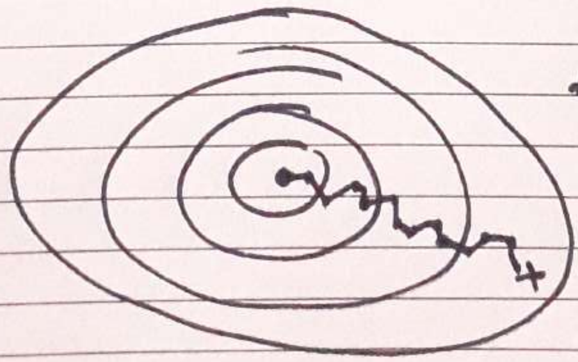
1 random input at a time

Batch Gradient Descent



entire data as an input at a time

Mini - Batch Gradient Descent

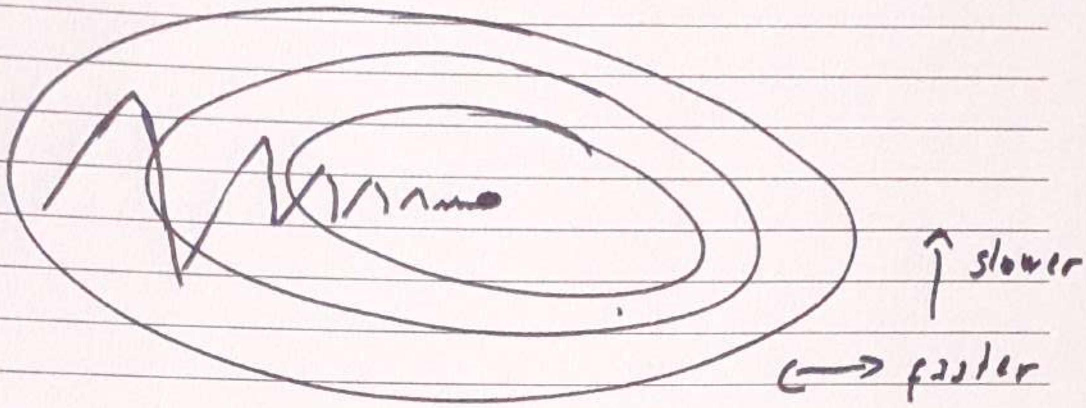


if I have 1.000.000 data, I would use 100.000 data as an input at a time

Subject :

Date :

Momentum

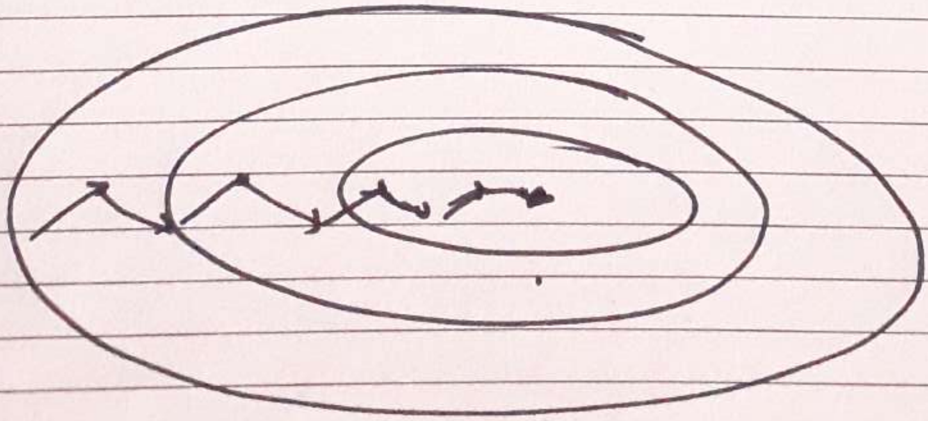


Momentum :

Compute dw, db on current batch / mini-batch

$$V_{dw} = \beta V_{dw} + (1-\beta) dw$$
$$V_{db} = \beta V_{db} + (1-\beta) db$$

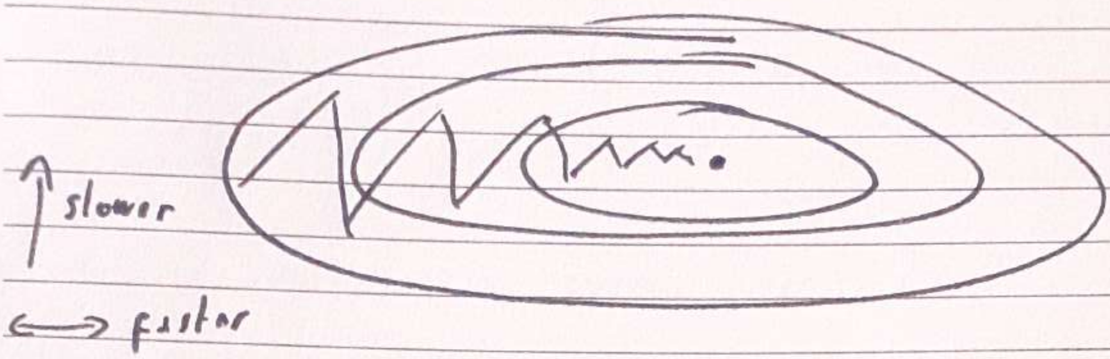
$$w := w - \alpha V_{dw} \quad , \quad b := b - \alpha V_{db}$$



Subject :

Date :

RMS Prop

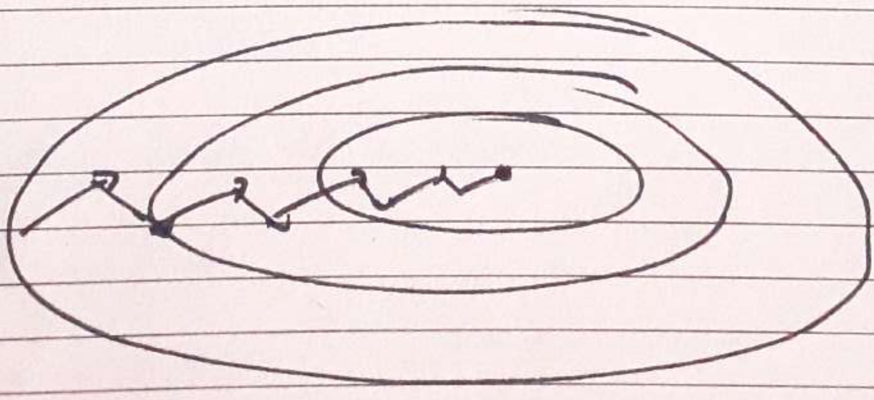


RMS (Root ~~Square~~ Mean Square):

Compute dw, db

$$S_{dw} = \beta S_{dw} + (1-\beta) dw^2 \text{ (element wire square)}$$
$$S_{db} = \beta S_{db} + (1-\beta) db^2$$

$$w := w - \alpha \frac{dw}{\sqrt{S_{dw}+\varepsilon}}, \quad b := w - \alpha \frac{dw}{\sqrt{S_{db}+\varepsilon}}$$



31 lines, 7mm

Subject :

Date :

Adam

$V_{dw} = 0$, $S_{dw} = 0$, $V_{db} = 0$, $S_{db} = 0$

On iteration t:

Compute dw, db using current mini-batch:

$V_{dW} = \beta_1 V_{dw} + (1-\beta_1) dW$, $V_{d} = \beta_1 V_{db} + (1-\beta_1) db$

$S_{dw} : \beta_2 S_{dw} + (1-\beta_2) dw^2$, $S_{db} = \beta_2 S_{db} + (1-\beta_2) db$

$V_{dw}^{corr} = V_{dw} / (1-\beta_1^t)$, $V_{db}^{corr} = V_{db} / (1-\beta_1^t)$

$S_{dw}^{corr} = S_{dw} / (1-\beta_2^t)$, $S_{db}^{corr} = S_{db} / (1-\beta_2^t)$

$$w := w - \alpha \frac{V_{dw}^{corr}}{\sqrt{S_{dw}^{corr}} + \varepsilon}, \quad b := b - \alpha \frac{V_{db}^{corr}}{\sqrt{S_{db}^{corr}} + \varepsilon}$$

α : needs to be tuned

β_1 : 0.9

$\beta_2 = 0.999$

$\varepsilon = 10^{-8}$

31 lines, 7mm

Subject :

Date :

Convolutional

$$\begin{pmatrix} 3 & 0 & 1 & 2 & 7 & 4 \\ 1 & 5 & 8 & 9 & 3 & 1 \\ 2 & 7 & 2 & 5 & 1 & 3 \\ 0 & 1 & 3 & 1 & 7 & 8 \\ 4 & 2 & 1 & 6 & 2 & 8 \\ 2 & 4 & 5 & 2 & 3 & 9 \end{pmatrix} * \begin{pmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{pmatrix}$$

6×6

filter / kernel
3×3

$$\begin{pmatrix} 3 & 0 & 1 \\ 1 & 5 & 8 \\ 2 & 7 & 2 \end{pmatrix} \times \begin{pmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{pmatrix}$$

$$3 \times 1 + 1 \times 1 + 2 \times 1 + 0 \times 0 + 5 \times 0 + 7 \times 0 + -1 \times -1 + 8 \times -1 + 2 \times -1$$
$$= -5$$

$$\begin{pmatrix} -5 & -4 & 0 & 8 \\ -10 & -2 & 2 & 3 \\ 0 & -2 & -4 & -7 \\ -3 & -2 & -3 & -16 \end{pmatrix}$$

4×4

$$\text{Output dim} = n - f + 1 \times n - f + 1$$
$$= 6 - 3 + 1 \times 6 - 3 + 1$$
$$= 4 \times 4 \text{ dim}$$

$\left\{ \begin{array}{l} n = \text{dim input} \\ f = \text{dim filter} \end{array} \right.$

Subject :

Date :

Prewitt Filters

$$\begin{pmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{pmatrix}$$

Horizontal Vertical

Sobel Filters

$$\begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{pmatrix}$$

Horizontal Vertical

Laplacian Filters

$$\begin{pmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{pmatrix}$$

Robinson Compass Masks

$$\begin{pmatrix} -2 & -1 & 0 \\ -1 & 0 & 1 \\ 0 & 1 & 2 \end{pmatrix} \quad \begin{pmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{pmatrix} \quad \begin{pmatrix} 0 & -1 & -2 \\ 1 & 0 & -1 \\ 2 & 1 & 0 \end{pmatrix}$$

north-west north north-east

$$\begin{pmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{pmatrix} \quad \begin{pmatrix} 2 & 1 & 0 \\ 1 & 0 & -1 \\ 0 & -1 & -2 \end{pmatrix} \quad \begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{pmatrix}$$

east south-east south

Subject :

Date :

$$\begin{pmatrix} 0 & 1 & 2 \\ -1 & 0 & 1 \\ -2 & -1 & 0 \end{pmatrix}$$

south - east

$$\begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix}$$

west

Krisch Compass Mask

$$\begin{pmatrix} -3 & 5 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & -3 \end{pmatrix}$$

north - west

$$\begin{pmatrix} -3 & -3 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & 5 \end{pmatrix}$$

north

$$\begin{pmatrix} -3 & -3 & -3 \\ -3 & 0 & 5 \\ -3 & 5 & 5 \end{pmatrix}$$

north - east

$$\begin{pmatrix} -3 & -3 & -3 \\ -3 & 0 & -3 \\ 5 & 5 & 5 \end{pmatrix}$$

east

$$\begin{pmatrix} -3 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & 5 & -3 \end{pmatrix}$$

south - east

$$\begin{pmatrix} 5 & 5 & -3 \\ 5 & 0 & -3 \\ -3 & -3 & -3 \end{pmatrix}$$

south - west

$$\begin{pmatrix} 5 & 5 & 5 \\ -3 & 0 & -3 \\ -3 & -3 & -3 \end{pmatrix}$$

west

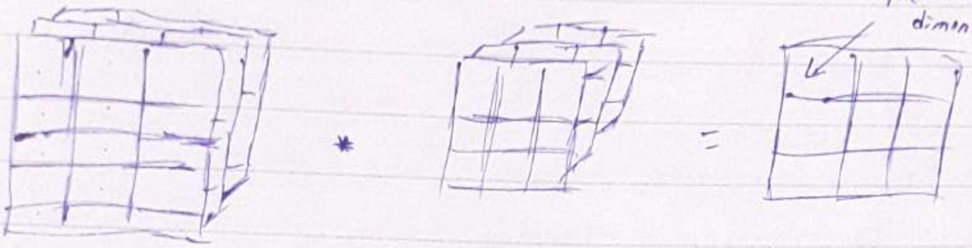
$$\begin{pmatrix} 5 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & -3 & -3 \end{pmatrix}$$

south

Hafidz

Convolution over Volume

the sum of combined dimension



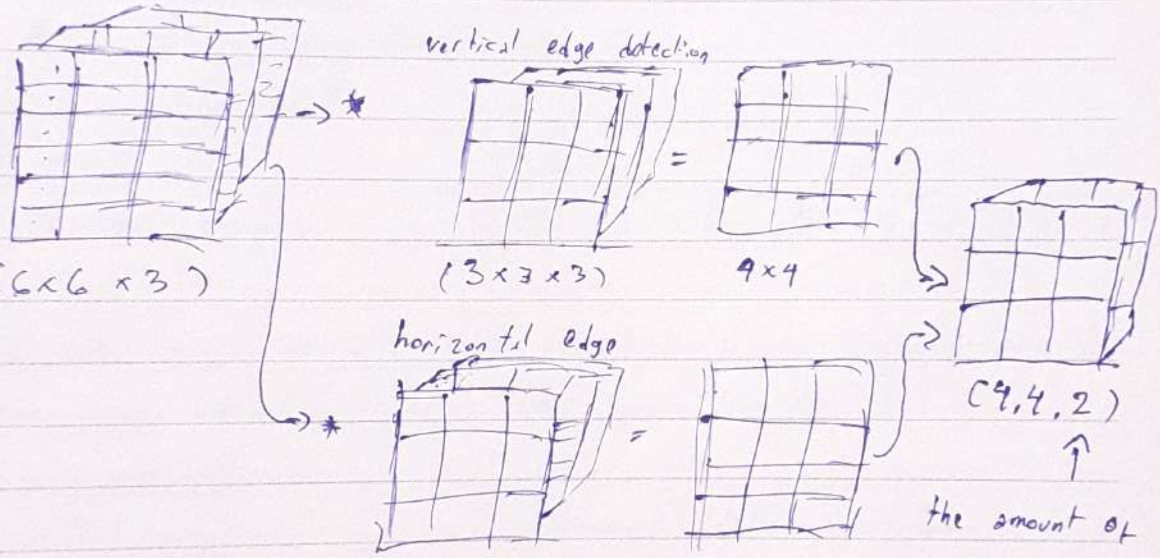
(6×6×3) (3×3×3) (4×4)

RGB dimension

always the same as input dimension

Multiple filters :

vertical edge detection



(6×6×3)

(3×3×3) 4×4

horizontal edge

(4,4,2)

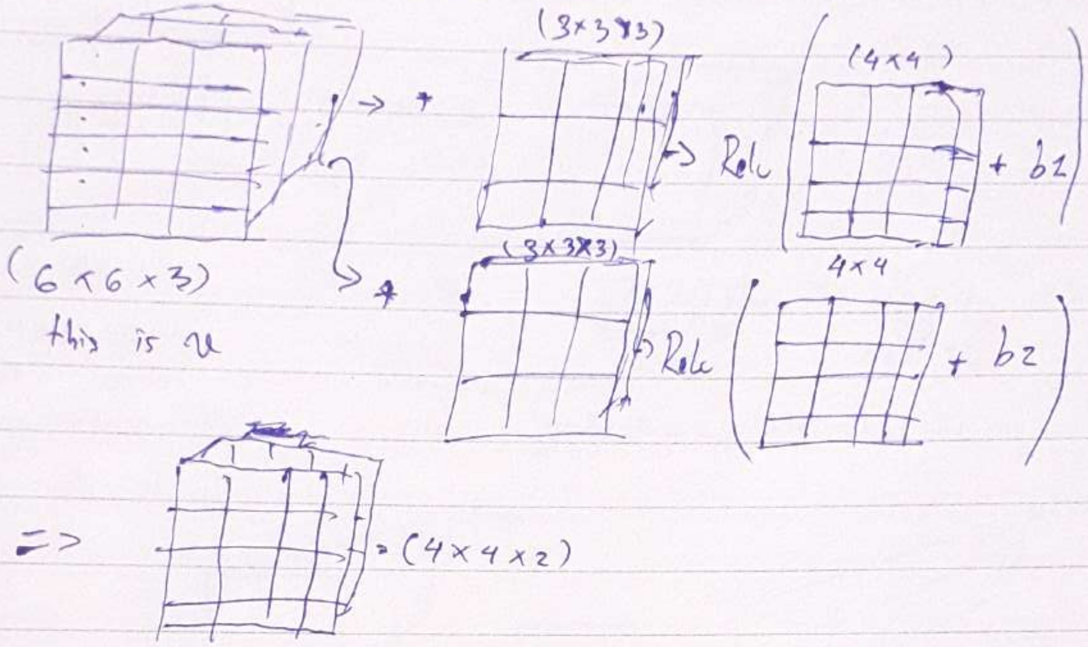
the amount of filter applied

(3×3×3) 4×4

Summari ; $n \times n \times n_{channel}$ * $f \times f \times n_{channel}$ → $n-f+1 \times n-f+1 \times n_{filter}$ (2 in here)

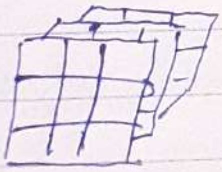
$6 \times 6 \times 3$ $\quad$ $3 \times 3 \times 3$

Example of a layer



if you remember $Z^{[l]} = w^{[l]} x^{[l-1]} + b^{[l]}$, then $a^{[l]} = g(Z^{[l]})$

if you have 10 filters that are 3 x 3 x 3 in 1 layer of a NN. how many parameters does that layer have?



(3x3x3)

= 27 param + bias

= 28 params x 10 filters

= 280 parms

Summary Notation in Conv

$f^{(l)}$ = filter size in layer

$p^{(l)}$ = padding size in layer

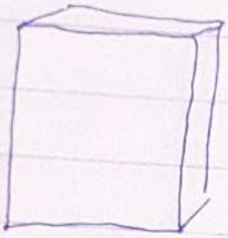
$s^{(l)}$ = Stride size in layer

input : $n_{height}^{(l-1)} \times n_{weight}^{(l-1)} \times n_{channel}^{(l-1)}$

output : $n_h^{(l)} \times n_w^{(l)} \times n_c^{(l)}$

$$n^{(l)} = \left\lfloor \frac{n + 2p - f}{s} + 1 \right\rfloor$$

Example of ConvNet



x

$39 \times 39 \times 3$

$n_{height}^{[0]} \times n_{width}^{[0]} = 39$

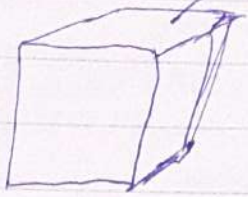
$n_c^{[0]} = 3$

$f^{(1)} = 3 \times 3$

$s^{(1)} = 1$

$p^{(1)} = 0$

10 filters



$a^{(1)}$

$37 \times 37 \times 10 \Rightarrow n_h^{(1)} = n_w^{(1)} = 37$

$n_c^{(1)} = 10$

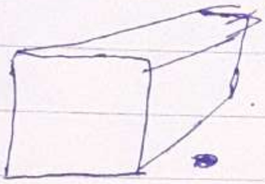
$$\frac{n+2p-f}{s} + 1 = \frac{39+0-3}{1} + 1 = 37$$

$f^{(2)} = 5 \times 3$

$s^{(2)} = 2$

$p^{(2)} = 0$

filters $= 20$



$a(2)$

$17 \times 17 \times 20$

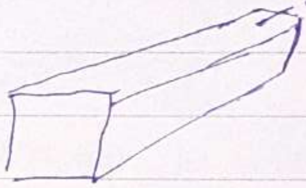
$n_h^{(2)} = n_w^{(2)} = 17$

$n_c^{(2)} = 20$

filter $= 5 \times 5$

$s^{(3)} = 2$

40 filters



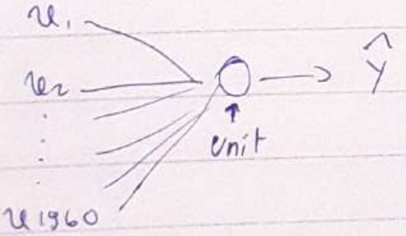
$a(3)$

$7 \times 7 \times 40$

$n_h^{(3)} \times n_w^{(3)} = 7$

$n_c^{(3)} = 40$

$7 \times 7 \times 40 = 1960$



Pooling layer : Max Pooling

1	3	2	1
2	9	1	1
1	3	2	3
5	6	1	2

4×4

find the highest number →

$f = 2 \times 2$

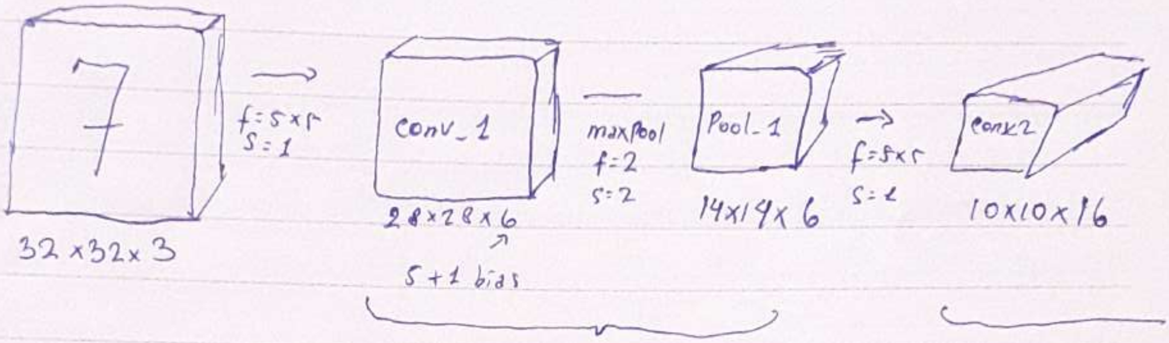
$s = 2$

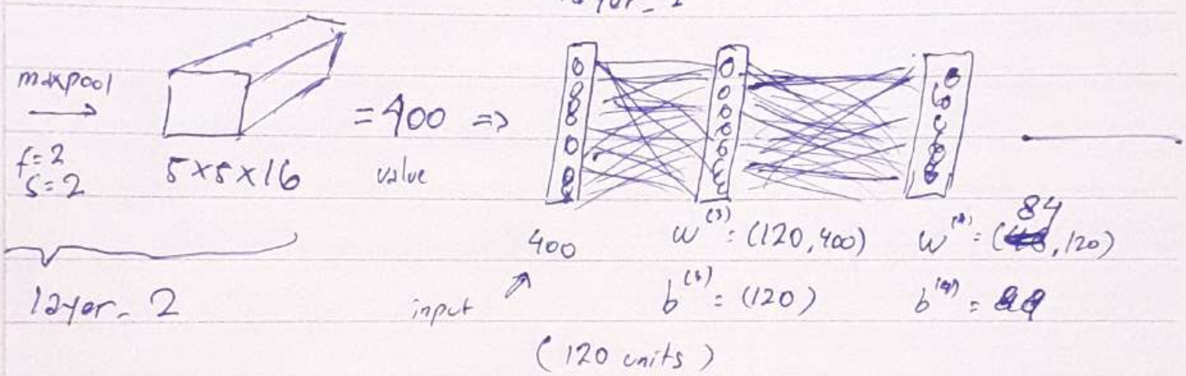
9	2
6	3

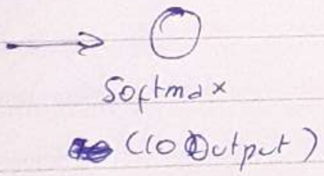
2×2

or you can find the average value, wich is called average pooling

Conv LeNet-5 Architecture





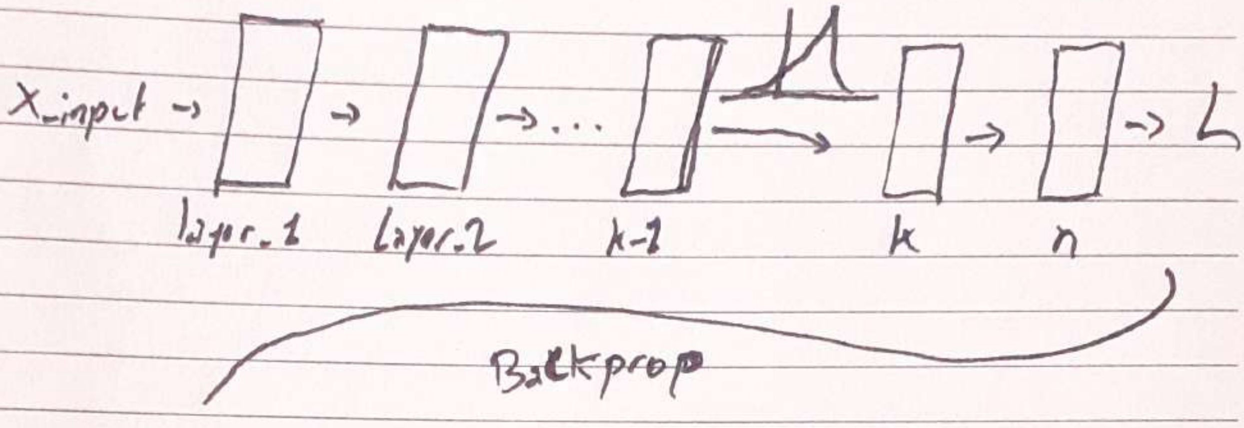


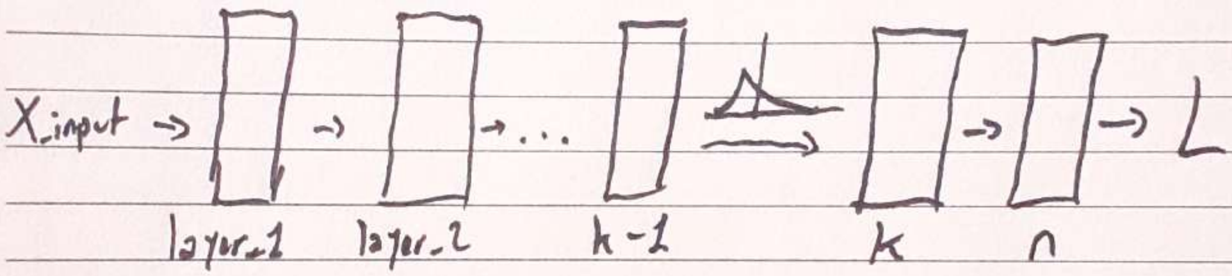
Subject :

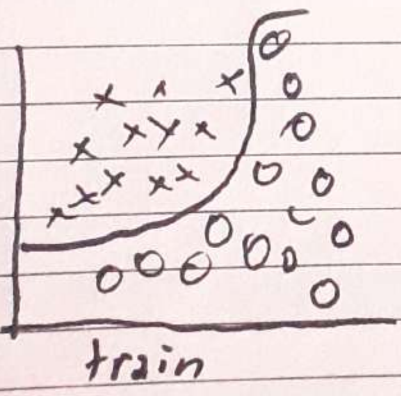
Date :

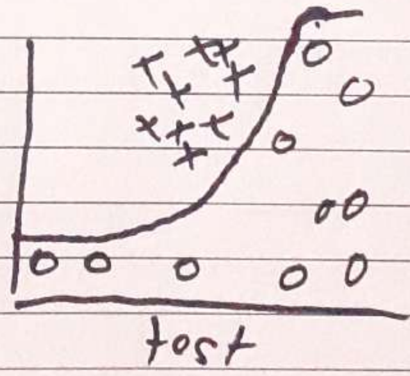
Batch - Normalization

Fixing internal Covariate shift problem









Subject :

Date :

$$\mu_\beta = \frac{1}{n}\sum_{i=1}^{n} x_i$$, mean

$$\sigma_\beta^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu_\beta)^2$$, variance

$$\hat{x}_i = \frac{x_i - \mu_\beta}{\sqrt{\sigma_\beta^2 + \epsilon}}$$, normalize

$$y_i = \gamma \hat{x}_i + \beta$$, scale & shift

input example : [3, 5, 8, 9, 11, 24]

$$\mu_\beta = \frac{1}{6}(3+5+8+9+11+24) = 10$$

$$\sigma^2 = \frac{1}{6}((3-10)^2 + (5-10)^2 + \dots (24-10)^2) = 46$$

$$\hat{x}^{(0)} = \frac{3-10}{\sqrt{46^2 + 0.00001}} = -1.03$$

$$\hat{x}^{(1)} = \frac{5-10}{\sqrt{46^2 + 0.00001}} = -0.75$$

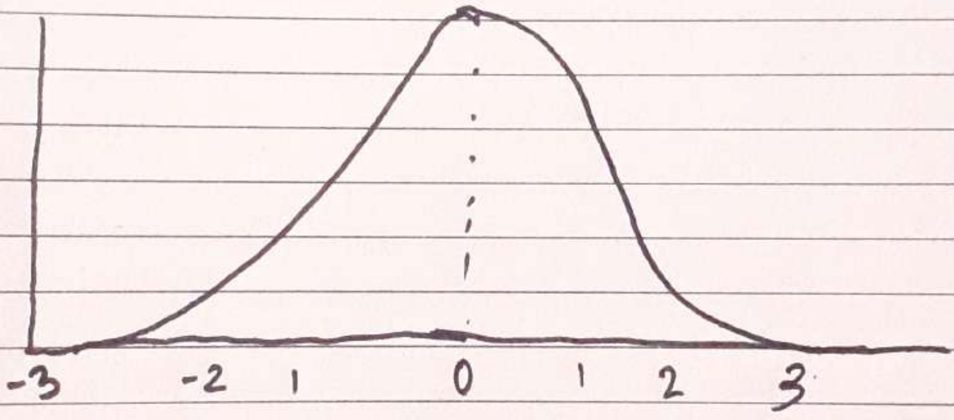
⋮

31 lines, 7mm

$$x^{(5)} = \frac{24-10}{\sqrt{46^2 + 0.00001}} = 2.06$$

output example = [-1.03, -0.74, -0.29, -0.15, 0.15, 2.06]

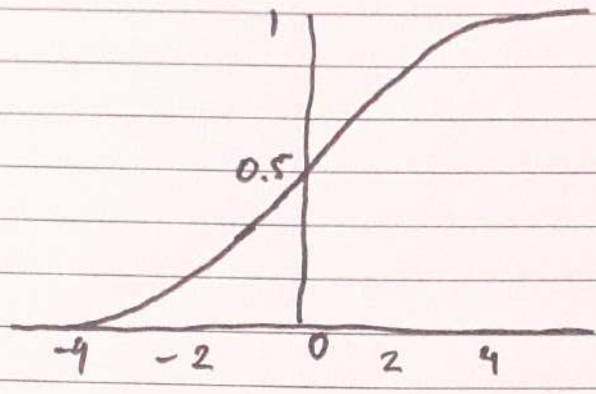
mean = 0
Std = 0.998



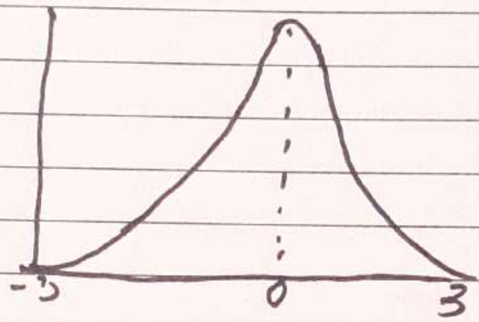
Subject :

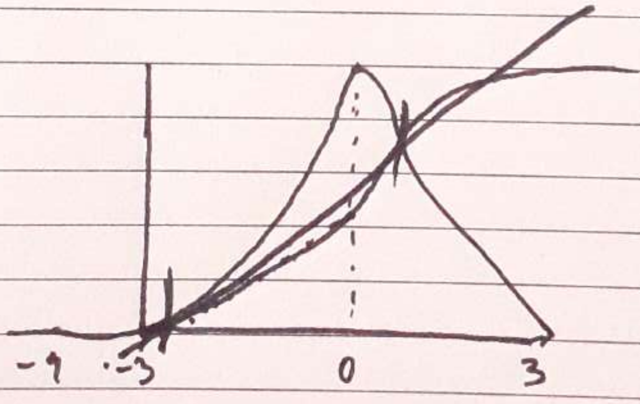
Date :

Sigmoid



output example





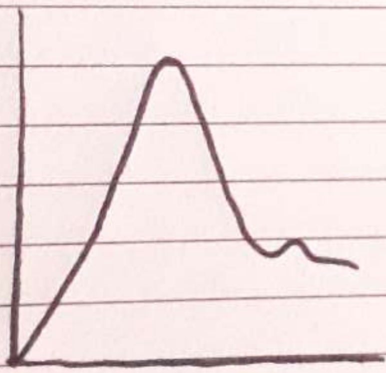
The distribution makes the non-linear function become useless

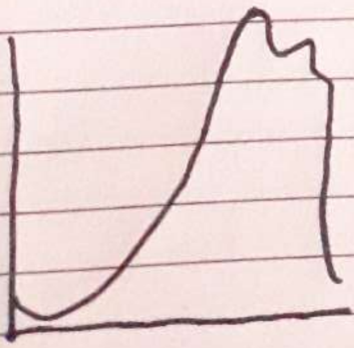
31 lines, 7mm

Subject :

Date :

$\downarrow$ scale & shifting



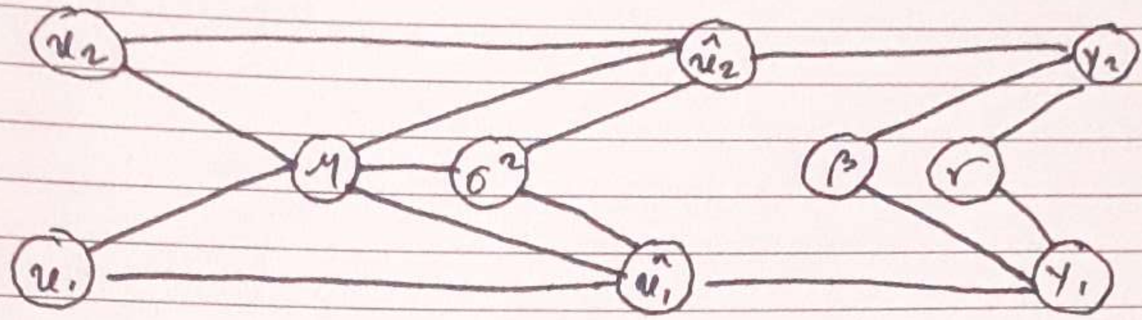


etc.

Subject :

Date :

BN Computation Graph



$$\frac{\partial L}{\partial \hat{x}_i} = ? \qquad \frac{\partial L}{\partial \mu} = ? \qquad \frac{\partial L}{\partial \gamma} = ?$$

$$\frac{\partial L}{\partial \sigma^2} = ? \qquad \frac{\partial L}{\partial x_i} = ? \qquad \frac{\partial L}{\partial \beta} = ?$$

1. $$\frac{\partial L}{\partial \beta} = \underbrace{\frac{\partial L}{\partial y_1} \cdot \frac{\partial y_1}{\partial \beta} + \frac{\partial L}{\partial y_2}}_{\text{known}} \cdot \frac{\partial y_2}{\partial \beta}$$

$$\frac{\partial y_1}{\partial \beta} = \frac{\partial (\gamma \hat{x}_1 + \beta)}{\partial \beta}$$

$$\frac{\partial y_1}{\partial \beta} = 1 \quad ; \quad \frac{\partial y_2}{\partial \beta} = 1$$

$$\frac{\partial L}{\partial \beta} = \frac{\partial L}{\partial y_1} \cdot 1 + \frac{\partial L}{\partial y_2} \cdot 1$$

$$= \frac{\partial L}{\partial y_1} + \frac{\partial L}{\partial y_2}$$

$$\boxed{\frac{\partial L}{\partial \beta} = \sum_{i=1}^{n} \frac{\partial L}{\partial y_i}}$$

31 lines, 7mm

Subject :

Date :

2. $$\frac{\partial L}{\partial r} = \underbrace{\frac{\partial L}{\partial y_1} \cdot \frac{\partial y_1}{\partial r} + \frac{\partial L}{\partial y_2}}_{\text{known}} \cdot \frac{\partial y_2}{\partial r}$$

$$\frac{\partial y_1}{\partial r} = \frac{\partial}{\partial r}(r\hat{u}_1 + \beta)$$

$$\frac{\partial y_1}{\partial r} = \hat{u}_1 \quad ; \quad \frac{\partial y_2}{\partial r} = \hat{u}_2$$

$$\frac{\partial L}{\partial r} = \frac{\partial L}{\partial y_1} \cdot \hat{u}_1 + \frac{\partial L}{\partial y_2} \cdot \hat{u}_2$$

$$\boxed{\frac{\partial L}{\partial r} = \sum_{i=1}^{n} \frac{\partial L}{\partial y_i} \cdot \hat{u}_i}$$

3. $\frac{\partial L}{\partial \hat{u}_1}$ and $\frac{\partial L}{\partial \hat{u}_2}$

$$\frac{\partial L}{\partial \hat{u}_1} = \frac{\partial L}{\partial y_1} \cdot \frac{\partial y_1}{\partial \hat{u}_1}$$

$$= \frac{\partial L}{\partial y_1} \cdot \frac{\partial}{\partial \hat{u}_1}(r\hat{u}_1 + \beta)$$

$$\frac{\partial L}{\partial \hat{u}_1} = \frac{\partial L}{\partial y_1} \cdot r \quad ; \quad \frac{\partial L}{\partial \hat{u}_2} = \frac{\partial L}{\partial y_2} \cdot r$$

$$\boxed{\frac{\partial L}{\partial \hat{u}_i} = \frac{\partial L}{\partial y_i} \cdot r}$$

31 lines, 7mm

Subject :

Date :

4. $$\frac{\partial L}{\partial \sigma^2} = \underbrace{\frac{\partial L}{\partial \hat{u}_1} \cdot \frac{\partial \hat{u}_1}{\partial \sigma^2} + \frac{\partial L}{\partial \hat{u}_2}}_{\text{known}} \cdot \frac{\partial \hat{u}_2}{\partial \sigma^2}$$

$$\frac{\partial \hat{u}_1}{\partial \sigma^2} = \frac{\partial}{\partial \sigma^2}\left(\frac{u_1 - \mu}{\sqrt{\sigma^2+\varepsilon}}\right)$$

$$= (u_1 - \mu) \cdot \frac{-1}{(\sqrt{\sigma^2+\varepsilon})^2} \cdot \frac{\partial}{\partial \sigma^2}(\sqrt{\sigma^2+\varepsilon})$$

$$= \frac{-(u_1 - \mu)}{\sigma^2+\varepsilon} \cdot \frac{1}{2\sqrt{\sigma^2+\varepsilon}} \cdot \frac{\partial}{\partial \sigma^2}(\sigma^2+\varepsilon)$$

$$= -\frac{1}{2} \cdot \frac{(u_1 - \mu)}{\sigma^2+\varepsilon} \cdot \frac{1}{(\sigma^2+\varepsilon)^{1/2}}$$

$$= -\frac{1}{2} \cdot (u_1 - \mu) \cdot \frac{1}{(\sigma^2+\varepsilon)^{3/2}}$$

$$\frac{\partial \hat{u}_1}{\partial \sigma^2} = -\frac{1}{2} \cdot (u_1 - \mu) \cdot (\sigma^2+\varepsilon)^{-3/2}$$

$$\frac{\partial L}{\partial \sigma^2} = \frac{\partial L}{\partial \hat{u}_1} \cdot \frac{-1}{2} \cdot (u_1 - \mu) \cdot (\sigma^2+\varepsilon)^{-3/2} + \frac{\partial L}{\partial \hat{u}_2} \cdot \frac{-1}{2} \cdot (u_2 - \mu) \cdot (\sigma^2+\varepsilon)^{-3/2}$$

$$\frac{\partial L}{\partial \sigma^2} = \sum_{i=1}^{n} \frac{\partial L}{\partial \hat{u}_i} \cdot \frac{-1}{2} \cdot (u_i - \mu) \cdot (\sigma^2+\varepsilon)^{-3/2}$$

31 lines, 7mm

Subject :

Date :

5. $$\frac{\partial L}{\partial \mu} = \frac{\partial L}{\partial \hat{x}_1} \cdot \frac{\partial \hat{x}_1}{\partial \mu} + \frac{\partial L}{\partial \hat{x}_2} \cdot \frac{\partial \hat{x}_2}{\partial \mu} + \frac{\partial L}{\partial \sigma^2} \cdot \frac{\partial \sigma^2}{\partial \mu}$$

known

$$\frac{\partial \hat{x}_1}{\partial \mu} = \frac{\partial}{\partial \mu}\left(\frac{x_1 - \mu}{\sqrt{\sigma^2+\varepsilon}}\right)$$

$$= \frac{1}{\sqrt{\sigma^2+\varepsilon}} \cdot (-1)$$

$$\frac{\partial \hat{x}_1}{\partial \mu} = \frac{-1}{\sqrt{\sigma^2+\varepsilon}} \quad ; \quad \frac{\partial \hat{x}_2}{\partial \mu} = \frac{-1}{\sqrt{\sigma^2+\varepsilon}}$$

$$\frac{\partial L}{\partial \hat{x}_1} \cdot \frac{-1}{\sqrt{\sigma^2+\varepsilon}} + \frac{\partial L}{\partial \hat{x}_2} \cdot \frac{-1}{\sqrt{\sigma^2+\varepsilon}}$$

$$\Rightarrow \sum_{i=1}^{n} \frac{\partial L}{\partial \hat{x}_i} \cdot \frac{-1}{\sqrt{\sigma^2+\varepsilon}}$$

$$\frac{\partial \sigma^2}{\partial \mu} = \frac{\partial}{\partial \mu}\left[\frac{(x_1-\mu)^2 + (x_2-\mu)^2}{2\,(n \text{ amount})}\right]$$

$$= \frac{\partial}{\partial \mu}\left(\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \mu)^2\right)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \mu} (x_i - \mu)^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} 2 \cdot (x_i - \mu) \cdot \frac{\partial}{\partial \mu} (x_i - \mu)$$

Subject :

Date :

$$= \frac{1}{n} \sum_{i=1}^{n} 2 \cdot (x_i - \mu) \cdot (-1)$$

$$\frac{\partial \sigma^2}{\partial \mu} = \frac{1}{n} \cdot \sum_{i=1}^{n} -2 \cdot (x_i - \mu)$$

$$\frac{\partial L}{\partial \mu} = \sum_{i=1}^{n} \frac{\partial L}{\partial \hat{x}_i} \cdot \frac{-1}{\sqrt{\sigma^2 + \varepsilon}} + \frac{\partial L}{\partial \sigma^2} \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} -2 \cdot (x_i - \mu)$$

6. $\frac{\partial L}{\partial x_1}$ and $\frac{\partial L}{\partial x_2}$

$$\frac{\partial L}{\partial x_1} = \underbrace{\frac{\partial L}{\partial \hat{x}_1}}_{\text{known}} \cdot \frac{\partial \hat{x}_1}{\partial x_1} + \underbrace{\frac{\partial L}{\partial \sigma^2}}_{\text{known}} \cdot \frac{\partial \sigma^2}{\partial x_1} + \underbrace{\frac{\partial L}{\partial \mu}}_{\text{known}} \cdot \frac{\partial \mu}{\partial x_1}$$

$$\frac{\partial \hat{x}_1}{\partial x_1} = \frac{\partial}{\partial x_1} \left(\frac{x_1 - \mu}{\sqrt{\sigma^2 + \varepsilon}} \right)$$

$$= \frac{1}{\sqrt{\sigma^2 + \varepsilon}} \cdot \frac{\partial}{\partial x_1} (x_1 - \mu)$$

$$= \frac{1}{\sqrt{\sigma^2 + \varepsilon}} \cdot 1$$

$$\frac{\partial \hat{x}_1}{\partial x_1} = \frac{1}{\sqrt{\sigma^2 + \varepsilon}}$$

31 lines, 7mm

Subject :

Date :

$$\frac{\partial \sigma^2}{\partial x_1} = \frac{\partial}{\partial x_1}\left(\frac{(x_1-\mu)^2+(x_2-\mu)^2}{2(n \text{ amount})}\right)$$

$$= \frac{1}{2}\cdot\frac{\partial}{\partial x_1}(x_1-\mu)^2$$

$$= \frac{1}{2}\left(2\cdot(x_1-\mu)\cdot\frac{\partial}{\partial x_1}(x_1-\mu)\right)$$

$$= \frac{1}{2}\left(2\cdot(x_1-\mu)\cdot 1\right)$$

$$\frac{\partial \sigma^2}{\partial x_1} = \frac{1}{2}\left(2\cdot(x_1-\mu)\right)$$

n amount

$$\frac{\partial \sigma^2}{\partial x_1} = \frac{2(x_i-\mu)}{n}$$

$$\frac{\partial \mu}{\partial x_1} = \frac{\partial}{\partial x_1}\left(\frac{x_1+x_2}{2}\right)$$

n

$$= \frac{1}{2}(1)$$

$$\frac{\partial \mu}{\partial x_1} = \frac{1}{n}$$

$$\frac{\partial L}{\partial x_i} = \frac{\partial L}{\partial \hat{x}_i}\cdot\frac{1}{\sqrt{\sigma^2+\epsilon}} + \frac{\partial L}{\partial \sigma^2}\cdot\frac{2(x_i-\mu)}{n} + \frac{\partial L}{\partial \mu}\cdot\frac{1}{n}$$

Subject :

Date :

Derivatives of BN

$$\frac{\partial L}{\partial \hat{x}_i} = \frac{\partial L}{\partial y_i} \cdot \gamma$$

$$\frac{\partial L}{\partial \sigma^2_\beta} = \sum_{i=1}^{n} \frac{\partial L}{\partial \hat{x}_i} \cdot (x_i - \mu_\beta) \cdot \frac{-1}{2} (\sigma^2_\beta + \varepsilon)^{-3/2}$$

$$\frac{\partial L}{\partial \mu_\beta} = \left(\sum_{i=1}^{n} \frac{\partial L}{\partial \hat{x}_i} \cdot \frac{-1}{\sqrt{\sigma^2 + \varepsilon}} \right) + \frac{\partial L}{\partial \sigma^2_\beta} \cdot \sum_{i=1}^{n} -2(x_i - \mu_\beta) \cdot \frac{1}{n}$$

$$\frac{\partial L}{\partial x_i} = \frac{\partial L}{\partial \hat{x}_i} \cdot \frac{1}{\sqrt{\sigma^2_\beta + \varepsilon}} + \frac{\partial L}{\partial \sigma^2_\beta} \cdot \frac{2(x_i - \mu_\beta)}{n} + \frac{\partial L}{\partial \mu_\beta} \cdot \frac{1}{n}$$

$$\frac{\partial L}{\partial \gamma} = \sum_{i=1}^{n} \frac{\partial L}{\partial y_i} \cdot \hat{x}_i$$

$$\frac{\partial L}{\partial \beta} = \sum_{i=1}^{n} \frac{\partial L}{\partial y_i}$$

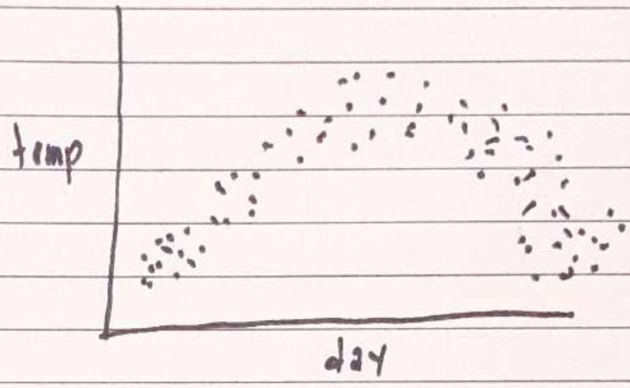
Subject :

Date :

Exponentially Moving Average

day & temp = θ

$\theta_1 = 4°C$
$\theta_2 = 9°C$
$\theta_3 = 5°C$
$\vdots$
$\theta_{180} = 15°C$
$\vdots$
θ_n



$$\boxed{EMA \Rightarrow V_t = \beta V_{t-1} + (1-\beta)\theta_t}$$

$V_0 = 0$
$V_1 = 0.9\, V_0 + 0.1\, \theta_1$
$V_2 = 0.9\, V_1 + 0.1\, \theta_2$
$V_3 = 0.9\, V_2 + 0.1\, \theta_3$
$\vdots$

$V_t = 0.9\, V_{t-1} + 0.1\, \theta_t$

31 lines, 7mm

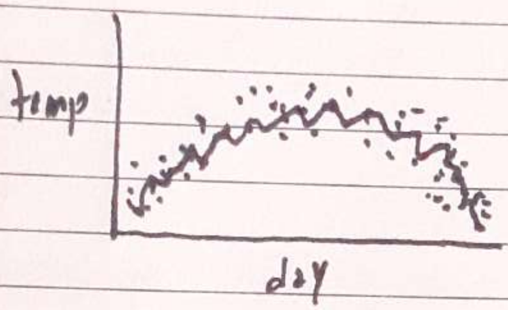
Subject :

Date :

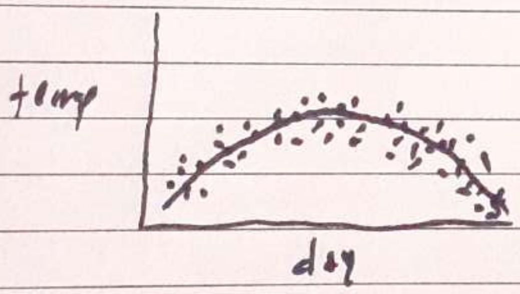
$V_t = \beta V_{t-1} + (1-\beta)\theta_t$

$\boxed{V_t \approx \frac{1}{1-\beta}}$

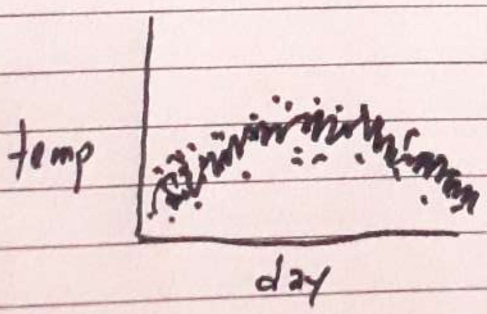
$\beta = 0.9$: $\approx$ 10 day



$\beta = 0.98$: $\approx$ 50 days $\approx> \frac{1}{1-0.98} = 50$



$\beta = 0.5$: $\approx$ 2 days



31 lines, 7mm

Subject :

Date :

EmA in BN

- Calculate mean & variance

  - Run entire training dataset once and calculate mean & variance at each neuron and save them

  - Calculate EmA of mean & variance during training, and save the last values after training

moving_mean = moving_mean · momentum + batch_mean · (1-momentum)

moving_var = moving_batch · momentum + batch_var · (1-momentum)

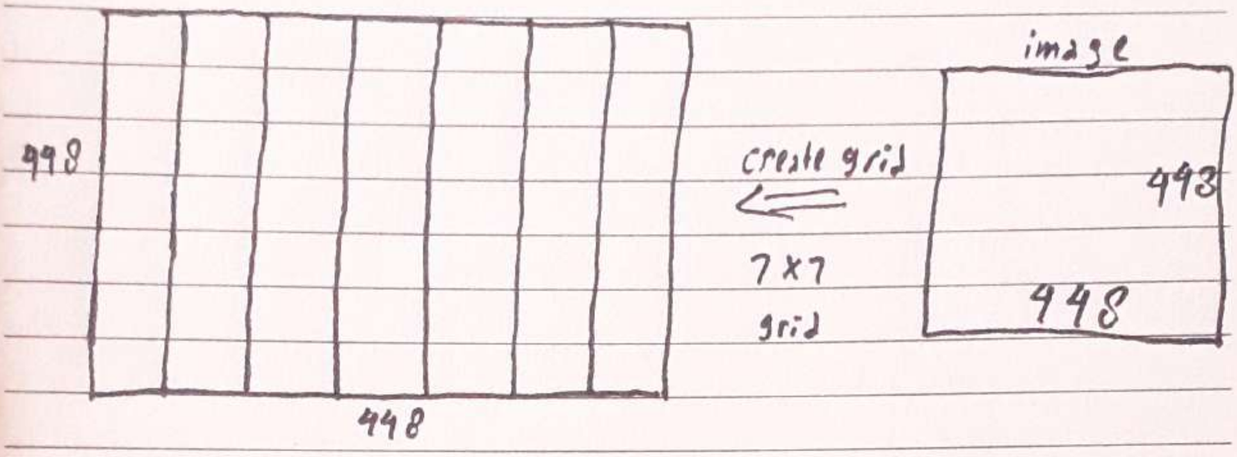
$$\boxed{V_t = \beta V_{t-1} + (1-\beta)\theta_t}$$

1 lines, 7mm

Subject :

Date :

YOLO



448×448 image with 7×7 grid => 1 grid has 64 pixel

Bounding Box

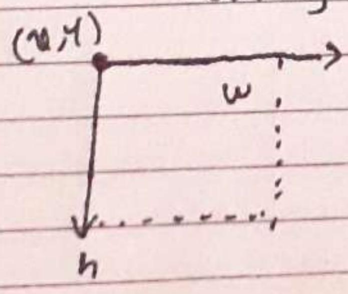
- Center Point (x, y) : relatives to BB

$$\Delta x = (x - x_a) / 64$$
$$\Delta y = (y - y_a) / 64$$

- Width / height (w, h) : relative to the whole image (448×448)

$$\Delta w : w / 448$$
$$\Delta h : h / 448$$

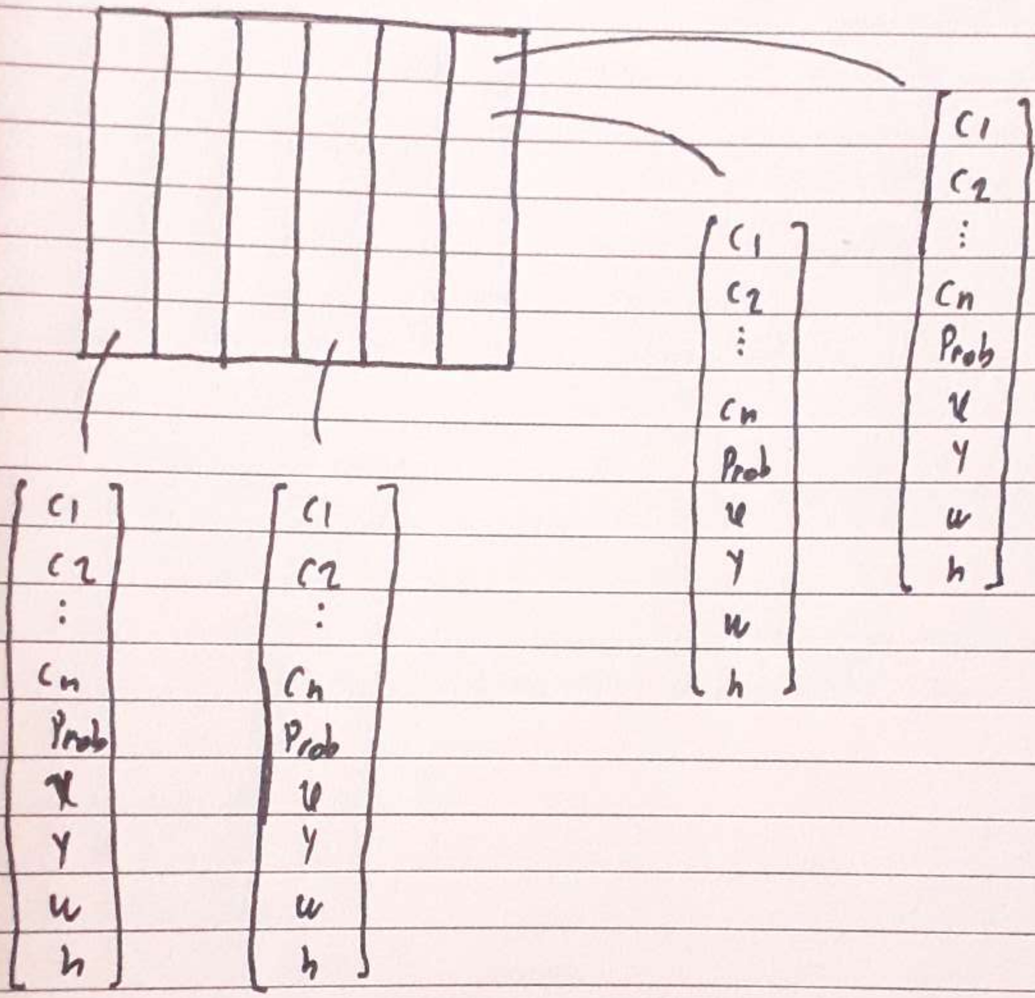
Bounding Box



Subject :

Date :

Yolo Label



Each grid has it's own label !!

Subject :

Date :

YOLO LOSS !!!

$$L = \underbrace{\sum_{i=1}^{S\times S} \mathbb{1}_i^{obj} \cdot L_{i,obj}}_{\text{object exist}} + \underbrace{\lambda_{no\_obj} \sum_{i=1}^{S\times S} \mathbb{1}_i^{no\_obj} , L_{i,no\_obj}}_{\text{no object}}$$

$$L_{i,obj} = L_{i,obj}^{box} + L_{i,obj}^{conf} + L_{i,obj}^{cls} \cdot \lambda$$

$$L = \lambda_{co} \sum_{i=0}^{S\times S} \sum_{j=0}^{B} \mathbb{1}_{ij}^{obj} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right]$$

$$+ \lambda_{co} \sum_{i=0}^{S\times S} \sum_{j=0}^{B} \mathbb{1}_{ij}^{obj} \left[\left(\sqrt{w_i} - \sqrt{\hat{w}_i}\right)^2 + \left(\sqrt{h_i} - \sqrt{\hat{h}_i}\right)^2 \right]$$

$$+ \sum_{i=0}^{S\times S} \sum_{j=0}^{B} \mathbb{1}_{ij}^{obj} (C_i - \hat{C}_i)^2$$

$$+ \lambda_{no\_obj} \sum_{i=0}^{S\times S} \sum_{j=0}^{B} \mathbb{1}_{i,j}^{no\_obj} (C_i - \hat{C}_i)^2$$

$$+ \sum_{i=0}^{S\times S} \mathbb{1}_i^{obj} \sum_{c \in class} (p_i(c) - \hat{p}_i(c))^2$$

pred

⇓

$$\begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_{20} \\ P_{c_1} \\ x \\ y \\ w \\ h \end{bmatrix}$$

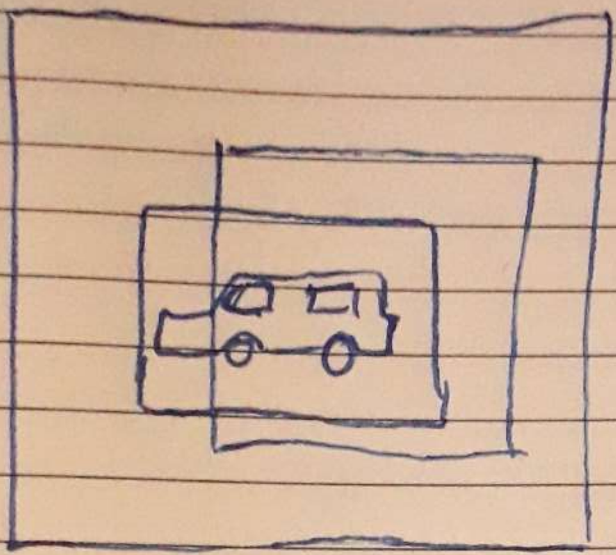
$$\text{Prediction} = [\underbrace{c1, c2, c3, \ldots c20}_{\text{classes}}, P_{c1}, \underbrace{x, y, w, h}_{\text{box 1}}, P_{c2}, \underbrace{x, y, w, h}_{\text{box 2}}]$$

Probability

Subject :

Date :

Intersection Over Union (IoU)



$$\frac{\text{Intersection}}{\text{Union}} = \frac{\text{Size of intersection}}{\text{Size of union}}$$

"Correct" if $\text{IoU} \geq 0.5$

31 lines, 7mm

Mean Average Precision (mAP)

- Confusion Matrix
  - True positive => prediction label is equal to the class, and the ground truth label is equal to class
  - True negative => prediction label is the class while the ground truth is not the class
  - False positive => prediction label is the class and the ground truth is not the class
  - False Negative => The prediction label is not the class and the ground truth label is the class

- Precision & Recall

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive} + \text{False positive}} \qquad \text{Recall} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}}$$

s, 7mm

Date

$$\text{Mean Average Precision} = \frac{1}{n} \sum_{k=1}^{k=n} AP_k$$

n = num classes
AP_k: the average precision of class k

s, 7mm