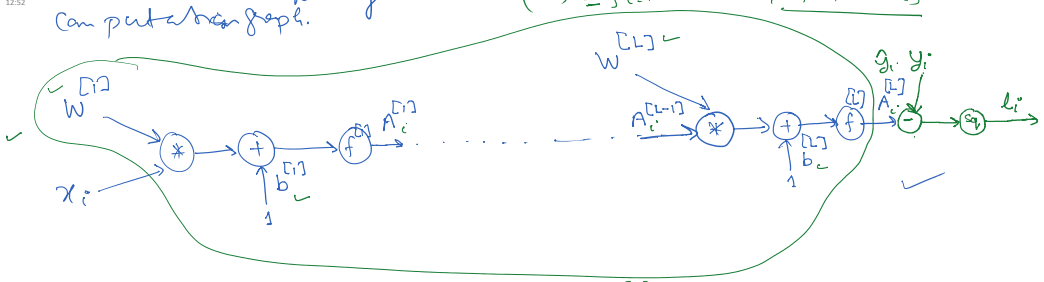
31 July 2025 12:52

Computation graph. L - layer $X = \{x_i, y_i\}_{i=1}^{N}$ $C: \{C_1, C_2, \cdots C_3\}$



$h(x_i) \rightarrow A_i^{[L]} = \hat{y}_i$

$l_i = (y_i - \hat{y}_i)^2$

Learning parameters

$\theta^* = \langle W^{[l]}, b^{[l]} \rangle$ .. for all $l = 1$ to L.

hyper-parameters.

Avg Absolute Error $= \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i|$

— Mean Square Error $= \frac{1}{2N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2$

If y_i is discrete then Cross Entropy $= -\sum_{c \in C_i}\sum_{i} y_i \cdot \log \hat{y}_i$

$$\underset{\theta}{\text{minimize}} \; L_\theta(y_i, \hat{y}_i)$$

$\nabla_\theta L = \frac{\partial L_\theta}{\partial \theta} \rightarrow 0 \Rightarrow$ closed form soln

$$W^{[l]} \leftarrow W^{[l]} - \eta \cdot \frac{\partial L_{W,b}}{\partial W}$$

$$b^{[l]} \leftarrow b^{[l]} - \eta \cdot \frac{\partial L_{W,b}}{\partial b^{[l]}}$$

Iterative optimization.

→ start with random $\theta = \theta^0$ (iteration i=0)

→ Untill Convergence { $\theta \leftarrow \theta - \eta \frac{\partial L_\theta}{\partial \theta}$ } (learning rate, parameter update)

→ Now you have θ^*.

If L_θ is convex in θ optimal soln

Otherwise: approximate soln = Acceptable soln

$g(x, y)$ $g(x_1, \ldots)$ Inputs

$\frac{\partial g}{\partial x_p} = g(x)$

$g(x) = x^2$ $g'(x) = 2x$ $x \in f(x)$

$\frac{dg}{dx} = \lim_{\epsilon \to 0} \frac{g(x+\epsilon) - g(x-\epsilon)}{2\epsilon} \rightarrow 0$

$x + \epsilon$: 1 .01 .01

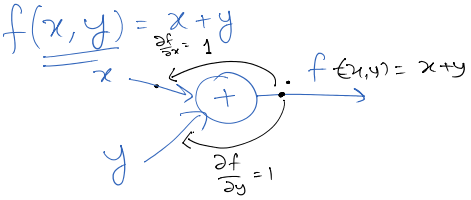
Gradient Checking

$\left\langle \frac{dg}{dx}, \frac{\partial g}{\partial y} \right\rangle$

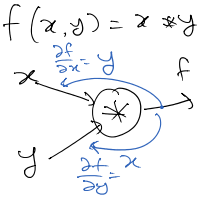
$\lim_{\epsilon \to 0} \frac{g(x+\epsilon) - g(x-\epsilon)}{2\epsilon}$ rated close using limit

x	ϵ	$g'(x)$	
2	.01	4	$\frac{(2.01)^2 - (1.99)^2}{2 \times 0.01} = 4$
2	.001	4	$= 4$

$f(x, y) = x + y$ $\frac{\partial f}{\partial x} = 1$ $\frac{\partial f}{\partial y} = 1$



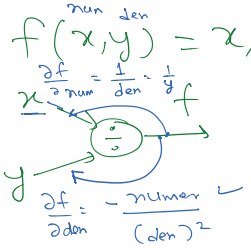
$f(x, y) = x * y$ $\frac{\partial f}{\partial x} = y$ $\frac{\partial f}{\partial y} = x$



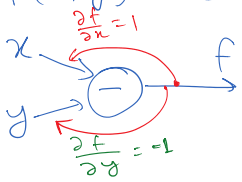
$\frac{\partial f}{\partial x} = 1$

$\frac{\partial f}{\partial y} = 1$

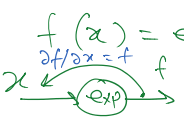
$f(x,y) = x/y = x \cdot y^{-1}$



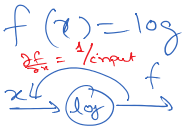
$f(x,y) = x - y$



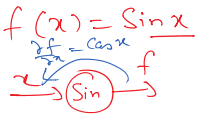
$f(x) = e^x$



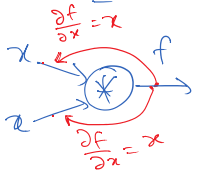
$f(x) = \log x$



$f(x) = \sin x$

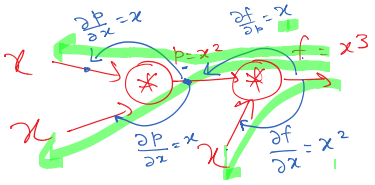


$f(x) = x^2 = x * x$



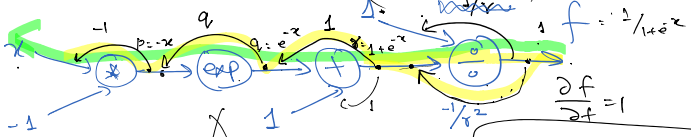
$\frac{\partial f}{\partial x} = [x + x] = 2x$

$f(x) = x^3 = x^2 * x = x * x * x$



$\frac{\partial f}{\partial x} = x^2 + x^2 + x^2 = 3x^2$

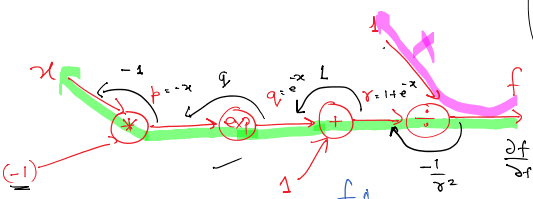
$f(x) = \frac{1}{1+e^{-x}}$ $\frac{\partial f}{\partial x}$



$\frac{\partial f}{\partial x} = 1 \times \left(\frac{1}{g^2}\right) \times 1 \times q \times (-1) = \frac{e^{-x}}{(1+e^{-x})^2} = f(1-f)$

$= 1 \times \left(-\frac{1}{g^2}\right) \times 1 \times q \times (-1) = \left(\frac{1}{1+e^{-x}}\right)\left(1 - \frac{1}{1+e^{-x}}\right)$

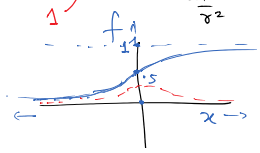
$f(x)$ $1 - f(x)$



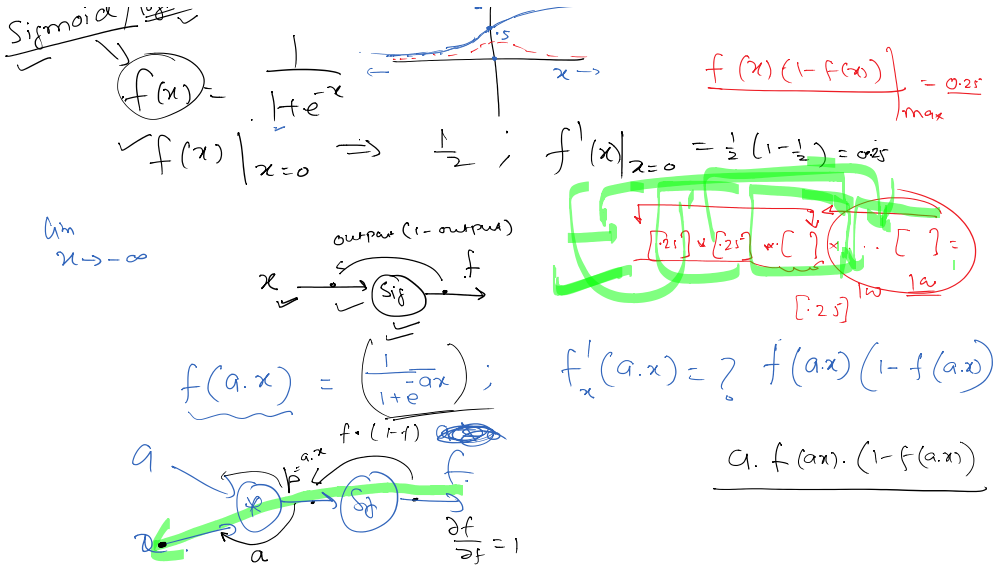
$\frac{e^{-x}}{1+e^{-x}}$

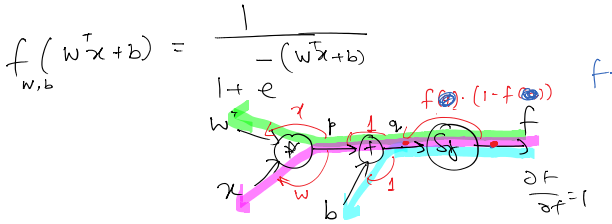
Sigmoid / logit

$f(x)$ $\frac{1}{1+e^{-x}}$



$f(x)(1 - f(x)) = 0.25$

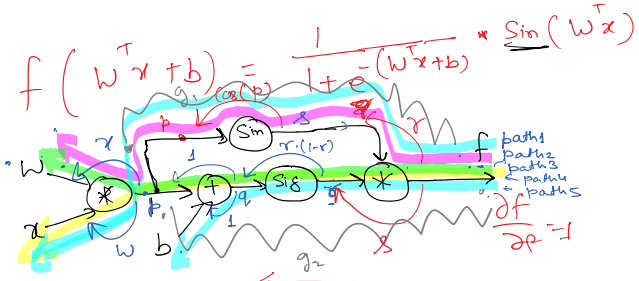


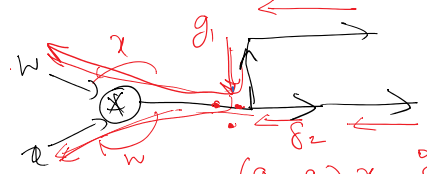


$f'_b = 1 \times f(\quad)(1-f(\quad)) \times 1$

$f'_x = 1 \times f(\quad)(1-f(\quad)) \times 1 \times w$

$f'_w = 1 \times f(\quad)(1-f(\quad)) \times 1 \times x$ → $Sig(q)$





$\frac{(g_1+g_2).x}{(g_1+g_2).w}$

$g_1.x + g_2 x = (1 \times s \times r(1-r) \times 1 \times x + r.\cos(p).x)$

$\frac{\partial f}{\partial w} = path3 + path2$

$\frac{\partial f}{\partial x} = path1 + path4$

$\frac{\partial f}{\partial b} = path5$

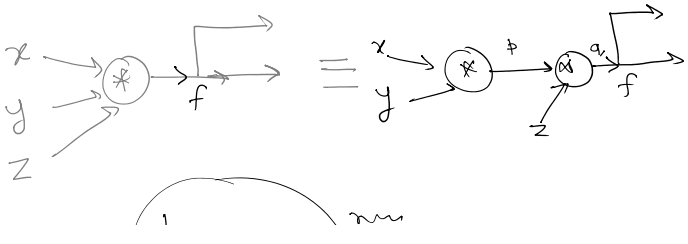
path 1 = $1 \times r \times \cos(p) \times w$

path 2 : $1 \times r.\cos(p).x$

path 3 : $1 \times s \times r(1-r) \times 1 \times x$

path 4 : $1 \times s \times r(1-r) \times 1 \times w$

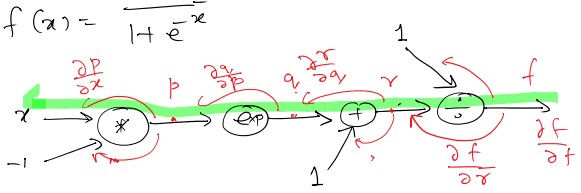
path 5 : $1 \times s \times r(1-r) \times 1$



$$f(w^T x + b) = \frac{1}{1 + e^{-(w^T x + b)}}$$

$$= \frac{\partial f}{\partial z} \cdot \frac{\partial z}{\partial q} \cdot \frac{\partial q}{\partial p} \cdot \frac{\partial p}{\partial x}$$

$$f(x) = \frac{1}{1 + e^{-x}}$$



$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial f} \cdot \frac{\partial f}{\partial r} \cdot \frac{\partial r}{\partial q} \cdot \frac{\partial q}{\partial p} \cdot \frac{\partial p}{\partial x}$$

Chain rule

Class Test-1: 3+2 points — Mid Sem

$$f(x,y) = x^y \cdot \log(x+y)$$

Compute $f_x'(x,y) = y \cdot x^{y-1} \log(x+y) + x^y \cdot \frac{1}{(x+y)}$

$$f_y'(x,y) = x^y \cdot \ln(x) \cdot \log(x+y) + \frac{x^y}{(x+y)}$$

$a \geq 0$ $\quad \lim_{a \to 0}$

$f(x) = a^x \Rightarrow f_x'(x) = ?$

$a^x \cdot \ln(a)$ incorrect

Correlation — Convolution

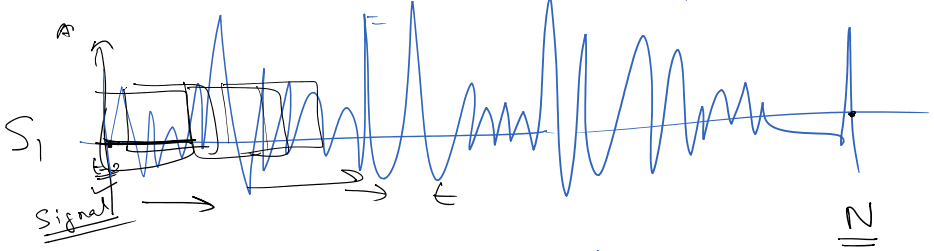
Sliding Vector (filter / kernel) Base Vector Signal

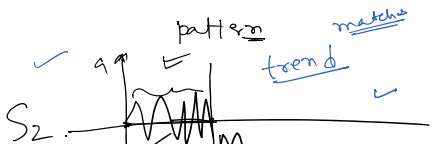
$=$ $\quad -$

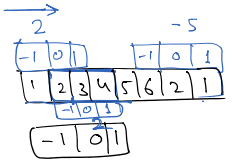
$|V_1| = |V_2|$

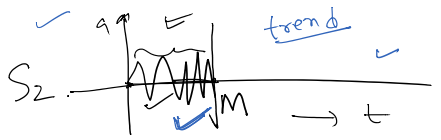
$$\text{Pearson Corrcd}(V_1, V_2) = \frac{1}{N} \left(\frac{V_1 - \mu_{V_1}}{\sqrt{\sigma_{V_1}^2}} \right) \cdot \left(\frac{V_2 - \mu_{V_2}}{\sqrt{\sigma_{V_2}^2}} \right)^T$$

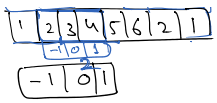
V_1	V_2
1	3
2	4
3	5
4	6
2	3
1	2
μ_{V_1}	μ_{V_2}
$\sigma_{V_1}^2$	$\sigma_{V_2}^2$











$M \leq N$

(Stride) Step size = 1. $N - M + 1$

Output vector size $= \left\lfloor \frac{N-M}{s} \right\rfloor + 1$

Stride = 2 $\left(\left\lfloor \frac{N-M}{2} \right\rfloor + 1 \right)$

Stride = s $= \left(\left\lfloor \frac{N-M}{s} \right\rfloor + 1 \right)$

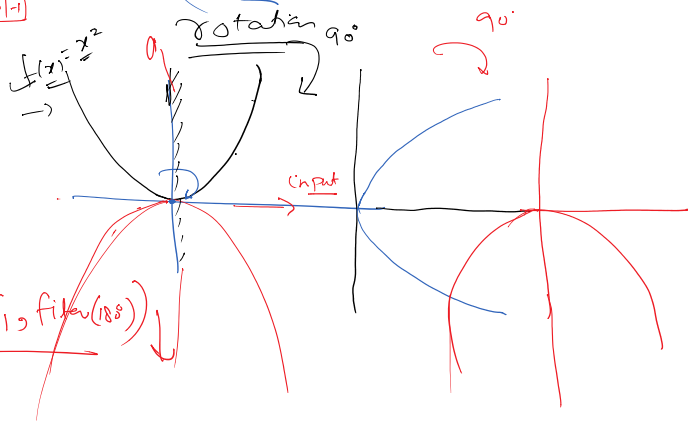
Convolution:

padding

rotate filter (180°)

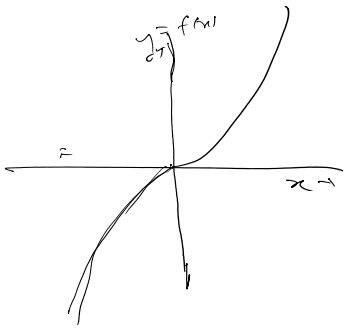
compute correlation $(S_1, \text{filter}(180°))$

$=$



mirror Image (x-y plane → x-axis is the mirror)

$y = f(x)$ rotate 180° $\Rightarrow y = -f(x)$



$f(x) = x^3$

$f(-x) = -f(x)$

$f(x)$ mirror at y-axis $f(-x)$

$= f(x) = x^2 \Rightarrow f(-x) = (-x)^2$

$f(x) = \log(x)$ x>0 $\Rightarrow f(-x) = \log(-x)$ X

$f(x) = \sin(x)$; $f(-x) = \sin(-x) = -\sin x$

$f(x) = \cos(x)$ $f(-x) = \cos(-x) = \cos(x)$

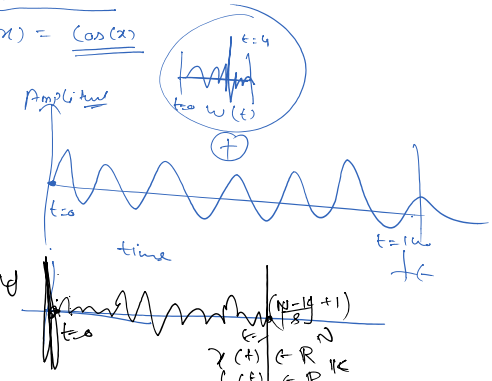
Signal $f(t)$ ·· filter $w(t)$

$N = 10$, $M = 5$

Stride = 1

Correl → $f(t) \cdot w(t)$

Convo → $f(t) \cdot w(-t)$



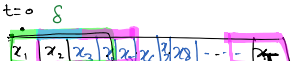
$y = R^{N-M+1}$ $\left(\frac{N-M}{s}+1\right)$

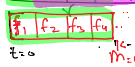
$x(t) \in R^N$ $f(t) \in R^M$

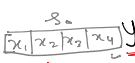
$s = 1$

y = Correlate $(x(t), f(t))$

1D:

$x(t) =$ 

$f(t) =$ 

$\frac{\partial y[0]}{\partial F_{t=0 \to t=k-1}} =$ 

$y[0] = x|_{t=0 \text{ To } t=k-1} \cdot f(t)|_{t=0 \text{ To } k-1}$

$= x_1 x_2 x_3 x_4$ $\begin{matrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{matrix}$ $= x_1 f_1 + x_2 f_2 + x_3 f_3 + x_4 f_4$

$s = 1$

$y[1] = x|_{t=1 \text{ to } t=k} \; f(t)|_{t=0 \text{ to } k-1} =$

$s=1$

$\frac{\partial y[1]}{\partial F} = [x_2 | x_3 | x_4 | x_5]$

$y[i] = x\Big|_{t=1 \text{ to } t=k} f(t)\Big|_{t=0 \text{ to } k-1}$

$[x_2 | x_3 | x_4 | x_5] \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{bmatrix} = x_2 \cdot f_1 + x_3 \cdot f_2 + x_4 \cdot f_3 + x_5 \cdot f_4$ ($t=1$ $t=0$, $t=2$ $t=1$, $t=3$ $t=2$, $t=4$ $t=3$)

$y[2] = x\Big|_{t=2 \text{ to } t=k+1} \cdot f(t)\Big|_{t=0 \text{ to } k-1}$

$\frac{\partial y[i]}{\partial F}\Big|_{t=0 \text{ to } k-1} = [x_{i+1}, x_{i+2} \cdots x_{i+k-1}]$

$y[i] = x\Big|_{t=i \text{ to } t=i+k-1} \cdot f(t)\Big|_{t=0 \text{ to } k-1}$

$= [x_{i+1} | x_{i+2} | x_{i+3} | x_{i+4}] \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{bmatrix} =$

$= x_{i+1} \cdot f_1 + x_{i+2} \cdot f_2 + x_{i+3} \cdot f_3 + x_{i+4} \cdot f_4$

$t=i$

$$y[i] = \sum_{j=1}^{k} x_{i+j} \cdot f_j$$

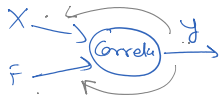
$y = \text{Corr}(X, F)$ $\quad 0 \le i \le \left\lfloor \frac{N-k}{s} \right\rfloor + 1$

$f(t) = [f_1 | f_2 | f_3 | f_4] \Rightarrow f(-t) = [f_4 | f_3 | f_2 | f_1]$

Convolution: $y[i] = \sum_{j=1}^{k} x_{i-j} \cdot f_j$

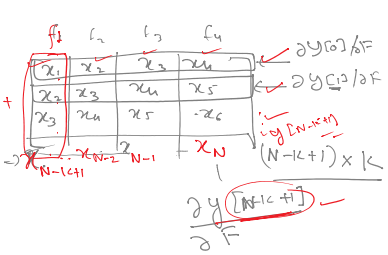
$\text{Cov}(x(t) \cdot f(t)) = \text{Correl}(x(t), f(-t))$

$y = \text{Corr}(X, F)$



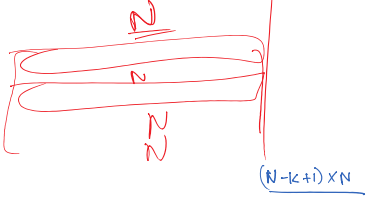
Jacobian $= \frac{\partial y}{\partial x} = ? \in \mathbb{R}^{N-k+1}$

$\frac{\partial y}{\partial F \in \mathbb{R}^k} = ?$



$\frac{\partial y[N-k+1]}{\partial F}$?

$\frac{\partial y}{\partial x}$ $\quad \frac{\partial y[0]}{\partial x}\Big|_{t=0 \text{ to } t=k-1} = F$?



$\frac{\partial y[0]}{\partial x_1} = f_1$, $\frac{\partial (y[0])}{\partial x_2} = f_2$ ... $\frac{\partial (y[0])}{\partial x_4} = f_4$, $\frac{\partial (y[0])}{\partial x_5} = 0$

$\frac{\partial (y[1])}{\partial x_1} = 0$ $\quad \frac{\partial (y[1])}{\partial x_2} = F_1$ ... $\frac{\partial (y[1])}{\partial x_5} = F_4$, $\frac{\partial y[1]}{\partial x_6} = 0$, $\frac{\partial y[1]}{\partial x_N} = 0$

$\frac{\partial (y[2])}{\partial x_1} = 0$ $\quad \frac{\partial (y[2])}{\partial x_2} = 0$ $\quad \frac{\partial (y[3])}{\partial x_3} = F_1$ ... $\frac{\partial [y[2]]}{\partial x_N} = 0$

$\frac{\partial (y[N-k+1])}{\partial x_1} = 0$ ... $= 0$ $\quad \frac{\partial y[N-k+1]}{\partial x_{N-k+1}}$, $\frac{\partial y[N-k+1]}{\partial x_N} = F_4$

$$\frac{\partial y(N-k+1)}{\partial x_1} = 0 \quad \cdots = 0 \quad \frac{\partial y[N-k+1]}{\partial x_{N-k+1}} \cdots \frac{\partial y[N-k+1]}{\partial x_N} = F_4$$

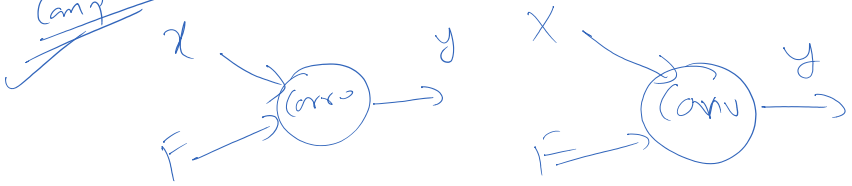
$$\frac{\partial y}{\partial x} = \text{Corr}(x(t), \text{flip } 180^\circ f(t))$$

$$= \text{Corr}(x(t), f(-t))$$

$$= \text{Conv}(x(t), f(t))$$

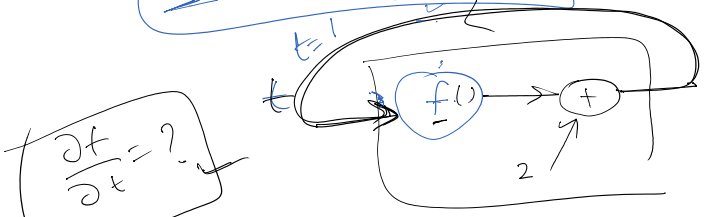
N-k+1

Commutative



$f(0) = 1$

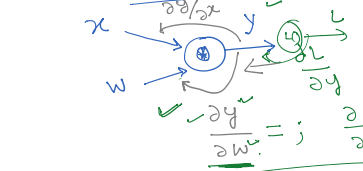
$f(t) = f(t-1) + 2$



$\frac{\partial f}{\partial t} = ?$

Convolution : (Deeplearning)

Edges are the basic building



$$\frac{\partial L}{\partial w} = \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial w}$$

$$\frac{\partial L}{\partial x} = \frac{\partial L}{\partial y} \cdot \frac{\partial y}{\partial x}$$

$x \in \mathbb{R}^N$

$w \in \mathbb{R}^M$

$y \in \mathbb{R}^{N-M+1}$

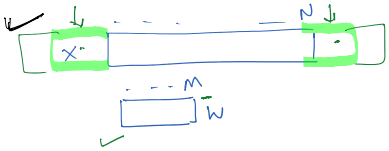
Stride = 1

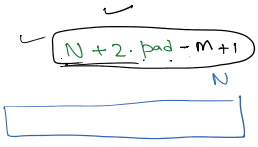
Valid Convolution (Stride = 1) = $N - M + 1$

Same Convolution => output vector should have same size as input signal vector.

Stride = 1

Pad = 1 => $N + 2 \cdot$ padding





If Stride = S:

$$\left\lfloor \frac{N + 2\,pad - M}{S} \right\rfloor + 1$$

Stride = 1

$$N + 2\,pad - M + 1 = N$$

$$pad = \left\lceil \frac{M-1}{2} \right\rceil$$: odd size kernels

Stride = S

$|N + 2\,pad - M| \cdot \cdots$

[1 2 1 kernels]

Stride = s

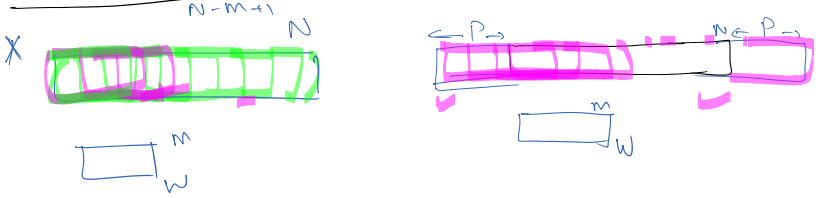
$$\left\lfloor \frac{N + 2\,pad - M}{s} \right\rfloor + 1 = N$$

$$Pad = \frac{s(N-1) + M - N}{2}$$

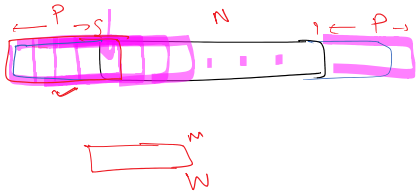
$$= \left\lceil \frac{N(s-1) - s + M}{2} \right\rceil$$

$P = \frac{M-1}{2}$

Full Convolution :



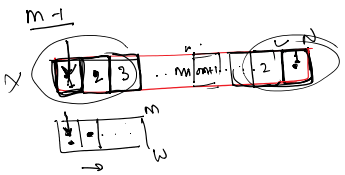
$P = M - 1$



$N + 2(M-1) - M + 1$

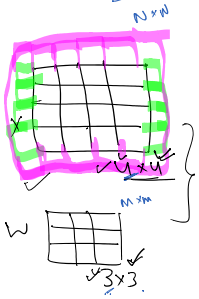
$\Rightarrow N + M - 1$

$y \in \mathbb{R}$



$s = 1$

$p = \frac{3-1}{2}, \frac{3-1}{2}$



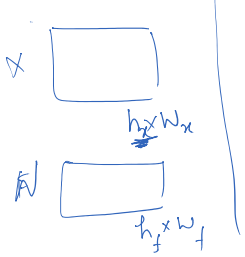
Valid Same Full

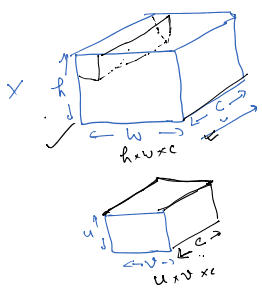
$(4-3+1) \times (4-3+1)$ | 4×4 | 6×6

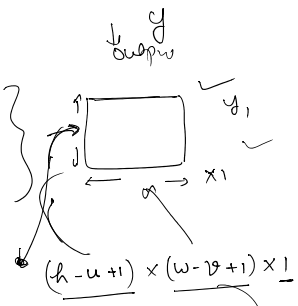
2×2

$(N-M+1) \times (N-M+1)$ N×N $(N+M-1) \times (N+M-1)$

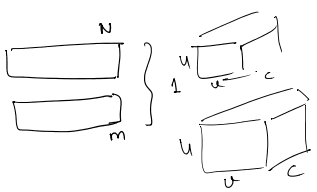
$\left(\frac{h_x + 2p - h_f}{s} + 1\right) \left(\frac{w_x + 2p - w_f}{s} + 1\right)$ $h_x \times w_x$ $(h_x + h_f - 1) \times (w_x + w_f - 1)$

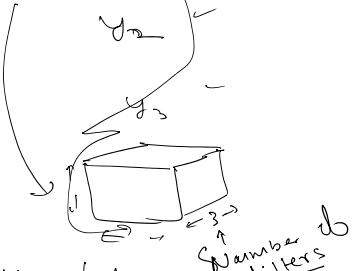




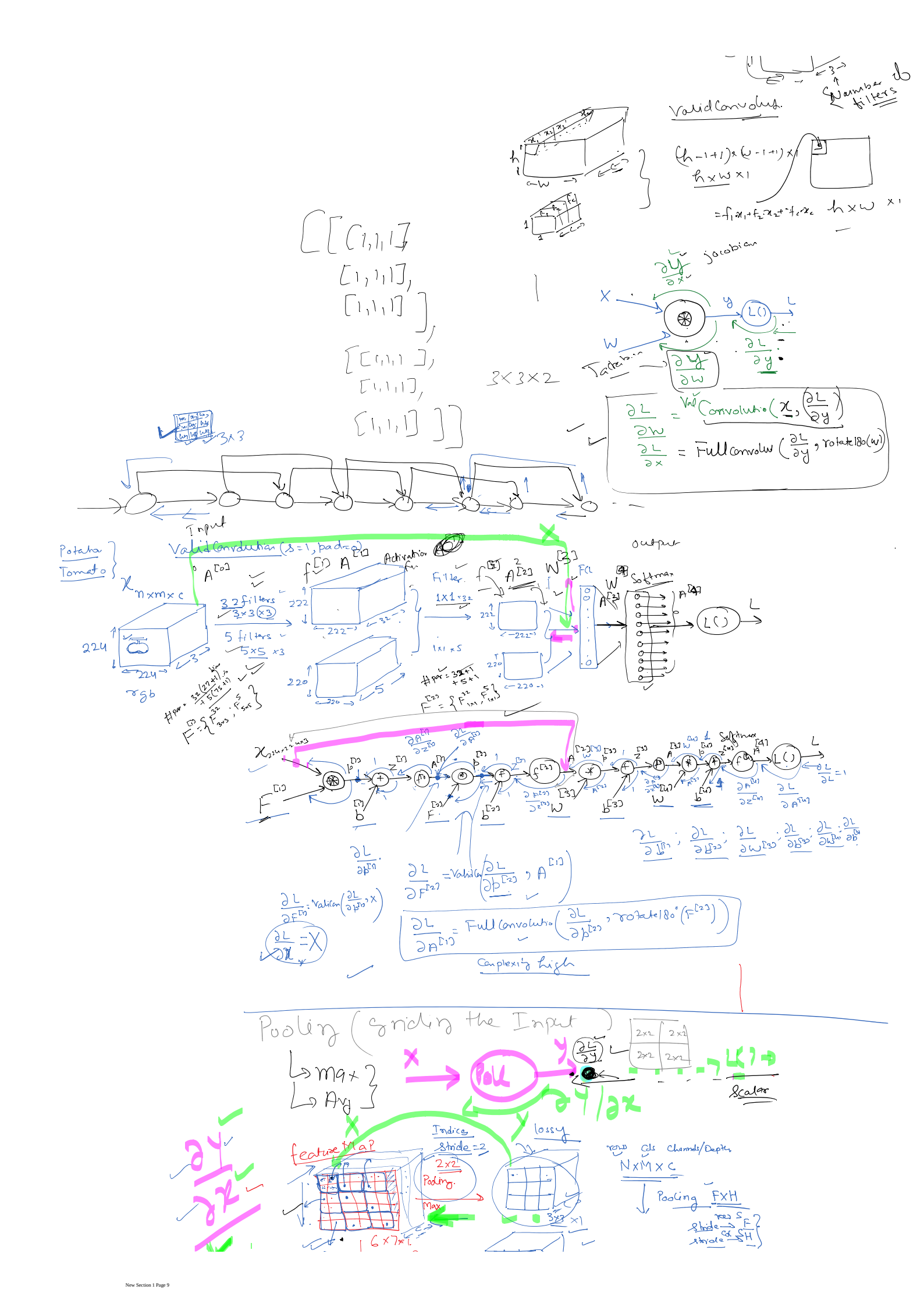


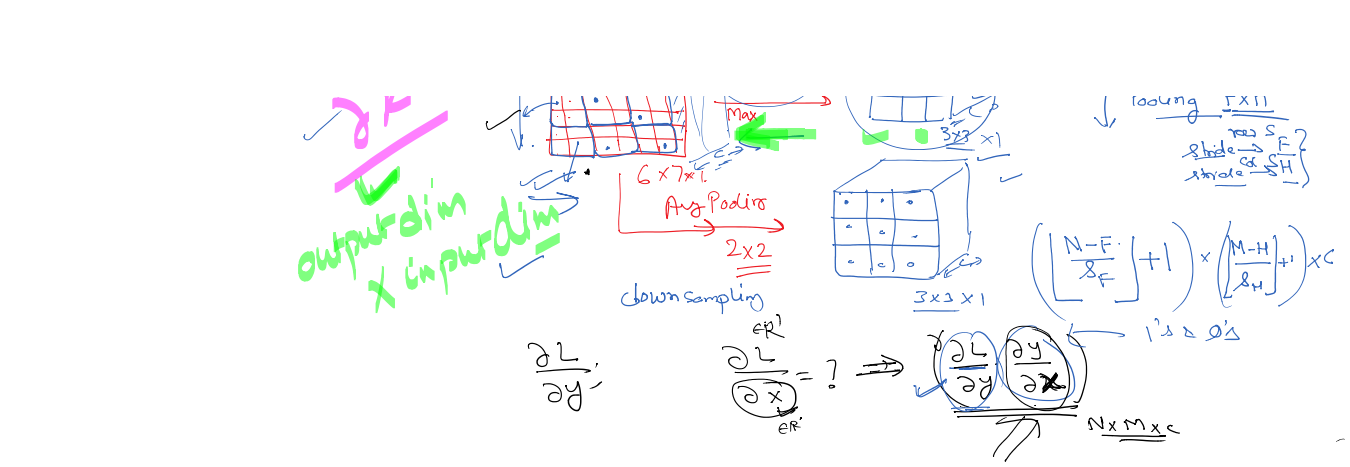
$N = M$





Valid Convolution


Valid Convolution
Jacobian
Input
Output
Softmax
Pooling (striding the Input)
Max
Avg
Scalar
lossy
Indices
Stride = 2
Complexity high



$\frac{\partial L}{\partial y}$ $\frac{\partial L}{\partial x} = ?$ $\Rightarrow \frac{\partial L}{\partial y} \frac{\partial y}{\partial x}$

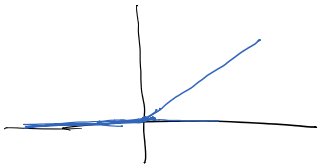
$x \xrightarrow{\text{Pooling}} y$

Relu

$f(x) = \max(x, 0) \Rightarrow$

$\frac{\partial f}{\partial x} = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x < 0 \\ \text{Undefined if } x = 0 \end{cases}$

$x \in \mathbb{R}$

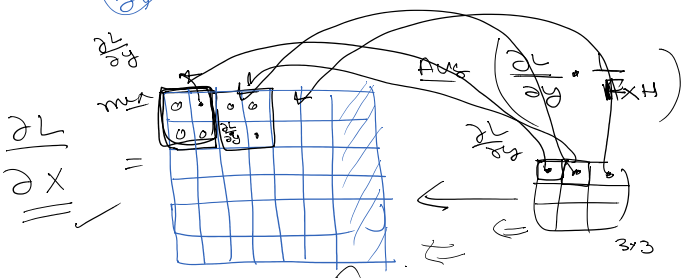


$f(x,y) = \max(x,y)$

$\frac{\partial f}{\partial x} = 1$ if $x > y$ otherwise 0

$\frac{\partial f}{\partial y} = 1$ if $y > x$ otherwise 0

$\frac{\partial L}{\partial x} =$



$= \frac{\text{upsampling}}{\text{de-pooling}}$

$\frac{\partial y}{\partial x} =$ Jacobian

$\frac{\partial L}{\partial x} = \dim \partial x$

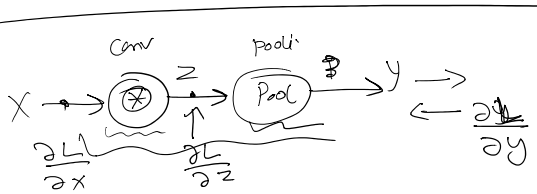
$\frac{\partial L}{\partial x} =$

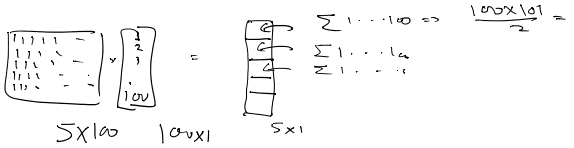
$\frac{\partial L}{\partial x} = \left(\frac{\partial L}{\partial y}\right)\left(\frac{\partial y}{\partial x}\right)$

$\left(\left\lfloor \frac{N-F}{S_F} \right\rfloor + 1 \times \left\lfloor \frac{M-H}{S_H} \right\rfloor + 1\right) \times (\quad)$

$1 \times \dim y \qquad \dim y \times \dim x$

$= 1 \times \dim x$





$f(x) = a \cdot f(x-1) + b$ (Recurrence Equation)

Unrolling: x is discrete time — $a \in \mathbb{R}$ $b \in \mathbb{R}$

$f(x) = a \cdot (a \cdot f(x-2) + b) + b$

$= a \cdot (a \cdot (a \cdot f(x-3) + b) + b) + b$

n steps

$= a^n \cdot f(x-n) + n \cdot b$ — $f(x-n)$ is independent of x

$f'(x) = \frac{\partial f}{\partial x} = a^n \cdot f'(x-n)$

$f(x) = a \cdot f(x-1) + b$

$f'(x) = a \cdot \left(\frac{\partial f(x-1)}{\partial x}\right) + \frac{\partial b}{\partial x}^{0}$

$= a \cdot a \left(\frac{\partial f(x-2)}{\partial x}\right)$

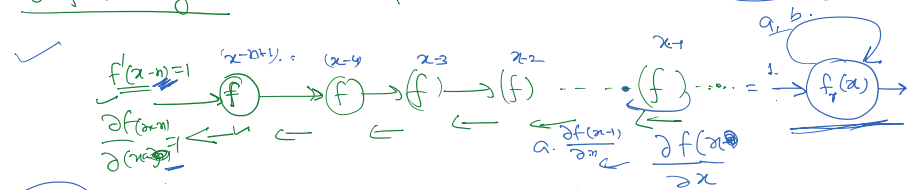
n ↓

$= a^n$

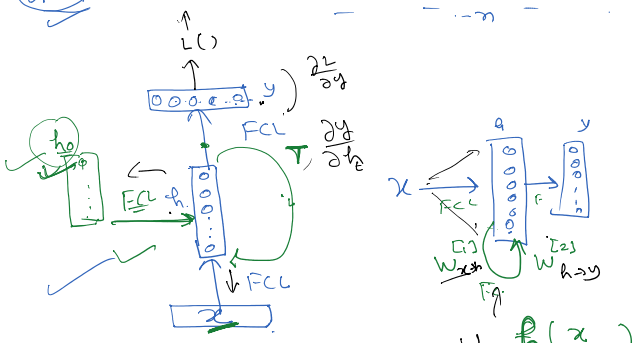
$f'(x-n) = 1$

$\left(\frac{\partial f(x)}{\partial f(x-1)} = a\right)$

Unfolding Recurrence Eq^n



a^n . . ← ← a ← a ← a ← a



$W_{h \to h} f(x)$

$h_t = f_{FCL}(h_{t-1}, x_t)$

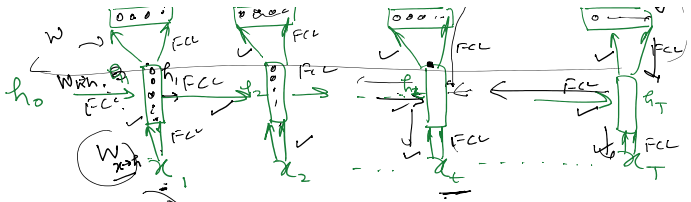
$h(x-1)$

3

h_t R h_{t-1}

FCL

$h_t = W_h \cdot h_{t-1}$ (4)



$f_1, f_2 := (\text{Sigmoid}, \text{tanh}, \text{ReLu})$

$(f_1 = \text{tanh}, f_2 = \text{tanh})$

$$h_t = f_1\left(\underbrace{w_{x\to h}\cdot x_t + w_{h\to h}\cdot h_{t-1}}_{z_t} + b_h\right)$$

$$y_t = f_2\left(w_{h\to y}\cdot h_t + b_y\right)$$

⇒ Exercise Computation Graph

$$\frac{\partial h_t}{\partial w_{h\to h}} = f_1'(z_t)\cdot\frac{\partial z_t}{\partial w_{hh}}$$

$$=$$

$$z_t = w_{x\to h}\cdot x_t + w_{h\to h}\cdot h_{t-1} + b_h$$

$$\frac{\partial z_t}{\partial w_{h\to h}} = \frac{\partial\left(w_{x\to h}\cdot x_t\right)}{\partial w_{h\to h}} + \frac{\partial\left(w_{h\to h}\cdot h_{t-1}\right)}{\partial w_{h\to h}} + \frac{\partial b_h}{\partial w_{h\to h}}$$

$$= \left(\frac{\partial w_{x\to h}}{\partial w_{h\to h}}\cdot x_t + w_{x\to h}\cdot\frac{\partial x_t}{\partial w_{h\to h}}\right) + \left(w_{h\to h}\cdot\frac{\partial h_{t-1}}{\partial w_{h\to h}} + \frac{\partial w_{h\to h}}{\partial w_{h\to h}}\cdot h_{t-1}\right)$$

$$= 0 + 0 + w_{h\to h}\cdot\frac{\partial h_{t-1}}{\partial w_{h\to h}} + 1\cdot h_{t-1}$$

$\frac{\partial z_t}{\partial w_{h\to h}} = \left(x_t + w_{hh}\cdot\frac{\partial h_{t-1}}{\partial w_{h\to h}}\right)$

(Note: $h_{t-1} = f_1\left(\underbrace{w_{x\to h}\cdot x_{t-1} + w_{h\to h}\cdot h_{t-2} + b_h}_{z_{t-1}}\right)$)

$$\frac{\partial z_t}{\partial w_{h\to h}} = h_{t-1} + w_{h\to h}\cdot\left(\frac{\partial h_{t-1}}{\partial w_{h\to h}}\right)$$

$$\frac{\partial z_{t-1}}{\partial w_{h\to h}} = h_{t-2} + w_{h\to h}\frac{\partial h_{t-2}}{\partial w_{h\to h}}$$

$$h_1 = f_1\left(\underbrace{w_{x\to h}\cdot x_1 + w_{h\to h}\cdot h_0 + b_h}_{z_1}\right)$$

$$\frac{\partial z_1}{\partial w_{h\to h}} = h_0$$

$f(z) = z$, $\frac{\partial f}{\partial z} = 1$ linear $= 1$

$\frac{\partial h_t}{\partial w_{h\to h}}$

$f(z) = z$, $\frac{\partial f}{\partial z} = 1$ time = 1

$$\frac{\partial h_t}{\partial w_{h\to h}} = f_1'(z_t)\cdot\left(h_{t-1} + w_{h\to h}\left(f_1'(z_{t-1})\cdot\left(h_{t-2} + w_{h\to h}\left(f_1'(z_{t-2})\cdot(h_{t-3} + w_{hx})\right)\right)\right)\right)$$

$$= \cdots f_1'(z_2)\cdot\left(h_1 + w_{h\to h}\left(f_1'(z_1)\cdot(h_0)\right)\right)\cdots$$

$$\frac{\partial h_t}{\partial w_{x\to h}} = (\quad ? \quad \cdot$$

$$\frac{\partial y_t}{\partial h_t} = f_2'(\quad)\cdot w_{h\to y}$$

$$\frac{\partial y_t}{\partial w_{h\to y}} = f_2'(\quad)\cdot h_t$$

Computation Graph

$P(y_T \mid x_1, \cdots x_T)$

RNN

$f(0) = x_0$

$f(x) = a\cdot f(x-1) + b$

$x = 3$:

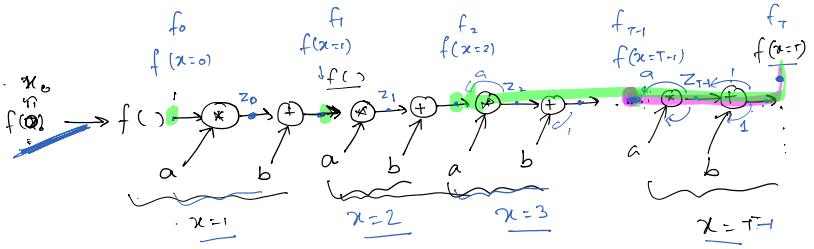
$f(3) = a^3 x_0 + 3b$

$a * f(2) + b$

$a * f(1) + b$

$a * f(0) + b$

$f(0) = x_0$

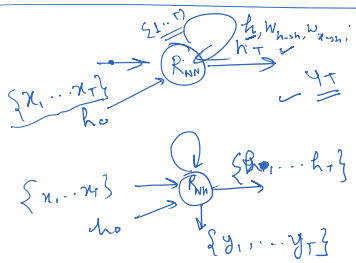


$\frac{\partial f_T}{\partial z_{T-1}} = 1 \cdot \frac{\partial z_{T-1}}{\partial f_{T-1}} = a \cdot \frac{\partial f_T}{\partial f_{T-1}} = 1\times a$

$\frac{\partial f_T}{\partial f_2} = (1\times a)\times(1\times a)\times(1\times a)\cdots\cdots(1\times a)$ ($T-2$ times) $= a^{T-2}$

$\frac{\partial f_T}{\partial f_1} = a^{T-1}$

$\frac{\partial f_T}{\partial f_0} = a^T \cdot (\quad)$



(only 'T'th output) Many to one

many input & one output

$\epsilon \in W_{h \to h}$ $\forall \epsilon$

$\epsilon \ll 1$

$$\frac{\partial h_t}{\partial W_{h \to h}} = h_{t-1} + W_{h \to h}\left(h_{t-2} + W_{h \to h}\left(h_{t-3} + \cdots \right.\right.$$

$$\left.\left. W_{h \to h}(h_0)\right)\right)$$

$$= h_{t-1} + W_{h \to h} \cdot h_{t-2} + (W_{h \to h} \cdot) W_{h \to h} \cdot h_{t-3} + \cdots$$

$$(W_{h \to h})^{t-1} h_0$$

→ RNN, LSTM, GRU

Bi-directional ; //

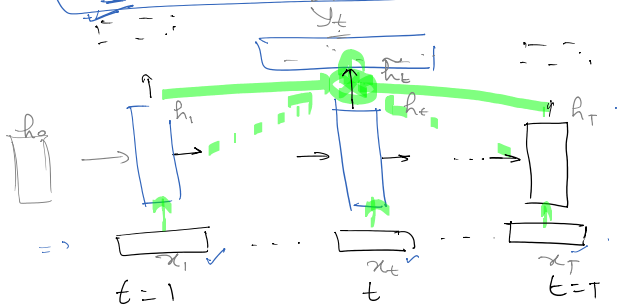
Attention Mechanism

Weighted Aggregation Method }

$V = [v_1\ v_2\ \cdots\ v_n]$ probabilistic Mapping

$V_{agg} = \alpha_1 v_1 + \alpha_2 v_2 + \cdots \alpha_n v_n$ → Normalize = Softmax (U_{if})

$\sum \alpha_i = 1$



$$\hat{y}_t = f_o\left(W_{ho}^T \tilde{h}_T + b_o\right)$$

$$\tilde{h}_t = \alpha_1^t h_1 + \alpha_2^t h_2 + \cdots \alpha_T^t h_T$$

$h_{t-1} =$

for one-hot vectors: [0 ... 1 ... 0]

	$n=10$	$n=100$	$n=1000$
$M_{one\text{-}hot}$	0	0	0
$\mathbb{R}^2$ one-hot	.	.	.

v_1, v_2

$v_1 \sim N(0,1)$

$v_2 \sim N(0,1)$

Correlation $= v_1^T v_2$

d = # neurons in the embedding $d \times 1$

DNN (one-hidden layer)

x_i → Embedding Module → d

one-hot encoding

$f\left(W_E^T x_i + \begin{pmatrix} b \\ \vdots \\ b \end{pmatrix}\right)$

Correlation — $V_1^T V_2$

one hot encoding

$\forall\ i = 1$ to T

x_i = Embedding Vector

$S_1, S_2, \ldots, S_N$ < $x_1\ x_2\ x_3\ x_4\ x_5\ x_6 \ldots x_t \ldots x_T$ >

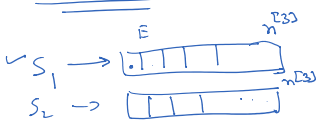
(0 1 2 3 4 . . . T-1)

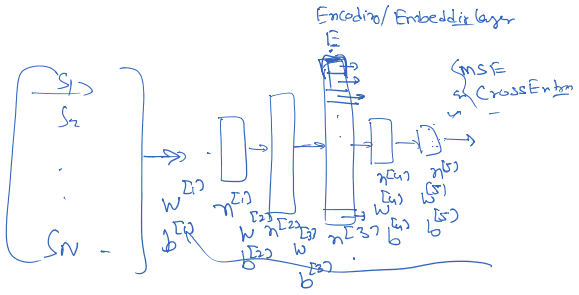
W = 1

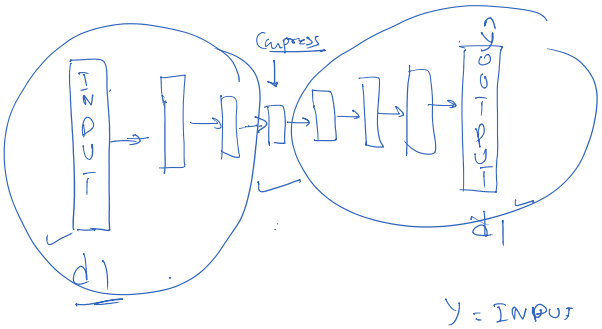
= $[x_1\ x_2\ x_3\ x_{5000}\ x_4]$

→ x_i = Embedding Vector

Similarity = $x_i^T x_j$







y = INPUT

one-hot

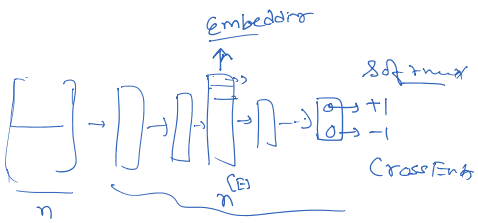
n^{2n}

S1 (x3, x10000) +1

(x3, x4) +1

(x3 x3000) -1

(x3 x7000) -1



→ token (word in text, value at certain time steps, frames, . . .)

Embedding Weights

→ token (word in text, value at certain time steps, frames, ...)

token → [ ] Embedding Vector. 1xd Embedding dimension

Embedding Space: dim1, dim2, dim3

Positional Encoding + Attention

text: S1: < x_1 x_2 , . . x_t , . . . x_T > (1, 2, t, t-1, T, T-1)

$x_1 \rightarrow Ex_1$ [ ]$_{1\times d}$? Position b_1 0 [ ] [ ]$_{1\times d}$

$x_2 \rightarrow Ex_2$ [ ]$_{1\times d}$? b_2 1 [ ] [1]

$x_T \rightarrow Ex_T$ [ ]$_{1\times d}$? b_T T-1 [ ] [T-1]

binary → Sinusoids

position in Positional Embedding vector → i = 0 to d/2

even: $2i$ → $\sin\left(\frac{\text{position of token}}{\delta^{2i/\text{Embedding dim of the token}}}\right)$

odd: $2i+1$ → $\cos\left(\frac{\text{position of token}}{\delta^{2i/\text{Embedding of the token}}}\right)$

$\delta = 10000$

S1: < x_1 x_2 x_3 > (0, 1, 2)

S2: < x_4 x_5 x_6 > (0, 1, 2)

d: 4 (Embedding dim)

Pos 0: x_1 = [ ]$_{1\times 4}$

1: x_2 = [ ]$_{1\times 4}$

2: x_3 = [ ]$_{1\times 4}$

2: x_6 = [ ]$_{1\times 4}$

Position Emb [ ]

0	1	2	3
0	1	0	1

i = 0, pos = 0

$\sin\left(\frac{0}{10000^{0/4}}\right) = 0$

$\cos\left(\frac{0}{10000^{0/4}}\right) = 1$

i = 1, pos = 0

$\sin\left(\frac{0}{c}\right) = 0$

$\cos\left(\frac{0}{c}\right) = 1$

i = 0, pos = 1

$\sin\left(\frac{1}{10000^{0/4}}\right) = \sin(1)$

$\cos\left(\frac{1}{10000^{0/4}}\right) = \cos(1)$

i = 1, pos = 1

$\sin\left(\frac{1}{10000^{2/4}}\right) = \sin\left(\frac{1}{\sqrt{100}}\right)$

$\cos\left(\quad\right) = \cos\left(\frac{1}{\sqrt{100}}\right)$

Assignment:

For a fixed Embedding dimension d,

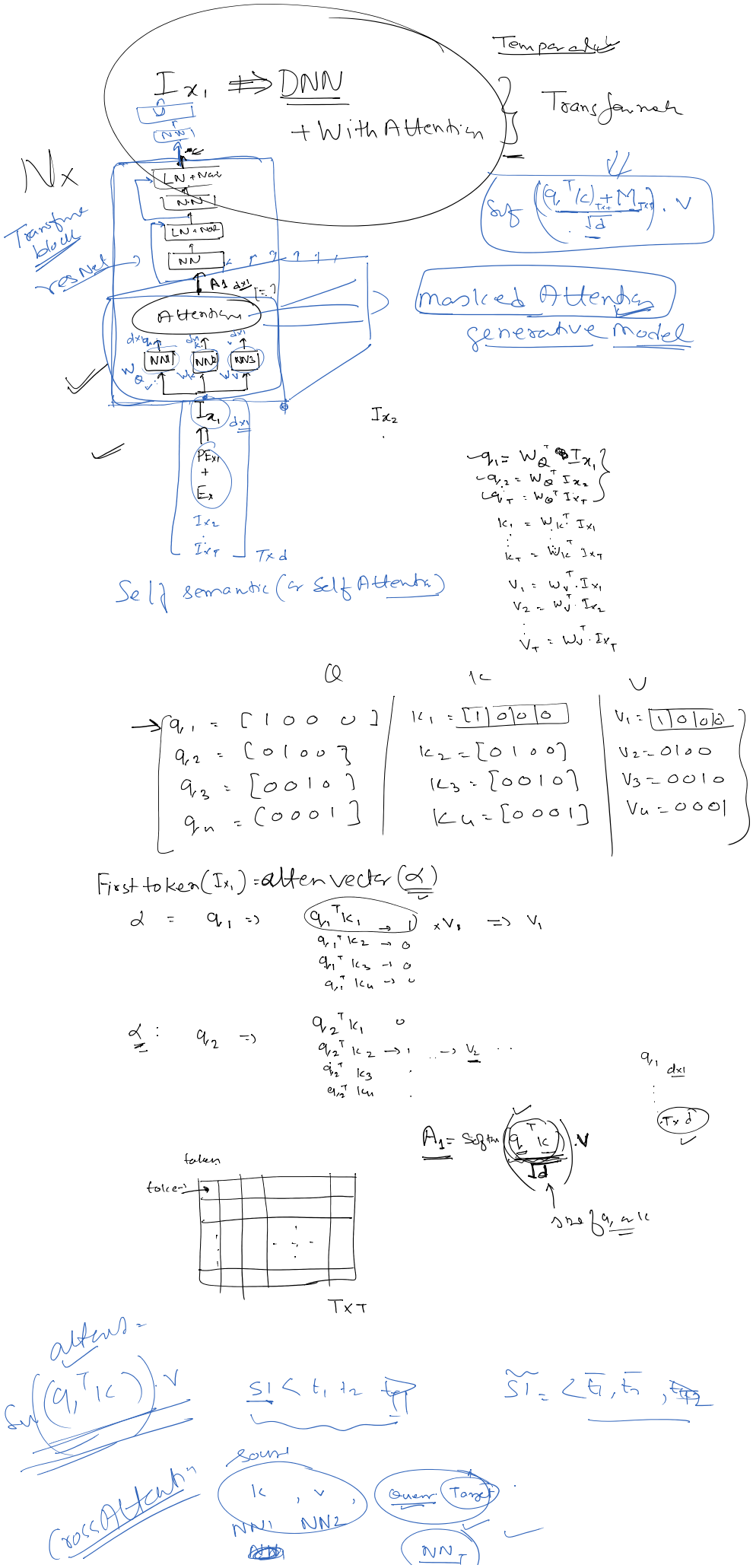
What is the maximum position that can be uniquely encoded in PE?

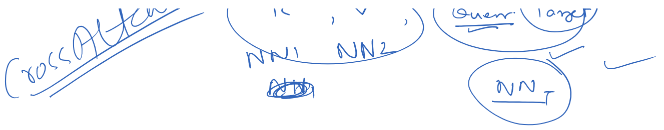
Visualize by varying 'd'

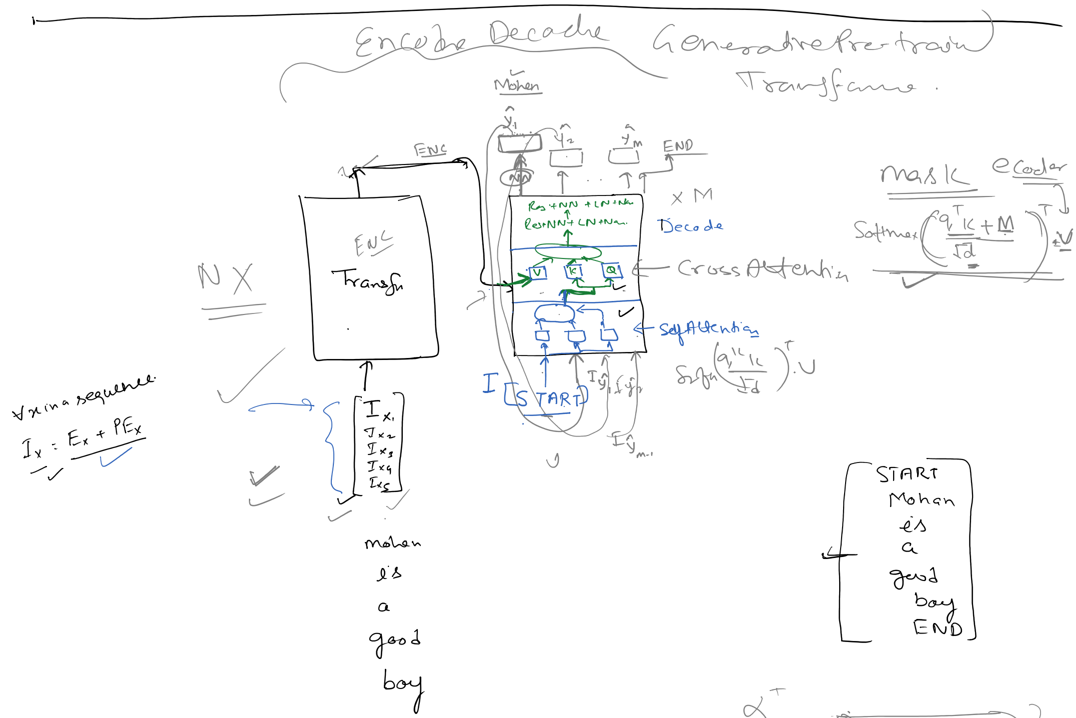
$I_{x_1} = Ex_1 + PE_{x_1}$ → Semantic + Position

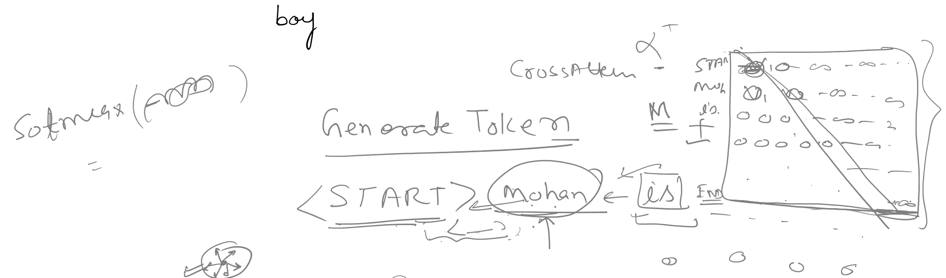
→ Simple Classification (Sentiment)

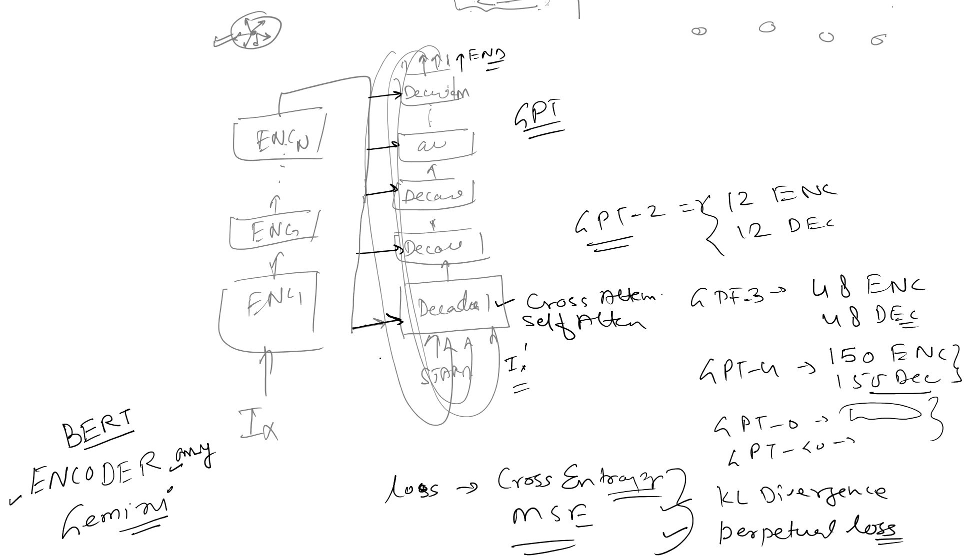
DNN


Temporal
Transformer
Ix1 ⇒ DNN
+ With Attention
Nx
Transformer block
LN + Mask
NN
LN + Mask
NN
Add
Attention
NN1
NN2
NN3
Ix1
Ix2
PEx1
+
Ex1
Ix2
IxT
T x d
Self semantic (or Self Attention)
masked Attention
generative model
q1 = WQT Ix1
q2 = WQT Ix2
qT = WQT IxT
k1 = WkT Ix1
kT = WkT IxT
V1 = WVT Ix1
V2 = WVT Ix2
VT = WVT IxT
Q
K
V
q1 = [1 0 0 0]
q2 = [0 1 0 0]
q3 = [0 0 1 0]
qn = [0 0 0 1]
k1 = [1 0 0 0]
k2 = [0 1 0 0]
k3 = [0 0 1 0]
kn = [0 0 0 1]
V1 = 1000
V2 = 0100
V3 = 0010
Vn = 0001
First token(Ix1) = atten vector (A)
A = q1 => q1T k1 → 1 x V1 => V1
q1T k2 → 0
q1T k3 → 0
q1T kn → 0
A : q2 => q2T k1 0
q2T k2 → 1 → V2
q2T k3
q2T kn
q1 dx1
Tx d
A1 = Softmax (q1T k / √d) . V
sim of q1 & k
token
token
T x T
atten = Softmax (q1T k) . V
S1 < t1 t2 tT
S1 = < t1, t2, tT2
cross Attention
source
k , v ,
NN1 NN2
Query Target
NNT









10,000

Top k outputs

I_x → [ ]

Human Feed back

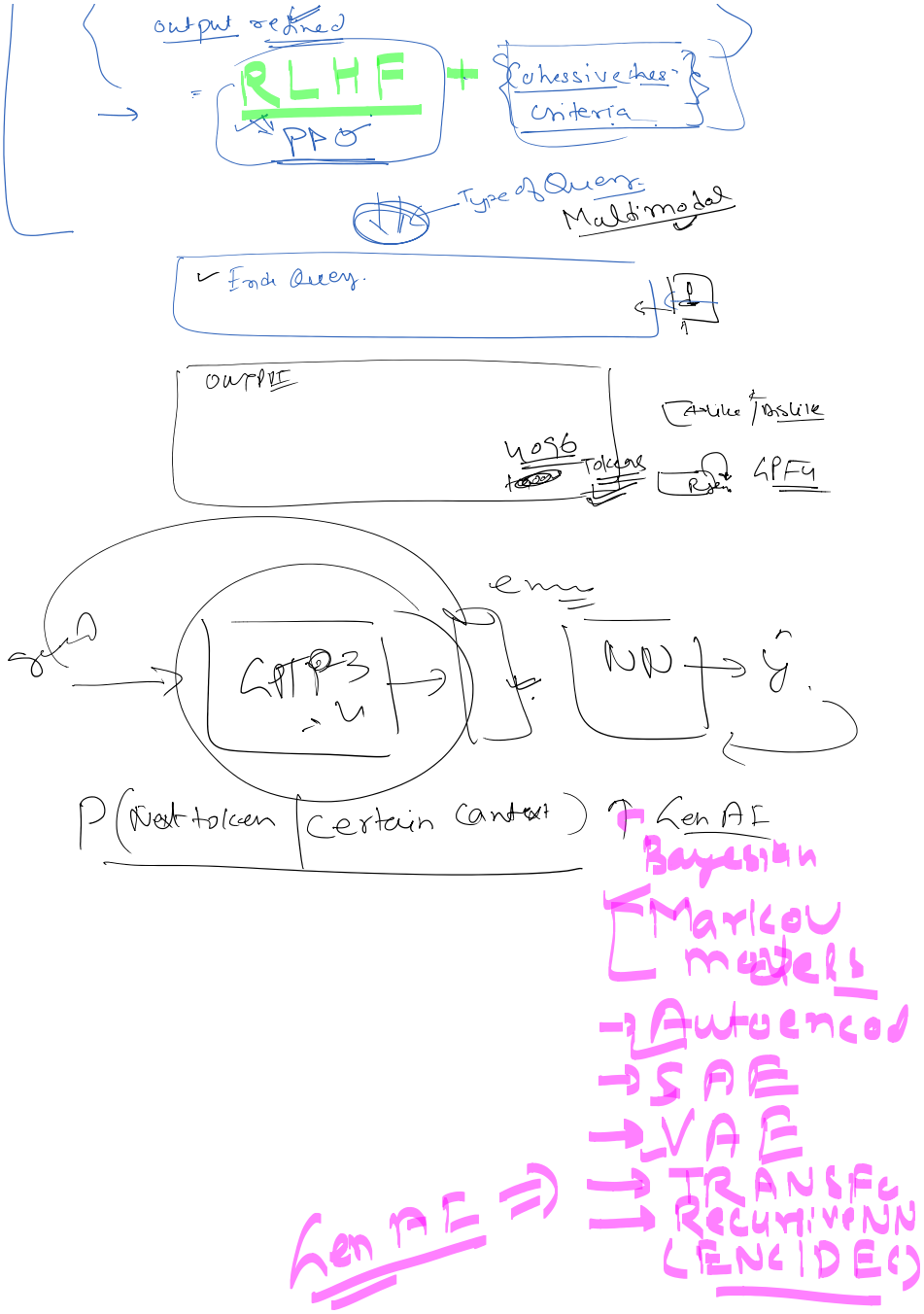
q sentence on good → { Mohan is a good boy → S1
He is good → S2
y = Cat is good → S3
Envir is good → S4 }

Use

{ Learning Reinforcement Learning (RL) }

output refined

RLHF + {Cohessiveness}


output refined
RLHF
PPO
Cohessive ness
criteria
Type of Query:
Multimodal
Fresh Query.
OUTPUT
4096
Tokens
GPT4
NN
P(Next token | certain context)
Gen AI
Bayesian
Markov models
Autoencod
SAE
VAE
TRANSFo
Recurrent NN
(ENCODED)
Gen AI