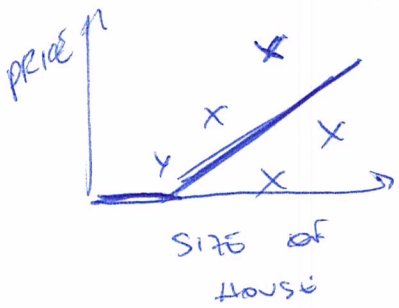
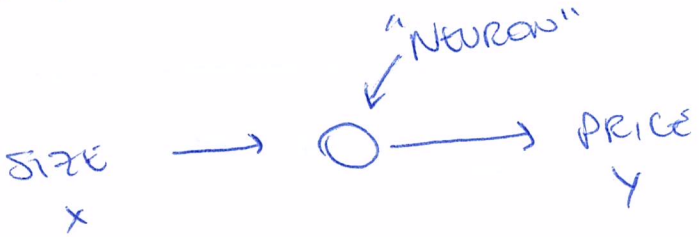
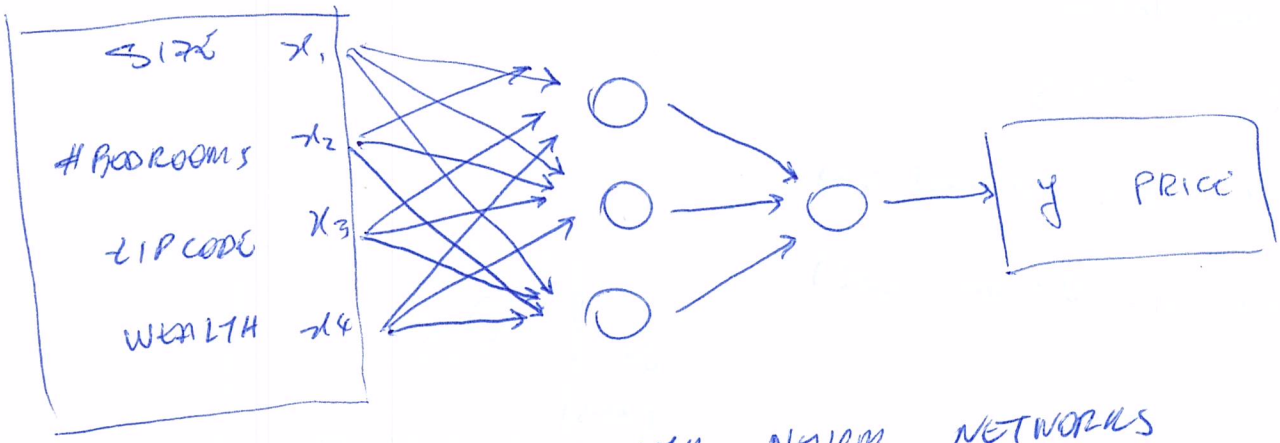
# DEEP LEARNING . AI

## COURSE 1 - NEURAL NETWORKS AND DEEP LEARNING

### - WHAT IS A NEURAL NETWORK?



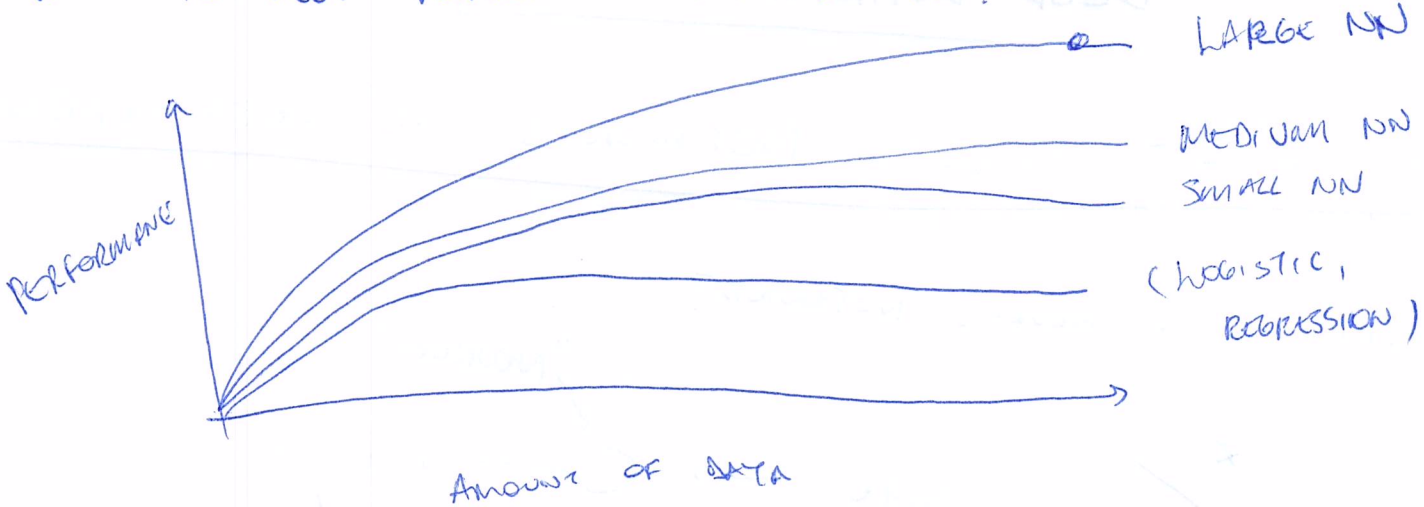




### - SUPERVISED LEARNING WITH NEURAL NETWORKS

| INPUT (x) | OUTPUT (y) | APPLICATION | |
|---|---|---|---|
| HOME FEATURES AD, USER INFO | PRICE | REAL ESTATE | } NN |
| | CLICK ON AD | ONLINE ADVERTISING | |
| IMAGE | OBJECT (1, ..., 1000) | PHOTO TAGGING | } CNN |
| AUDIO | TEXT TRANSCRIPT | SPEECH RECOGNITION | } RNN |
| ENGLISH | CHINESE | MACHINE TRANSLATION | |
| IMAGE, RADAR INFO | POSITION OF OTHER CARS | AUTONOMOUS DRIVING | } CUSTOM / HYBRID NN |

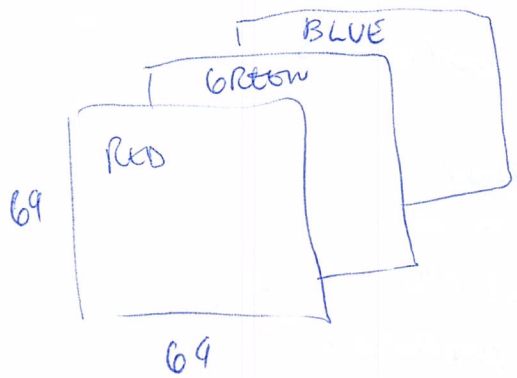
- WHY IS DEEP LEARNING TAKING OFF?




* SCALE DRIVES DEEP LEARNING PROGRESS
  - DATA
  - COMPUTATION
  - ALGORITHMS

- BINARY CLASSIFICATION

1 (CAT) VS. 0 (NON CAT)


$$x = \begin{bmatrix} 255 \\ 231 \\ \vdots \\ 255 \\ 134 \\ \vdots \\ 255 \\ 134 \\ \vdots \end{bmatrix} \quad n = n_x = 12288$$

$64 \times 64 \times 3 = 12288$

$(x, y) \Rightarrow x \in \mathbb{R}^{n_x}$

$y \in \{0, 1\}$

TRAINING EXAMPLES: $\{ (x^{(1)}, y^{(1)}), \dots, (x^{(m)}, y^{(m)}) \}$

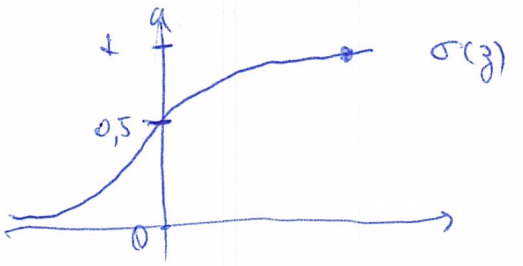
$$X = \begin{bmatrix} | & | & & | \\ x^{(1)} & x^{(2)} & \dots & x^{(m)} \\ | & | & & | \end{bmatrix} \quad \downarrow n_x$$

- LOGISTIC REGRESSION

$x$, $\hat{y} = P(y=1|x)$, $0 \le \hat{y} \le 1$, $x \in \mathbb{R}^{n_x}$

PARAMETERS: $w \in \mathbb{R}^{n_x}$, $b \in \mathbb{R}$

OUTPUT: $\hat{y} = \sigma(\underbrace{w^T x + b}_{z})$, $z = w_1 x_1 + w_2 x_2 + b$ (FOR EXAMPLE)




$\sigma(z) = \frac{1}{1+e^{-z}}$ (SIGMOID FUNCTION)

- LOGISTIC REGRESSION COST FUNCTION

LOSS (ERROR) FUNCTION:

$L(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$ $\Rightarrow$  (LOCAL MAXIMA)

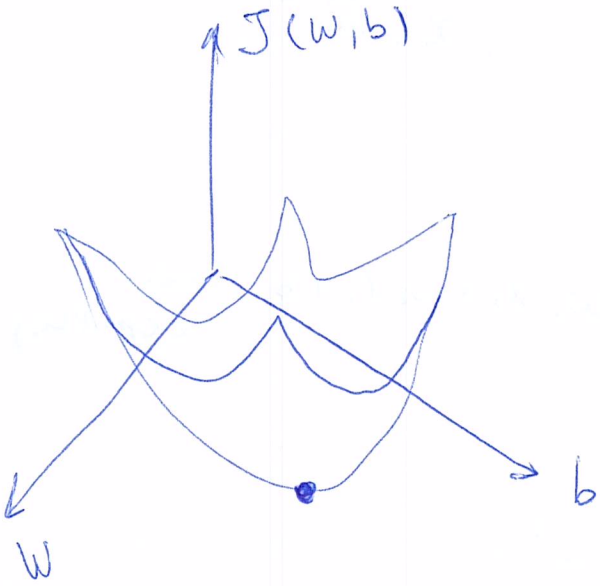
$L(\hat{y}, y) = -\left[ y \log \hat{y} + (1-y) \log(1-\hat{y}) \right]$

IF $y=1$: $L(\hat{y}, 1) = -\log \hat{y}$ { WANT $\log \hat{y}$ LARGE $\Rightarrow$ $\hat{y}$ LARGE

IF $y=0$: $L(\hat{y}, 0) = \log(1-\hat{y})$ { WANT $\log 1-\hat{y}$ LARGE $\Rightarrow$ $\hat{y}$ SMALL

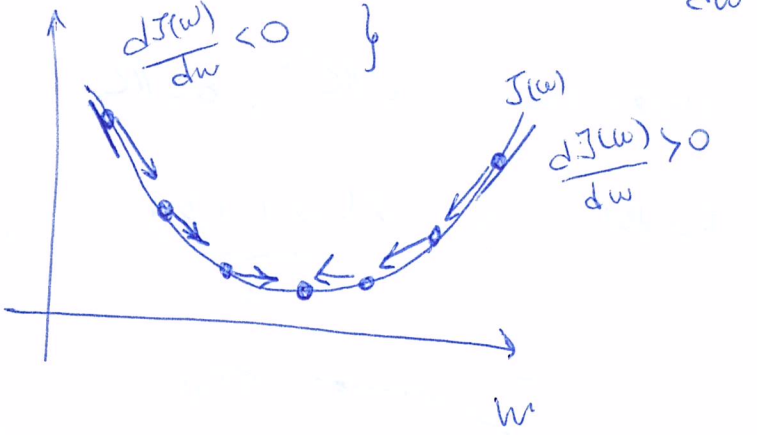
$\therefore$ COST FUNCTION: $J(w,b) = \frac{1}{m} \sum_{i=1}^{m} L(\hat{y}^{(i)}, y^{(i)})$

- GRADIENT DESCENT




REPEAT {

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




- LOGISTIC REGRESSION ON m EXAMPLES

$J=0 \; ; \; dw_1 = 0 \; ; \; dw_2 = 0 \; ; \; db = 0$

FOR $i = 1$ TO $m$

$$z^{(i)} = w^T x^{(i)} + b$$

$$a^{(i)} = \sigma(z^{(i)})$$

$$J \mathrel{+}= -\left[ y^{(i)} \log a^{(i)} + (1 - y^{(i)}) \log (1 - a^{(i)}) \right]$$

$$dz^{(i)} = a^{(i)} - y^{(i)}$$

$$dw_1 \mathrel{+}= x_1^{(i)} dz^{(i)}$$

$$dw_2 \mathrel{+}= x_2^{(i)} dz^{(i)}$$

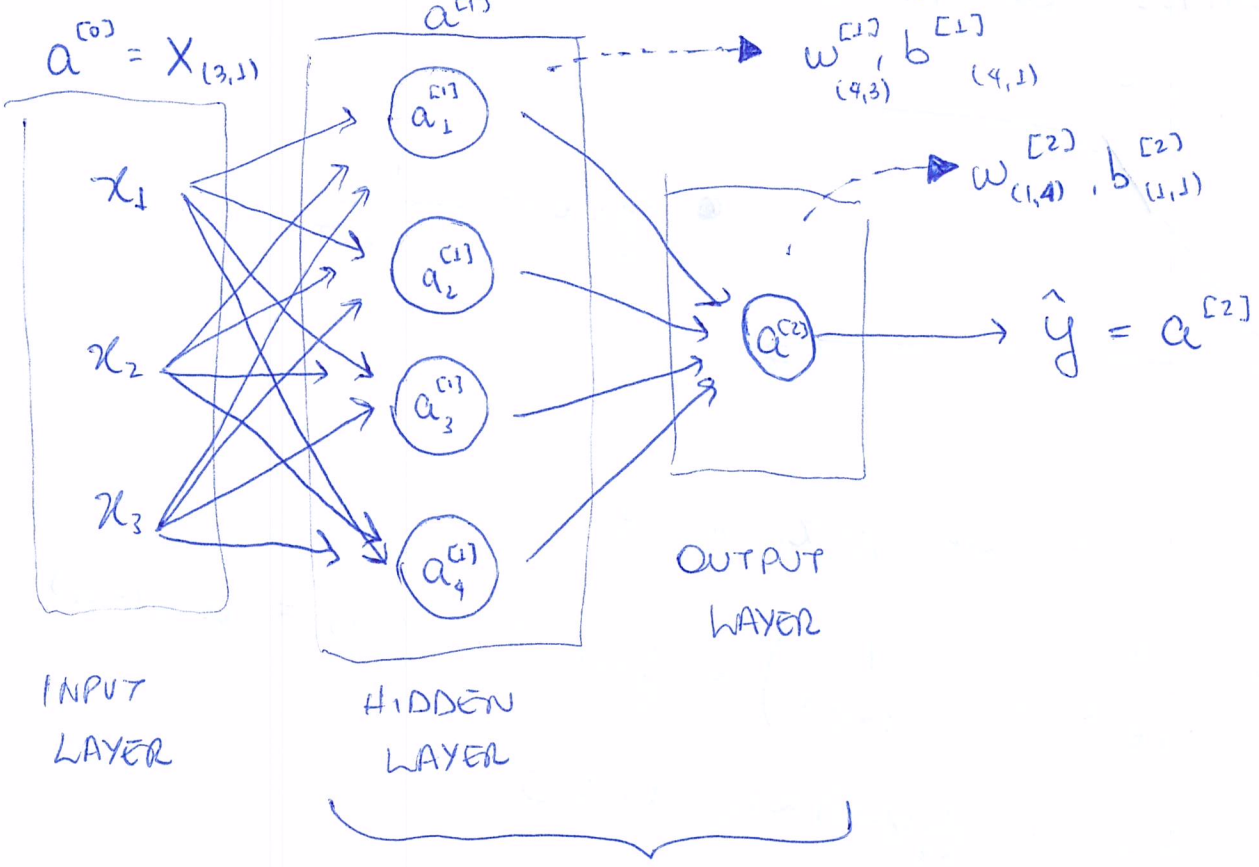
$n = 2$

$$db \mathrel{+}= dz^{(i)}$$

$J \mathrel{/}= m$

$dw_1 \mathrel{/}= m \; ; \; dw_2 \mathrel{/}= m \; ; \; db \mathrel{/}= m$

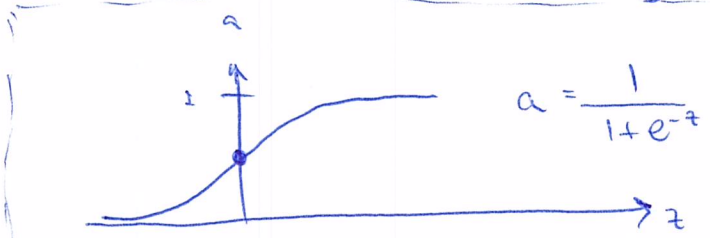
# - NEURAL NETWORK REPRESENTATION



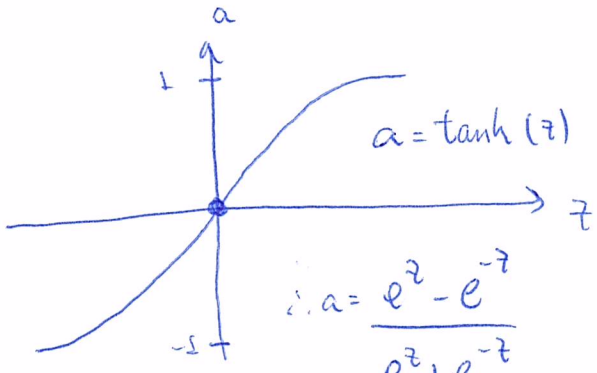

"2 LAYER NN"

$$z^{[1]} = \begin{bmatrix} -w_1^{[1]T}- \\ -w_2^{[1]T}- \\ -w_3^{[1]T}- \\ -w_4^{[1]T}- \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} b_1^{[1]} \\ b_2^{[1]} \\ b_3^{[1]} \\ b_4^{[1]} \end{bmatrix} = \begin{bmatrix} w_1^{[1]T}x + b_1^{[1]} \\ w_2^{[1]T}x + b_2^{[1]} \\ w_3^{[1]T}x + b_3^{[1]} \\ w_4^{[1]T}x + b_4^{[1]} \end{bmatrix} = \begin{bmatrix} z_1^{[1]} \\ z_2^{[1]} \\ z_3^{[1]} \\ z_4^{[1]} \end{bmatrix}$$

$$\therefore a^{[1]} = \begin{bmatrix} a_1^{[1]} \\ a_2^{[1]} \\ a_3^{[1]} \\ a_4^{[1]} \end{bmatrix} = \begin{bmatrix} \sigma(z_1^{[1]}) \\ \sigma(z_2^{[1]}) \\ \sigma(z_3^{[1]}) \\ \sigma(z_4^{[1]}) \end{bmatrix} = \sigma(z^{[1]})$$

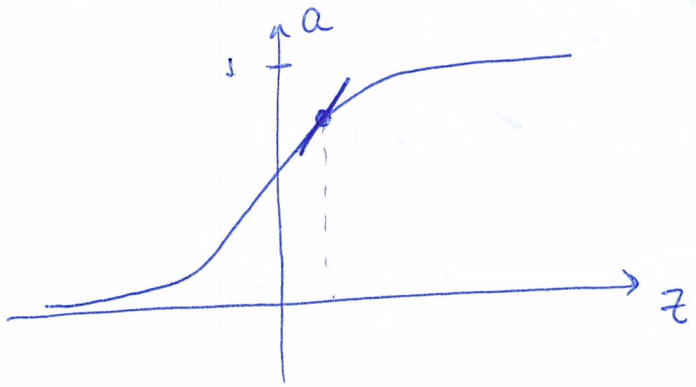



$0 \leq a \leq 1$




$-1 \leq a \leq 1$

# - DERIVATIVES OF ACTIVATION FUNCTIONS




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

$\frac{d}{dz} g(z)$ = Slope of $g(z)$ at $z$

$$= \frac{1}{1+e^{-z}} \left(1 - \frac{1}{1+e^{-z}}\right) = g(z) \cdot (1 - g(z)) = a(1-a)$$

$z = 10$ : $g(z) \approx 1 \Rightarrow \frac{d}{dz}(g(z)) \approx 1 \cdot (1-1) \approx 0$

$z = -10$ : $g(z) \approx 0 \Rightarrow \frac{d}{dz}(g(z)) \approx 0 \cdot (1-0) \approx 0$

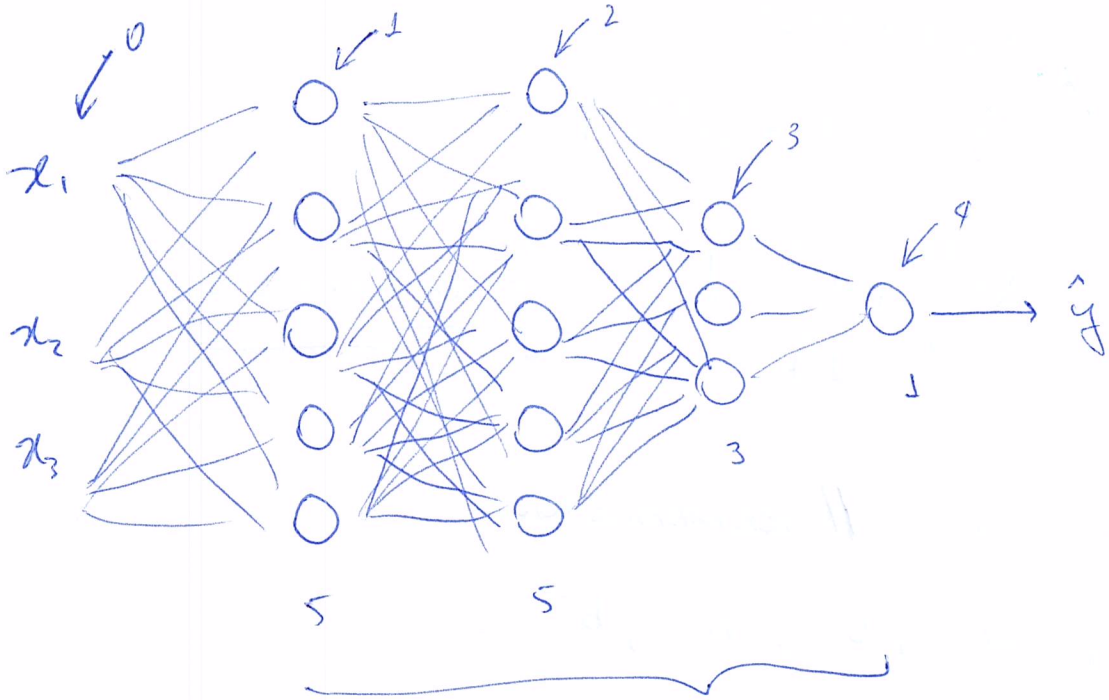
$z = 0$ : $g(z) = 1/2 \Rightarrow \frac{d}{dz}(g(z)) = \frac{1}{2} \cdot (1 - 1/2) = \frac{1}{4}$

$g'(z) = \frac{d}{dz} g(z) = a(1-a) \Rightarrow g'(z) = a(1-a)$

- $g(z) = \tanh(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$ $\quad g'(z) = \frac{d}{dz} g(z) = 1 - (\tanh(z))^2$

$a = g(z)$, $g'(z) = 1 - a^2$

- DEEP L-LAYER NETWORK




$L = 4$ (# LAYERS)

$n^{[l]}$ = # UNITS IN LAYER $l$

$a^{[l]}$ = activations in LAYER $l$

$n^{[1]} = 5, \quad n^{[2]} = 5, \quad n^{[3]} = 3, \quad n^{[4]} = n^{[L]} = 1, \quad n^{[0]} = 3$

- FORWARD PROPAGATION IN A DEEP NETWORK

USING THE LAST EXAMPLE: $x = a^{[0]}$

$$z^{[1]} = w^{[1]} x + b^{[1]}$$
$$a^{[1]} = g^{[1]}(z^{[1]})$$
$$z^{[2]} = w^{[2]} \cdot a^{[1]} + b^{[1]}$$
$$a^{[2]} = g^{[2]}(z^{[2]})$$
$$\vdots$$
$$z^{[4]} = w^{[4]} \cdot a^{[3]} + b^{[4]}$$
$$a^{[4]} = g^{[4]}(z^{[4]}) = \hat{y}$$

$$z^{[l]} = w^{[l]} \cdot a^{[l-1]} + b^{[l]}$$
$$a^{[l]} = g^{[l]}(z^{[l]})$$

$$z^{[1]} = W^{[1]} \cdot x + b^{[1]}$$

$$(3,1) \leftarrow (3,2) \cdot (2,1) + (3,1)$$

$$(n^{[1]},1) \leftarrow (n^{[1]},n^{[0]})(n^{[0]},1) + (n^{[1]},1)$$

$\therefore$ $dW^{[l]} = (n^{[l]}, n^{[l-1]})$

$db^{[l]} = (n^{[l]}, 1)$ } DIMENSIONS

- PARAMETERS VS. HYPERPARAMETERS

PARAMETERS: $W^{[1]}, b^{[1]}, W^{[2]}, b^{[2]}, \ldots$

HYPERPARAMETERS:

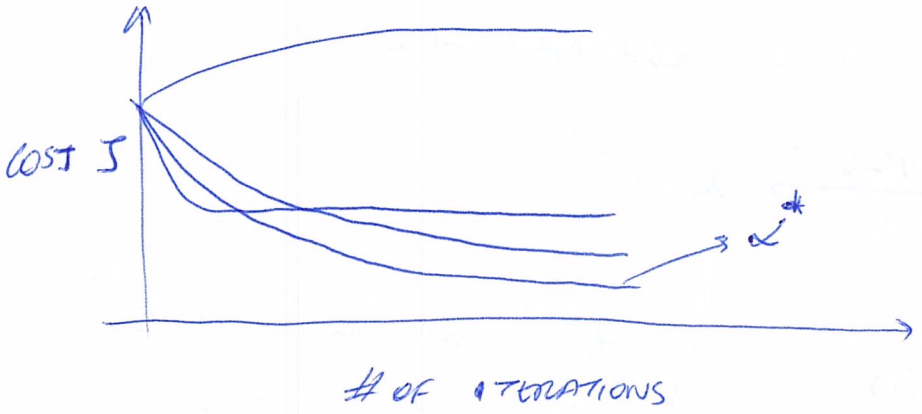
LEARNING RATE $\alpha$

# ITERATIONS

# HIDDEN LAYERS $L$

# HIDDEN UNITS $n^{[1]}, n^{[2]}, \ldots$

CHOICE OF ACTIVATION FUNCTION




APPLIED DEEP LEARNING IS A VERY EMPIRICAL PROCESS

# COURSE 2 - IMPROVING DEEP NEURAL NETWORKS

- TRAIN / VALID / TEST SETS SIZES

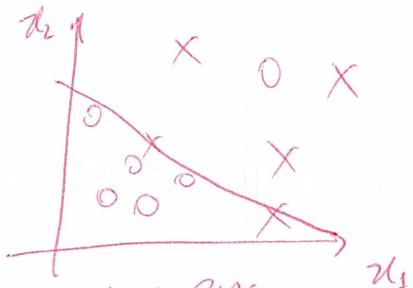
BEFORE: 70% (TRAIN) / 30% (TEST)

OR

60% (TRAIN) / 20% (VALID) / 20% (TEST)

BIG DATA ERA: 1,000,000 + => 98% / 1% / 1%

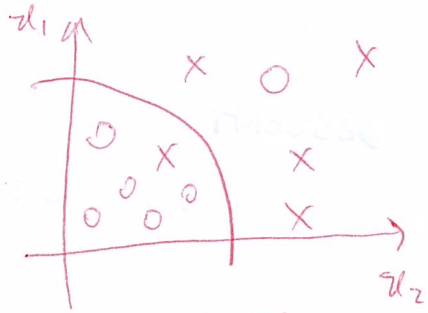
- BIAS AND VARIANCE



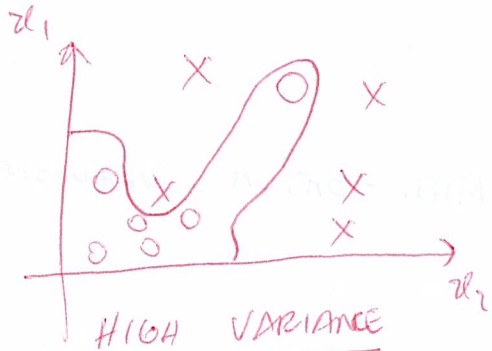

HIGH BIAS

||

UNDERFITTING




JUST RIGHT




HIGH VARIANCE

||

OVERFITTING

- SOLVE BIAS / VARIANCE PROBLEMS

HIGH BIAS? ⟶
- BIGGER NETWORK
- BIGGER TRAINING SET

HIGH VARIANCE? ⟶
- MORE DATA (MORE VARIABILITY)
- REGULARIZATION

- REGULARIZATION (L2)

$$J(w^{[1]}, b^{[1]}, \dots, w^{[L]}, b^{[L]}) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(\hat{y}^{(i)}, y^{(i)}) + \frac{\lambda}{2n} \sum_{l=1}^{L} \|w^{[l]}\|_F^2$$

COST FUNCTION — LOSS FUNCTION — REGULARIZATION FUNCTION

ELIMINATES SOME "EXTRA" NEURONS.

($\lambda$ := REGULARIZATION PARAMETER)

- NORMALIZING INPUTS

$x_1$ : 1 ... 100
$x_2$ : 0 ... 1

} $w_1 \gg w_2$ => MORE TIME / ITERATIONS FOR GRADIENT DESCENT

- REGULARIZATION WITH DROPOUT

DURING TRAINING ELIMINATES (RANDOMLY) NODES, USING A VARIABLE OF KEEP-PROB.

* IMPORTANT : ONLY IN TRAINING

- MINI-BATCH GRADIENT DESCENT

CREATES BATCHS FOR TRAINING SET (M OBSERVATIONS), AND DIVIDES SETS.

MINI-BATCH SIZE NOT TOO BIG / SMALL

⇓

FASTEST LEARNING
- VECTORIZATION
- MAKE PROGRESS WITHOUT PROCESSING ENTIRE SET

OR
- BATCH GRADIENT DESCENT (SIZE=m)
  ⇓
  TOO LONG PER ITERATION
- STOCHASTIC GRADIENT DESCENT (SIZE=1)
  ⇓
  LOSE SPEED

IF SMALL TRAINING SET : USE BATCH GRADIENT DESCENT
($m \leq 2000$)

TYPICAL SIZES OF MINI-BATCHS: 64, 128, 256, 512
$2^6$, $2^7$, $2^8$, $2^9$

- ADAM OPTIMIZATION ALGORITHM

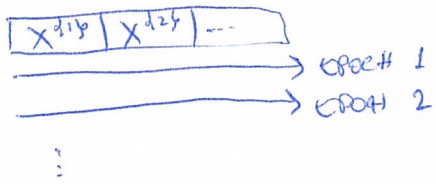
MOMENTUM ⊕ RMSPROP = ADAM

"GRADIENT DESCENT MOMENTUM"

ADAM := ADAPTED MOMENT ESTIMATION

- LEARNING RATE DECAY

1 EPOCH = 1 PASS THROUGH THE DATA




$$\alpha = \frac{1}{1 + \text{DECAY\_RATE} * \text{EPOCH\_NUM}} \alpha_0$$

$\alpha_0 = 0.2$

DECAY_RATE = 1

| EPOCH | $\alpha$ |
|---|---|
| 1 | 0.1 |
| 2 | 0.67 |
| 3 | 0.5 |
| ⋮ | ⋮ |

- HYPERPARAMETER TUNING

1° $\alpha$

2° $\beta$

$\beta_1$, $\beta_2$, $\varepsilon$
0.9, 0.999, $10^{-8}$

3° # LAYERS

2° # HIDDEN UNITS

3° LEARNING RATE DECAY

2° MINI-BATCH SIZE

BABYSITTING ONE MODEL vs. TRAINING MANY MODELS IN PARALLEL

- BATCH NORMALIZATION

FEATURE NORMALIZATION $\Rightarrow$ SPEED UP LEARNING

USING BATCH NORMALIZATION :

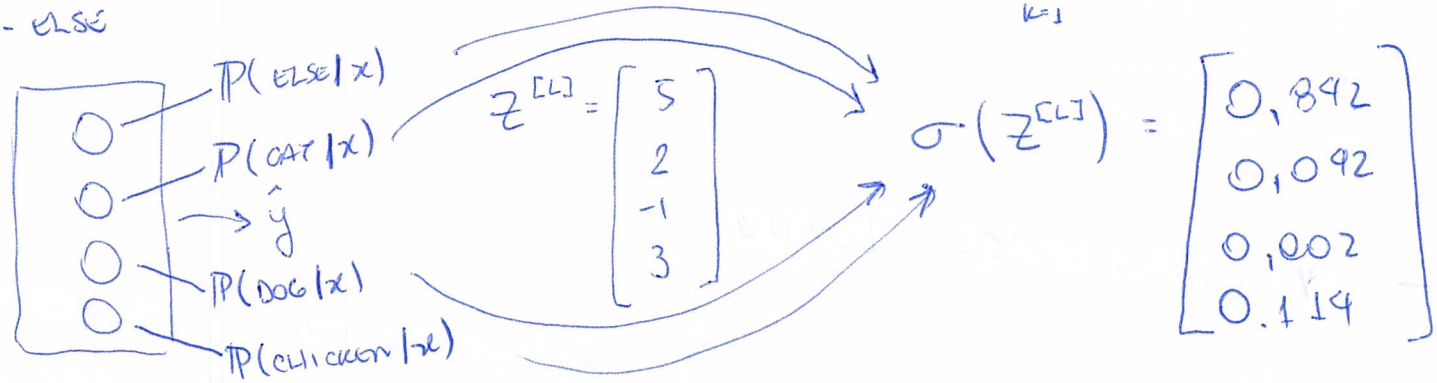
- WE CAN USE HIGHER LEARNING RATES, CAUSE ACTIVATION VALUES ARE NOW
- IT REDUCES OVERFITTING BECAUSE IT HAS A SLIGHT REGULARIZATION EFFECTS (ADD NOISE TO EACH HIDDEN LAYER)

- SOFTMAX REGRESSION (MULTI-CLASS CLASSIFICATION)

1- CAT
2- DOG
3- CHICKEN
0- ELSE

$C = \#\text{CLASSES} = 4$

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}, \quad j = 1, \dots, K$$




"HARD MAX" $\Rightarrow \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$ AND IF $C = 2$ $\Rightarrow$ SOFTMAX = LOGISTIC REGRESSION

- loss function

$$\mathcal{L}(\hat{y}, y) = -\sum_{j=1}^{4} y_j \cdot \log \hat{y}_j \qquad y = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \ \hat{y} = \begin{bmatrix} 0.3 \\ 0.2 \\ 0.1 \\ 0.4 \end{bmatrix}$$

$$= -\log \hat{y}_2$$

$$\text{MIN } \mathcal{L}(\hat{y}, y) = \text{MIN} - \log \hat{y}_2 \Rightarrow \text{MAX } \hat{y}_2$$

# COURSE 3 - STRUCTURING ML PROJECTS

WAYS TO ANALYZE PROBLEMS OF ML IN A WAY TO FIND SOLUTIONS (GOOD ONES).

## CHAIN OF ASSUMPTIONS IN ML:

- FIT TRAINING SET WELL
- " DEV " "
- " TEST " "
- PERFORMS WELL IN REAL WORLD

## USING A SINGLE NUMBER EVALUATION METRIC:

| CLASSIFIER | PRECISION* | RECALL* | F1 SCORE |
|---|---|---|---|
| A | 95% | 90% | 92.4% |
| B | 98% | 85% | 91% |

$$F1\ SCORE = \frac{2}{\frac{1}{P} + \frac{1}{R}}$$ "HARMONIC MEAN"

↳ TO COMPARE PRECISION AND RECALL IN ONLY ONE NUMBER

* 

| | | PREDICTED: NEGATIVE | PREDICTED: POSITIVE |
|---|---|---|---|
| ACTUAL | NEGATIVE | TN | FP |
| ACTUAL | POSITIVE | FN | TP |

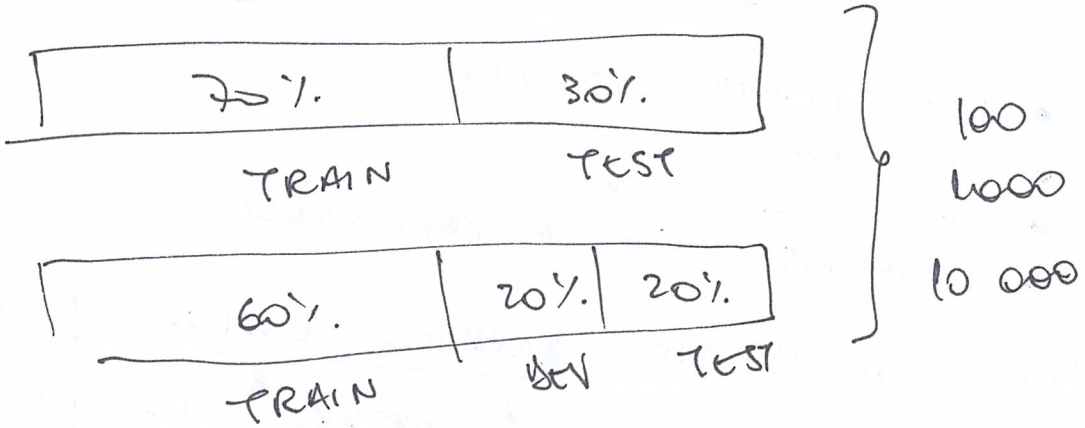
$$PRECISION = \frac{TP}{TP + FP}$$

$$RECALL = \frac{TP}{TP + FN}$$ (SENSITIVITY)

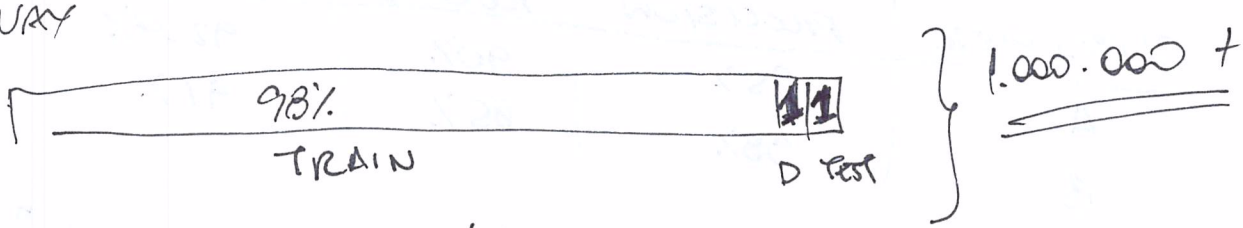
$$SPECIFICITY = \frac{TN}{TN + FP}$$

## SIZE OF DEV/TRAIN/TEST SET

- OLD WAY OF SPLITTING




- NEW WAY




## WHEN TO CHANGE DEV/TESTS SETS AND METRICS

→ IMPORTANT TO ANALYZE ERROR PUBLIC OVER TAX ERROR

$$\text{ERROR} = \frac{1}{\sum w^{(i)}} \frac{1}{M_{dev}} \sum_{i=1}^{M_{dev}} w^{(i)} \, \mathbb{1}\{y_{PRED}^{(i)} \neq y^{(i)}\}$$

$$w^{(i)} = \begin{cases} 1, & \text{IF } x^{(i)} \text{ IS NON-PORN} \\ 10, & \text{IF } x^{(i)} \text{ IS PORN} \end{cases}$$

## UNDERSTANDING HUMAN-LEVEL PERFORMANCE

- HUMAN-LEVEL ERROR (PROXY FOR BAYES ERROR)

↕ "AVOIDABLE BIAS"

- TRAINING ERROR

↕ "VARIANCE"

- DEV ERROR

## SURPASSING HUMAN-LEVEL PERFORMANCE

- TEAM OF HUMANS — 0.5% → PROBABLY IS THE BAYES ERROR

1% — 0.1% "AVOIDABLE BIAS"

- ONE HUMAN

- TRAINING ERROR — 0.6%

0.2% "VARIANCE"

- DEV ERROR — 0.8%

# IMPROVING YOUR MODEL PERFORMANCE

HUMAN-LEVEL

↕ AVOIDABLE BIAS →
- TRAIN BIGGER MODEL
- TRAIN LONGER/BETTER OPTIMIZATION ALGORITHMS
- NN ARCHITECTURE / HYPERPARAMETERS SEARCH (RNN, CNN) *

TRAINING ERROR

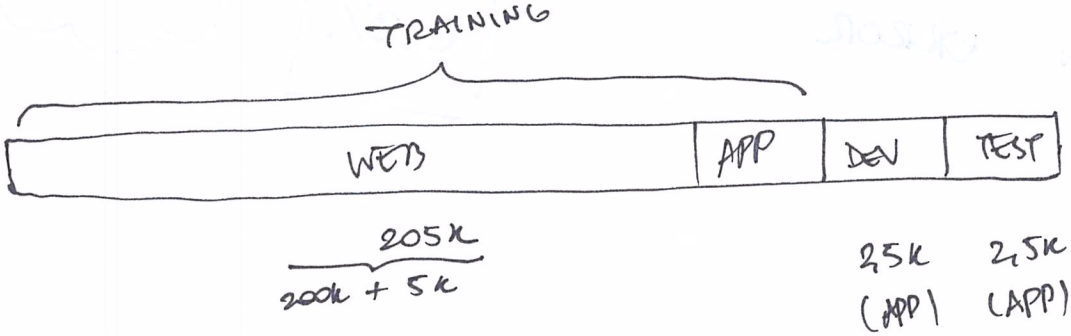
↕ VARIANCE →
- MORE DATA
- REGULARIZATION L2, DROPOUT, DATA AUGMENTATION

DEV ERROR

- (*)

## MISMATCHED TRAINING AND DEV/~~SELF~~ TEST SET

- 200k obs from WEB
- 10k obs from APP (THE ONES YOU WANT TO IDENTIFY)

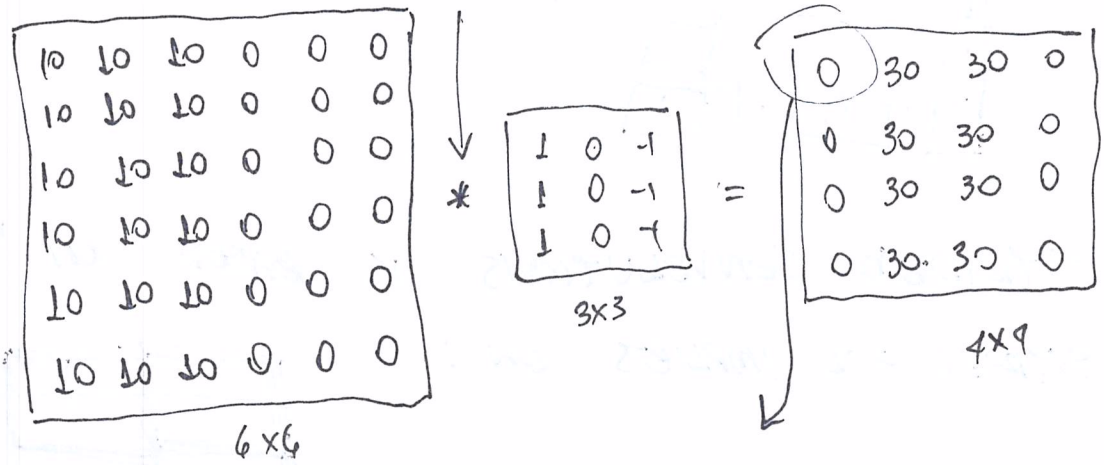


# COURSE 4 - CONVOLUTIONAL NEURAL NETWORKS

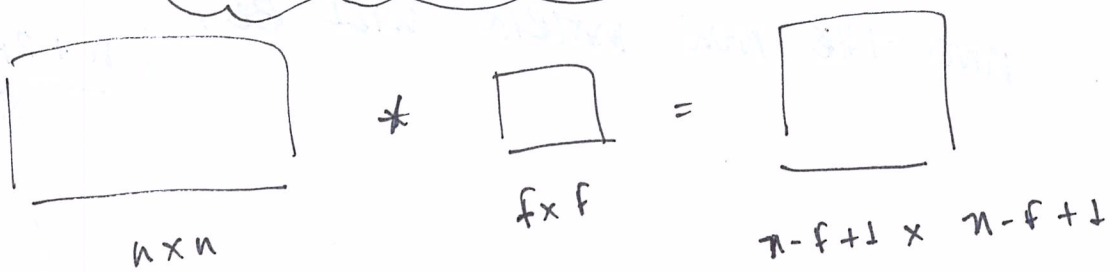
VERY USED ON COMPUTER VISION

How convolution works?

CONVOLUTION OPERATOR




$10 \cdot 1 + 10 \cdot 1 + 10 \cdot 1 + 10 \cdot 0 + 10 \cdot 0 + 10 \cdot 0 + 10 \cdot (-1) + 10 \cdot (-1) + 10 \cdot (-1) = 0$



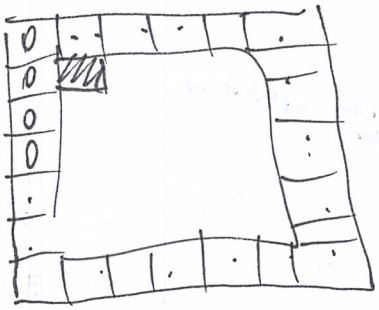

"VALID" AND "SAME" CONVOLUTIONS

- "VALID": $n \times n \quad * \quad f \times f \rightarrow n-f+1 \times n-f+1$
- "SAME": PAD (1) SO THE OUTPUT SIZE IS THE SAME AS THE INPUT SIZE.

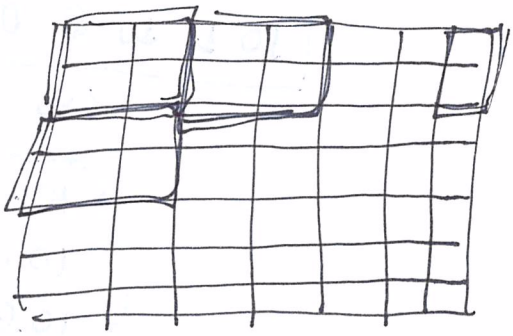
$$n + 2p - f + 1 = n \Rightarrow 2p = f - 1 \Rightarrow$$

$$\boxed{p = \frac{f-1}{2}}$$

(1) PADDING IS TO ADD COLUMNS TO THE IMAGES IN ORDER TO ENHANCE INFLUENCE OF CERTAIN VALUE OF THE IMAGE.

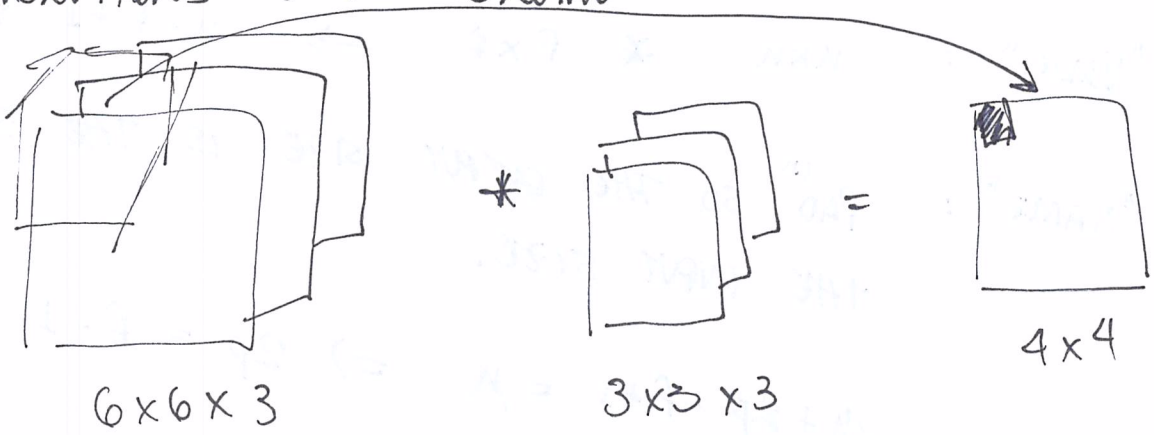


STRIDED CONVOLUTIONS IS BASED ON "JUMPS", SO STRIDE = 2 IMPLIES ON:



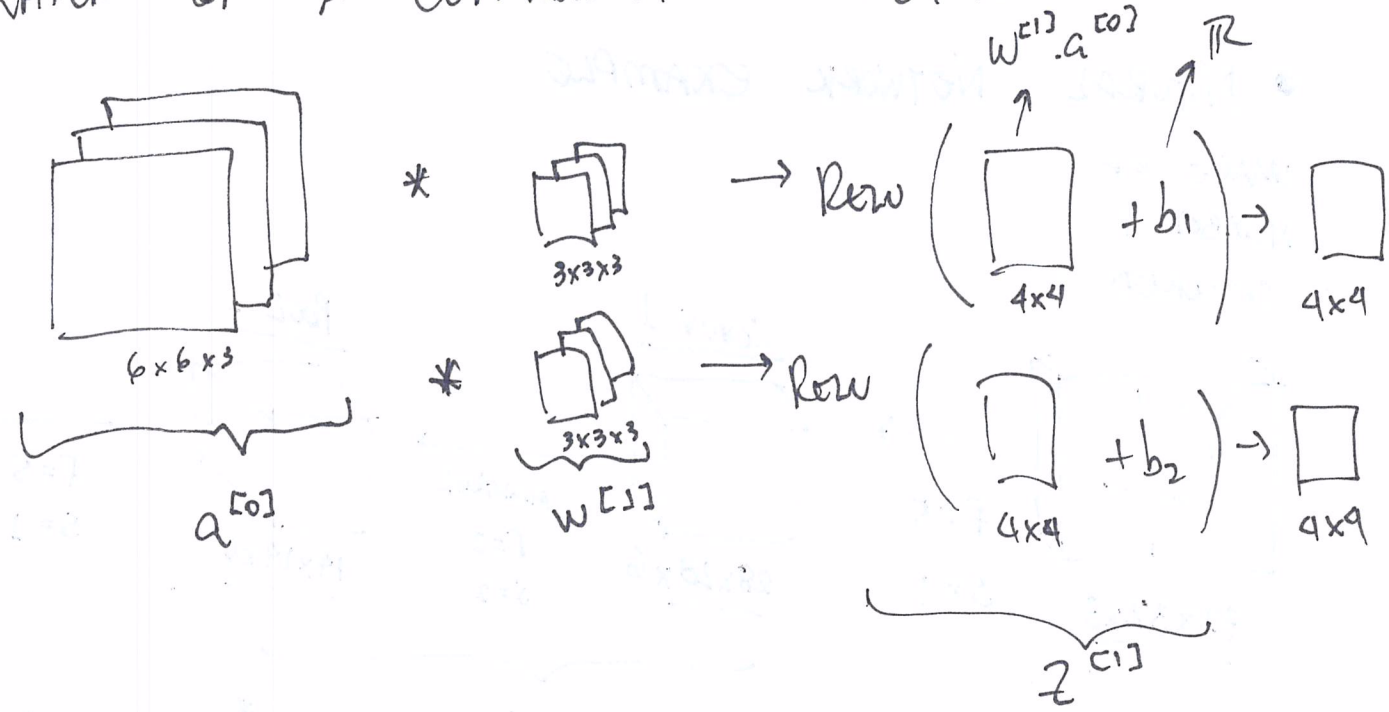
AND THE NEW MATRIX WILL BE: $\left(\frac{n+2p-f}{s}+1\right) \times \left(\frac{n+2p-f}{s}+1\right)$

CONVOLUTIONS OVER VOLUME:



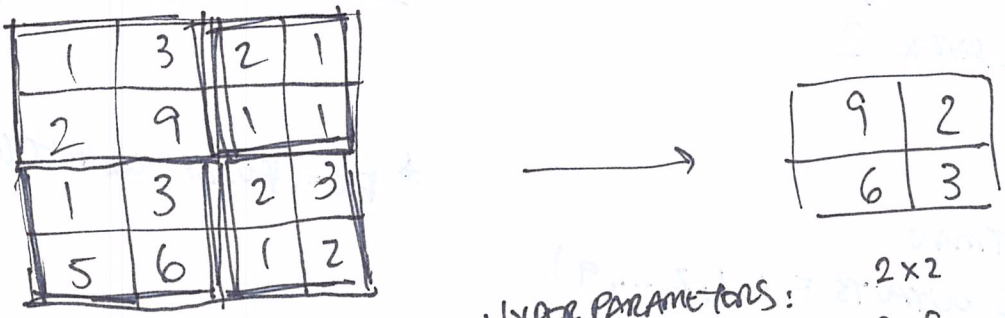

"MULTIPLE FILTERS"

- LAYER OF A CONVOLUTIONAL NETWORK




$$\begin{cases} z^{[1]} = W^{[1]} a^{[0]} + b^{[1]} \\ a^{[1]} = g(z^{[1]}) \end{cases}$$

- POOLING LAYERS : MAX POOLING

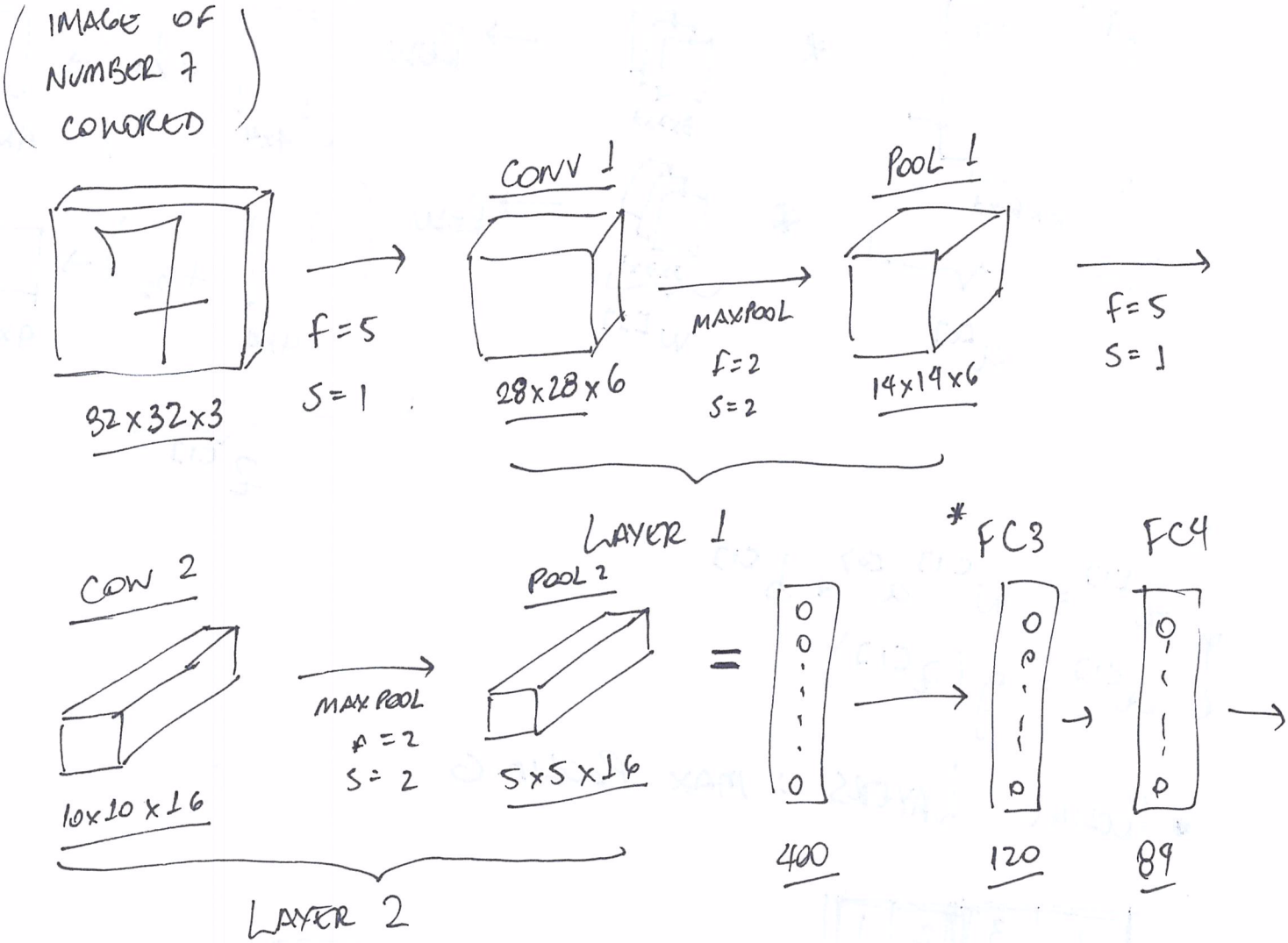



HYPERPARAMETERS: f×f

f = 2 (FILTER DIMENSION)

S = 2 (STRIDE)

- NEURAL NETWORK EXAMPLE




* FC = FULLY CONNECTED LAYER

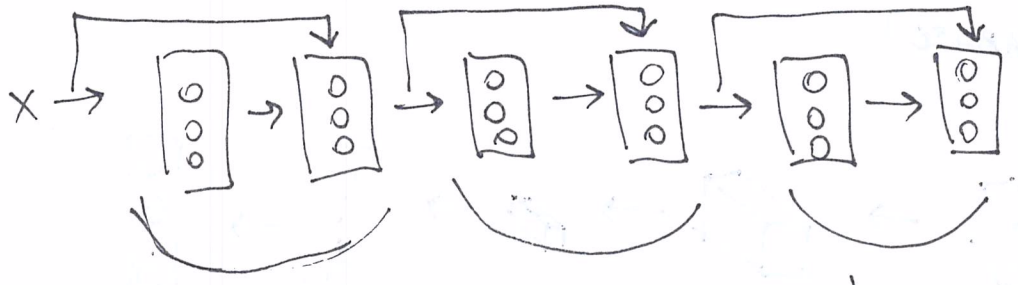
→ ○ SOFTMAX
(10 OUTPUTS = 0, 1, 2, ..., 9)

- WHY CONVOLUTIONS?

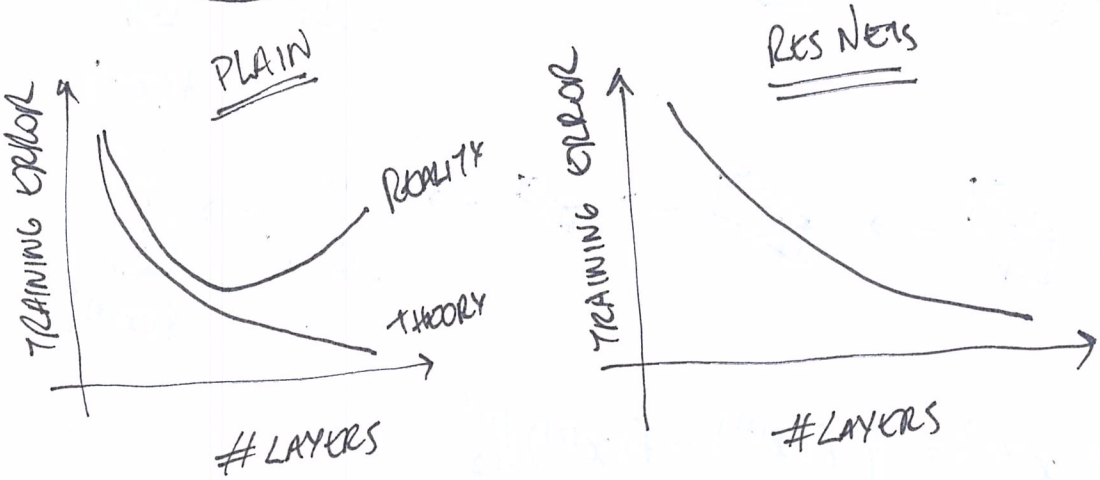
→ PARAMETER SHARING: A FEATURE DETECTOR THAT'S USEFUL IN ONE PART OF THE IMAGE MIGHT BE USEFUL IN ANOTHER PART OF THE IMAGE.

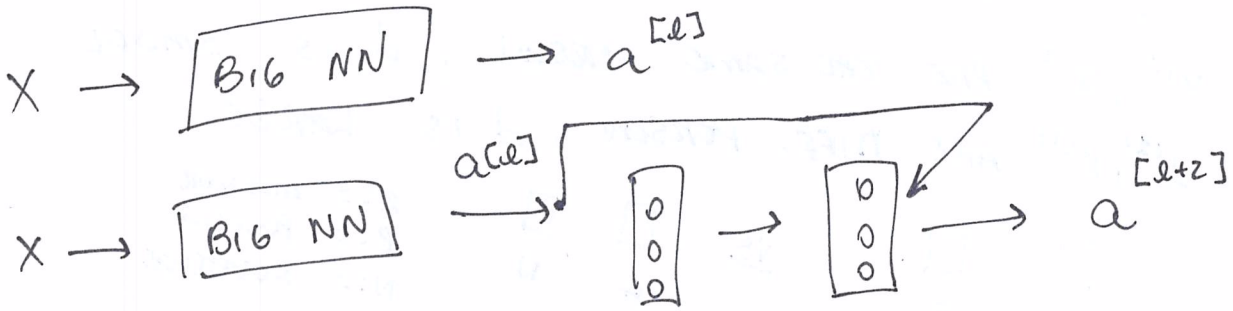
→ SPARSITY CONNECTIONS: IN EACH LAYER, EACH OUTPUT VALUE DEPENDS ONLY ON A SMALL NUMBER OF INPUTS.

- RESIDUAL NETWORK (RESNETS)




3 RESIDUAL BLOCKS




$$a^{[l+2]} = g\left(z^{[l+2]} + a^{[l]}\right)$$

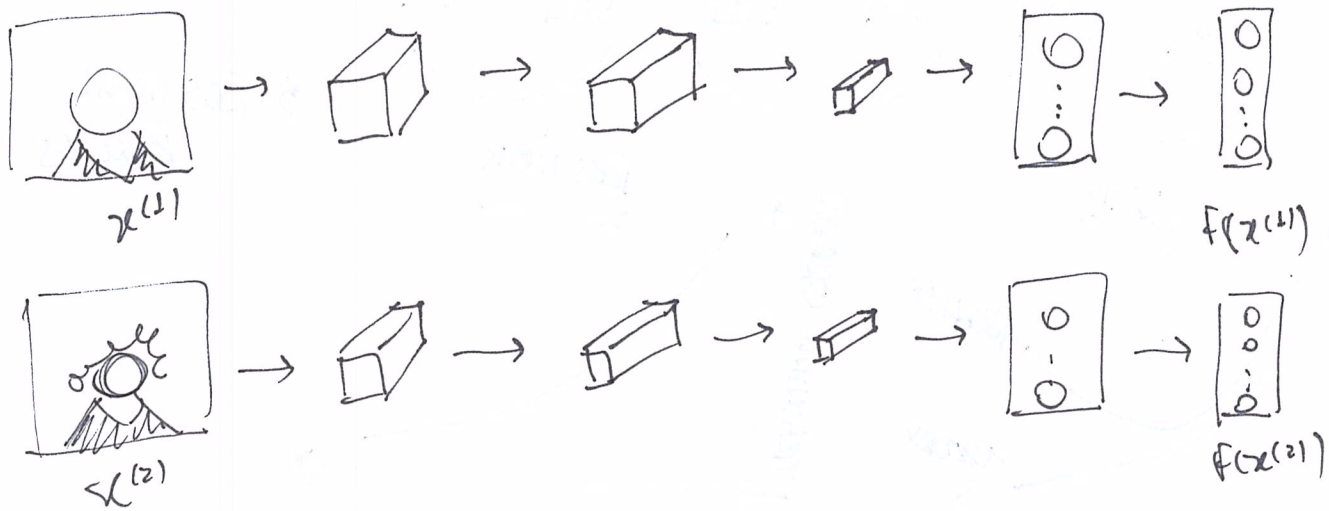
$$= g\left(\underbrace{w^{[l+2]} a^{[l+1]} + b^{[l+2]}}_{=0} + a^{[l]}\right) \quad \overset{(*)\ \text{RELU}}{=} g(a^{[l]}) = a^{[l]}$$

RESIDUAL NETWORKS SKIP CONNECTIONS OR SHORT-CUTS TO JUMP OVER SOME LAYERS, THE MOTIVATION IS TO AVOID THE PROBLEM OF VANISHING GRADIENTS BY REUSING ACTIVATION FROM A PREVIOUS LAYER UNTIL THE LAYER NEXT TO THE CURRENT ONE HAVE LEARNED ITS WEIGHTS.

THE INTUITION ON WHY THIS WORKS IS THAT THE NEURAL NETWORK COLLAPSES INTO FEWER LAYERS INITIALLY, AND AFTER EXPAND

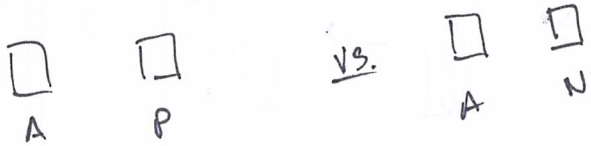
- SIAMESE NETWORK

USED TO SOLVE ONE SHOT LEARNING (IDENTIFY IF PHOTO EXISTS ON DATABASE)




$$d(x^{(1)}, x^{(2)}) = \| f(x^{(1)}) - f(x^{(2)}) \|_2^2$$

- IF $x^{(1)}, x^{(2)}$ ARE THE SAME PERSON, $d$ IS SMALL
- IF $x^{(2)}, x^{(2)}$ ARE DIFF. PERSON, $d$ IS LARGE




A := ANCHOR
P := POSITIVE
N := NEGATIVE

LOSS FUNCTION (= "TRIPLET LOSS") : $\mathcal{L}(A,P,N) = \max(\|f(A) - f(P)\|^2 - \|f(A) - f(N)\|^2 + \alpha, 0)$

$\|f(A) - f(P)\|^2 + \alpha \leq \|f(A) - f(N)\|^2$

$d(A,P)$ $\qquad$ $d(A,N)$

# COURSE 5 - SEQUENCE MODELS

## RECURRENT NEURAL NETWORKS

### WHY RNNs?

SPEECH RECOGNITIONS  => "YOU ARE THE SUNSHINE of MY LIF

SENTIMENT CLASSIFICATIONS "HATED THIS MOVIE" => ★ ☆ ☆ ☆ ☆

DNA SEQ. ANALYSIS AGCCCCFGTGACC => PATTERN RECOG.

MACHINE TRANSLATION "HOW ARE YOU?" => "COMO VOCÊ ESTÁ?"

NAME RECOGNITION "HARRY POTTER INVENTED A SPELL" => HARRY POTTER

E.G.: RECOGNIZING NAMES ON PHRASES

HARRY POTTER INVENTED A SPELL

NOTATION

x: $x^{<1>}$ $x^{<2>}$ $x^{<3>}$ ... $x^{<t>}$ .. $x^{<5>}$

y: 1 1 0 .. ... 0

$y^{<1>}$ $y^{<2>}$ $y^{<3>}$ . . . . $y^{<5>}$

$X^{(i)<t>}$ := WORD IN POSITION t OF SAMPLE i

$Y^{(i)<t>}$ := ANSWER IN POSITION t OF SAMPLE i

$T_x^{(i)}$ := NUMBER OF WORDS OF SAMPLE i

$T_y^{(i)}$ := NUMBER OF ANSWER OF SAMPLE i

SOME CASES $T_x = T_y$ (LIKE THIS EXAMPLE ABOVE). BUT FOR SENTIMENT ANALYSIS, FOR EXAMPLE, $T_x$ = # OF WORDS IN TEXT, $T_y$ = #ANSWERS (0 OR 1) = GOOD OR BAD.

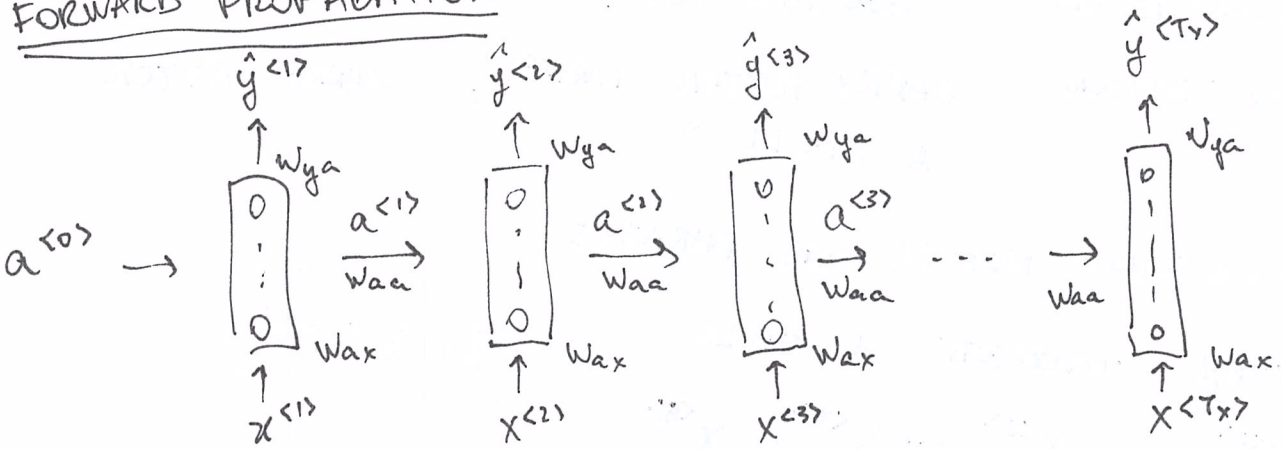
EXPLANATORY FEATURES TO INPUT :

VOCABULARY = [ A : 1 ; HARRY : 3675 ; POTTER : 6800 ; ZULU : 10000 ]

$x^{<1>}$ = "HARRY" → 3675

$x^{<2>}$ = "POTTER" → 6800

## FORWARD PROPAGATION




$a^{<0>} = \vec{0}$

$a^{<1>} = g_1(W_{aa}\, a^{<0>} + W_{ax}\, x^{<1>} + b_a)$ $\quad g_1 \rightarrow$ TANH/RELU

$\hat{y}^{<1>} = g_2(W_{ya}\, a^{<1>} + b_y)$ $\quad g_2 \rightarrow$ SIGMOID

$\therefore$

$$a^{<t>} = g(W_{aa}\, a^{<t-1>} + W_{ax}\, x^{<t>} + b_a)$$

$$\hat{y}^{<t>} = g(W_{ya}\, a^{<t>} + b_y)$$

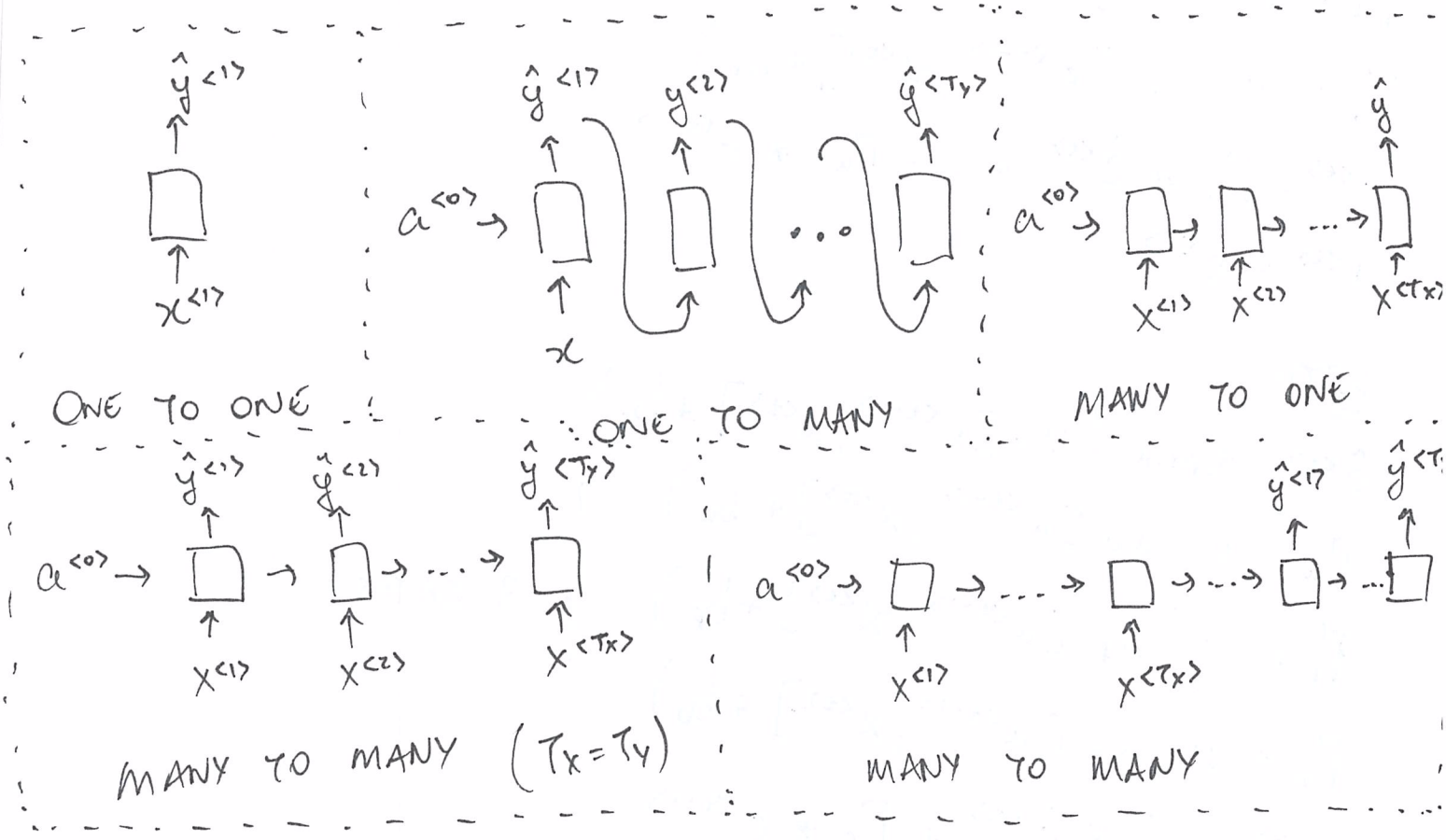
FOR SIMPLIFICATION : $W_a = [\, W_{aa} \vdots W_{ax} \,]$

FOR BACK PROPAGATION :

$$\mathcal{L}^{<t>}(\hat{y}^{<t>}, y^{<t>}) = -y^{<t>} \cdot \log \hat{y}^{<t>} - (1 - y^{<t>}) \log (1 - \hat{y}^{<t>})$$

$$\therefore \mathcal{L}(\hat{y}, y) = \sum_{t=1}^{T_y} \mathcal{L}^{\langle t \rangle}(\hat{y}^{\langle t \rangle}, y^{\langle t \rangle})$$

## RNN TYPES




## GRU & LSTM

MOTIVATION: "THE (DOGS), WHO ATE A HOT YESTERDAY, (WERE) TIRED TODAY"

BUILD A MODEL THAT RELATES WORDS DISTANT FROM EACH OTHER.

GRU: GATED RECURRENT UNIT

LSTM: LONG SHORT TERM MEMORY

## GRU

$$\tilde{c}^{\langle t \rangle} = \tanh\left(W_c\left[\Gamma_r * c^{\langle t-1 \rangle}, x^{\langle t \rangle}\right] + b_c\right)$$

$$\Gamma_u = \sigma\left(W_u\left[c^{\langle t-1 \rangle}, x^{\langle t \rangle}\right] + b_u\right)$$

$$\Gamma_r = \sigma\left(W_r\left[c^{\langle t-1 \rangle}, x^{\langle t \rangle}\right] + b_r\right)$$

} 2 GATES

$$c^{\langle t \rangle} = \Gamma_u * \tilde{c}^{\langle t \rangle} + (1 - \Gamma_u) * c^{\langle t-1 \rangle}$$

$$a^{\langle t \rangle} = c^{\langle t \rangle}$$

## LSTM

$$\tilde{c}^{\langle t \rangle} = \tanh\left(W_c\left[a^{\langle t-1 \rangle}, x^{\langle t \rangle}\right] + b_c\right)$$

$$\Gamma_u = \sigma\left(W_u\left[a^{\langle t-1 \rangle}, x^{\langle t \rangle}\right] + b_u\right)$$

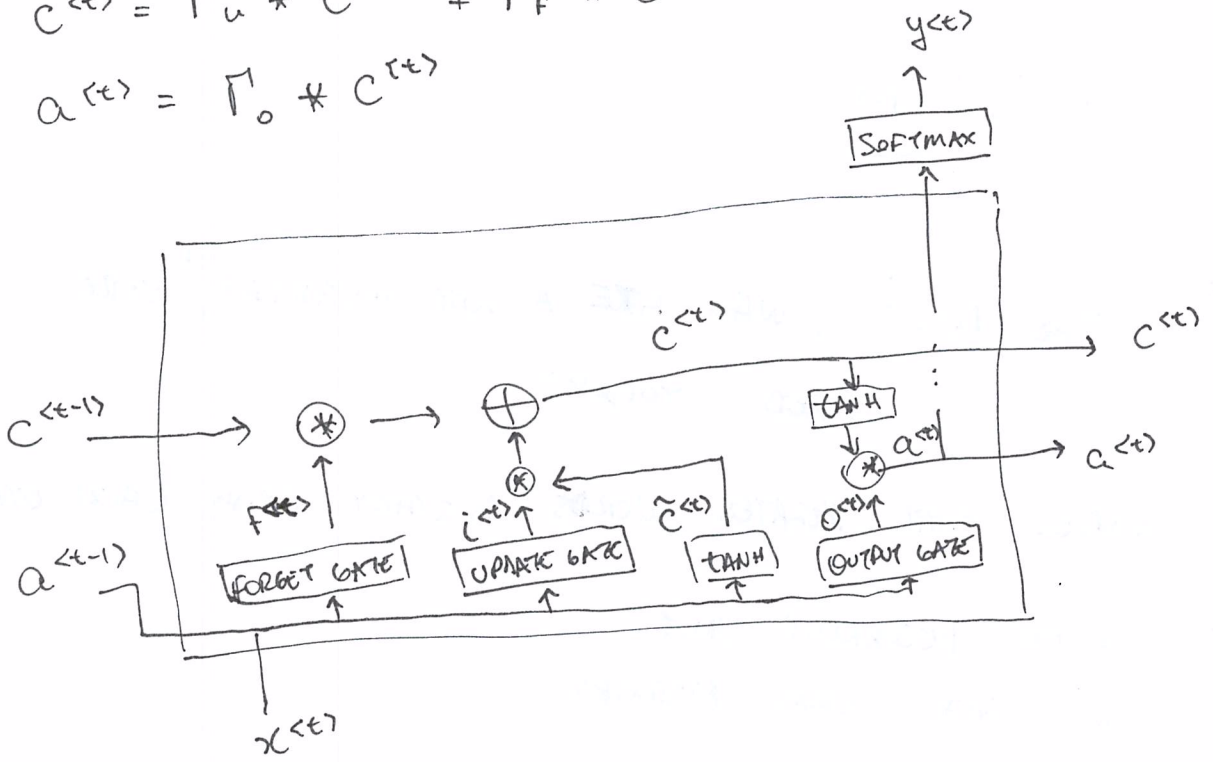
$$\Gamma_f = \sigma\left(W_f\left[a^{\langle t-1 \rangle}, x^{\langle t \rangle}\right] + b_f\right)$$

$$\Gamma_o = \sigma\left(W_o\left[a^{\langle t-1 \rangle}, x^{\langle t \rangle}\right] + b_o\right)$$

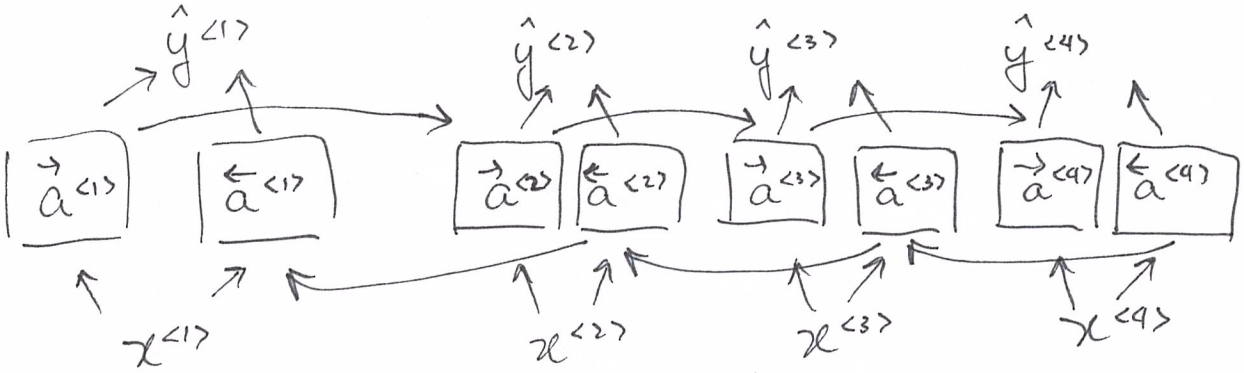
} 3 GATES

$$c^{\langle t \rangle} = \Gamma_u * \tilde{c}^{\langle t \rangle} + \Gamma_f * c^{\langle t-1 \rangle}$$

$$a^{\langle t \rangle} = \Gamma_o * c^{\langle t \rangle}$$



# BIDIRECTIONAL RNN



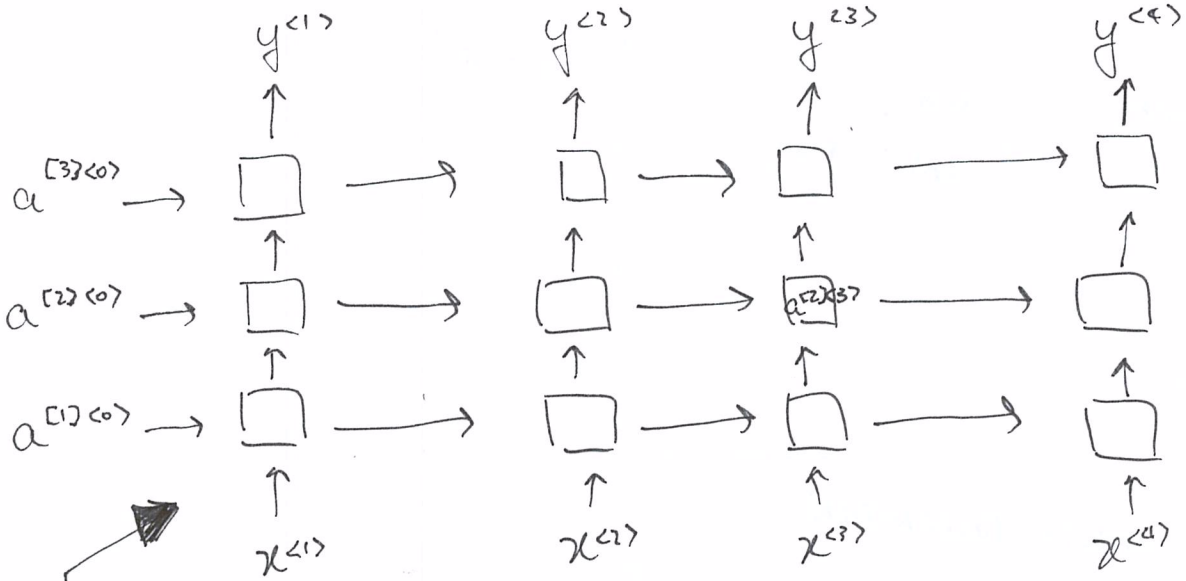
$$\therefore \hat{y}^{<t>} = g\left(W_y\left[\vec{a}^{<t>}, \overleftarrow{a}^{<t>}\right] + b_y\right)$$

THESE BLOCKS CAN BE GRU AND LSTM

ADVANTAGE: CAN PREDICT ANY TYPE OF TEXT (MIDDLE, BEGINNING OR END OF TEXT)

DISAVANTAGE: YOU NEED THE ENTIRE SEQUENCE

# DEEP RNNs



$$a^{[2]<3>} = g\left(W_a^{[2]}\left[a^{[2]<2>}, a^{[1]<3>}\right] + b_a^{[2]}\right)$$

RNN, GRU, LSTM, BRNN

## WORD REPRESENTATION

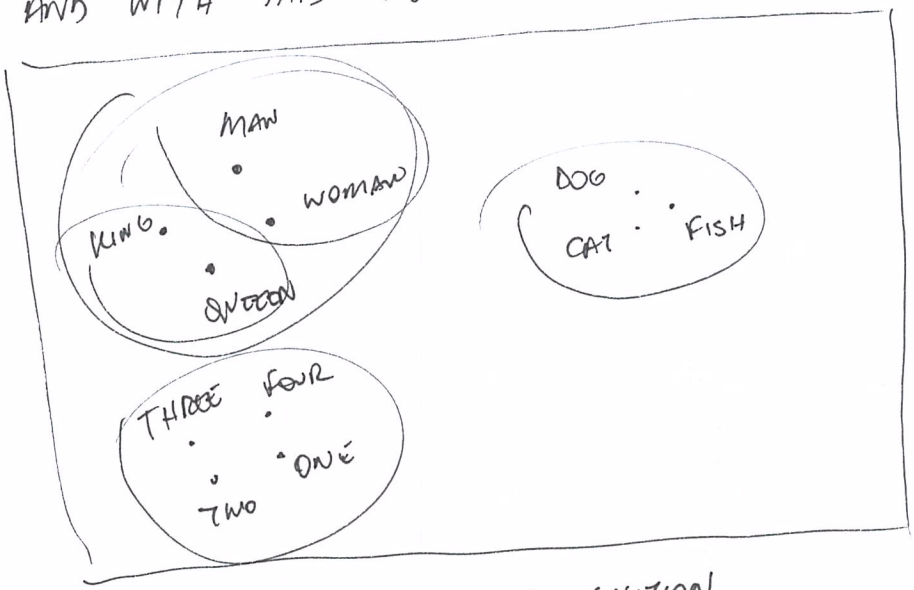
$V = [A, AARON, \ldots, ZULU, \langle UNKNOWN \rangle]$ , $|V| = 10\,000$

## FEATURE REPRESENTATION

| | MAN | WOMAN | KING | QUEEN | APPLE | ORANGE |
|---|---|---|---|---|---|---|
| GENDER | -1 | 1 | -0.95 | 0.97 | 0.00 | 0.01 |
| ROYAL | 0.01 | 0.02 | 0.93 | 0.95 | -0.01 | 0.00 |
| AGE | 0.03 | 0.02 | 0.7 | 0.69 | 0.03 | -0.02 |
| FOOD | 0.04 | 0.01 | 0.02 | 0.01 | 0.95 | 0.97 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| SIZE | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| COST | | | | | | |

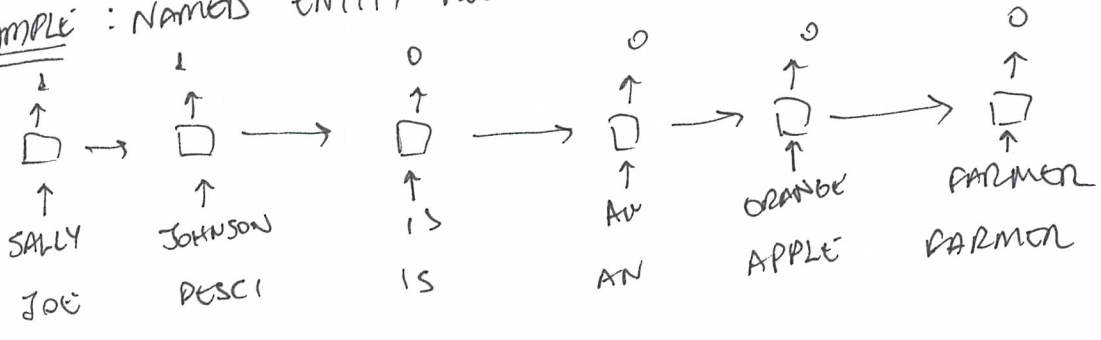
$\hookrightarrow e_{5391}$ $\hookrightarrow e_{9853}$

AND WITH THIS REPRESENTATION CAN CREATE CLUSTERS:




t-SNE

EXAMPLE: NAMED ENTITY RECOGNITION

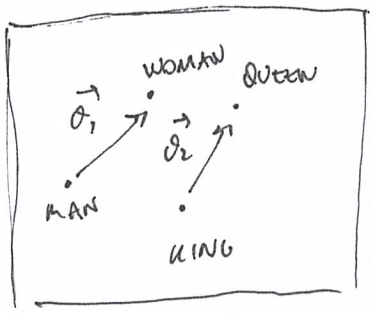


IT IS POSSIBLE FOR THE MODEL TO IDENTIFY CLOSURE FROM OTHER EXAMPLES (ANALOGIES). IN THAT MANNER:

$$e_{MAN} - e_{WOMAN} \approx \begin{bmatrix} -2 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$e_{KING} - e_{QUEEN} \approx \begin{bmatrix} -2 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$$e_{MAN} - e_{WOMAN} \approx e_{KING} - e_? \Rightarrow e_? = e_{QUEEN}$$

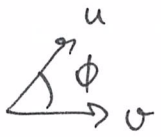



$\vec{v}_1$ AND $\vec{v}_2$ ARE VERY SIMILAR

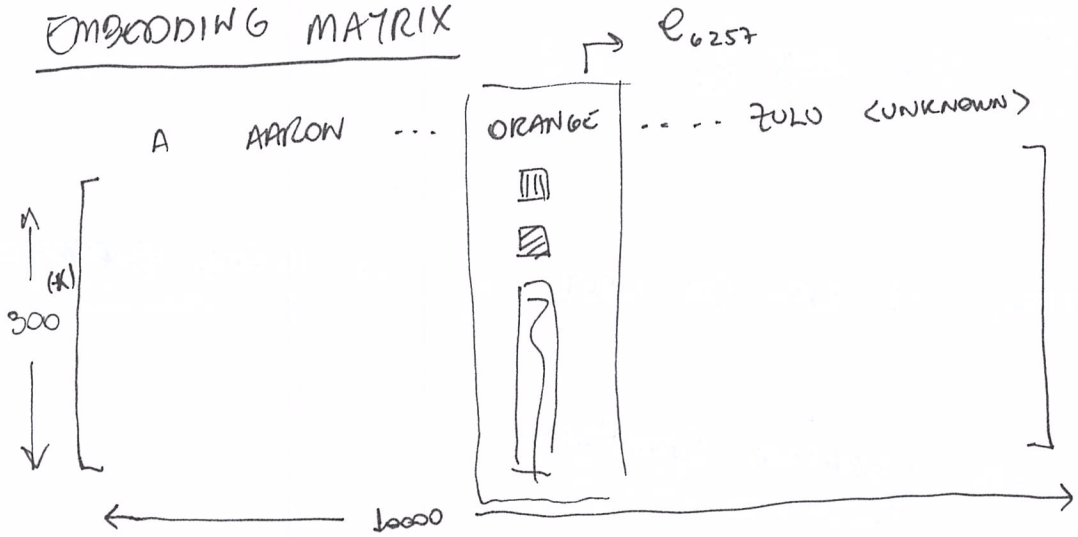
TO FIND WORD w : $\arg\max_w \text{sim}(e_w, e_{KING} - e_{MAN} + e_{WOMAN})$

FOR FUNCTION sim(·) : COSINE FUNCTION OF SIMILARITY

$$\text{sim}(u, v) = \frac{u^T v}{\|u\|_2 \|v\|_2}$$




EMBEDDING MATRIX




$$O_{6257} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \rightarrow 6257 \quad \Big\downarrow 10000 \qquad E_{(300, 10000)} \cdot O_{6257}{}_{(10000, 1)} = \begin{bmatrix} \\ \\ \\ \end{bmatrix} = e_{6257}$$

THAT MEANS: $E \cdot O_j = e_j$ := EMBEDDING FOR WORD $j$.

## WORD 2 VEC

"I WANT A GLASS OF ORANGE JUICE TO GO ALONG WITH MY CEREAL"

↻ RANDOMLY PICK CONTEXT

| CONTEXT | TARGET |
|---|---|
| ORANGE | JUICE |
| ORANGE | GLASS |
| ORANGE | MY |

TARGET → BY CHANCE CHOOSE +N OR -N WORD POSITION TO BE THE TARGET. ($N \in [1, 10]$)

$x \longrightarrow y$

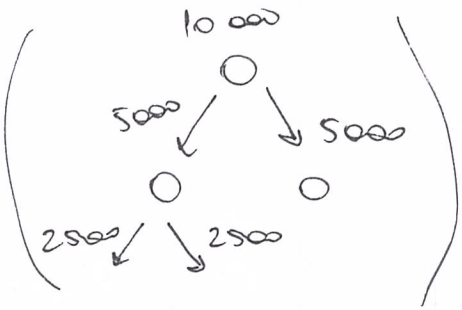
CONTEXT (c) → TARGET (t) ∴ $O_c \rightarrow E \rightarrow e_c \rightarrow \text{SOFTMAX} \rightarrow \hat{y}$

$E \cdot O_c = e_c$

SOFTMAX: $\mathbb{P}(t|c) = \frac{e^{\theta_t^T e_c}}{\sum_{j=1}^{10000} e^{\theta_j^T e_c}}$, $\theta_t$ = PARAMETER ASSOCIATED WITH OUTPUT $t$

$$\mathcal{L}(\hat{y}, y) = -\sum_{i=1}^{10000} y_i \cdot \log \hat{y}_i \quad , \quad y = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

TO TRAIN THE MODEL IT CAN BE USED HIERARCHICAL SOFTMAX




OR NEGATIVE SAMPLING.

# NEGATIVE SAMPLING

"I WANT A GLASS OF ORANGE JUICE TO GO ALONG WITH MY CEREAL"

| x | | y |
|---|---|---|
| CONTEXT | WORD | TARGET |
| ORANGE | JUICE | 1 |
| " | KING | 0 |
| " | BOOK | 0 |
| " | THE | 0 |
| " | OF | 0 |

K

THESE ARE ALL NEG. SAMPLES, THAT IS, NON TRUE EXAMPLES

$O_{6257} \rightarrow E \rightarrow e_{6257}$ → 10000 (0, 0, ⋮, 0)

BUT ONLY CLASSIFY K+1 AND THAT'S WHY IS FAST THAN HIERARCHICAL SOFTMA

## SELECTING NEG. EXAMPLES

$$P(w_i) = \frac{f(w_i)^{3/4}}{\sum_{j=1}^{10000} f(w_j)^{3/4}}$$, WHERE $f(w_i)$ IS THE FREQUENCY OF WORDS $i$.

## OBSERVATION

THERE ARE TWO MAIN WORD 2VEC MODELS:

- CONTINUOUS BAG OF WORDS (CBOW)
- SKIP-GRAM

IN THE CBOW MODEL, WE PREDICT A WORD GIVEN A CONTEXT (A CONTEXT CAN BE SOMETHING LIKE A SENTENCE)

IN THE SKIP-GRAM IS THE OPPOSITE: PREDICT THE CONTEXT GIVEN AN INPUT WORD.