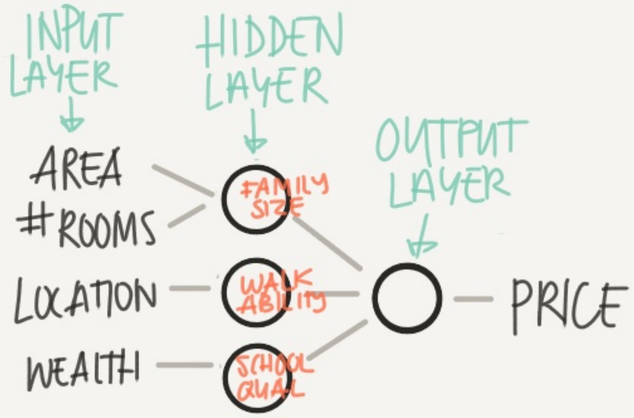
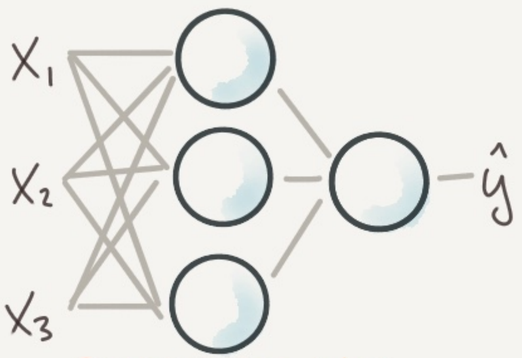
INTRO TO DEEP LEARNING

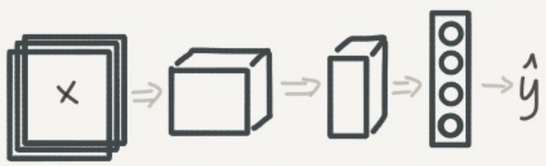


SUPERVISED LEARNING

INPUT: X	OUTPUT: Y	NN TYPE
HOME FEATURES	PRICE	STANDARD NN
AD + USER INFO	WILL CLICK ON AD (0/1)	
IMAGE	OBJECT (1...1000)	CONV. NN (CNN)
AUDIO	TEXT TRANSCRIPT	RECURRENT NN (RNN)
ENGLISH	CHINESE	
IMAGE/RADAR	POS OF OTHER CARS	CUSTOM/HYBRID

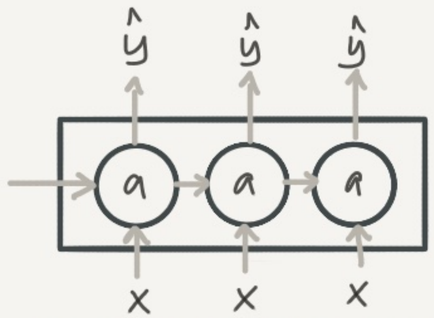


STANDARD NN



CONVOLUTIONAL NN

NETWORK ARCHITECTURES



RECURRENT NN

NNs CAN DEAL WITH BOTH STRUCTURED & UNSTRUCTURED DATA

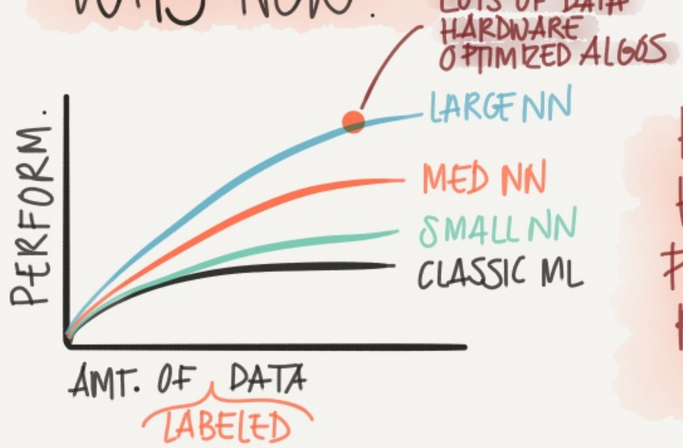
STRUCTURED

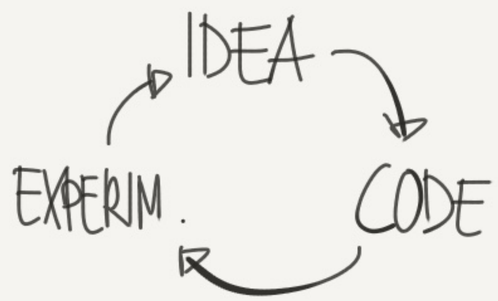
"THE QUICK BROWN FOX"

UNSTRUCTURED

HUMANS ARE GOOD AT THIS

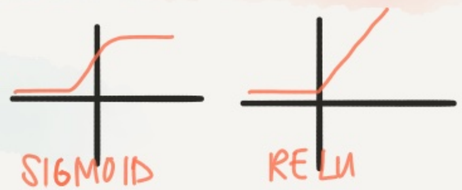
WHY NOW?





FASTER COMPUTATION IS IMPORTANT TO SPEED UP THE ITERATIVE PROCESS

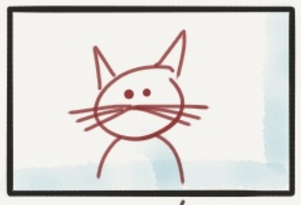
ONE OF THE BIG BREAKTHROUGHS HAS BEEN MOVING FROM SIGMOID TO RELU FOR FASTER GRADIENT DESCENT



LOGISTIC REGRESSION

AS A NEURAL NET

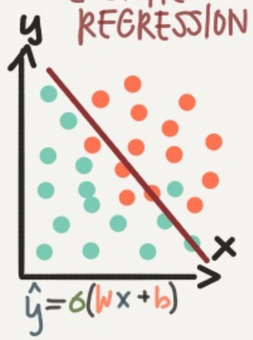
BINARY CLASSIFICATION



1 : CAT
0 : NOT CAT

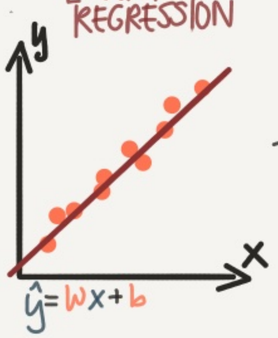
X → y

LOGISTIC REGRESSION



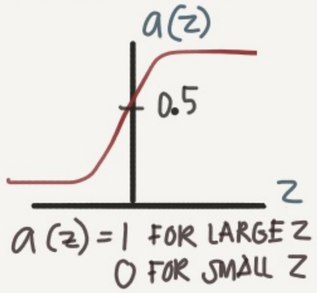
=

LINEAR REGRESSION

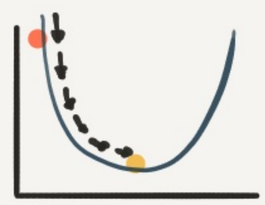


+

σ SIGMOID



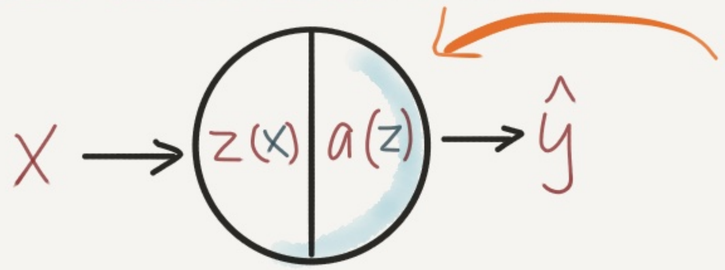
FINDING THE MINIMUM WITH GRADIENT DESCENT



1. FIND THE DOWNHILL DIRECTION (USING DERIVATIVES)
2. WALK (UPDATE W & B) AT A α LEARNING RATE

REPEAT UNTIL YOU REACH BOTTOM (CONVERGE)

PUTTING IT ALL TOGETHER



$$z(x) = wx + b$$

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

1. FORWARD PROPAGATION • CALCULATE $\hat{y}$
2. BACKWARD PROPAGATION • GRADIENT DESCENT + UPDATE W & b

REPEAT UNTIL IT CONVERGES

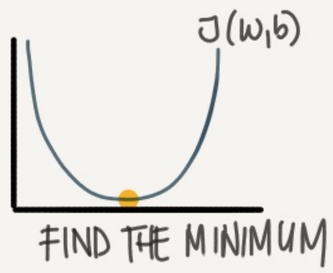
THE TASK IS TO LEARN w & b BUT HOW?

A : OPTIMIZE HOW GOOD THE GUESS IS BY MINIMIZING THE DIFF BETWEEN GUESS ($\hat{y}$) AND TRUTH (y)

$$\text{LOSS} = \mathcal{L}(\hat{y}, y)$$

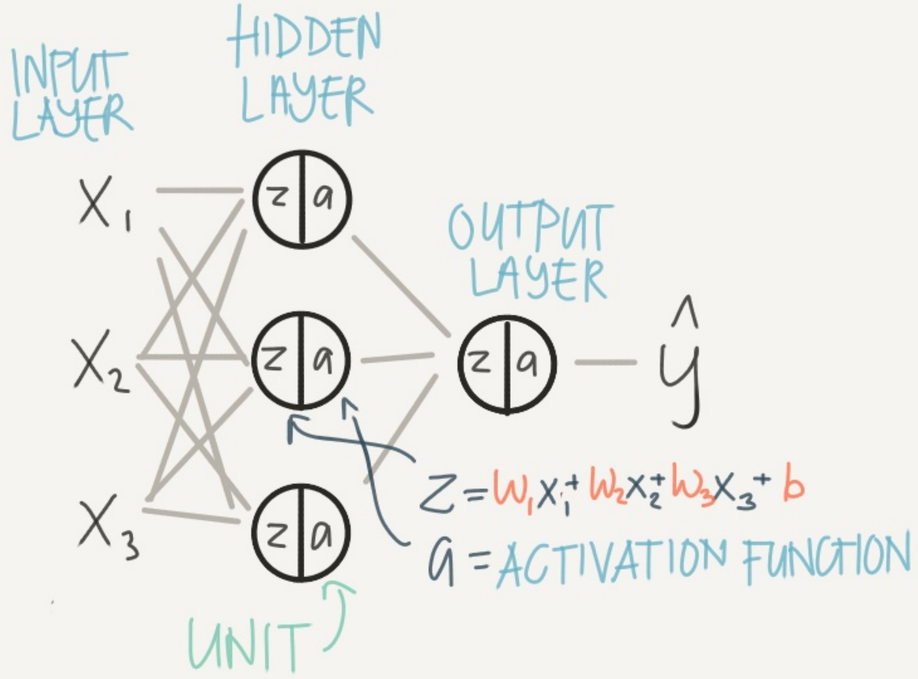
$$\text{COST} = J(w,b) = \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}(\hat{y}^{(i)}, y^{(i)})$$

COST = LOSS FOR THE ENTIRE DATASET



SHALLOW NEURAL NETS

2 LAYER NEURAL NET



WHY ACTIVATION FUNCTIONS?

EX. WITH NO ACTIVATION – $a = z$

$$a^{[1]} = z^{[1]} = w^{[1]}x + b^{[1]} \quad \text{LAYER 1}$$
$$a^{[2]} = z^{[2]} = w^{[2]}a^{[1]} + b^{[2]} \quad \text{LAYER 2}$$

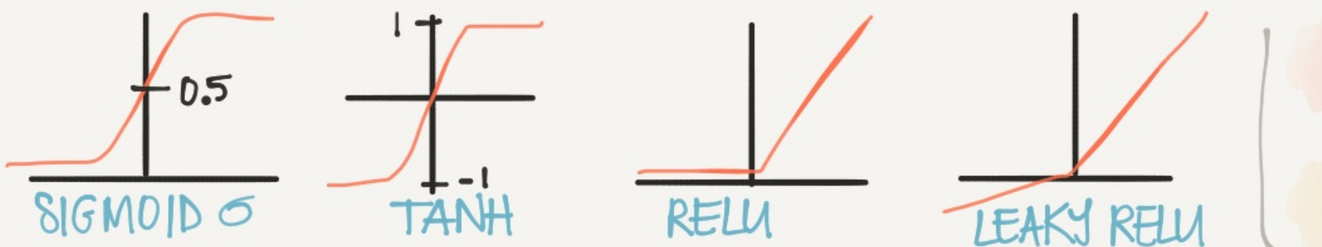
PLUG IN $a^{[1]}$

$$a^{[2]} = w^{[2]}(w^{[1]}x + b^{[1]}) + b^{[2]}$$
$$= w^{[2]}w^{[1]}x + w^{[2]}b^{[1]} + b^{[2]}$$
$$w'x + b'$$

← LINEAR FUNCTION

WE COULD JUST AS WELL HAVE SKIPPED THE WHOLE NEURAL NET & USED LIN. REGR.

ACTIVATION FUNCTIONS



SIGMOID σ
- BINARY CLASSIFIER - ONLY USED FOR OUTPUT LAYER
- SLOW GRAD DESCENT SINCE SLOPE IS SMALL FOR LARGE/SMALL VAL

TANH
- NORMALIZED ⇒ GRADIENT DESCENT IS FASTER

RELU
- DEFAULT CHOICE FOR ACTIVATION
- SLOPE = 1/0

LEAKY RELU
- AVOIDS UNDEF SLOPE AT 0 BUT RARELY USED IN PRACTICE

INITIALIZING w+b

WHAT IF: INIT TO 0

THIS WILL CAUSE ALL THE UNITS TO BE THE SAME AND LEARN EXACTLY THE SAME FEATURES

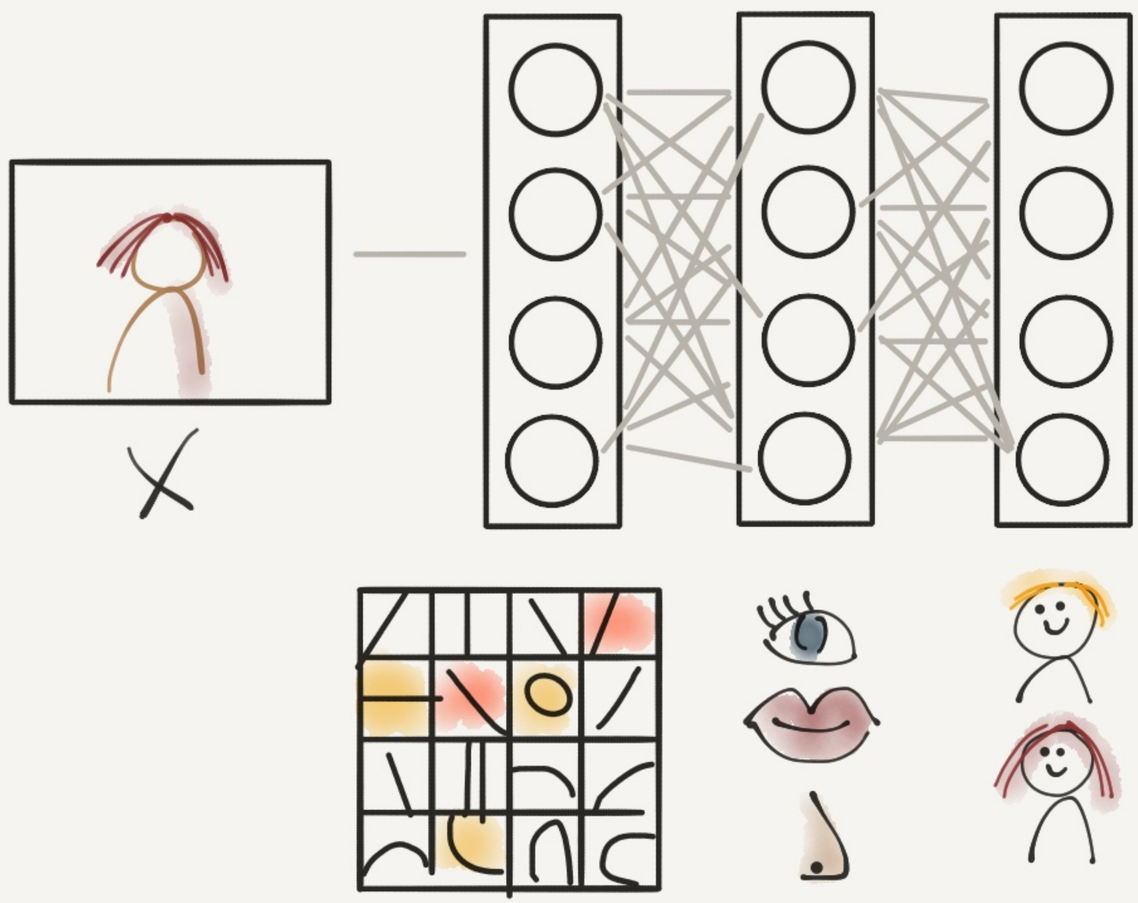
SOLUTION: RANDOM INIT

BUT ALSO WANT THEM SMALL SO RAND * 0.01 ← HYPERPARAM

DEEP NEURAL NETS

THERE ARE FUNCTIONS A SMALL DEEP NET CAN COMPUTE THAT SHALLOW NETS NEED EXP. MORE UNITS TO COMP.

WHY DEEP NEURAL NETS?



LOW LEVEL AUDIO WAVE FEATURES ↗↘ PITCH — PHONEMES — C A T — WORDS — SENTENCES

VERY DATA HUNGRY

NEED LOTS OF COMPUTER POWER

ALWAYS VECTORIZE

VECTOR MULT · CHEAPER THAN FOR LOOPS

COMPUTE ON GPUs

LOTS OF HYPERPARAMS

- LEARNING RATE α
- # ITERATIONS
- # HIDDEN LAYERS
- # HIDDEN UNITS
- CHOICE OF ACTIVATION
- MOMENTUM
- MINI-BATCH SIZE
- REGULARIZATION

RECURRENT
NEURAL NETWORKS

SEQUENCE MODELS · COURSERA

SEQUENCE PROBLEMS

IN	OUT	PURPOSE
	THE QUICK BROWN FOX JUMPED...	SPEECH RECOGNITION
∅		MUSIC GENERATION
THERE IS NOTHING TO LIKE IN THIS MOVIE		SENTIMENT CLASSIFICATION
AGCCCCTGTG AGGAACTAG	AGCCCCTGTG AGGAACTAG	DNA SEQUENCE ANALYSIS
Voulez-vous chanter avec moi?	Do you want to sing with me?	MACHINE TRANSLATION
	RUNNING	VIDEO ACTIVITY RECOGNITION
Yesterday Harry Potter met Hermoine Granger	Yesterday Harry Potter met Hermoine Granger	NAME ENTITY RECOGNITION

NAME ENTITY RECOGNITION

X = HARRY POTTER AND HERMOINE $T_x = 9$ (9 WORDS)
$x^{<1>}$ $x^{<2>}$ ...

GRANGER INVENTED A NEW SPELL

y = 1 1 0 1 $T_y = T_x$
$y^{<1>}$ $y^{<2>}$ ...
1 0 0 0 0

EXAMPLE OF A PROBLEM WHERE EVERY $x^{<i>}$ HAS AN OUTPUT $y^{<i>}$

HOW DO WE REPRESENT WORDS?

CREATE A VOCABULARY (EG 10K MOST COMMON WORDS IN YOUR TEXTS · OR DOWNLOAD EXISTING)

a	1
aaron	2
and	367
harry	4075
potter	6830
zulu	10000

EACH WORD IS A ~~ONE-HOT~~ VECTOR

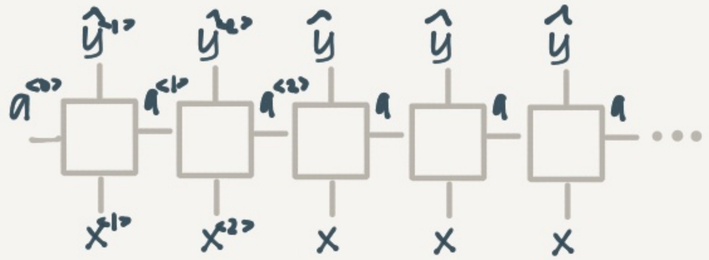
HARRY = [0 0 ... 1 (4075) ... 0]

WE COULD USE A STANDARD NETWORK BUT...

(A) INPUT & OUTPUTS CAN HAVE DIFFERENT LENGTHS IN DIFF EXAMPLES

(B) WE DON'T SHARE FEATURES LEARNED ACROSS DIFFERENT POSITIONS

RECURRENT NEURAL NET (RNN)



PREVIOUS RESULTS ARE PASSED IN AS INPUTS SO WE GET CONTEXT.

$a^{<1>} = g_1(W_a[a^{<0>}, x^{<1>}] + b_a)$ TANH/RELU

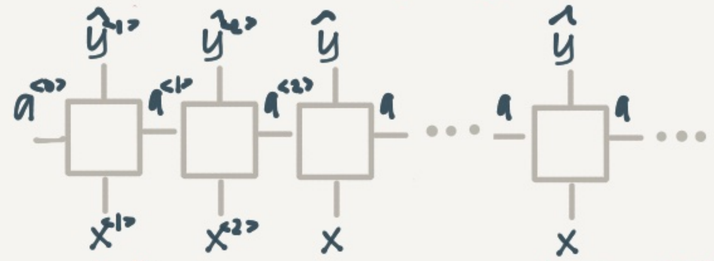
$\hat{y}^{<1>} = g_2(W_{ya}\, a^{<1>} + b_y)$ SIGMOID

THE SAME W & b ARE USED IN ALL TIME STEPS

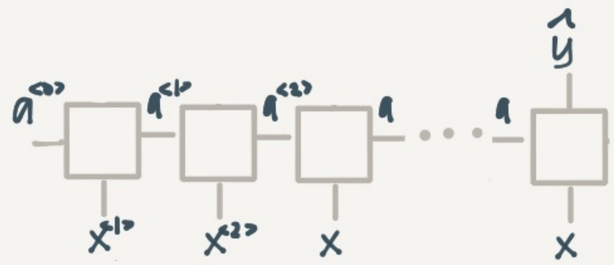
THE LOSS WE OPTIMIZE IS THE SUM OF $\mathcal{L}(\hat{y}, y)$ FROM 1–T

DIFFERENT TYPES OF RNN

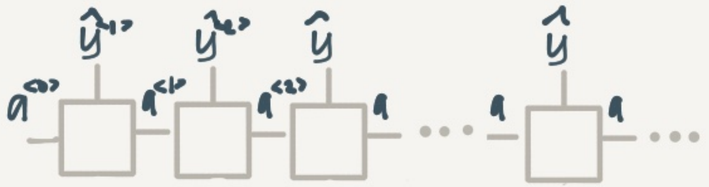
MANY-TO-MANY $T_x = T_y$



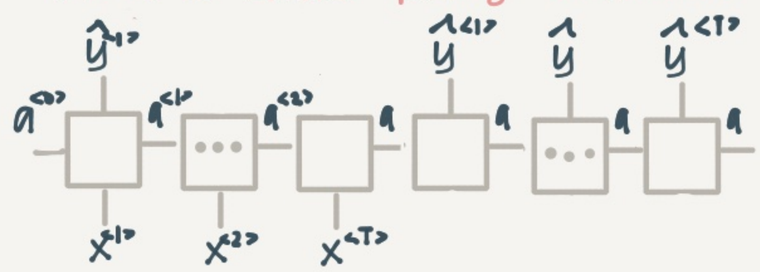
MANY-TO-ONE EX. SENTIMENT ANALYSIS



ONE-TO-MANY · MUSIC GENERATION



MANY-TO-MANY $T_x \neq T_y$ TRANSLATION



SEQUENCE MODELS · COURSERA

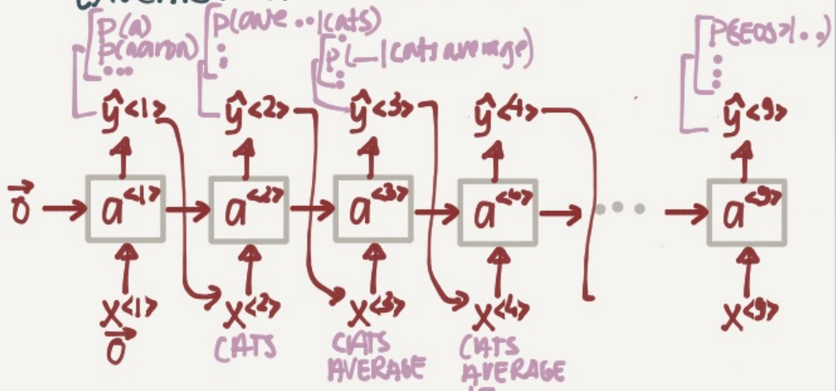
MORE ON RNNs

LANGUAGE MODELLING

HOW DO YOU KNOW IF SOMEONE SAID
THE APPLE AND PAIR SALAD OR
THE APPLE AND PEAR SALAD?

THE PURPOSE OF A LANG. MODEL IS TO CALCULATE THE PROBABILITIES

EX. CATS AVERAGE 15 HOURS OF SLEEP A DAY



SO GIVEN: CATS AVERAGE 15 WHAT IS THE PROB. THE NEXT WORD IS HOURS?

SAMPLING SENTENCES

1. TRAIN ON ALL HARRY POTTER BOOKS.
2. RANDOMLY SELECT A WORD (ONE OF THE TOP WORDS) (EX. THE)
3. PASS THIS INTO THE NEXT TIMESTAMP AND SAMPLE A NEW WORD
4. REPEAT UNTIL X WORDS OR YOU REACHED <EOS>

CAN DO AT CHARACTERLEVEL AS WELL

YAY! YOU ARE NOW YOUR OWN J.K. ROWLING

VANISHING GRADIENTS

THE CAT, WHO ALREADY ATE APPLES AND ORANGES AND A FEW MORE THINGS BLA BLA WAS FULL
THE CATS, WHO ALREADY ATE ... WERE FULL

NEED TO REMEMBER SING/PLURAL FOR A LONG TIME

SINCE LONG SENTENCE ⇒ DEEP RNN
WE GET THE VANISHING GRADIENTS PROB WE HAVE IN STANDARD NNs – I.E THE GRADIENTS FOR CAT/CATS HAVE LITTLE OR NO EFFECT ON WAS/WERE.

NOTE SOMETIMES YOU SEE EXPLODING GRAD (AS OVERFLOW NaN) BUT THIS IS EASILY FIXED WITH GRADIENT CLIPPING

GATED RECURRENT UNIT GRU

HELPS RECALL IF CAT WAS SING. OR PLURAL

GRU

THE GRU ACTS AS A MEMORY — AT EVERY TIMESTEP IT CALCULATES A NEW $\tilde{c}$ TO STORE AND A GATE Γ_u DECIDES TO UPDATE c TO $\tilde{c}$ OR NOT

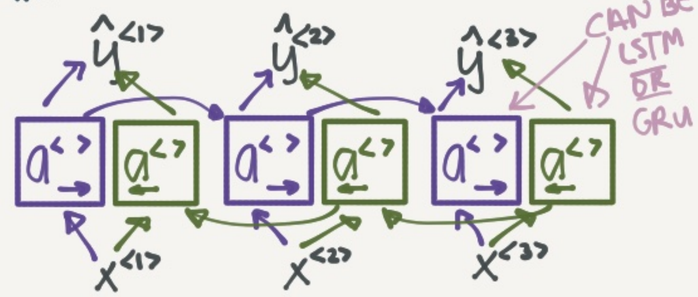
LONG SHORT TERM MEMORY (LSTM)

THE LSTM IS A VARIATION ON THE SAME THEME AS GRU BUT WITH AN ADDITIONAL Γ_F FORGET GATE

BI-DIRECTIONAL RNNs (BRNN)

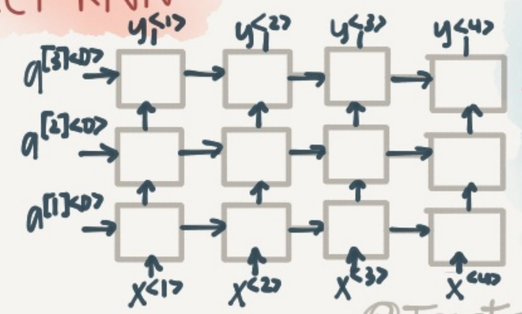
HE SAID, 'TEDDY BEARS ARE ON SALE'
HE SAID, 'TEDDY ROOSEVELT WAS A GREAT PRESIDENT'

PROBLEM: WITHOUT LOOKING FORWARD WE CANT SAY IF TEDDY IS A TOY OR A NAME



ONE DISADVANTAGE IS THAT YOU NEED THE FULL SENTENCE BEFORE YOU BEGIN - SO NOT SUITABLE FOR LIVE SPEECH RECO

DEEP RNN



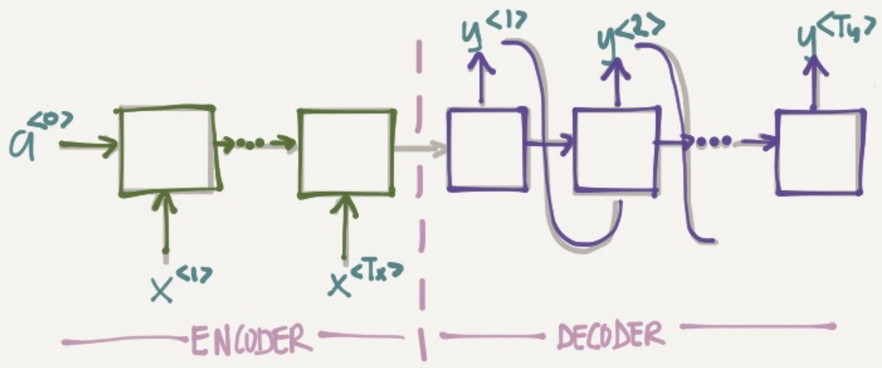
SINCE THEY ARE ALREADY TEMORALLY DEEP THEY USUALLY DON'T HAVE A LOT OF LAYERS

SEQUENCE TO SEQUENCE

SEQUENCE MODELS · COURSERA

BASIC MODELS

JANE VISITE L'AFRICE EN SEPTEMBRE → JANE IS VISITING AFRICA IN SEPTEMBER



 → THIS IS A CAT ON A CHAIR

CNN → RNN

HOW DO YOU PICK THE MOST LIKELY SENTENCE?

$$P(y^{<1>}, \dots, y^{<T_y>} | x)$$

(ENGLISH, FRENCH)

WE DON'T WANT A RANDOMLY GENERATED SENTENCE
(WE WOULD SOMETIMES GET A GOOD, SOMETIMES BAD)
INSTEAD WE WANT TO MAXIMIZE

$$\text{ARG MAX } P(y^{<1>}, \dots, y^{<T_y>} | x)$$

IDEA: USE GREEDY SEARCH

1. PICK THE WORD WITH THE BEST PROBABILITY
2. REPEAT UNTIL <EOS>

WITH THIS WE COULD GET
- JANE IS GOING TO BE VISITING AFRICA THIS SEPTEMBER

INSTEAD OF
- JANE IS VISITING AFRICA THIS SEPTEMBER

SOLUTION: OPTIMISE THE PROB OF THE WHOLE SENTENCE INSTEAD

BEAM SEARCH

1. PICK THE FIRST WORD
   PICK THE B (EX 3) BEST ALTERNATIVES
   (IN, JANE, SEPTEMBER)
2. FOR EACH B WORDS PICK THE NEXT WORD AND EVALUATE THE PAIRS TO END UP W B PAIRS
   $P(y^{<1>}, y^{<2>} | x) = P(y^{<1>} | x)\, P(y^{<2>} | x, y^{<1>})$

(IN SEPTEMBER, JANE IS, JANE VISITS)

NOTE: KEEP TRACK OF THE P FOR THE SENTENCES OF EACH LENGTH - AFTER X ITER (X MAX WORDS) PICK THE BEST

3. REPEAT TIL DONE

$$\text{ARG MAX} \prod_{t=1}^{T_y} P(y^{<t>} | x, y^{<1>}, \dots y^{<t-1>})$$

PROBLEM: MULTIPLYING PROBABILITIES (0<P<1) RESULTS IN A VERY SMALL NUMBER → OVERFLOWS

PROBLEM II: IF WE OPTIMIZE FOR THE MULT WE WILL PREFER SHORT SENTENCES. SINCE EACH WORD WILL REDUCE PROB

INSTEAD WE CAN OPTIMIZE FOR THIS

$$\frac{1}{T_y^{\alpha}} \sum_{t=1}^{T_y} \log(P(y^{<t>} | x, y^{<1>}, \dots, y^{<t-1>}))$$

HOW DO WE PICK B?

LARGE B: BETTER RESULT, SLOWER
SMALL B: WORSE RESULT, BETTER

IN PROD YOU MIGHT SE B=10.
100 IS PROBABLY A BIT TOO HIGH - BUT ITS DOMAIN DEPENDENT

ERROR ANALYSIS IN BEAM S.

HUMAN: JANE VISITS AFRICA IN SEPT... y^*
ALGO: JANE VISITED AFRICA LAST SEPTEMBER $\hat{y}$

HOW DO WE KNOW IF ITS OUR RNN OR OUR BEAM SEARCH WE SHOULD WORK ON?

LET THE RNN GIVE $Py^* = P(y^*|x)$ & $P\hat{y} = P(\hat{y}|x)$

IF $Py^* > P\hat{y}$:
BEAM PICKED THE WRONG ONE
TRY A HIGHER B

ELSE:
THE RNN PICKED THE WRONG PROBS · SO FOCUS ON THE RNN

SEQUENCE TO SEQUENCE

FRENCH: LE CHAT EST SUR LE TAPIS
HUMAN1: THE CAT IS ON THE MAT
HUMAN2: THERE IS A CAT ON THE MAT

HOW DO YOU EVALUATE THE MACHINE TRANSLATION WHEN MULTIPLE ARE RIGHT?

BLEU SCORE

IDEA: CHECK IF THE WORDS (MT) APPEAR IN THE REAL TRANSLATION

THE THE THE THE THE THE THE
SCORE: 7/7

IDEA: ONLY GIVE CREDIT FOR A WORD THE MAX # TIMES IT APPEARS IN A TARGET SENTENCE
SCORE: 2/7 ← COUNT CLIP

THE CAT THE CAT ON THE MAT

BIGRAMS	COUNT	COUNT CLIP
THE CAT	2	1
CAT THE	1	0
CAT ON	1	1
ON THE	1	1
THE MAT	1	1
	6	4/6

BI-GRAM SCORE: 4/6

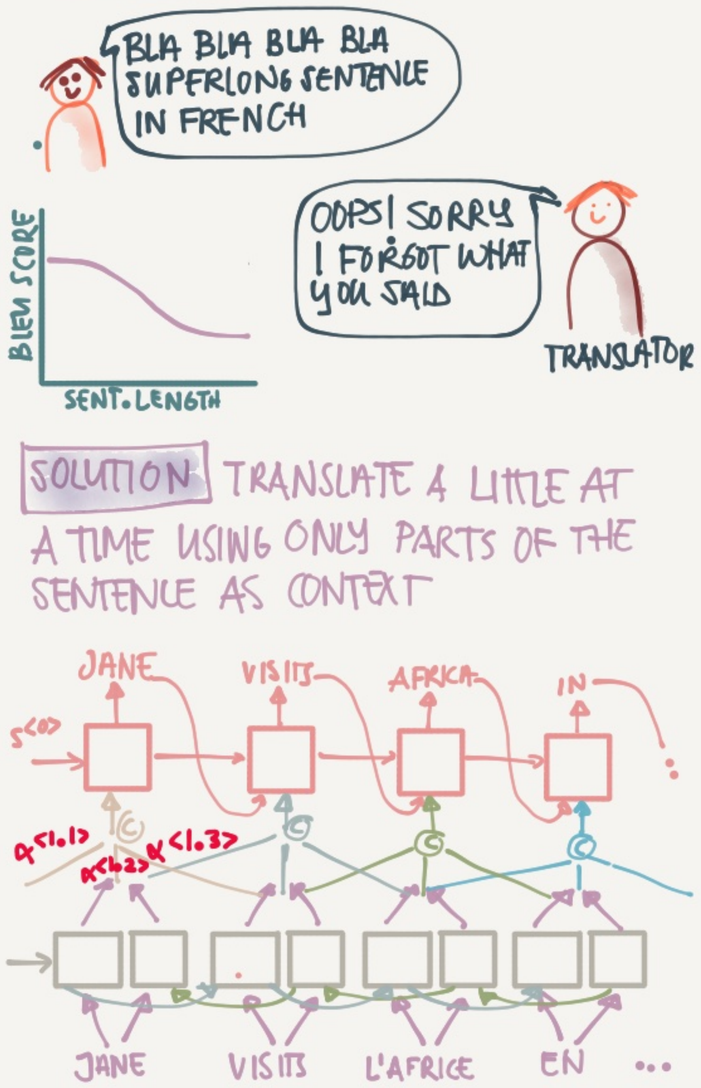
COMBINED BLEU SCORE

$BP \cdot \exp\left(\frac{1}{4}\sum_{n=1}^{4} P_n\right)$

P_1 = SCORE SINGLE WORD
P_2 = SORE BIGRAMS
...
BP = BREVITY PENALTY ← PENALIZES SENTENCES SHORTER THAN THE TARGET

↑ A USEFUL SINGLE NUMBER EVAL METRIC

ATTENTION MODEL



SOLUTION TRANSLATE A LITTLE AT A TIME USING ONLY PARTS OF THE SENTENCE AS CONTEXT

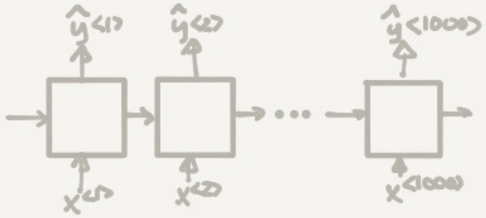
$\alpha^{<t,t'>}$ = HOW MUCH ATTENTION $y^{<t>}$ SHOULD PAY TO $a^{<t'>}$

$$c^{<2>} = \sum_{t'} \alpha^{<2,t'>} \cdot a^{<t'>} \quad \Big| \quad \sum_{t'} \alpha^{<1,t'>} = 1$$

α IS CALCULATED USING A SMALL NEURAL NETWORK

$$\alpha^{<t,t'>} = \frac{\exp(e^{<t,t'>})}{\sum_{t'=1}^{T_x} \exp(e^{<t,t'>})}$$

SPEECH RECOGNITION



PROBLEM: 10s CLIP AT 100Hz = 1000 INPUTS BUT ONLY ≈ 20 OUTPUTS

SOLUTION USE CTC COST (CONNECTION TEMPORAL CLASSIFICATION)

ttt_h_eee___u___ggg__ → COLLAPSE REPEATED CHARS NOT SEP BY BLANK

TRIGGER WORD DETECTION

SEQUENCE MODELS · COURSERA