



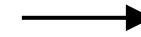
deeplearning.ai

Recurrent Neural Networks

Why sequence models?

Examples of sequence data

Speech recognition



“The quick brown fox jumped
over the lazy dog.”

Music generation

∅



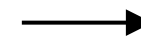
Sentiment classification

“There is nothing to like
in this movie.”



DNA sequence analysis

AGCCCCTGTGAGGAACTAG



AG**CCCCTGTGAGGAACT**AG

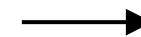
Machine translation

Voulez-vous chanter avec
moi?



Do you want to sing with
me?

Video activity recognition



Running

Name entity recognition

Yesterday, Harry Potter
met Hermione Granger.



Yesterday, **Harry Potter**
met **Hermione Granger**.



deeplearning.ai

Recurrent Neural Networks

Notation

Motivating example

x: Harry Potter and Hermione Granger invented a new spell.

Representing words

x: Harry Potter and Hermione Granger invented a new spell.

$$x^{<1>} \quad x^{<2>} \quad x^{<3>} \quad \dots \quad x^{<9>}$$

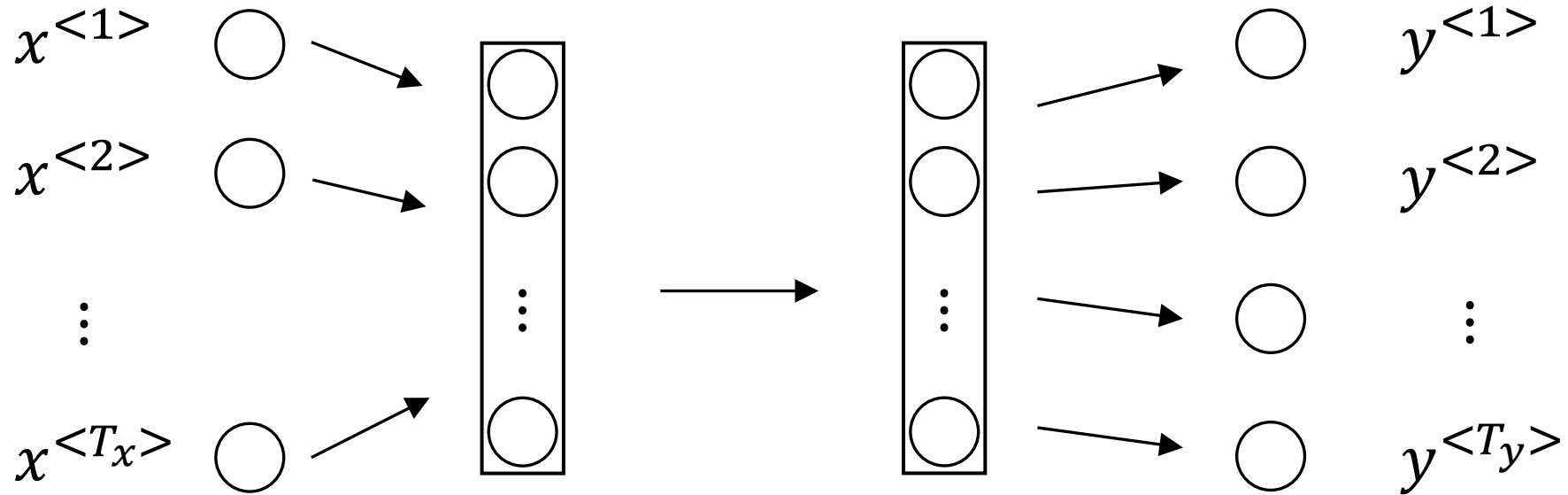


deeplearning.ai

Recurrent Neural Networks

Recurrent Neural Network Model

Why not a standard network?



Problems:

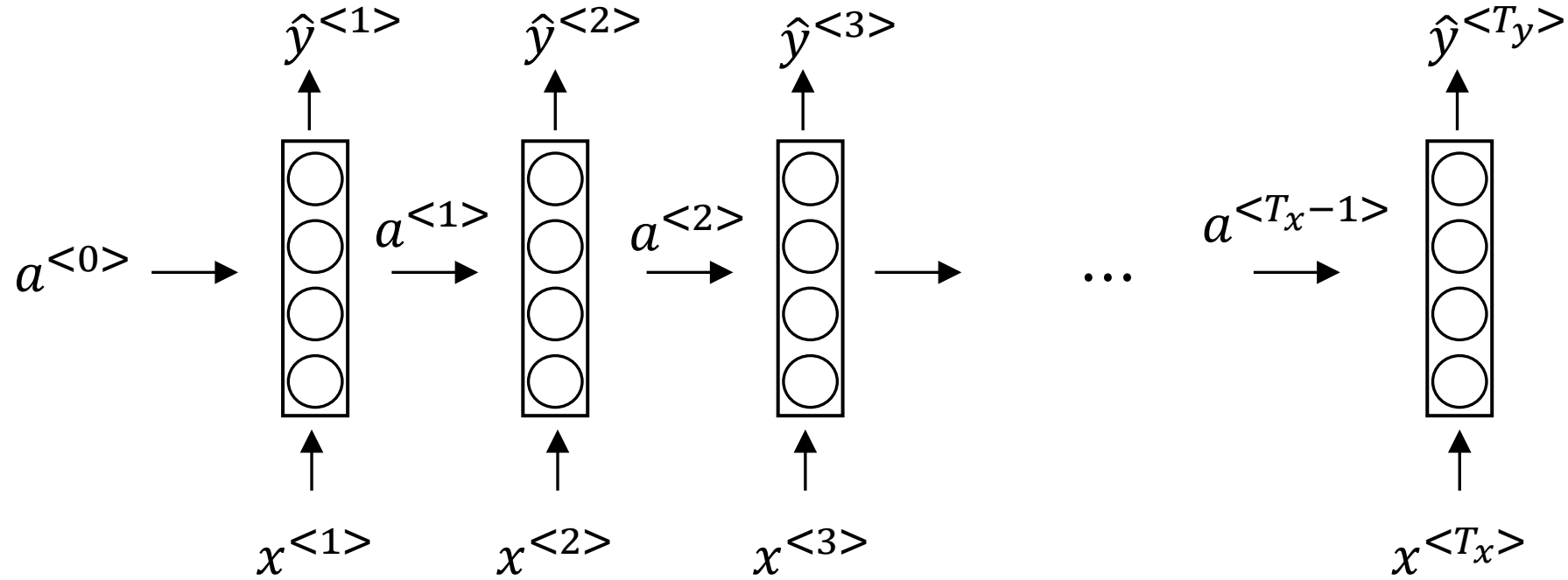
- Inputs, outputs can be different lengths in different examples.
- Doesn't share features learned across different positions of text.

Recurrent Neural Networks

He said, “Teddy Roosevelt was a great President.”

He said, “Teddy bears are on sale!”

Forward Propagation



Simplified RNN notation

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

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

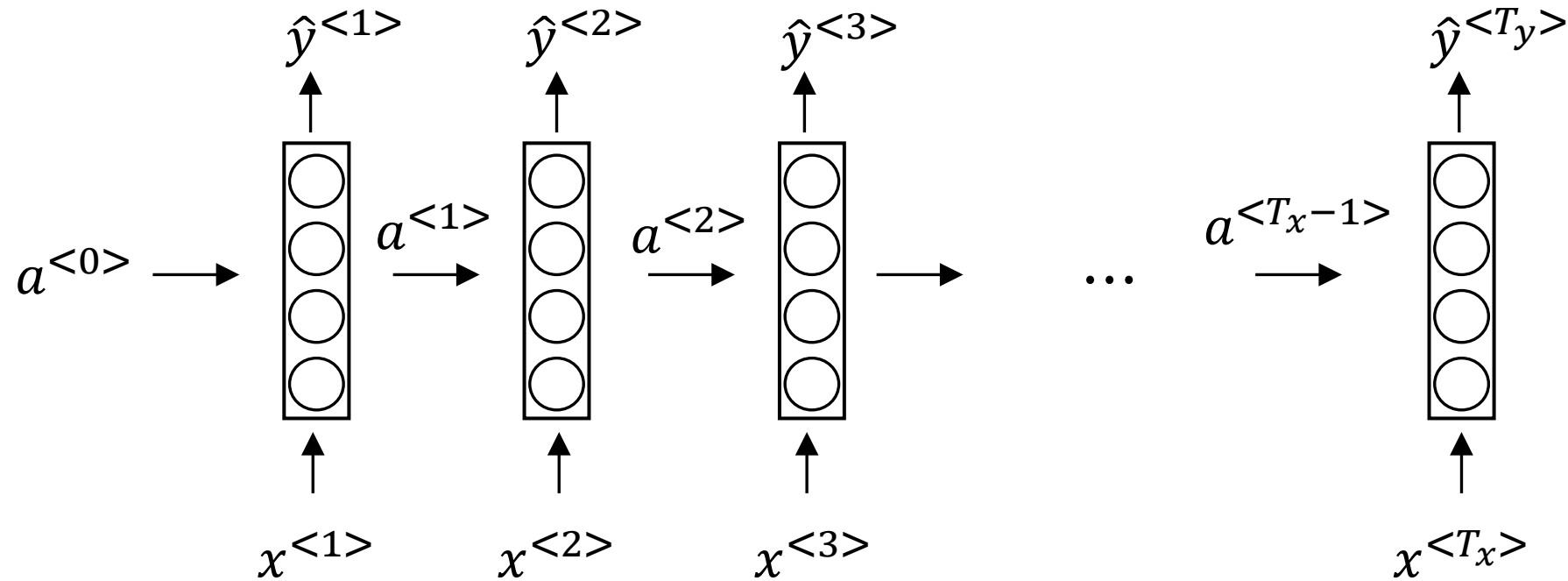


deeplearning.ai

Recurrent Neural Networks

Backpropagation through time

Forward propagation and backpropagation



Forward propagation and backpropagation

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

Backpropagation through time
13/117



deeplearning.ai

Recurrent Neural Networks

Different types of RNNs

Examples of sequence data

Speech recognition



“The quick brown fox jumped over the lazy dog.”

Music generation

∅



Sentiment classification

“There is nothing to like in this movie.”



DNA sequence analysis

AGCCCCTGTGAGGAACTAG



AG**CCCCTGTGAGGAACT**AG

Machine translation

Voulez-vous chanter avec moi?



Do you want to sing with me?

Video activity recognition



Running

Name entity recognition

Yesterday, Harry Potter met Hermione Granger.

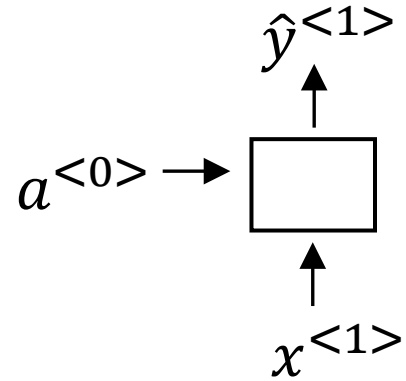


Yesterday, **Harry Potter** met **Hermione Granger**.

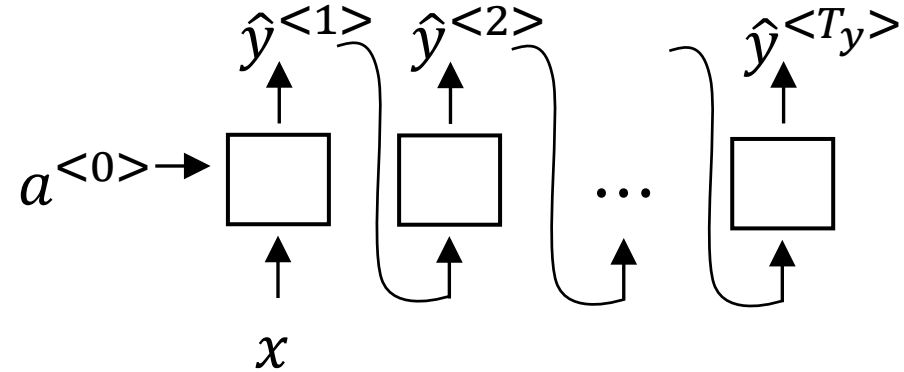
Examples of RNN architectures

Examples of RNN architectures

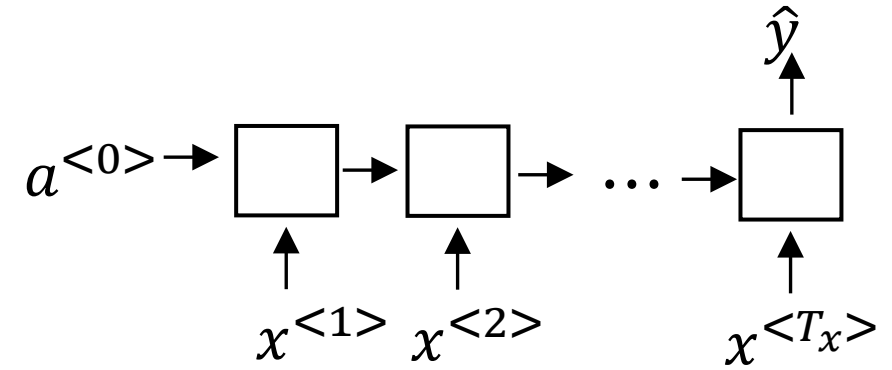
Summary of RNN types



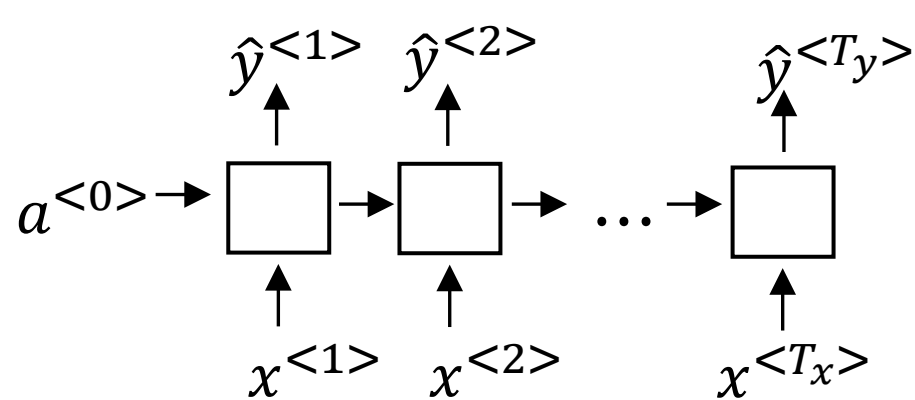
One to one



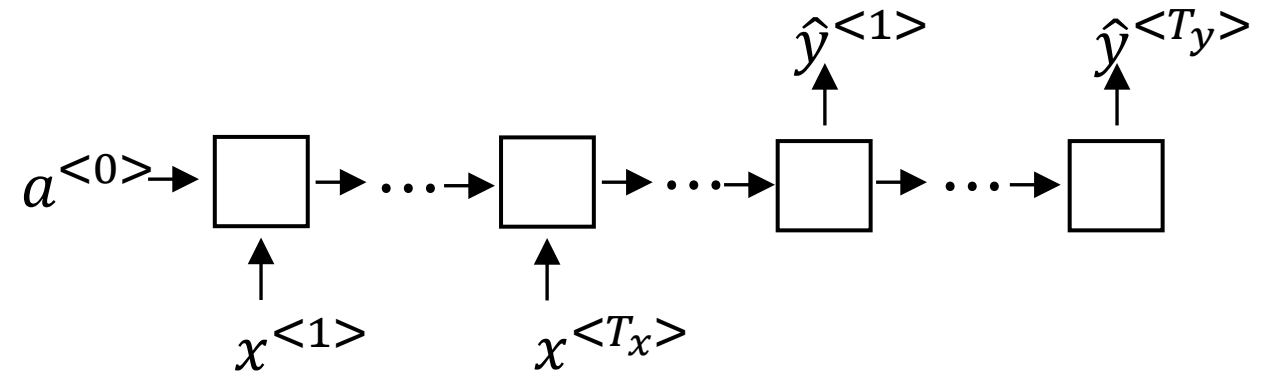
One to many



Many to one



Many to many



Many to many



deeplearning.ai

Recurrent Neural Networks

Language model and
sequence generation

What is language modelling?

Speech recognition

The apple and pair salad.

The apple and pear salad.

$P(\text{The apple and pair salad}) =$

$P(\text{The apple and pear salad}) =$

Language modelling with an RNN

Training set: large corpus of english text.

Cats average 15 hours of sleep a day.

The Egyptian Mau is a breed of cat. <EOS>

RNN model

Cats average 15 hours of sleep a day. <EOS>

$$\mathcal{L}(\hat{y}^{<t>}, y^{<t>}) = - \sum_i y_i^{<t>} \log \hat{y}_i^{<t>}$$

$$\mathcal{L} = \sum_t \mathcal{L}^{<t>}(\hat{y}^{<t>}, y^{<t>})$$

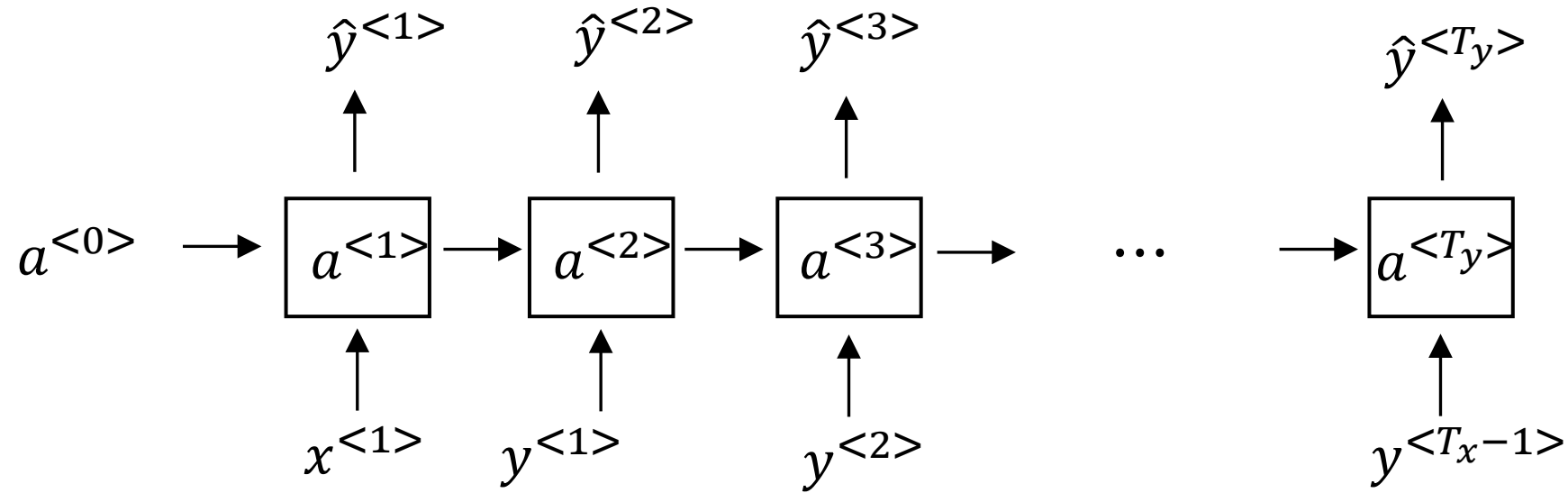


deeplearning.ai

Recurrent Neural Networks

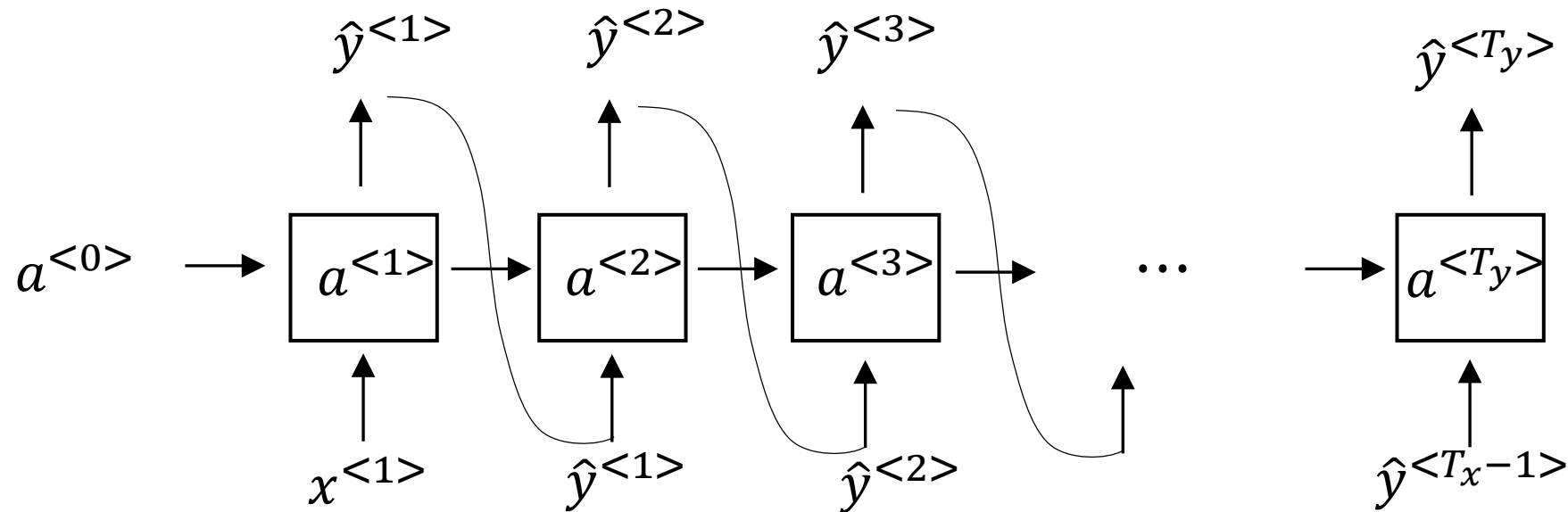
Sampling novel
sequences

Sampling a sequence from a trained RNN



Character-level language model

Vocabulary = [a, aaron, ..., zulu, <UNK>]



Sequence generation

News

President enrique peña nieto, announced
sench's sulk former coming football langston
paring.

"I was not at all surprised," said hich langston.

"Concussion epidemic", to be examined.

The gray football the told some and this has on
the uefa icon, should money as.

Shakespeare

The mortal moon hath her eclipse in love.

And subject of this thou art another this fold.

When besser be my love to me see sabl's.

For whose are ruse of mine eyes heaves.

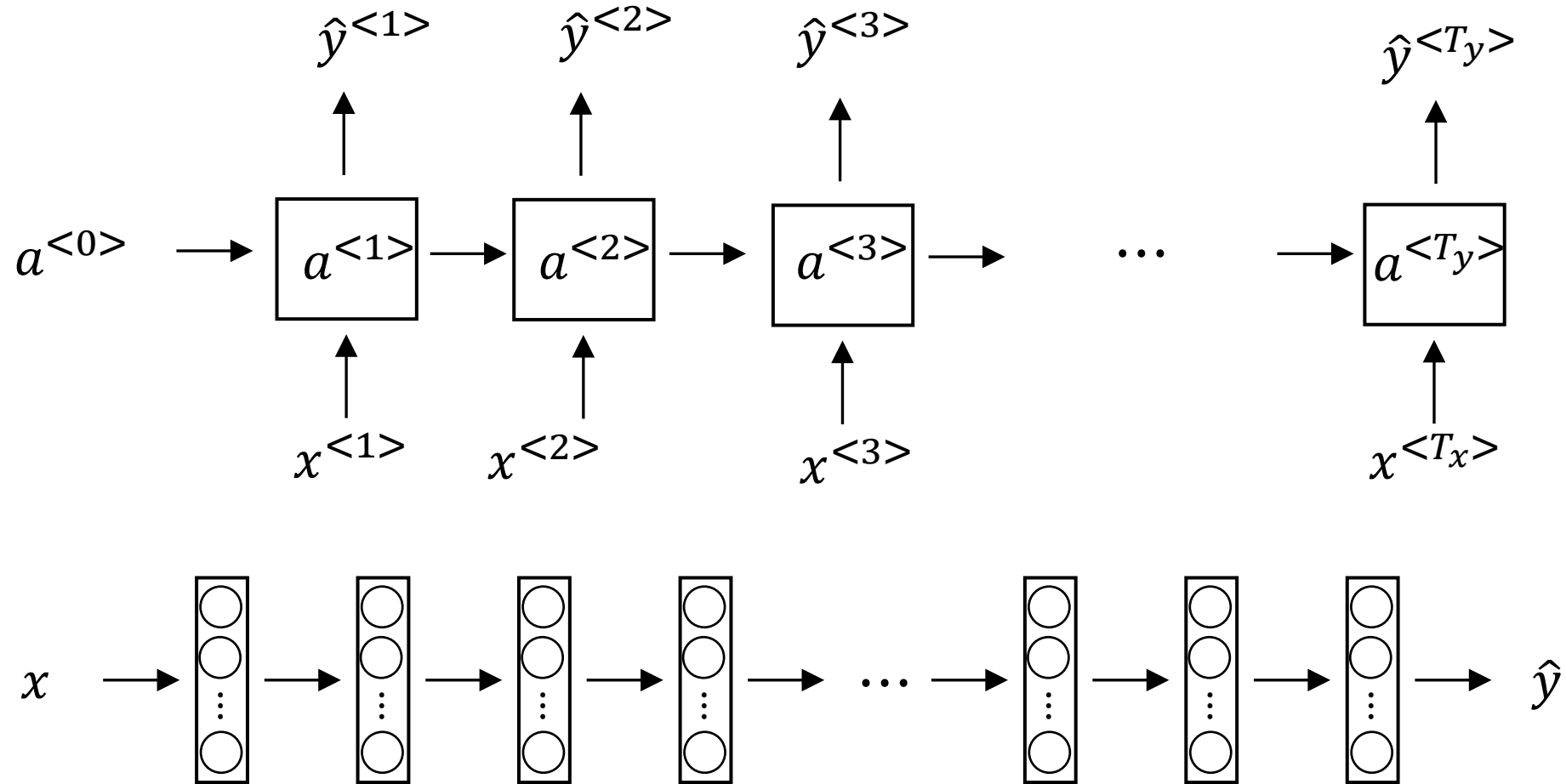


deeplearning.ai

Recurrent Neural Networks

Vanishing gradients with RNNs

Vanishing gradients with RNNs



Exploding gradients.



deeplearning.ai

Recurrent Neural Networks

Gated Recurrent Unit (GRU)

RNN unit

$$a^{<t>} = g(W_a[a^{<t-1>}, x^{<t>}] + b_a)$$

GRU (simplified)

The cat, which already ate ..., was full.

[Cho et al., 2014. On the properties of neural machine translation: Encoder-decoder approaches]

[Chung et al., 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling]

Full GRU

$$\tilde{c}^{<t>} = \tanh(W_c[c^{<t-1>}, x^{<t>}] + b_c)$$

$$\Gamma_u = \sigma(W_u[c^{<t-1>}, x^{<t>}] + b_u)$$

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

The cat, which ate already, was full.



deeplearning.ai

Recurrent Neural Networks

LSTM (long short
term memory) unit

GRU and LSTM

GRU

LSTM

$$\tilde{c}^{<t>} = \tanh(W_c[\Gamma_r * c^{<t-1>}, x^{<t>}] + b_c)$$

$$\Gamma_u = \sigma(W_u[c^{<t-1>}, x^{<t>}] + b_u)$$

$$\Gamma_r = \sigma(W_r[c^{<t-1>}, x^{<t>}] + b_r)$$

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

$$a^{<t>} = c^{<t>}$$

LSTM in pictures

$$\tilde{c}^{<t>} = \tanh(W_c[a^{<t-1>}, x^{<t>}] + b_c)$$

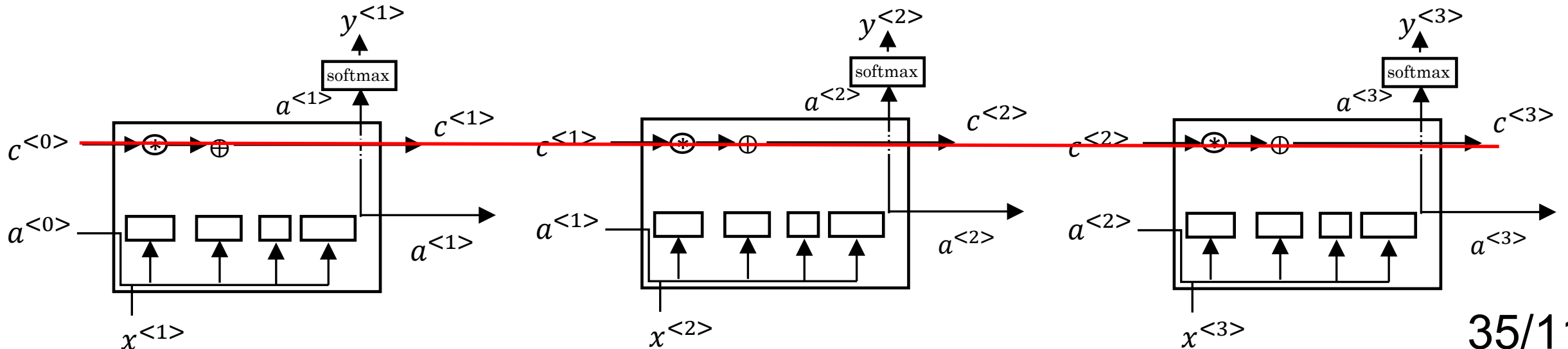
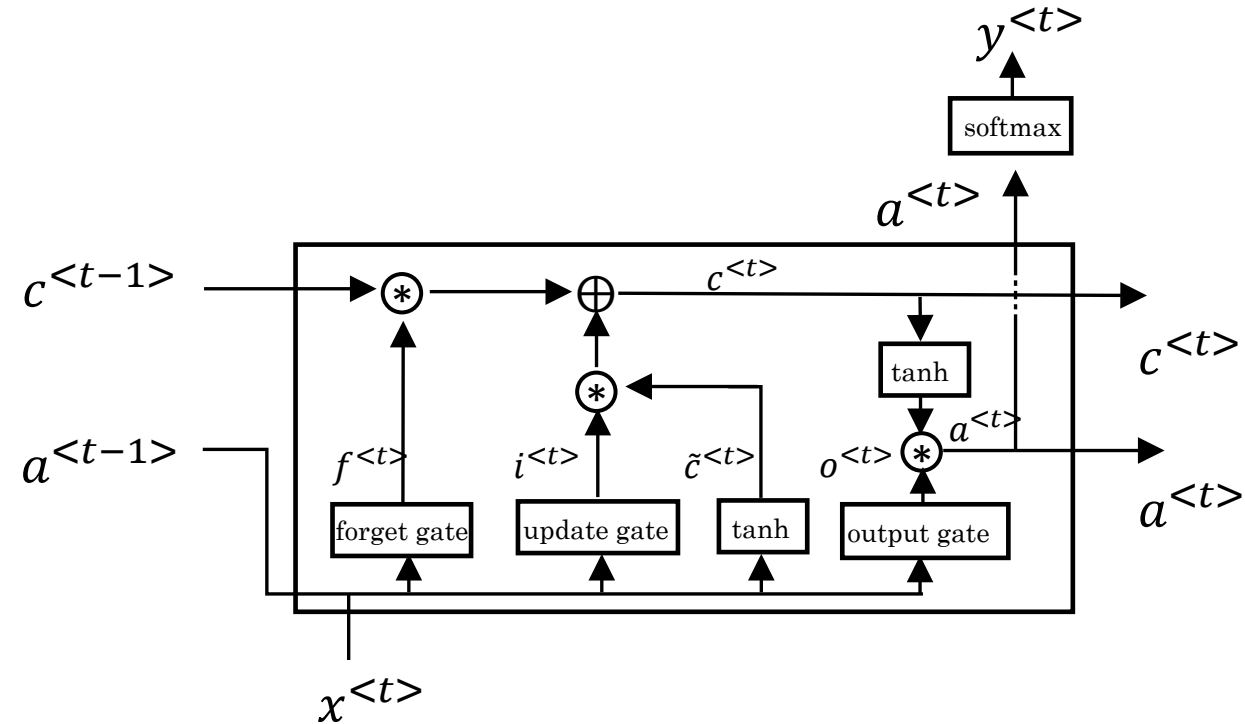
$$\Gamma_u = \sigma(W_u[a^{<t-1>}, x^{<t>}] + b_u)$$

$$\Gamma_f = \sigma(W_f[a^{<t-1>}, x^{<t>}] + b_f)$$

$$\Gamma_o = \sigma(W_o[a^{<t-1>}, x^{<t>}] + b_o)$$

$$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + \Gamma_f * c^{<t-1>}$$

$$a^{<t>} = \Gamma_o * c^{<t>}$$





deeplearning.ai

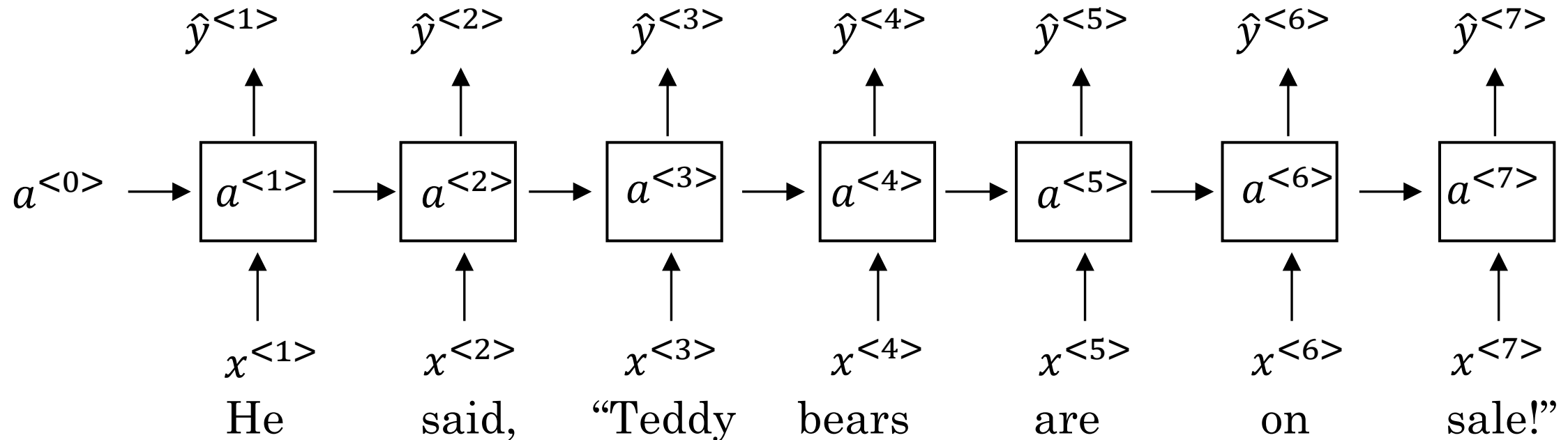
Recurrent Neural Networks

Bidirectional RNN

Getting information from the future

He said, “Teddy bears are on sale!”

He said, “Teddy Roosevelt was a great President!”



Bidirectional RNN (BRNN)

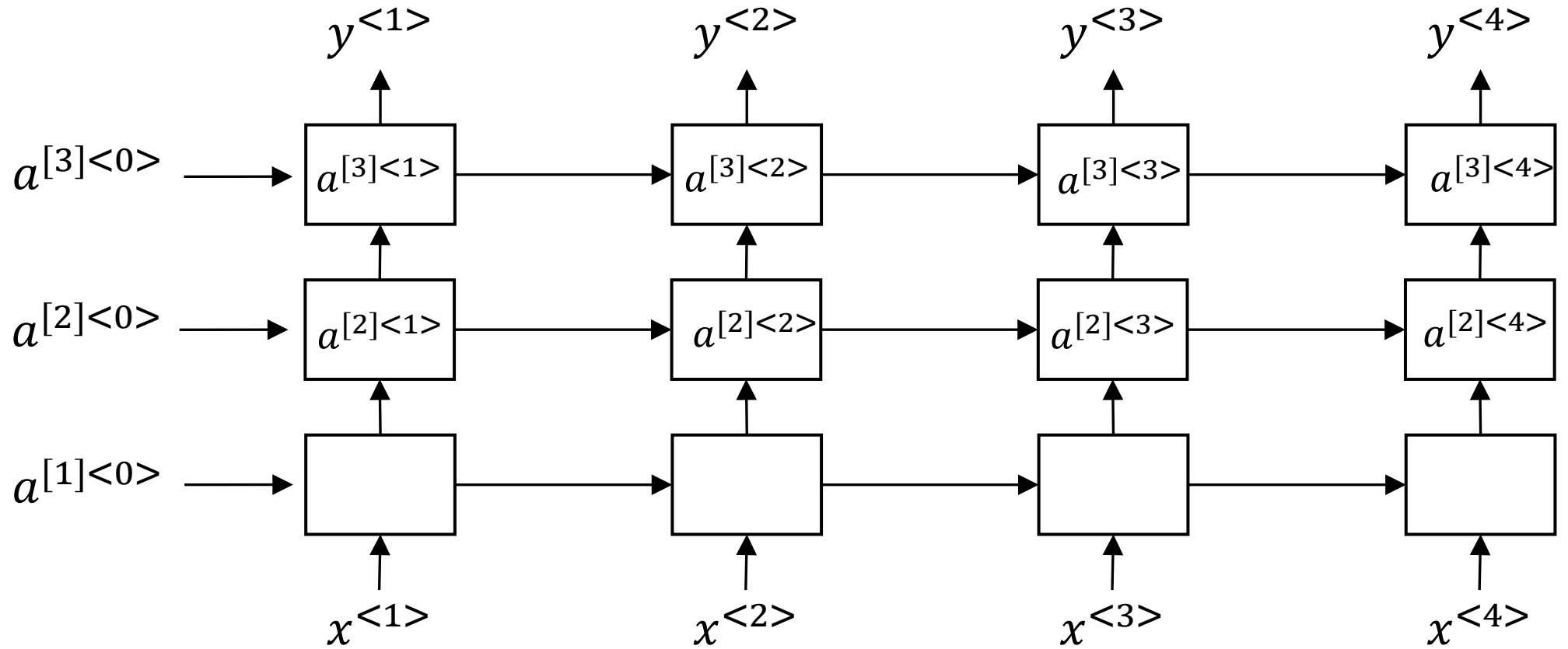


deeplearning.ai

Recurrent Neural Networks

Deep RNNs

Deep RNN example





deeplearning.ai

NLP and Word Embeddings

Word representation

Word representation

$V = [a, aaron, \dots, zulu, <UNK>]$

1-hot representation

Man (5391)	Woman (9853)	King (4914)	Queen (7157)	Apple (456)	Orange (6257)
$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$

I want a glass of orange _____.

I want a glass of apple_____.

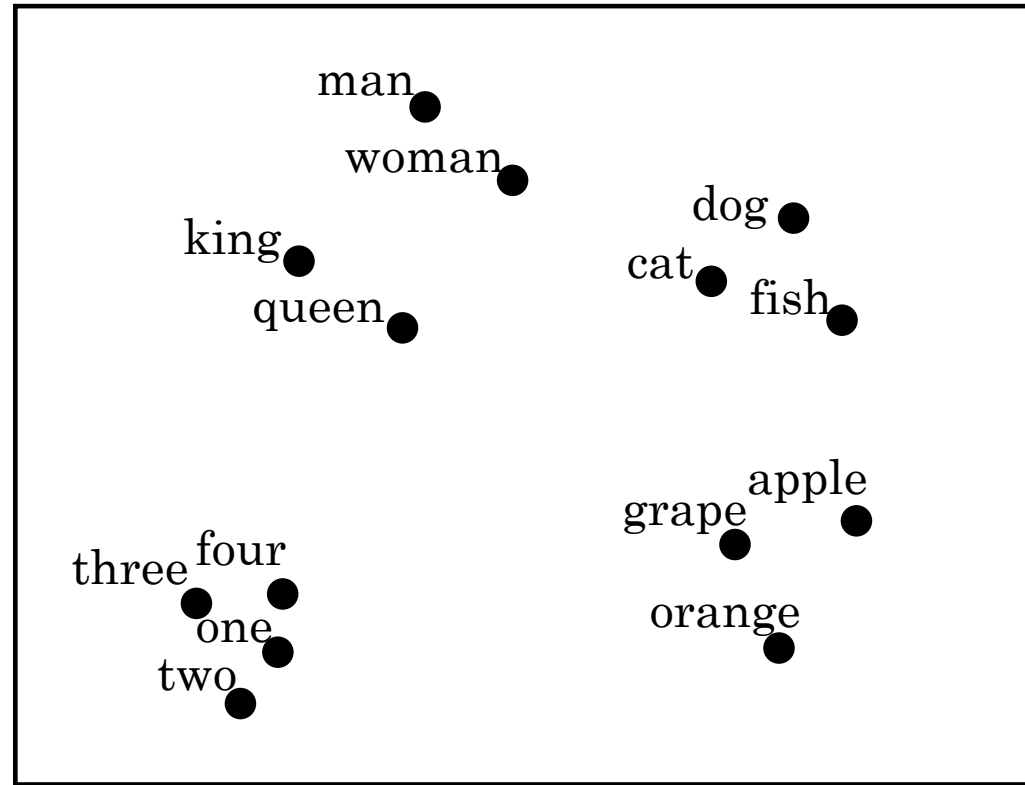
Featurized representation: word embedding

	Man (5391)	Woman (9853)	King (4914)	Queen (7157)	Apple (456)	Orange (6257)
			-0.95	0.97	0.00	0.01
			0.93	0.95	-0.01	0.00
			0.7	0.69	0.03	-0.02
			0.02	0.01	0.95	0.97

I want a glass of orange _____.

I want a glass of apple_____.

Visualizing word embeddings



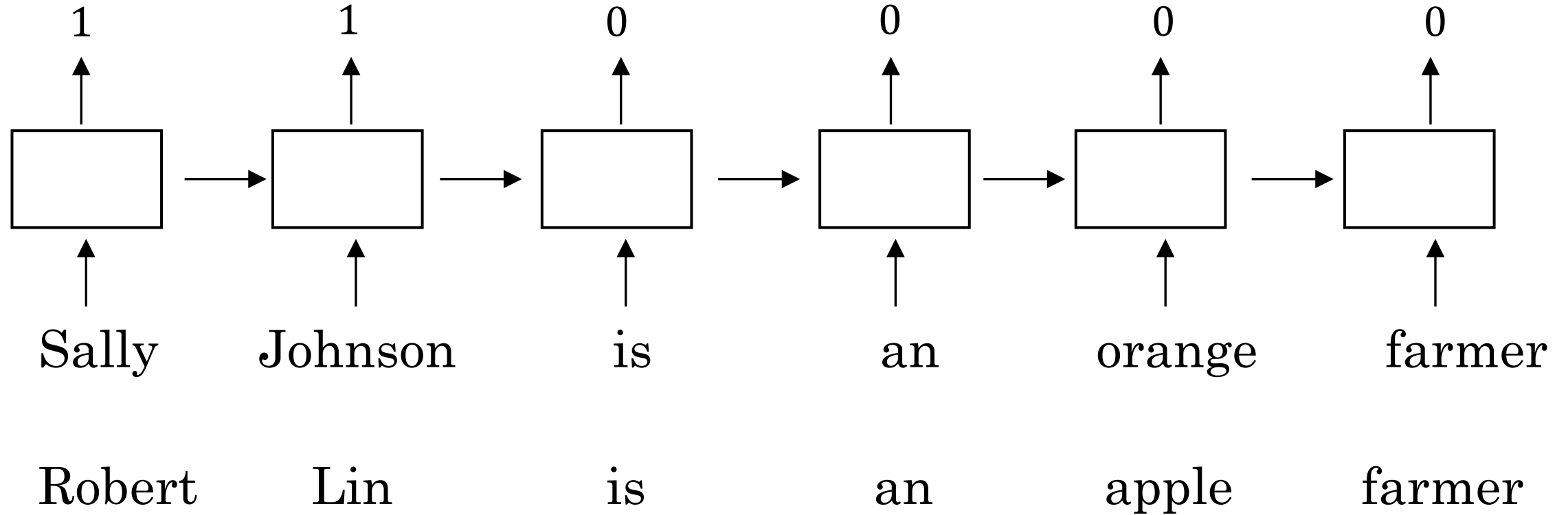


deeplearning.ai

NLP and Word Embeddings

Using word embeddings

Named entity recognition example



Transfer learning and word embeddings

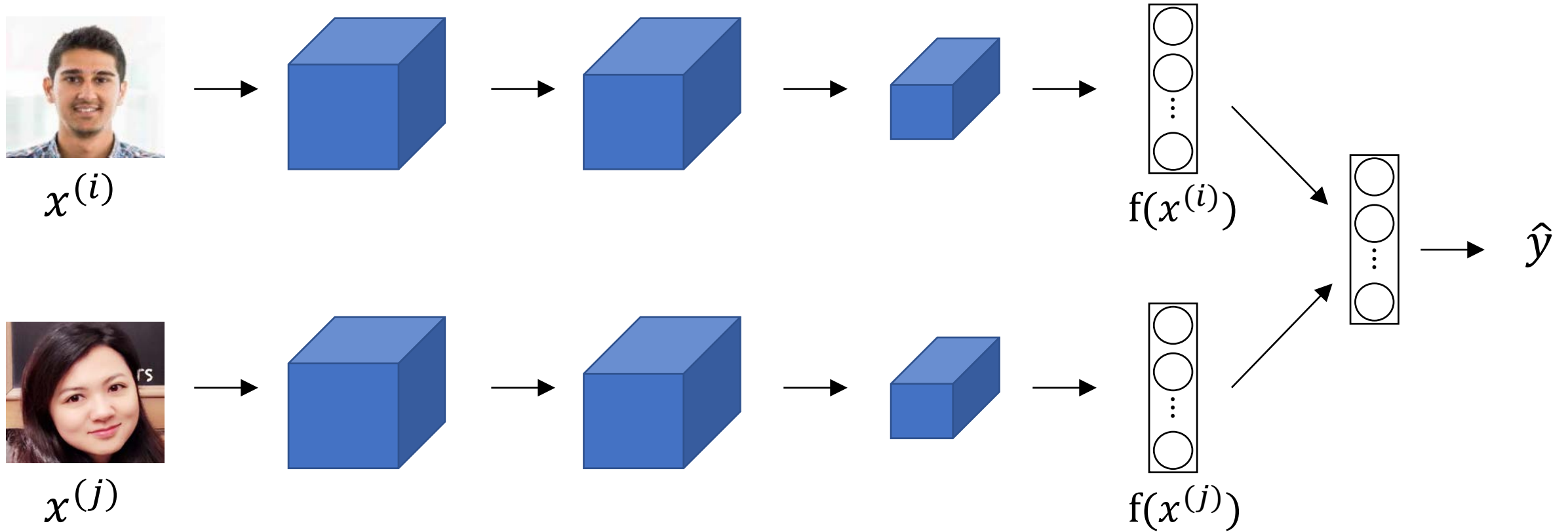
1. Learn word embeddings from large text corpus. (1-100B words)

(Or download pre-trained embedding online.)

2. Transfer embedding to new task with smaller training set.
(say, 100k words)

3. Optional: Continue to finetune the word embeddings with new data.

Relation to face encoding





deeplearning.ai

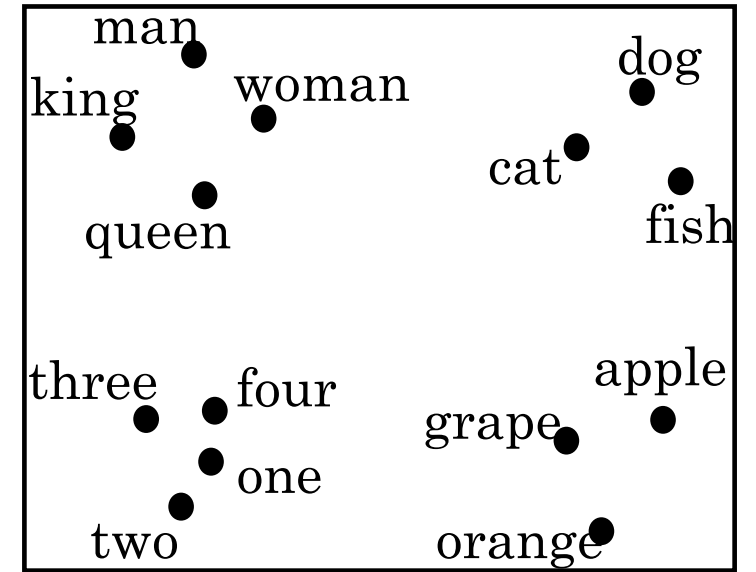
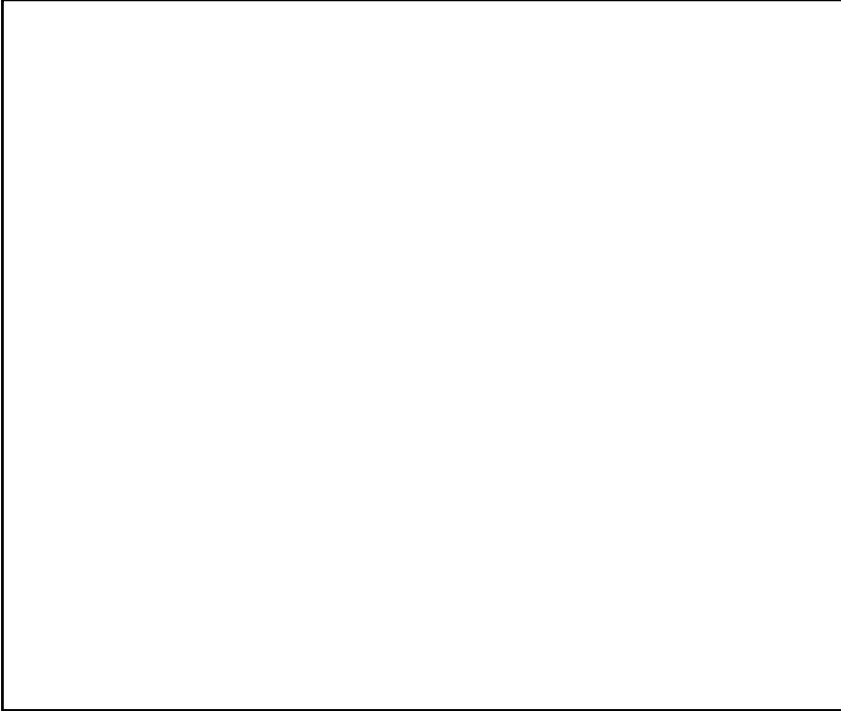
NLP and Word Embeddings

Properties of word embeddings

Analogies

	Man (5391)	Woman (9853)	King (4914)	Queen (7157)	Apple (456)	Orange (6257)
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	0.70	0.69	0.03	-0.02
Food	0.09	0.01	0.02	0.01	0.95	0.97

Analogies using word vectors



$$e_{man} - e_{woman} \approx e_{king} - e_?$$

Cosine similarity

$$\text{sim}(e_w, e_{king} - e_{man} + e_{woman})$$

Man:Woman as Boy:Girl

Ottawa:Canada as Nairobi:Kenya

Big:Bigger as Tall:Taller

Yen:Japan as Ruble:Russia



deeplearning.ai

NLP and Word Embeddings

Embedding matrix

Embedding matrix

In practice, use specialized function to look up an embedding.



deeplearning.ai

NLP and Word Embeddings

Learning word embeddings

Neural language model

I want a glass of orange _____.

4343 9665 1 3852 6163 6257

I o_{4343} \longrightarrow E \longrightarrow e_{4343}

want o_{9665} \longrightarrow E \longrightarrow e_{9665}

a o_1 \longrightarrow E \longrightarrow e_1

glass o_{3852} \longrightarrow E \longrightarrow e_{3852}

of o_{6163} \longrightarrow E \longrightarrow e_{6163}

orange o_{6257} \longrightarrow E \longrightarrow e_{6257}

Other context/target pairs

I want a glass of orange juice to go along with my cereal.

Context: Last 4 words.

4 words on left & right

Last 1 word

Nearby 1 word



deeplearning.ai

NLP and Word Embeddings

Word2Vec

Skip-grams

I want a glass of orange juice to go along with my cereal.

Model

Vocab size = 10,000k

Problems with softmax classification

$$p(t|c) = \frac{e^{\theta_t^T e_c}}{\sum_{j=1}^{10,000} e^{\theta_j^T e_c}}$$

How to sample the context c ?



deeplearning.ai

NLP and Word Embeddings

Negative sampling

Defining a new learning problem

I want a glass of orange juice to go along with my cereal.

Model

Softmax:
$$p(t|c) = \frac{e^{\theta_t^T e_c}}{\sum_{j=1}^{10,000} e^{\theta_j^T e_c}}$$

<u>context</u>	<u>word</u>	<u>target?</u>
orange	juice	1
orange	king	0
orange	book	0
orange	the	0
orange	of	0

Selecting negative examples

<u>context</u>	<u>word</u>	<u>target?</u>
orange	juice	1
orange	king	0
orange	book	0
orange	the	0
orange	of	0



deeplearning.ai

NLP and Word Embeddings

GloVe word vectors

GloVe (global vectors for word representation)

I want a glass of orange juice to go along with my cereal.

Model

A note on the featurization view of word embeddings

	Man (5391)	Woman (9853)	King (4914)	Queen (7157)
Gender	-1	1	-0.95	0.97
Royal	0.01	0.02	0.93	0.95
Age	0.03	0.02	0.70	0.69
Food	0.09	0.01	0.02	0.01

$$\text{minimize } \sum_{i=1}^{10,000} \sum_{j=1}^{10,000} f(X_{ij}) (\theta_i^T e_j + b_i - b'_j - \log X_{ij})^2$$



deeplearning.ai

NLP and Word Embeddings

Sentiment classification

Sentiment classification problem

x

y

The dessert is excellent.



Service was quite slow.



Good for a quick meal, but nothing special.



Completely lacking in good taste, good service, and good ambience.



Simple sentiment classification model

The dessert is excellent
8928 2468 4694 3180



The o_{8928} \longrightarrow E \longrightarrow e_{8928}

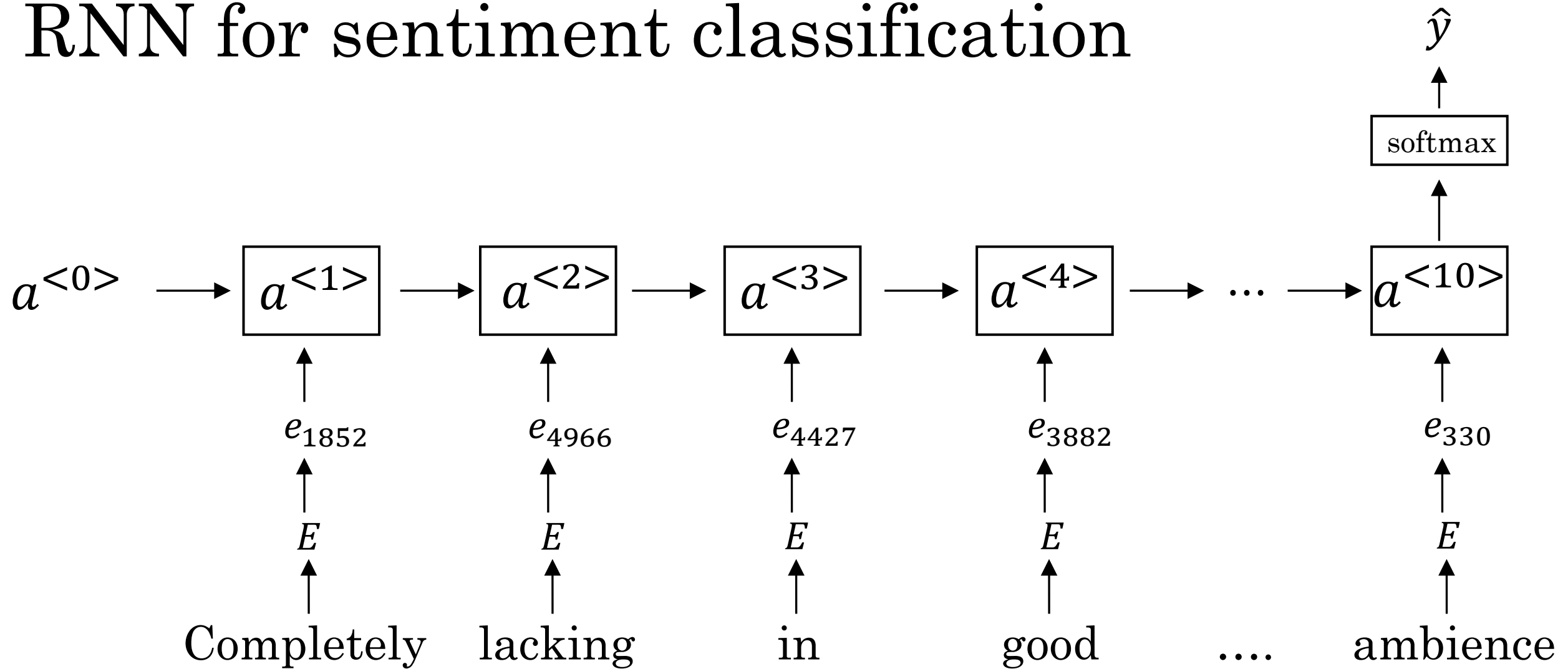
dessert o_{2468} \longrightarrow E \longrightarrow e_{2468}

is o_{4694} \longrightarrow E \longrightarrow e_{4694}

excellent o_{3180} \longrightarrow E \longrightarrow e_{3180}

“Completely lacking in good
taste, good service, and good
ambience.”

RNN for sentiment classification





deeplearning.ai

NLP and Word Embeddings

Debiasing word embeddings

The problem of bias in word embeddings

Man:Woman as King:Queen

Man:Computer_Programmer as Woman:Homemaker

Father:Doctor as Mother:Nurse

Word embeddings can reflect gender, ethnicity, age, sexual orientation, and other biases of the text used to train the model.

Addressing bias in word embeddings

1. Identify bias direction.

2. Neutralize: For every word that is not definitional, project to get rid of bias.

3. Equalize pairs.



deeplearning.ai

Sequence to sequence models

Basic models

Sequence to sequence model

$x^{<1>} \quad x^{<2>} \quad x^{<3>} \quad x^{<4>} \quad x^{<5>}$
Jane visite l'Afrique en septembre

→ Jane is visiting Africa in September.

$y^{<1>} \quad y^{<2>} \quad y^{<3>} \quad y^{<4>} \quad y^{<5>} \quad y^{<6>}$

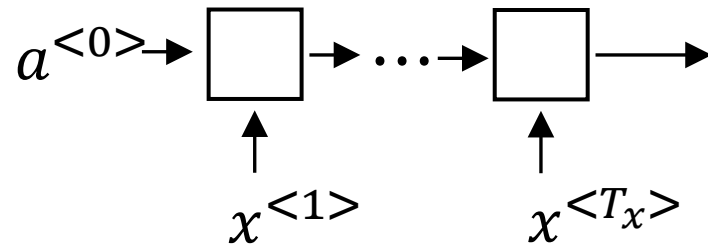
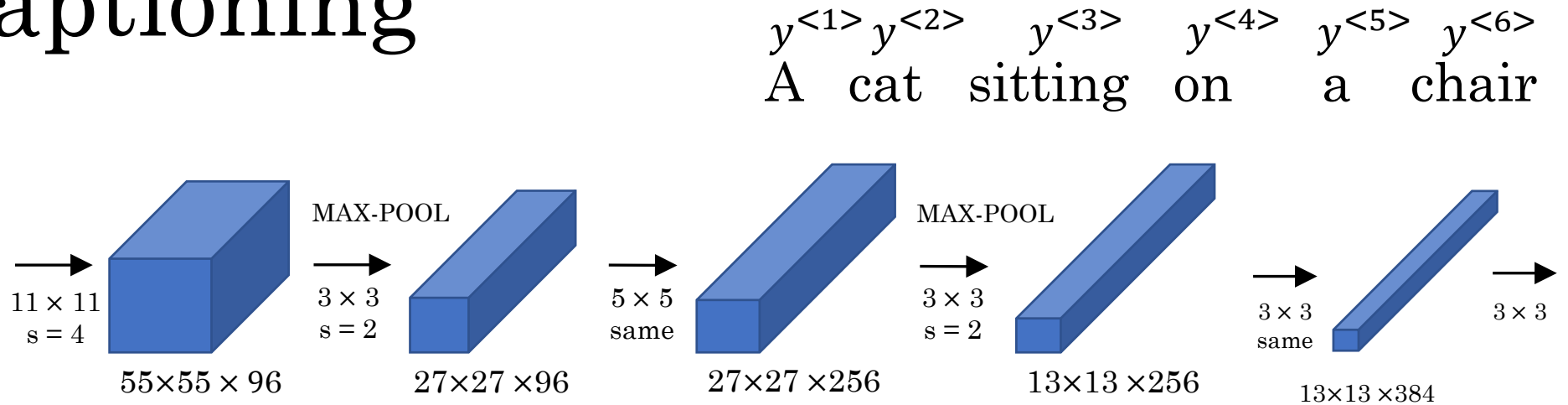
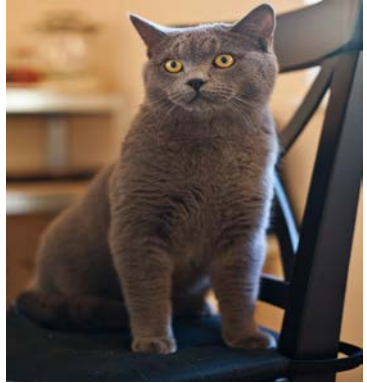
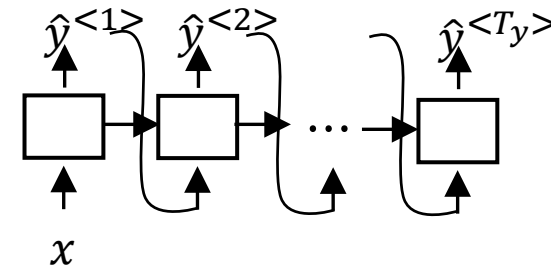
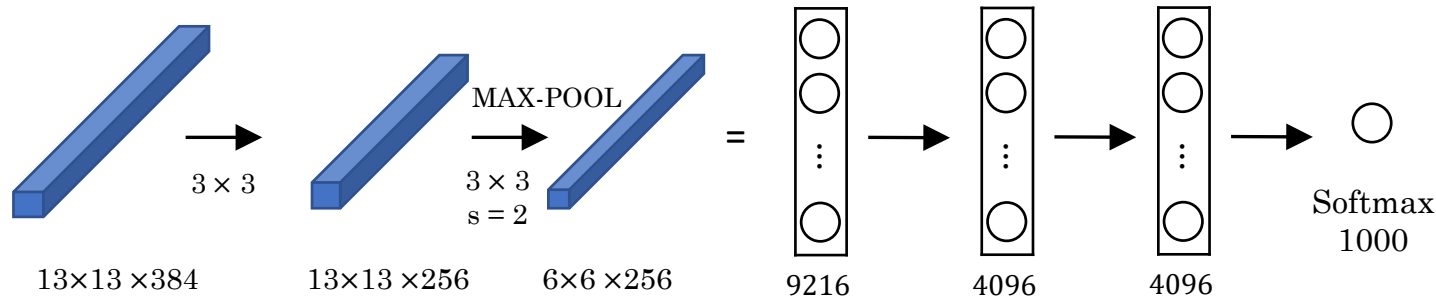


Image captioning



$y^{<1>}$ $y^{<2>}$ $y^{<3>}$ $y^{<4>}$ $y^{<5>}$ $y^{<6>}$
A cat sitting on a chair



[Mao et. al., 2014. Deep captioning with multimodal recurrent neural networks]

[Vinyals et. al., 2014. Show and tell: Neural image caption generator]

[Karpathy and Li, 2015. Deep visual-semantic alignments for generating image descriptions]

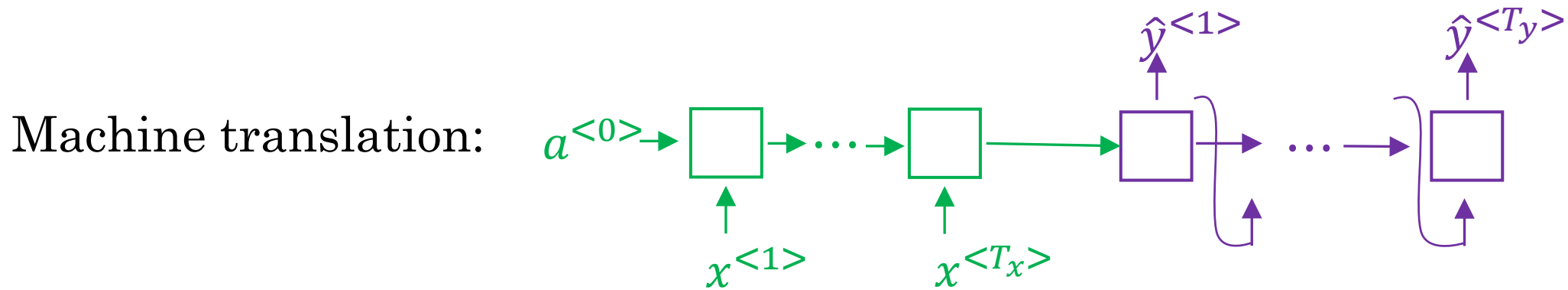
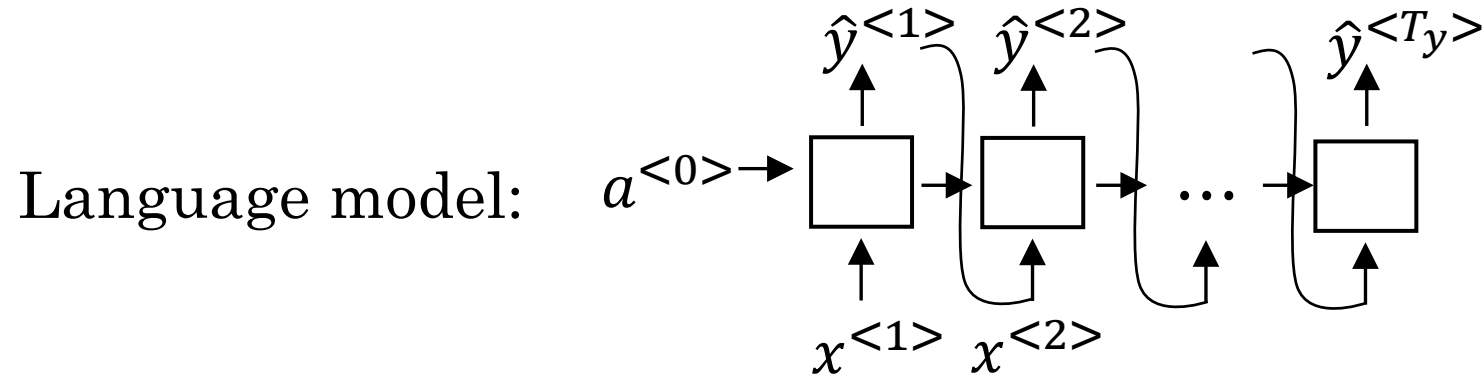


deeplearning.ai

Sequence to sequence models

Picking the most likely sentence

Machine translation as building a conditional language model



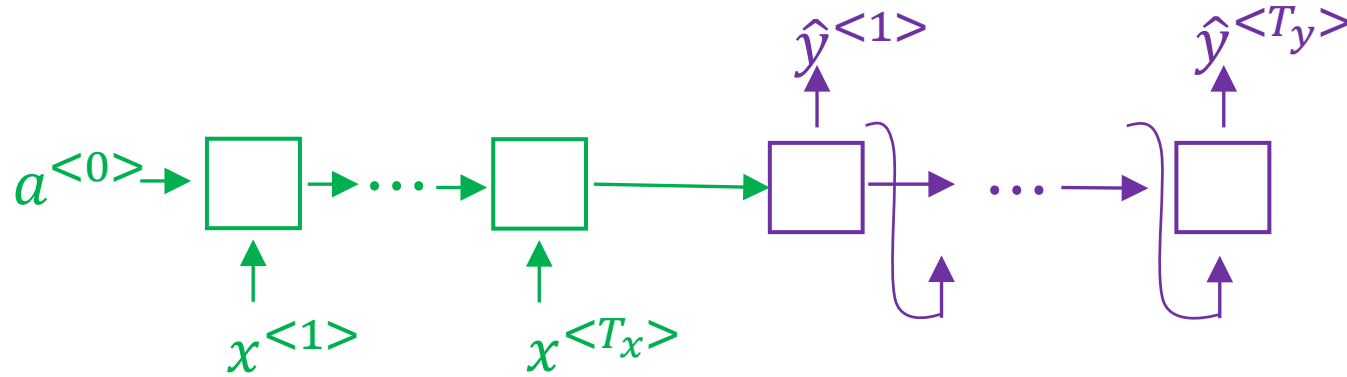
Finding the most likely translation

Jane visite l'Afrique en septembre. $P(y^{<1>}, \dots, y^{<T_y>} | x)$

- Jane is visiting Africa in September.
- Jane is going to be visiting Africa in September.
- In September, Jane will visit Africa.
- Her African friend welcomed Jane in September.

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

Why not a greedy search?



- Jane is visiting Africa in September.
- Jane is going to be visiting Africa in September.



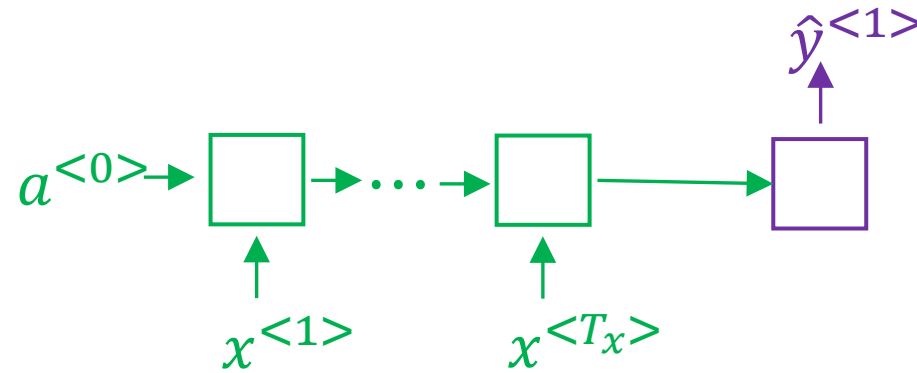
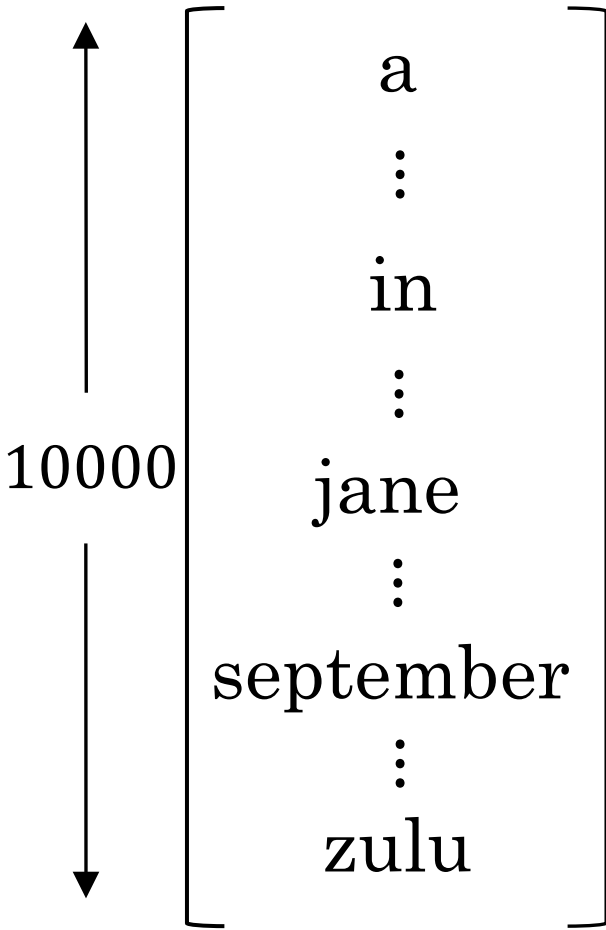
deeplearning.ai

Sequence to sequence models

Beam search

Beam search algorithm

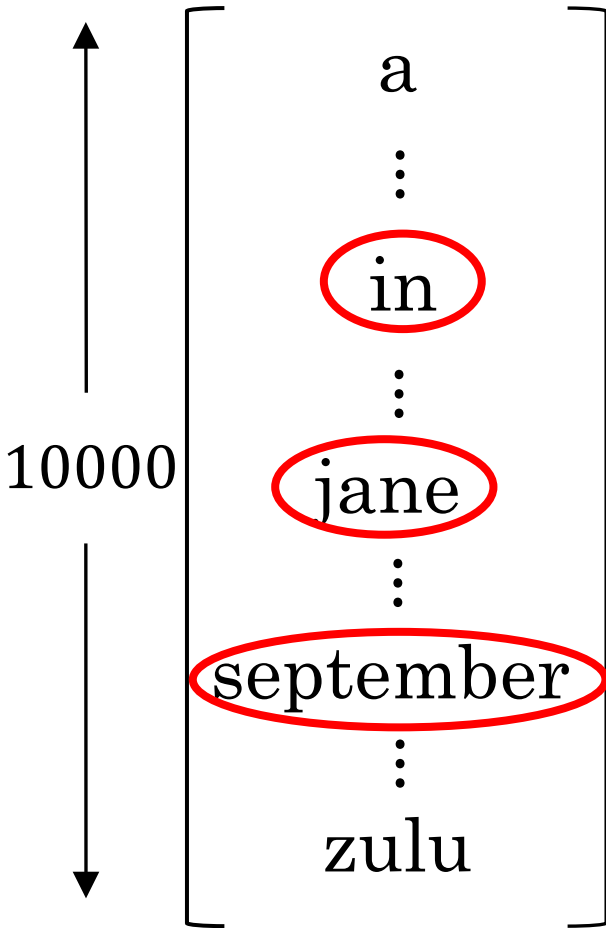
Step 1



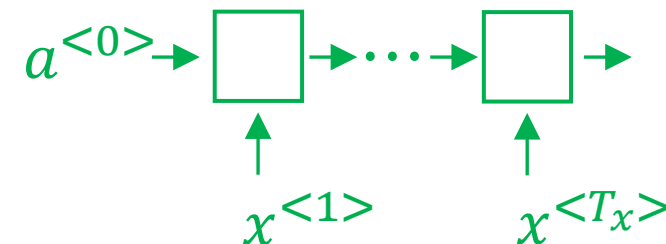
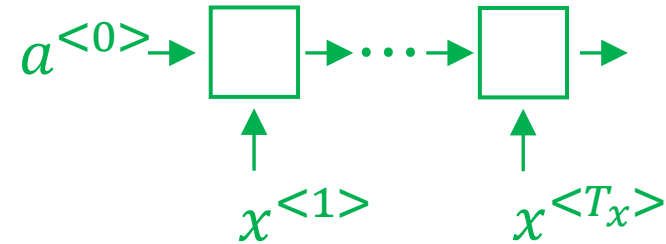
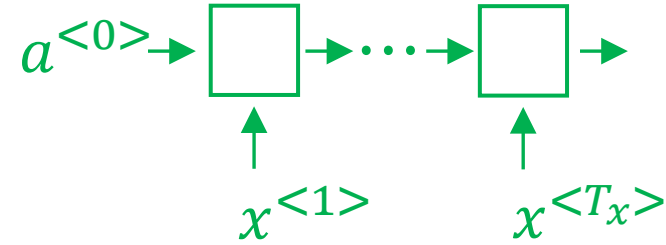
$$P(y^{<1>} | x)$$

Beam search algorithm

Step 1

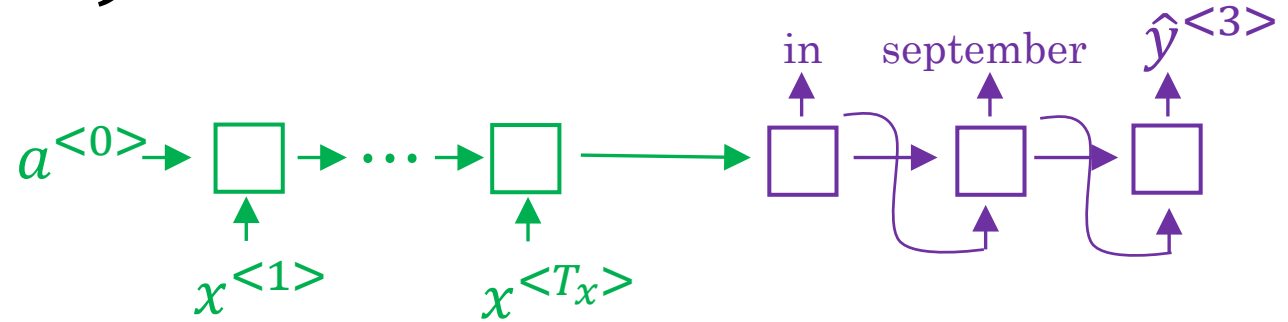


Step 2

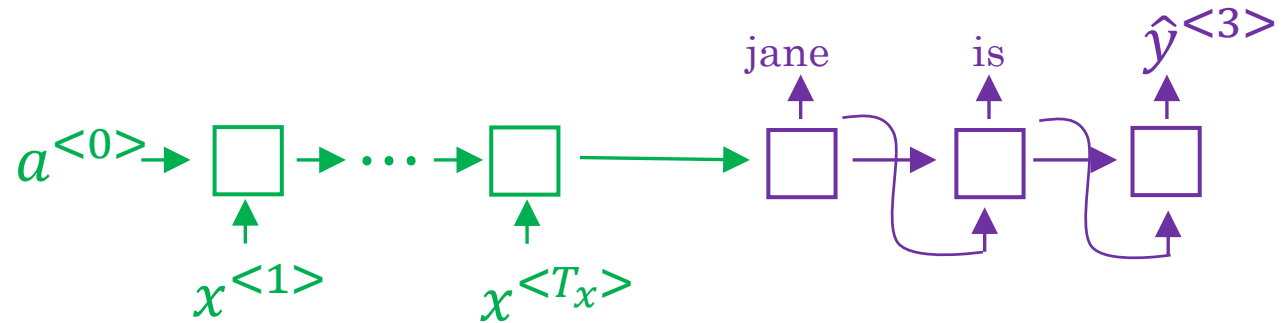


Beam search ($B = 3$)

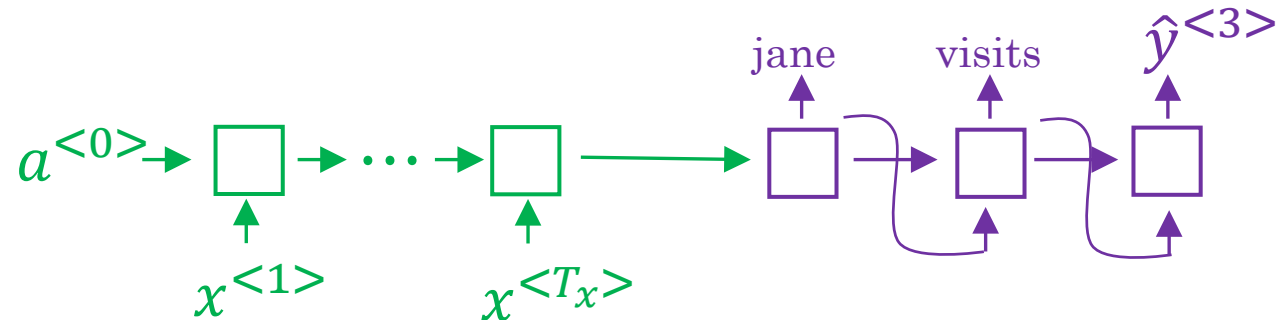
in september



jane is



jane visits



$$P(y^{<1>}, y^{<2>} | x)$$

jane visits africa in september. <EOS>



deeplearning.ai

Sequence to sequence models

Refinements to beam search

Length normalization

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

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

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

Beam search discussion

Beam width B?

Unlike exact search algorithms like BFS (Breadth First Search) or DFS (Depth First Search), Beam Search runs faster but is not guaranteed to find exact maximum for $\arg \max_y P(y|x)$.



deeplearning.ai

Sequence to sequence models

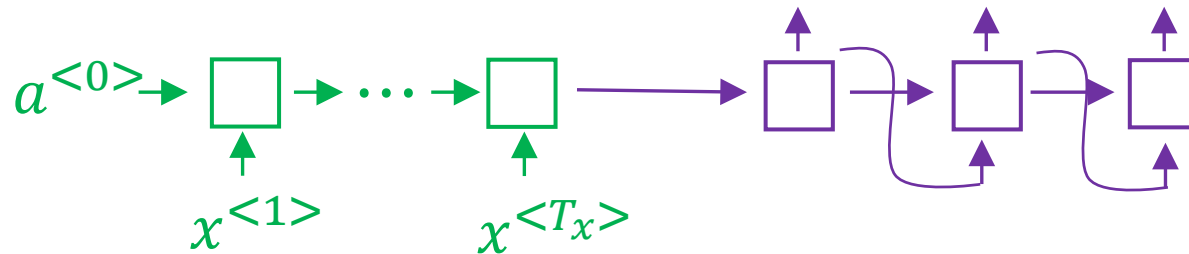
Error analysis on beam search

Example

Jane visite l'Afrique en septembre.

Human: Jane visits Africa in September.

Algorithm: Jane visited Africa last September.



Error analysis on beam search

Human: Jane visits Africa in September. (y^*)

Algorithm: Jane visited Africa last September. (\hat{y})

Case 1:

Beam search chose \hat{y} . But y^* attains higher $P(y|x)$.

Conclusion: Beam search is at fault.

Case 2:

y^* is a better translation than \hat{y} . But RNN predicted $P(y^*|x) < P(\hat{y}|x)$.

Conclusion: RNN model is at fault.

Error analysis process

Human	Algorithm	$P(y^* x)$	$P(\hat{y} x)$	At fault?
Jane visits Africa in September.	Jane visited Africa last September.			

Figures out what fraction of errors are “due to” beam search vs. RNN model



deeplearning.ai

Sequence to sequence models

Bleu score (optional)

Evaluating machine translation

French: Le chat est sur le tapis.

Reference 1: The cat is on the mat.

Reference 2: There is a cat on the mat.

MT output: the the the the the the the.

Precision:

Modified precision:

Bleu score on bigrams

Example: Reference 1: The cat is on the mat.

Reference 2: There is a cat on the mat.

MT output: The cat the cat on the mat.

the cat

cat the

cat on

on the

the mat

Bleu score on unigrams

Example: Reference 1: The cat is on the mat.

Reference 2: There is a cat on the mat.

MT output: The cat the cat on the mat.

$$p_1 = \frac{\sum_{unigram \in \hat{y}} count_{clip}(unigram)}{\sum_{unigram \in \hat{y}} count(unigram)}$$

$$p_n = \frac{\sum_{ngram \in \hat{y}} count_{clip}(ngram)}{\sum_{ngram \in \hat{y}} count(ngram)}$$

Bleu details

p_n = Bleu score on n-grams only

Combined Bleu score:

$$\text{BP} = \begin{cases} 1 & \text{if MT_output_length} > \text{reference_output_length} \\ \exp(1 - \text{MT_output_length}/\text{reference_output_length}) & \text{otherwise} \end{cases}$$

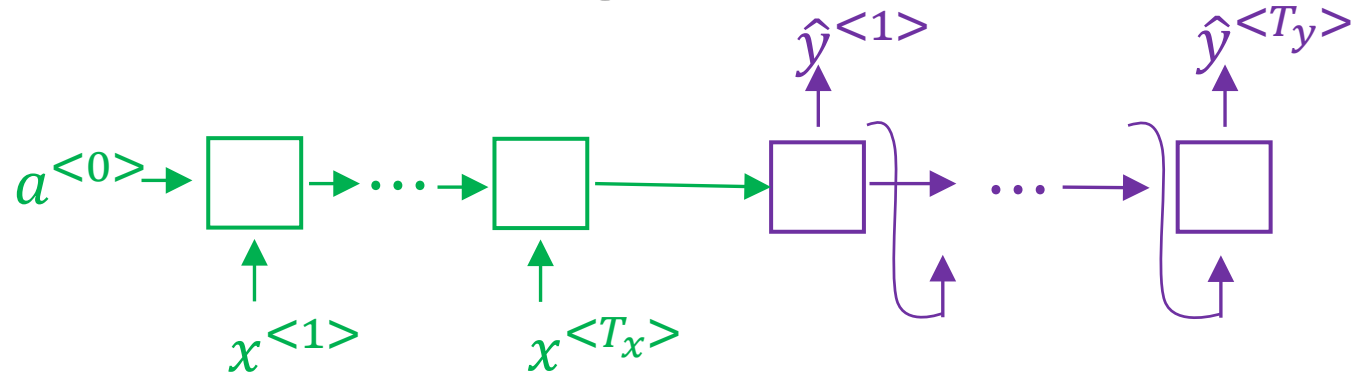


deeplearning.ai

Sequence to sequence models

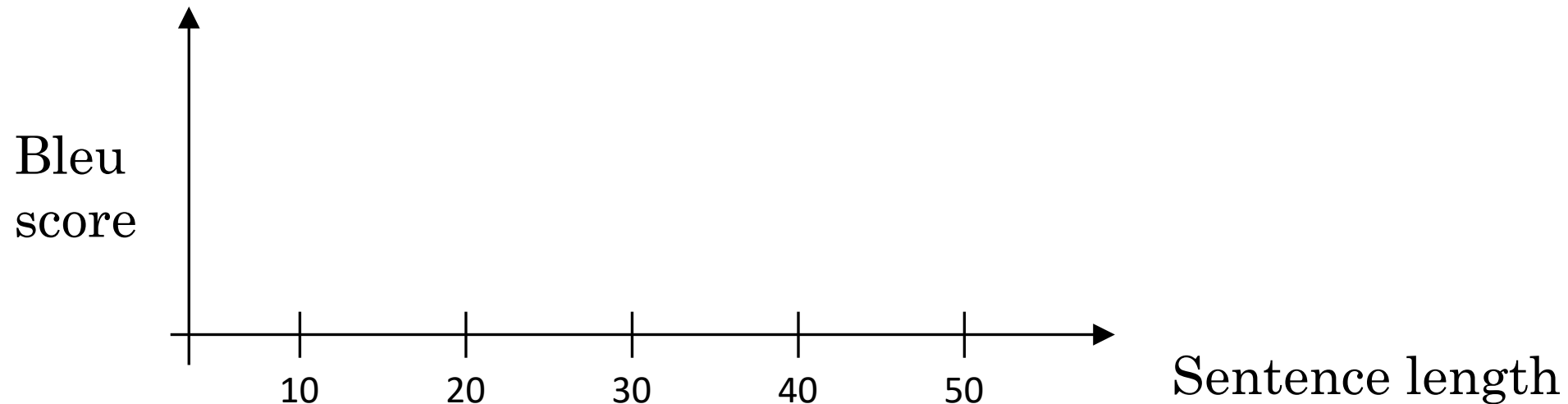
Attention model intuition

The problem of long sequences

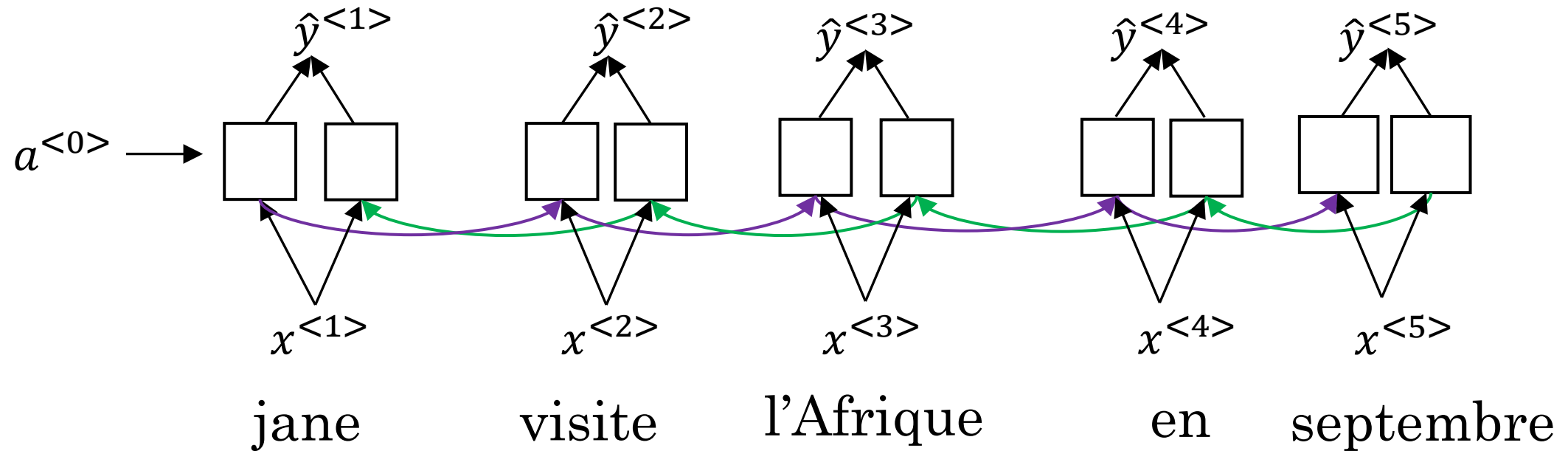


Jane s'est rendue en Afrique en septembre dernier, a apprécié la culture et a rencontré beaucoup de gens merveilleux; elle est revenue en parlant comment son voyage était merveilleux, et elle me tente d'y aller aussi.

Jane went to Africa last September, and enjoyed the culture and met many wonderful people; she came back raving about how wonderful her trip was, and is tempting me to go too.



Attention model intuition



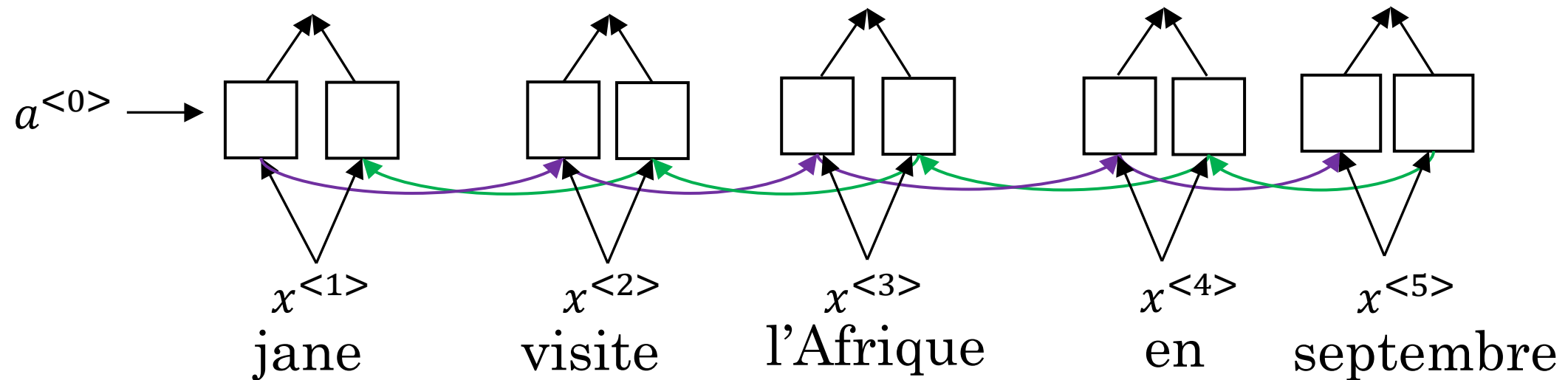


deeplearning.ai

Sequence to sequence models

Attention model

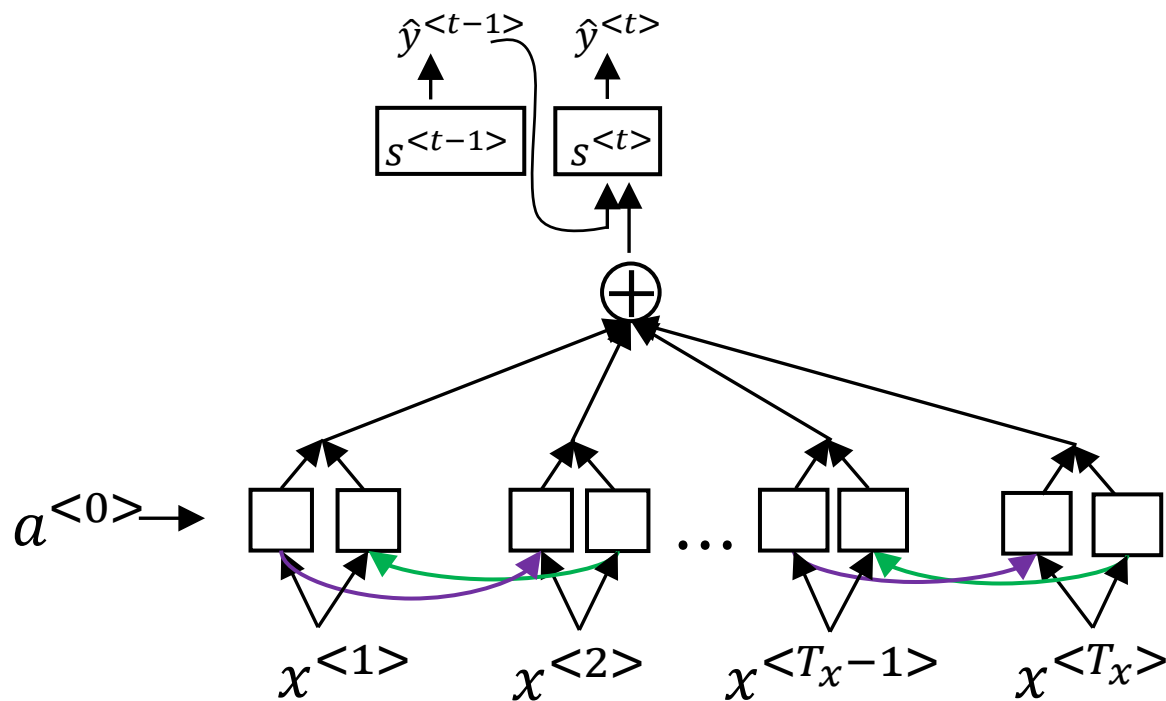
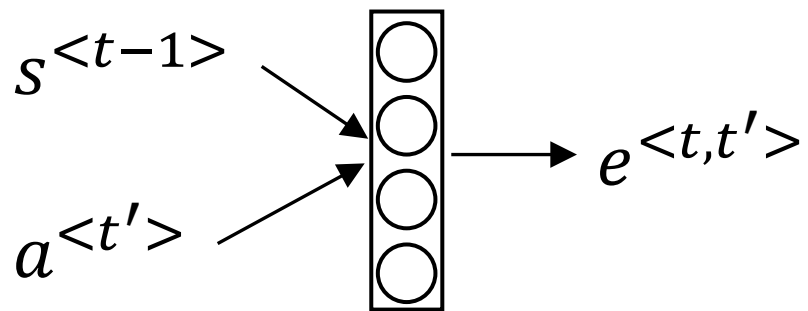
Attention model



Computing attention $\alpha^{<t,t'>}$

$\alpha^{<t,t'>}$ = amount of attention $y^{<t>}$ should pay to $a^{<t'>}$

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

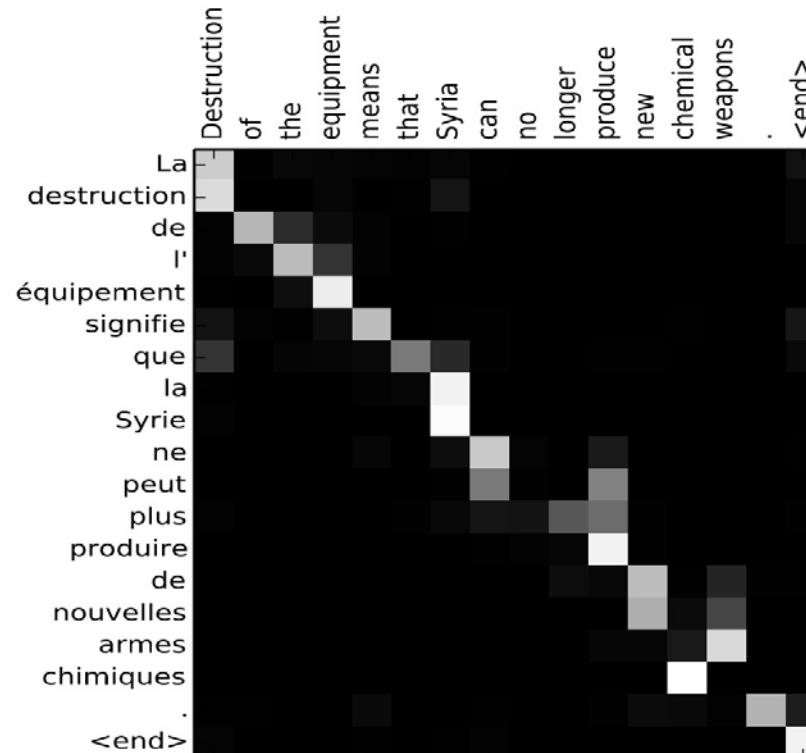


Attention examples

July 20th 1969 → 1969 – 07 – 20

23 April, 1564 → 1564 – 04 – 23

Visualization of $\alpha^{<t,t'>}$:





deeplearning.ai

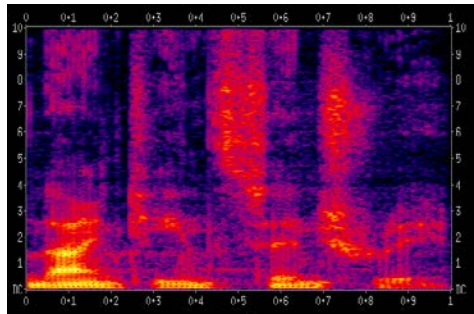
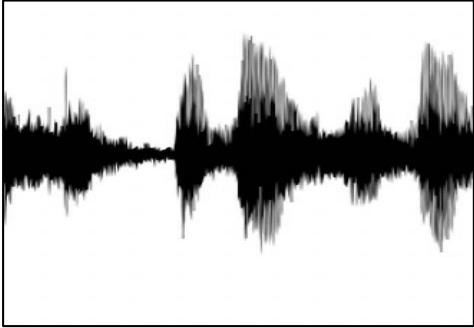
Audio data

Speech recognition

Speech recognition problem

 x

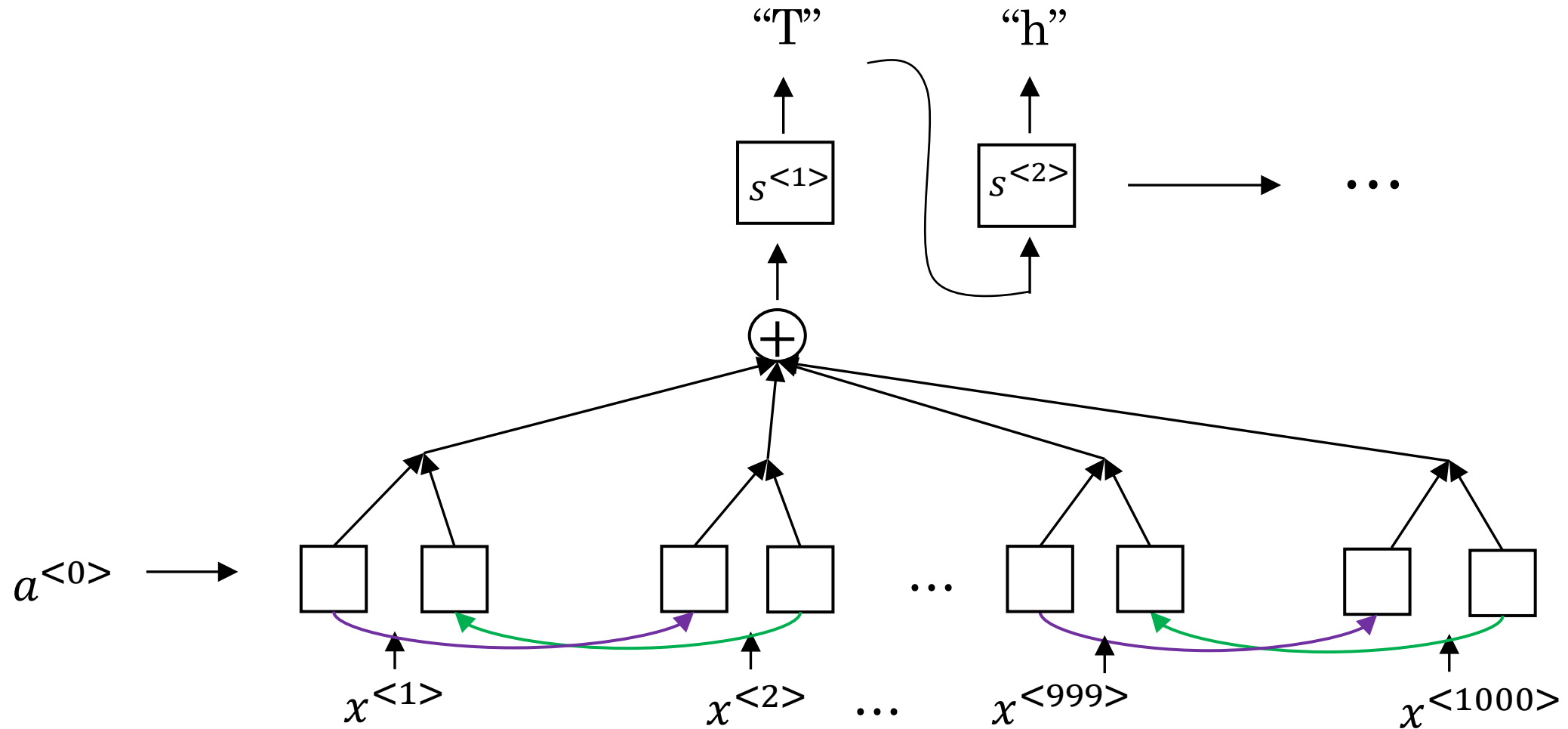
audio clip

 y

transcript

“the quick brown fox”

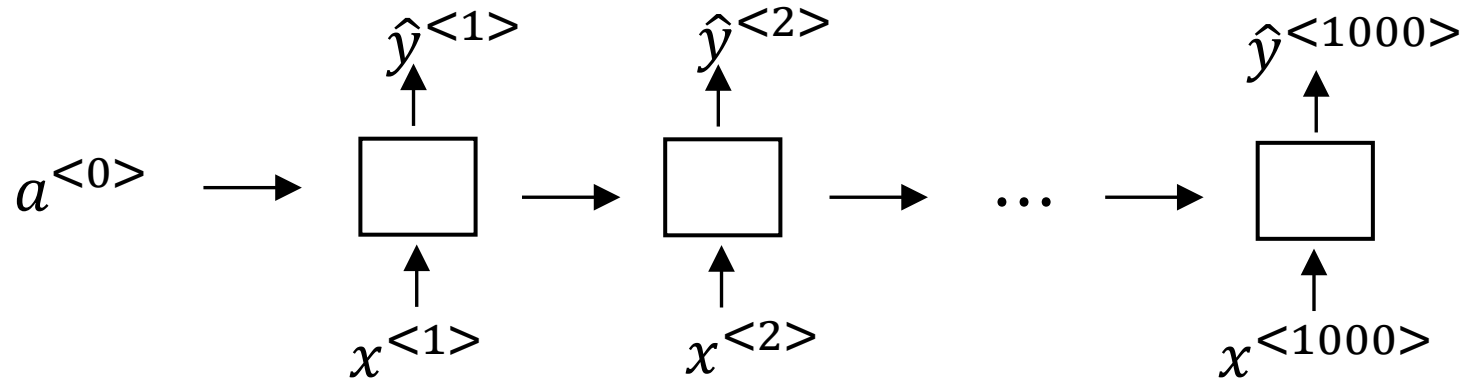
Attention model for speech recognition



CTC cost for speech recognition

(Connectionist temporal classification)

“the quick brown fox”



Basic rule: collapse repeated characters not separated by “blank”

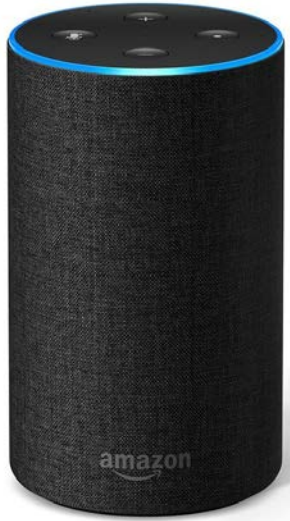


deeplearning.ai

Audio data

Trigger word detection

What is trigger word detection?



Amazon Echo
(Alexa)



Baidu DuerOS
(xiaodunihao)

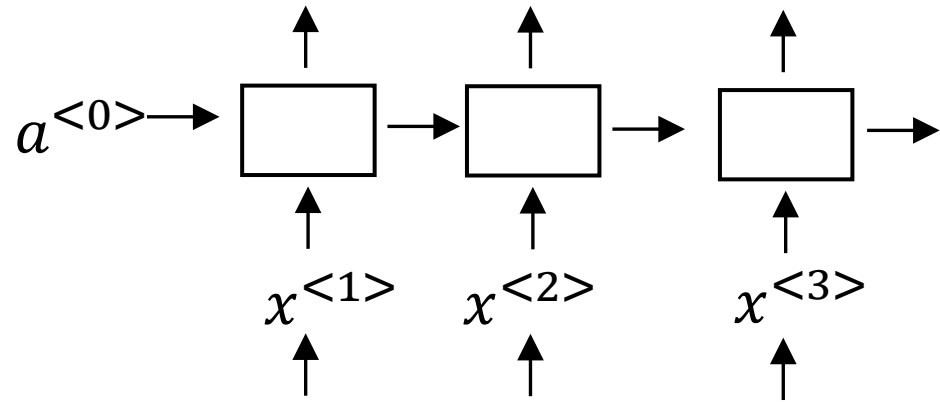


Apple Siri
(Hey Siri)



Google Home
(Okay Google)

Trigger word detection algorithm





deeplearning.ai

Conclusion

Summary and thank you

Specialization outline

1. Neural Networks and Deep Learning
2. Improving Deep Neural Networks: Hyperparameter tuning, Regularization and Optimization
3. Structuring Machine Learning Projects
4. Convolutional Neural Networks
5. Sequence Models

Deep learning is a super power

Please buy this
from shutterstock
and replace in
final video.



www.shutterstock.com · 331201091

Thank you.

- Andrew Ng