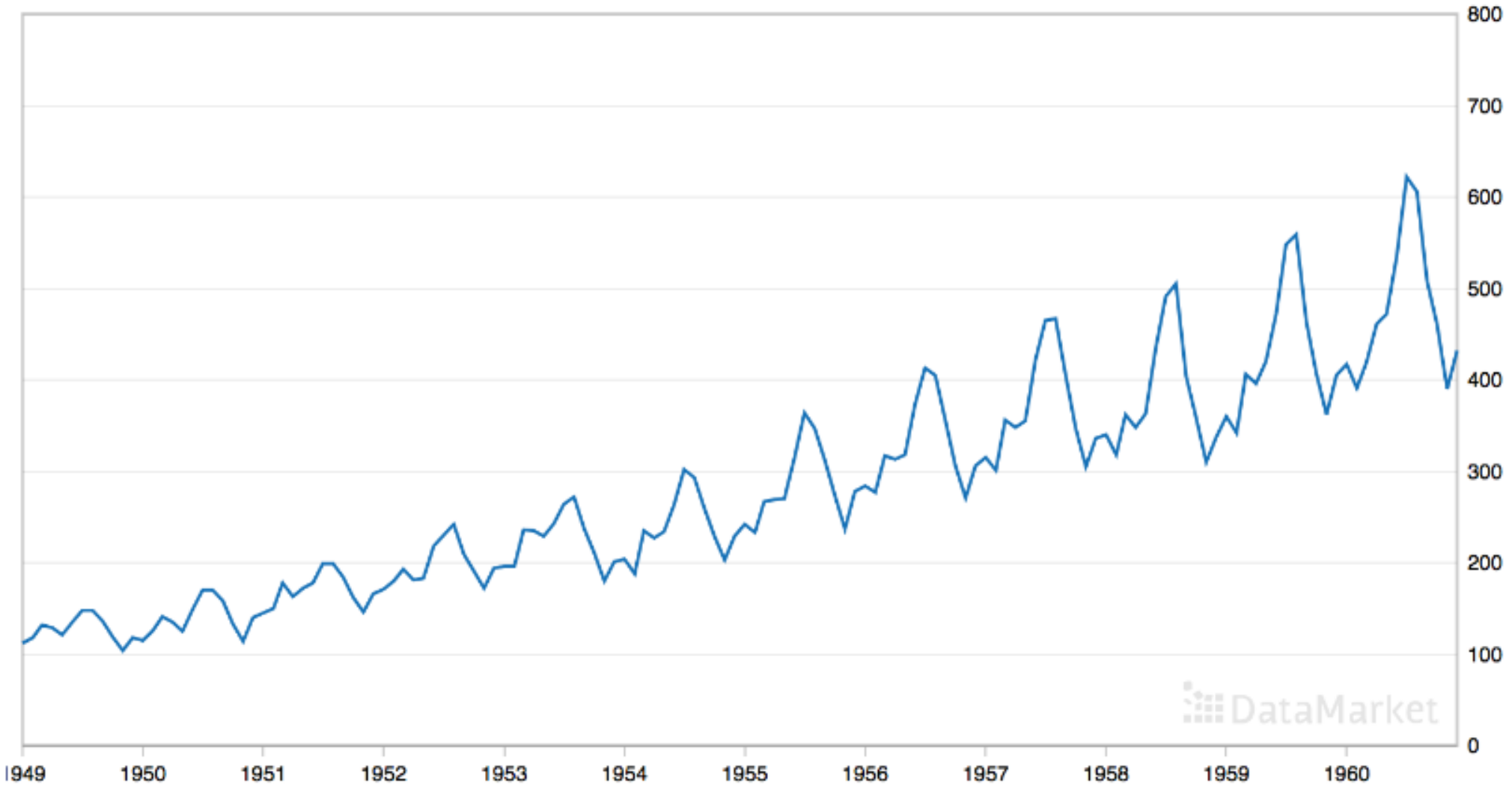
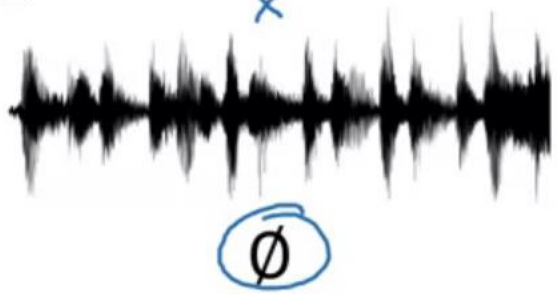
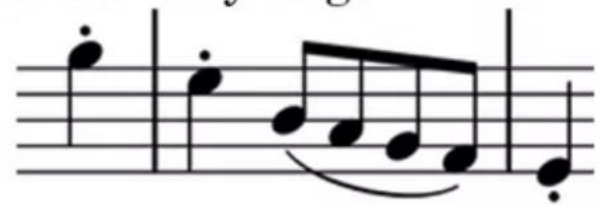
Timeseries (Airline Sales)



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	→	AGCCCCTGTGAGGAACTAG
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.

Andrew Ng

RNN/LSTM is heavily used in stock price forecasting researches

Table 1: Stock Price Forecasting Using Only Raw Time Series Data

Art.	Data Set	Period	Feature Set	Lag	Horizon	Method	Performance Criteria	Env.
[80]	38 stocks in KOSPI	2010-2014	Lagged stock returns	50min	5min	DNN	NMSE, RMSE, MAE, MI	-
[81]	China stock market, 3049 Stocks	1990-2015	OCHLV	30d	3d	LSTM	Accuracy	Theano, Keras
[82]	Daily returns of 'BRD' stock in Romanian Market	2001-2016	OCHLV	-	1d	LSTM	RMSE, MAE	Python, Theano
[83]	297 listed companies of CSE	2012-2013	OCHLV	2d	1d	LSTM, SRNN, GRU	MAD, MAPE	Keras
[84]	5 stock in NSE	1997-2016	OCHLV, Price data, turnover and number of trades.	200d	1..10d	LSTM, RNN, CNN, MLP	MAPE	-
[85]	Stocks of Infosys, TCS and CIPLA from NSE	2014	Price data	-	-	RNN, LSTM and CNN	Accuracy	-
[86]	10 stocks in S&P500	1997-2016	OCHLV, Price data	36m	1m	RNN, LSTM, GRU	Accuracy, Monthly return	Keras, Tensorflow
[87]	Stocks data from S&P500	2011-2016	OCHLV	1d	1d	DBN	MSE, norm-RMSE, MAE	-
[88]	High-frequency transaction data of the CSI300 futures	2017	Price data	-	1min	DNN, ELM, RBF	RMSE, MAPE, Accuracy	Matlab
[89]	Stocks in the S&P500	1990-2015	Price data	240d	1d	DNN, GBT, RF	Mean return, MDD, Calmar ratio	H2O
[90]	ACI Worldwide, Staples, and Seagate in NASDAQ	2006-2010	Daily closing prices	17d	1d	RNN, ANN	RMSE	-
[91]	Chinese Stocks	2007-2017	OCHLV	30d	1..5d	CNN + LSTM	Annualized Return, Mxm Retracement	Python
[92]	20 stocks in S&P500	2010-2015	Price data	-	-	AE + LSTM	Weekly Returns	-
[93]	S&P500	1985-2006	Monthly and daily log-returns	*	1d	DBN+MLP	Validation, Test Error	Theano, Python, Matlab
[94]	12 stocks from SSE Composite Index	2000-2017	OCHLV	60d	1..7d	DWNN	MSE	Tensorflow
[95]	50 stocks from NYSE	2007-2016	Price data	-	1d, 3d, 5d	SFM	MSE	-

Art.	Data Set	Period	Feature Set	Lag	Horizon	Method	Performance Criteria	Env.
[98]	U.S. low-level disaggregated macroeconomic time series	1959-2008	GDP, Unemployment rate, Inventories, etc.	-	-	DNN	R^2	-
[99]	CDAX stock market data	2010-2013	Financial news, stock market data	20d	1d	LSTM	MSE, RMSE, MAE, Accuracy, AUC	TensorFlow, Theano, Python, Scikit-Learn
[100]	Stock of Tsugami Corporation	2013	Price data	-	-	LSTM	RMSE	Keras, Tensorflow
[101]	Stocks in China's A-share	2006-2007	11 technical indicators	-	1d	LSTM	AR, IR, IC	-
[102]	SCI prices	2008-2015	OCHL of change rate, price	7d	-	EmotionalAnalysis + LSTM	MSE	-
[103]	10 stocks in Nikkei 225 and news	2001-2008	Textual information and Stock prices	10d	-	Paragraph Vector + LSTM	Profit	-
[104]	TKC stock in NYSE and QQQQ ETF	1999-2006	Technical indicators, Price	50d	1d	RNN (Jordan-Elman)	Profit, MSE	Java
[105]	10 Stocks in NYSE	-	Price data, Technical indicators	20min	1min	LSTM, MLP	RMSE	-
[106]	42 stocks in China's SSE	2016	OCHLV, Technical Indicators	242min	1min	GAN (LSTM, CNN)	RMSRE, DPA, GAN-F, GAN-D	-
[107]	Google's daily stock data	2004-2015	OCHLV, Technical indicators	20d	1d	$(2D)^2$ PCA + DNN	SMAPE, PCD, MAPE, RMSE, HR, TR, R^2	R, Matlab
[108]	GarantiBank in BIST, Turkey	2016	OCHLV, Volatility, etc.	-	-	PLR, Graves LSTM	MSE, RMSE, MAE, RSE, R^2	Spark
[109]	Stocks in NYSE, AMEX, NASDAQ, TAQ intraday trade	1993-2017	Price, 15 firm characteristics	80d	1d	LSTM+MLP	Monthly return, SR	Python,Keras Tensorflow in AWS
[110]	Private brokerage company's real data of risky transactions	-	250 features: order details, etc.	-	-	CNN, LSTM	F1-Score	Keras, Tensorflow
[111]	Fundamental and Technical Data, Economic Data	-	Fundamental , technical and market information	-	-	CNN	-	-
[112]	The LOB of 5 stocks of Finnish Stock Market	2010	FI-2010 dataset: bid/ask and volume	-	*	WMTR, MDA	Accuracy, Precision, Recall, F1-Score	-
[113]	Returns in NYSE, AMEX, NASDAQ	1975-2017	57 firm characteristics	*	-	Fama-French n-factor model DL	R^2, RMSE	Tensorflow

Source Paper: Financial time series forecasting with deep learning : A systematic literature review: 2005–2019

pad_sequence

	0	1	2	3	4	5	6	7	8	9	...	490	491	492	493	494	495	496	497	498	499
0	0	0	0	0	0	0	0	0	0	0	...	4472	113	103	32	15	16	2	19	178	32
1	0	0	0	0	0	0	0	0	0	0	...	52	154	462	33	89	78	285	16	145	95
2	0	0	0	0	0	0	0	0	0	0	...	106	607	624	35	534	6	227	7	129	113
3	687	23	4	2	2	6	3693	42	38	39	...	26	49	2	15	566	30	579	21	64	2574
4	0	0	0	0	0	0	0	0	0	0	...	19	14	5	2	6	226	251	7	61	113

5 rows × 500 columns

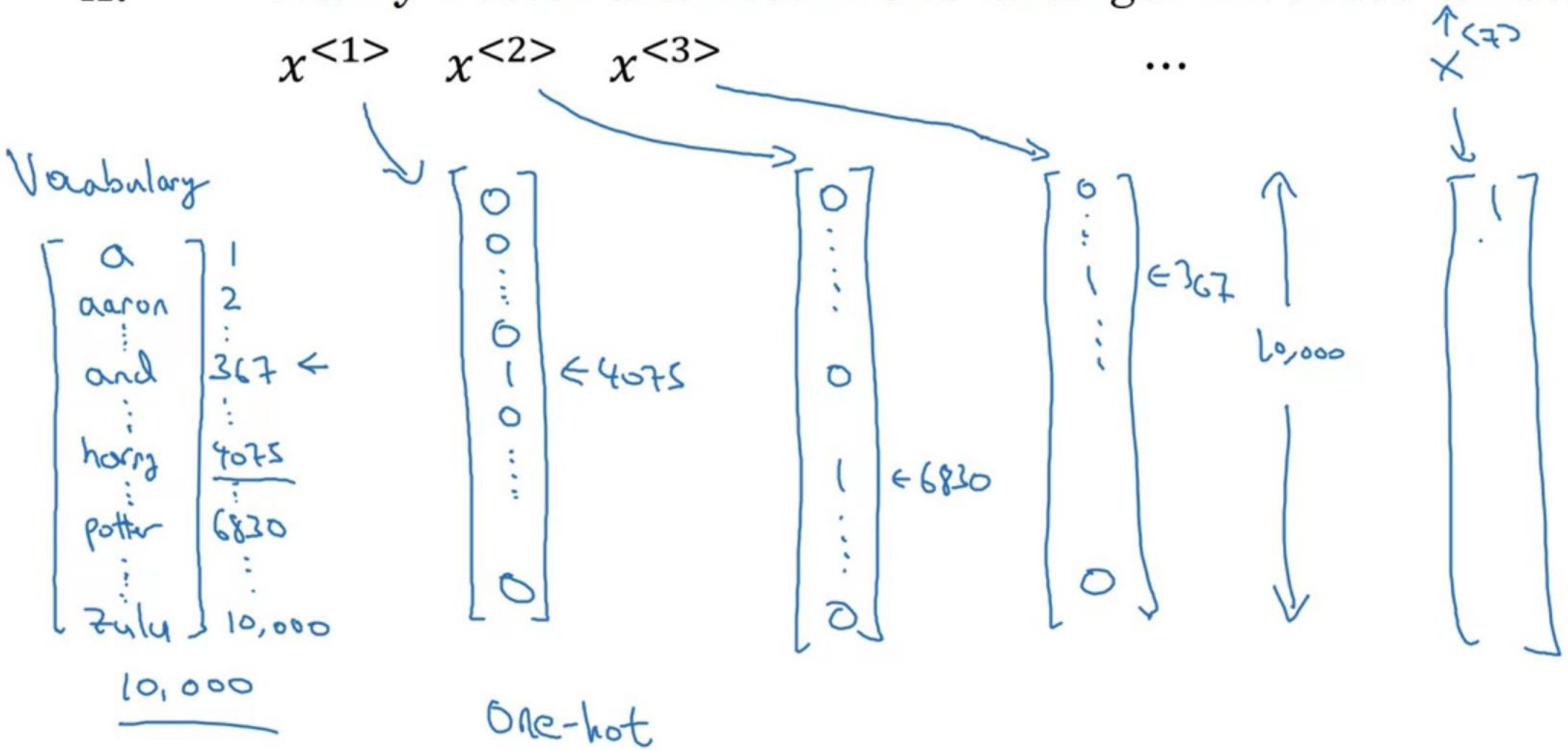
documentation
https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/sequence/pad_sequences

Representing words

One Hot Encoding

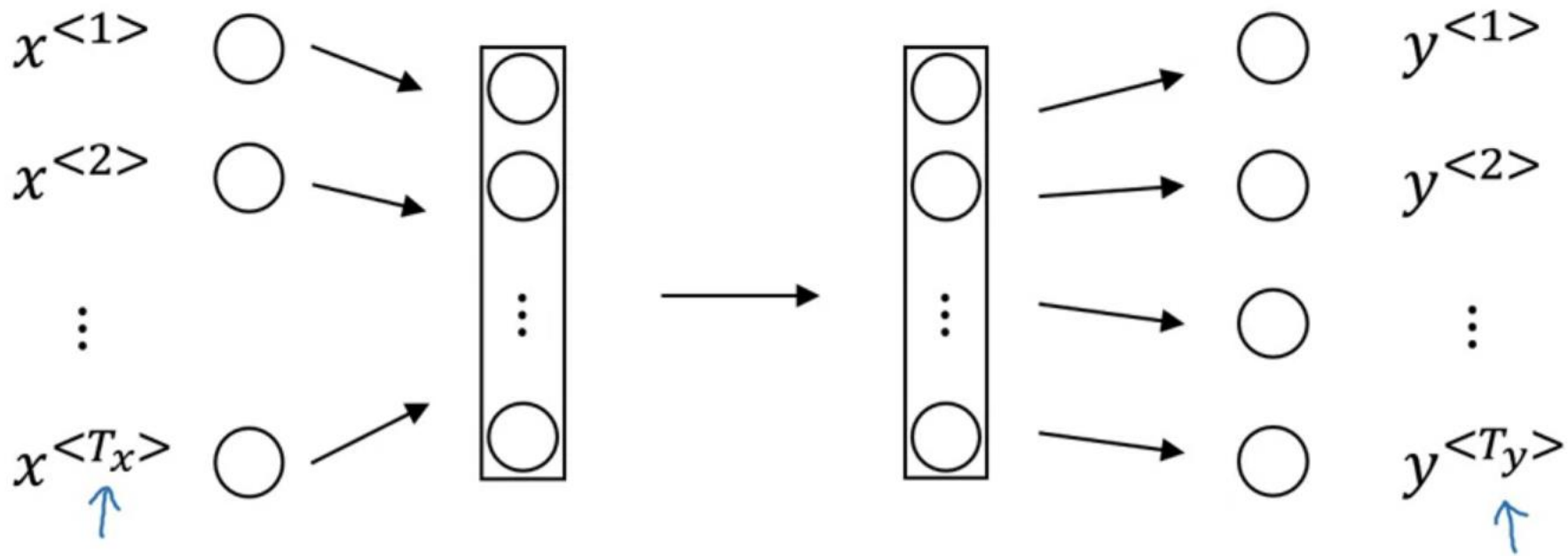
x: Harry Potter and Hermione Granger invented a new spell.

$x^{<1>}$ $x^{<2>}$ $x^{<3>}$ ... $x^{<7>}$ $x^{<9>}$



Andrew Ng

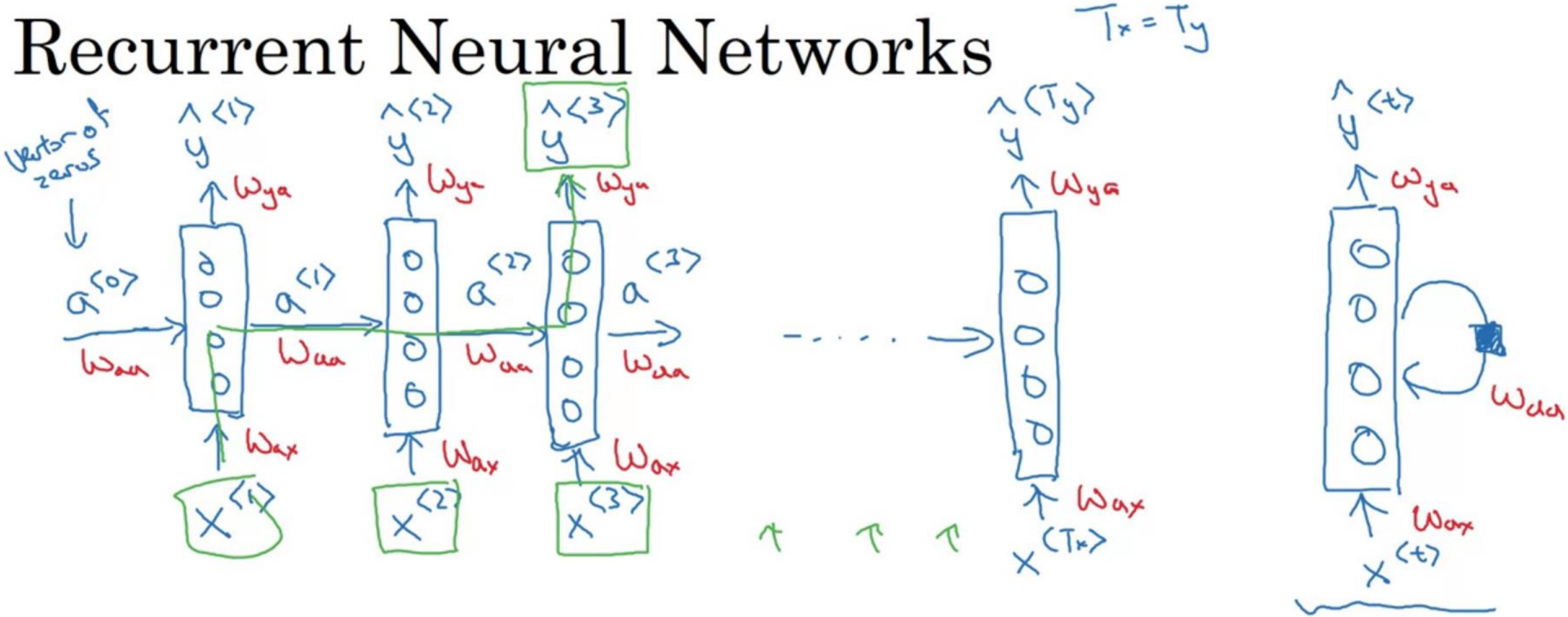
Why not a standard network?



Problems:

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

Recurrent Neural Networks

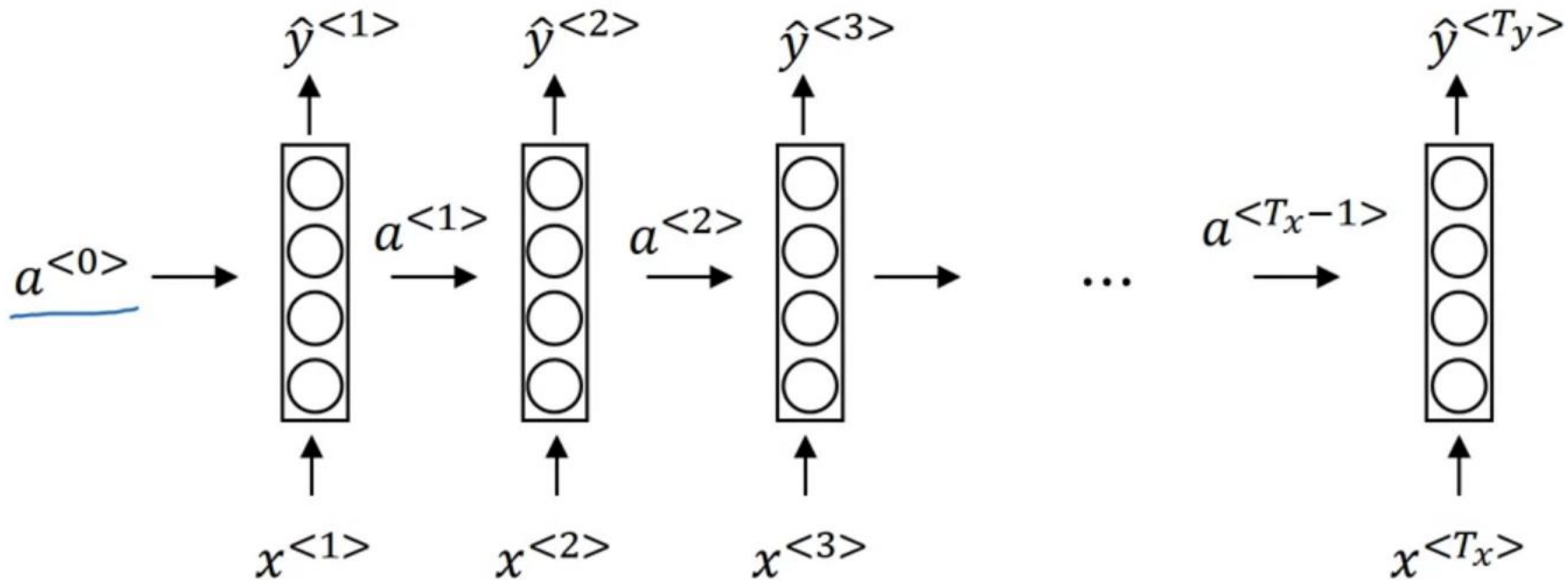


He said, “Teddy Roosevelt was a great President.”

He said, “Teddy bears are on sale!”

Andrew Ng

Forward Propagation



$a^{<0>}$.

Simplified RNN notation

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

W_{aa}: (100, 100) — 100

W_{ax}: (100, 10,000) — 10,000

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

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

$$\begin{bmatrix} W_{aa} \vdots W_{ax} \end{bmatrix} = W_a$$

100 × (100 + 10000)

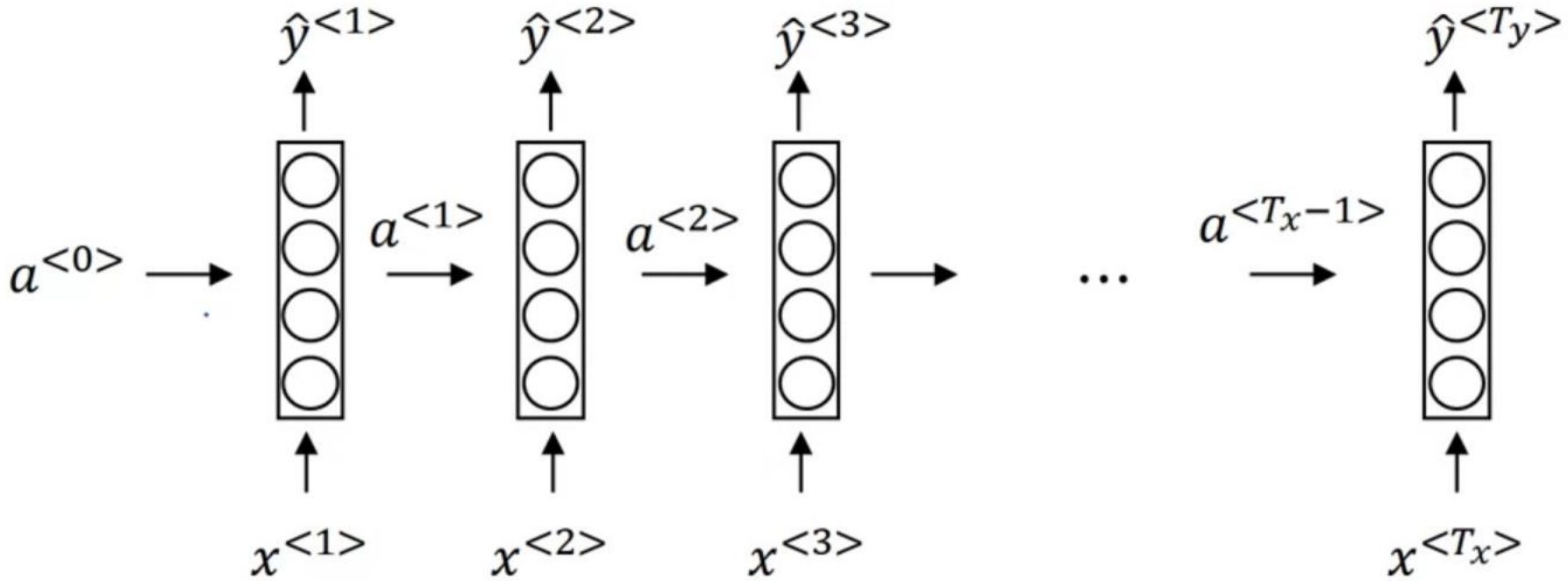
(100, 10100)

$$[a^{<t-1>}, x^{<t>}] = \begin{bmatrix} a^{<t-1>} \\ x^{<t>} \end{bmatrix}$$

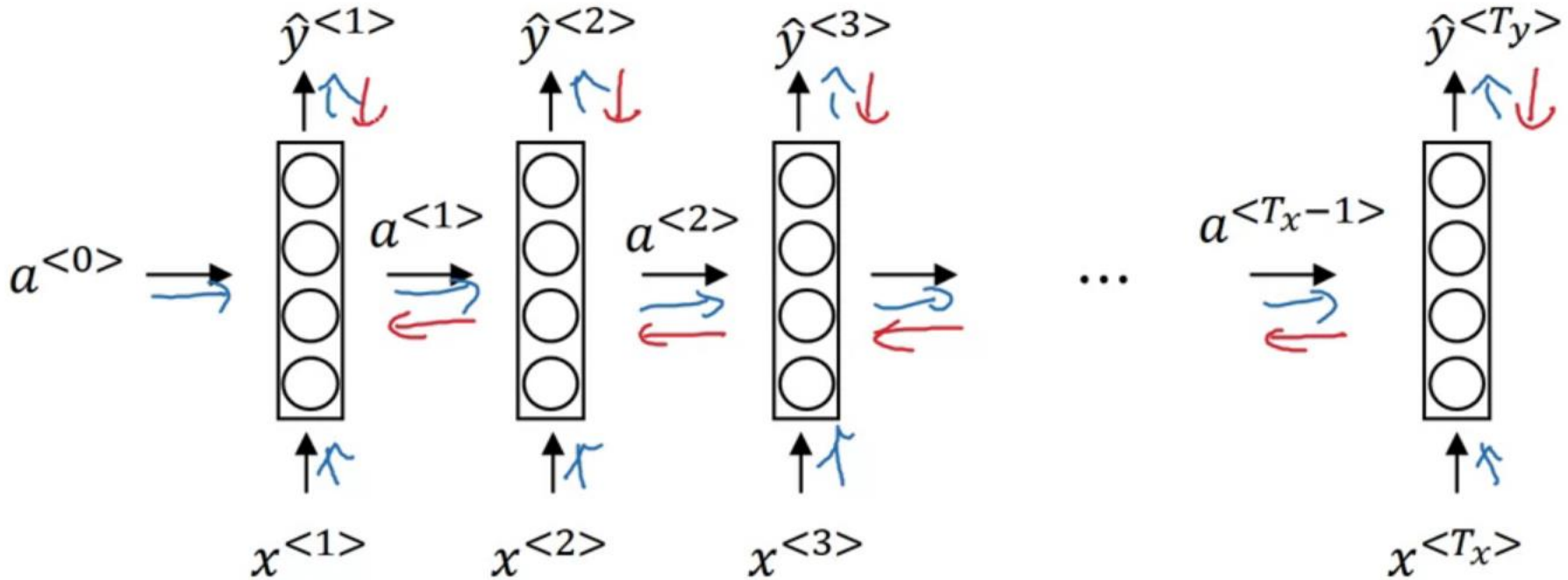
100 + 10000 = 10100

Andrew Ng

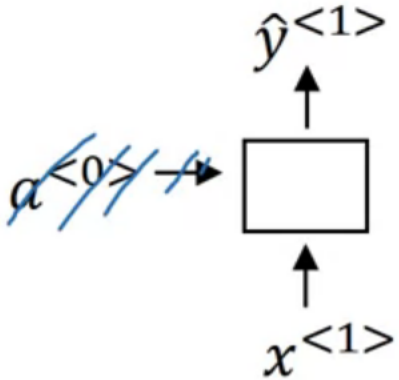
Forward propagation and backpropagation



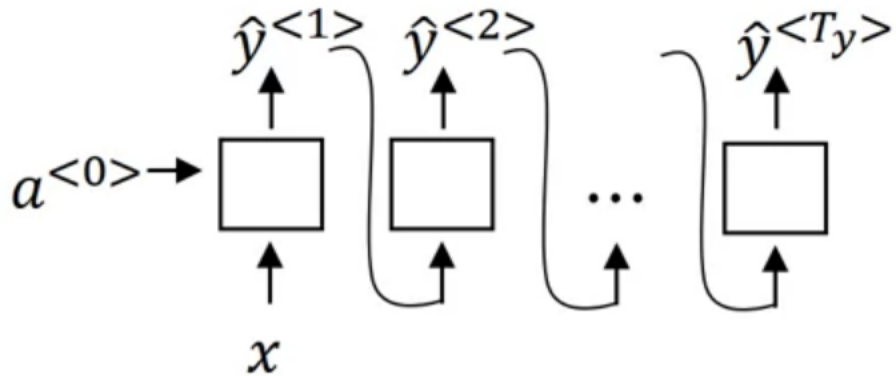
Forward propagation and backpropagation



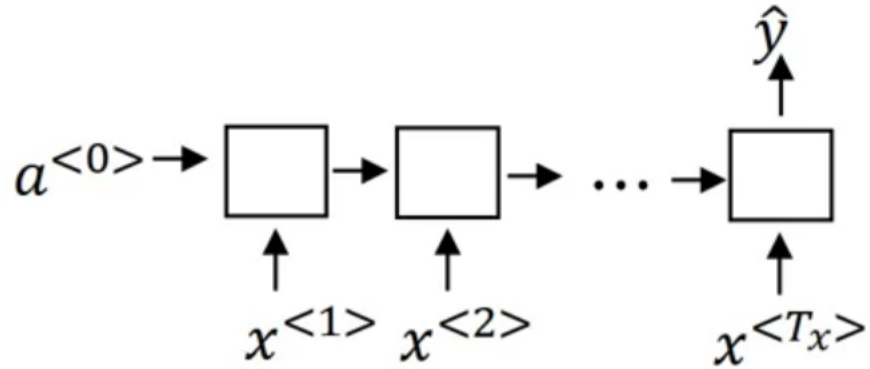
Summary of RNN types



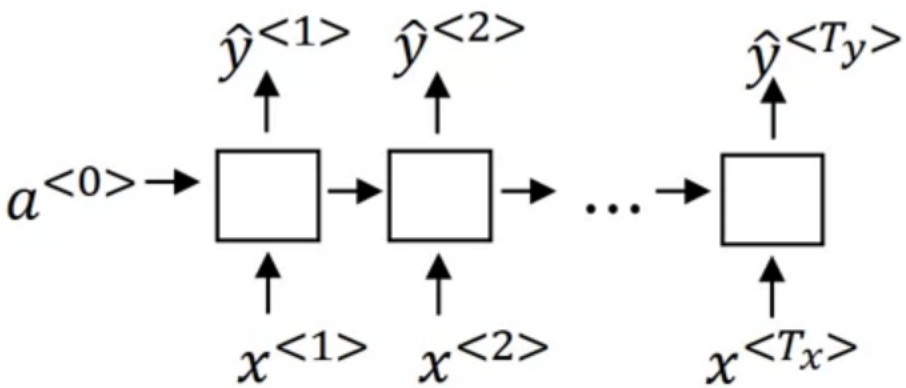
One to one



One to many

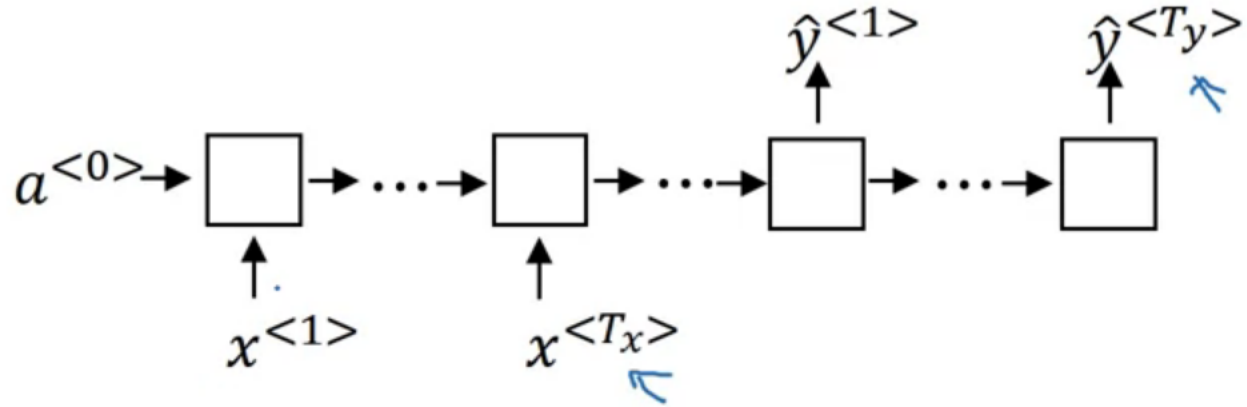


Many to one



Many to many

$T_x = T_y$



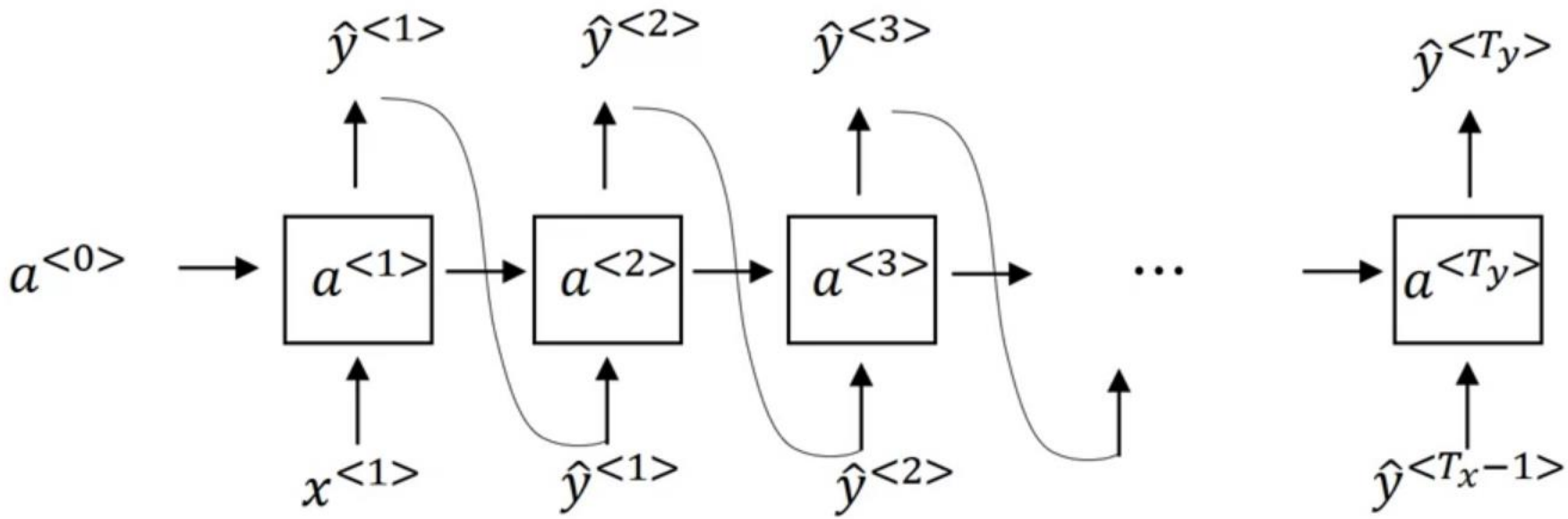
Many to many

Andrew Ng

Character-level language model

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

Vocabulary = [a, b, c, … , z, ␣, ., ,, ;, 0, …, 9, A, …, Z]



Andrew Ng

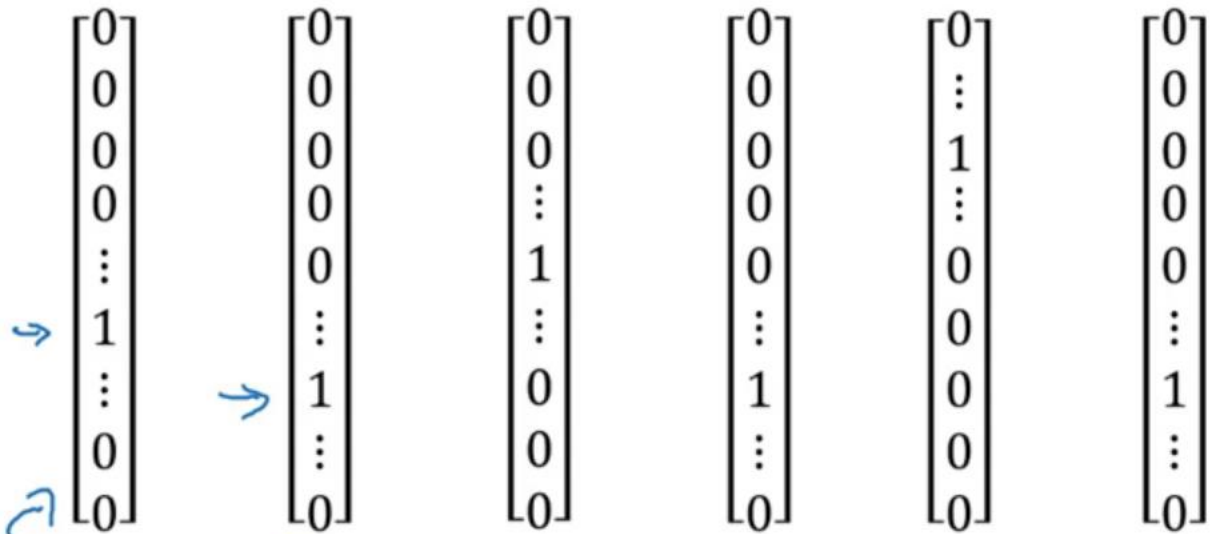
Word representation

V = [a, aaron, …, zulu, <UNK>]

$|V| = 10{,}000$

1-hot representation

Man (5391)	Woman (9853)	King (4914)	Queen (7157)	Apple (456)	Orange (6257)
O_{5391}	O_{9853}				



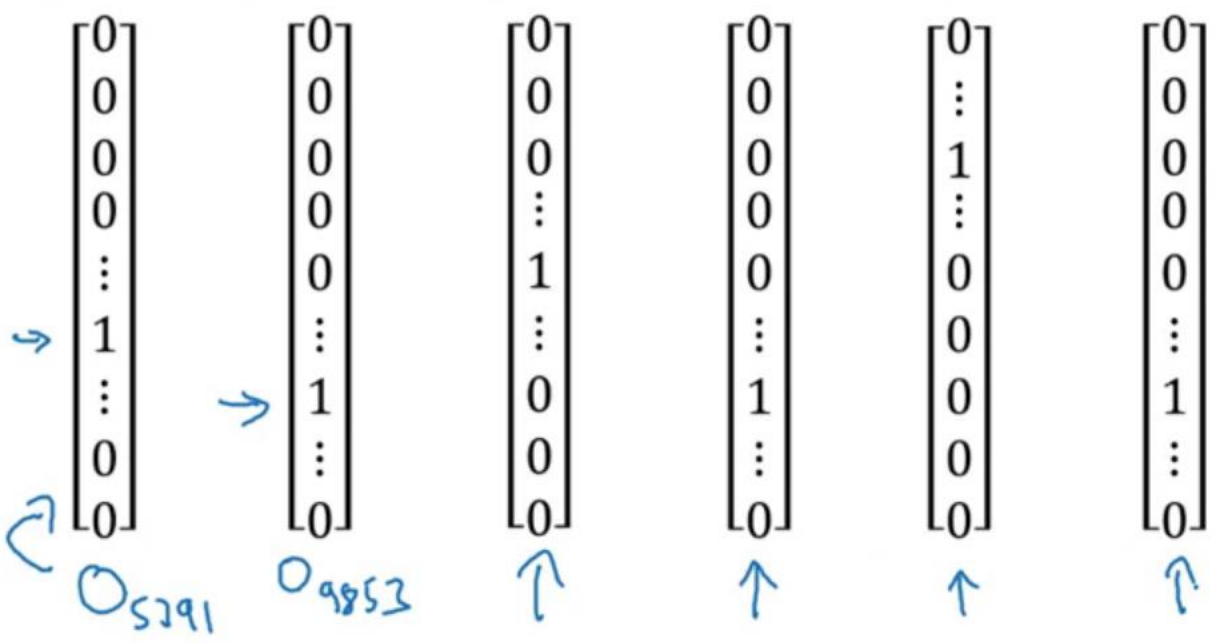
Word representation

V = [a, aaron, …, zulu, <UNK>]

|V| = 10,000

1-hot representation

Man	Woman	King	Queen	Apple	Orange
(5391)	(9853)	(4914)	(7157)	(456)	(6257)



I want a glass of orange juice.

I want a glass of apple ?.

Andrew Ng

Featurized representation: word embedding

	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	0.7	0.69	0.03	-0.02
Food	0.04	0.01	0.02	0.01	0.95	0.97
⋮						
size						
cost						
alive						
verb						

300

Andrew Ng

Featurized representation: word embedding

	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	0.7	0.69	0.03	-0.02
Food	0.04	0.01	0.02	0.01	0.95	0.97
⋮	⋮	⋮				
size						
cost						
alive						
verb						

300

e_{5391} e_{9853}

I want a glass of orange _____.

I want a glass of apple_____.

Andrew Ng

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

e_{5391}
e_{man}

e_{woman}

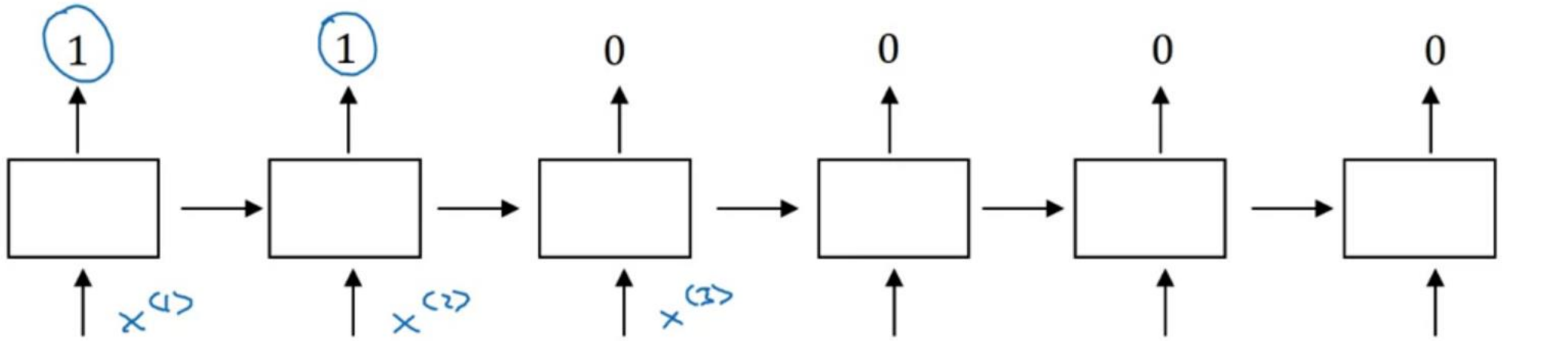
$e_{man} - e_{woman} \approx \begin{bmatrix} -2 \\ 0 \\ 0 \\ 0 \end{bmatrix}$

Man → Woman as King → ? Queen

$e_{king} - e_{queen} \approx \begin{bmatrix} -2 \\ 0 \\ 0 \\ 0 \end{bmatrix}$

[Mikolov et. al., 2013, Linguistic regularities in continuous space word representations]

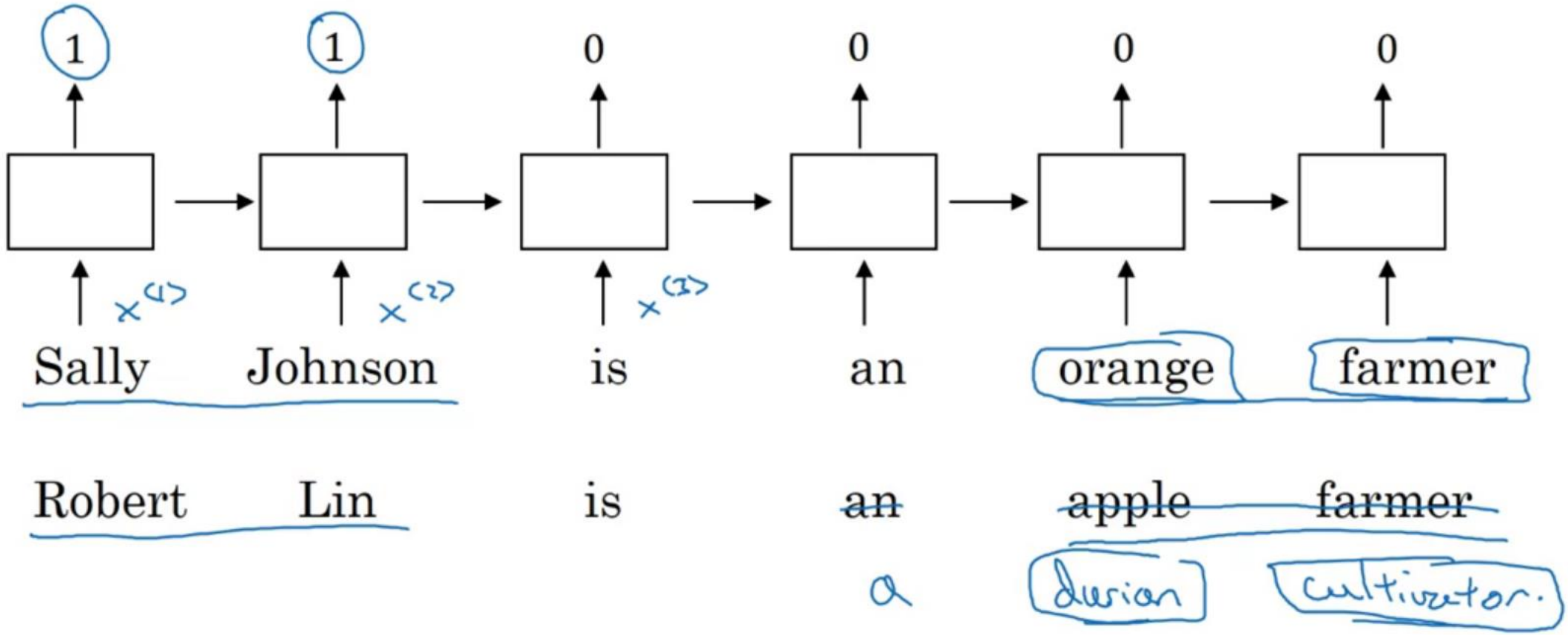
Named entity recognition example



Robert Lin is an apple farmer

Andrew Ng

Named entity recognition example



Andrew Ng

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) → 10,000 → 300

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

Andrew Ng

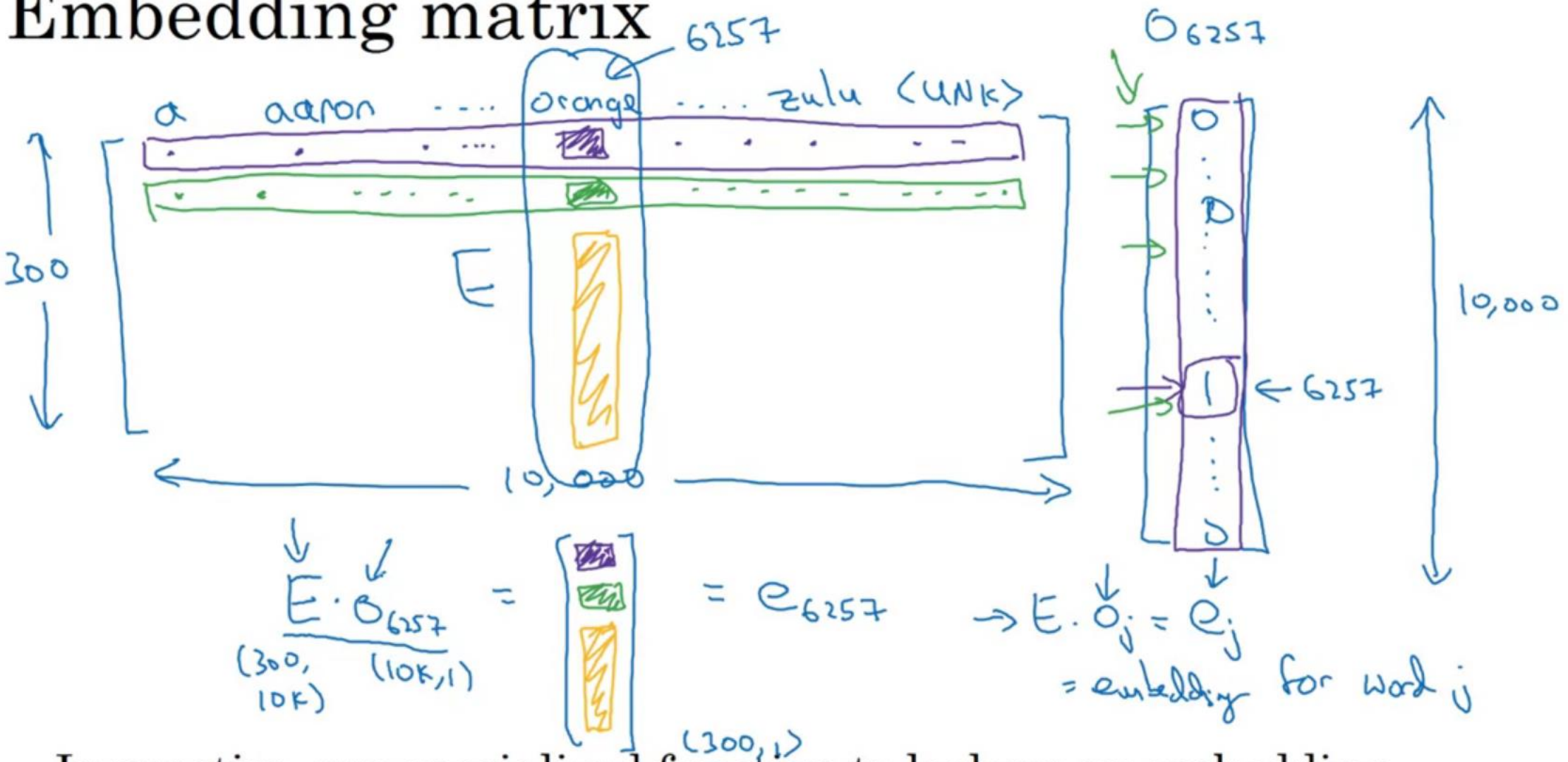
Analogies

	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

Man → Woman as King → ?

[Mikolov et. al., 2013, Linguistic regularities in continuous space word representations]

Embedding matrix



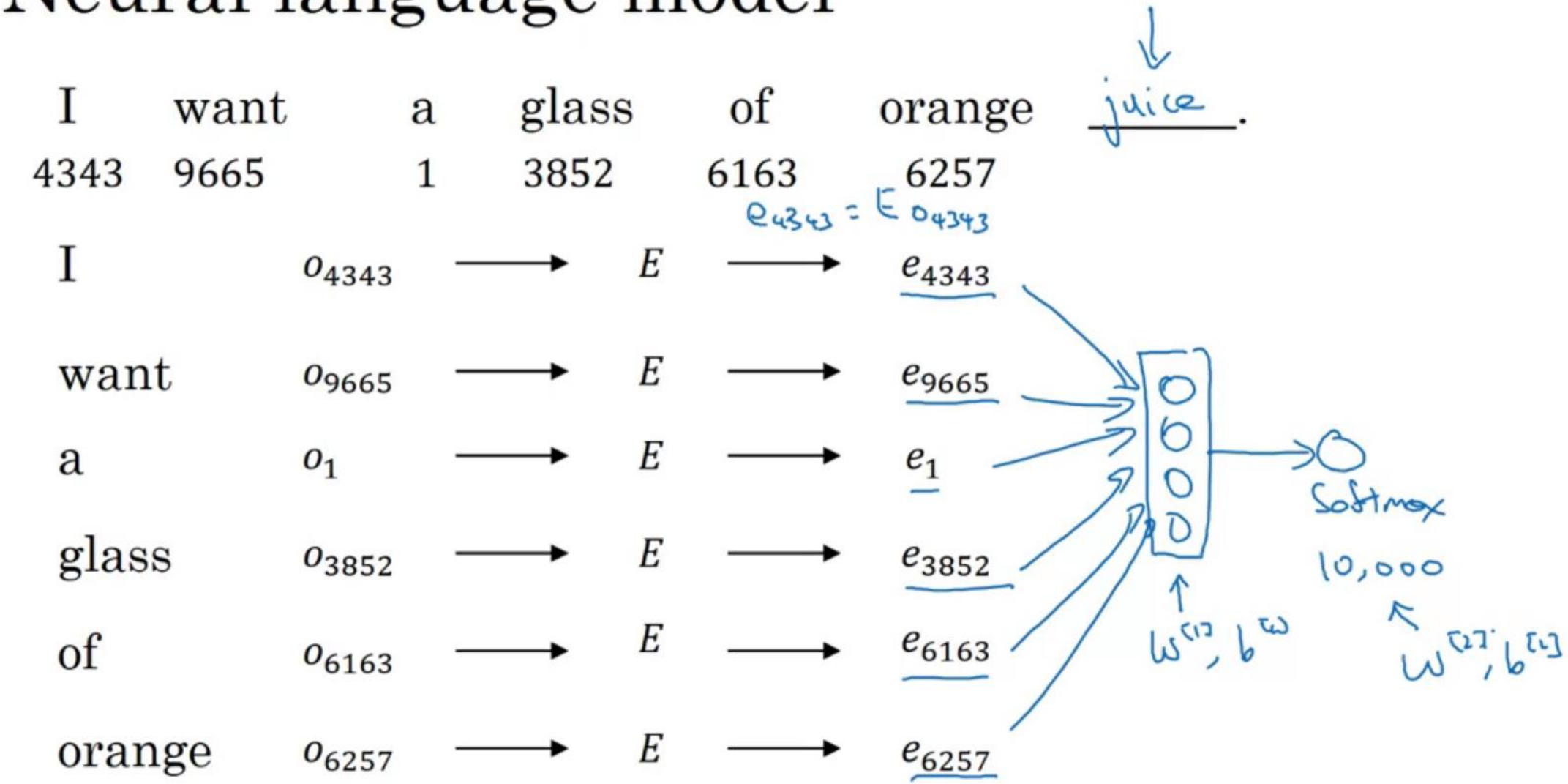
In practice, use specialized function to look up an embedding.

Andrew Ng

Neural language model

I want a glass of orange juice.

4343 9665 1 3852 6163 6257

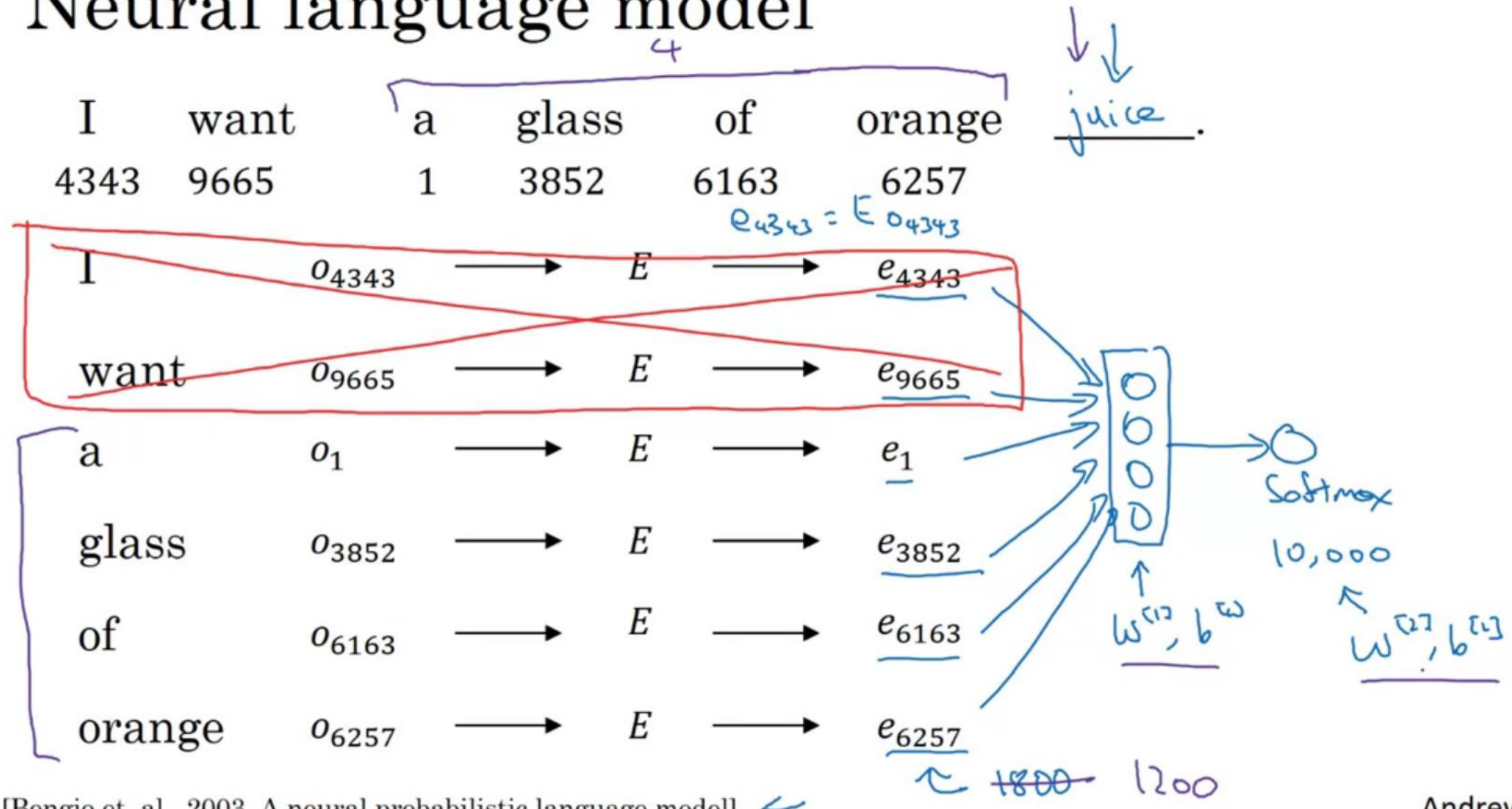


[Bengio et. al., 2003, A neural probabilistic language model]

Neural language model

I want a glass of orange juice.

4343 9665 1 3852 6163 6257

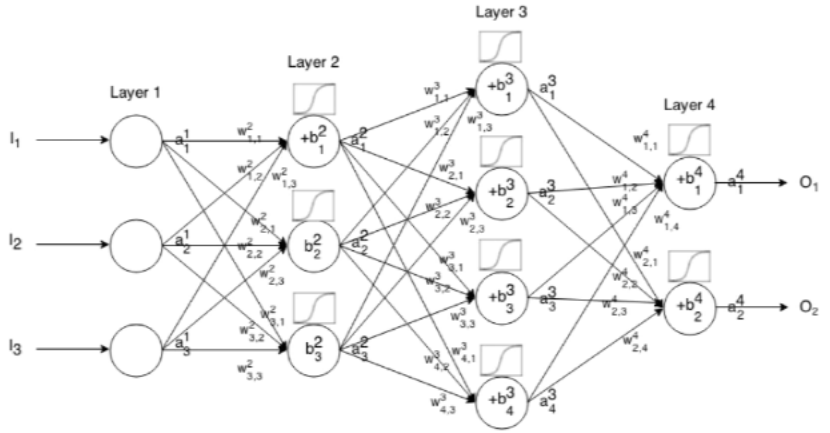


[Bengio et. al., 2003, A neural probabilistic language model]

Andrew Ng

Why do we need embeddings?

- What is NLP? Machine learning applied to text / speech
- Text is represented as a string inside the computer
- Deep Learning expects numbers as input (first operation is multiplication)
- Text is not a number → problem!
- What happens when you multiply a string by a number?

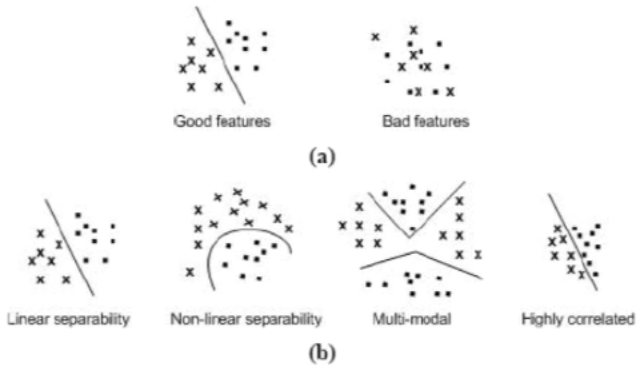


Why do we need embeddings?

- One solution is just to one-hot encode each word
- If we have only 3 words “apple”, “banana”, “carrot”, we can use:
  - Apple = [1, 0, 0]
  - Banana = [0, 1, 0]
  - Carrot = [0, 0, 1]
- Quiz: Why is this not a good representation for words?

What is a feature vector?

- Ex. machine learning problem: I take a survey of my students, I would like to predict whether or not they will succeed in this course
- I ask some questions:
  - “Do you meet the prerequisites of this course?”
  - “Are you an independent learner and are you able to do independent research if you come across a topic you don’t know?”
- Each student will answer the question based on a scale from 1...10



What is a feature vector?

- Each row is a feature vector

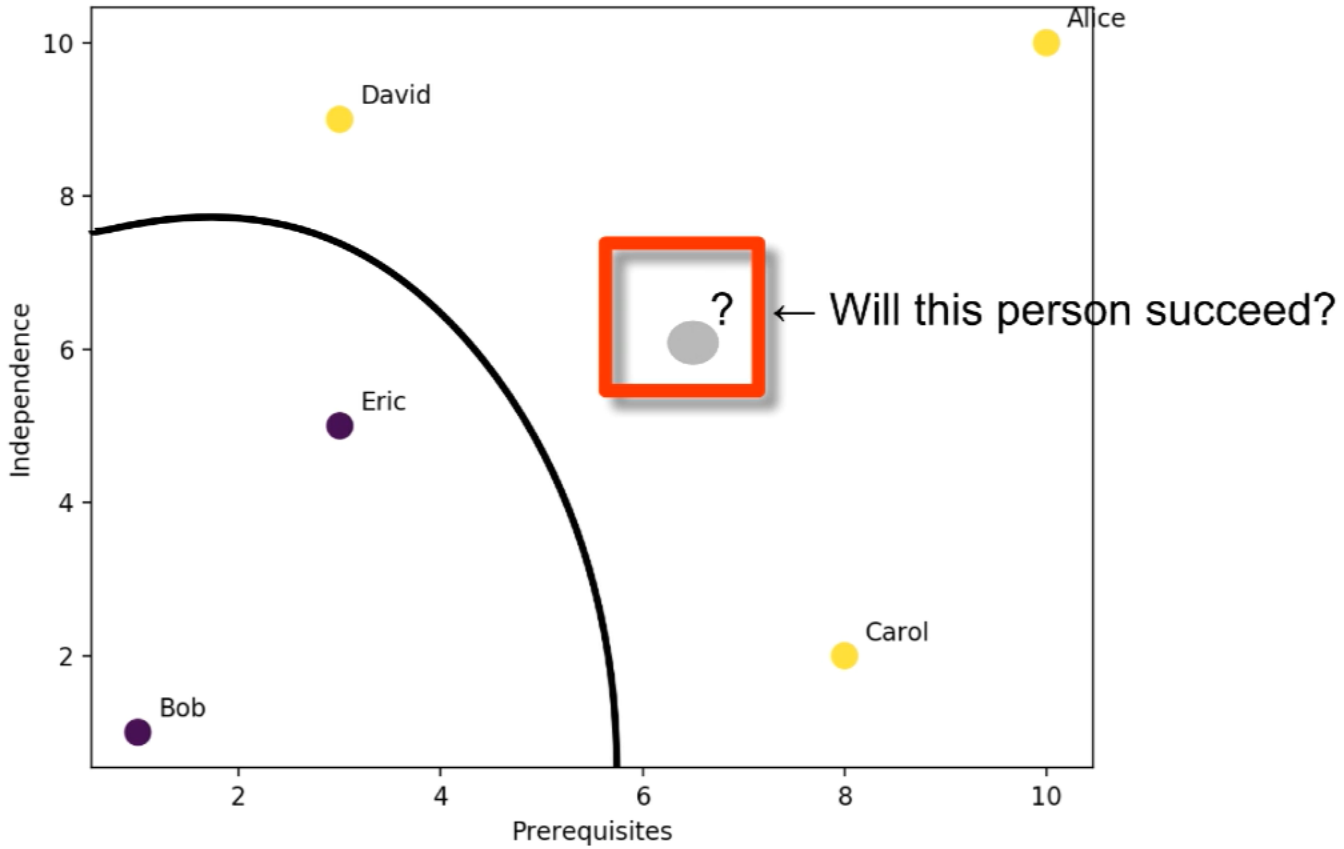
	Meets prerequisites	Independent learner
Alice	10	10
Bob	1	1
Carol	8	2
David	3	9
Eric	3	5

This is a vector

This is also a vector

	Student succeeded
Alice	Yes
Bob	No
Carol	Yes
David	Yes
Eric	No

How do we use the data?



Now an NLP example

These are feature vectors corresponding to words, so we'll call them: "Word vectors"

	Gender	Age	Place
King	+1	+1	0
Queen	-1	+1	0
Prince	+0.9	-1	0
Princess	-0.8	-0.9	0.0000001
Britain	0	0	0.9
Spain	0	0	1

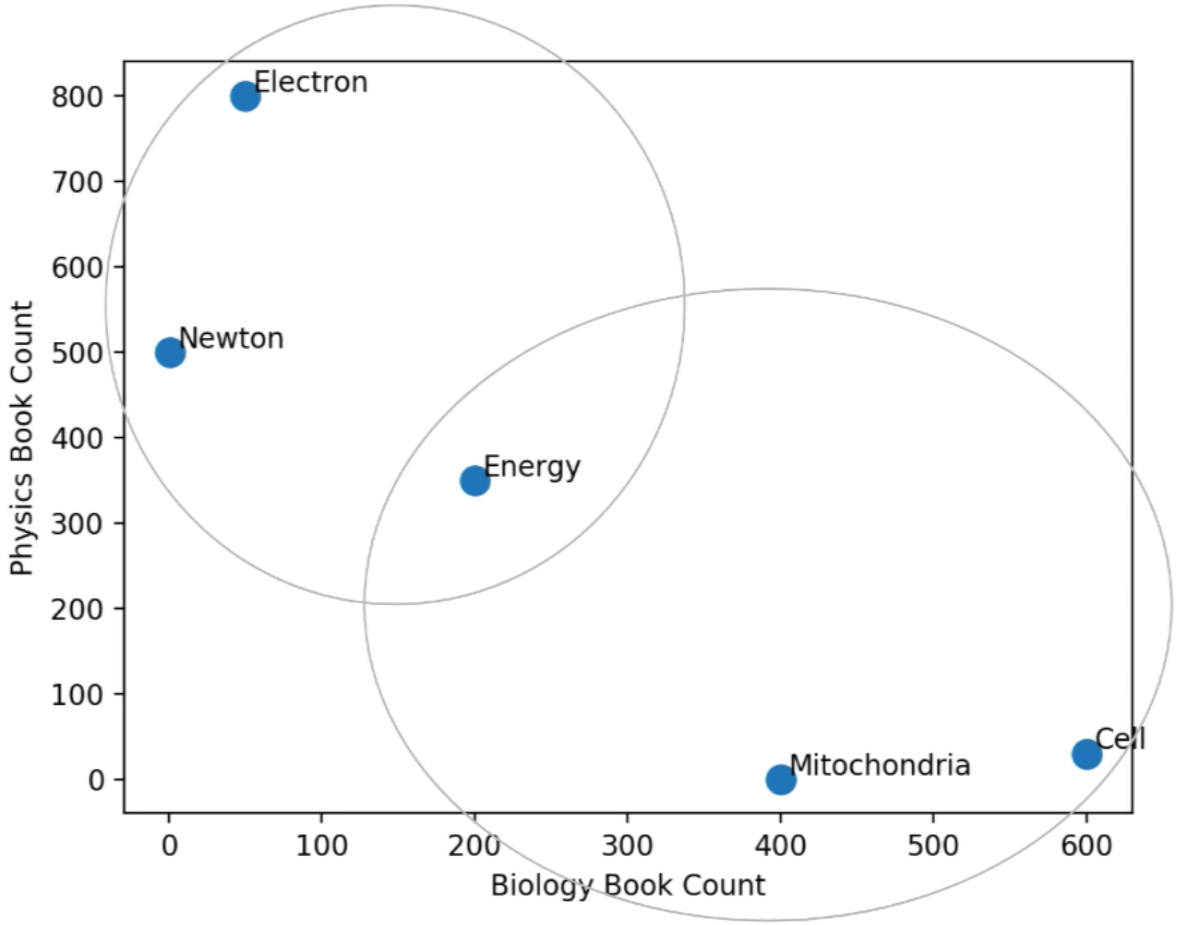
Automation

- In real life, I can't ask each word what "age" it feels like - in fact, I can't ask a word anything (plus, there are way too many words in the dictionary to survey individually)
- Thus, we need some automated way to create word vectors
- The simplest way is by counting

How to count

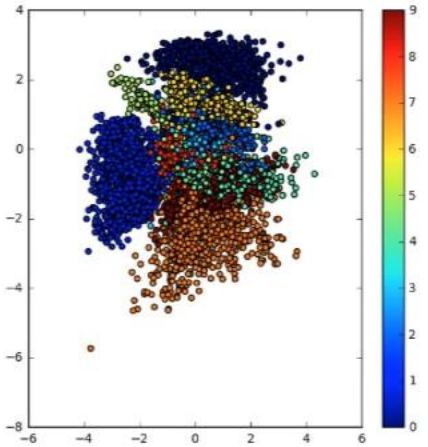
	Favorite biology book	Favorite physics book
Electron	50	800
Newton	1	500
Energy	200	350
Mitochondria	400	0
Cell	600	30

How to count


Electron
Newton
Energy
Mitochondria
Cell
Physics Book Count
Biology Book Count
0
100
200
300
400
500
600
700
800

In Modern Times...

- We have more interesting ways to find word vectors
- Since they use unsupervised learning, the features don't necessarily have to make sense to us humans
- They only have to make sense geometrically
- We can think of them as "latent vectors" or "hidden vectors" (like the inner layers of a neural network)
- 3 popular methods: Word2Vec, GLoVe, FastText



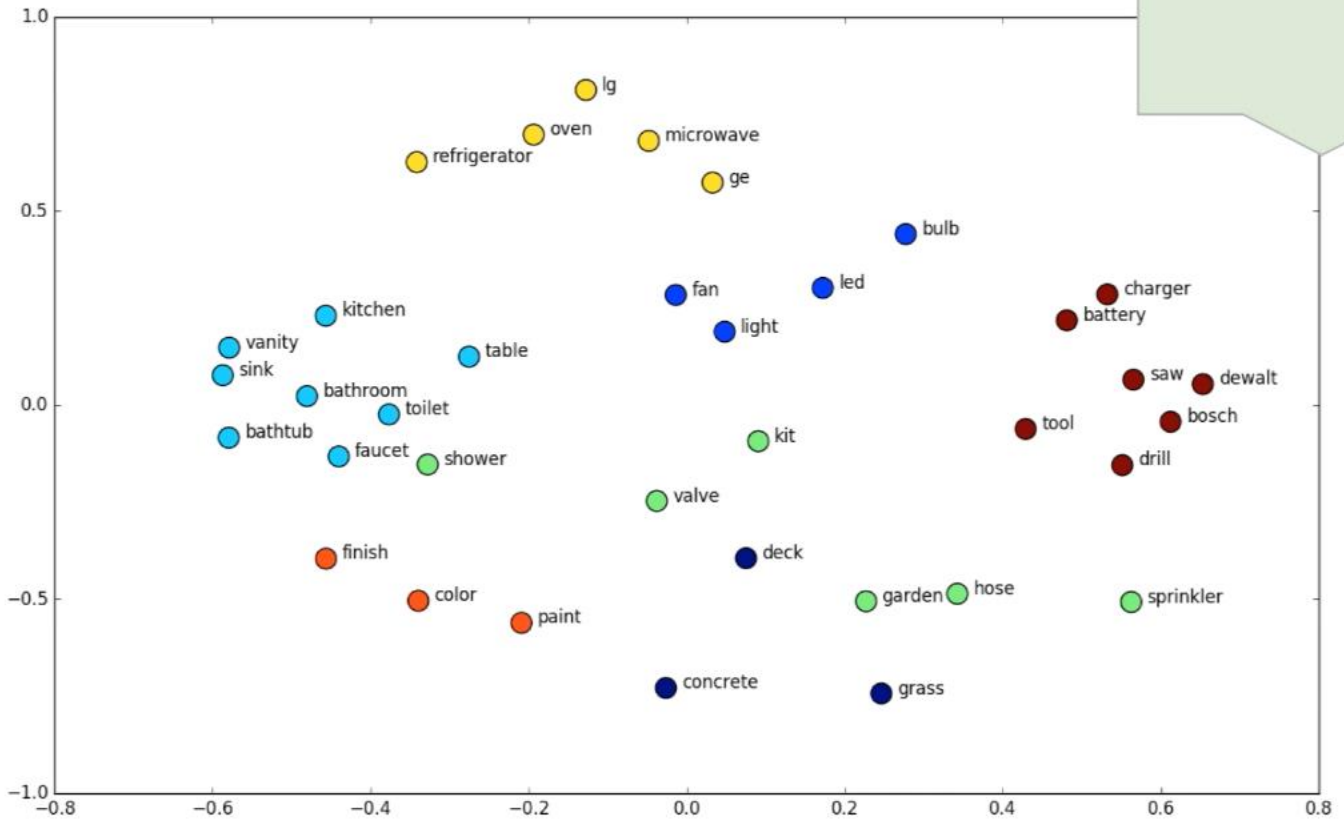
Word Analogies

- Modern algorithms are very good at relationships between words
- E.g.

Vec(“king”) - Vec(“man”) ~= Vec(“queen”) - Vec(“woman”)

Word Similarity

Nearby words are similar



What does this have to do with embeddings?

- In Deep Learning, we work with matrices, e.g. every dense layer is:
  f(input.dot(weights) + bias)
- A word embedding is simply a matrix of stacked word vectors

	Favorite biology book	Favorite physics book
Electron	50	800
Newton	1	500
Energy	200	350
Mitochondria	400	0
Cell	600	30

Single word vector

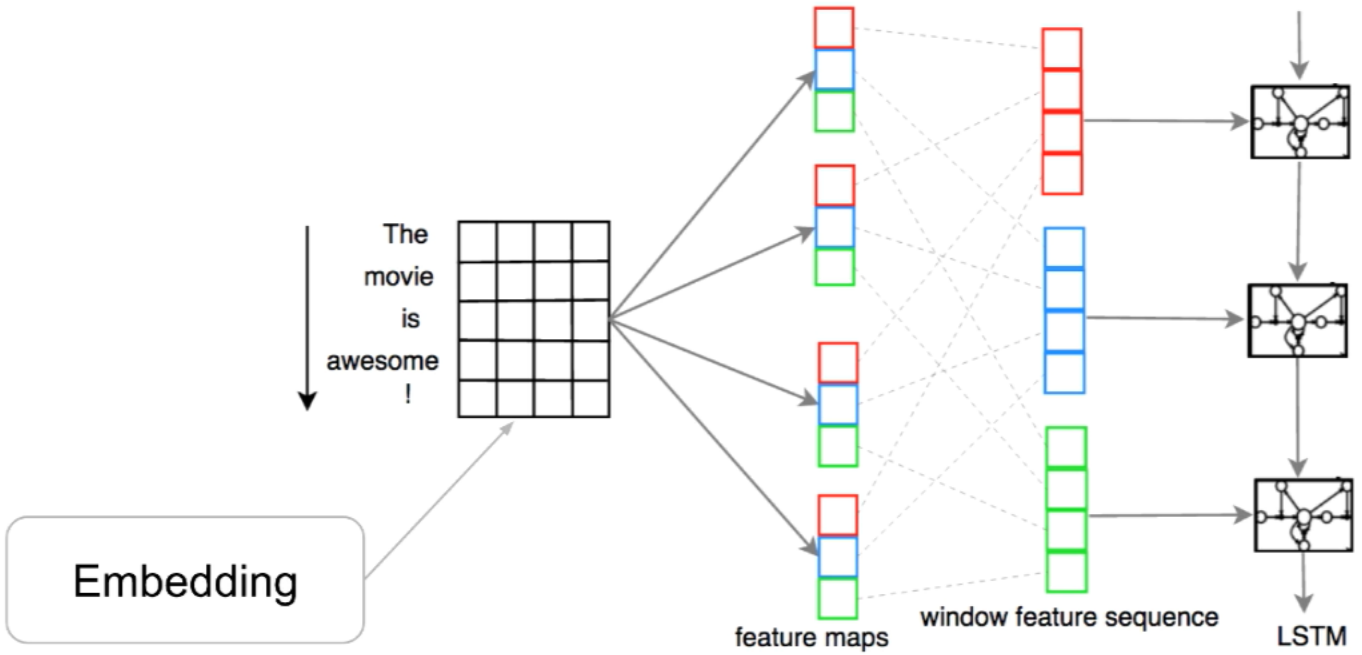
Word embedding

Conventions

- V = # of rows = vocabulary size = # of distinct words
- D = embedding dimension = feature vector size

Computational Trick

- Note: The first layer of a neural network with one-hot inputs is always an embedding



Computational Trick

- What's an easier way to get a row of a matrix?
- To get the kth row of a matrix W → W[k]
- You should never multiply a one-hot vector by a matrix → inefficient
- In Theano: W[k]
- In Tensorflow: tf.nn.embedding_lookup
- In Keras: Embedding() (never use Dense() for an embedding)

Using Word Embeddings

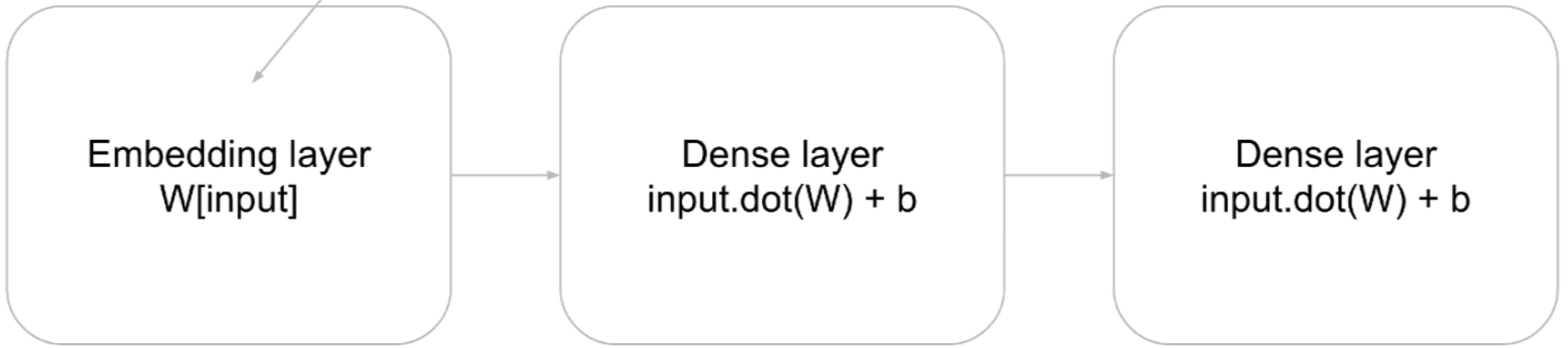
- In previous Deep NLP, we looked at how to train Word2Vec / GLoVe (you don't need to know that for this course)
  - Please, don't be intimidated by this! I repeat: you will do just fine in this course not knowing how Word2Vec / GLoVe vectors are trained
- We will do something like transfer learning, where we set the embedding layer with pre-trained weights (making use of good hyperparameters found by others!)
  - No such thing as a "formula" for calculating the best hyperparameters

Using Word Embeddings

Pretrained Embedding

“a”	0.5	-0.3	...
“an”	-0.1	1.5	...
...	1.2	0.75	...

```
W = preloaded from CSV
```
instead of:
```
W = np.random.randn(V, D)*0.01
```



Train pre-trained embeddings?

- model.fit(X, Y) automatically updates all parameters in the model
- Should you let it train the embedding layer or not?
- It might be necessary to train embeddings that were initialized randomly (not found in pre-trained embedding)
  - But most libraries don't have such fine-grained control (updating only a subset of a matrix)
  - Thus, you either train the entire embedding, or none of it
- Typically we don't bother to fine-tune