DEEP LEARNING

David Kauchak
CS158 – Fall 2016

Admin

Assignment 7

Assignment 8

No office hours Thursday

Wednesday office hours extended 2-5pm

Deep learning

WIKIPEDIA

Deep learning is a branch of machine learning based on a set of algorithms that attempt to model high level abstractions in data by using a deep graph with multiple processing layers, composed of multiple linear and non-linear transformations.

Deep learning is part of a broader family of machine learning methods based on learning representations of data.

Deep learning

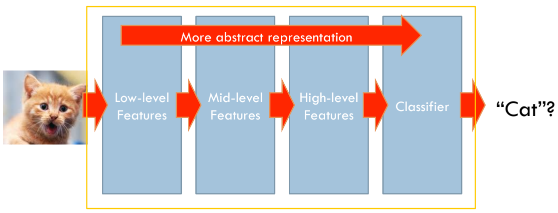
Key: learning better features that abstract from the "raw" data

Using learned feature representations based on large amounts of data, generally unsupervised

Using classifiers with multiple layers of learning

Deep learning

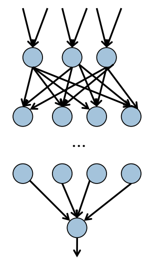
- Train *multiple layers* of features/abstractions from data.
- Try to discover *representation* that makes decisions easy.



Deep Learning: train layers of features so that classifier works well.

Slide adapted from: Adam Coates

Deep learning for neural networks



Traditional NN models: 1-2 hidden layers

Deep learning NN models: 3+ hidden layers

Geoffrey Hinton

I now work part-time for Google as an Engineering Fellow and part-time for the University of Toronto as an Emeritus Distinguished Professor. For much of the year, I work at the University in the morning and at the Google Toronto office at 111 Richmond Street from 2.00pm to 6.00pm. I also spend several months per year working full-time for Google in Mountain View, California.

http://www.cs.toronto.edu/~hinton/

Geoffrey Hinton

Geoffrey Everest Hinton FRS[6] (born 6 December 1947) is a British-born cognitive psychologist and computer scientist, most noted for his work on artificial neural networks. As of 2015 he divides his time working for Google and University of Toronto.[7] He was one of the first researchers who demonstrated the use of generalized backpropagation algorithm for training multi-layer neural nets and is an important figure in the deep learning community.[8][9][10]

Hinton is the great-great-grandson both of logician George Boole whose work eventually became one of the foundations of modern computer science, and of surgeon and author James Hinton.[22] His father is Howard Hinton.[23]

https://en.wikipedia.org/wiki/Geoffrey_Hinton

Importance of features: Hinton

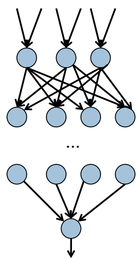
Once you have the right features, the algorithm you pick is relatively unimportant

Normal process = hand-crafted features

Deep learning: find algorithms to automatically discover features from the data

Challenges

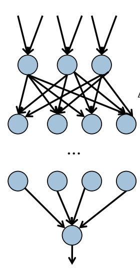
What makes "deep learning" hard for NNs?



$$w = w + input * \Delta_{current}$$
$$\Delta_{current} = f'(current\_input)\sum w_{output}\Delta_{output}$$

Challenges

What makes "deep learning" hard for NNs?



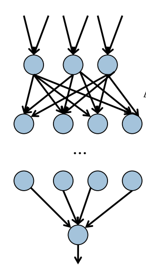
$$w = w + input * \Delta_{current}$$
$$\Delta_{current} = f'(current\_input)\sum w_{output}\Delta_{output}$$
$$w = w + input * \Delta_{current}$$
$$\Delta_{current} = f'(current\_input)\sum w_{output}\Delta_{output}$$

…

What will happened to the modified errors further up?

Challenges

What makes "deep learning" hard for NNs?



$$w = w + input * \Delta_{current}$$
$$\Delta_{current} = f'(current\_input)\sum w_{output}\Delta_{output}$$
$$w = w + input * \Delta_{current}$$
$$\Delta_{current} = f'(current\_input)\sum w_{output}\Delta_{output}$$

…

Modified errors tend to get diluted as they get combined with many layers of weight corrections

Deep learning

Growing field

Driven by:
- Increase in data availability
- Increase in computational power
- Parallelizability of many of the algorithms

Involves more than just neural networks (though, they're a very popular model)

word2vec

How many people have heard of it?

What is it?

Word representations

Wine data uses word occurrences as a feature

What does this miss?

Word representations

Wine data uses word occurrences as a feature

What does this miss?

"The wine had a dark red color"	Zinfandel
"The wine was a deep crimson color"	label?
"The wine was a deep yellow color"	label?

Would like to recognize that words have similar meaning even though they aren't lexically the same

Word representations

Key idea: project words into a multi-dimensional "meaning" space

word ➡ $[x_1, x_2, \ldots, x_d]$

Word representations

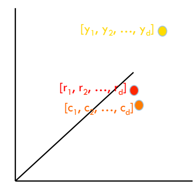
Key idea: project words into a multi-dimensional "meaning" space

word ➡ $[x_1, x_2, \ldots, x_d]$

red ➡ $[r_1, r_2, \ldots, r_d]$

crimson ➡ $[c_1, c_2, \ldots, c_d]$

yellow ➡ $[y_1, y_2, \ldots, y_d]$



Word representations

Key idea: project words into a multi-dimensional "meaning" space

word ➡ $[x_1, x_2, \ldots, x_d]$

The idea of word representations is not new:
- Co-occurrence matrices
- Latent Semantic Analysis (LSA)

New idea: learn word representation using a task-driven approach

A prediction problem

I like to eat bananas with cream cheese

Given a context of words

Predict what words are likely to occur in that context

A prediction problem

Given text, can generate lots of positive examples:

I like to eat bananas with cream cheese

input	prediction
___ like to eat	I
I ___ to eat bananas	like
I like ___ eat bananas with	to
I like to ___ bananas with cream	eat
…	…

A prediction problem

Use data like this to learn a distribution:

$$p(word \mid context)$$

$$p(w_i \mid \underbrace{w_{i-2} w_{i-1}}_{\text{words before}} \underbrace{w_{i+1} w_{i+2}}_{\text{words after}})$$

A prediction problem

Any problems with using only positive examples?

$$p(w_i \mid w_{i-2} w_{i-1} w_{i+1} w_{i+2})$$

input	prediction
___ like to eat	I
I ___ to eat bananas	like
I like ___ eat bananas with	to
I like to ___ bananas with cream	eat
…	…

A prediction problem

Want to learn a distribution over **all** words

$$p(w_i \mid w_{i-2} w_{i-1} w_{i+1} w_{i+2})$$

input	prediction
___ like to eat	I
I ___ to eat bananas	like
I like ___ eat bananas with	to
I like to ___ bananas with cream	eat
…	…

A prediction problem

Negative examples?

I like to eat bananas with cream cheese

input	prediction
___ like to eat	I
I ___ to eat bananas	like
I like ___ eat bananas with	to
I like to ___ bananas with cream	eat
…	…

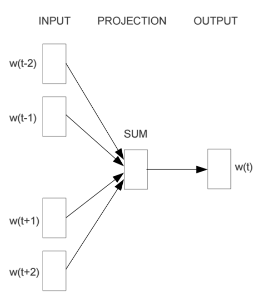
A prediction problem

Use random words to generate negative examples

I like to eat bananas with cream cheese

input	prediction (negative)
___ like to eat	car
I ___ to eat bananas	snoopy
I like ___ eat bananas with	run
I like to ___ bananas with cream	sloth
…	…

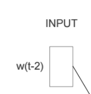
Train a neural network on this problem



https://arxiv.org/pdf/1301.3781v3.pdf

Encoding words

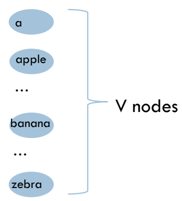
How can we input a "word" into a network?



"One-hot" encoding

For a vocabulary of V words, have V input nodes

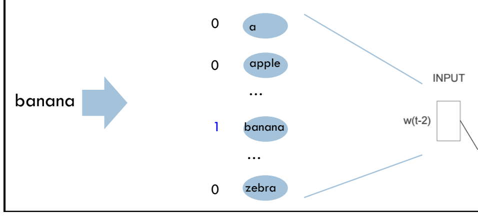
All inputs are 0 except the for the one corresponding to the word



"One-hot" encoding

For a vocabulary of V words, have V input nodes

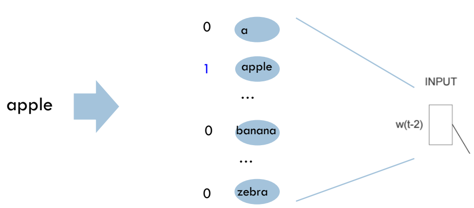
All inputs are 0 except the for the one corresponding to the word

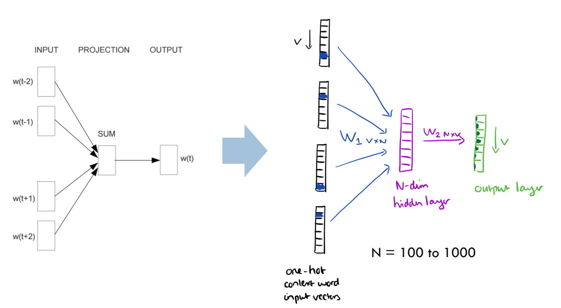


"One-hot" encoding

For a vocabulary of V words, have V input nodes

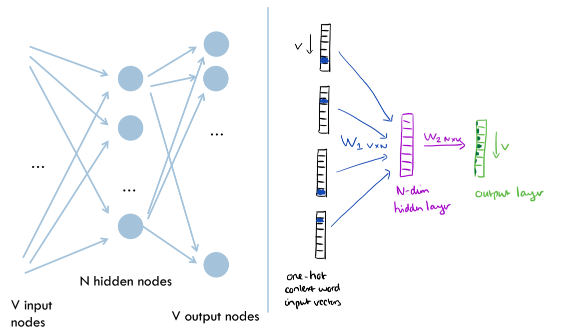
All inputs are 0 except the for the one corresponding to the word



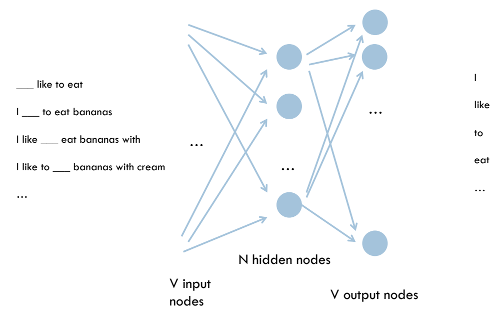


https://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/

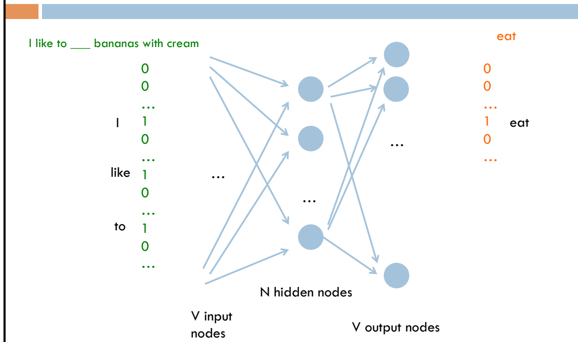
Another view



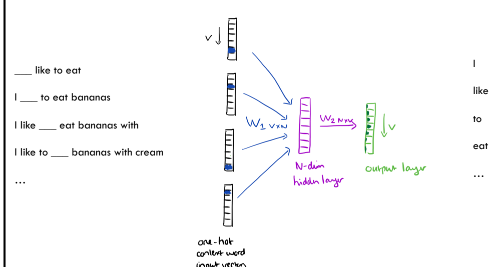
Training: backpropagation



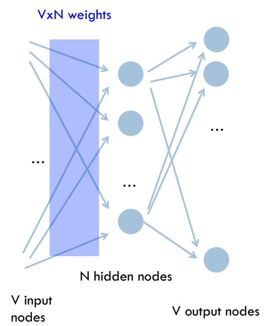
Training: backpropagation



Training: backpropagation



Word representation



The weights for each word provide an N dimensional mapping of the word

Words that predict similarly should have similar weights

Results

vector(word1) – vector(word2) = vector(word3) - X

word1 is to word2 as word3 is to X

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter

Results

vector(word1) – vector(word2) = vector(word3) - X

word1 is to word2 as word3 is to X

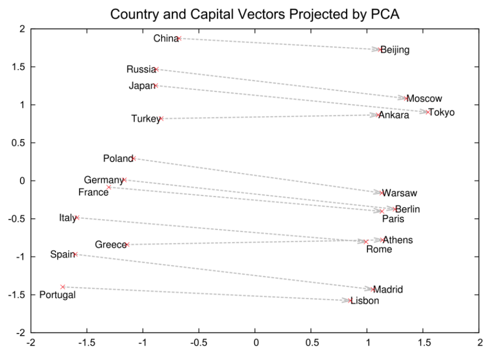
Type of relationship	Word Pair 1		Word Pair 2	
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks

Results

vector(word1) – vector(word2) = vector(word3) - X

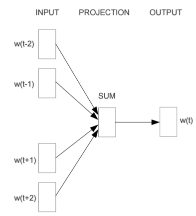
word1 is to word2 as word3 is to X

Newspapers			
New York San Jose	New York Times San Jose Mercury News	Baltimore Cincinnati	Baltimore Sun Cincinnati Enquirer
NHL Teams			
Boston Phoenix	Boston Bruins Phoenix Coyotes	Montreal Nashville	Montreal Canadiens Nashville Predators
NBA Teams			
Detroit Oakland	Detroit Pistons Golden State Warriors	Toronto Memphis	Toronto Raptors Memphis Grizzlies
Airlines			
Austria Belgium	Austrian Airlines Brussels Airlines	Spain Greece	Spainair Aegean Airlines
Company executives			
Steve Ballmer Samuel J. Palmisano	Microsoft IBM	Larry Page Werner Vogels	Google Amazon



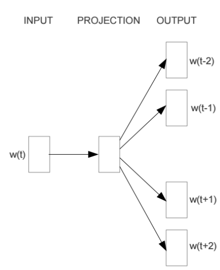
2-Dimensional projection of the N-dimensional space

Continuous Bag Of Words



CBOW

Other models: skip-gram



word2vec

A model for learning word representations from large amounts of data

Has become a popular pre-processing step for learning a more robust feature representation

Models like word2vec have also been incorporated into other learning approaches (e.g. translation tasks)

word2vec resources

- https://blog.acolyer.org/2016/04/21/the-amazing-power-of-word-vectors/
- https://code.google.com/archive/p/word2vec/
- https://deeplearning4j.org/word2vec
- https://arxiv.org/pdf/1301.3781v3.pdf

Big Data

What is "big data"?

What are some sources of big data?

What are the challenges of dealing with big data?

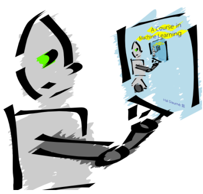
What are some of the tools you've heard of?

Big data and ML

Why talk about it in a course like this?

Machine Learning is…

Machine learning is about predicting the future based on the past.
-- Hal Daume III



Machine Learning is…

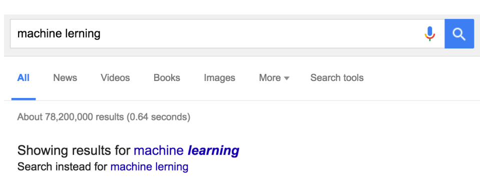
Machine learning is about predicting the future based on the past.
-- Hal Daume III

If the "past" has lots of data, then we need tools to process it!

Big data and ML

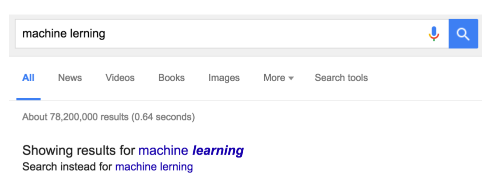
Why talk about it in a course like this?

Many "machine learning" problems become much easier when you have lots of data



Big data and ML

How would you do it?



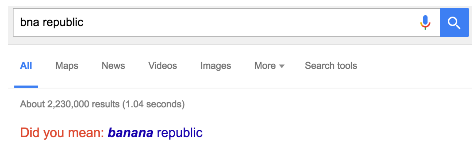
Big data and ML

How would you do it?

edit distance

Big data and ML

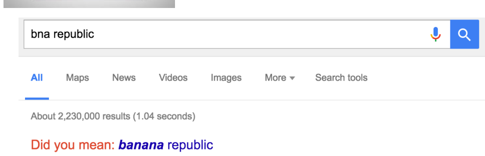
How would you do it?



May not get example like this!

Big data and ML

How would they do it?
(small company)



May not get example like this correct!

Big data and ML

How would they do it?
(small company)

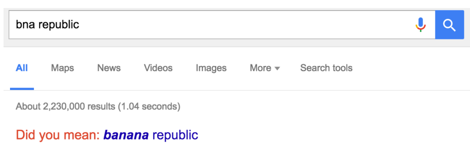
machine learning

model/
predictor

text corpus

Big data and ML

How does Google do it?



May not get example like this!

Big data and ML

Google now handles at least 2 trillion searches per year

The search giant won't say exactly how many trillions of queries it processes, other than it's now two or more. It last claimed 1.2 trillion in 2012.

http://searchengineland.com/google-now-handles-2-999-trillion-searches-per-year-250247

Big data and ML

Search logs

user_id	time	query
	…	
131524	t	bna republic
	…	
131524	t+5s	banana republic
	…	

Many problems get easy when you have lots of data!

Big data and ML

Many **problems** get easy when you have lots of data!

Challenge: processing all this data in an efficient way