Frontiers in NLP I

From BoW to Word2Vec

Agenda

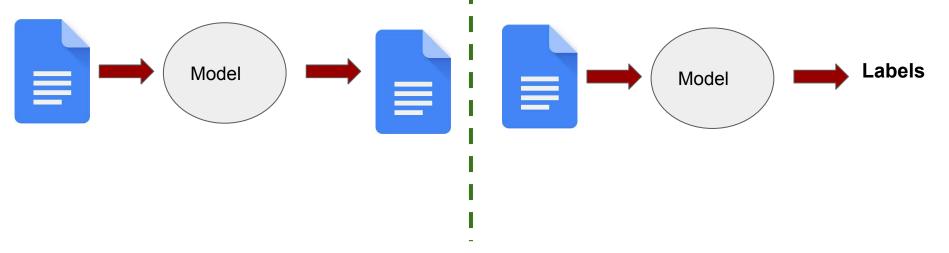
- 1. NLP and Its Tasks
- 2. Representation Learning
- 3. Word Embeddings
- 4. Neural Networks for NLP
 - a. CNN
 - b. RNN

NLP and Its Tasks

Natural Language Processing

 NLP is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data.

In NLP, there are various NLP tasks such as sentiment analysis, search engine, POS Tagging, translation, chatbot and so one.



Sentiment Analysis

The movie: batman begins was spectacular.

The movie: batman begins was borning in the middle.

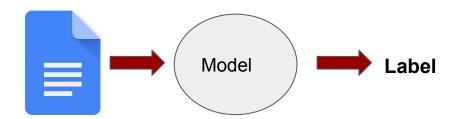


• Even the movie: batman begins was boring in the middle, it was spectacular.



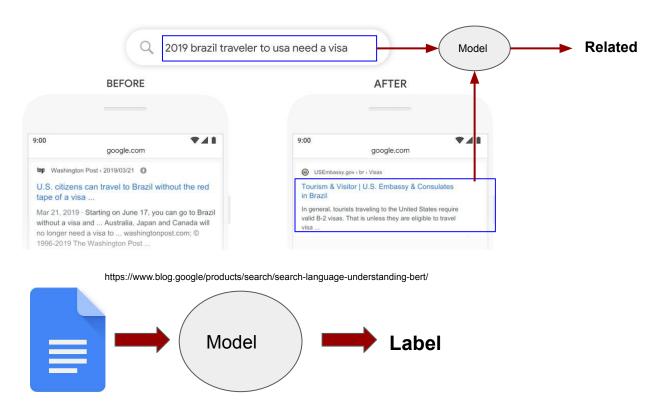
Even the movie: batman begins was spectacular, it was boring in the middle





One Sequence of Words

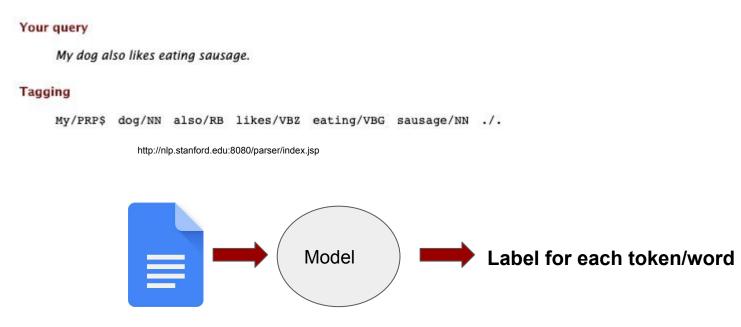
Search Engine



Two Sequences of Words

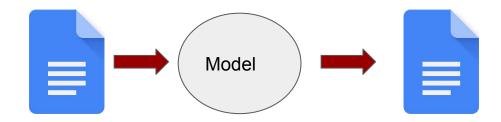
Part-of-Speech (POS) Tagging

Label each word in a sentence with a part-of-speech (e.g. Verb, Adjective, Noun).



One Sequence of Words

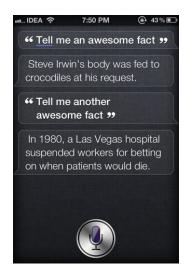
Machine Translation



One Sequences of Words

General Sequences of Words

Dialogue Systems



Chatting

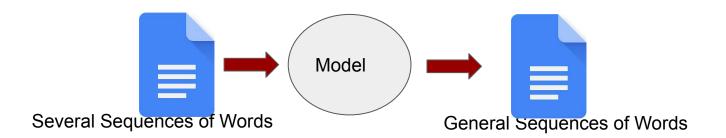
If I don't have a job or certification of account balance

Sorry, I'm not sure about your meaning, please rephrase it or leave a message on our website.

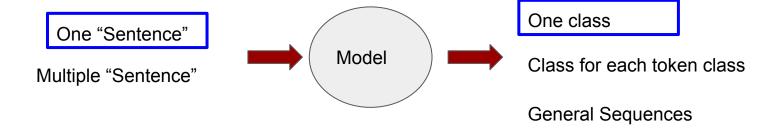
but I do have credit cards from other banks, will this application be approved?

Which kinds of credit card reward do you prefer?

Task-oriented

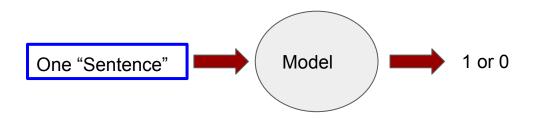


NLP Tasks



In practice, most of NLP tasks are text classification.

Fraudulent Financial Statement



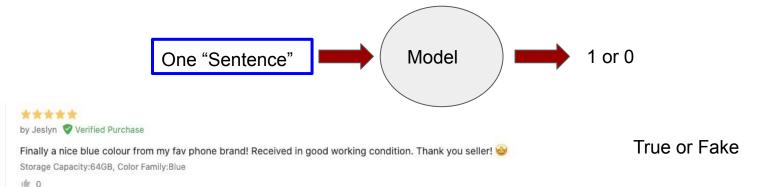
		UNITED STATES		
	SECURI	TIES AND EXCHANGE COM	MISSION	
		Washington, D.C. 20549		
		FORM 10-K		
dark One)				
ANNUAL REPORT	PURSUANT TO SECTION 13 OR 18(6) OF THE SE	CURITIES EXCHANGE ACT OF 1934		
r the fixed year ended December 31,	2004			
TRANSITION REP	ORT PURSUANT TO SECTION 13 OR 15(4) OF TH	_		
r the transition period from to	ORI PURSUENT TO SECTION IN OR INDION IN			
		Commission File No. 000-22533		
		AMAZON.COM, INC.		
		(Exact name of registreast as specified in its chorter)		
	Delaware		90-1646960	
	(force other jurisdiction of hompowing or organization)		(IRS Engineer Identification No.)	
		458 Tarry Armon North Searth, Washington 95589-8200 (200) 264-1000		
	(Million as	belegione number, including man cole, of regionant's principal ex	cutra effica:	
	TO ATTACKS	Securities registered pursuant to Section 12(k) of the Act:	Page of Each Endoare in Whith Resident	
	Common Stock, par value 5.01 per share		NASDEQ Global Select Makes	
		Securities registered pursuant to Section 12(g) of the Act. None		
Indicate by check mark if the registr	set is a well-leaven seasoned soose, as defined in Rule 400 of th	Securities Act. No. 00 No. Cl		
	ant is not required to file reports pursuant to Section 13 or Sectio			
Indicate by check mark robether the en subject to such filing requirements fi		T or 1700 of the Securities Enchange Act of 1904 during the proce	ing 13 months (or fir such shorter period that the registrant to	in required to file such reports), and (2) I
ried that the registrant was required to a	almit and post such files). No. 50 No. Cl	Web site, if any, every lateractive Data File required to be colonist		
Indicate by check mark if disclosure one 13-K or any assessment to this For	of follower files pursued to hers 400 of Explicition S-K is to is 10-K. ID	r contained bensin, and will not be contained, to the best of registrat	(a knowledge, in definitive printy or information statements i	occupanted by reference in Part II of this
Indicate by shock mark whether the st.	registrant is a large accelerated filter, an accelerated filter, a some	celested film, or a smaller reporting company. See definitions of "I	age asselected file;" "asselected file" and "smaller report	ng company" in Rule 126-2 of the Books
args accelerated filer	=		Applicated Size	
no accelerated filer	□ (Do not check if a smaller reporting compa	63	Smaller reporting company	D
Indicate by shock mark relation the	registrant is a shell company (as defined in Rule 126-2 of the Ex	Sange Act). Yes 🗆 No 🖾		
graphs market rules of voting stock is unless of shares of common stock outsi	aid to non-affiliates of the registreet as of 7em 31, 2004 anding as of January 25, 2017		\$	280,128,378,534 477,179,618
		DOCUMENTS INCORPORATED BY REPERENCE		
The information received by East TT	of this Report, to the extent not set forth herein, is incorporated in manission within 129 days other the end of the fiscal year to which	ereis by reduced from the resistant's definitive occur extrement o	dating to the Assessed 2 Desting of Shareholders to be held in 20	07, which definitive poors expressed shall

Fraudulent OR Not?

SEC 10-k

Fake Review Detection

"User" Generated Review



Samsung probed in Taiwan over 'fake web reviews'

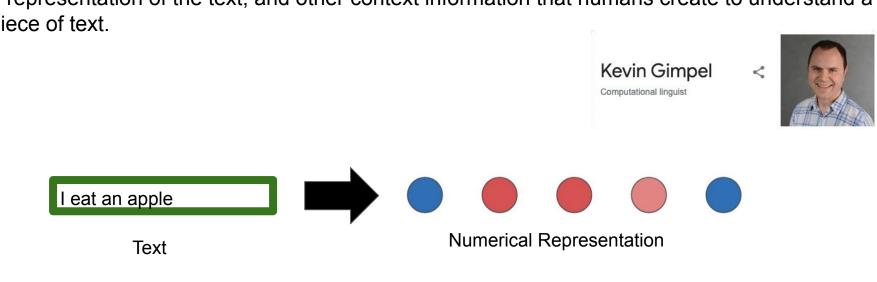
() 16 April 2013

Posting Positive Reviews We post positive and good reviews on-line and in all the biggest search engines. We are the best in the review management business.						
Post Good Reviews	Name					
Posting Positive Reviews	Email *					
Posting Positive Reviews: Prides itself on being an ethical search engine optimization firm, and our team of search engine marketing specialists works hard to stay abreast of the most effective ways to get your site to the top of the search engines. Our search engine marketing firm will never get your website penalized, and, in fact, we have never had a client site penalized - and we've been providing search engine optimization services for over six wers.	Message *					
Post Good Reviews: Corporate reputation repair is a sensitive area. Companies, both large and small, need to develop a strategic approach when dealing with a bad reputation online; whether it's negative search results or being "Imane" by a blogger. Often times, even an honest effort to repair a reputation can cause an unwanted	Send					

Representation Learning

Key Task: Representation Learning

 We need to develop systems that read and understand text the way a person does, by forming a representation of the text, and other context information that humans create to understand a piece of text.



The learned representation should capture

high-level semantic and syntactic information.

History of NLP

- Now, neural nlp models are able to achieve state-of-arts results in all tasks.
- Before neural nlp:
 - Symbolic NLP: rule-based system (derived from linguistic)
 - Statistical NLP: data-driven and use statistical methods



1950 - early 1990s 1990s - 2010s Present Future

Statistical NLP

- Starting from Document-Term Matrix
 - It contains the co-occurrence information
 - Bag-of-Words: n-gram as features
 - TF-IDF: frequency of words to measure importance
 - Matrix Decomposition:
 - SVD -> Latent Semantic Analysis
 - Probabilistic model-> Topic Model

D0: I eat an apple every day D1: I eat an orange every day Corpus D2: I like driving my car to work **Document-Term Matrix** driving eat every like my orange to work 0 1





Bag-of-Words

TF-IDF

Latent Semantic Topic Models Analysis

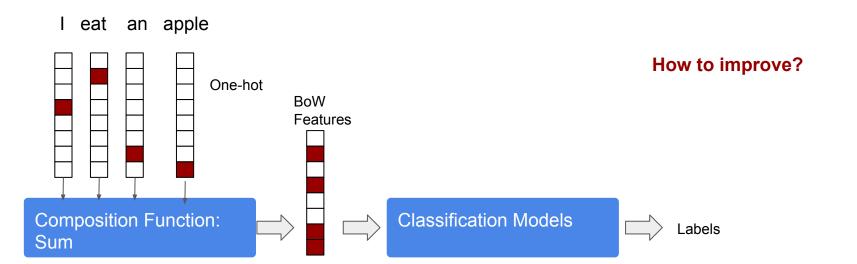
Limitations of Document-Term Matrix

- Too strong assumption: all words are independent of each other
 | orange peach | < | orange car |
- Can not capture the order information in the sequence
- High dimensionality due to large size of vocabulary

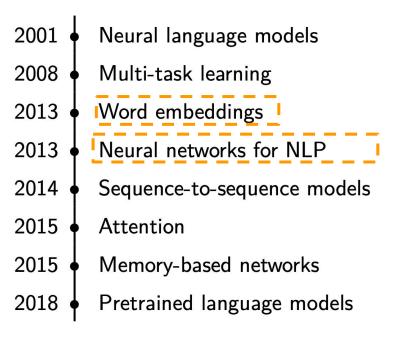
	an	apple	car	day	driving	eat	every	like	my	orange	to	work	
0	1	1	0	1	0	1	1	0	0	0	0	0	
1	1	0	0	1	0	1	1	0	0	1	0	0	
2	0	0	1	0	1	0	0	1	1	0	1	1	

A New Perspective on BoW

- Each word in vocab is represented in one-hot embedding
- Sum one-hot vectors of the words in a sentence
- The final vector is the representation for the given sentence and then fed into a classifier.



Neural NLP

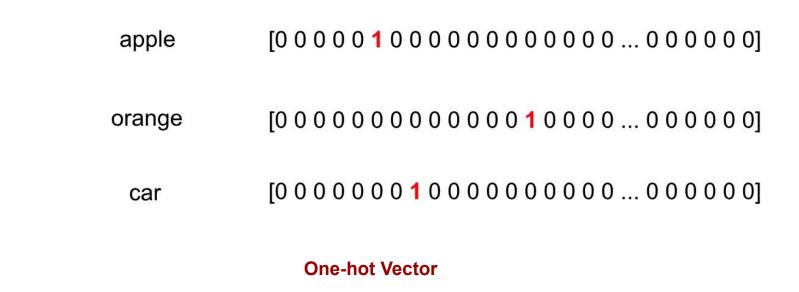


https://www.kamperh.com/slides/ruder+kamper_indaba2018_talk.pdf

Word Embeddings

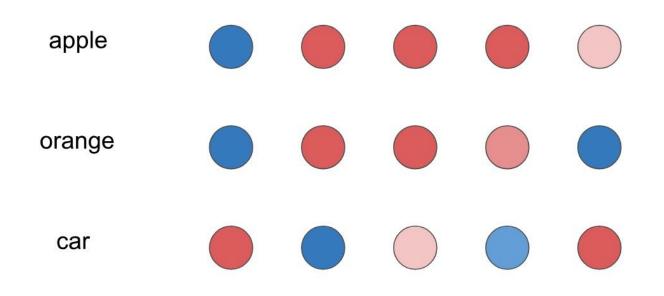
Word Representation

How to represent word in a vector space



Distributed Representation

Words should be encoded into a low-dimensional and dense vector



Word Vectors

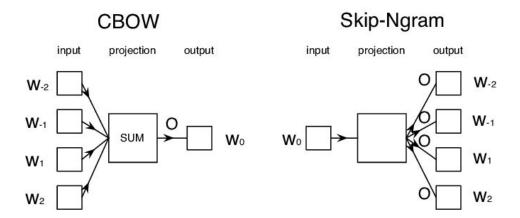
Project word vectors in a two-dimensional space. And visualize them!

juice milk bus car train

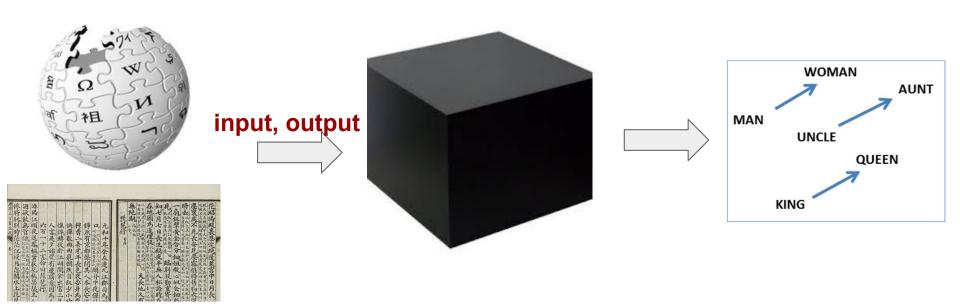
Similar words are close to each other.

Word2Vec

- · A method of computing vector representation of words developed by Google.
- Open-source version of Word2Vec hosted by Google (in C)
- Train a simple neural network with a single hidden layer to perform word prediction tasks
- Two structures proposed Continuous Bag of Words (cbow) vs skip-gram:



Word2Vec as BlackBox



Corpus Word2Vec Tool Word Embeddings

A Good Visualization for Word2Vec

https://ronxin.github.io/wevi/

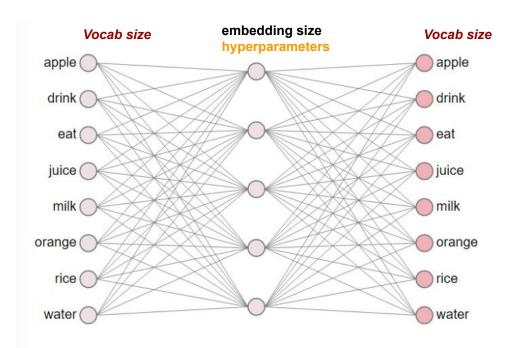
Target

- Given a training corpus, we prepare a list of N (input_word, output_word).
- Objective Function: Maximize probability of all the output words given the corresponding input words.

$$\mathbf{J}(heta) = \prod_{i=1}^N p(w^i_{output}|w^i_{input}, heta)$$
 Neural network parameters that will

be optimized

Model Architecture



Structure Highlights:

- input layer
 - one-hot vector
- hidden layer
 - linear (identity)
- output layer
 - softmax

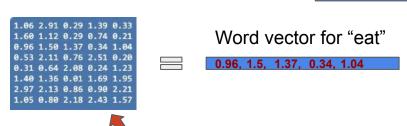
Hidden Layer

One-hot vector

Index of eat

[0,0,1,0,0,0,0,0]

- linear-activation function here
- 5 neurons are the word vec. dimensions
- This layer is operating as a 'lookup' table
- Input word matrix denoted as IVec



Hidden Layer Weights Matrix 5 neurons inputs

Word Vector Look Up

5 features

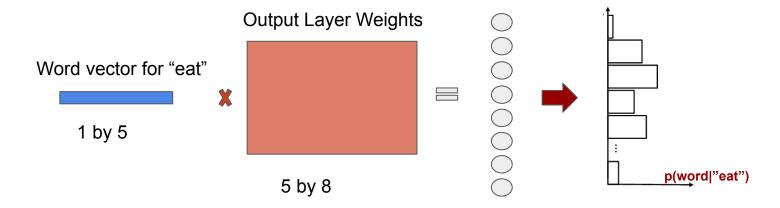
Table

words

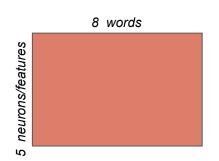
This is a **projection/look up** process: given the index of the word, we take the ith row in the word vector matrix out

Output Layer

- Softmax classifier
- Output word matrix denoted as OVec



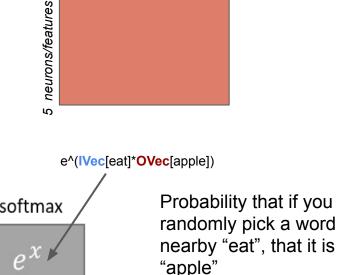
Output Layer Weights Matrix A.K.A Output word vectors



Scores over 8 words

Output Layer

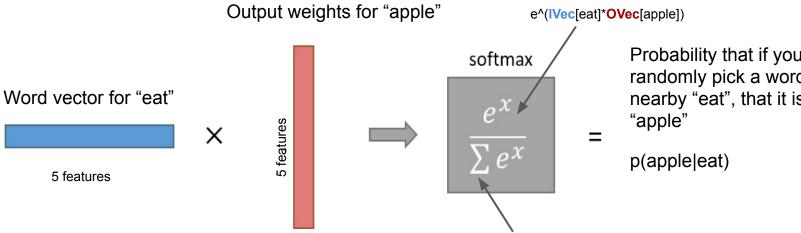
- Softmax classifier
- Output word matrix denoted as OVec



Output Layer Weights Matrix

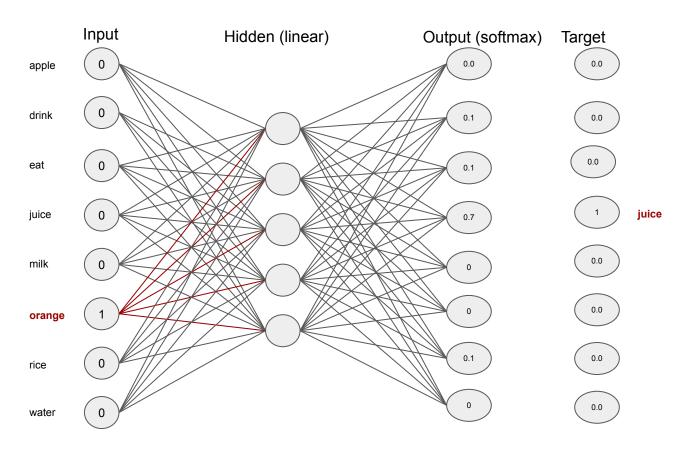
A.K.A Output word vectors

8 words



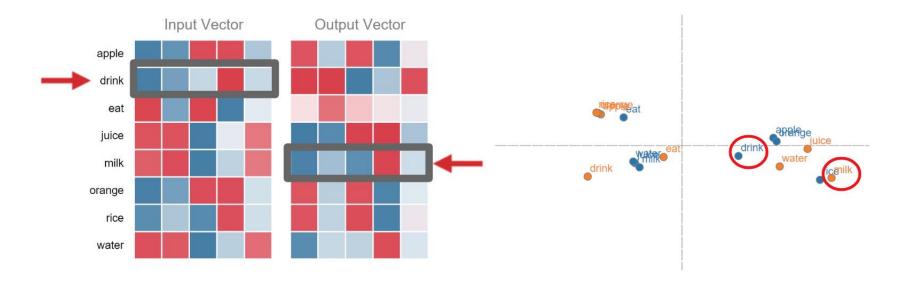
e^(IVec[eat]*OVec[apple]) + e^(IVec[eat]*OVec[juice]) + e^(IVec[eat]*OVec[drink])+e^(IVec[eat]*OVec[other vocab words)

Word2Vec Network



Then, we can compute the loss and call gradient descent to update model parameters.

Updating Word Vectors



A force-directed graph



What decides the strength of the string?

Idea behind Word2Vec

• Feature vector assigned to a word will be adjusted if it can not be used for accurate prediction of that word's context.

 Each word's context in the corpus is the teacher sending error signals back to modify the feature vector.

It means that words with similar context will be assigned similar vectors!

Distributional Semantics

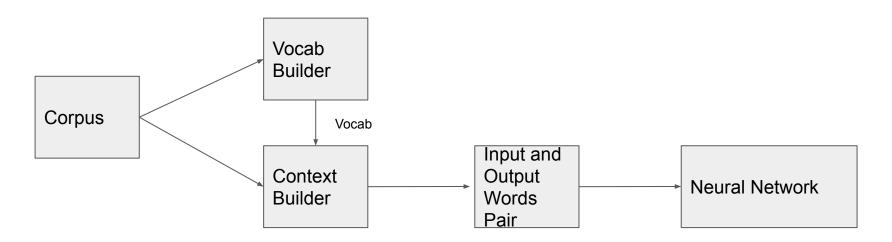
"You shall know a word by the company it keeps" - by Firth (1957)



Input and Output Words

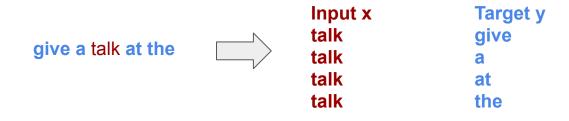
How to select them from corpus

Skip-gram and CBoW differ here.



Skip-Gram

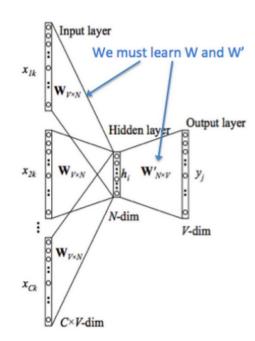
- Task Definition: given a specific word, predict its nearby word (probability output)
- Model input: source word, Model output: nearby word
- Input is one word, output is one word
- The output can be interpreted as prob. scores, which are regarded as how likely it is that each vocabulary word can be nearby your input word.



CBoW

- Task Definition: given context, predict its target word
- Model input: context (several words), Model output: center word
- Input is several words, output is one word
- Core Trick: average these context vectors for prob score computing

give a talk at the Input x (give,a,at,the) Target y talk



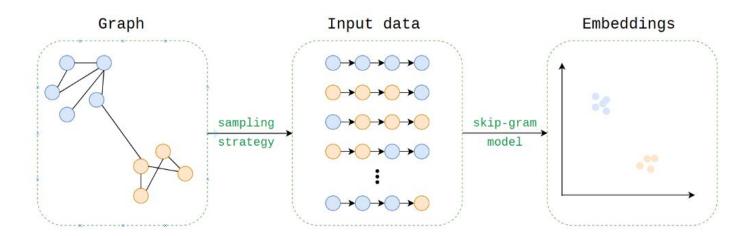
Skip-Gram Vs CBoW

- CBoW: learning to predict the word by the context
- Skip-gram: learning to predict the context by the center word

- ?: several times faster to train the ?
- ?: works well with small amount of the training data, represents well even rare words or phrases.

Embedding for Graph Data

- Embeddings can be extended beyond NLP domain
- Embeddings can be learned for any nodes in a graph

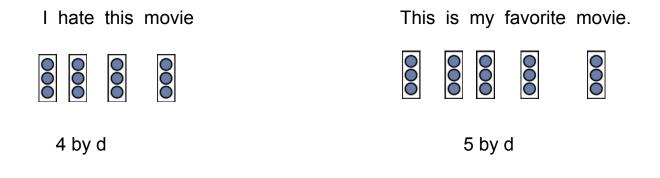


- Nodes can be items, web pages and so on in user clicked stream data
- Embeddings can be learned for any group of discrete and co-occurring states.

Neural Networks for NLP

Sequence of Words

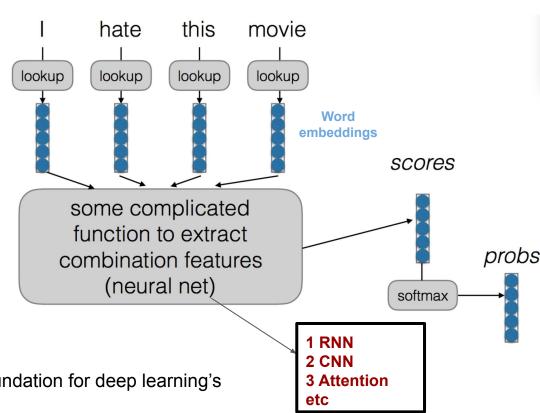
• Each sentence or document can be regarded as a sequence of vectors.



- The shape of matrix depends on the length of sequence. However, the majority of ML systems need fixed-length feature vectors.
- One simple solution: average the sequence of vectors, just like bag-of-words (abandon order information).

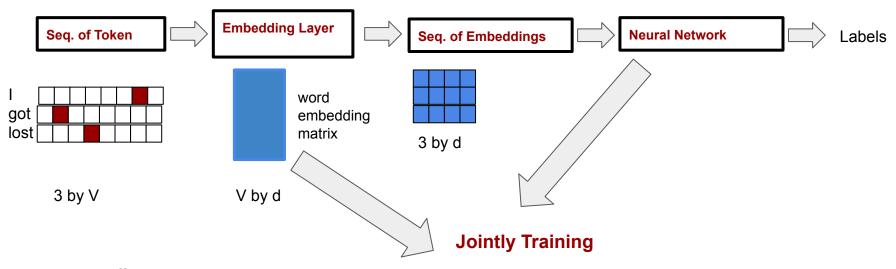
Complex Semantic

- 1. Input Text: a sequence of words;
- 2. Through Word Embedding Look-up: a sequence of word vectors;
- 3. Neural networks is applied upon the vector sequences to learn semantic **composition** for final prediction;
- Human understand the word meaning firstly, then get the whole sentence meaning by composing these words' meaning together.



Word Embeddings is the foundation for deep learning's applications on NLP

Neural Networks for NLP



Three different approaches to build word embedding layer

Learn from Scratch

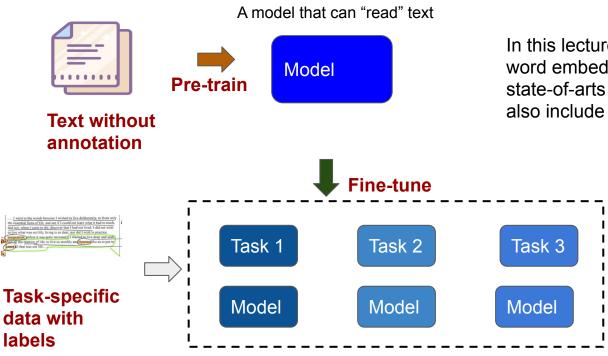
Random initialize the word embedding matrix and update the matrix and neural network parameters in the specific task

2. Pre-train

Got pre-trained word embeddings as the embedding layer and only update neural network parameters in the specific task

3. Pre-train then fine tune

Pre-train then Fine-tune



In this lecture, the model is only referred to word embedding matrix. However, in state-of-arts NLP techniques, the model will also include the following neural network.

Is word2vec good enough?



Word Embeddings used for downstream tasks

- Can not capture different senses of words (context independent)
 - Solution: Take the word order into account

Word2Vec Tool

- Can not address Out-of-Vocabulary words
 - Solution: Use characters or subwords

Contextualized Word Embeddings: BERT



1 use transformer to capture word order 2 input is subwords and output is subwords embeddings

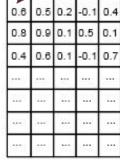
CNN for Text

Convolution Operation

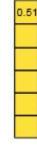
Word Vectors

l 0.6 0.5 0.2 -0.1 0.4 0.2 0.1 0.2 0.1 like 0.8 0.9 0.1 0.5 0.1 0.1 0.1 0.1 0.4 0.1

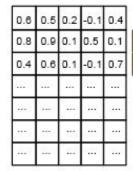
this movie very much

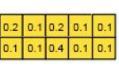


Filters updated during training



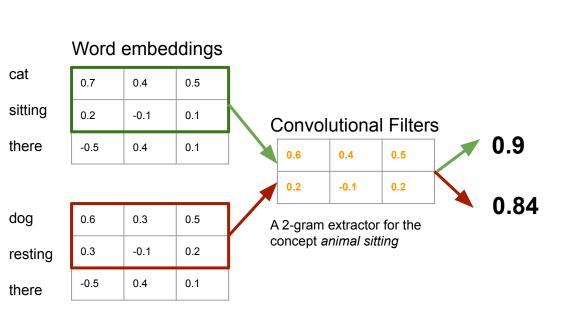
like this movie very much





Feature Maps

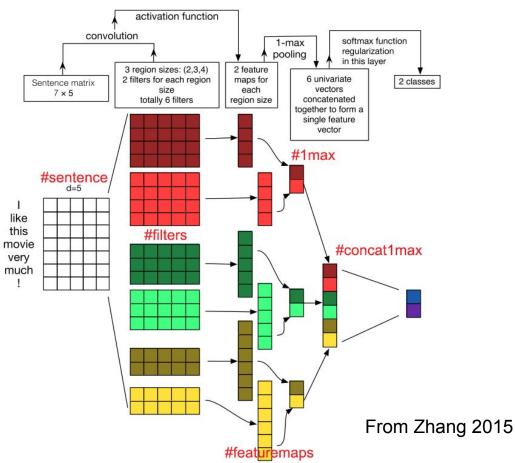
Toy Example



- This convolution provides high activations for 2-grams with certain meaning
- Can be extended to 3-grams,
 4-grams, etc.
- Can have various filters, need to track many n-grams.
- They are called 1D since we only slice the windows only in one direction

Why is it better than BoW?

CNN Framework



Multiple Channels

- Like image, CNN is applied on R-G-B channels
- For NLP, different word embeddings can be regarded as different channels

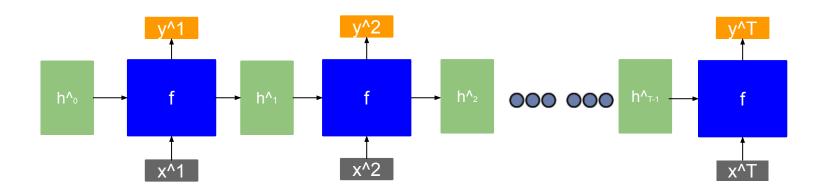
CNN for NLP

- 1. n-grams features are important (window size)
- 2. Location of key n-grams are trivial (pooling)
- Stack of Convolutional layer or large window size can also capture long-range information

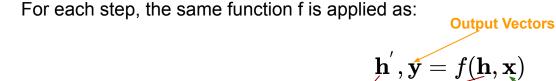
RNN for Text

Recurrent Neural Network

Given a sequence of T steps: x^1, x^2, ..., x^T



Input Vectors

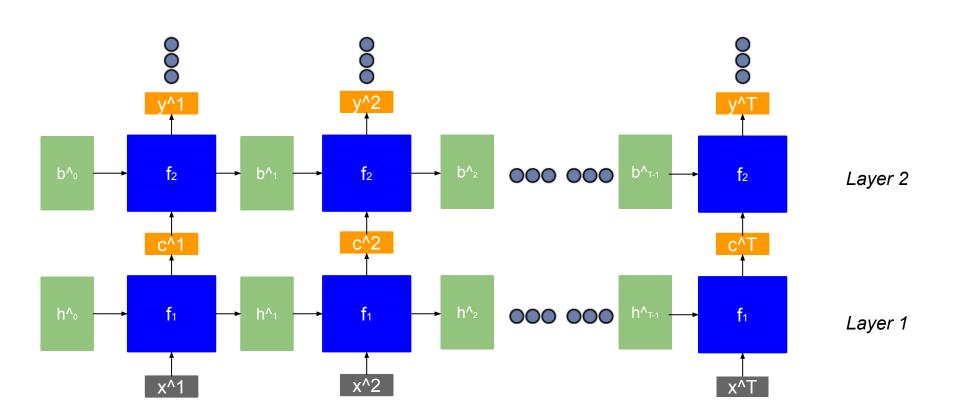


Hidden vectors with the same dimension

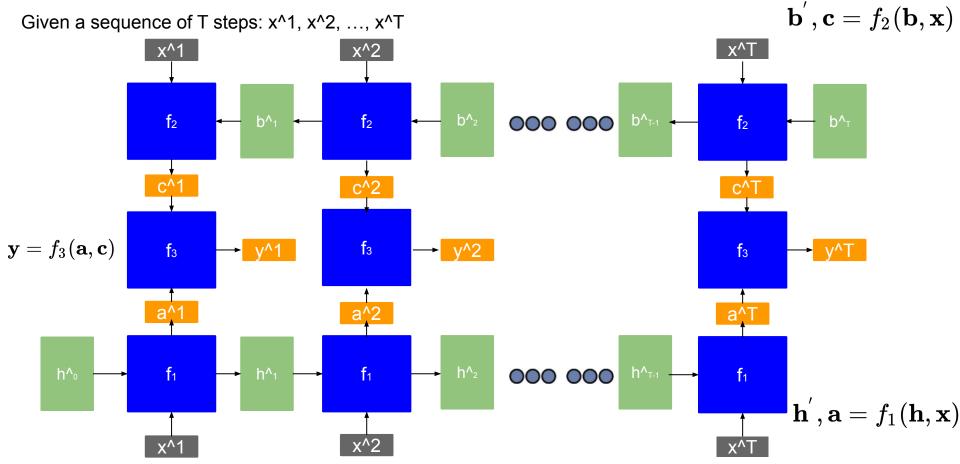
Deep RNN

 $\mathbf{h}^{'},\mathbf{c}=f_{1}(\mathbf{h},\mathbf{x})~~\mathbf{b}^{'},\mathbf{y}=f_{2}(\mathbf{b},\mathbf{c})$

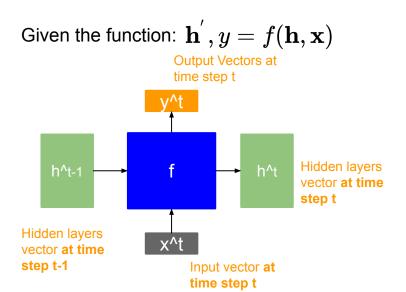
Given a sequence of T steps: x^1, x^2, ..., x^T



Bidirectional RNN



Naive RNN

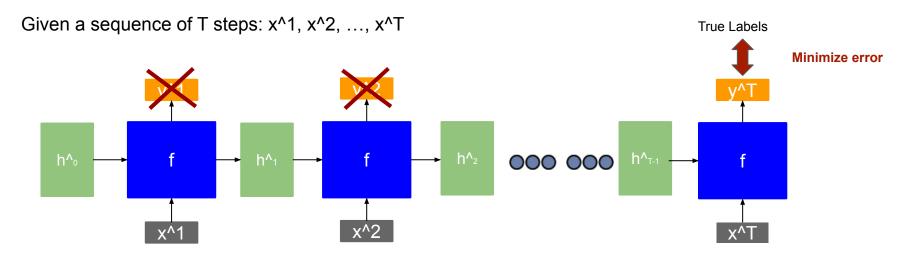


$$\mathbf{h}_t = \sigma(\mathbf{W}^{hh}\mathbf{h}_{t-1} + \mathbf{W}^i\mathbf{x}_t)$$

$$\mathbf{y}_t = softmax(\mathbf{W}^o\mathbf{h}_t)$$

 $\mathbf{W}^{hh}, \mathbf{W}^i, \mathbf{W}^o$ Time-Invariant Model Parameters

RNN for Sequence Classification



Only apply classification layer on the hidden output of the last time step.

RNN's Bottleneck

- RNN is not suitable for parallel computation.
- RNN's training is not easy
 - Gradient Vanishing
 - Gradient Exploding
 - Reminder on gradient descent algorithms:

W = W - alpha dL/dW

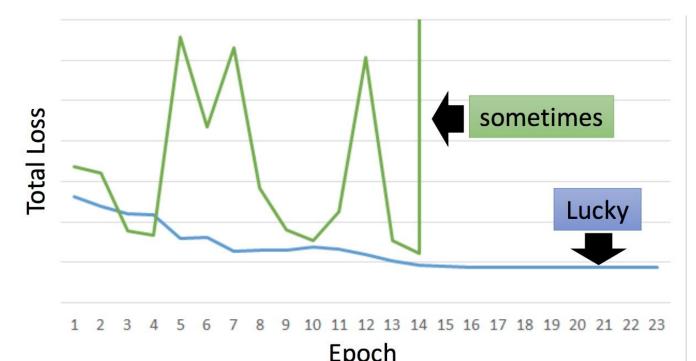
W is the layer weights;

L is the loss function;

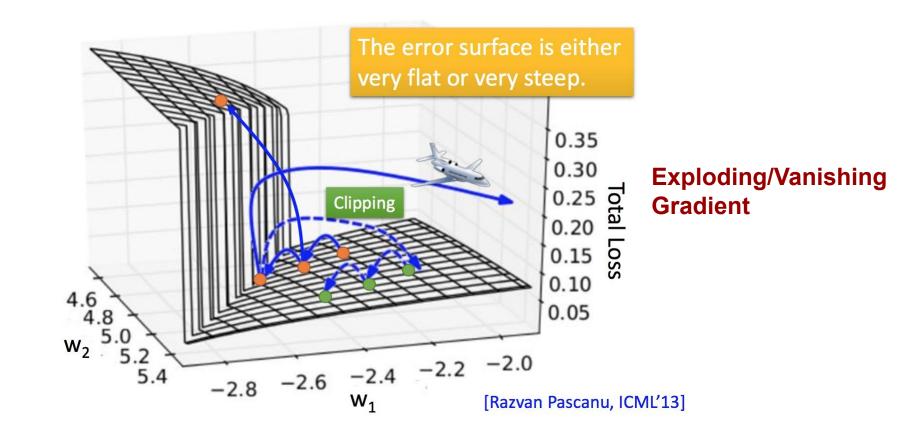
Alpha is the learning rate, which represents how aggressive that our model parameters are updated.

RNN Training is Hard

Real experiments on Language Models

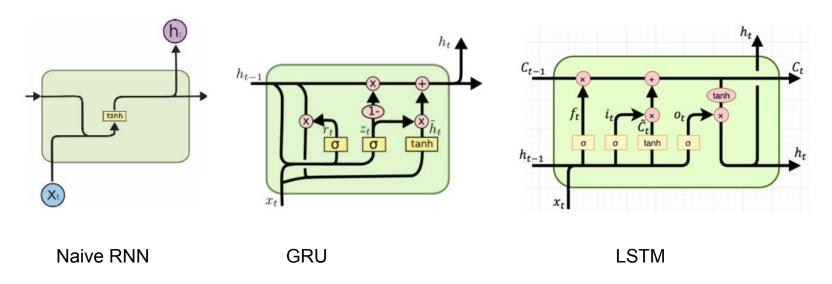


Rough Error Surface of RNN



LSTM/GRU

• LSTM/GRU are using gates in cell computation to control information flow



Source: https://colah.github.io/posts/2015-08-Understanding-LSTMs/

 Neural language models Multi-task learning Word embeddings Neural networks for NLP Sequence-to-sequence models Attention Memory-based networks Pretrained language models 		
 Word embeddings Neural networks for NLP Sequence-to-sequence models Attention Memory-based networks 	001	Neural language models
 Neural networks for NLP Sequence-to-sequence models Attention Memory-based networks 	008	Multi-task learning
 Sequence-to-sequence models Attention Memory-based networks 	013	Word embeddings
015 Attention 015 Memory-based networks	013	Neural networks for NLP
015 Memory-based networks	014	Sequence-to-sequence models
	015	Attention
018 Pretrained language models	015	Memory-based networks
	018	Pretrained language models

Frontiers in NLP II