Statistical Language Models

Week 8

- http://www.primaryobjects.com/kory-becker/

> Here's an interesting bit of background. One of my products was a software tool, which in 2002, earned a PC Magazine Editors' Choice Award. This was one of my first products that included artificial intelligence technology in a commercial software product. The particular piece used Markov chains to identify malware file-names in the Windows system folder (at that time, malware usually took on randomly arranged letters for the file-names). The Markov chain was trained to recognize likely combinations of valid letters that file-names typically hold. Any file-name that strayed too far from the pattern was identified as malware. This worked surprisingly well!

- https://gist.github.com/primaryobjects/8038d345aae48ae48988906b0525d175

Word2Vec

- 2013 Google research work

Efficient Estimation of Word Representations in Vector Space

Tomas Mikolov
Google Inc., Mountain View, CA
tmikolov@google.com

Kai Chen
Google Inc., Mountain View, CA
kaichen@google.com

Greg Corrado
Google Inc., Mountain View, CA
gcorrado@google.com

Jeffrey Dean
Google Inc., Mountain View, CA
jeff@google.com

Abstract

We propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing techniques based on different types of neural networks. We observe large improvements in accuracy at much lower computational cost, i.e. it takes less than a day to learn high quality word vectors from a 1.6 billion words data set. Furthermore, we show that these vectors provide state-of-the-art performance on our test set for measuring syntactic and semantic word similarities.

Distributed Representations of Words and Phrases and their Compositionality

Tomas Mikolov
Google Inc.
Mountain View
mikolov@google.com

Ilya Sutskever
Google Inc.
Mountain View
ilyasu@google.com

Kai Chen
Google Inc.
Mountain View
kai@google.com

Greg Corrado
Google Inc.
Mountain View
gcorrado@google.com

Jeffrey Dean
Google Inc.
Mountain View
jeff@google.com

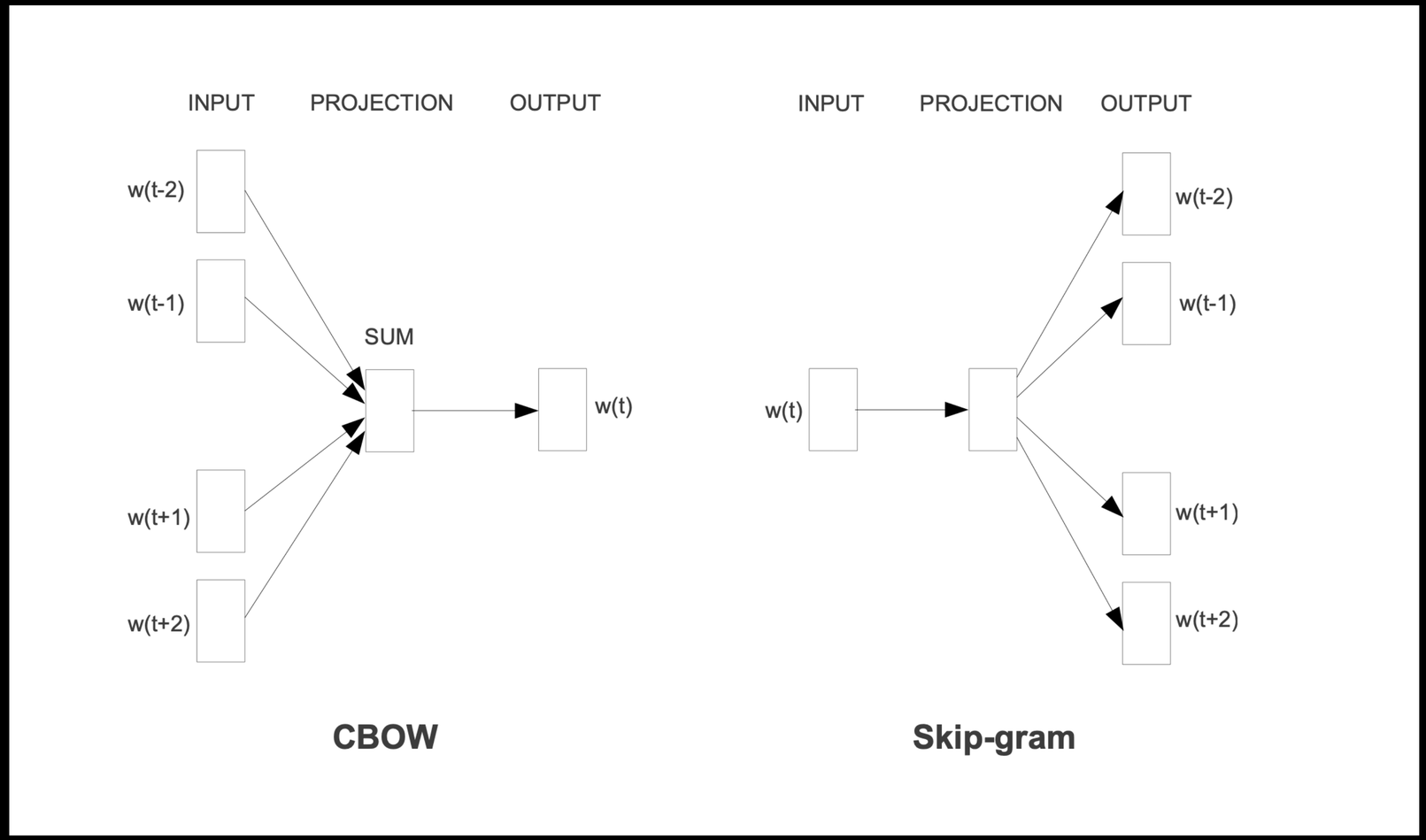
Abstract

The recently introduced continuous Skip-gram model is an efficient method for learning high-quality distributed vector representations that capture a large number of precise syntactic and semantic word relationships. In this paper we present several extensions that improve both the quality of the vectors and the training speed. By subsampling of the frequent words we obtain significant speedup and also learn more regular word representations. We also describe a simple alternative to the hierarchical softmax called negative sampling.

An inherent limitation of word representations is their indifference to word order and their inability to represent idiomatic phrases. For example, the meanings of "Canada" and "Air" cannot be easily combined to obtain "Air Canada". Motivated by this example, we present a simple method for finding phrases in text, and show that learning good vector representations for millions of phrases is possible.

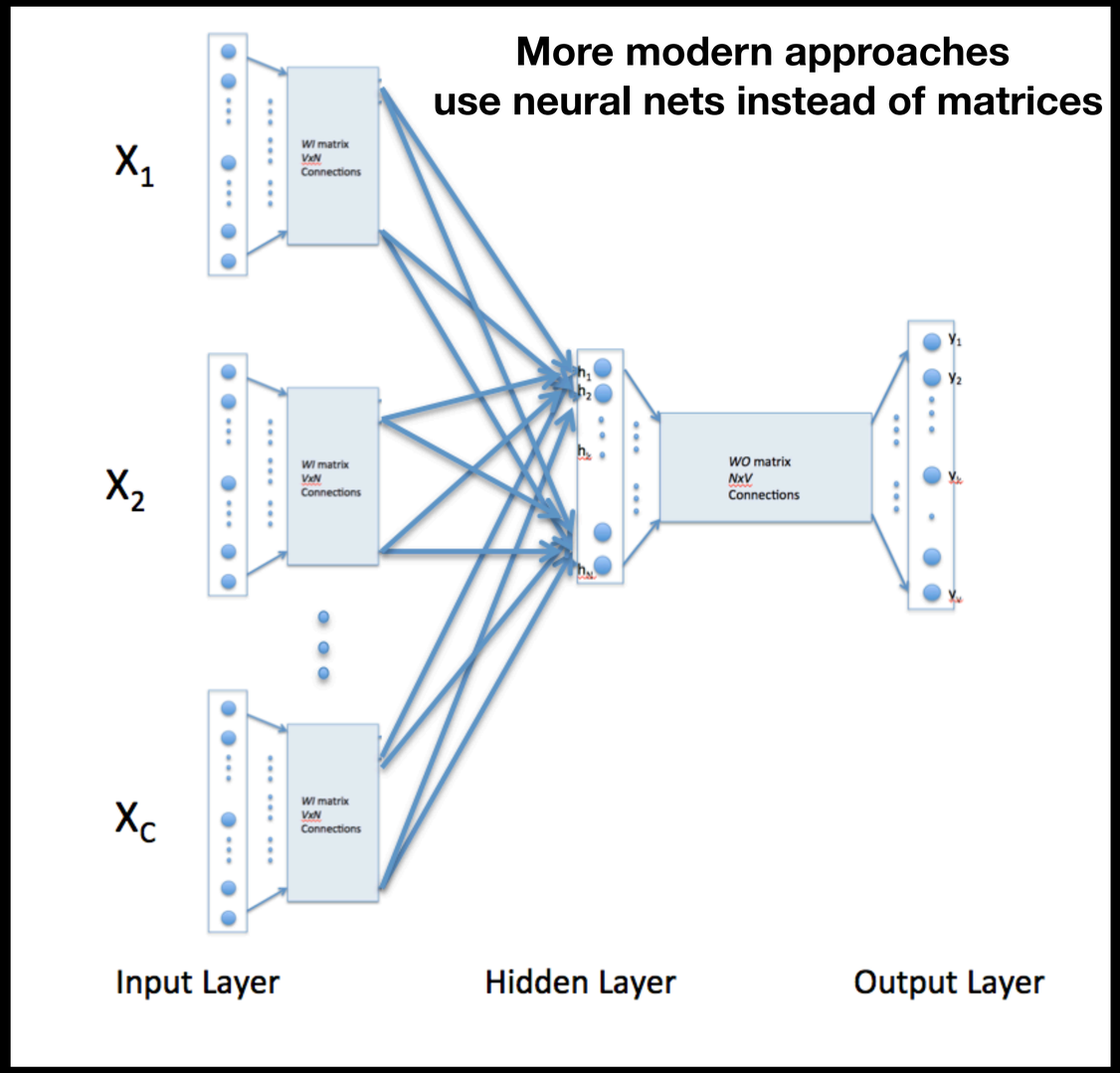
2 strategies

- From context predict the middle (CBOW)
- From word predict context (Skip-gram)



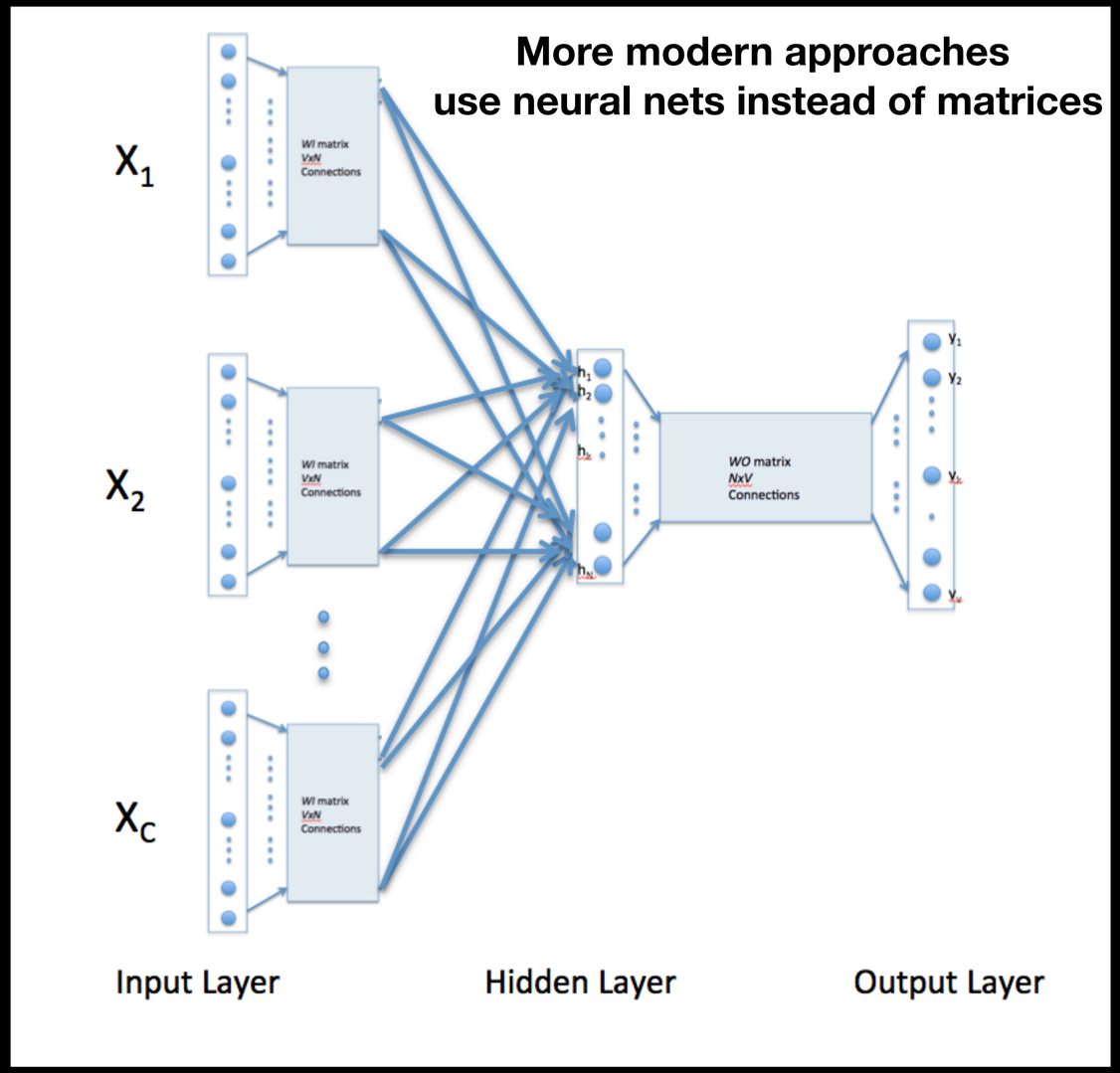
Word2Vec

- Goal: predict missing word(t)
- Input: {word(t-1), word(t-2), word(t+1), word(t+2)}
- aka: "Continuous Bag of Words" model for predicting centre word from context; BUT word order does not matter



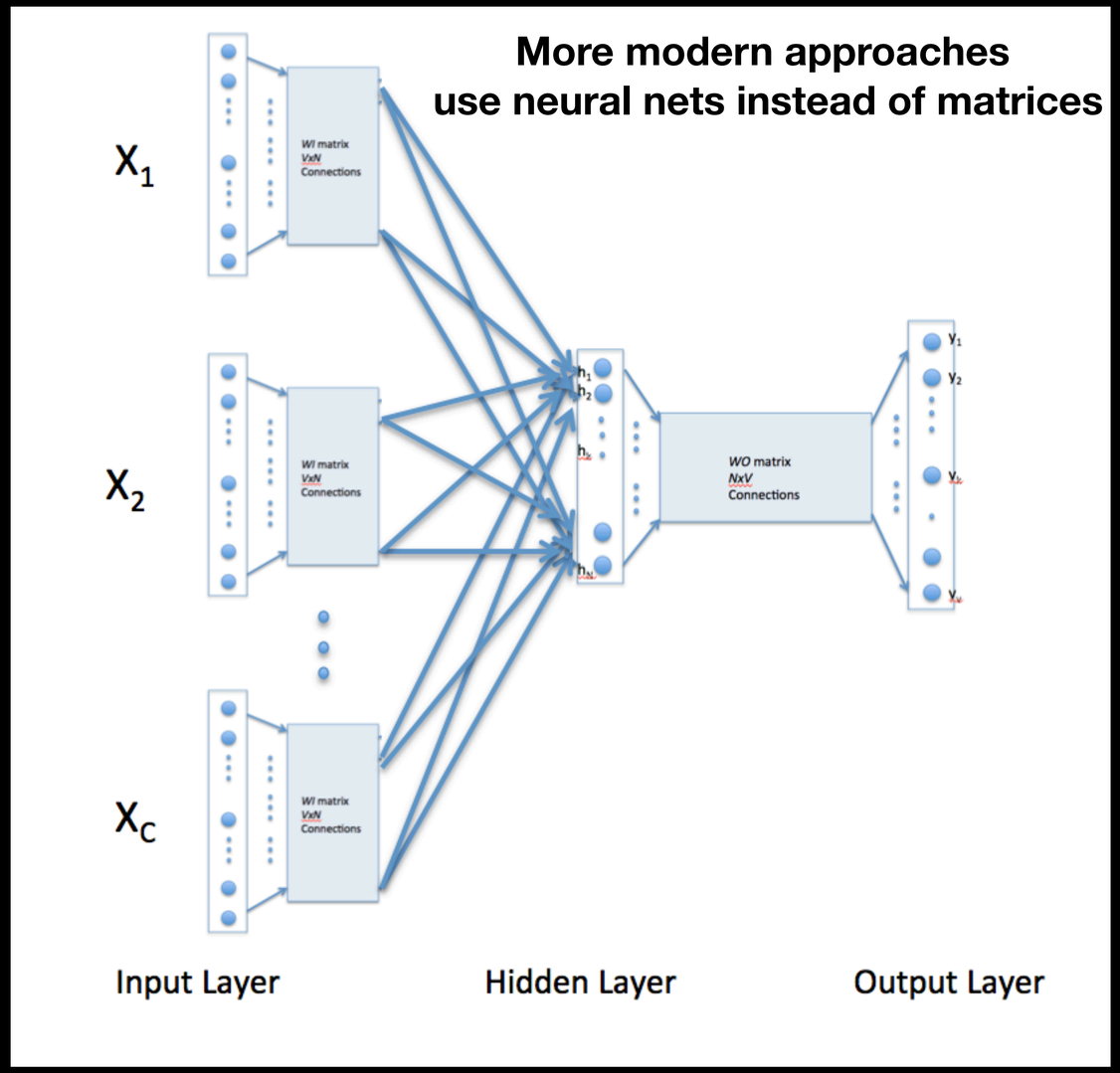
Word2Vec

- Goal: predict missing word(t)
- Input: {word(t-1), word(t-2), word(t+1), word(t+2)}
- Hidden layer: numerically combines weights and one-hot-encoded covariates



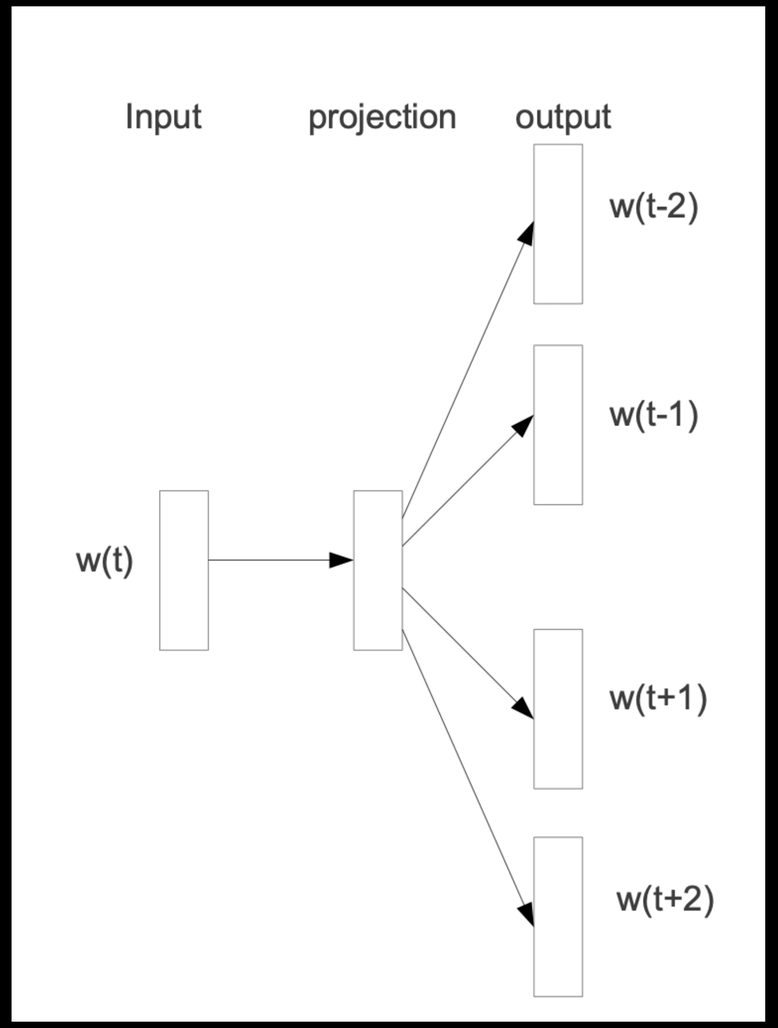
Word2Vec

- Goal: predict missing word(t)
- Input: {word(t-1), word(t-2), word(t+1), word(t+2)}
- Hidden layer: numerically combines weights and one-hot-encoded covariates
- Output: best word(t)
- Note input dimension and output dimension are the same. Internal dimension (hidden) determines the dimension of the embedding space



Skip-gram architecture

- Input is a single (middle) word
- Output goal is the prediction of all of the neighbouring words



Skip-gram architecture

- Maximize the average log probability
- Training context size c (word window)

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t)$$

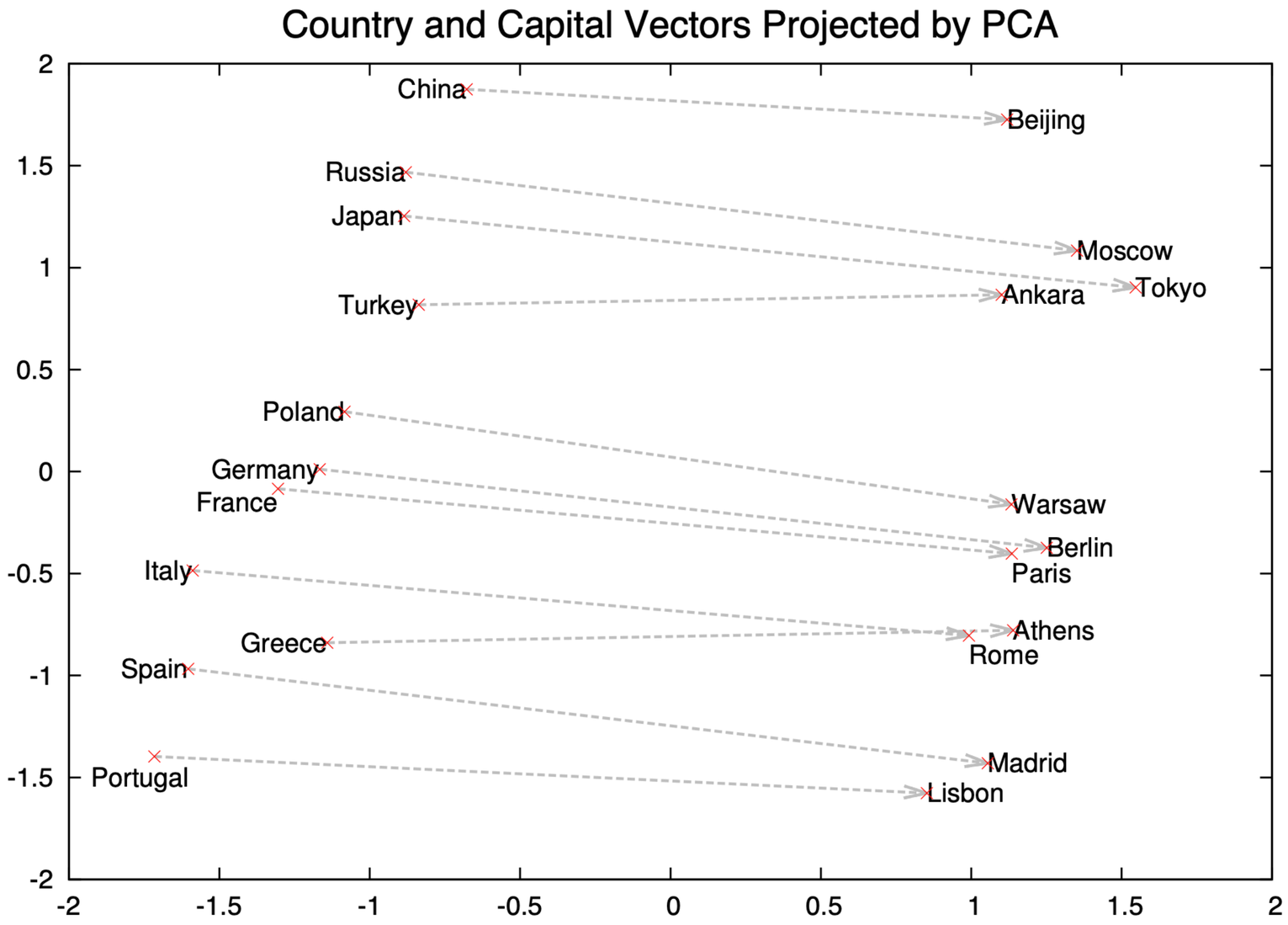


Figure 2: Two-dimensional PCA projection of the 1000-dimensional Skip-gram vectors of countries and their capital cities. The figure illustrates ability of the model to automatically organize concepts and learn implicitly the relationships between them, as during the training we did not provide any supervised information about what a capital city means.

Distributed Representations of Words and Phrases and their Compositionality

Tomas Mikolov
Google Inc.
Mountain View
mikolov@google.com

Ilya Sutskever
Google Inc.
Mountain View
ilyasu@google.com

Kai Chen
Google Inc.
Mountain View
kai@google.com

Greg Corrado
Google Inc.
Mountain View
gcorrado@google.com

Jeffrey Dean
Google Inc.
Mountain View
jeff@google.com

Abstract

The recently introduced continuous Skip-gram model is an efficient method for learning high-quality distributed vector representations that capture a large number of precise syntactic and semantic word relationships. In this paper we present several extensions that improve both the quality of the vectors and the training speed. By subsampling of the frequent words we obtain significant speedup and also learn more regular word representations. We also describe a simple alternative to the hierarchical softmax called negative sampling.

An inherent limitation of word representations is their indifference to word order and their inability to represent idiomatic phrases. For example, the meanings of "Canada" and "Air" cannot be easily combined to obtain "Air Canada". Motivated by this example, we present a simple method for finding phrases in text, and show that learning good vector representations for millions of phrases is possible.

Closeness of terms

- Cosine similarity:
- $cos(\theta) = \frac{\sum_{i=1}^{N} A_i B_i}{\sqrt{\sum_{i=1}^{N} A_i^2}\sqrt{\sum_{i=1}^{N} B_i^2}}$ for vectors A, B with components A_i, B_i\
- -1 = exact opposite, 1 = exactly the same, 0 = orthogonal

Word2Vec

- library(devtools)
- library(httr)
- library(tm)
- install_github("bmschmidt/wordVectors") # yup install from GitHub
- library(wordVectors)
- vignette("introduction", "wordVectors")

Model examination

- Working through examples:
- https://gist.github.com/primaryobjects/8038d345aae48ae48988906b0525d175
- Input is a 100MB text file, the first 10^8 bytes of the XML text dump of the English version of Wikipedia on Mar. 3, 2006
- http://mattmahoney.net/dc/text8.zip

Model examination

- # Train word2vec model and explore.
- model <- word2vec('text8')
- Embedding dimension is 200, window is size 12

anarchism originated as a term of abuse first used against early working class radicals including the diggers of the english revolution and the sans culottes of the french revolution whilst the term is still used in a pejorative way to describe any act that used violent means to destroy the organization of society it has also been taken up as a positive label by self defined anarchists the word anarchism is derived from the greek without archons ruler chief king anarchism as a political philosophy is the belief that rulers are unnecessary and should be abolished although there are differing interpretations of what this means anarchism also refers to related social movements that advocate the elimination of authoritarian institutions particularly the state the word anarchy as most anarchists use it does not imply chaos nihilism or anomie but rather a harmonious anti authoritarian society in place of what are regarded as authoritarian political structures and coercive economic institutions anarchists advocate social relations based upon voluntary association of autonomous individuals mutual aid and self governance while anarchism is most easily defined by what it is against anarchists also offer positive visions of what they believe to be a truly free society however ideas about how an anarchist society might work vary considerably especially with respect to economics there is also disagreement about how a free society might be brought about origins and predecessors kropotkin and others argue that before recorded history human society was organized on anarchist principles most anthropologists follow kropotkin and engels in believing that hunter gatherer bands were egalitarian and lacked division of labour accumulated wealth or decreed law and had equal access to resources william godwin anarchists including the the anarchy organisation and rothbard find anarchist attitudes in taoism from ancient china kropotkin found similar ideas in stoic zeno of citium according to kropotkin zeno repudiated the omnipotence of the state its intervention and regimentation and proclaimed the sovereignty of the moral law of the individual the anabaptists of one six th century europe are sometimes considered to be religious forerunners of modern anarchism bertrand russell in his history of western philosophy writes that the anabaptists repudiated all law since they held that the good man will be guided at every moment by the holy spirit from this premise they arrive at communism the diggers or true levellers were an early communistic movement during the time of the english civil war and are considered by some as forerunners of modern anarchism in the modern era the first to use the term to mean something other than chaos was louis armand baron de lahontan in his nouveaux voyages dans l am rique septentrionale one seven zero three where he described the indigenous american society which had no state laws prisons priests or private property as being in anarchy russell means a libertarian and leader in the american indian movement has repeatedly stated that he is an anarchist and so are all his ancestors in one seven nine three in the thick of the french revolution william godwin published an enquiry concerning political justice although godwin did not use the word anarchism many later anarchists have regarded this book as the first major anarchist text and godwin as the founder of philosophical anarchism but at this point no anarchist movement yet existed and the term anarchiste was known mainly as an insult hurled by the bourgeois girondins at more radical elements in the french revolution the first self labelled anarchist pierre joseph proudhon it is commonly held that it wasn t until pierre joseph proudhon published what is property in one eight four zero that the term anarchist was adopted as a self description it is for this reason that some claim proudhon as the founder of modern anarchist theory in what is property proudhon answers with the famous accusation property is theft in this work he opposed the institution of decreed property propri t where owners have complete rights to use and abuse their property as they wish such as exploiting workers for profit in its place proudhon supported what he called possession individuals can have limited rights to use resources capital and goods in accordance with principles of equality and justice proudhon s vision of anarchy which he called mutualism mutuellisme involved an exchange economy where individuals and groups could trade the products of their labor using labor notes which represented the amount of working time involved in production this would ensure that no one would profit from the labor of others workers could freely join together in co operative workshops an interest free bank would be set up to provide everyone with access to the means of production proudhon s ideas were influential within french working class movements and his followers were active in the revolution of one eight four eight in france proudhon s philosophy of property is complex it was developed in a number of works over his lifetime and there are differing interpretations of some of his ideas for more detailed discussion see here max stirner s egoism in his the ego and its own stirner argued that most commonly accepted social institutions including the notion of state property as a right natural rights in general and the very notion of society were mere illusions or ghosts in the mind saying of society that the individuals are its reality he advocated egoism and a form of amoralism in which individuals would unite in associations of egoists only when it was in their self interest to do so for him property simply comes about through might whoever knows how to take to defend the

Model examination

- #NearBy Terms
- model %>% closest_to("communism")

- # Plot similar terms to 'computer' and 'internet'.
- computers <- model[[c("computer","internet"),average=F]]

- # model[1:3000,] here restricts to the 3000 most common words in the set.
- computer_and_internet <- model[1:3000,] %>% cosineSimilarity(computers)

- # Filter to the top 20 terms.
- computer_and_internet <- computer_and_internet[
- rank(-computer_and_internet[,1])<20 | rank(-computer_and_internet[,2])<20,]

- plot(computer_and_internet,type='n')
- text(computer_and_internet,labels=rownames(computer_and_internet))

Cookbook skip grams

- vignette("introduction", "wordVectors")
- collection of cookbooks from Michigan State University:
- "The Feeding America: The Historic American Cookbook dataset contains transcribed and encoded text from 76 influential American cookbooks held by MSU Libraries Special Collections. Features encoded within the text include but are not limited to recipes, types of recipes, cooking implements, and ingredients. The 76 texts were chosen among more than 7000 cookbooks that MSU Libraries holds as representative of periods and themes in American cookbook history spanning the late 18th to early 20th century."
- Feeding America: The Historic American Cookbook Dataset. East Lansing: Michigan State University Libraries Special Collections. https://www.lib.msu.edu/feedingamericadata/

whipped_cream over the jelly in each basket serve in a bed of orange or laurel_leaves no 2 with a vegetable_cutter cut out several small portions of the peel in the basket and handle to give an open work effect and fill with a mixture of orange wine and lemon jelly cut into inch_dice and piled lightly in the baskets or the baskets may be filled with bavarian_cream orange sections cut off a small portion from the end of the orange and scoop_out the pulp and juice be careful_not to break through the skin fill them with orange jelly which is thoroughly cold but not hard and place them upright in a pan of broken ice when hard cut each orange in quarters and serve garnished with green leaves imperial cream make the rule for lemon jelly page 350 and color part of it pink with cochineal or cranberry juice harden it in a shallow_pan make snow pudding page 347 and when nearly stiff enough to drop stir in small squares of the pink and lemon jelly mould and when ready to serve turn out on a dish garnish the base and top with macaroons soaked in wine pour rich boiled custard round the dish and put macaroons and cubes of the jellies in the custard whipped_cream many wholesome delicious and attractive dishes may be made with whipped_cream to those_who can obtain plenty of cream these dishes afford a cheaper more easily prepared and far more satisfactory course than pie and many forms of hot puddings many of them are equally suitable for tea very rich cream should be diluted and well mixed with an equal_quantity of milk the best_quality of cream obtained from the milkman is usually of the proper_consistency thin cream will become liquid after whipping and thick cream will turn to butter the cream should always be icy cold when it is to be served as a garnish or for cream whips it should be sweetened and flavored before it is whipped a whip_churn is the best utensil for whipping cream this is a tin cylinder perforated at the bottom and sides and having a perforated_dasher when the churn is placed in a bowl of cream and the dasher worked up and down the air is forced from the cylinder into the cream causing it to become light and frothy a dover_egg beater will make the cream light but it has a different consistency from that obtained by churning to whip cream place a bowl half filled with cream in a pan of broken ice when very cold put the churn into the cream hold the cylinder firmly and keep the cover in place with the left_hand tip the churn slightly that the cream may flow out at the bottom work the dasher with a light short stroke up and a hard pushing stroke down when the froth appears stir it

- Main tool is:

- train_word2vec(InputFileName,OutputFileName,vectors=LatentDimension, threads=CPUCores,window=ContextWindow,iter=ObviouslyMoreIsBetter ButSlower)

- train_word2vec("cookbooks.txt","cookbook_vectors.bin",vectors=200,threads=4,window=12,iter=5)

- #Skipgram default
- model = train_word2vec("cookbooks.txt","cookbook_vectors.bin",vectors=200,threads=4,cbow=1, window=12,iter=5,negative_samples=0)
- model %>% closest_to("fish")
- some_fish = **closest_to**(model,model[[c("fish","salmon","trout","shad","flounder","carp","roe","eels")]], 150)
- fishy = model[[some_fish$word,average=F]]
- plot(fishy,method="pca")

- #CBOW
- model = train_word2vec("cookbooks.txt","cookbook_vectors_CBOW.bin",vectors=200,threads=4,**cbow=1**,window=12,iter=5,negative_samples=0)
- model %>% closest_to("beet")