word2vec

Utility of General and Specific Word Embeddings for Classifying Translational Stages of Research

Conventional text classification models make a bag-of-words assumption reducing text into word occurrence counts per document. Recent algorithms such as word2vec are capable of learning semantic meaning and similarity between words in an entirely unsupervised manner using a contextual window and doing so much faster than previous methods. Each word is projected into vector space such that similar meaning words such as "strong" and "powerful" are projected into the same general Euclidean space. Open questions about these embeddings include their utility across classification tasks and the optimal properties and source of documents to construct broadly functional embeddings. In this work, we demonstrate the usefulness of pre-trained embeddings for classification in our task and demonstrate that custom word embeddings, built in the domain and for the tasks, can improve performance over word embeddings learnt on more general data including news articles or Wikipedia.  
 **main\_author: Yindalon Aphinyanaphongs**  
 **published\_date: 2017-05-17 17:08:11+00:00**  
 *pdf\_url: http://arxiv.org/pdf/1705.06262v2*

hep-th

We apply techniques in natural language processing, computational linguistics, and machine-learning to investigate papers in hep-th and four related sections of the arXiv: hep-ph, hep-lat, gr-qc, and math-ph. All of the titles of papers in each of these sections, from the inception of the arXiv until the end of 2017, are extracted and treated as a corpus which we use to train the neural network Word2Vec. A comparative study of common n-grams, linear syntactical identities, word cloud and word similarities is carried out. We find notable scientific and sociological differences between the fields. In conjunction with support vector machines, we also show that the syntactic structure of the titles in different sub-fields of high energy and mathematical physics are sufficiently different that a neural network can perform a binary classification of formal versus phenomenological sections with 87.1% accuracy, and can perform a finer five-fold classification across all sections with 65.1% accuracy.  
 **main\_author: Brent D. Nelson**  
 **published\_date: 2018-06-27 11:44:35+00:00**  
 *pdf\_url: http://arxiv.org/pdf/1807.00735v1*

A neural network system for transformation of regional cuisine style

We propose a novel system which can transform a recipe into any selected regional style (e.g., Japanese, Mediterranean, or Italian). This system has two characteristics. First the system can identify the degree of regional cuisine style mixture of any selected recipe and visualize such regional cuisine style mixtures using barycentric Newton diagrams. Second, the system can suggest ingredient substitutions through an extended word2vec model, such that a recipe becomes more authentic for any selected regional cuisine style. Drawing on a large number of recipes from Yummly, an example shows how the proposed system can transform a traditional Japanese recipe, Sukiyaki, into French style.  
 **main\_author: Yoshiki Ishikawa**  
 **published\_date: 2017-05-06 18:35:35+00:00**  
 *pdf\_url: http://arxiv.org/pdf/1705.03487v2*

The Corpus Replication Task

In the field of Natural Language Processing (NLP), we revisit the well-known word embedding algorithm word2vec. Word embeddings identify words by vectors such that the words' distributional similarity is captured. Unexpectedly, besides semantic similarity even relational similarity has been shown to be captured in word embeddings generated by word2vec, whence two questions arise. Firstly, which kind of relations are representable in continuous space and secondly, how are relations built. In order to tackle these questions we propose a bottom-up point of view. We call generating input text for which word2vec outputs target relations solving the Corpus Replication Task. Deeming generalizations of this approach to any set of relations possible, we expect solving of the Corpus Replication Task to provide partial answers to the questions.  
 **main\_author: Tobias Eichinger**  
 **published\_date: 2018-06-20 20:37:28+00:00**  
 *pdf\_url: http://arxiv.org/pdf/1806.07978v1*

SubGram: Extending Skip-gram Word Representation with Substrings

Skip-gram (word2vec) is a recent method for creating vector representations of words ("distributed word representations") using a neural network. The representation gained popularity in various areas of natural language processing, because it seems to capture syntactic and semantic information about words without any explicit supervision in this respect. We propose SubGram, a refinement of the Skip-gram model to consider also the word structure during the training process, achieving large gains on the Skip-gram original test set.  
 **main\_author: Ondřej Bojar**  
 **published\_date: 2018-06-18 09:31:38+00:00**  
 *pdf\_url: http://arxiv.org/pdf/1806.06571v1*

Unsupervised Deep Image Hashing through Tag Embeddings

Many approaches to semantic image hashing have been formulated as supervised learning problems that utilize images and label information to learn the binary hash codes. However, large-scale labelled image data is expensive to obtain, thus imposing a restriction on the usage of such algorithms. On the other hand, unlabelled image data is abundant due to the existence of many Web image repositories. Such Web images may often come with images tags that contains useful information, although raw tags in general do not readily lead to semantic labels. Motivated by this scenario, we formulate the problem of image hashing as an unsupervised learning problem. We utilize the information contained in the user-generated tags associated with the images to learn the hash codes. More specifically, we extract the word2vec semantic embeddings of the tags and use the information contained in them for constraining the learning. Accordingly, we name our model Unsupervised Deep Hashing using Tag Embeddings (UDHT). UDHT is tested for the task of semantic image retrieval and is compared against several state-of-art unsupervised models. Results show that our approach sets a new state-of-art in the area of unsupervised image hashing.  
 **main\_author: Baoxin Li**  
 **published\_date: 2018-06-15 05:24:30+00:00**  
 *pdf\_url: http://arxiv.org/pdf/1806.05804v1*

Spectral Network Embedding: A Fast and Scalable Method via Sparsity

Network embedding aims to learn low-dimensional representations of nodes in a network, while the network structure and inherent properties are preserved. It has attracted tremendous attention recently due to significant progress in downstream network learning tasks, such as node classification, link prediction, and visualization. However, most existing network embedding methods suffer from the expensive computations due to the large volume of networks. In this paper, we propose a $10\times \sim 100\times$ faster network embedding method, called Progle, by elegantly utilizing the sparsity property of online networks and spectral analysis. In Progle, we first construct a \textit{sparse} proximity matrix and train the network embedding efficiently via sparse matrix decomposition. Then we introduce a network propagation pattern via spectral analysis to incorporate local and global structure information into the embedding. Besides, this model can be generalized to integrate network information into other insufficiently trained embeddings at speed. Benefiting from sparse spectral network embedding, our experiment on four different datasets shows that Progle outperforms or is comparable to state-of-the-art unsupervised comparison approaches---DeepWalk, LINE, node2vec, GraRep, and HOPE, regarding accuracy, while is $10\times$ faster than the fastest word2vec-based method. Finally, we validate the scalability of Progle both in real large-scale networks and multiple scales of synthetic networks.  
 **main\_author: Ming Ding**  
 **published\_date: 2018-06-07 11:38:34+00:00**  
 *pdf\_url: http://arxiv.org/pdf/1806.02623v2*

fMRI Semantic Category Decoding using Linguistic Encoding of Word Embeddings

The dispute of how the human brain represents conceptual knowledge has been argued in many scientific fields. Brain imaging studies have shown that the spatial patterns of neural activation in the brain are correlated with thinking about different semantic categories of words (for example, tools, animals, and buildings) or when viewing the related pictures. In this paper, we present a computational model that learns to predict the neural activation captured in functional magnetic resonance imaging (fMRI) data of test words. Unlike the models with hand-crafted features that have been used in the literature, in this paper we propose a novel approach wherein decoding models are built with features extracted from popular linguistic encodings of Word2Vec, GloVe, Meta-Embeddings in conjunction with the empirical fMRI data associated with viewing several dozen concrete nouns. We compared these models with several other models that use word features extracted from FastText, Randomly-generated features, Mitchell's 25 features [1]. The experimental results show that the predicted fMRI images using Meta-Embeddings meet the state-of-the-art performance. Although models with features from GloVe and Word2Vec predict fMRI images similar to the state-of-the-art model, model with features from Meta-Embeddings predicts significantly better. The proposed scheme that uses popular linguistic encoding offers a simple and easy approach for semantic decoding from fMRI experiments.  
 **main\_author: Bapi Raju S**  
 **published\_date: 2018-06-13 10:59:33+00:00**  
 *pdf\_url: http://arxiv.org/pdf/1806.05177v1*

Speech2Vec: A Sequence-to-Sequence Framework for Learning Word Embeddings from Speech

In this paper, we propose a novel deep neural network architecture, Speech2Vec, for learning fixed-length vector representations of audio segments excised from a speech corpus, where the vectors contain semantic information pertaining to the underlying spoken words, and are close to other vectors in the embedding space if their corresponding underlying spoken words are semantically similar. The proposed model can be viewed as a speech version of Word2Vec. Its design is based on a RNN Encoder-Decoder framework, and borrows the methodology of skipgrams or continuous bag-of-words for training. Learning word embeddings directly from speech enables Speech2Vec to make use of the semantic information carried by speech that does not exist in plain text. The learned word embeddings are evaluated and analyzed on 13 widely used word similarity benchmarks, and outperform word embeddings learned by Word2Vec from the transcriptions.  
 **main\_author: James Glass**  
 **published\_date: 2018-03-23 20:59:09+00:00**  
 *pdf\_url: http://arxiv.org/pdf/1803.08976v2*

Text Classification based on Word Subspace with Term-Frequency

Text classification has become indispensable due to the rapid increase of text in digital form. Over the past three decades, efforts have been made to approach this task using various learning algorithms and statistical models based on bag-of-words (BOW) features. Despite its simple implementation, BOW features lack semantic meaning representation. To solve this problem, neural networks started to be employed to learn word vectors, such as the word2vec. Word2vec embeds word semantic structure into vectors, where the angle between vectors indicates the meaningful similarity between words. To measure the similarity between texts, we propose the novel concept of word subspace, which can represent the intrinsic variability of features in a set of word vectors. Through this concept, it is possible to model text from word vectors while holding semantic information. To incorporate the word frequency directly in the subspace model, we further extend the word subspace to the term-frequency (TF) weighted word subspace. Based on these new concepts, text classification can be performed under the mutual subspace method (MSM) framework. The validity of our modeling is shown through experiments on the Reuters text database, comparing the results to various state-of-art algorithms.  
 **main\_author: Kazuhiro Fukui**  
 **published\_date: 2018-06-08 12:55:37+00:00**  
 *pdf\_url: http://arxiv.org/pdf/1806.03125v1*