Distributed Representations of Words and Phrases and their Compositionality

Summary

Thomas Mikolov et. al., in "Efficient Estimation of Word Representations in Vector Space" and in this paper proposed a neural-network architecture to create low-dimensional word embeddings in an unsupervised fashion. In this paper, the authors provided extensions on the recently introduced Skip – gram models that improved the quality of the vectors as well as the training speed. They applied hierarchical softmax on the output, using a binary Huffman tree, to the word2vec hypothesis (skip-gram model) and showed considerable improvements in speedup and quality of the embeddings generated. Mikolov et. al. proposed a new technique called Negative Sampling, which acts as an alternative to hierarchical softmax approach, inspired by the Noise Contrastive Estimation (NCE) technique introduced earlier in literature. The main difference in their model with respect to NCE was that unlike NCE, which required both samples and numerical probabilities of the noise distribution, negative sampling uses only the samples. The authors further proposed another technique to counter the imbalance between rare and frequent words. For example, although the cooccurrences of the words "France" and "Paris" led to new information being learnt by the Skip gram model; the frequent cooccurrences of the "France" and "the" did not lead to any valuable information.

Comparing the results of each of the approaches described above, Mikolov et. al. showed that negative sampling outperforms the Herarchical Softmax on the analogical reasoning task, and even performed better that NCE. They claim that the subsampling of the frequent words improves the training speed by multiple orders of magnitude and also makes the word representations significantly more accurate.

The paper also explains the importance of modelling/embedding phrases rather than individual words. For example, "New York Times" and "Toronto Maple Leafs" are considered a single word and replaced by unique tokens, while other bigrams like "this is" remain unchanged. It also showed contrasting results, where Hierarchical Subsampling outperformed the proposed Negative Sampling approach, when the corpus the frequent words were downsampled.

The paper concluded by stating that hyperparameter tuning like model architecture, the size of the vectors, the subsampling rate and the size of the training window had a significant effect on the performance of the word2vec skip-gram model. The entire project has been open-sourced by the authors and is available online to be used as a "plug and play" tool.

The problem:

The paper proposed new methods (Negative Sampling with frequent word downsampling) as well as applied other approaches (like Hierarchical Softmax) to the existing algorithms. The methods were useful as it could really harness the power of the word2vec hypothesis by reducing the time to train and create low dimensional word embeddings.

Solution:

The solutions to the problem were elegant and reasonable, and significantly cut down on the training speed and even improved the accuracy of the word2vec model.

Evaluation:

The paper used a large dataset consisting of various news articles (an internal Google dataset with one billion words) to train the Skip-gram models. The performance of various skip-gram models are reported using the word-analogy test set hosted by Google (freely available). The evaluation tasks included finding syntactic as well as sematic accuracy, using different methods proposed in the paper. One of the major limitation is that the dataset used to train the models are not publicly available and thus testing/validating the results is difficult.

Related Work:

The related work is the field of NLP is extensive, adequate and the paper cites great articles to get started from. It starts by redefining the Skip-gram model proposed in a previous paper and explaining the alternative approaches briefly before proceeding to the novel methods. The paper can be more explanatory is this domain, by having more figures and explanation on the related work either by the same or any different author.

Writing:

The paper writing is clear and easy to follow and is structurally reasonable.

Take-home messages:

1. Word2Vec is a single layer feedforward neural network architecture with one hidden layer which acts as linear activation function.
2. The loss function proposed is cross-entropy
3. Speedup of 50x and greater can be achieved by using either Hierarchical Softmax or negative sampling.
4. The concept of skip-gram models in Word2vec can be extended to phrases to capture information from a phrase, that would have otherwise been lost in a simple word embedding

Alternative Solutions:

There exist many alternative solutions to word2vec. The most popular one is GloVe: Global Vectors for Word Representations which is a count-based model, unlike word2vec, which is predictive model. There also exist alternate solutions to cut the training speed in word2vec like Hierarchical Sampling, Noise Contrastive Enhancement (NCE) etc.

Open Questions:

1. How does word2vec create linguistic analogies using vector arithmetic?
2. Why does Hierarchical Sampling outperforms Negative Sampling in some cases, but performs poorly in others?

Future Work:

Word Embeddings are very vulnerable to hyperparameter tuning making the model not as robust as desirable. It would be interesting to make the model more robust to parameter tuning.

Deep Walk: Online Learning of Social Representations

Summary

The paper proposes a novel approach to learn social representation, i.e. latent features which capture neighborhood similarity and community membership, of a graph's vertices by modelling a stream of short random walks. DeepWalk tackles the problem of data sparsity in graphs to learn latent representations of features that can be used for real-world applications such as network classification and anomaly detection. It is an unsupervised feature learning technique, that uses the advancements make in the field of natural language processing, into network analysis for the first time. The algorithm generalizes neural language models to process a new language generated entirely by random walks.

The paper demonstrates the scalability of DeepWalk by building representation of web-scale graphs, such as Flickr, YouTube etc. using a parallel implementation. Using the power-law connection between a scale-free graph and word frequency in natural language (wherein both vertex frequency in short random walks and word frequency both have a power law distribution), the paper uses word2vec to create embeddings for the nodes appearing in truncated random walks.

The authors claim that DeepWalk outperforms other latent representation methods for creating social dimensions. They show extensive results showing the accuracy of DeepWalk with other baseline methods on large web-scale graphs.

The Problem:

The problem to make a scalable algorithm to capture network topology is very useful as it can be utilized across all graphs, for different applications. Specifically, the paper targets feature learning in sparse graphs, which is a useful problem as all real-world graphs are sparse and are continuously evolving.

The Solution:

It proposes a new approach for the problem of relational classification (also called as the collective classification). The solution is a novel and excellent idea to model the advancements in nlp and utilize it to generate low-dimensional vertex embeddings in graphs. It has a significant impact, leading to an online scalable algorithm that outperforms all existing methods when dealing with sparse graphs. One of the disadvantages with this solution is that it loses the topological structure in the latent space. It lacks a clear objective function to articulate how to preserve the network structure and is prone to preserve only the second order complexity.

Evaluation:

The authors use multiple real word graphs to present results. The use, BlogCatalog – network of social relationships provided by blogger authors, Flickr – network of contacts between users of the photo sharing website, and YouTube – network between users of the popular video sharing website. The performance of the algorithm was validated against the following baseline methods: SpectralClustering, Modularity, EdgeClusterm, wvRN and Majority.

Related Work:

The paper mentions related work adequately, discussing both the random walk approach and the language modelling approach. However, the authors do not discuss about other work in the literature similar to theirs.

Writing:

The writing of the paper is clear and structurally reasonable. It is easy to understand and follow the paper.

Take Home Messages:

The key take away message from this paper is that language modeling techniques, although in an entirely different domain, can be used for creating a low-dimensional embeddings of nodes in a network and thus learn to an online learning of network representations.

Alternative Solutions:

Other network embeddings solutions exist like Node2Vec, LINE etc. The idea is to generate random walks in a different manner. LINE on the other hand, targets to find the first and second order proximity between the nodes determined by the observed links.

Open Questions Left:

1. How do we preserve both the local and the global structure of the graph?
2. How to handle the noises introduced due to randomness (for vertices with high degrees) when using random walks to enrich the neighboring vertices?

Future Work:

Design a new random walk to preserve both the local and the global structure of the graph, and retain the topological structure in the latent space. This could be done by doing some sort of discriminative learning approaches.