Discriminative Deep Random Walk for Network Classification

Juzheng Li, Jun Zhu, Bo Zhang

Dept. of Comp. Sci. & Tech., State Key Lab of Intell. Tech. & Sys. 
Tsinghua University, Beijing, 100084, China

Reporter: Juzheng Li
2016.6.17
Outline

• Motivation
• Algorithm
• Experiments
• Conclusion
Motivation

Discriminative Deep Random Walk (DDRW)
Motivation

- A large amount of the linguistic materials present a network structure.
• A large amount of the linguistic materials present a network structure.
Motivation

- A large amount of the linguistic materials present a network structure.

- One common challenge of statistical network models is to deal with the sparsity of networks.
• Learn a latent space representation
  (Hoff et al., 2002; Tang and Liu, 2011; Zhu, 2012; Perozzi et al., 2014; Tang et al., 2015)
Motivation

DeepWalk (Perozzi et al., 2014)
DeepWalk (Perozzi et al., 2014)

• Capture entity features like neighborhood similarity and represents them by Euclidean distances
DeepWalk (Perozzi et al., 2014)

- Capture entity features like neighborhood similarity and represents them by Euclidean distances
- Separate embedding and classification
Motivation

DeepWalk (Perozzi et al., 2014)

Karate Graph (Macskassy and Provost, 1977) and DeepWalk embedding
Motivation

DeepWalk Implementation
DeepWalk Implementation

Motivation

Power-law distribution of vertices and words

YouTube Social Graph  Wikipedia Article Text
DeepWalk Implementation

a) Truncated Random Walks
DeepWalk Implementation

a) Truncated Random Walks

b) Word Embedding using Word2vec (Mikolov et al., 2013)
Motivation

DeepWalk Implementation

a) Truncated Random Walks

b) Word Embedding using Word2vec  
(Mikolov et al., 2013)

c) Linear Classifier
Comments on DeepWalk
Comments on DeepWalk

Advantages:
• Effective on learning embeddings of the topological structure
• Find common attributes between networks and natural language and introduce NLP methods to solve the problem
Comments on DeepWalk

Advantages:
• Effective on learning embeddings of the topological structure
• Find common attributes between networks and natural language and introduce NLP methods to solve the problem

Disadvantages:
• Can be suboptimal as it lacks a mechanism to optimize the objective of the target task
Discriminative Deep Random Walk
Discriminative Deep Random Walk

- A novel method for relational network classification
- Jointly optimize representation and discriminative objectives
- Outperform baseline methods on multi-label network classification tasks
- Retain the topological structure in the latent space
• Random Walk

\[
\begin{align*}
W_i & : \ldots 4 \ 16 \ 18 \ 3 \ 5 \ldots \\
W_{i+1} & : \ldots 16 \ 12 \ 11 \ 5 \ 9 \ 18 \ldots \\
\ldots
\end{align*}
\]
Algorithm

- **Word2Vec** (Mikolov et al., 2013)

\[
L_T(\theta, \alpha) = -\sum_{i=1}^{\tau} \frac{1}{s} \sum_{t=1}^{s} \sum_{-R \leq j \leq R, j \neq 0} \log p(\omega_{i,t+j} | \omega_{i,j})
\]

- **Skip-gram**

\[
p(\omega_0 | \omega_I) = \frac{\exp(\theta_T^I \hat{\omega}_I)}{\sum_{i=1}^{|V|} \exp(\theta_T^I \hat{\omega}_I)}
\]

- Hierarchical Softmax: the Huffman binary tree is employed as an alternative representation for the vocabulary.
• **L2-regularized and L2-loss SVC** (Fan et al., 2008)

\[
\mathcal{L}_c(\theta, \beta, y) = C \sum_{i=1}^{|V|} (\sigma(1 - y_i \beta^T \theta_i))^2 + \frac{1}{2} \beta^T \beta
\]
Joint Learning

\[
\mathcal{L}(\theta, \beta, \alpha, y) = \eta \mathcal{L}_r(\theta, \alpha) + \mathcal{L}_c(\theta, \beta, y)
\]

In each SGD step,

\[
\theta \leftarrow \theta - \delta \frac{\partial \mathcal{L}}{\partial \theta} = \theta - \delta (\eta \frac{\partial \mathcal{L}_r}{\partial \theta} + \frac{\partial \mathcal{L}_c}{\partial \theta}),
\]

\[
\beta \leftarrow \beta - \delta \frac{\partial \mathcal{L}}{\partial \beta} = \beta - \delta \frac{\partial \mathcal{L}_c}{\partial \beta},
\]
Experiments

Experimental Setup
Experiments

• Database

a) BlogCatalog: a network of social relationships provided by blog authors. The labels of this graph are the topics specified by the uploading users.

b) Flickr: a network of the contacts between users of the Flickr photo sharing website. The labels of this graph represent the interests of users towards certain categories of photos.

c) YouTube: a network between users of the YouTube video sharing website. The labels stand for the groups of the users interested in different types of videos.
Experiments

- **Baselines**
  - LINE (Tang et al., 2015)
  - DeepWalk (Perozzi et al., 2014)
  - SpectralClustering (Tang and Liu, 2011)
  - EdgeCluster (Tang and Liu, 2009)
  - Majority
Experimental Results
Experiments

- Classification Task
• **Classification Task**
Experiments

- Classification Task
Experiments

- Parameter Sensitivity

(a) BlogCatalog, $\eta$

(b) Flickr, $\eta$

(c) BlogCatalog, $d$

(d) Flickr, $d$
Experiments

- Representation Task

Top-K adjacency predict accuracy(%) in BlogCatalog
Conclusion
Conclusion

- By simultaneously optimizing embedding and classification objectives, DDRW gains significantly better performances in network classification tasks than baseline methods.
Conclusion

- By simultaneously optimizing embedding and classification objectives, DDRW gains significantly better performances in network classification tasks than baseline methods.
- Experiments on different real-world datasets represent adequate stability of DDRW.
Conclusion

• By simultaneously optimizing embedding and classification objectives, DDRW gains significantly better performances in network classification tasks than baseline methods.
• Experiments on different real-world datasets represent adequate stability of DDRW.
• The representations produced by DDRW is both an intermediate variable and a by-product.
Conclusion

• By simultaneously optimizing embedding and classification objectives, DDRW gains significantly better performances in network classification tasks than baseline methods.

• Experiments on different real-world datasets represent adequate stability of DDRW.

• The representations produced by DDRW is both an intermediate variable and a by-product.

• DDRW is also naturally an online algorithm and thus easy to parallel.
Future Work

• Introducing semi-supervised learning
Conclusion

Future Work

• Introducing semi-supervised learning

• A better form of random walk
Reference

[1] Peter D. Hoff, Adrian E. Raftery, and Mark S. Handcock. 2002. Latent space approaches to social network analysis. Journal of the American Statistical Association, 97:1090–1098.
[2] Lei Tang and Huan Liu. 2011. Leveraging social media networks for classification. Data Mining and Knowledge Discovery, 23:447–478.
[3] Jun Zhu. 2012. Max-margin nonparametric latent feature models for link prediction. In Proceedings of the 29th International Conference on Machine Learning, pages 719–726.
[4] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. DeepWalk: online learning of social representations. In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 701–710.
[5] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. LINE: Large-scale information network embedding. In Proceedings of the 24th International Conference on World Wide Web, pages 1067–1077.
[6] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems, pages 3111–3119.
[7] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research, 9:1871–1874.
[8] Lei Tang and Huan Liu. 2009. Scalable learning of collective behavior based on sparse social dimensions. In Proceedings of the 18th ACM Conference on Information and Knowledge Management, pages 1107–1116.
Thank You!