Using Convolutional Embeddings of Large Graphs to Improve Category Classification for Amazon Items

Fengjiao Lyu (fengjiao@stanford.edu), **Joseph Lee** (wejlee@stanford.edu), **Yaqing Li** (yaqing2@stanford.edu)

Overview

In recent years, huge advances have been made in both NLP techniques as well as representation learning on graphs. However, to our knowledge, these two fields remain largely distinct. We thus aim to synergize these two fields by using the hidden states from NLP techniques (CNN, RNN (with GRU/LSTM)) as node embeddings for our inductive graph techniques (GraphSAGE). On our Amazon dataset, we were able to predict which category an item is classified under (node classification) with an accuracy of 0.410. This task of predicting categories is significant as it can suggest to users likely categories for new items listed on Amazon.com. In addition, we experimented with non-uniform sampling which has improved the existing GraphSAGE algorithm on its reported baseline.

Dataset Overview

We tested our hypothesis on items listed on Amazon.com, extracted by SNAP. Our dataset consists of 548,552 products (393,561 books, 19,828 DVDs, 103,114 music CDs and 26,132 videos) as well as the co-purchasing network (products that are frequently co-purchased with each other based on the "Customers who Bought this Item Also Bought…" feature). After pre-processing the data, there are 127 different categories an item can belong to.

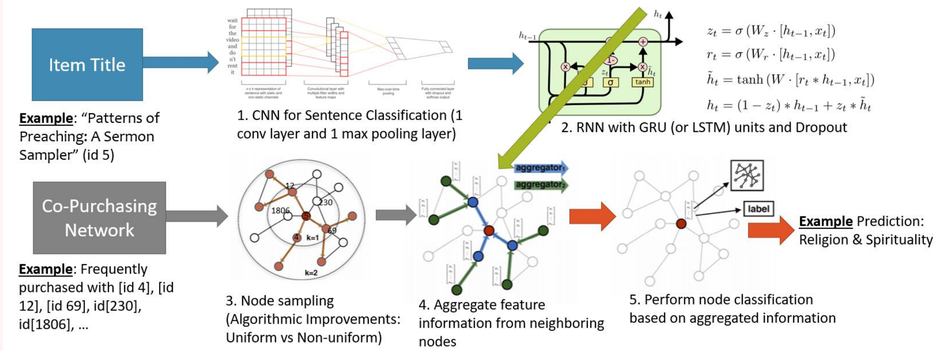
Features

Our two main features are the title of the item and the co-purchasing network. A title often contains some information about its category (e.g. "Patterns of Preaching: A Sermon Sampler" can be categorized as "Religion and Spirituality"). Nevertheless, this is still a very difficult problem – without context, it's difficult to know "Irish Stew!" should be classified under "Literature and Fiction". Performing graph algorithms on a co-purchasing network might thus exploit information about similar items and thus provide the needed context to classify such titles.

Methodology

Our model

We combine our two data features by using the hidden state after the RNN (GRU/LSTM) step from the Item Title as the node embedding for the GraphSAGE algorithm which is performed on the co-purchasing network, as shown below:



Sentence Classification with CNN / RNN (GRU/LSTM)

CNNs has widely been used for sentence classifications where the temporal ordering of words do not matter (Kim, 2014). Combining this with RNN (GRU/LSTM) represent the cutting-edge techniques for text classification problems, and have shown state-of-the-art results in other text classification problems.

Node Classification with GraphSAGE

We used GraphSAGE (Hamilton et al, 2017) largely due to its inductive representation, which allows us to generate node embeddings for unseen data. This is crucial for the Amazon dataset since an inductive algorithm allows us to generalize and classify new items, which are "unseen nodes" in the graph.

Algorithmic Improvements for GraphSAGE

The current GraphSAGE model uniformly samples a fixed-size set of neighbors. However, non-uniform sampling methods might utilize more information in the graph by choosing the most "valuable" neighbors to sample. In particular, we experiment by sampling based on the neighbor's node degree.

Results

Baseline Comparisons on Amazon Dataset

We took a 0.6 – 0.2 – 0.2 train-dev-test split for GraphSAGE, and achieved the following results on our models:

Methodology	Test Accuracy
Baseline: CNN, RNN (LSTM)	0.233
Baseline: CNN, RNN (GRU)	0.267
GraphSAGE (Uniform Sampling)	0.427
GraphSAGE (Non-Uniform Sampling)	0.428
CNN, RNN (LSTM) + GraphSAGE (Uniform Sampling)	0.405
CNN, RNN (GRU) + GraphSAGE (Uniform Sampling)	0.411
CNN, RNN (LSTM) + GraphSAGE (Non-Uniform Sampling)	0.409
CNN, RNN (GRU) + GraphSAGE (Non-Uniform Sampling)	0.410
Improvement over baseline	54% - 76%

Other Comparisons / Experiments

Non-Uniform sampling on Reddit Dataset

(Dataset used in GraphSAGE Original Paper)

Sampling Methods	Uniform Sampling	Highest degree Sampling	Highest degree + Random	Lowest degree Sampling	Lowest degree + Random
F1 Score	0.947	0.919	0.943	0.946	**0.950**

Naïve Bayes on Amazon Dataset

Surprisingly, the Naïve Bayes algorithm boasts an accuracy of 0.362 on the Amazon dataset (Top 5 Accuracy: 0.546), which outperforms some of the complex algorithms used.

Discussion

Our new model on Amazon Dataset: Comparing with our baseline of CNN, RNN (GRU/LSTM), incorporating the co-purchasing information adds a 54% - 76% increase in accuracy, suggesting that the co-purchasing network indeed provides the necessary context for item titles. However, there is little improvement in using the hidden states over Word2Vec as node embeddings, which is likely due to the short titles of the Amazon items. We hypothesize that this method might prove more useful for document classification rather than for titles.

Non-Uniform Sampling: The intuition behind lowest-degree sampling is that the neighbors with lower degree are more valuable for node classification since they are less likely to be "general" items that can be linked to many different classes of nodes. However, while this provided some improvements in the Reddit dataset, we found the Amazon dataset's degree had too little variance for this to be a huge factor. Overall, we have still not achieved extremely high levels of accuracy. Upon performing Error Analysis, we found that Amazon items often lend itself to multiple categories (out of 127 possible categories). For example, our NLP classification predicted "On Ethics and Economics" as "Social Sciences" when the ground truth label was "Reference". Further work would include cutting down to 20 general categories.

Conclusion

The focus of our project was to introduce a new way of synthesizing NLP techniques with state-of-the-art inductive graph techniques. Future work would include trying these techniques on other datasets to analyse if this method is suited for certain types of data, as well as experimenting with different NLP techniques (RCNN, Hierarchical Attention Networks) and other graph algorithms (Graph Attention Networks by Veličković et al (2017)). Nevertheless, our current algorithm has the potential to suggest to users likely categories for new items listed, saving time for sellers on Amazon.

References: [1]Jure Leskovec and Andrej Krevl (Jun 2014), SNAP Datasets: Stanford Large Network Dataset Collection, http://snap.stanford.edu/data [2]J. Leskovec, L. Adamic and B. Adamic. The Dynamics of Viral Marketing. ACM Transactions on the Web (ACM TWEB), 1(1), 2007. [3]Code for CNN/RNN(GRU/LSTM) largely adapted from: https://github.com/jiegzhan/multi-class-text-classification-cnn-rnn [4]Inductive Representation Learning on Large Graphs. W.L. Hamilton, R. Ying, and J. Leskovec arXiv:1706.02216 [cs.SI], 2017.