GePE: Generalizable Paper Embedding with Language-Driven Biased Random Walk

Kailing Wang [*†] Weiji Xie [*] Xiangyuan Xue [*]

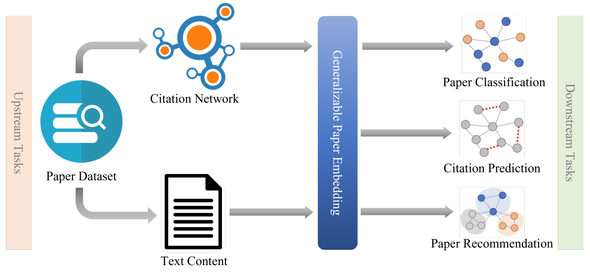
[*]Equal Contribution [†]Project Leader

Motivation: Incorporating Textual Information in Paper Embedding

- Today, especially in the domain of computer science, almost all the research papers are freely available from arXiv, which allows us to exploit the rich information in the citation network.
- Learning representation of papers is crucial for downstream tasks such as paper classification and recommendation, which helps researchers keep up with the latest research progress.
- Although data mining on citation networks has been well studied, most existing methods only consider the structure information while ignoring the textual information in the papers [2], which limits the performance and hinders the embeddings from generalizing to unseen papers.
- With the development of language models, it is possible to encode the textual information in the papers and incorporate it into the representation learning process on the graph.

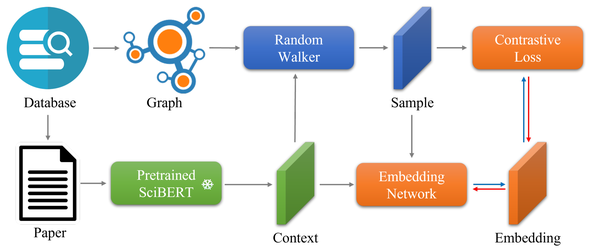
Pipeline: Extending from Upstream Tasks to Downstream Tasks

- We propose a **Ge**neralizable **P**aper **E**mbedding (GePE) which learns from both the structure information in the citation network and the textual information in the paper content.
- The paper embedding is learned as an upstream task, which can be applied to downstream tasks, such as paper classification, citation prediction, and paper recommendation.



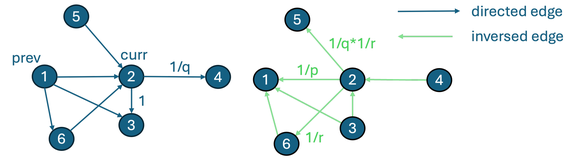
Method: Combining Language Encoder with Biased Random Walk

- The title and abstract of the papers are encoded by a pretrained SciBERT [1] model.
- The embedding network converts the context vectors into corresponding paper embeddings.
- The contrastive loss of biased random walk is used to train the embedding network.



Implementation: Reversible Biased Random Walk on OGBN-arXiv

- The OGBN-arXiv [4] dataset consists of a directed graph including 169343 paper nodes from 1971 to 2020 and 1166243 citation edges, which is used for training and evaluation.
- Although a citation describes a unidirectional relationship, we adapt the random walk [3] process to be reversible so that the algorithm can better represent the link information.
- We introduce the inverse ratio r as another hyperparameter, so that an additional probability factor of $\frac{1}{r}$ will be assigned when traversing in the opposite direction of the citation edge.



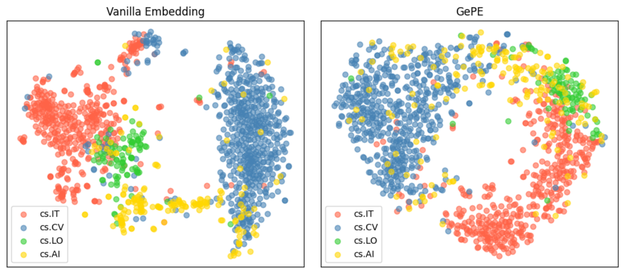
Evaluation: Performance of Node Classification and Link Prediction

- We compare our methods with Hash Mapping, Vanilla Embedding, and Language Encoding on two downstream tasks: Node Classification (NC) and Link Prediction (LP).
  - Hash Mapping: An MLP which hashes the node ID and converts it to an embedding.
  - Vanilla Embedding: The classic learnable embedding as used in Node2Vec.
  - Language Encoding: A BERT which directly encodes the title and abstract into embedding.
- For node classification, we train another classifier on embeddings to compute the ACC. For link prediction, we use inner product as the predicted score to compute the AUC.

Method	# Parameters	Generalizable	NC (ACC)	LP (AUC)
Hash Mapping	3.7M	No	9.8%	0.558
Vanilla Embedding	130.0M	No	60.5%	0.934
Language Encoding	30.5M	Yes	26.9%	0.733
GePE (ours)	23.1M	Yes	68.3%	0.859

Visualization: Embedding Space of Papers with Specific Classes

- We manually select 4 classes to visualize the embedding space of the papers by t-SNE.
- GePE can better separate the papers of different classes compared to Vanilla Embedding.



Recommendation: Balancing Content Relevance and Paper Quality

- We use nearest neighbor algorithm based on cosine distance for paper recommendation.
- To recommend k papers from the dataset, we first fetch λk papers by KNN and then select the top-k most cited papers, so that content relevance and paper quality are balanced.

User Specification

Attention is all you need cs.CL 2017
... transduction models are based on complex recurrent or convolutional neural networks in an encoder-decoder configuration. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms ...

System Recommendation

Data Recombination for Neural Semantic Parsing cs.CL 2016
... captures important conditional independence properties commonly found in semantic parsing. We then train a sequence-to-sequence recurrent network (RNN) model with a novel attention-based copying mechanism on data-points sampled from this grammar, thereby teaching the model about these structural properties ...

Neural language correction with character based attention cs.CL 2016
... Motivated by these issues, we present a neural network-based approach to language correction. The core component of our method is an encoder-decoder recurrent neural network with an attention mechanism. By operating at the character level, the network avoids the problem of out-of-vocabulary words ...

Adding interpretable attention to neural translation models improves word alignment cs.CL 2019
... Multi-layer models with multiple attention heads per layer provide superior translation quality compared to simpler and shallower models, but determining what source context is most relevant to each target word is more challenging as a result. Therefore, deriving high-accuracy word alignments from the activations of a state-of-the-art ...

Conclusion: Valuable Findings and Critical Insights

- GePE integrates textual information from paper content and structural information from citation network, thus enhancing the quality and generalizability of paper embeddings. Compared to traditional methods, language-driven biased random walk can better capture semantic relationships between papers, thus offering deeper academic insights.
- Our experiments indicate that GePE can help improve the performance of paper classification and citation prediction, which significantly outperforms traditional methods. The number of parameters is also significantly reduced and will not increase with the size of dataset.
- Both of the language encoder and random walker can be highly parallelized, so it is possible to scale up our method to much larger datasets. With its strong generalization ability, GePE can be directly applied to unseen papers without any retraining or finetuning.

References

[1] Iz Beltagy, Kyle Lo, and Arman Cohan.
Scibert: A pretrained language model for scientific text.
arXiv preprint arXiv:1903.10676, 2019.

[2] Soumyajit Ganguly and Vikram Pudi.
Paper2vec: Combining graph and text information for scientific paper representation.
In *Advances in Information Retrieval: 39th European Conference on IR Research, ECIR 2017, Aberdeen, UK, April 8-13, 2017, Proceedings 39*, pages 383–395. Springer, 2017.

[3] Aditya Grover and Jure Leskovec.
node2vec: Scalable feature learning for networks.
In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864, 2016.

[4] Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec.
Open graph benchmark: Datasets for machine learning on graphs.
Advances in neural information processing systems, 33:22118–22133, 2020.