Learning Contextual Representations of Citations via Graph Transformer *

Hyeon-Ju Jeon[1,2][0000−0002−2400−8360], Gyu-Sik Choi[3], Se-Young Cho[3], Hanbin Lee[4], Hee Yeon Ko[5], Jason J. Jung[1,**][0000−0003−0050−7445], O-Joun Lee[6][0000−0001−8921−5443], and Myeong-Yeon Yi[7]

[1] Chung-Ang University, Dongjak-gu, Seoul, Korea
{hyeonju,j3ung}@cau.ac.kr
[2] Korea Institute of Atmospheric Prediction Systems, Dongjak-gu, Seoul, Korea
hjjeon@kiaps.org
[3] Sogang University, Mapo-gu, Seoul, Korea
{gyusik19,wnl383}@sogang.ac.kr
[4] Incheon National University, Yeonsu-gu, Incheon, Korea
gksqls@inu.ac.kr
[5] Soongsil University, Dongjak-gu, Seoul, Korea
0525ojkmt@soongsil.ac.kr
[6] Catholic University of Korea, Bucheon-si, Gyeonggi-do, Korea
ojlee@catholic.ac.kr
[7] NAVER Corp., Seongnam-si, Gyeonggi-do, Korea
myeongyeon.yi@navercorp.com

Abstract. This study aims at representing the citation based on the citation context extracted from the citation network. Researchers cite papers for various purposes to describe their arguments in a logical structure. Thus, citations have different roles depending on what structure they are cited in the paper. In this paper, we first present a definition of the citation context and initialize the embedding vector based on the citation order and location. Then, based on the graph transformer model, we learn contextual citation embeddings. To represent citation context, we consider the following three parts: (i) textual features of paper, (ii) positional features of the citation context, and (iii) structural features of the citation network by applying the self-attention mechanism.

Keywords: Citation Context · Citation Network · Graph Transformer · Network Embedding · Positional Embedding

** Corresponding Author.

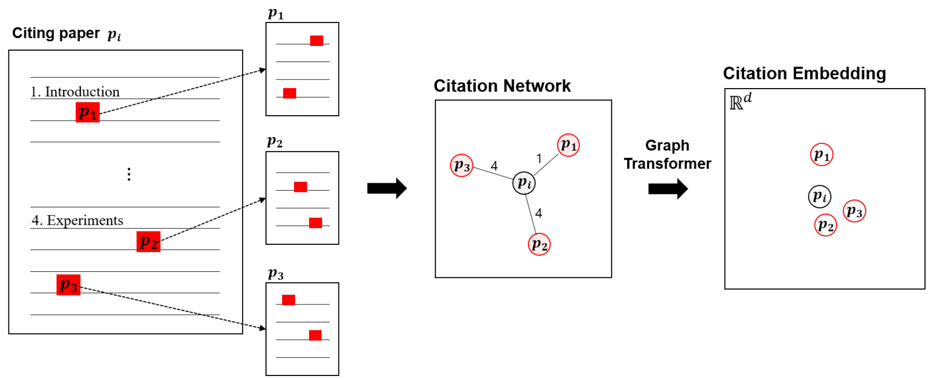

Fig. 1: Illustration of the citation embedding. There are citing paper p_i and cited papers p_1, p_2, p_3. In the citation network, each node represents a paper. The citation embedding refers to the context based embedding of citation. If the citation location is different, the distinguish meaning is reflected in the citation embedding, even if cited in the same paper.

1 Introduction

The exponentially increment academic papers cause various services (e.g., citation recommendation [3, 7, 13], bibliographical retrieval [15], and so on). Such services need exquisite analysis of the scientific impact and content of papers [9].

There have been various studies [14, 16] on citation analysis to assess the quality of the paper and understand the context. These studies have mostly applied citation frequency-based and content-based approaches. The frequency-based approaches was only given the same weight regardless of the purposes of citation. As shown in Fig. 1, when two papers p_1 and p_2 are cited by p_i, suppose that p_1 is located in introduction section, and p_2 is located in evaluation section. In this case, p_1 and p_2 are cited for different purposes, and their importance is also different.

To solve this problem, it is necessary to understand the overall context of the citation in the paper. The content-based approaches [6, 19] attempted to learn the contextual features of the paper using a language model based on RNN/LSTM. Nevertheless, these studies only concentrated on not discovering a citation context or their roles but measuring contents similarity between two papers.

Thereby, in this paper, we define and extract the citation context in citation networks. First of all, we assume that the cited papers compose the contents of the citing paper, and the order and location of the cited papers reflect the role of each paper in the citing paper. To represent citations, we propose an embedding method considering (i) textual features of paper, (ii) positional features of the citation context, and (iii) structural features of the citation network by applying the self-attention mechanism [18]. The proposed method can represent global citation features using fewer layers than the convectional GCN model. It is also efficient to learn the context of long papers.

Finally, based on the graph-transformer [20], the proposed method generates pre-training citation vectors considering the influence and correlation between citation papers. This result can be used in various tasks such as citation classification, research topic discovery, and paper evaluation in the future.

2 Related work

This section introduces the existing methods for analyzing the citation relationship in the citation network. To deal with the large citation network, various studies investigated the co-citation frequency.

Boyack and Klavans [2] focused on the network theory which can measure node importance and weight to analyze co-citation relationship and bibliographic coupling. Although this approach reflects the feature of network structure level, it is difficult to say that the different roles of citations are considered. To solve this problem, Habib and Afzal [5] exploited the distribution of citations in sections to capture the citation context. Nevertheless, it is necessary to analyze the distinguishing characteristics of co-citation papers at the content level. The proximity based methods [4, 12] was proposed for weighting edges of the co-citation network by using contexts. The edge weight was based on the strength of co-citation context in the sentence level. Also, Ahmad and Afzal [1] showed that traditional co-citation analysis can produce better results when combined with metadata information of the paper (e.g., author, affiliation, venue, and so on.)

The above approaches focused on comparing content-based similarities in consideration of the relationship between cited papers. While these are effective for application to specific tasks such as citation recommendations and searches, it is difficult to generate widely used representation by unsupervised learning. Thereby, a few studies conducted network representation learning [11] for embedding the paper node based on the citation context in network structure level. VOPRec [10] learned vector representation of paper by combining text information with structural identity in the citation network. DocCit2Vec [21] which represents paper based on the citation context at the document level is used for the recommendation system by applying the attention mechanism.

However, it is difficult to consider the contextual features reflected in the structure of papers. From this perspectives, we extract the context of a citation through citation networks constructed according to the citation section. After that, initial embedding is performed considering the network structure and textual features so that the transformer model can learn various features of citations.

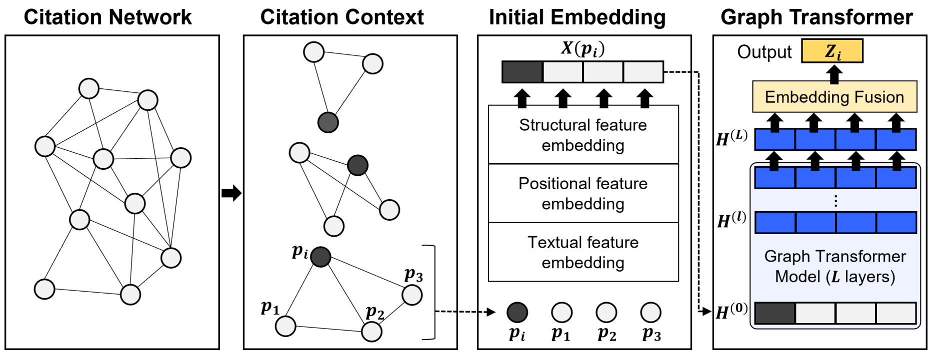

Fig. 2: Architecture of the contextual citation embedding model.

3 Learning representation of citation context

In this section, we will introduce the detailed approach about the contextual citation embedding model. As illustrated in Fig. 2, the model composes three components: (1) extracting citation context, (2) initializing the citation embedding, (3) graph-transformer based encoder. Therefore, the graph transformer model learns a representation a target citation by fusing the input initial embedding vectors. To extract the context of the citation in the first component, we define our citation network as follows.

Definition 1 (Citation Network). *The citation networks* ($\mathcal{N}$) *contains paper node* ($\mathbb{P}$). *There are citation relationship* ($\mathcal{C} \in \mathbb{R}^{|\mathbb{P}|\times|\mathbb{P}|}$) *between paper nodes. When paper* p_i *cites paper* p_j *in the* n^{th} *section, the citation relationship has weights* ($w \in \{0, \cdots, n, \cdots, N\}$). *This can be formulated as follows:*

$$\mathcal{N} = \langle \mathbb{P}, \mathcal{C}, w, \mathbf{t} \rangle, \tag{1}$$

where $\mathbf{t}$ *refers to a textual feature vector of* $\mathbb{P}$.

To consider the different compositions of the sections of the paper, we rearrange the paper into four sections from 0 to 3: 0 represents an introduction, 1 represents a related task, 2 represents a methodology, and 3 represents a result. In this case, the maximum section number is 3. Also, for the text features of each paper p_i, different word embedding models can be used.

3.1 Extracting citation context

Instead of working on the entire citation network $\mathcal{N}$, we extract the citation context from the citation network. The existing network embedding method uses a node sampling approach that is weighted according to the importance of the node. However, since the importance of cited papers is determined by the purpose of citations, analysis of the purpose and characteristics of each cited paper is necessary.

As stated in Sect. 1, we assume that the citation order and location of the paper relates to the purpose of the citation paper. Thus, we extract various subgraphs for the target paper by sampling the cited paper for each section rather than sampling the entire citation paper. In this section, we define the subgraphs as citation context;

Definition 2 (Citation Context). *Given an input citation network $\mathcal{N}$, for paper p_i in the network, citation context is a set of sampled paper at each section $n \in [0, N]$ This can be formulated as follows:*

$$\Gamma(p_i) = \langle \Gamma(p_{i,0}), \cdots, \Gamma(p_{i,N}) \rangle, \tag{2}$$

where $\Gamma(p_{i,n})$ represents the contextual citation in section n. This can be formulated as follows:

$$\Gamma(p_{i,n}) = \{p_j | p_j \in \mathbb{P} \setminus \{p_i\} \wedge w(i,j) = n\}. \tag{3}$$

To efficiently extract citation context for a batch of papers during the training of the embedding model, we extend a node sampling algorithm to enable node sampling for each section. The sampling method iteratively samples a list of papers for a target paper p_i using adaptive sampling depth K_n by section. Let $\mathcal{S}_{p_i}^{k_n-1}$ refer to the bag of papers sampled at the $(k-1)^{th}$ step in n^{th} section. For each paper node p_i in $\mathcal{S}_{p_i}^{k_n-1}$, we randomly sample cited papers in citation network with replacement from p_i's one-hop neighbors at the k^{th} step. Through this process, the papers in p_i's citation context $\Gamma(p_i)$ can cover both local neighbors of p_i and papers far away.

3.2 Initializing the citation embedding

Based on the citation context concept, we obtain the set of sampled subgraph batches for all the nodes as get $\mathcal{G} = \{g_1, g_2, \cdots, g_{|\mathbb{P}|}\}$, where g_i represents the subgraph sampled for target paper p_i. Different from general graph in which the nodes are orderless, in the paper, cited papers are logically constructed, so the order of citations is meaningful. Therefore, the citation context is serialized in the order cited in the paper. Formally, we concatenate the target paper p_i and its ordered contextual citation g_1, denoted by $\mathcal{I}_{p_i} = [p_i, p_{i,1}, p_{i,2}, \cdots, p_{i,S}]$, where $p_{i,j}$ is the j^{th} node in g_j, and $1 \leq j \leq S$. In this section, we define paper embeddings along the citation order quoted in a paper. The paper embeddings will be the input to the graph-transformer model.

For textual embedding, we can embed textual feature vector $\mathbf{t}_j$ into a shared feature space for each paper $p_j \in \mathcal{I}_{(p_i)}$ in the citation context g_i. Simple fully connected layers can be used for the textual input. This can be formulated as follows:

$$\mathbf{x}_{text}(p_j) = Embedding(\mathbf{t}_j) \in \mathbb{R}^d, \tag{4}$$

where d indicates dimension of the shared feature space.

The position of a paper in the citation context $\mathcal{I}_{p_i}$ reflects the purpose and characteristic of the citation to the target paper p_i. Thus, we suggest that the order of papers in $\mathcal{I}_{p_i}$ is significant in learning citation representations. The following position-id embedding is used to identify the cited paper order information of an input list,

$$\mathbf{x}_{pos}(p_j) = Embedding[p(j)] \in \mathbb{R}^d, \tag{5}$$

where $p(j)$ indicates the position-id of paper p_j in $\mathcal{I}_i$.

Our main objective is to obtain the representation of the target paper p_i based on the structural roles. To identify the role of each paper, we use the embedding method based on Weisfeiler-Lehman (WL) algorithm [17]. This can be formulated as follows:

$$\mathbf{x}_{role}(p_j) = Embedding[r(j)] \in \mathbb{R}^d, \tag{6}$$

where $r(t)$ refers to the role label.

After computing the three terms of embedding, we aggregate them to be the initial input paper embedding of the graph transformer model. The embedding fusion is formalized as follows:

$$\mathbf{x}(p_j) = \mathbf{x}_{text}(p_j) + \mathbf{x}_{pos}(p_j) + \mathbf{x}_{role}(p_j) \in \mathbb{R}^d. \tag{7}$$

We define the embedding fusion function as the summation of three embedding terms.

Finally, given a target paper p_i, we obtain the initial paper embedding of each paper in its substructure cited paper set. The initial paper embedding for the paper in the citation context $\mathcal{I}_{p_i}$ can be stacked to a embedding matrix. The embedding matrix is represented by $\mathbf{X}(p_i) = [x(p_1), x(p_2), \cdots, x(p_S)] \in \mathbb{R}^{S \times d}$.

3.3 Graph-transformer based encoder

The target of the graph-transformer model is to aggregate the initial embedding of each paper and generate a low-dimensional embedding vector for each of paper. A numbers of attention layers are stacked to compose the transformer module. A single layer can be formulated as:

$$\mathbf{H}^{(l)} = attention\big(\mathbf{H}^{(l-1)}\big) = softmax\Big(\frac{\mathbf{Q}^{(l)}\mathbf{K}^{(l)\top}}{\sqrt{d}}\Big)\mathbf{V}^{(l)}, \tag{8}$$

where $\mathbf{H}^{(l)}$ and $\mathbf{H}^{(l-1)}$ denote the output embedding of the l and $(l-1)$ layer, $\mathbf{Q}^{(l)}$, $\mathbf{K}^{(l)}$, and $\mathbf{V}^{(l)}$ are the query matrix, key matrix, and value matrix respectively, and d is the dimension of paper embedding. Specifically, $\mathbf{Q}^{(l)}$, $\mathbf{K}^{(l)}$, and $\mathbf{V}^{(l)}$ are calculated as follows:

$$\begin{cases} \mathbf{Q}^{(l)} = \mathbf{H}^{(l-1)}\mathbf{W}_{\mathbf{Q}}^{(l)}, \\ \mathbf{K}^{(l)} = \mathbf{H}^{(l-1)}\mathbf{W}_{\mathbf{K}}^{(l)}, \\ \mathbf{V}^{(l)} = \mathbf{H}^{(l-1)}\mathbf{W}_{\mathbf{V}}^{(l)}, \end{cases} \tag{9}$$

where $\mathbf{W}_{\mathbf{Q}}^{(l)}$, $\mathbf{W}_{\mathbf{K}}^{(l)}$, and $\mathbf{W}_{\mathbf{V}}^{(l)}$ are the weight matrices of the l^{th} attention layer.

The input of the graph-transformer model $\mathbf{H}^{(0)}$ is denoted as the embedding matrix of the target paper $\mathbf{X}(p_i)$. The output of the last attention layer $\mathbf{H}^{(L)}$ is defined as the output paper embedding matrix $\mathbf{Z}$ of the transformer model.

4 Conclusion and future work

In this paper, we have proposed the learning representation of contextual citation network. We have defined the citation context by sampling a different number of papers per section. Using a graph transformer model, paper vectors were output based on salient citations within the citation context. According to our initial assumption, the results of the embedding model can reflect the role of each citations in the paper.

The citation purpose of a paper can change dynamically [8]. As future work, we can represent the paper with the meaning of citations that change over time. In addition, various bibliographic entities such as high reputed journals and authors affect the citation. If the graph transformer model is extended to heterogeneous networks in the future, rich interactions between bibliographic information are able to analyze. Finally, we intend to examine the proposed embedding model in a large contextual citation network.

Acknowledgements This work was supported by Korea Foundation for Women In Science, Engineering and Technology (WISET) grant funded by the Ministry of Science and ICT(MSIT) under the team research program for female engineering students. (WISET Contract No. 2021-178)

References

1. Ahmad, S., Afzal, M.T.: Combining co-citation and metadata for recommending more related papers. In: 2017 International Conference on Frontiers of Information Technology (FIT 2017). pp. 218–222. IEEE (dec 2017). https://doi.org/10.1109/fit.2017.00046
2. Boyack, K.W., Klavans, R.: Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? Journal of the American Society for Information Science and Technology **61**(12), 2389–2404 (dec 2010). https://doi.org/10.1002/asi.21419
3. Cai, X., Zheng, Y., Yang, L., Dai, T., Guo, L.: Bibliographic network representation based personalized citation recommendation. IEEE Access **7**, 457–467 (Dec 2019). https://doi.org/10.1109/access.2018.2885507
4. Eto, M.: Extended co-citation search: Graph-based document retrieval on a co-citation network containing citation context information. Information Processing & Management **56**(6), 102046 (nov 2019). https://doi.org/10.1016/j.ipm.2019.05.007
5. Habib, R., Afzal, M.T.: Sections-based bibliographic coupling for research paper recommendation. Scientometrics **119**(2), 643–656 (mar 2019). https://doi.org/10.1007/s11192-019-03053-8

6. Huang, W., Kataria, S., Caragea, C., Mitra, P., Giles, C.L., Rokach, L.: Recommending citations: translating papers into references. In: Chen, X., Lebanon, G., Wang, H., Zaki, M.J. (eds.) Proceedings of the 21st ACM international conference on Information and knowledge management (CIKM 2012). pp. 1910–1914. ACM Press, Maui, HI, USA (Oct 2012). https://doi.org/10.1145/2396761.2398542
7. Huang, W., Wu, Z., Liang, C., Mitra, P., Giles, C.L.: A neural probabilistic model for context based citation recommendation. In: Bonet, B., Koenig, S. (eds.) Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI 2015). pp. 2404–2410. AAAI Press, Austin, Texas, USA (Jan 2015)
8. Jeon, H.J., Jung, J.J.: Discovering the role model of authors by embedding research history. Journal of Information Science **0**(0), 01655515211034407 (2021). https://doi.org/10.1177/01655515211034407
9. Jeon, H.J., Lee, O.J., Jung, J.J.: Is performance of scholars correlated to their research collaboration patterns? Frontiers in Big Data **2**(39) (Nov 2019). https://doi.org/10.3389/fdata.2019.00039
10. Kong, X., Mao, M., Wang, W., Liu, J., Xu, B.: VOPRec: Vector representation learning of papers with text information and structural identity for recommendation. IEEE Transactions on Emerging Topics in Computing **9**(1), 226–237 (jan 2021). https://doi.org/10.1109/tetc.2018.2830698
11. Lee, O.J., Jeon, H.J., Jung, J.J.: Learning multi-resolution representations of research patterns in bibliographic networks. Journal of Informetrics **15**(1), 101126 (Feb 2021). https://doi.org/10.1016/j.joi.2020.101126
12. Liu, S., Chen, C.: The proximity of co-citation. Scientometrics **91**(2), 495–511 (dec 2011). https://doi.org/10.1007/s11192-011-0575-7
13. Ma, S., Zhang, C., Liu, X.: A review of citation recommendation: from textual content to enriched context. Scientometrics **122**(3), 1445–1472 (jan 2020). https://doi.org/10.1007/s11192-019-03336-0
14. MacRoberts, M.H., MacRoberts, B.R.: The mismeasure of science: Citation analysis. Journal of the Association for Information Science and Technology **69**(3), 474–482 (nov 2017). https://doi.org/10.1002/asi.23970
15. Raamkumar, A.S., Foo, S., Pang, N.: Using author-specified keywords in building an initial reading list of research papers in scientific paper retrieval and recommender systems. Information Processing & Management **53**(3), 577–594 (May 2017). https://doi.org/10.1016/j.ipm.2016.12.006
16. Roman, M., Shahid, A., Uddin, M.I., Hua, Q., Maqsood, S.: Exploiting contextual word embedding of authorship and title of articles for discovering citation intent classification. Complexity **2021**, 5554874:1–5554874:13 (apr 2021). https://doi.org/10.1155/2021/5554874
17. Shervashidze, N., Schweitzer, P., van Leeuwen, E.J., Mehlhorn, K., Borgwardt, K.M.: Weisfeiler-lehman graph kernels. Journal of Machine Learning Research **12**, 2539–2561 (Sep 2011)
18. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30th Annual Conference on Neural Information Processing Systems (NIPS 2017). pp. 5998–6008. Long Beach, CA, USA (Dec 2017), https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

19. Wang, J., Zhu, L., Dai, T., Wang, Y.: Deep memory network with bi-LSTM for personalized context-aware citation recommendation. Neurocomputing **410**, 103–113 (oct 2020). https://doi.org/10.1016/j.neucom.2020.05.047
20. Zhang, J., Zhang, H., Xia, C., Sun, L.: Graph-bert: Only attention is needed for learning graph representations (2020), https://arxiv.org/abs/2001.05140, abs/2001.05140
21. Zhang, Y., Ma, Q.: Citation recommendations considering content and structural context embedding. In: Lee, W., Chen, L., Moon, Y., Bourgeois, J., Bennis, M., Li, Y., Ha, Y., Kwon, H., Cuzzocrea, A. (eds.) 2020 IEEE International Conference on Big Data and Smart Computing (BigComp 2020). pp. 1–7. IEEE, Busan, Korea (South) (Feb 2020). https://doi.org/10.1109/bigcomp48618.2020.0-109