Research on Dense Enhanced Document Retrieval Based on G-mixup

Jiawei Tang, Junping Liu

*Corresponding author. Email: 2359451809@qq.com

Wuhan Textile University, Wuhan, 430200, China

ABSTRACT: The dense document retrieval model based on Mixup regards words as independent individuals, splits the connection between words, ignores the semantic information of the text, and also has the problem of insufficient labeled training data. In view of the above problems, this paper proposes a G-mixup graph based data intensive enhanced document retrieval model GDAR (Graph Data Augment Retrieval). The model first uses the graph convolutional neural network to convert queries and documents into graph data; then, uses the same type of document graph to construct a graph generator Graphon; finally, mixes the graph generator Graphon in the Euclidean space to obtain The new graph generator Graphons performs linear interpolation and perturbation operations on Graphons to obtain new training data with soft labels, solving the problem of lack of labeled data in dense document retrieval models. Experiments on the Natural Questions and TriviaQA datasets show that the method improves the accuracy index of T-1 by 4.12% and 4.88% respectively compared with the best baseline method.

Keywords: Mixup, Dense Document Retrieval, Graph Convolutional Neural

1.Introduntion

With the rapid development of the Internet, massive document information has been accumulated on the network, how to realize fast and effective document re-trieval in massive document information has become an urgent problem to be solved[1]. Traditional document retrieval techniques rely on term matching, ignoring the semantic information of the text. In order to make better use of the semantic information of text, researchers have conducted in-depth research on dense document retrieval[2].

Dense document retrieval models require a large amount of labeled training data [3]. However, datasets for dense document retrieval such as NQ (Natural Question) datasets, where the amount of labeled data is much smaller than unlabeled data. It is therefore almost impossible to rely on manual annotation of large corpora of docu-ments, which would waste a lot of human effort.

In view of the shortage of marked data in dense document retrieval, data expan-sion and data enhancement are generally used to solve the problem. Data augmenta-tion is often used to expand the number of training samples. Qu et al.[4] proposed to expand the training samples by generating additional query document pairs, and Liang and Ma et al. [5,6] proposed to use the generative model to generate query to expand additional training data. The above data expansion methods need to consume a large amount of computing resources, and the cost is

ICIDC 2023, June 02-04, Nanchang, People's Republic of China

DOI 10.4108/eai.2-6-2023.2334610

high. Data augmentation is to obtain new training data through linear interpolation or perturbation. Wei et al. [7] performed data enhancement by means of word perturbation, and Srivastava et al. [8] performed data enhancement by combining supervised learning with word perturbation. Ding et al. [9] proposed to use graph diffusion perturbation to generate enhanced graphs for data augmentation. Jeong et al. [10] proposed the Mixup data enhancement method to obtain training samples with soft labels through linear interpolation and perturbation, so that the dense document retrieval model achieved excellent performance.

However, the dense document retrieval model based on Mixup has the following problems: First, the model is a vector space model, which treats words as independent individuals in retrieval, splits the connection between words, and ignores the semantic information of the text ;Secondly, Mixup's data enhancement method mixes any two training samples, destroying the structural information of the text; finally, the model does not solve the problem of insufficient labeled training data. To address the above issues, this paper proposes a dense document retrieval method based on G-mixup graph data augmentation. Use graphs to represent documents and queries, retain the semantic information of text, and use G-mixup graph data enhancement method to enhance the data of documents to solve the problem of insufficient labeled training data. The specific process is as follows: the first step, use the graph convolutional neural network GCN (Graph Convolution Neural Network) to map the query data and document data from the non-Euclidean space to the Euclidean space, use the graph to represent the query and document, and obtain the query and document data Vector representation; the second step is to construct a graph genera-tor Graphon by using the document graph belonging to the same category; the third step is to randomly select two graph generators Graphon in the Euclidean space for mixing to generate a new graph generator Graphons, and then perform linear interpolation Int (Interpolation) and perturbation Per (Perturbation) operations to obtain new graph data with soft labels, and solve the problem of insufficient labeled training data for dense document retrieval.

2.Dense Augmented Retrieval Models for Graph Data

The model consists of three parts: vector representation module, graph data en-hancement module, and correlation calculation module.

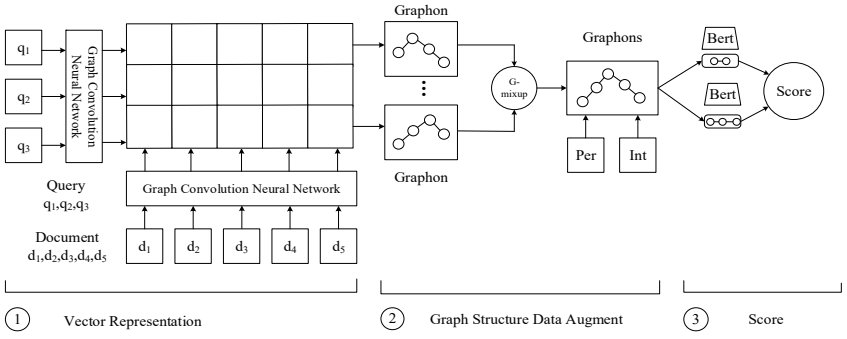

Fig. 1. The overall structure of the model

2.1Graph Data Augmentation And Similarity Score Calculation

Since the graph is irregular, misaligned, non-Euclidean data. Therefore, it is necessary to use graph convolutional neural network to map graph data from non-Euclidean space to Euclidean space to generate vector representations of queries and documents, such as show in formula (1):

$$Q \to W_Q, D \to W_D \tag{1}$$

where Q and D denote query and document respectively, W_Q and W_D are their vector representations. Then, use the same type of graph to construct a graph generator Graphon, then mix the graph generator Graphon to get Graphons, and perform linear interpolation operation to mix to get the new first enhanced graph data, as shown in formula (2):

$$\tilde{W} = \lambda W_D + (1-\lambda) W_D \tag{2}$$

$\tilde{W}$ represents the generated mixed document, $\lambda \in [0,1]$, to construct a mixed document with soft label λ . The second way to enhance graph data is to modify the dropout hyperparameters and perturb to form new graph data.

Regarding the calculation of the similarity between the query and the document, first a query q and a document d are given, and the encoded word sequence is $q = [w_1^q, ..., w_m^q]$, $d = [w_1^d, ..., w_n^d]$ .Where w_i^q represents the i-th word in the query, w_i^d represents the i-th word in the document, and m and n represent the lengths of the query and the document, respectively. Regarding the calculation of the correlation score $f_{i,j}(q,d)$, between query words and document words, as shown in formula (3):

$$f_{i,j}(q,d) = \cos(e_i^q, e_j^d) \tag{3}$$

3.Experiment and result analysis

This paper conducts experiments on two datasets NQ (Natural Questions) [11] and TQA (TriviaQA) [12] of OpenQA, and divides the dataset into training set, test set and validation set. In addition, this paper uses the TOP-K accuracy rate T-K as the evaluation index, and uses it to evaluate the accuracy rate of relevant documents re-trieved. This article compares the four stages of TOP-K accuracy rate, which are T-1, T-5, T-20, T-100. The T-K value in Table 1 is the result of multiplying it by 100, that is, in Table 1. The value a represents a %.

Table 1. Retrieval experiment results

	Natural Questions(NQ)				TriviaQA(TQA)			
	T-100	T-20	T-5	T-1	T-100	T-20	T-5	T-1
BM25	78.25	62.94	43.77	22.11	83.15	76.41	66.28	46.30
DPR	83.77	72.94	54.02	27.45	80.50	71.07	57.74	33.63
DPR w/QA	83.46	72.13	55.46	27.67	80.76	71.88	59.14	35.39
DPR w/DA	83.68	72.83	55.51	29.31	80.42	71.55	58.64	35.85
DAR	84.18	75.04	57.62	30.42	81.30	72.66	59.88	36.94
QAR	84.21	74.76	57.51	31.25	80.91	72.12	59.94	36.92
GDAR	**82.58**	**73.20**	**58.14**	**34.54**	**82.31**	**73.62**	**63.25**	**41.82**

It can be observed from Table.1 that the GDAR model in the T-1 stage on the two data sets is com-pared with that of the DAR model. It has been increased by 4.12% and 4.88% respectively, which shows that after the query and document are converted into graph data, the connection between words is strengthened, so that the query and document do not lose a lot of semantic information in the encoding stage, and The G-mixup graph data enhancement method also solves the problem of insufficient labeled training data on the basis of retaining document structure information, and effectively improves the performance of the GDAR model, making the GDAR model better than the baseline of other dense document retrieval models. In addition, this paper observes that the GDAR model per-forms well in the T-1 and T-5 stages of the two datasets, but performs worse than the DAR model in the T-20 and T-100 stages of the NQ dataset. This paper believes that this is due to The larger the retrieval granularity, the lower the retrieval accuracy.

In order to prove the effectiveness of the proposed model, this paper conducts ab-lation experiments on the NQ dataset. It is used to distinguish the impact of the two modules of linear interpolation Int (Interpolation) and perturbation Per (Perturbation) on the results of GDAR model retrieval experiments. The details are shown in Table 2. Table 2 shows the results of the model's ablation experiment on the NQ dataset by removing the two modules of linear interpolation Int and perturbation per. It can be seen from the data in the table 2 that when the perturbed Per module is removed, the accuracy of T-1 and T-5 drops by 3.52% and 1.30% respectively, which verifies the importance of the perturbed Per module to the dense enhanced retrieval model GDAR. When the linear interpolation Int module is removed, the accuracy of T-1 and T-5 drops by 1.69% and 0.51%, respectively, indicating the contribution of linear interpolation Int to the dense augmented retrieval model GDAR.

Table 2. GDAR model ablation experiments

	T-5	T-1
GDAR	58.14	34.54
w/o Per	56.84	31.02
w/o Int	57.63	32.85
DAR	57.62	30.42
DPR	54.02	27.45

4.Conclusion

This paper proposes to introduce G-mixup, a graph data enhancement method, into the dense document retrieval model. In this paper, graphs are used to represent queries and documents. On the basis of preserving the semantic and structural information of the text, it also solves the problem of insufficient labeled training data. For existing methods, there is a problem of information bloat, which increases the computational cost of dense document retrieval models. Since not all information in the document is highly relevant to the query, you can consider deleting document information to improve the computational efficiency of the model. Therefore, in future work, how to effectively delete and query the text information of irrelevant documents is a direction worthy of research.

References

[1] Chen, J., Chen, Q., Li, D., & Huang, Y. (2022). SeDR: Segment Representation Learning for Long Documents Dense Retrieval. arXiv preprint arXiv:2211.10841. https://doi.org/10.48550/arXiv.2211.10841.

[2] Liu, Y., Hashimoto, K., Zhou, Y., Yavuz, S., Xiong, C., & Yu, P. S. (2021). Dense hierarchical retrieval for open-domain question answering. arXiv preprint arXiv:2110.15439. https://doi.org/10.48550/arXiv.2110.15439.

[3] Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., ... & Yih, W. T. (2020). Dense passage retrieval for open-domain question answering. arXiv preprint arXiv:2004. 04906. https://aclanthology.org/2020.emnlp-main.550.

[4] Qu, Y., Ding, Y., Liu, J., Liu, K., Ren, R., Zhao, W. X., ... & Wang, H. (2020). RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering. arXiv preprint arXiv:2010.08191. https://aclanthology.org/2021.naacl-main.466.

[5] Liang, D., Xu, P., Shakeri, S., Santos, C. N. D., Nallapati, R., Huang, Z., & Xiang, B. (2020). Embedding-based zero-shot retrieval through query generation. arXiv preprint arXiv:2009.10270.https://doi.org/10.48550/arXiv.2009.10270.

[6] Ma, J., Korotkov, I., Yang, Y., Hall, K., & McDonald, R. (2020). Zero-shot neural passage retrieval via domain-targeted synthetic question generation. arXiv preprint arXiv:2004.14 503.https://doi.org/10.48550/arXiv.2004.14503.

[7] Wei, J., & Zou, K. (2019). Eda: Easy data augmentation techniques for boosting performance on text classification tasks. arXiv preprint arXiv:1901.11196. https://aclanthology. org/D19-1670.

[8] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929-1958. https://dl.acm.org/doi/10.5555/2627435.267 0313.

[9] Ding, K., Xu, Z., Tong, H., & Liu, H. (2022). Data augmentation for deep graph learning: A survey. ACM SIGKDD Explorations Newsletter, 24(2), 61-77.https://doi.org/10.485 50/arXiv.2202.08235.

[10] Jeong, S., Baek, J., Cho, S., Hwang, S. J., & Park, J. C. (2022). Augmenting document representations for dense retrieval with interpolation and perturbation. arXiv preprint arXiv:2203.07735.https://doi.org/10. 48550/arXiv.2203.07735.

[11] Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., ... & Petrov, S. (2019). Natural questions: a benchmark for question answering research. Transactions of the Association for Computational Linguistics, 7, 453-466. https://aclantho logy.org/Q19-1026.
[12] Joshi, M., Choi, E., Weld, D. S., & Zettlemoyer, L. (2017). Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. arXiv preprint arXiv:1 705.03551. https://aclanthology.org/P17-1147.