The 6th International Conference on Interdisciplinary research on Computer science, Psychology, and Education (ICICPE' 2022)
December 26-28, 2022. Pattaya, Thailand.

You Need More Data? Data Augmentation using Retrieval Technique

Seungjun Lee[1], Jaehyung Seo[1], Jungseob Lee[1],
Myunghoon Kang[1], Hyeonseok Moon[1],
Jaewook Lee[1], and Heuiseok Lim[1*],

[1] Korea University, Computer Science. 145, Anam-ro, Seongbuk-gu,
Seoul, Republic of Korea
{dzzy6505, seojae777, omanma1928, chaos8527, glee889, jaewook133,
limhseok}@korea.ac.kr

Abstract. In general, data augmentation techniques help improve low-resource environments. Recent research have attempted augmenting data by generating synthetic data using generative models. These methods aim to increase lexical and structural diversity without compromising the semantic similarity with the original sentence. This paper proposes a task-oriented data augmentation method. We use a retrieval model and a pre-trained generative model. We use retrieval model for paring setnteces which are similar to the input sentences in the training dataset. Then, a generative model is trained to generate synthetic data. The methodology of this paper is able to improve the baseline performance by up to 4% or more in a low-resource environment, and it shows a higher performance improvement than the existing data augmentation method.

Keywords: Data Augmentation, Information Retrieval, Generation, NLP

1 Introduction

Collecting data and labeling it is expensive and time consuming. Data Augmentation (DA) is a method of effectively acquiring data without additional labeling in many fields, including computer vision and speech recognition. In addition, increasing the size of the dataset has the effect of reducing overfitting of the model and enhancing the robustness of the model in low-resource tasks.

This paper proposes DART (Data Augmentation using Retrieval Technique), a data augmentation technique using a retrieval model and a natural language generation model for a text classification task. In this paper, the generative model is used to construct effective synthetic data in consideration of the lexical and structural variability and diversity of the language model.

In addition, we use a search model to generate synthetic data that considers task-sentences similar to the sentences of the training dataset from external data.

Then, a generative model is trained using the similar pair data constructed in this way. Synthetic data is generated using the training dataset to be augmented as an input to the generative model. The proposed methodology enables the generative model to learn by considering vocabulary and structural diversity for semantically similar sentences. This paper has conducted an experiment on the text classification task and proves that it has high performance compared to the existing data augmentation method.

2 DART

The latest generative model research focuses on the generation of synthetic data in which the sequence-to-sequence (Seq2Seq) model considers various vocabularies and sentences for semantically similar pairs. This paper utilizes this methodology to fine-tune using a pre-trained generative model for sentence pairs similar to the training dataset. By using the learned generative model, we aim to generate synthetic data that is semantically similar to the training dataset, but with modified vocabulary and sentence structure.

In previous studies, generative models were trained by conditioning the label information of the training dataset. Although this method has a significant performance improvement, the lexical and sentence transformation of the generated synthetic data has hardly been achieved. This is because only label information is conditioned, and only related lexical patterns are learned. In order to increase the diversity of synthetic data, there have been attempts to learn generative models by conditioning the training data. This study uses the [CLS] embedding of BERT [1] trained in the training dataset to find similar sentence pairs within the training dataset. Fine-tune generative models using similar sentence pairs. Thereafter, the training data is passed through the fine-tuned generative model to generate synthetic data. This methodology should divide the training dataset into three categories: data for classifier learning, data for search target, and data for synthetic data generation. This method is difficult to properly train a generative model using only a small amount of data in a low-resource environment, and it is difficult to increase diversity because it relies only on internal data to search for similar pairs.

Therefore, this paper proposes DART (Data Augmentation using Retrieval Technique) that can guarantee the diversity of synthetic data without partitioning the dataset. A search model is used to utilize similar pairs, and various synthetic data are generated by fine-tuning the searched similar pairs to the generative model.

3 Experiment

This paper conducts an experiment on the low-data regime setting and text classification task. We follow experiment settings of previous data augmentation research. Implementation for arbitrarily low-resource environment Classification task Use 100 data per class in dataset. The text classification experiment is conducted on the topic classification (TC) task of the KLUE [2] benchmark, a Korean dataset.

Table 1 shows the classifier performance according to data augmentation for the topic classification task of KLUE. In the case of the search model, except for BM25, both the cross-encoder and the bi-encoder showed the effect of data augmentation. On the out-domain basis, there was a performance improvement of 3.43 and 3.51, respectively, compared to no augmentation.

Table 1. Comparison of Existing Data Augmentation Methodology and Retrieval Performance on the KLUE Topic Classification Dataset.

	F1
No augmentation	78.36
EDA	81.58
Back Translation	81.17
In-domain	80.48

4 Conclusion

This paper proposes Dart, a data augmentation technique that utilizes a search model and a generative model. We trained by fine-tuning the generative model to generate synthetic data similar to the training data. The data used to train the generative model are sentence pairs similar to the training dataset retrieved using the search model from an external dataset. These sentence pairs are considered for semantic similarity so that the generative model can learn vocabulary and sentence transformations. In addition, through the experiment, the existing data augmentation method, comparison experiment, and search model performance experiment were conducted. Quantitatively, the search model using the out-domain as external data showed higher performance than the existing method. Qualitatively, various vocabularies and sentence transformations were possible while maintaining semantic similarity with the original data. In a future study, we intend to apply Dart to various tasks such as question answering.

5 Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2020-0-00368, A Neural-Symbolic Model for Knowledge Acquisition and Inference Techniques). Also, this research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2022-2018-0-01405) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation).

References

1. Devlin, J., Chang, M.W., Lee, K. and Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. (2018).
2. Park, S., Moon, J., Kim, S., Cho, W. I., Han, J., Park, J., ... & Cho, K.: Klue: Korean language understanding evaluation. arXiv preprint arXiv:2105..(2021)