Hybrid Model of Data Augmentation Methods for Text Classification Task

Jia Hui Feng and Mahsa Mohaghegh
Engineering, Computer & Mathematical Sciences, Auckland University of Technology, Auckland, New Zealand
{jiahui.feng, mahsa.mohaghegh}@autuni.ac.nz

Keywords: Data Augmentation, Hybrid Models, Machine Learning, Natural Language Processing.

Abstract: Data augmentation techniques have been increasingly explored in natural language processing to create more textual data for training. However, the performance gain of existing techniques is often marginal. This paper explores the performance of combining two EDA (Easy Data Augmentation) methods, random swap and random delete for the performance in text classification. The classification tasks were conducted using CNN as a text classifier model on a portion of the SST-2: Stanford Sentiment Treebank dataset. The results show that the performance gain of this hybrid model performs worse than the benchmark accuracy. The research can be continued with a different combination of methods and experimented on larger datasets.

1 INTRODUCTION

Artificial intelligence has changed the world in many ways, from unlocking our digital devices with just a glance to automating our daily manual tasks, all to make our lives more efficient. One of the fundamental tasks to achieve such technological capabilities is being able to identify and label information. For the machine to be perform classification tasks such as seeing an image of a dog and the machine to identify it as "dog", it requires large amounts of training data for the machine to learn. In this particular case, multiple different images of dogs would be required to train the machine on what a dog looks like. But often in the real world, resources of such quality data are expensive and time consuming to obtain. This issue is of significant personal interest to the author, as they have worked as an artificial intelligence consultant and often faced constraints in collecting training data due to limited resources.

One of the solutions is a method called data augmentation. This method helps to increase the dataset by introducing varied distributions and increases the performance of the model for different tasks (Giridhara et al, 2021). It consists of generating synthetic data for the machine to be trained on, which will reduce the time and resources spent to obtain the training data from the real world. The earliest demonstrations showing the effectiveness of this technique are from computer vision (Shorten & Khoshgoftaar, 2019). For instance, if we only have one good quality image of a dog, simple transformations such as horizontal flipping, changing colour, and enlarging the dog's face generates four more images for the training dataset. In general, it is agreed that the more data a neural network gets trained on, the more effective it becomes (Giridhara et al, 2021).

The field of computer vision has already seen the success and benefits of data augmentation on images, however, the same fundamental tasks of labeling and classifying is also a need in languages and text, or what is also known as the field of natural language processing (NLP). The field of NLP in artificial intelligence consists of chatbots, SIRI, autocompletion when searching, speech recognition and more. In order to achieve the abilities of these types of technologies, the same task of training and classifying information for image data is also fundamental for text data. Just like images, textual datasets are also expensive and time consuming to obtain.

Recent studies have begun to explore the idea of applying the technique of data augmentation in NLP, but due to the complexity of languages it has been a challenging topic (Kobayashi, 2018). Some studies show hybrid machine learning algorithms for text classifications to improve performance, for instance combing Naïve Bayes and SVM models has been shown to improve accuracy substantially (Asogwa et al, 2021). Since hybrid models for text classifications have been shown to be effective, this paper explores

the idea of combining textual data augmentation techniques for text classification tasks. The use of a hybrid model with data augmentation techniques has not previously been used for text classification tasks.

This research topic explores the effectiveness of a hybrid model of data augmentation techniques for text classification tasks. We will be exploring this by combining the individual methods of a novel data augmentation technique called Easy Data Augmentation (EDA) (Wei & Zhou, 2019). Recent data augmentation technique, EDA, is a novel method of using four strategies for text classification tasks. The four strategies are: synonym replacement, random insertion, random swap and random deletion (Wei & Zhou, 2019).

Although these strategies have been shown to increase performance on text classification tasks, it is only minimal. Closer examination of this novel technique shows that the two best performing techniques are random swap and random deletion, where they have the highest performance gain out of all the four at the most optimal parameter, 0.1. Random deletion has the highest out of all four, but performance gain becomes close to 0% when the data augmentation alpha parameter is at 0.5 (Wei & Zhou, 2019), which is the worst performing out of all the four at this parameter. Random swap however has the best performance at 0.5. The hypothesis is by combining random deletion and random swap together, the performance gain would increase at two chosen parameters. The first chosen parameter is the most optimal parameter, 0.05, and the second is the worst performing parameter at 0.5. Thus, the sub-questions are: **1) How effective is RS&RD as a hybrid model evaluated by performance gain?** And **2) How does this hybrid model compare to the techniques of RS and RD individually?**

2 LITERATURE REVIEW

Although there has been significant success with data augmentation in image data, the usage of data augmentation for NLP is quite different. For instance, taking one sentence and augmenting it, for example, by changing the words of that sentence to generate a new sentence would make the meaning of the sentence entirely different. Previous studies on the common approach for data augmentation is replacing words with their synonyms such as Word-Net (Miller, 1995). However, as reported by Kobayashi (2018) these studies using the techniques of synonym-based augmentation can only be applied to a small percentage of a vocabulary.

There has been research on using lexical and semantic embeddings as a data augmentation technique (Wang and Yang, 2015) which has shown improvements of accuracy, however this was used to enhance computational behavior analysis, and it does not show the effectiveness of applying it to text classification. Another popular technique is back-translation, and showed the effects of generating new data by translating sentences from French back to English, but the major benefit was to bring more syntactical diversity to the enhanced data (Yu et al., 2018). A more recent data augmentation technique, EDA (Easy Data Augmentation) is a novel technique of using 4 strategies (Wei & Zhou, 2019). However, performance gain is only marginal and declines when dataset becomes larger. There is still a significant amount of research in this field of data augmentation for text classification techniques.

3 RESEARCH DESIGN & IMPLEMENTATION

This experiment is conducted by using one benchmark text classification task and one network architecture to evaluate the hybrid model.

3.1 Benchmark Dataset

In this study, the dataset used is one of the benchmark datasets Wei and Zhou used for their EDA experiments. The original dataset is the SST-2: Stanford Sentiment Treebank (Yu et al., 2018). For Wei and Zhou's dataset, they have extracted 500 sentences of movie reviews from the SST-2 dataset which are then cleaned and processed. Each sentence also has a label of 0 or 1 indicating the sentiment type, 0 being negative or neutral and 1 being positive.

3.2 Hybrid Model

The hybrid model will be a combination of the two EDA techniques, random deletion and random swap. The techniques individually will take one sentence and randomly delete words or randomly swap two words. For the hybrid model, it will take one sentence and perform both a random swap and a random deletion. The hybrid model can be combined in two ways, the first method is randomly deleting words first, then randomly swapping two words. The second method is randomly swapping two words first, then randomly deleting the words.

3.3 Text Classification Model

The text classification model used is Convolutional neural networks (CNNs), as CNNs have achieved high performance for text classifications (Wei & Zhou, 2019). The benchmark data has 500 sentences. It will be split into 80/20 for training and validation. Thus, the 500 sentences will be split into 400 sentences and 100 sentences. The 400 sentences will be the training data, and the 100 sentences will be the validation data. The CNN model will run on the benchmark dataset to obtain the benchmark accuracy.

The two methods of hybrid models will each have two different α parameters, which indicates the amount of augmentation per sentence. Each will have parameter for 0.05 and 0.5, which is 5% of augmentation and 50% of augmentation, respectively. For instance, if α is 0.05 then the model will randomly swap 5% of the sentence and randomly delete 5% of the sentence. Each is run on the 400 sentences training set and will generate 16 augmentations per sentence, as 16 is the recommended number of augmentations for 500 sentences, based on Wei and Zhou's EDA experiment. Thus, it would generate a dataset of 6400 sentences. Then the CNN model will be trained on the augmented dataset and validate it against the validation training set. There will be a total of 5 results, 1 being the benchmark accuracy and the other 4 the accuracy results of training on the different augmented data.

4 FINDINGS & DISCUSSION

Table 1: Results of Augmentation.

Model	% of augmentation	Accuracy	Result
Baseline (CNN)	0	77%	0%
CNN + RD & RS	5% 50%	73% 70%	-4% -7%
CNN + RS & RD	5% 50%	73% 67%	-7% -10%

As shown, the baseline accuracy is 77% using CNN. The first hybrid model (random delete first then random swap) at 5% augmentation has achieved an accuracy of 73%, and at 50% augmentation has achieved an accuracy of 70%.

Both results have performed worse than the baseline. Then for the second type of hybrid model (random swap then random delete) at 5% augmentation has achieved 70% accuracy andat 50% augmentation has achieved 67%. This is the worst performance of all models. It follows that applying the hybrid model has decreased the accuracy between 4% to 10%, and that random swap and random deletion yield better results individually.

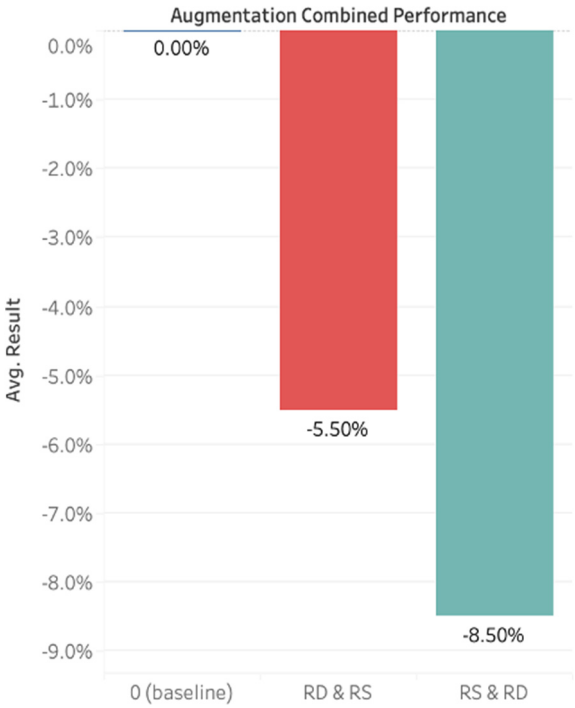

Figure 1: Average performance of hybrid models 5% and 50%.

5 LIMITATIONS

The limitation is that the study was only experimented on a very limited and small dataset so we could only see the results of this type of data, in this case it was a small dataset of 500 sentences which was a sentiment classification of either 0 or 1.

6 CONCLUSIONS & FUTURE WORK

Training data is a crucial component in the process for most kinds of machine learning systems, and being able to obtain more training data for machines to learn becomes extremely helpful in saving time and resources. Synthetically generating more training data for computer vision has been proven to be

extremely effective, and there is more research to be explored in the field of NLP.

In this experiment, we looked at a potential data augmentation technique of combining two existing methods from EDA. We combined the two best performing methods, random swap and random deletion, and evaluated the performance by the accuracy results from a CNN model. There are two types of this hybrid model by combing the two methods. The first type is to random delete then random swap; the second type is to random swap first then random delete. It was then used to generate synthetic textual data for a text sentiment classification task generating 5% of augmentation and 50% of augmentation. Both have shown to perform worse than the baseline results. This is perhaps because the generated text has become so different from the original sentence that the original sentiment has been changed, since each sentence has contents swapped and deleted.

Thus, in the future, more experiments should be conducted on different sized large datasets, with which different results are anticipated. In addition, since the hybrid model of random swap and random delete arranges the content and deletes a portion of it, there can be an addition to this two-method hybrid model of adding more content such as 'random insertion' from EDA. Since this experiment is also only a hybrid model of two methods, there are many existing different methods available other than EDA, which could be further experimented to explore more potential in hybrid models for textual data augmentation.

REFERENCES

Asogwa, D. C., Anigbogu, S. O., Onyenwe, I. E., & Sani, F. A. (2021). Text Classification Using Hybrid Machine Learning Algorithms on Big Data. *ArXiv:2103.16624 [Cs]*. http://arxiv.org/abs/2103.16624

Ganapathibhotla, M., & Liu, B. (2008). Mining opinions in comparative sentences. *Proceedings of the 22nd International Conference on Computational Linguistics - COLING '08*, *1*, 241–248. https://doi.org/10.3115/1599081.1599112

Giridhara, P., Mishra, C., Venkataramana, R., Bukhari, S., & Dengel, A. (2021). *A Study of Various Text Augmentation Techniques for Relation Classification in Free Text*. 360–367. https://www.scitepress.org/PublicationsDetail.aspx? ID=R8tLLz6nUJk=&t=1

Hu, M., & Liu, B. (n.d.). *Mining and Summarizing Customer Reviews*. 10.

Kobayashi, S. (2018). Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), 452–457. https://doi.org/10.18653/v1/N18-2072

Li, X., & Roth, D. (2002). Learning question classifiers. *Proceedings of the 19th International Conference on Computational Linguistics -*, *1*, 1–7. https://doi.org/10.3115/1072228.1072378

Liesting, T., Frasincar, F., & Trusca, M. M. (2021). Data Augmentation in a Hybrid Approach for Aspect-Based Sentiment Analysis. *ArXiv:2103.15912 [Cs]*. http://arxiv.org/abs/2103.15912

Miller, G. A. (n.d.). *A LEXICAL DATABASE FOR ENGLISH*. 1.

Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics - ACL '04*, 271-es. https://doi.org/10.3115/1218955.1218990

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, *6*(1), 60. https://doi.org/10.1186/s40537-019-0197-0

Wang, W. Y., & Yang, D. (2015). That's So Annoying!!!: A Lexical and Frame-Semantic Embedding Based Data Augmentation Approach to Automatic Categorization of Annoying Behaviors using #petpeeve Tweets. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2557–2563. https://doi.org/10.18653/v1/D15-1306

Wei, J., & Zou, K. (2019). EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. *ArXiv:1901.11196 [Cs]*. http://arxiv.org/abs/1901.11196

Yu, A. W., Dohan, D., Luong, M.-T., Zhao, R., Chen, K., Norouzi, M., & Le, Q. V. (2018). QANet: Combining Local Convolution with Global Self- Attention for Reading Comprehension. *ArXiv:1804.09541 [Cs]*.