Generative AI Authorship Verification based on ChatGLM

Notebook for the PAN Lab at CLEF 2024

Haotian Lei, Xiangyu Liu*, Guo Niu, Yan Zhou and Yuexia Zhou

Foshan University, Foshan, China

Abstract

In this paper, we use the LoRA method to fine-tune the large language model ChatGLM. To balance the data distribution in the dataset, we modified the labels and transformed it into a multi-classification task. This enables the large language model to better learn the differences in expression among different authors on the same topic, thereby learning the writing styles of humans and machines. During inference, we modify its final output by remapping it into a binary classification task, distinguishing whether the text was authored by a human or a machine. This approach aims to achieve the task of Generative AI Authorship Verification. The evaluation results on the PAN corpus test dataset indicate that this method is effective, with an mean score greater than 0.7.

Keywords

Generative AI Authorship Verification, Large Language Models, LoRA

1. Introduction

As artificial intelligence-generated content (AIGC) technology continues to advance, large language models (LLMs) such as ChatGPT, ChatGLM [1], and Qwen [2] are improving at an astonishing rate and being increasingly adopted across various sectors. The text generated by these models has reached a level comparable to that of human peers, enabling them to provide highly fluent and meaningful responses to a wide variety of user queries. The rapid development and widespread adoption of LLMs highlight their potential to revolutionize how we interact with technology, offering significant improvements in efficiency and user experience.

However, with these advancements, several issues have also surfaced. One major concern is the rapid spread of fake news, as LLMs can generate realistic and convincing false information that can be quickly disseminated across various platforms. Additionally, there is the manipulation of public opinion through social media comments, where LLMs are used to produce a large volume of persuasive and biased posts, swaying public perception and discourse. Another significant problem is academic dishonesty, with students using LLMs to complete their assignments, which undermines the integrity of the educational process and presents challenges for educators in assessing genuine student performance.

This paper presents our approach for Generative AI Authorship Verification task [3, 4] on PAN 2024. For this task, our approach is to use LLMs to counteract LLMs. Our work is based on ChatGLM, utilizing the LoRA [5] method for fine-tuning. Additionally, to better fine-tune the LLM, we modified the training dataset content and its labels to make it easier to distinguish between human and machine writing, thereby improving its reasoning ability on the test set. Finally, we submitted our results on TIRA [6].

2. Related Work

As large language models rapidly advance in generating extremely high-quality text, powerful LLMs provide unprecedented convenience to people. These models not only understand and process complex language inputs but also excel in generating coherent and contextually appropriate text, making

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

928718890@qq.com (H. Lei); xyliu1805@fosu.edu.cn (X. Liu*); niuguo@mail2.sysu.edu.cn (G. Niu); zhouyan791266@fosu.edu.cn (Y. Zhou); fs_zyx@fosu.edu.cn (Y. Zhou)

0009-0004-8490-8880 (H. Lei); 0009-0006-4760-6837 (X. Liu*); 0000-0002-1552-7310 (G. Niu)

them valuable tools for applications such as customer service, content creation, educational support, and more.However, the false text generated by powerful LLMs is raising ethical and legal concerns. Moreover, it has become increasingly difficult for people to rely on their own experience to determine whether a piece of text was written by a human or a machine. RoFT [7] attempted to involve users in detecting machine-generated text. However, only 15.8% of annotations correctly identified the detection boundary. This has led researchers to consider using more accurate methods to combat and detect false text. Consequently, various methods have been developed to detect and differentiate these generated texts.Zellers [8] illustrated the generation of machine-produced fake news by proposing a GPT-based news generator called GROVER. They also used GROVER itself to classify and detect fake news. GLTR [9] detects generated text in a zero-shot manner by utilizing token prediction probabilities from available pre-trained NLP models, such as BERT [10] and GPT-2 [11]. OpenAI recently released an AI text classifier by fine-tuning a GPT model [12], using LLMs to counteract the misuse of LLMs. Similarly, we fine-tune a large language model, ChatGLM, using the LoRA method to achieve Generative AI Authorship Verification.

3. System Overview

For this task, we utilize the dataset provided by PAN, which includes both genuine and fake news articles spanning multiple headlines from the United States in 2021. It consists of a JSON file written by 13 different machine authors and one human author. Each file contains articles on the same topic. The IDs and line order of the articles are the same, so the same line always corresponds to the same topic. Each document contains 24 topics and 1087 articles.

Although this could be considered as a binary classification task, to determine whether the text is written by a human or not.However, due to the extremely high quality of text generated by large language models, we believe that simply dividing this task into a binary classification task may not yield satisfactory results. Since we are also using a large language model to perform this task, in order to enable the model to learn different textual features, we have modified the dataset labels to implement a multi-class classification task.

Table 1
Type,Token Range and Label of Generative AI Authorship Verification training datasets

Type	Token Range	Label
Human	[25,7989]	0
alpaca-7b	[0,3141]	1
bigscience-bloomz-7b1	[96,3557]	2
chavinlo-alpaca-13b	[0,5505]	3
gemini-pro	[1205,5881]	4
gpt-3.5-turbo-0125	[75,5961]	5
gpt-4-turbo-preview	[1189,6931]	6
meta-llama-llama-2-7b-chat-hf	[367,5865]	7
meta-llama-llama-2-70b-chat-hf	[1209,5957]	8
mistralai-mistral-7b-instruct-v0.2	[1446,6274]	9
mistralai-mixtral-8x7b-instruct-v0.1	[811,6928]	10
qwen-qwen1.5-72b-chat-8bit	[1404,3917]	11
text-bison-002	[0,5613]	12
vicgalle-gpt2-open-instruct-v1	[52,3653]	13

Table 1 shows the specific modifications made to the labels in the dataset. Fortunately, the dataset provided by PAN is very well-organized, with each type of "author" providing the same number of articles, which is 1087.In the form of multi-class tasks, there won't be exaggerated data proportions like "human:machines=1:13", which could lead to the issue of unbalanced learning data. Finally, during the inference test, we limit the output result to a number between 0 and 1. The closer the number is to

1, the more likely the text is written by a human. Conversely, the closer the number is to 0, the more likely it is written by a machine. When the probabilities are equal, we default to considering the first comment as written by a human and the second comment as written by a machine.

This paper employs LoRA technology to fine-tune ChatGLM. LoRA is a method used for fine-tuning large language models, aimed at enhancing the model's performance on specific tasks. The core idea of LoRA is to expand the language representation of the model by introducing domain-specific corpora, making it more specialized and adaptable, as shown in Equation 1

$$H = Wx + \Delta Wx = Wx + \mathrm{BA}x = (W + BA)x \tag{1}$$

Where $W \in R_{d\times k}$ represents the weight matrix of the pre-trained model. ΔW represents the change in the weights. $\Delta W = BA$ represents the update part obtained through fine-tuning, which is updated using low-rank decomposition. Where $B \in R_{d\times r}$, $A \in R_{d\times k}$, and the rank $r \ll \min(d, k)$.

During fine-tuning, the weight parameters W of the pre-trained model are frozen, and only the parameters A and B are trained. As shown in Figure 1, LoRA integrates the trained bypass weight parameters with the pre-trained model weights without introducing additional pathways for inference, making it suitable for real-time requirements in vertical domains.

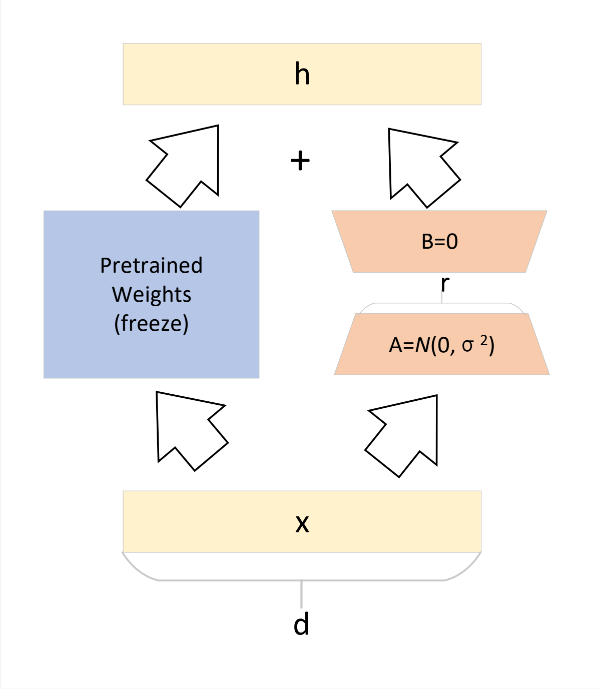

Figure 1: Framework diagram of LoRA

In this work, for LoRA, we set the LoRA rank to 4, the batch size to 8, and the learning rate to 1e-4, using FP16 for training. We completed this fine-tuning on a single A800 GPU. For ChatGLM, we set the top p to 0.7, max length to 2, and temperature to 0.2.

4. Results

We submitted our system to TIRA and utilized the evaluation metrics provided by TIRA, which specifically include the following:

ROC-AUC: The area under the ROC (Receiver Operating Characteristic) curve.

Brier: The complement of the Brier score (mean squared loss).

C@1: A modified accuracy score that assigns non-answers (score = 0.5) the average accuracy of the remaining cases.

F1: The harmonic mean of precision and recall.

F0.5u: A modified F0.5 measure (precision-weighted F measure) that treats non-answers (score = 0.5) as false negatives.

Mean:The arithmetic mean of all the metrics above

Table 2
Overview of the accuracy in detecting if a text is written by an human in task 4 on PAN 2024 (Voight-Kampff Generative AI Authorship Verification). We report ROC-AUC, Brier, C@1, F_1, $F_{0.5u}$ and their mean.

Approach	ROC-AUC	Brier	C@1	F_1	$F_{0.5u}$	Mean
brownian-architect	0.727	0.727	0.727	0.656	0.778	0.723
Baseline Binoculars	0.972	0.957	0.966	0.964	0.965	0.965
Baseline Fast-DetectGPT (Mistral)	0.876	0.8	0.886	0.883	0.883	0.866
Baseline PPMd	0.795	0.798	0.754	0.753	0.749	0.77
Baseline Unmasking	0.697	0.774	0.691	0.658	0.666	0.697
Baseline Fast-DetectGPT	0.668	0.776	0.695	0.69	0.691	0.704
95-th quantile	0.994	0.987	0.989	0.989	0.989	0.990
75-th quantile	0.969	0.925	0.950	0.933	0.939	0.941
Median	0.909	0.890	0.887	0.871	0.867	0.889
25-th quantile	0.701	0.768	0.683	0.657	0.670	0.689
Min	0.131	0.265	0.005	0.006	0.007	0.224

Table 3
Overview of the mean accuracy over 9 variants of the test set. We report the minumum, median, the maximum, the 25-th, and the 75-th quantile, of the mean per the 9 datasets.

Approach	Minimum	25-th Quantile	Median	75-th Quantile	Max
brownian-architect	0.219	0.691	0.725	0.776	0.907
Baseline Binoculars	0.342	0.818	0.844	0.965	0.996
Baseline Fast-DetectGPT (Mistral)	0.095	0.793	0.842	0.931	0.958
Baseline PPMd	0.270	0.546	0.750	0.770	0.863
Baseline Unmasking	0.250	0.662	0.696	0.697	0.762
Baseline Fast-DetectGPT	0.159	0.579	0.704	0.719	0.982
95-th quantile	0.863	0.971	0.978	0.990	1.000
75-th quantile	0.758	0.865	0.933	0.959	0.991
Median	0.605	0.645	0.875	0.889	0.936
25-th quantile	0.353	0.496	0.658	0.675	0.711
Min	0.015	0.038	0.231	0.244	0.252

We evaluated the performance of our model on the new test set provided by PAN, and the test results are shown in Table 2. Our test results are higher than Baseline Unmasking and Baseline Fast-DetectGPT, but lower than Baseline Binoculars, Baseline Fast-DetectGPT (Mistral), and Baseline PPMd. Our approach performs poorly on the new test dataset. This suggests that our model approach's generalization ability is not satisfactory.

Table 3 further shows the average accuracy of ours model on different dataset variants, particularly on the test sets of nine variants. Our model's minimum value across all variants was 0.219,with the 25th and 75th percentiles at 0.691 and 0.776, respectively, a median of 0.725, and a maximum value of 0.907. Our method surpasses PPMd, Unmasking, and Fast-DetectGPT on the 25-th quantile,75-th quantile and Max. Compared to the quantile results of other participants, our model is surpasses the models in the 25-th percentile in most metrics. This indicates that there is still a significant gap between our approach and the current state-of-the-art methods.

5. Conclusion

In this paper, we propose a method for Generative AI Authorship Verification on PAN 2024. We modified the labels in the training dataset to transform it into a multi-classification task. We fine-tuned ChatGLM with the aim of enabling the large language model to better understand the writing styles of different authors, thus learning the differences between robot and human writing. We utilized LoRA technology for fine-tuning, as LoRA method can extend the language representation of the model, making it more professional and adaptive. From the results, it appears that the method performs poorly on the new test dataset, indicating that our approach lacks some degree of generalization ability.In subsequent work, it is advisable to employ more effective methods to augment the data and enhance the model's classification ability on open sets.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 61972091), Natural Science Foundation of Guangdong Province of China (No. 2022A1515010101, No. 2021A1515012639).

References

[1] A. Zeng, X. Liu, Z. Du, Z. Wang, H. Lai, M. Ding, Z. Yang, Y. Xu, W. Zheng, X. Xia, et al., Glm-130b: An open bilingual pre-trained model, arXiv preprint arXiv:2210.02414 (2022).

[2] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, et al., Qwen technical report, arXiv preprint arXiv:2309.16609 (2023).

[3] J. Bevendorff, X. B. Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korenčić, M. Mayerl, A. Mukherjee, A. Panchenko, M. Potthast, F. Rangel, P. Rosso, A. Smirnova, E. Stamatatos, B. Stein, M. Taulé, D. Ustalov, M. Wiegmann, E. Zangerle, Overview of PAN 2024: Multi-Author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, L. Soulier, G. M. D. Nunzio, P. Galuščáková, A. G. S. de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2024.

[4] J. Bevendorff, M. Wiegmann, E. Stamatatos, M. Potthast, B. Stein, Overview of the Voight-Kampff Generative AI Authorship Verification Task at PAN 2024, in: G. F. N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.

[5] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, arXiv preprint arXiv:2106.09685 (2021).

[6] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241. doi:10.1007/978-3-031-28241-6_20.

[7] L. Dugan, D. Ippolito, A. Kirubarajan, C. Callison-Burch, Roft: A tool for evaluating human detection of machine-generated text, arXiv preprint arXiv:2010.03070 (2020).

[8] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, Y. Choi, Defending against neural fake news, Advances in neural information processing systems 32 (2019).

[9] S. Gehrmann, H. Strobelt, A. M. Rush, Gltr: Statistical detection and visualization of generated text, arXiv preprint arXiv:1906.04043 (2019).

[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[11] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[12] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).