Meta-Contrastive Learning for Generative AI Authorship Verification

Notebook for the PAN Lab at CLEF 2024

Jiajun Lv, Yong Han and Leilei Kong

Foshan University, Foshan, China

Abstract

This paper proposes a method that combines meta-learning and contrastive learning to address the task of Generative AI Authorship Verification. Our motivation is to leverage supervised contrastive learning to enhance the model's discriminative ability by optimizing the relationships between samples. Additionally, we employ the meta-learning algorithm Reptile to improve the generalization ability on out-of-domain data. Finally, we select the model weights that achieve the best performance on the validation set. We obtained an average score of 0.949 on the test set.

Keywords

Authorship Verification, Contrastive Learning, Meta-learning

1. Introduction

With the widespread application of generative AI and large language models (LLMs), complex issues have emerged, such as the spread of misinformation[1], facilitating plagiarism[2], particularly in academic writing using LLMs[1]. This creates an urgent need to develop detectors capable of identifying LLM-generated text. Since LLMs are trained on extensive datasets of text and code, they can produce content that closely resembles human-written text[3]. As a result, distinguishing between human and machine-written text has become increasingly challenging. In this study, we propose a method that combines contrastive learning and the Reptile meta-learning algorithm to address the PAN: Voight-Kampff Generative AI Authorship Verification task in CLEF 2024[4]. This task requires identifying the human-written text from two given texts.

In this research, we propose a combination of comparative learning and Reptile[5] meta-learning based approach to address the CLEF 2024 task PAN:Voight-Kampff Generative AI Authorship Verification which requires identifying human-written texts in a given two texts[6]

2. Related work

Since 2011, the PAN organization has been continuously organizing authorship verification tasks[7]. Unlike previous focuses on cross-discourse type authorship verification, PAN 2024 Authorship Verification[4] aims to address whether generative AI authorship verification can be solved[8]. The task requires participants to design classification methods to distinguish between human and machine-written texts.

In recent work on generative AI detectors, fine-tuning language models and zero-shot learning methods are predominant [3]. Zero-shot detectors do not require additional training through supervised signals. Major methods include perplexity (PPL) [9], probability curvature [10], and likelihood ratio ranking (LRR) [11]. Currently, supervised fine-tuning of pre-trained language models is very powerful in natural language understanding [12]. Recent works [3][12][13] further confirm that fine-tuning with pre-trained language models from the BERT family can outperform zero-shot methods in-domain.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France
lvjiajun.96@gmail.com (J. Lv); hanyong2005@fosu.edu.cn (Y. Han); kongleilei@fosu.edu.cn (L. Kong)
0000-0002-8755-5310 (J. Lv); 0000-0002-9416-2398 (Y. Han); 0000-0002-4636-3507 (L. Kong)

To further improve the detection capability of unknown models, contrastive learning has also been applied to LLMs text checking. ConDA [13] proposed a contrastive domain adaptation framework that combines domain adaptation with contrastive learning representations, enhancing the detector's performance on out-of-domain data. Reviewing last year's authorship verification task, the first-place team Ibrahim, M.et al[14] and the second-place team Guo, M. et al[15]. both adopted feature encoding and contrastive learning concepts. From these methods, it is evident that contrastive learning might be key to the authorship verification task.

Inspired by [13][16][17], we propose a method that combines contrastive learning and Reptile meta-learning[18]. Contrastive learning, by learning the relative distances between samples, avoids mapping texts to a single label. Unlike conventional fine-tuning methods, we use Reptile meta-learning to help the model learn better feature representations, enhancing its generalization ability.

3. Method

The goal of our model is to allow the model to learn the relative distance between samples on the same topic, with different authors. Feeding x text into the model yields a soft label y that encodes the text, the smaller the label value the more likely the text is to be judged as human-authored, and conversely the more likely it is to be judged as AI-generated text.

3.1. Contrastive Learning

Our method revolves around constructing a training task τ, where τ_n is represented as a collection of texts on the same topic written by different authors, denoted as $\{x_0^+, x_1^-, x_2^-, \ldots, x_n^-\}$. In this collection, x_0^+ is the only positive example, representing a human author, while $x_1^-, \ldots, x_n^-$ are negative examples, representing AI-generated authors.

The text x_i is input to the encoder, and the $[CLS]$ markers of the output vector of the last layer of the encoder are taken as the representation E_i of the text, and we feed the obtained vector E_i to the $ReLU$ activation function and the linear layer to obtain the soft labels y_i of the input text x_i.

$$E_i = encoder(x_i) \tag{1}$$

$$\hat{y}_i = \sigma(E_i W_h^T + b_h) \tag{2}$$

where $E_i \in \mathbb{R}^{batch\_size \times h}$,$W_h \in \mathbb{R}^{h \times 1}$, h is the dimension of the hidden layer of the encoder, and b_n is the bias of the fully connected layer. The $\sigma()$ is the nonlinear activation function $ReLU$. We compute the MarginRankingLoss loss function between numerical labels:

$$loss = max(0, margin - (\hat{y}_i^+ - \hat{y}_i^-)) \tag{3}$$

Where $\hat{y}_i^+$ is the soft label for positive examples, $\hat{y}_i^-$ is the soft label for negative examples, and $margin$ spacing boundaries, which indicates the minimum gap between two scores, and if the value is larger, it means that it is expected that $\hat{y}_i^+$ is further away from $\hat{y}_i^-$.

3.2. Reptile Meta-Learning

We use the batch version of the algorithm, define slow weight as ϕ, first copy ϕ model parameters as fast weight denoted as θ, use fast weight to sample n groups of training tasks on the training set to train the updated model, get the updated $\hat{\theta}$, calculate the difference between $\hat{\theta}$ and the difference of parameter ϕ as the gradient direction of updating ϕ, and carry out updating ϕ to get ϕ_1 by repeated iterations,During training, we adjust the parameter weights of DeBERTa and the linear classification layer,reptile training algorithm1

Algorithm 1 Reptile training algorithm

Input: Dataset τ, margin m, Model ϕ, N number of AI author categories
Output: Model parameter $\phi^{'}$

Initialising model parameters ϕ
for iteration = 1,2,... t **do**
  copy model parameters ϕ to θ
  Sample task $\tau_1, \tau_2, \tau_3 .... \tau_n$ in τ
  for i = 1,2,... n **do**
    $\hat{y}_i = \theta(\tau_i)$
    $L_n = \frac{1}{N}\sum_{j=1}^{j=N} max(0, m - (\hat{y}_0^{+} - \hat{y}_j^{-}))$
    $\theta^{'} \leftarrow L_n + \theta$
  end for
  $\phi^{'} \leftarrow \phi - \eta(\theta^{'} - \phi)$
  Deletion of parameters $\theta^{'}$
end for

4. Experiments

4.1. Dataset statistics

We perform sequence length statistics for each author's data in the training dataset, as shown in Figure1.

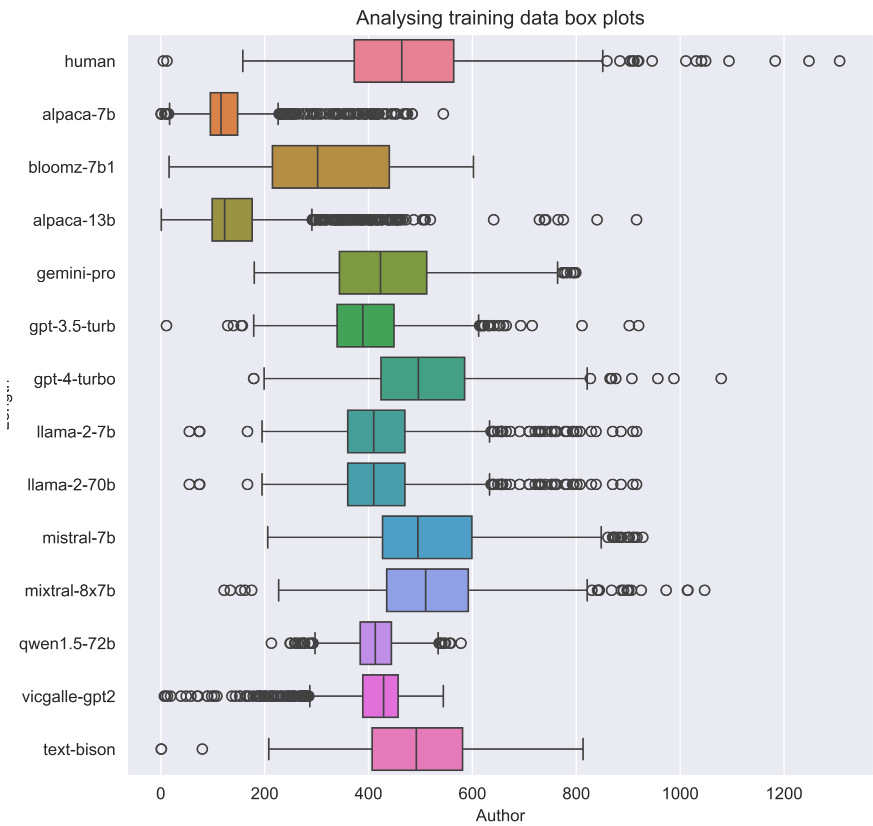

Figure 1: Dataset statistics Analyzing the length of text sequences on a dataset.

From the chart, it can be seen that the sequence length of the training dataset is around 500. Among them, the sequence lengths of the alpaca-7b, chavinlo-alpaca-13b, and bigscience-bloomz-7b datasets are significantly below the average.

4.2. Experimental setup

In this study, we chose the DeBERTa-base[19] model as our pre-trained base model. We set the hyperparameters as follows: the batch size is set to 16, the maximum sequence length is set to 512 (with sequences longer than this being truncated), and the margin is set to 0.5. The initial learning rate is set to 2e-5, and we train for 3 epochs. We use AdamW for optimization during each training session. During the training phase, we use the officially provided labeled dataset to train the model. To evaluate the model's performance across different domains, we use the HC3 dataset [20] during the validation phase. The results of our model on our validation set Table1

Table 1
Results of our model on the validation set we used ROC-AUC, Brier, C@1, F_1, $F_{0.5u}$ and their mean.

ROC-AUC	Brier	C@1	F_1	$F_{0.5u}$	Mean
0.998	0.972	0.991	0.974	0.973	0.981

4.3. Result

We selected the model with the best performance in validation, tested it on TIRA [9], and scored all test tasks separately. The combined results for the test dataset are presented in the following Table3 and Table 2.

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
merciless-broth	0.98	0.945	0.954	0.932	0.935	0.949
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 2 shows the results, initially pre-filled with the official baselines provided by the PAN organizers and summary statistics of all submissions to the task (i.e., the maximum, median, minimum, and 95-th, 75-th, and 25-th percentiles over all submissions to the task).

Table 3 shows the summarized results averaged (arithmetic mean) over 10 variants of the test dataset. Each dataset variant applies one potential technique to measure the robustness of authorship verification approaches, e.g., switching the text encoding, translating the text, switchign the domain, manual obfuscation by humans, etc. Please focus your description on the discussion of the results on the main dataset, e.g., Table 2. I.e., Table 3 is only here for your completeness, please discuss only the details on the main dataset (i.e., Table 2). A detailed description of all dataset variants will be available in the overview notebook.

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
merciless-broth	0.601	0.859	0.945	0.978	0.987
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

5. Conclusions

In this paper, we propose a method combining contrastive learning and meta-learning to address the task set by PAN: Voight-Kampff Generative AI Authorship Verification. Our proposed method achieved scores of roc-auc: 0.98, brier: 0.945, c@1: 0.954, F1: 0.93, F0.5u: 0.935, and Mean: 0.949 on the leaderboard. These results validate the effectiveness of our proposed method in the task of Generative AI Authorship Verification.

Acknowledgments

This research was supported by the Natural Science Platforms and Projects of Guangdong Province Ordinary Universities (KeyField Special Projects) (No. 2023ZDZX1023)

References

[1] A. Extance, Chatgpt has entered the classroom: how llms could transform education, Nature 623 (2023) 474–477.

[2] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, et al., Ethical and social risks of harm from language models, arXiv preprint arXiv:2112.04359 (2021).

[3] J. Wu, S. Yang, R. Zhan, Y. Yuan, D. F. Wong, L. S. Chao, A survey on llm-gernerated text detection: Necessity, methods, and future directions, arXiv preprint arXiv:2310.14724 (2023).

[4] 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.

[5] T. Hospedales, A. Antoniou, P. Micaelli, A. Storkey, Meta-learning in neural networks: A survey, IEEE transactions on pattern analysis and machine intelligence 44 (2021) 5149–5169.

[6] 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.

[7] 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.

[8] A. A. Ayele, N. Babakov, J. Bevendorff, X. B. Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korenčić, M. Mayerl, D. Moskovskiy, A. Mukherjee, A. Panchenko, M. Potthast, F. Rangel, N. Rizwan, P. Rosso, F. Schneider, A. Smirnova, E. Stamatatos, E. Stakovskii, B. Stein, M. Taulé, D. Ustalov, X. Wang, M. Wiegmann, S. M. Yimam, 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.

[9] Y. Arase, M. Zhou, Machine translation detection from monolingual web-text, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2013, pp. 1597–1607.

[10] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, C. Finn, Detectgpt: Zero-shot machine-generated text detection using probability curvature, in: International Conference on Machine Learning, PMLR, 2023, pp. 24950–24962.

[11] J. Su, T. Y. Zhuo, D. Wang, P. Nakov, Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text, arXiv preprint arXiv:2306.05540 (2023).

[12] 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.

[13] A. Bhattacharjee, T. Kumarage, R. Moraffah, H. Liu, Conda: Contrastive domain adaptation for ai-generated text detection, arXiv preprint arXiv:2309.03992 (2023).

[14] M. Ibrahim, A. Akram, M. Radwan, R. Ayman, M. Abd-El-Hameed, N. El-Makky, M. Torki, Enhancing Authorship Verification using Sentence-Transformers, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2023, pp. 2640–2651. URL: https://ceur-ws.org/Vol-3497/paper-216.pdf.

[15] M. Guo, Z. Han, H. Chen, H. Qi, A contrastive learning of sample pairs for authorship verification, Working Notes of CLEF (2023).

[16] M. Boudiaf, J. Rony, I. M. Ziko, E. Granger, M. Pedersoli, P. Piantanida, I. B. Ayed, A unifying mutual information view of metric learning: cross-entropy vs. pairwise losses, in: European conference on computer vision, Springer, 2020, pp. 548–564.

[17] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, 2020. arXiv:2002.05709.

[18] A. Nichol, J. Achiam, J. Schulman, On first-order meta-learning algorithms, arXiv preprint arXiv:1803.02999 (2018).

[19] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, arXiv preprint arXiv:2006.03654 (2020).

[20] B. Guo, X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, Y. Wu, How close is chatgpt to human experts? comparison corpus, evaluation, and detection, arXiv preprint arXiv:2301.07597 (2023).