Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ30590

Research on Performance Optimization of BERT Model in AI-Generated Text Detection

Jiarui Yu

College of Computer Science and Technology (College of Data Science), Taiyuan University of Technology, Taiyuan, China
yujiarui998@126.com

Abstract. With the rapid development of large language models, the text generated by artificial intelligence (AI) is increasingly approaching the human level in terms of fluency, coherence and factuality. It brings unprecedented challenges to many fields such as information security and academic integrity. Therefore, the research on AI-generated text detection technology has become crucial. In recent years, detection methods based on pre-trained language models have shown significant advantages and become a research hotspot in this field. This paper reviews the representative research progress on the performance optimization of Bidirectional Encoder Representations from Transformers (BERT) model in the field of AI-generated text detection in recent years. It also discusses the limitations of the current BERT detection model, and looks forward to future research directions. Research analysis shows that the detection performance and robustness of BERT model in cross-domain and AI generation model can be significantly improved by using optimization strategies such as feature fusion and adversarial training. These results provide a valuable reference for the further development of AI-generated text detection technology.

Keywords: BERT, AI-generated text detection, Performance optimization, Deep learning, Natural language processing

1. Introduction

The breakthrough of AI-generated content technology, especially the wide application of large language models represented by GPT series, ChatGPT, Gemini, etc., has greatly improved the quality and diversity of machine-generated text [1]. These models can generate news, comments, academic abstracts and even programming codes that are almost fake. Although this has brought improvement in production efficiency, it has also caused a series of serious social problems such as the spread of false information, academic misconduct, and online fraud [2]. Therefore, the development of mature AI-generated text detection technology is of great practical significance for maintaining a clear cyberspace, ensuring academic integrity, and building a credible AI ecosystem.

To cope with this challenge, researchers have turned their attention to the pre-trained Transformer model represented by Bidirectional Encoder Representations from Transformers (BERT) [3]. Through self-supervised pre-training on large-scale corpus, BERT provides new possibilities for human-computer text discrimination due to its powerful contextual semantic understanding ability

DOI: 10.54254/2755-2721/2026.TJ30590

[4]. Current research progress shows that detection models based on BERT and its variants have achieved leading performance in multiple benchmarks.

This paper adopts the research method of literature review to systematically sort out and deeply analyze the technical route and optimization strategy of BERT-based detection models. By summarizing the advantages and disadvantages of the existing optimization schemes, this paper puts forward predictions and prospects for the development of general and robust next generation AI text detection technology in the future. It provides technical reference and framework for subsequent researchers to develop more powerful detector models, and

2. Theoretical basis

The fundamental reason for detecting AI-generated text lies in the fact that there are subtle yet identifiable differences in statistical features and language patterns between machine-generated text and human-written text. Human writing is usually more creative, emotionally volatile, and logically occasionally inconsistent, while machine-generated text may exhibit abnormal smoothness, specific word distribution bias, and excessive regularity in certain grammatical structures [4]. It is based on these assumptions that traditional detection methods capture differences by hand-engineered features, such as perplexity, word frequency distribution (e.g., Boers' law), grammatical error rate, etc. However, when these features are faced with large and carefully tuned language models, the discrimination becomes very fuzzy.

Compared with traditional methods, end-to-end methods based on deep learning, especially BERT-based models, are able to automatically learn high-level, abstract feature representations from raw text without relying on tedious feature engineering. Through its multi-layer Transformer encoder structure, BERT is able to understand the meaning of each word in both directions in the context. To capture the text more effectively long-range dependence and complex semantic model [3]. This ability allows it to identify "deep fingerprints" of subtle deficits in semantic consistency, anaphora resolution, and commonsense reasoning in machine text.

Following BERT, a number of enhanced pre-trained models were put out and used for detecting tasks, establishing the current standard research techniques. For example, RoBERTa further improves the semantic understanding ability of the model by removing the next sentence prediction task and using larger batches and more data for training [5]. DeBERTa, on the other hand, introduces a decoupled attention mechanism and an enhanced mask decoder, which outperforms BERT on multiple natural language understanding tasks [6]. These models obtain strong language priors by pre-training on general corpus, and then adapt them into efficient binary classification detectors by supervised fine-tuning on a specific human-machine text pair dataset.

3. Application of BERT model in AI-generated text detection

3.1. Model training

In long text processing, the piecewise fine-tuning strategy can be used to cope with BERT's 512-word input limit, and the DocBERT model is derived from it. In the specific implementation, the long text is divided into multiple sub-segments according to an natural paragraph or a fixed length (such as 512 words), and each sub-segment is independently input into the BERT model to obtain the hidden state vector of [CLS] token [7]. Then, the features of each sub-segment are aggregated through the attention pooling mechanism. This aggregation layer dynamically assigns the importance of different sub-segments based on learnable weight parameters, and finally fuses global

DOI: 10.54254/2755-2721/2026.TJ30590

semantic information to generate the document-level representation. This design is verified to be effective in the Reuters-21578 multi-label classification task. Compared with the traditional sliding window method, it reduces the amount of calculation by about 40% while maintaining the classification accuracy. Figure 1 shows the comparative performance curve of the knowledge distillation compression model KD-LSTMreg and BERTbase/BERTlarge.

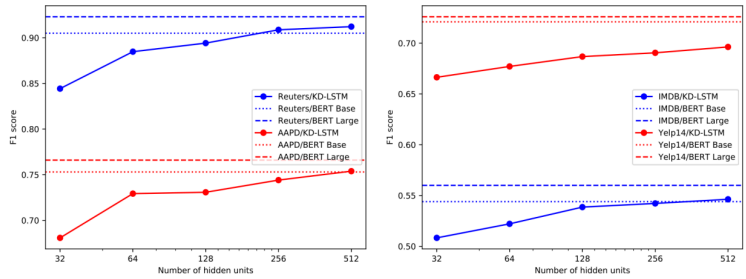

Figure 1. Effectiveness of KD-LSTMreg vs. BERTbase and BERTlarge [7]

Second, the most commonly used fine-tuning objective is the cross-entropy loss function. In order to improve the robustness and generalization ability of the model, researchers have introduced a variety of advanced training strategies. For example, adversarial training methods can significantly improve the resistance of the model to slight rewriting and adversarial attacks. This is achieved by introducing small perturbations to the input embedding layer and allowing the model train on both adversarial samples and original samples [8]. A contrastive learning method constructs positive and negative sample pairs. This allows the model to learn that the similarity between human texts and AI texts is greater than that between human-computer texts, thereby enhancing the discrimination of the feature space [9]. Moreover, the detection task can be jointly trained with auxiliary tasks, such as text genre classification and sentiment analysis. This forces the model to learn more generalized feature representations and prevents it from overfitting to the specific biases of the training data.

3.2. Datasets

BERT models rely on reliable datasets to train and evaluate AI-generated text detection models. Researchers can use commonly used datasets to train BERT models on AI text detection. The HC3 dataset is a large-scale bilingual Chinese and English dataset. It contains the answers of human experts and ChatGPT to the same series of questions. This database is an important benchmark for evaluating the ability of models to detect ChatGPT's generated text [10]. The GPT-2 Output dataset, published by Radford et al., includes news articles generated by GPT-2 models of different scales and written by humans. It is a widely used benchmark in early studies [1]. The TURINGBench dataset is a comprehensive benchmark containing text generated from multiple models (e.g., GPT-2, GPT-3, Grover) in different domains, which facilitates the system to evaluate the generalization ability of detectors [11]. The M4 dataset is a large-scale multilingual detection dataset covering 10 languages. It is created by a variety of text generation models and is crucial for research on cross-language detection [12]. Other notable datasets include the GLUE dataset [3].

3.3. Model limitations and improvements

The BERT model also has limitations. Relying solely on the semantic features of BERT is not enough to capture all the differences. The reason is that the computational cost is high, and the fine-

DOI: 10.54254/2755-2721/2026.TJ30590

tuning and inference of large BERT models require a lot of computing resources, which makes it difficult to deploy in edge devices or scenarios that require real-time detection. The performance comparison between DistilBERT and BERT is shown in Table 1 [13]. Secondly, the upper limit of BERT model ability is limited by pre-training, and its own semantic space is not optimal for modeling human-computer text differences. BERT is also very vulnerable to adversarial attacks, and many BERT-based detectors can be bypassed by slight perturbations such as simple word substitution and sentence pattern reconstruction on AI-generated text [14]. Finally, the detectors trained for one dataset or for a certain type of generative model have insufficient generalization ability, and their performance deteriorates seriously when applied to other fields or new large models.

Table 1. Comparison results of the development set in the GLUE benchmark test [13]

Model	Score	CoLA	MNLI	MRPC	QNLI	QQP	RTE	SST-2	STS-B	WNLI
ELMo	68.7	44.1	68.6	76.6	71.1	86.2	53.4	91.5	70.4	56.3
BERT-base	79.5	56.3	86.7	88.6	91.8	89.6	69.3	92.7	89.0	53.5
DistilBERT	77.0	51.3	82.2	87.5	89.2	88.5	59.9	91.3	86.9	56.3

Furthermore, optimizing model training paradigm is also a key way to improve generalization ability. It is an effective strategy to learn text source features through multi-class pre-training. For example, instead of adopting a simple human-machine binary classification, Zellers et al. trained models to distinguish whether text came from a human or a specific multiple generative model (e.g. GPT-2, Grover, etc.) [15]. This training method forces the model to learn more general and essential machine text features instead of memorizing model-specific biases, which significantly improves the zero-shot and few-shot detection ability of the model when facing unknown generative models.

Another important method, is improving the model architecture and training objective. For example, in the M4 benchmark model, by using the contrastive learning framework, the model can effectively learn the "source" features of the text by narrowing the representation of the same source text and moving the representation of the different source text far away, thus significantly improving its zero-shot and few-shot detection ability in the face of unknown generative models [12].

4. Case study of BERT model performance optimization

The GLTR tool developed by Gehrmann et al. is a classic detection case of fusing shallow statistical features. Although it is not directly fine-tuning BERT, its idea has profoundly influenced the subsequent optimization direction [4]. This tool analyzes the probability rank of each word in the text in the prediction results of large language models (such as GPT-2), and visualizes it to assist human judgment. The statistical features revealed by GLTR, such as the distribution of word units in the Top-1, Top-10, and Top-100 intervals, provide strong discriminative clues for distinguishing human-computer texts.

Inspired by this, subsequent studies have extensively explored the fusion strategy of concatenate the statistical features provided by GLTR with the [CLS] semantic representation vector of BERT and other models, and then input it into the classifier. The essence of this optimization strategy is the complementarity and fusion of shallow statistical features and deep semantic features. BERT excels at understanding contextual semantics and logical consistency, while GLTR features directly expose the statistical bias of the generative model in word-level selection. Multiple studies have proved that this kind of fusion model usually has significantly better F1 score and AUC value than a single

DOI: 10.54254/2755-2721/2026.TJ30590

model when detecting texts generated by GPT-2, Grover and other models. In particular, when facing slightly rewritten adversarial texts, it can show stronger robustness.

5. Conclusion

This paper systematically reviews the research on the performance optimization of BERT model in AI-generated text detection. In general, the pre-trained Transformer model represented by BERT shows an overwhelming advantage over traditional methods on this task due to its deep context understanding ability, and has become the current mainstream technology route. By adopting advanced fine-tuning strategies, constructing large-scale and diverse datasets, and implementing various feature and architecture optimization schemes, researchers continue to push the boundaries of detection performance.

However, this article also has some limitations, such as the limited number of literatures covered, not exhaustive research, and insufficient discussion depth for some emerging optimization techniques (e.g., reinforcement learning-based optimization). These are also directions that can be supplemented and improved by future research.

Looking ahead, AI-generated text detection is an ongoing technological game. Relying solely on the passive upgrade of BERT model is difficult to achieve a decisive victory. Future development will rely on cross-domain integration and innovation, including the design of more robust model architectures, the construction of dynamic evolution of the evaluation system, and the in-depth exploration of the essence of human-computer text differences. Only through the continuous joint efforts of academia and industry can we jointly maintain the authenticity and security of digital information.

References

[1] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

[2] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

[3] Kenton, J. D. M. W. C., & Toutanova, L. K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of naacL-HLT (Vol. 1, No. 2).

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

[5] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A prototype optimized bert pretraining approach. arXiv preprint arXiv: 1907.11692.

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

[7] Adhikari, A., Ram, A., Tang, R., & Lin, J. (2019). Docbert: Bert for document classification. arXiv preprint arXiv: 1904.08398.

[8] Miyato, T., Maeda, S. I., Koyama, M., & Ishii, S. (2018). Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE transactions on pattern analysis and machine intelligence, 41(8), 1979-1993.

[9] Yang, X., Cheng, W., Wu, Y., Petzold, L., Wang, W. Y., & Chen, H. (2023). Dna-gpt: Divergent n-gram analysis for training-free detection of gpt-generated text. arXiv preprint arXiv: 2305.17359.

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

[11] Uchendu, A., Ma, Z., Le, T., Zhang, R., & Lee, D. (2021). Turingbench: A benchmark environment for turing test in the age of neural text generation. arXiv preprint arXiv: 2109.13296.

[12] Wang, Y., Mansurov, J., Ivanov, P., Su, J., Shelmanov, A., Tsvigun, A., ... & Nakov, P. (2023). M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. arXiv preprint arXiv: 2305.14902.

DOI: 10.54254/2755-2721/2026.TJ30590

[13] Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv: 1910.01108.
[14] Abdelnabi, S., & Fritz, M. (2021, May). Adversarial watermarking transformer: Towards tracing text provenance with data hiding. In 2021 IEEE Symposium on Security and Privacy (SP) (pp. 121-140). IEEE.
[15] Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., & Choi, Y. (2019). Defending against neural fake news. Advances in neural information processing systems, 32.