Uncertainty Estimation Methods for Countering Attacks on Machine-Generated Text Detectors

Valeriy Levanov[1], **Anastasia Voznyuk**[1], and **Andrey Grabovoy**[1]

[1]Moscow Institute of Physics and Technology, Moscow

April 10, 2025

Abstract

In this work, we explore the use of uncertainty estimation methods to enhance the quality of machine-generated text detectors against various attacks, such as homoglyphs, paraphrasing and noise injection. These attacks are not only used to bypass detection but also to test the resilience of detectors. We will test the hypothesis that uncertainty estimation can provide a more sustainable approach, eliminating the need for constant retraining. The research will evaluate this hypothesis in two scenarios: when only the text is available and when access to the model's internal states is also provided. The planned experiments aim to validate the results and compare them with current state-of-the-art solutions.

1 Introduction

Recent advancements in large language models (LLMs) allow for the easy creation of coherent texts that are almost have not diffence from those written by humans. Despite the wide range of efficient applications of generative models for society, there is also a high risk of their misuse for spreading misinformation or solving students' homework. Therefore, there is a need for effective methods to discern machine-generated text from human-written text. This task is highly challenging due to the diversity of models, their generation styles and different texts domains. Also various types of attacks pose particular difficulties for detection, and even the simplest of these attacks can significantly reduce the accuracy of effective detectors.

The main approach to finding differences between texts will be estimation of uncertainty. This method has already been investigated for various NLP tasks such as machine translation (MT), text summarization (TS), and question answering (QA)[1]. Uncertainty estimation is also proven effective for detecting generated images by analyzing the distributions of natural images [2].

In this paper, we aim to combine an approach with uncertainty estimation for the task of detecting machine-generated text. We will examine whether the representations of the machine and human text models differ from each other. This research could serve as a foundation for developing attack-resistant detectors capable of identifying machine-generated text with high accuracy.

2 Problem statement

Formally, the problem can be described as follows:

Given a set of texts $\{x_n\}$ with binary labels $\{y_n\}$, where the labels indicate whether the text is machine-generated or human-written. The research will explore two approaches to calculating uncertainty.

2.1 White-box methods

In this case, the model $F(x)$ can be represented as a composition of mappings $f_1(x)$ and $f_2(x)$, where $f_1(x)$ maps the text to some internal representation of the model, and $f_2(x)$ maps this representation to the model's prediction. In the white-box method, for a specific text, we can observe both $f_1(x)$ and $F(x)$, which means that the model allows us to examine its internal states during prediction. The goal is to use uncertainty estimation methods to study whether texts with different labels cluster together. In this case, uncertainty can be estimated in various ways[1], which require some knowledge of the internal workings of the model.

2.2 Black-box methods

In many modern models, internal states and architecture are not available for study, but even in this case there are many methods for estimating uncertainty based only on the response model $F(x)$. Black-box methods were also used in the work[1]. The task is to study these methods and compare their effectiveness with the white-box methods.

3 Computational Experiment

The primary objective is to compare binary classification based on uncertainty estimates with alternative methods for detecting machine-generated text. To do this, we will use part of the M4GT[3] dataset as data.

We processed texts from both labels using the LLM (Llama-3.1-8B-Instruct) and utilized contextual logits to compute four metrics: perplexity, mean token entropy, monte carlo sequence entropy and mahalanobis distance. For consistency of the experiment, all data was taken from Reddit. The results can be seen in Table 1. Some metrics show pronounced variation between human- and machine-generated texts. This finding is promising, as it suggests that the distinction between MG and HW text distributions can potentially be detected using uncertainty estimation (UE). In the following sections, we explore this idea further to develop a more robust detection framework.

Metric	Label	Mean	Median	Std	Min	Max
mean entropy	Human	0.0045	0.0045	0.0005	0.0023	0.0060
mean entropy	Machine	0.0034	0.0035	0.0009	0.0004	0.0067
perplexity	Human	2.6851	2.6658	0.3378	1.6291	3.8900
perplexity	Machine	2.1756	2.1370	0.4377	1.0716	4.5481
mc entropy	Human	429.206	452.375	115.155	164.375	695.500
mc entropy	Machine	216.063	190.125	98.2335	19.0625	674.500
mahalanobis	Human	1465.920	1400.880	350.174	357.836	4817.470
mahalanobis	Machine	2029.660	1983.840	469.339	996.327	6088.490

Table 1: Metrics for human and machine-generated texts

The next step involves training a simple neural network classifier. A lightweight architecture comprising several linear layers was selected and implemented. We will compare this model with RoBERTa for sequence classification. The model will be fine-tuned on the same training dataset, and the results will be compared.

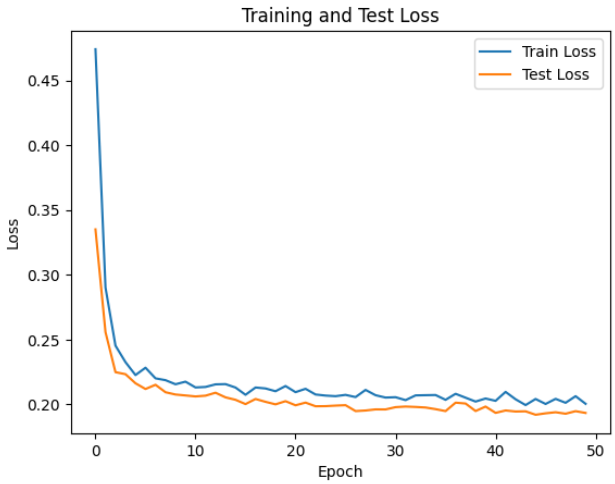

Figure 1: Loss for Simple Classifier

Model	Accuracy	Precision	Recall	Train Time (s)
RobertaForSequenceClassification	0.9901	0.9901	0.9992	594.4377
neural network classifier with uncertainty	0.9136	0.9455	0.9542	30.1410

Table 2: Comparison of model performance metrics

Table 2 presents the results. While our model achieves a good accuracy, it does not outperform RoBERTa. This discrepancy may be attributed to limitations in the classifier itself. Therefore, by selecting a different classifier model, performance could be improved. However, it is noteworthy that the training time of our classifier is significantly shorter.

4 Method

4.1 Uncertainty Calculation

Uncertainty estimation is a key method we use to analyze texts in our paper. One of the main approaches is information-based methods that work with the model's output probabilities. These methods include:

1) **Perplexity (PPL)** - measures how surprised the model is by the text:

$$PPL = \exp\left(-\frac{1}{L}\sum_{l=1}^{L}\log P(w_l|w_{<l})\right)$$

2) **Mean token entropy** - calculates average uncertainty per token:

$$H = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j} P(w_j|w_{<i})\log P(w_j|w_{<i})$$

3) **Monte Carlo Sequence Entropy** - we run the model multiple times with small random changes and average the results:

$$H_S(x;\theta) = -\frac{1}{K}\sum_{k=1}^{K}\log P(y^{(k)}|x,\theta)$$

Another approach uses **density-based methods**. We first analyze the model's hidden states from the training data - calculating their mean and variance. Then, we compare these learned patterns to the hidden states of new test texts to measure how different they are.

The main method here is **Mahalanobis Distance**:

$$MD(x) = (h(x)-\mu)^T\Sigma^{-1}(h(x)-\mu)$$

4.2 LLM

It is very important to choose the model through which the logits of the text contexts will be obtained. The choice fell on **Llama-3-8B-Instruct**. Its main advantages are its relative lightness and access to the internal representations of the context. Text will be sent to the input of the model without additional requests. The interesting output of the model is the context logits, and we will usually focus on the k most likely logits to simplify calculations.

4.3 Datasets

The basis of the research is the availability of extensive datasets. An important requirement for them will be the presence of texts from multiple domains and several generation models. These qualities help to increase the robustness and accuracy of the detectors.

M4GT[3] is designed for the task of binary classification and contains 65,177 human-written texts and 73,288 machine-generated texts. It includes several domains (Reddit, wikiHow, Wikipedia) and generative models (GPT-4, Cohere, Dolly).

RAID[4] is a massive dataset of 6 million text examples from a wide range of domains and models. During generation, various decoding strategies and repetition penalties were used, significantly increasing text diversity. Its key feature is the inclusion of adversarial attacks on texts, which aids in training robust detectors.

5 Conclusion

References

[1] A. Tsvigun A. Vazhentsev S. Petrakov E. Fadeeva, R. Vashurin. Lm-polygraph: Uncertainty estimation for language models, 2023.

[2] T. Liu Y. Cheung Bo Han X. Tian J. Nie, Y. Zhang. Detecting discrepancies between ai-generated and natural images using uncertainty, 2024.

[3] Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohanned Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, et al. M4gt-bench: Evaluation benchmark for black-box machine-generated text detection. *to appear in ACL 2024*, 2024.

[4] F. Trhlik J. M. Ludan A. Zhu H. Xu D. Ippolito C. Callison-Burch L. Dugan, A. Hwang. Raid: A shared benchmark for robust evaluation of machine-generated text detectors., 2024.