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

March 13, 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 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.