Adversarial Learning for AI Text Detection

Aryan Vats (avats31@gatech.edu), Deeksha Manjunath (deekshamanjunath@gatech.edu), Reshma Ramachandra (reshmaram@gatech.edu)

Georgia Institute of Technology

SUMMARY

Lately, Large Language Models (LLMs) are being explored quite extensively for text generation. However, with the advent of such technologies, it is very important to prevent misinformation.

Thus, detection of text generated by Artificial Intelligence (AI) is necessary to distinguish between actual and fake information.

This project proposes a model that can identify if the given text was AI-generated or not using **adversarial learning**.

Inspired by the RADAR framework, we train two networks, **an ensemble paraphraser** and a **detector**, that compete with one another to learn and enhance the efficiency of the detection model.

The proposed approach **outperforms** the RADAR implementation on the Wiki Intro Dataset with an **AUROC score of 0.78**.

OBJECTIVES

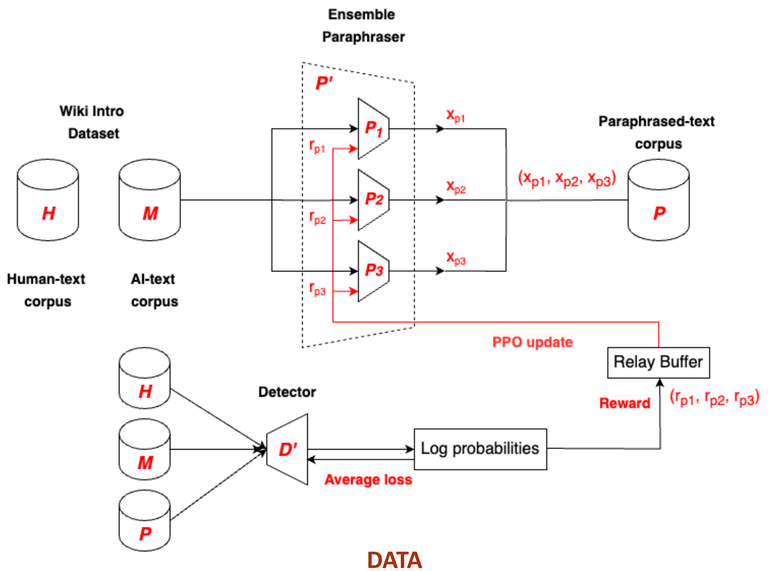
The main objectives include:

- **RoBERTa-based Detector**: Design a robust AI-text detection model trained using Adversarial Learning that provides accurate results across various text-generation models.
- **Ensemble Paraphraser**: Combine the outputs of the fine-tuned Text-to-Text Transfer Transformer (**T5-large**) model, the Bidirectional and Auto-Regressive Transformers (**BART**) model and the **Pegasus** model to enhance the quality of the information fed to the detector model.
- The detector and the ensemble paraphraser are trained in an **adversarial** manner, where each model influences the training of the other.
- The **Proximal Policy Optimization** (PPO) algorithm is used to reward the paraphraser based on the detector's prediction outputs.
- Compare this approach against a baseline method that solely focuses on training the detector network without employing adversarial learning.

APPROACH

- The Wiki Intro Dataset constitutes the *H* corpus, and the corresponding AI-generated text constitutes the *M* corpus.
  - Samples x_m from *M* are fed into the Ensemble Paraphraser **P'** .
  - The human-written text samples x_h, AI-generated text samples x_m and the paraphrased texts x_p are fed into the Detector **D'** which aims to predict the probability of given input text being human-generated.
  - Finally, the paraphraser and detector models are updated until there is no improvement in the AUROC.
  - With these updates, the new reward can be defined as:

$$R(x_p, \phi) = \mathcal{D'}_{\phi}(x_p) \in [0, 1]$$



DATA

The **Wiki Intro Dataset** is used to train the model. These samples cover various domains, collected from Wikipedia.

It comprises of 150k topics, with a varied **distribution** of **machine-generated text**, created by GPT (Curie) model and **human-written text**.

A prompt was used to generate the GPT response using the title of the Wikipedia page and the first seven words from the introduction paragraph. Prompt used for generating text:

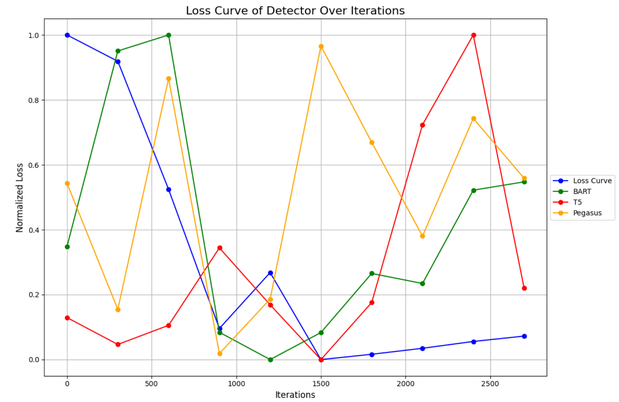
200 word Wikipedia-style introduction on '\{title\}'

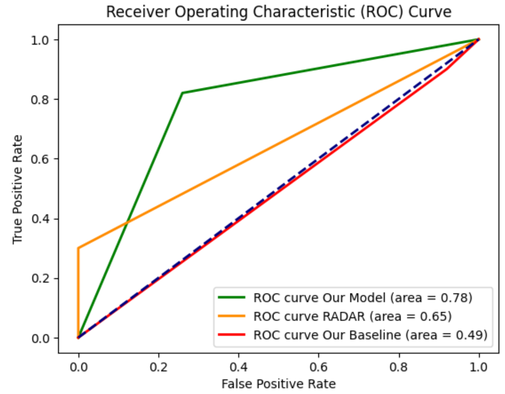
\{starter_text\}

where \{title\} is the title for the Wikipedia page, and \{starter_text\} is the first seven words of the Wikipedia introduction.

With **equal amounts** of data belonging to both the human-written and AI-generated classes, the Wiki Intro Dataset seems promising to effectively train models for AI text detection.

RESULTS





CONCLUSION

- In this research, we delved into AI text detection using adversarial learning, leveraging the RADAR framework as our guiding research.
- The proposed approach demonstrated **good generalization** across different domains within the dataset, encompassing factual introductions and scientific explanations.
- From experimentation shown above, the proposed approach outperforms the baseline implementation which did not involve any adversarial learning and the RADAR approach.
- This work holds potential for combating misinformation and enhancing reliability on online information.

REFERENCES

RADAR Framework: Hu, X., Chen, P. Y., & Ho, T. Y. (2023). Radar: Robust ai-text detection via adversarial learning. arXiv preprint arXiv:2307.03838.**Baseline RoBERTa approach**: Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Wiki Intro Dataset: https://huggingface.co/datasets/aadityaubhat/GPT-wiki-intro