Deep learning methods for AI-Generated Text Detection

Introduction

Because of the rise of ChatGPT and GPT-4, people started using generative models in different fields, including writing essays and giving different ideas. However, some people are over relying ChatGPT and just use it to do the essay writing without any human work inside, which may have the problem of using the other content unintendedly due to the nature of training the LLM. Meanwhile, detecting if the article is AI or human is a huge topic for a long time. However, most of the research only use the training dataset which consist of purely human written text and AI-generated text, which ignore the case of mixture of them happening more often in real life. In view of this, the project aims to investigate different deep learning methods' performance on detecting the AI-generated text in different cases: 1. Length of text 2. Power of the LLM 3. Purely / mixed AI-text

The idea is form reading the paper A Survey on LLM-generated Text Detection: Necessity, Methods, and Future Directions and the Kaggle competition LLM - Detect AI Generated Text. And the aim is to build a small model having a relative good performance when doing the same problem.

While the mixing AI and human text problem is from the paper MixSet: Official code repository for Mixset (github page of the paper).

Data set

The data for this project will consist of a balanced dataset containing AI-generated texts and human-written texts. For the training process, I will use M4 dataset: it is an expansive benchmark corpus spanning multiple generators, domains and languages. While containing from diverse sources like wiki pages, reddit and arXiv pages. The LLM-generated texts in M4 are also crafted from different models like ChatGPT, LLaMa, BLOOMz, FlanT5, and Dolly. In total, it has 122k for both human and AI text for training. Nonetheless, in this project, only the English data will be used. Meanwhile, the Mixset dataset will be used for the testing process. It is a relatively small dataset for mixed text case involving both AI and human-generated content. It involves 3.6k instances involves in email, news, game review and in total 6 categories where both the human written text and the machine generated text are further been employed either one of the six operations by GPT-4 or Llama2-70b for mixing the content. For the AI-text generated by GPT4, the dataset from MGTBench will be

used. The project will use the 1000 sample of essay and 1000 samples of crafted stories generated by GPT4all to test the model's performance.

In this project, various learning methods will be experimented with to detect AI-generated text, aiming to enhance the performance of small-scale LLMs to match that of larger LLMs. The methodologies include:

Traditional Methods:
Hybrid CNN-RNN approach: Aim to implement a text classification model by combining CNN and RNN (i.e. Long Short-term Memory LSTM) networks. To capture both local features using CNN and sequential nature using RNN. The idea is mainly from the paper "Fake News Identification on Twitter with Hybrid CNN and RNN Models". Modification: changing the layer architectures and structures.
SVMs: I will use SVMs with RBF kernels for baseline text classification. To improve the performance, I will also apply the TF-IDF vectorization to transform texts into feature vectors. Modification: Using other kernel functions

The idea of CNN-RNN approach models from the following papers:
Stacked CNN - LSTM approach for prediction of suicidal ideation on social media
Indonesian Sentiment Analysis: An Experimental Study of Machine Learning and Deep Learning Methods Fake news detection: A hybrid CNN-RNN based deep learning approach
Combine Convolution with Recurrent Networks for Text Classification

Zero-Shot Learning: I will also utilize the OpenAI API to use ChatGPT3.5 for text classification using zero-shot learning, mainly to act as baseline to compare to the fine-tune LLMs.
The idea comes from ZeroGPT for detecting the AI-text

Fine-Tuning LLM: Due to the limited resources, I will try to fine-tune a smaller language model first, like miniCPM 2B. The paper shows the small scall model somehow be comparable to other 7B and even 20B model in certain field. I will use with bf16 training to improve efficiency, DeepSpeed and QLoRa for low-resource environments, coupled with mixed-precision training to reduce computational demands and training time. The pre-trained models can be downloaded from HuggingFace. (If allowed, maybe I would fine-tune RoBERTa instead). Modification: Original model did not test for text classification so training on a customize dataset.

Results:
In the experiment, I will first evaluate all the model's performance on 1. Detection of text generated by GPT-3.5 2. Detection of text generated by GPT-4 3. Detection of Mixed AI-Human texts using AUC and F1-scores. Then I will plot the curve of their performance in detection AI-text in different length. The project assumes that 1. CNN-RNN is better than SVM, fine-tune LLM better than zero-shot GPT. Training the models with purely AI text will have a worse performance when testing with mixed AI-generated text. Also, the project will invest the performance in different types of field of data (news, blog…) and the out-of-distribution dataset, and will use attention maps to understand the decision-making process.

(This is the original written version. I just feed in this document to gpt to refine the proposal based on the guidelines provided in lecture notes. Then, I further add the points in the refined proposal that I find they are missing / need to mentioned. Finally, I summarized the dataset information into the table in the refined version.)