Proposal for SpotGPT: Variational Autoencoder-Based Machine-Generated Text Detection

Ahmed Abdellatif & Achinth Bharadwaj (as part of CPSC 440)

March 24, 2023

Project Abstract

Generative Pre-trained Transformers (GPTs) are large language models (LLMs) trained on large amounts of data using unsupervised learning methods. Their transformer architectures enable them to capture long-range dependencies in text and generate highly coherent and contextually relevant language. As machine-generated text becomes increasingly prevalent, distinguishing between human and machine-generated text has become crucial. While some methods exist for detecting machine-generated text (e.g., GPTZero, DetectGPT), none have yet utilized Bayesian-based approaches to approximate GPT detection. This project proposes a novel machine learning method that combines variational autoencoders, natural language processing algorithms, and pre-trained transformers to analyze the text's syntactic and semantic features. We aim to develop a robust scoring mechanism that yields a likelihood or score of the possibility of being a body of machine-generated text.

1 Background & Motivating Problems

The misuse of generative pre-trained transformers has become a growing concern in journalism, academic integrity, and social media misinformation campaigns. GPTs are powerful language models that can generate high-quality text, making it difficult to distinguish between human-generated and machine-generated content. This has led to an increase in the use of GPTs for unethical purposes such as academic fraud and spreading misinformation.

According to a report by the MIT Technology Review, 'We may be witnessing, in real time, the birth of a snowball of [expletive]' as a result of the introduction of more machine-generated text, which are capable of regurgitating falsehoods, and in turn, ultimately skew the dissemination and perception of information (Heikkilä, 2022) [3]. Similarly, in academic circles, there have been instances of students using GPTs to generate fake essays and academic papers 'spark[ing] significant academic integrity concerns in higher education' (Sullivan et al., 2023) [6].

Since the inception of ChatGPT in November 2022, novel research methods have been released, with the creation of GPTZero by Edward Tian and the Princeton NLP lab, and DetectGPT (Mitchell et al., 2023) [4]. The latter is a zero-shot method which relies on probability curvature of the transformer and text pertubation to estimate a body of text that has been generated by a GPT-based machine or not. However, we were unable to find any research projects which use unsupervised learning or probablistic assumptions with available training data that were published in recent discourse.

The DetectGPT paper will be heavily cited throughout our project, and their assumptions and research experiments have given us inspiration to lead our investigation, basing off this team's research.

2 Hypothesis

Similar to the DetectGPT paper, we cite a similar hypothesis:

Machine-Generated Text (MGT) algorithms generate the most likely body of text given their input (based on the principle of Maximum Likelihood Estimation). This means - given some likelihood function (and its parameters), the generated output is likely to sit at some optimum of their respective functions. Human-generated text (HGT) does not obey such strict rules of maxima and do not exhibit perfect function-approximable behaviour. Exploiting these two tendencies allows us learn the parameters of the text-generating distribution through some approximator. If we can empirically verify our hypothesis, we can then use the parameters of these functions to reconstruct our data as see if it is likely to be MGT or not.

3 Methods

3.1 Requirements

1. **English GPT-generated text data**: In order to conduct our research, we will be needing voluminous bodies of text from a variety of topics and writing styles that has been generated by a family of GPTs (including but not limited to GPT3.5 and beyond). A certain Aaditya Bhat on HuggingFace has published a GPT-generated dataset of articles that were written by humans about a variety of topics on Wikipedia, along with their ChatGPT-generated counterparts [1]. This will be included in our experiments, but will certainly not be the only source of data. All sources of data will be thoroughly cleaned and standardized before encoding and model training. For purposes of reliability and timespan, we will be limiting our model to the English language.

2. **Variational Autoencoders**: VAEs are a family of autoencoders which faithfully generate new data points that are similar to its inputs. It is a generative model that can learn the underlying distribution of the data and sample new data points successively. The encoder network maps the input data to a lower-dimensional latent space representation, while the decoder network maps the latent space representation back to the input space. From our hypothesis, the latent-space distribution we aim to learn in the VAE is the the parameters of the text-generating distribution that GPTs obey. If the VAE is trained on GPT data, we expect the model to faithfully regenerate the body of text since it will have successfully learned the underlying distribution parameters. We expect HGT to reconstruct poorly (since the learned parameters will be related to the MGT text distribution).

3. **OpenAI's logprobs API endpoint**: in order for us to evaluate our scoring metric:

   $\frac{\mathcal{S}(x)}{\mathcal{S}(\overline{x}}$ where $\mathcal{S}(i)$ is the log-probability of the candidate text i and $\overline{x}$ is the text generated by our VAE

   we need to be able to calculate the log-probability, which OpenAI thankfully provides as part of their GPT API specifications. Our expectation is that ratio ad mentioned above, if the VAE can learn the distribution and generate the reconstruction perfectly, this means that the log-odds score specified should be around one.

3.2 Experiments

1. Split GPT-generated text dataset into training and testing data. Collect test samples of human-generated text that are similar in length, topic, and dialect to GPT-generated text to augment our VAE. Our test data will also consist of GPT-generated test data and human-generated text.

2. Choose a suitable word embedding (if needed) for the training data. Candidates for trial include but are not limited to Bag of Words (BoW), Dict2Vec, Word2Vec and so forth.

3. Train variational autoencoders on GPT-generated training data. The architecture that we are proposing is a hybrid model, that was suggested in Bowman et. al., 2015 [2] and Semeniuta et al, 2017 [5].

We will be training our model on various sizes of architectures to achieve a variety of performances to compare.

4. Reconstruct our test data (human and GPT-generated) through trained VAEs.
5. Use OpenAI's logprobs API endpoint to generate the following log probabilities for each observation in our test set. We will be calculating two scores: The log probability for the original test data texts, and the log probability for VAE-reconstructed test data texts.
6. Calculate probability ratio for each observation by dividing log probability of original data by that of reconstructed data.
7. Compare probability ratio with text labels:
   - If probability ratio is close to 1, consider text to be machine-generated. If it is close to 0, consider it to be human-generated.
8. Lastly, if needed, transforming the log-probability radio by simply raising Euler s constant e to our odds-score, and a negative scaling:

$$-1 * e^{\frac{S(x)}{S(\bar{x})}} : [0,1] \to [0,\infty]$$

would change our output domain from $[0,1]$ to $[-\infty,0]$ to $[0,\infty]$ which would express our faith that the higher the transformed output score, the higher the likelihood that it was generated by a human. The lower the score, the more likely that it was generated by a large language model.

4 Contributions and food for thought

While VAEs have been widely used for tasks such as image generation, denoising, and compression, their application in detecting the origin of text data is relatively unexplored. This study showcases the potential of VAEs as a valuable tool for tackling the challenge of identifying GPT-generated content. We also would like our research to contribute to existing literature in detecting text generated by large language models, given that the future of such a niche and novel system has not been reliably tested, with its limits unknown and scalability unexplored with potential dangers to thwart the status quo of the human condition.

5 Bibliography

References

[1] Aaditya Bhat. Gpt-wiki-intro (revision 0e458f5), 2023.

[2] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. Generating sentences from a continuous space. *CoRR*, abs/1511.06349, 2015.

[3] Melissa Heikkilä. How ai-generated text is poisoning the internet. *MIT Technology Review*, 2022.

[4] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature, 2023.

[5] Stanislau Semeniuta, Aliaksei Severyn, and Erhardt Barth. A hybrid convolutional variational autoencoder for text generation. *CoRR*, abs/1702.02390, 2017.

[6] Miriam Sullivan, Andrew Kelly, and Paul McLaughlan. Chatgpt in higher education: Considerations for academic integrity and student learning. *Journal of Applied Learning Teaching*, 6(1):1, 2023.