Exploring Classifying Features for a RoBERTa-based Classifier by Calculating Attribution Scores

{CSC3160} {Default} Project

Cheng Yang (Student ID, 120090195)
Fintech, Chinese University of Hong Kong, Shenzhen
120090195@link.cuhk.edu.cn

Abstract

ChatGPT's powerful writing ability has been widely used in people's daily life. Thus, It is important to distinguish which text is generated by ChatGPT. This paper aims to study a RoBERTa-based [1] ChatGPT Classifier, about the features it captures to detect a GPT-generated text. We adopted DeepLift algorithm to calculate the attribution score[2] of each token. DeepLift is a modified version of Integrated Gradients (IG)[3]. Their results are approximate, but Deeplift has a higher speed of calculation. We first conducted a statistical analysis of the dataset in terms of sentiment, part of speech and word frequency, and obtained certain rules (e.g., ChatGPT's preference of repetition). Then we tried to verify these rules by calculating the attribution score and visualizing the distribution of the attribution score for each variable. However, the distribution of the attribution score did not show any certain rules. Regarding the difference between the results, we will summarize the observed rules and analyze the reasons for the poor effect of the attribution score calculated by DeepLift. The code is in URL: https://github.com/SLPcourse/CSC3160-120090195-ChengYang/blob/main/RoBERTa_Attribution_Analysis.ipynb

1 Key Information to include

- I will attend the poster session.

2 Introduction

ChatGPT is one of the most popular large language models (LLM) now. Because of its excellent writing ability and dialogue ability, it is widely used in daily study and work. However, we also need to prevent people from excessive use of ChatGPT in some situations, such as preventing student papers from copying ChatGPT and ensuring the originality of novels and papers. Detecting whether a text is generated by ChatGPT has evolved into an important task. There are already many kinds of ChatGPT detectors, such as GPT-Zero and GPT Detector [4]. However, since these classifiers are also based on deep learning models, few model publishers can explain which features of the text their models are based on. Some publishers may have counted sentiment and other features of texts in their models' data set, but it is difficult to explain the features based on the output of the model. Some scholars tried visualizing the attention to interpret the classifier, but the multi-layer results are hard to understand.

In order to explore the features captured by classifier by combining the input and output of classifier, we calculate the attribution score of each token. A positive attribution score means that the token has a positive effect on the predicted label of the whole text, while a negative number means a negative effect. The common calculation methods of attribution score include Integrated Gradients (IG), Saliency and DeepLift. We use DeepLift algorithm to calculate the attribution score, which is similar to IG but has a faster operation speed.

The model we focus is the baseline model offered by Dr. Jiang Feng which reproduces the ChatGPT Detector. Our project begins with a statistical analysis of the dataset in terms of sentiment, part of speech and word frequency. We summarize the differences between human generated text and ChatGPT generated text in these three aspects by drawing line plots and box plots. Then we visualized the distribution of attribution score with these three variables, hoping to get the same conclusion. Unfortunately, it is difficult to observe regularity in the results obtained by investigating attribution score. We have to rethink whether the attribution score calculated by deeplift can be used to explain the RoBERTa classification model and the reason of incompatibility.

3 Related Work

3.1 Detecting Methods

The two main current methods for detecting ChatGPT-generated text are the black-box method and the white-box method[5]. For external identity, black-box is preferred since white-box method needs high-level access permission for the language generating model. The baseline model applies a black-box method named RoBERTa to work on the task, an updated version of the BERT model. The transformers in the RoBERTa model involve some potential features that function well for detection.

3.2 Observed Rules

Some studies point out that human-written texts have a wider vocabulary. And ChatGPT uses more attributives, conjunctions and auxiliaries while keeping a neutral sentiment[4]. These observations motivate us to study the terms of sentiment, part of speech and word frequency.

3.3 Attribution Score

The attribution score we focus is calculated by the Integrated Gradients (IG) method initially. Later, a new method named DeepLIFT is published. It decomposes the output prediction of a neural network on a specific input by back-propagating the contributions of all neurons in the network to features of the input. The two methods are related well that their results are approximate while DeepLIFT has a higher speed. However, there are seldom researches adopting attribution score to explore BERT or RoBERTa model. Moreover, some scholars think that there are crucial logic traps in these evaluation methods are ignored, causing the unfair evaluation[6]. Based on these controversies and the results of this project, we will investigate the compatibility of attribution score calculated by DeepLIFT method.

4 Approach

In this section, we present our workflow for exploring the classifying features and the compatibility of DeepLIFT. We will first provide an overview of our approach and explain DeepLIFT algorithm in detail.

4.1 Overview

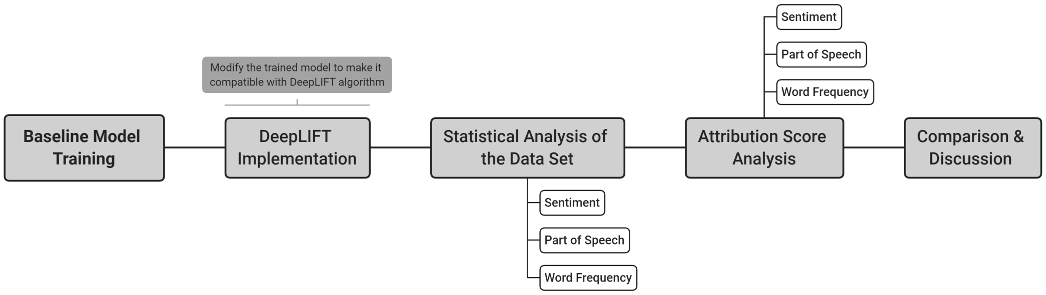


Figure 1: WorkFlow Overview

As shown in Figure.1, our approach begins with training the baseline model offered by Dr. Jiang Feng. Next, we adopt DeepLIFT algorithm to calculate the attribution scores. Then we analyze the dataset and the distribution of attribution scores in terms of sentiment, part of speech and word frequency. Finally, we compare the results and discuss our findings.

4.2 DeepLIFT Algorithm

DeepLIFT is an algorithm for decomposing the output prediction of a neural network on a specific input by backpropagating the contributions of all neurons in the network to every feature of the input. It compares the activation of each neuron to its 'reference activation' and assigns contribution scores according to the difference.[3]

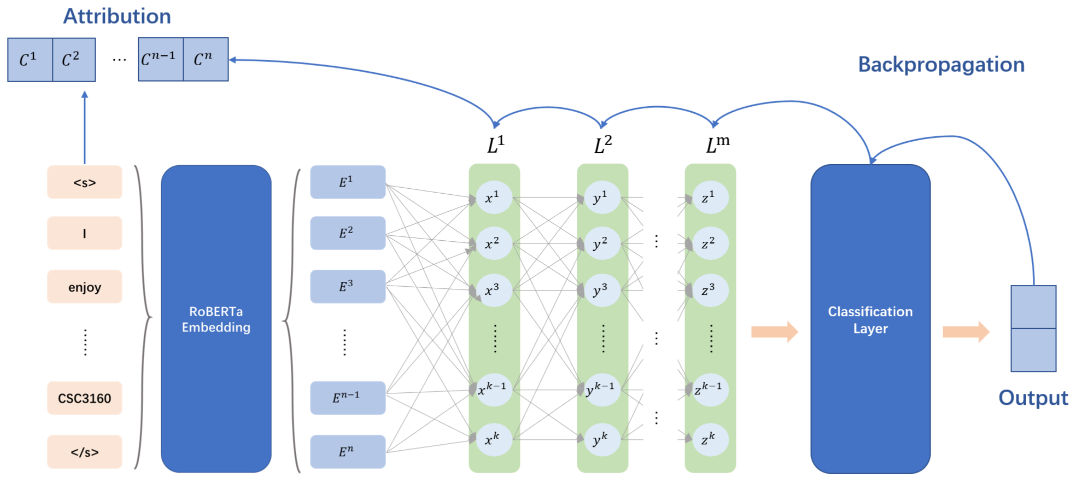


Figure 2: Combination of RoBERTa Classifier and DeepLIFT Algorithm

We provide a simplified version of DeepLIFT process. We use t to denote a target output which can be calculated through the set of layers in Figure.2. And we have a reference activation of t denoted by t_0. Δt is defined as the difference-from-reference that $\Delta t = t - t_0$. Assuming that x_1,x_2, and x_k are the neurons in the first input layer, we assign contribution score $C_{\Delta x_j \Delta t}$ to Δx_j s.t.

$$\sum_{j=1}^{k} C_{\Delta x_j \Delta t} = \Delta t$$

Next, for the convenience of subsequent usage of the Chain Rule, we define a multiplier $m_{\Delta x \Delta t}$ as:

$$m_{\Delta x \Delta t} = \frac{C_{\Delta x \Delta t}}{\Delta x}$$

According to the Chain Rule, we can get that:

$$m_{\Delta x_j \Delta t} = \sum_{h=1}^{k} m_{\Delta x_j \Delta y_h} m_{\Delta y_h \Delta t}$$

Finally, we can use backpropagation to derive the contribution score $C_{\Delta x_j}$ for each neuron j in first input layer. Then we transform the $C_{\Delta x_j}$ into attribution score C_i for each token i.

5 Experiments

This section contains experimental setups and results.

5.1 Data

Our dataset is sourced from HC3-English. Since there was no publicly available dataset split mentioned in the paper, we have divided the dataset into a typical 80% training set and 20% test set, with 10% of the training set reserved for validation set. To ensure the classifier's generalizability, the data sources are various and belongs to different domain as shown in Table.1. We only use the answers and their label (Human or ChatGPT generated).

5.2 Evaluation method

For the statistical analysis of the dataset, we split the dataset into human-generated and ChatGPT-generated. We use AFINN sentiment analysis to calculate the average sentiment score of each token in the two groups. And we evaluate the average proportion of different parts of speech and draw the box plot. Besides, we count the number of occurrences of the most frequent words in each text and split these words into different groups base on their part of speech. We only focus on NOUN, ADJ, PRON, ADV, and VERB since they are more importantant in a sentence.

For attribution score analysis, we visualize the distribution of attribution score with different sentiment score and part of speech. To explore the effect of word frequency, we calculate the average attribution score for those most frequent words.

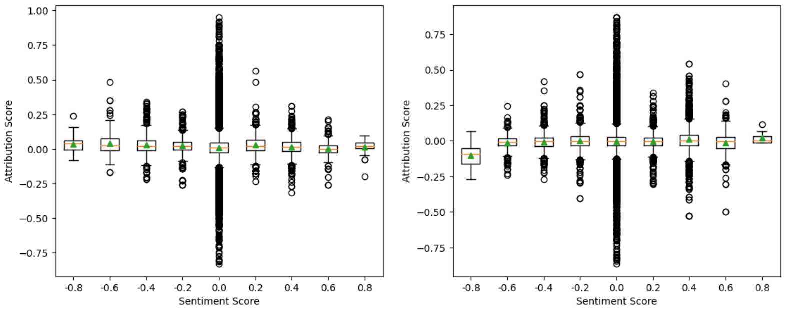
5.3 Experimental details

Foe the statistical analysis of the dataset, we invetigate the whole training set. We store every token's information about sentiment, part of speech and word frequency into several DataFrames. The aggregated information could be helpful for further visualization. We can combine them with each token's attribution score in the next part.

For the attribution score analysis, we adopt DeepLIFT algorithm to calculate the result. The parameter $target$ is automatically set as the target class since it is the common setting for classification problem. For the parameter $baselines$. we don't provide a custom tensor since we don't have enough background knowledge to assign an appropriate $baselines$. Thus, the $baselines$ is default as zero scalar corresponding to each input tensor. This may become a reason for further difficulty of interpreting the attribution score. We choose 1000 texts from each of the human-generated and ChatGPT-generated text training set to evaluate. The smaller sample scale is due to the limitation of Colab's GPU RAM. We think 1000 texts are enough to conclude the rules of whole training set.

5.4 Results

Sentiment The average sentiment score for human-generated text is 0.0058 while the average sentiment score for ChatGPT-generated text is 0.0015, which means **human tends to use more emotional words**. However, the distribution of attribution score with sentiment score did not show obvious difference between the two groups. As shown in Figure.3, the similar results indicate that non-neural token cannot contribute to predict the text sourced from either human or ChatGPT.

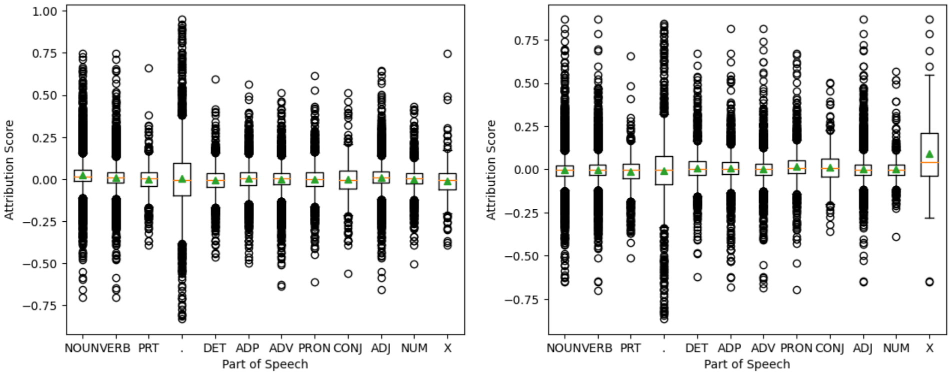


(a) Human

(b) ChatGPT

Figure 3: The Distribution of Attribution Scores with Sentiment Scores

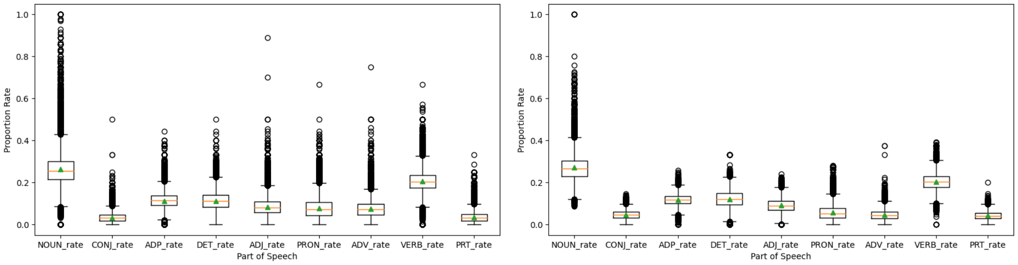
Part of Speech The average proportion rate for different part of speech for the two sources are approximate. However, we find that **human-generated text has more outliers while ChatGPT always allocate a stable proportion to each part of speech** based on Figure.4. However, it is hard to conclude the relation between attribution score with part of speech according to Figure.5. Besides, The Table.1 shows that **the repetitive pron and verb contribute to predict ChatGPT-detected text while repetitive noun and adjective contribute to predict human-detected text**.



(a) Human

(b) ChatGPT

Figure 4: The Distribution of Attribution Scores with Part of Speech



(a) Human

(b) ChatGPT

Figure 5: The Distribution of Proportion Rate with Part of Speech

Word Frequency According to the results in Table.1, ChatGPT are more likely to repeat a certain word than humans, especially for noun and verb. However, the attribution scores for these most frequent words show a difference regarding different part of speech. As mention before, the repetitive pron and verb contribute to predict ChatGPT-detected text while repetitive noun and adjective contribute to predict human-detected text. This is contradictory to the counting of occurrences.

Table 1: Statistics about the Most Frequent Word

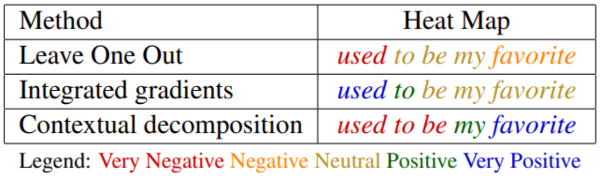
	Occurrencess of the Most Frequent Word					**Attribution Score of the Most Frequent Word**				
	NOUN	ADJ	PRON	ADV	VERB	Attri_NOUN	Attri_ADJ	Attri_PRON	Attri_ADV	Attri_VERB
Human	3.538	1.814	3.536	1.831	3.297	**6.56%**	**1.84%**	-0.38%	0.16%	**-2.86%**
ChatGPT	6.050	2.865	4.300	2.144	5.606	**-7.32%**	-0.11%	**4.13%**	1.43%	0.88%

6 Analysis

The key findings derived from the statistical analysis of the training set includes: 1) Human-generated text consists of more emotional words. 2) ChatGPT has a more stable allocation of proportion rate on different part of speech. 3) ChatGPT are more likely to repeat a certain word than human, especially for noun and verb.

However, the attribution score calculated by DeepLIFT algorithm can hardly reproduce these findings. The distributions of attribution score with token's sentiment are very similar. Thus, we cannot conclude any rules in terms of sentiment based on attribution scores. Thus, we need to disscuss why we cannot achieve the expected findings and rethink the reliability of attribution score.

Algorithm Selection One potential reason is the inappropriate selection of algorithms. We only adopt DeepLIFT algorithm because of its high speed rather than compare different algorithm's results. The performance of different algorithms could differ a lot. As shown in Figure.6, Ju et al. (2023) [6]provide a heat map to illustrate the different attribution score for each token based on different algorithms. Thus, the poor explaining power of the project's result may be due to the incompatibility of DeepLIFT with this problem.



Method	Heat Map
Leave One Out	*used to be my favorite*
Integrated gradients	*used to be my favorite*
Contextual decomposition	*used to be my favorite*

Legend: Very Negative Negative Neutral Positive Very Positive

Figure 6: Heat maps for a portion of a yelp review generated by different attribution algorithms.

Baseline Setting Another potential reason for the poor performance of DeePLIFT is the improvable setting of parameter $baseline$. Though Ancona et al. (2017) [7]claim that the defalut setting of zero is reasonable, alough arbitrary, they also point out that a great baseline could improve the performance of DeepLIFT. The current baseline improvement are mainly focus on images rather than texts. This is because capturing the background of a image is easier than exploring the baseline of a text.

Additionally, although the attribution scores of the most frequent word show differences in terms of noun, adjective, pronoun and verb, we cannot arbitrarily generalize any rules based on the current unreliable attribution score. We can try diffrent algorithms and settings to validate these findings.

7 Conclusion

In this work, we conclude that 1) human's answer consists of more emotional word; 2) ChatGPT has a more stable proportion for each part of speech; 3) ChatGPT is more likely to repeat a certain noun or adjective. Specially, our project provide a insight to analyze the most frequent word's attribution score. The limitation is that we fail to reproduce these findings by evaluating the attribution score calculated by DeepLIFT algorithm. However, this failure inspires us to rethink the setting of $baseline$ in DeepLIFT and other algorithm for NLP problem. For future work, we will adopt different algorithms on our evaluation framework. Moreover, we will customize the setting of $baseline$ by summarizing the statistical rules we concluded from the dataset. Such a exploration of $baseline$ could be extended to other NLP domains for better interpretability of models.

References

[1] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[2] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[3] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR, 2017.

[4] Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*, 2023.

[5] Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. The science of detecting llm-generated texts. *arXiv preprint arXiv:2303.07205*, 2023.

[6] Yiming Ju, Yuanzhe Zhang, Zhao Yang, Zhongtao Jiang, Kang Liu, and Jun Zhao. Logic traps in evaluating attribution scores. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5911–5922, 2022.

[7] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104*, 2017.