Comparison of the Robustness of Multimodal Models and Unimodal Models under Text-based Adversarial Attacks

Yizhi Liu

School of Foreign Languages and Literature, Wuhan University, Wuhan, China

2022301021003@whu.edu.cn

Abstract. With the popularization of artificial intelligence technology, adversarial attacks have become a major challenge in the field of machine learning. This paper explores the robustness of multimodal and unimodal models under textual adversarial attacks, and probes to understand their differences and commonalities. By comparing and analyzing the performance of the CLIP multimodal model and BERT unimodal model under different textual datasets, it is pointed out that the multimodal model does not perform better than the unimodal model under unimodal adversarial attack when the multimodal fusion advantage cannot be reflected. On the contrary, the CLIP model, which is a multimodal model, exhibits larger robustness fluctuations similar to the BERT model under single-modal adversarial attacks. The advantages of multimodal models do not automatically translate into better robustness in all scenarios but need to be optimized for specific tasks and adversarial strategies, and the multimodal models do not have better accuracy than the unimodal models without task-specific pre-training. Both exhibit significant robustness fluctuations in the face of textual adversarial attacks. The research in this paper provides research value and further research directions for future studies

Keywords: Multimodal model, Unimodal model, Adversarial attack, Robustness.

1. Introduction

With the development of artificial intelligence in recent years, adversarial attacks have gradually become an important field. Adversarial attacks refer to the construction of adversarial samples by an attacker by adding some small, imperceptible perturbations to the input data of a machine learning model, which can cause the model to output wrong classification or prediction results with high confidence. Initial research on adversarial attacks focused on unimodal models, especially image classification models. Ian Goodfellow et al. proposed in 2014 that Fast Gradient Sign Method (FGSM) generates adversarial samples by calculating the gradient of the input data and adding small perturbations [1]. Madry et al. proposed in 2017 that, Projected Gradient Descent (PGD) is an extension of FGSM, which is optimized through multiple iterations to generate stronger adversarial samples [2]. And with the development of multimodal models in recent years, people have to start focusing on the adversarial robustness of these models. Researchers have proposed a variety of adversarial attack methods for multimodal models. For example, the VLATTACK method attacks pre-trained visual language models by generating perturbations on visual and textual modalities [3]. Meanwhile, researchers have been exploring ways to deal with these adversarial attacks, such as several defense strategies summarized in the review by Wei such as adversarial training, input transformations, model enhancement, and detection mechanisms [4].

The main difference between multimodal and unimodal models in this respect is the type of data they handle and the way they are processed. Multimodal models are capable of fusing data from different perceptual modalities, such as vision, language, etc., to obtain more comprehensive information. This ability to fuse data gives multimodal models an advantage when dealing with complex tasks. However, this advantage does not mean that multimodal models are more stable than unimodal models in all cases. Whether multimodal models actually have greater adversarial robustness in practical applications, and what the sources and mechanisms of this robustness are, need to be further explored. If the multimodal model is unable to take advantage of multimodal fusion, the potential advantages and possible limitations of multimodal learning compared to unimodal models under single-modal adversarial attacks can be investigated through comparative analysis. With the wide application of Artificial Intelligence (AI) technology, especially the rapid development of multimodal technology in the fields of automatic driving, medical image analysis, smart home, etc., its adversarial robustness is directly related to the reliability and safety of the technology. Therefore, it is of great significance to deeply explore the performance of multimodal models under adversarial attacks to enhance the practical application of these technologies. In this study, a representative multimodal model CLIP and an unimodal model Bert are selected as the experimental objects for further experimental analysis by using the method of comparative analysis.

2. Research methodology

2.1. Model and dataset selection

Firstly, in this paper, the openai/clip-vit-large-patch14 model [5] and the google-bert/bert-base-uncased model [6] are selected as the multimodal and unimodal experimental models, respectively. the CLIP model is a multimodal pre-trained neural network model released by OpenAI in 2021 that is pre-trained with a large number of Internet image-text pairs to learn the alignment relationship between images and text. And BERT model is a pre-trained language model proposed by Google, which learns to get rich contextual representation by pre-training on large-scale text data. Since both CLIP model and BERT model are capable of text processing, some text datasets such as IMDB [7], AG's News [8], YELP [8] are selected for this paper, in which AG's News topic categorization dataset includes a large number of news sources collected from more than two thousand News corpus, containing four categories, namely world, sports, business, science/technology; IMDB is a document-level movie review sentiment classification dataset, containing both positive and negative categories; Yelp sentiment analysis dataset is constructed based on reviews on Yelp website, containing both positive and negative categories. both positive and negative categories. Here in this paper, after determining the experimental model and experimental dataset, the dataset is processed and then handed over to the CLIP model and BERT model respectively, and the corresponding predictions are output.

2.2. Visualization and classification

After obtaining the feature files corresponding to the dataset processed by the model respectively, this paper uses PCA to downscale and visualize them in 3D. At the same time as downscaling and visualization, this paper also uses the K-Nearest Neighbor algorithm to complete the classification of the feature files, which can evaluate the accuracy of the corresponding model in processing this dataset. By combining the visualized images and the runtime results this paper can get the accuracy performance of the model under the corresponding dataset and its antagonistic samples.

2.3. Repeated Experiments with Adversarial Samples

After having the evaluation results of the original dataset, this paper can generate the confrontation samples for the selected dataset. Here this paper adopts the type of confrontation sample of proximate word replacement, for example, a part of the data can be randomly extracted from IMDB, and proximate word replacement can be adopted in order to generate confrontation samples. When some of the important words in the original data are replaced with near-synonyms to get the confrontation samples, the model may get wrong classification results by accepting the confrontation samples as inputs while ensuring that the semantics are not greatly shifted. In terms of the BERT model, this paper uses three versions of BERT pre-trained by Di Jin et al. through the IMDB, AG, and YELP datasets respectively for experiments, and also uses the confrontation samples corresponding to the IMDB, AG, and YELP datasets obtained by them through TextFooler [9]. Meanwhile, in this study, considering that the CLIP model as a multimodal model is generally considered to be more advanced and should be more sensitive to the prediction and changes in the dataset, the same experiments were conducted on the three datasets and their adversarial samples using the original encapsulated CLIP model. Since the best uni-modal network often outperforms the multi-modal network, this observation is consistent across the different combinations of modalities and on the other hand, the CLIP model is used to conduct the same experiments. different combinations of modalities and on different tasks and benchmarks for video classification" [10]. Therefore, this study explores the robustness and also takes this opportunity to observe whether the CLIP model, as a more advanced multi-modal model, has a better classification performance on unfamiliar untrained datasets. Repeating the above process for these adversarial samples will be able to obtain the visualization of the adversarial samples under model processing and the stability of the model.

3. Findings

3.1. Bert-based visualization results

First is the visualization of the performance of the BERT model pre-trained with the corresponding datasets under the three datasets and their adversarial samples.

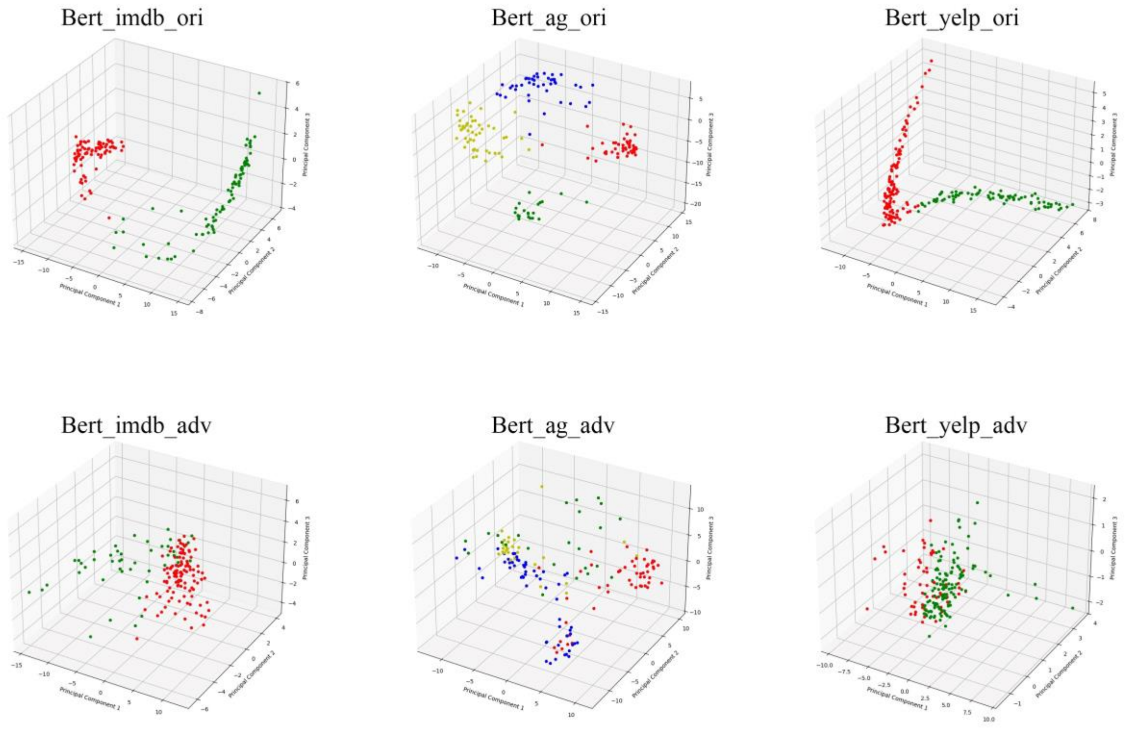

Figure 1. Bert-based visualization results

From Figure1, it can be seen that the BERT model pre-trained on the corresponding dataset shows a fairly high accuracy on the original dataset, and the labels can be clearly classified. After the attack, it can be clearly seen that the classification accuracy of the BERT model has decreased significantly, which is in line with the expectation of the attack in this paper.

3.2. Clip-based visualization results

Similarly, this paper also obtains the visualization performance of the CLIP model under three datasets and their adversarial samples.

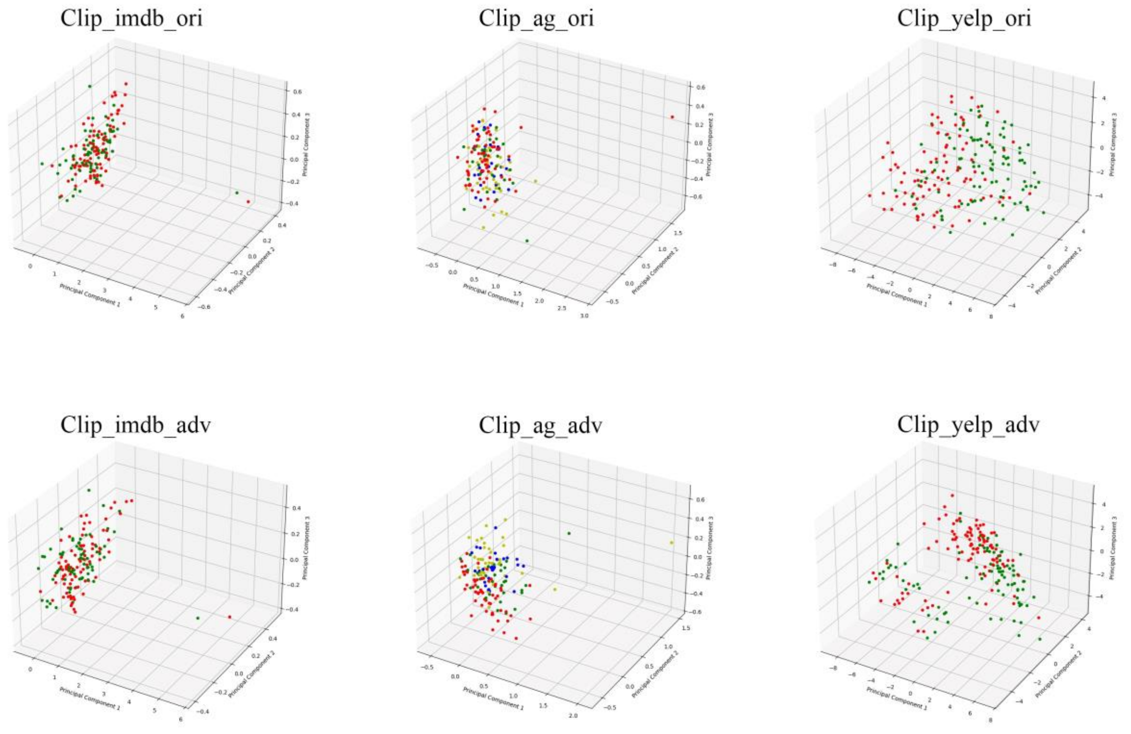

Figure 2. Clip-based visualization results

As shown in Figure 2, in fact, it can be seen through the scatter plot that the CLIP model is not good enough to classify the labels on the original dataset in the untrained state, so much so that the visualization image does not show the obvious change in the accuracy before and after the confrontation. So this paper also takes the method of comparing the accuracy rate obtained by the KNN algorithm to further explore the fluctuation of the robustness change of the two models.

3.3. Results

Table 1. Results

		BERT			CLIP	
	IMDB	AG	YELP	IMDB	AG	YELP
Original Accuracy	99	99	99	61	65	73
After-Attack Accuracy	41	49	60	44	25	28
Difference	58	50	39	17	40	45

As shown in Table 1, this paper clearly obtains the accuracy of the two models on the three datasets and their confrontation samples as well as the before and after changes. From this paper, it is found that even though the CLIP model is not pre-trained and thus does not perform as well as the pre-trained BERT model on the original dataset, the CLIP shows unexpectedly low accuracy on the confrontation samples, and it produces a robustness degradation before and after the confrontation that is similar to that of the BERT model.

4. Conclusion

First, the study in this paper found that the CLIP model does not classify as well as the pre-trained BERT model on the original dataset without task-specific pre-training. This suggests that while multimodal models have the potential to fuse information from multiple modalities, their performance on a specific task is still highly dependent on the relevance of the pre-trained dataset and the task.

Second, when faced with textual adversarial attacks, the CLIP model exhibits large robustness fluctuations similar to those of the BERT model. Specifically, although the CLIP model has the theoretical advantage of cross-modal fusion, this advantage does not translate into better robustness performance under adversarial attacks against textual modalities only. On the contrary, the accuracy of both models on the adversarial samples shows a significant decrease, indicating that under the current adversarial attack techniques, it is difficult for either multimodal or unimodal models to be completely immune to the interference of textual adversarial samples.

This finding has important implications for understanding the behavioral patterns of multimodal models in complex adversarial environments. It suggests that the advantages of multimodal fusion do not automatically translate into better robustness in all scenarios but need to be optimized for specific tasks and confrontation strategies. In addition, it also reminds us that in practical applications, people cannot solely rely on the cross-modal fusion capability of multimodal models to guarantee their adversarial robustness but need to take multiple means to improve the stability and reliability of the models in conjunction with specific task requirements.

Finally, this study provides useful insights for further improving the adversarial robustness of multimodal models in the future. On the one hand, researchers can try to better adapt multimodal models to specific tasks and adversarial environments by optimizing their pre-training strategies; on the other hand, researchers can also explore new adversarial defense techniques, such as adversarial training and defensive distillation, in order to improve the classification accuracy and stability of the models under adversarial samples.

In summary, this paper reveals the advantages and limitations of the CLIP multimodal model and BERT unimodal model in specific scenarios by comparing and analyzing their robustness performance under textual adversarial attacks. Future research can further explore the optimization methods and adversarial defense techniques of multimodal models to enhance their practical application in complex tasks.

References

[1] Ian, J. G., Jonathon, S., and Christian, S. (2015). Explaining and Harnessing Adversarial Examples, 2015. *arXiv:1412.6572v3.*

[2] Aleksander, M., Aleksandar, M., Ludwig, S., Dimitris, T., and Adrian, V. (2018). Towards Deep Learning Models Resistant to Adversarial Attacks, *arXiv:1706.06083v4*.

[3] Yin, Z. Y., Ye, M. C. et al., (2023). VLATTACK: Multimodal Adversarial Attacks on Vision-Language Tasks via Pre-trained Models, 2023. In *Advances in Neural Information Processing Systems,36* (pp. 52936-52956).

[4] Emma, Z., Quan, Z. S., Ahoud, A., Chen, L. (2020). Adversarial Attacks on Deep-learning Models in Natural Language Processing: A Survey, 2020. In *ACM Transactions on Intelligent Systems and Technology (TIST), Volume 11, Issue 3* (pp. 1-41)

[5] Alec, R., Wook, K., Chris, H., Aditya, R., Gabriel, G., Sandhini, A., Girish, S., Amanda, A., Pamela, M., Jack, C., Gretchen, K., Ilya, S. (2021). Learning Transferable Visual Models From Natural Language Supervision, 2021. In *Proceedings of the 38th International Conference on Machine Learning*, PMLR 139:8748-8763.

[6] Jacob, D., Ming, W. C., Kenton, L., Kristina, T. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019. *arXiv:1810.04805*.

[7] Andrew, L. M., Raymond, E. D., Peter, T. P., Dan, H., Andrew, Y., and Christopher, P. (2011). Learning Word Vectors for Sentiment Analysis, 2011. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 142-150).

[8] Xiang, Z., Jun, Z., Yann, L. (2015). Character-level Convolutional Networks for Text Classification, 2015. In *Advances in Neural Information Processing Systems 28 (NIPS 2015).*

[9] Jin, D., Jin, Z., Zhou, J. T., & Szolovits, P. (2020). Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment, 2020. In *Proceedings of the AAAI Conference on Artificial Intelligence,* 34(05), 8018-8025.

[10] Wei, W., Du, T., Matt, F. (2020). What Makes Training Multi-Modal Classification Networks Hard, 2020. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 12695-12705).