Assessing the Adversarial Robustness of Multimodal Medical AI Systems: Insights into Vulnerabilities and Modality Interactions

(Executive Summary)

Ekaterina Mozhegova, Asad Khattak*, Adil Khan, Roman Garaev, Bader Rasheed

*College of Technological Innovation, Zayed University, Dubai, UAE
Asad.Khattak@zu.ac.ae

1. Introduction

Deep learning systems have shown rapid development and are currently being extensively applied in a wide range of fields, including healthcare, where the reliance on diverse data types - texts, images, numeric recordings is essential. Task-specific models are commonly employed for analyzing these data types. Recently, general-purpose multimodal large models have emerged, offering the potential to process these different data simultaneously. Despite their potential, deep learning models often remain vulnerable to adversarial attacks. These attacks involve small, often imperceptible, perturbations to the input data, capable of misleading model predictions. Due to the healthcare realm being an area with high demands to systems accuracy and robustness, it is crucial to thoroughly understand the vulnerabilities of these models to ensure their reliability and safety in medical applications.

In our research, we contribute to the fields of multimodality and adversarial robustness, with a specific focus on applications in medicine. As we investigate the behavior of multimodal models under various attack scenarios, this research sheds light on how these models endure attacks and whether the multimodal nature of these models enhances their resilience. We conducted experiments by applying attacks to two modalities: images and texts. We applied attacks to single-modality models, combined these models with modality fusion techniques

and applied the same attacks on the combined model. This approach allowed us to validate our hypothesis that multimodality enhances robustness to adversarial attacks.
Our findings can be valuable for future research aimed at creating robust multimodal systems. Additionally, this study can provide foundation for further research focused on the data flow in multimodal systems.

2. Research Question(s)

This research investigates the robustness of multimodal systems within the healthcare domain. Specifically, we explore how different modalities are affected by adversarial attacks when applied separately and how the multimodal models are affected by attacks on different modalities simultaneously. We also explore how interactions between modalities influence the overall adversarial robustness of AI systems.

3. Research Methods

Experiment setup overview

In this section, we present an overview of our experimental setup.
We initially constructed two separate models: a vision model M_{V} and a language model M_{L}. We then combine M_{V} and M_{L} to create a multimodal model, M_{VL}, resulting in three distinct models. These models were used to apply adversarial attacks and evaluate their robustness under different attack scenarios.
First, we implement a Fast Gradient Sign Method (FGSM) attack on the visual model, denoted as $A_{\mathrm{I}}^{\mathrm{FGSM}}$.
We apply attacks on the language model, which include synonym substitution, denoted as $A_{\mathrm{T}}^{\mathrm{SYN}}$, and word deletion, $A_{\mathrm{T}}^{\mathrm{DEL}}$. $A_{\mathrm{T}}^{\mathrm{DEL}}$ involves removing a fraction of the words from the text caption. We tested half-word deletion, where 50% of the words were removed. $A_{\mathrm{T}}^{\mathrm{SYN}}$ involves replacing a fraction of the words in the text caption with their synonyms. We tested substitution fractions of 20% and 40%.
For the multimodal model M_{VL}, we test each of the mentioned attacks individually, then combinations of them, and finally apply all attacks simultaneously to target both modalities.

Dataset

We use a multimodal dataset incorporating chest X-ray images accompanied by text captions. This dataset was collected by Indiana University and includes patient IDs, frontal and lateral chest X-ray images, and text columns combining findings and impressions.
To retrieve the text description, we combined the *Impression*, *Findings*, and *Indication* columns and used them as a dataset for M_{L}.
We used both frontal and lateral chest X-ray images as a dataset for the CNN, M_{V}.

Models

CNN

The vision model M_V was built using transfer learning with a pre-trained SENet-154 architecture. It has been shown that ResNet-based architectures are effective for solving medical imaging tasks such as chest X-ray classification. Rajpurkar et al. in their study used ResNet-50, while we utilized a more advanced model, SENet-154, which incorporates a squeeze-and-excitation block and is expected to provide improved performance over ResNet-50 for this task. We utilized this model for the binary classification task for predicting whether a person has any disease based on chest X-ray images.

The following parameters for training were used:

- Optimizer: Adam
- Learning rate: 0.001
- Betas: (0.9, 0.999)
- Batch size: 16

Language model

We utilized the Bio_ClinicalBERT model as the language model.

This pre-trained model is based on the BERT architecture and has been fine-tuned specifically for clinical text, making it well-suited for analyzing medical data.

We post-trained the model for 5 epochs using AdamW with a learning rate of $2 \cdot 10^{-5}$, which is commonly used for fine-tuning transformer models. The CrossEntropyLoss function was applied for the loss calculation. This model solved the same binary classification task as M_V but with the text labels as inputs.

Multimodal model

For the multimodal model M_{VL}, we employed two fusion techniques: early fusion, where textual and image embeddings are concatenated, and late fusion. Accordingly, we implemented two models for classification: VisionBERT_EarlyFusion and VisionBERT_LateFusion. The multimodal model aimed to predict whether a person has a disease or is healthy based on chest X- ray images accompanied by text labels.

VisionBERT_EarlyFusion:

This model combines lateral and frontal images using the SE-ResNet architecture for feature extraction, excluding the final fully connected layer to obtain spatial features. These image features are concatenated and fused with the textual features from BERT's [CLS] token representation. The fused features are passed through a linear layer for classification.

Training Parameters:

- Optimizer: Adam
- Learning Rate: $1 \cdot 10^{-4}$
- Epochs: 5

VisionBERT_LateFusion

Similar to the first model, this architecture extracts features from both the image (via SE-ResNet) and text (via BERT). However, late fusion is applied: separate classifiers for each modality produce independent predictions, which are concatenated and passed to a final classifier for decision-making. This enables the model to learn the contributions of each modality before fusion.

Training Parameters:

- Loss Function: Binary Cross-Entropy Loss (BCEWithLogitsLoss)
- Optimizer: Adam
- Learning Rate: $1 \cdot 10^{-5}$
- Epochs: 5

4. Key Findings

Attack Type (%)	Vision Model (%)	Language Model (%)	Multimodal Model (%)
No Attack	**68**	**98.44**	**95.98**
$A_I^{FGSM_{01}}$ ($\epsilon = 0.01$)	63.06	-	96.39
$A_I^{FGSM_{05}}$ ($\epsilon = 0.05$)	63.06	-	97.22
$A_I^{FGSM_1}$ ($\epsilon = 0.1$)	63.06	-	95.56
$A_I^{FGSM_2}$ ($\epsilon = 0.2$)	63.06	-	63.61
A_T^{SYN}	-	85.44	92.93
A_T^{DEL}	-	85.58	81.28

Table 1: Accuracy of models under different attack types

Attack Type Legend:

- $A_I^{FGSM_{01}}$ ($\epsilon = 0.01$): Fast Gradient Sign Method (FGSM) with $\epsilon = 0.01$.
- $A_I^{FGSM_{05}}$ ($\epsilon = 0.05$): Fast Gradient Sign Method (FGSM) with $\epsilon = 0.05$.
- $A_I^{FGSM_1}$ ($\epsilon = 0.1$): Fast Gradient Sign Method (FGSM) with $\epsilon = 0.1$.
- $A_I^{FGSM_2}$ ($\epsilon = 0.2$): Fast Gradient Sign Method (FGSM) with $\epsilon = 0.2$.
- A_T^{SYN}: Text attack based on synonym replacement.

- $A_{\mathrm{T}}^{\mathrm{DEL}}$: Text attack based on word deletion.

As shown in the experiments, both single-modality models and multimodal models are vulnerable to adversarial attacks, though with different intensities. While even gentle attacks with small parameters significantly degraded the performance of single-modality models, the multimodal model only experienced significant accuracy drop under exceptionally strong attacks.

Another point we want to mention concerns the multimodality domain. Although our vision model alone exhibited poor performance, VisionBERT benefited from the strong performance of the effective language model, which contributed to its overall success.

The multimodal model VisionBERT demonstrated exceptional performance and relative robustness against various types of attacks on different modalities. Although attacks reduced the model's accuracy, it still outperformed singlemodality models under similar conditions. So, not only multimodality can enhance the overall performance by combining the strengths of the individual models it integrates but also increases the overall robustness to adversarial scenarios.

5. Implications

In our study, we observed interesting behavior in multimodal models and examined their resilience under different adversarial scenarios.

Our experiments demonstrate that all models can be attacked by adversarial examples, but the multimodal model appears to be slightly more resilient to such perturbations. We attribute this behavior to the multimodal nature of the models.

We propose that further research is needed in both the domain of multimodality AI models and adversarial attacks on such models. Understanding how information flows across modalities is particularly intriguing. This insight could enhance our understanding of how deep learning models work, which makes this study particularly significant.

6. Conclusion

This research focused on studying the vulnerabilities of medical models, with a primary emphasis on different modalities. Our findings suggest that multimodal models have varying degrees of vulnerability depending on the modality and fusion technique being used.

Our research has the following limitations:

1. We applied a limited number of attacks to each modality, with most being white-box attacks. Further investigation into black-box attacks is needed.
2. More attention could be given to medicine-specific attacks, such as lesion-specific attacks and other domain-specific attacks.
7. The fusion technique used to combine modalities can significantly influence the results. However, we only implemented two fusion techniques (early fusion and late fusion), excluding other promising approaches, such as cross-attention and others.

Thus, we define the following directions for future research:

1. A broader range of attacks will be explored for each modality, including both black-box and white-box attacks. Some of these attacks will be specific to the healthcare domain, beyond general-purpose ones.
2. Future research will investigate how cross-attention mechanisms influence adversarial attacks. This is particularly important since attention mechanisms are integral to state-of-the-art models.
8. Additionally, we aim to propose defensive strategies to mitigate the types of attacks explored in this study.