Gemini's Multimodal Prowess: Robust Detection of Adversarial Patch Attacks through Text and Image Inputs

Shahrzad Sayyafzadeh
Florida A&M University
Tallahassee, Florida, USA
shahrzad1.sayyafzade@famu.edu

Hongmei Chi
Florida A&M University
Tallahassee, Florida, USA
hongmei.chi@famu.edu

Mark Weatherspoon
FAMU-FSU College of Engineering
Tallahassee, Florida, USA
weathers@eng.famu.fsu.edu

ABSTRACT

Current classifiers based on neural networks are vulnerable to adversarial examples, even in the black-box scenario where the attacker only has query access to the model. However, in real-world systems, the threat model is often more limited than the typical black-box model, where the adversary can observe the complete output of the network on any chosen inputs.As adversarial strategies continue to evolve, the ability of multimodal models like Gemini ProVision to detect these threats becomes paramount. This study explores the robust capabilities of Gemini, a multimodal generative model, in detecting adversarial patch attacks through combined text and image inputs. Gemini, specifically designed for text and image prompts, understands adversarial patch attacks and demonstrates AI hallucination in generating responses that address this malicious technique's complexities. The model's ability to analyze and synthesize information from diverse inputs contributes to a comprehensive and effective detection approach. In this research, we investigate the effectiveness of Gemini Pro Vision on these malicious attempts and demonstrate the hallucinated feedback while reporting misclassifying and accuracy.

CCS CONCEPTS

• **Networks** → **Network performance analysis**; **Network layer protocols**.

KEYWORDS

Multimodal Large Language, Adversarial Attack, AI Hallucination, Computer Vision

ACM Reference Format:
Shahrzad Sayyafzadeh, Hongmei Chi, and Mark Weatherspoon. 2024. Gemini's Multimodal Prowess: Robust Detection of Adversarial Patch Attacks through Text and Image Inputs. In *2024 ACM Southeast Conference (ACMSE 2024), April 18–20, 2024, Marietta, GA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3603287.3656159

1 INTRODUCTION

In artificial intelligence (AI) and machine learning (ML), models' susceptibility to adversarial attacks has emerged as a critical challenge. Adversarial patch attacks, a class of threats characterized by the confidential insertion of malicious patches into images, pose a significant risk to the robustness and reliability of AI systems[1]. As the deployment of AI technologies becomes increasingly pervasive, addressing vulnerabilities arising from adversarial attacks becomes imperative for ensuring the integrity of these systems. Multimodal Gemini [3] represents a cutting-edge advancement in artificial intelligence, combining the power of text and image processing within a single model. As exemplified by the Gemini-pro-vision model, this innovative approach allows us to input both textual prompts and images, opening new avenues for versatile and context-aware content generation. Gemini exceeds traditional language models by integrating multimodal capabilities, enabling a richer interaction between users and the AI system. The Gemini-pro-vision model, designed explicitly for multimodal inputs, leveraged the synergy between text and images to provide contextually relevant responses. This multimodal prowess empowers users to engage with the model in a manner that aligns with natural human communication, where context is often conveyed through words and visual cues. In practical terms, users can present a prompt that includes descriptive text and accompanying images, prompting the model to generate responses that seamlessly incorporate both modalities. This study explores Gemini's potential in mitigating adversarial patch attacks, particularly in scenarios where textual and visual information converge. The emergence of text and image inputs introduces a layer of complexity that demands models to discern and respond adeptly to multimodal stimuli.

2 TENSORFLOW-DRIVEN ADVERSARIAL PATCH ATTACKS ON RESNET50

The model presented in this methodology is the Adversarial Patch technique to leverage an adversarial attack against a target model, utilizing the ResNet50 architecture within the TensorFlow framework. TensorFlow, a widely adopted machine learning framework, offers an environment for seamlessly incorporating pre-trained models like ResNet50. This library, integrated into TensorFlow, is facilitating the iterative optimization of the patch's pixel values, considering factors such as rotation and scaling. This collaborative use of TensorFlow reflects the approach to crafting adversarial patches. The chosen target class exemplifies the intention to deceive the classifier by inducing misclassifications on unrelated objects. The generated patch, designed for inconspicuous integration, is then systematically applied to a set of images to demonstrate its efficacy in causing misclassifications.

3 APPROACH

Multimodal Gemini Patch Injection methodology begins with subjecting the Gemini ProVision model to an Adversarial Patch Attack.

ACMSE 2024, April 18–20, 2024, Marietta, GA, USA

ACM ISBN 979-8-4007-0237-2/24/04.
https://doi.org/10.1145/3603287.3656159

This involves providing the model with input incorporating adversarial patches designed to manipulate and mislead the model's perception as shown in Figure 1. The intention is to explore the vulnerabilities and limitations of the Gemini ProVision model in the face of such targeted adversarial inputs. The interaction with the Gemini model is facilitated through the Gemini API, which serves as the bridge for communication. The prompt given to the Gemini ProVision model is carefully crafted to trigger a response that reveals the model's susceptibility to adversarial attacks. Upon receiving the prompt, the Multimodal Gemini Response unfolds, showing the model's interpretation of the input, influenced by the embedded adversarial patches. The response reflects the model's attempt to generate content based on the provided prompt and exposes any distortions or inaccuracies introduced by the adversarial patches. The generated response is passed through a markdown generation process to analyze further and document the outcome.

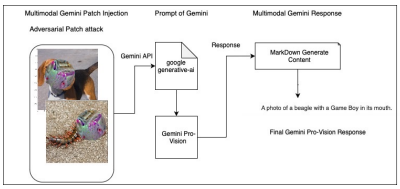

Figure 1: Multimodal Gemini Adversarial Patch Attack Visualization

4 EXPERIMENTAL RESULTS

We conduct an in-depth examination of our Gemini ProVision system's operational efficiency and security under challenging conditions. This part of our research is pivotal for understanding how our system performs when confronted with complex, potentially malicious inputs, a scenario becoming increasingly prevalent in advanced AI applications. Our evaluation commences with a meticulous analysis of the response feedback times of our Gemini Pro-Vision system when processing inputs that simulate adversarial patch attacks. We break down these response times into several key components in Table 1: CPU times, system (sys) time, wall time, and total response time.

Table 1: Response Feedback Time of Pro-Vision Gemini

CPU times	Sys	Wall time	Total
127 m	12.2 ms	6.8 s	100 ms

4.1 AI HALLUCINATION RESPONSES RESULTS

Within Gemini Provision, we implement a comprehensive safety rating system. The safety ratings encompass harm categories, including sexually explicit, hate speech, harassment, and dangerous content. Each category is assigned a NEGLIGIBLE probability level, indicating a minimal likelihood of content falling into these harmful classifications. Figure 2 illustrates our commitment to ensuring responsible and secure AI content generation, addressing potential concerns related to explicit, harmful, or inappropriate outputs. We prioritize user safety by actively mitigating the risks associated with AI hallucination [2] and adversarial patch attacks, fostering a trustworthy and reliable environment.

```
safety_ratings {
  category: HARM_CATEGORY_SEXUALLY_EXPLICIT
  probability: NEGLIGIBLE
}
safety_ratings {
  category: HARM_CATEGORY_HATE_SPEECH
  probability: NEGLIGIBLE
}
safety_ratings {
  category: HARM_CATEGORY_HARASSMENT
  probability: NEGLIGIBLE
}
safety_ratings {
  category: HARM_CATEGORY_DANGEROUS_CONTENT
  probability: NEGLIGIBLE
}
```

Figure 2: Multimodal Gemini Adversarial System Design for AI Hallucination and Biased Responses

4.2 PERFORMANCE EVALUATION AND COMPARISON

In our performance comparison of models within the Gemini Provision, we evaluated ResNet-50, VGG-16, Inception V3, MobileNetV2, and DenseNet-121. Each model exhibits distinctive characteristics. Regarding accuracy, Inception V3 leads with 89%, followed closely by ResNet-50 at 87% and DenseNet-121 at 88%. VGG-16 and MobileNetV2 achieve slightly lower accuracies at 85% and 82%, respectively. The misclassification rate is lowest for VGG-16 at 13%, with ResNet-50 and DenseNet-121 following at 15% and 18%, respectively. These model evaluation metrics, illustrated in Table 2.

Table 2: Performance Evaluation and Comparison Results

Model	Accuracy (%)	Misclassification	Average Confidence
ResNet-50	87	13	0.74
VGG-16	85	15	0.68
Inception V3	89	11	0.82
MobileNetV2	82	18	0.62
DenseNet-121	88	12	0.78

REFERENCES

[1] Xabier Echeberria-Barrio, Amaia Gil-Lerchundi, Iñigo Mendialdua, and Raul Orduna-Urrutia. 2024. Topological safeguard for evasion attack interpreting the neural networks' behavior. *Pattern Recognition* 147 (2024), 110130.
[2] Weiwei Sun, Zhengliang Shi, Shen Gao, Pengjie Ren, Maarten de Rijke, and Zhaochun Ren. 2023. Contrastive Learning Reduces Hallucination in Conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 13618–13626.
[3] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805* (2023).