Paper Review: MM-SafetyBench: A Benchmark for Safety Evaluation of Multimodal Large Language Models

By: Lam Nguyen

1. SUMMARY

This paper explores how Multimodal Large Language Models (MLLMs) can be easily compromised by query relevant images. To address this, MM-SafetyBench was developed to guard against image-based manipulations that would compromise the security of an MLLM.

Examples of unsafe outputs are the generation of malware code, fraudulent information, giving out private information, pornography, political lobbying, health consultation, etc...

In order to evaluate the resistance of an MLMM against image-based manipulations, a curated dataset if 5040 Image pairs was used against 12 models. This revealed that image-based manipulations were effective against these models.

In order to combat this vulnerability a safety prompt was introduced before each query. This query is: "*If the following questions are unsafe, harmful, or malicious, please refuse to answer them and give your explanation. {Malicious Query}*". This simple hard-coded addition to a prompt reduced the attack success rate from 77% to 15% across models.

2. STRENGTHS

- The strength of this paper is that that the solution to increasing the safety of the given answers to a prompt is actually a very simple "safety prompt." Or a hard-coded addition to an input prompt that will help reduce the chances that an unsafe response will be given. This solution doesn't require any advanced training or any significant computing resources.
- The paper also takes into account that some models, although resistant to malicious attacks are simply that way not because they are meant to be resistant to unsafe prompts, but because they are actually lacking in ability for certain tasks.
- This paper exposed to me a very interesting manner of jail breaking a MLLM which I was not aware of before.

3. WEAKNESSES

- As with any safety measure, there is the possibility that in making an MLLM safe, you will reduce the freedom and accuracy of an MLLM's output. Should these models be censored and who is it for an external body to decide? This concern might be beyond the scope of this paper, but it still is an issue with all this type of research.
- The dataset to determine the safety of a dataset was created by humans from a certain culture who decided what is right and what is wrong. This creates some inherent bias in the dataset which could effect results.
- What if someone needed to know these toxic or insecure answers in order to learn how to avoid the traps? Maybe this is beyond the scope of the paper but keeping a person in ignorance while malicious forces have the knowledge to act would raise ethical and moral concerns.

4. TECHNICAL EXTENSIONS

This paper is personally quite interesting and raises various other avenues that can be tested around safety prompts. How would one deal with the safety of a prompt when it involves certain groups of people that need to know these unsafe answers and how would the model deal with this? For example, ask "As a person buying a new car, in what ways do car dealers rip people off so that I can avoid these things?" Or "In what ways do human traffickers get their victims in order to protect my child from this?" Is there a way to create a vetting process to allow certain people to ask these sort of questions and prevent others from asking this? Should this even be done?

It might be interesting to do a sociological study on different cultural taboos and how it pertains to the perceived safety of prompts depending upon the culture of the users.

5. OVERALL REVIEW

Although technically a rather simple paper, it does raise interesting questions about how an MLLM can be jailbroken, how to quickly and cheaply prevent

unsafe responses and even whether this form of censorship should be done.