Paper Review: Safe-CLIP: Removing NSFW Concepts from Vision-and-Language Models

By: Lam Nguyen

1. SUMMARY

Vision Language models such as CLIP are trained on the Web. The web is full of NSFW content. This NSFW content can hinder the application of LLMs and VLMs in sensitive contexts such as in hospitals, law settings, educating minors, etc….

To address this issue, Safe-Clip was developed. This technique seeks to unlearn the links between unsafe concepts and embedding regions. This can be done by fine-tuning a CLIP model on synthetic data which was trained to convert between safe and unsafe sentences and a text to image generator.

Once the CLIP model is trained, it can ignore NSFW content.

Related work to Safe-Clip revolves around removing concepts from vision and language models. Detecting NSFW content, and finetuning LLMs with little data.

How does Safe-Clip work? The first approach that was proposed was to clean the data at a large scale. However, this is somewhat inefficient. The more efficient solution is to make the textual encoder and the visual encoder of CLIP safer.

To make CLIP safer, a custom dataset was used that contains a curated combination of safe and unsafe images and sentences.

The NSFW textual generator was trained on this dataset. The LLM used was Llama- 2-Chat. It was used to generate unsafe sentences starting from safe ones.

With the unsafe text generator created, the next thing to create was the unsafe image generator. This was a diffusion model trained on unsafe data.

In order to make CLIP safe, training was done to teach CLIP to ignore both the Unsafe texts and the Unsafe images.

2. STRENGTHS

- The strength of this approach is also it's weakness. Since the model seems to have been fine-tuned on a deep level to avoid unsafe data… It will be very difficult to jailbreak the model to do anything unsafe since the links themselves are broken.

3. WEAKNESSES

- The weakness of this approach to creating safe content is that it requires a lot of computing resources and time in order to make the model safe.
- What happens if new biases and things that need to be censored arise? This approach might not be enough.
- Also, there are the moral implications of censorship. What will stop malicious actors from using this technology to censor others?

4. TECHNICAL EXTENSIONS

- Expand the capabilities of this approach to allow for adaptation to changing needs for censoring the model. There is a technique called SAFREE that is relatively light weight and adaptable that can insure that the model is able to be adapted to new forms of bias and unsafe content.
- Is there a way to store the broken links in a separate dataset so that the model can be made unsafe again? Or what methods can be used to remember the unsafe tendencies?

5. OVERALL REVIEW

Safe-CLIP is a method that essentially involves finetuning the CLIP model on how to avoid NSFW data by inputting NSFW data into the model and performing almost a reverse loss function.

It requires a lot of computing resources but this makes it less likely to be jailbroken since the actual CLIP model's weights have been altered.