Paper Review: Forget-Me-Not: Learning to Forget in Text-to-Image Diffusion Models

By: Lam Nguyen

1. SUMMARY

Text-to-Image models have advanced very quickly. But as they've advanced, there has also been the problem of these models violating things like copyright, safety, privacy, etc…

There are many methods to censor these models. They mainly revolve around fine-tuning and pretraining a model; essentially modifying the knowledge base of the model. Modifying the knowledge base can affect the accuracy and reliability of a model negatively. Also, these pre-existing methods require hundreds of optimization steps.

The Forget-Me-Not method in contrast, is an efficient solution designed to remove identities, objects and styles from a model in as little as 30 seconds without significantly impairing a model's ability to generate other content. It does this by steering a model away from unwanted content.

The benefit of this is that NSFW content can be removed and also there will be an increase in diversity and inclusion.

Next, a discussion will be made on how the Forget-Me-Not method works. The method is based on the concept of concept forgetting. Concept forgetting is breaking the relationship between the concept and the visual representation. The Forget-Me-Not method gets rid of the correlation between image representation and the concept.

To do this, two new loss functions were created: The Attention Re-steering Loss and the Visual De-Noising loss. Also for scenarios where the prompt associated with a concept is not known, a technique called Concept Inversion is used to extract text embeddings of a concept directly from images.

2. STRENGTHS

- The strength of this method is that it is fast and cheap. It doesn't excessively alter the underlying structure of a model so that the model can stay as accurate as possible
- The forgetting concept can be applied to many different things besides NSFW concepts, so in this way, the model can evolve without much retraining.
- There is the possibility that in making a model more diverse and inclusive… The accuracy of a model can be increased as there is less bias.
- Fine tuning cross attention versus fine tuning the entire model is more robust and won't break the model as easily.

3. WEAKNESSES

- There is still the problem with all censorship of a model. In trying to break the correlation between concepts, there can be unexpected consequences to the accuracy and hallucinations can occur. These models are very complex an any manipulation can have unexpected and negative consequences to mode accuracy.
- There is the possibility that censorship can be abused. Giving this tool to society can allow large institutions to have the ability to manipulate the minds of other people by reprogramming concepts. This might not be a scientific concern… But it is more ethical and moral. In trying to prevent NSFW content, you could just remove free speech and awareness of the NSFW parts of life.
- Concept inversion is not consistent. Its effectiveness can change depending on the concept.

4. TECHNICAL EXTENSIONS

A technical extension could be to see if the concepts that are forgotten can be remembered again. Is there a way to store the forgotten concepts and remember again?

5. OVERALL REVIEW

The forget-me-not method is a cheap way of breaking correlations between a concepts and visual outputs. It uses two different loss functions and concept inversion. It can help to promote privacy, accuracy

and diversity. However, there is always the issue of what is NSFW and should large institutions be allowed to do such a thing? Also what is right and what is wrong?