Literature Review on Online Detection of AI-Generated Images (Epstein et al.) [1]

https://arxiv.org/abs/2310.15150

Literature review by: Jeffrey and Happy

Overview

- Existing AI-generated image detection methods are created based on the assumption that if a model detecting AI-generated images is trained on images created by *one* generator, that model is generalizable to images created by *other* generators
- However, this assumption cannot always be made due to the fluid environment of advancements in AI and the constant rolling out of new generators.
- Thus, in this paper, the authors look to test if a detection model trained on the *aggregation* of a series of existing generators can reliably generalize over a wider breadth of images

Datasets

- 14 datasets of images, each consisting of images generated by a different generator (i.e. DALLE2, GLIDE, etc.)
- LAION-400M: a dataset consisting of 400 million REAL text-image pairs, widely used by AI researchers

Model

- Given the speed of the advancement of AI and the seemingly unending rolling out of new generative AI applications, the assumption that a detection model trained on a single, potentially outdated generator can generalize to new and improved generators may not always be valid
- Using the 14 datasets derived from the 14 AI-image generators, the authors progressively trained the model by adding new information (i.e. images from newer generators) in the order in which the generators were released to the public
- CNN binary classifier (AI-generated or not AI-generated)
- The model also explored techniques for detecting the edited images. In other words, to detect the images edited by AI using augmentation techniques such as CutMix. This is a training technique that mixes parts of real images with AI-generated so that the model can recognize edited parts with limited training.
- They made three sets of pictures for filling in the missing parts using two State Diffusion and Adobe Firefly.

Results and Limitations

- This approach of training a detection model based on aggregating generators in the order in which they were released instead of relying on a single, potentially outdated generator

in fact works well PROVIDED THAT the architecture (i.e. Diffusion U-Net, Diffusion Decoder, etc.) of the generators remains constant
- Suggests that the best way of building a model like this would be to aggregate images from existing generators that are built on the same architecture
- Problems with obtaining data from other prominent generators, such as unavailable source code or an unavailable API, may have influenced the results in that a representative sample of images from popular, publicly available generators was not attainable

What we can use/implement in our project

- Can try to replicate the idea of training a detection model on images from a series of generators rather than the more usual approach of using images from just a single generator
- Can use a subset of the LAION-400M dataset's text-image pairs to help train the model we eventually want to build
- Can also use a subset of the 14 image datasets with the added constraint that the datasets we use all correspond to the same generator architecture, as the results of this paper suggest that a detection model trained by aggregating generators of the same architecture has a better performance than a detection model trained on generators of differing architectures.

References

[1] D. C. Epstein, I. Jain, O. Wang, and R. Zhang, 'Online Detection of AI-Generated Images', in *ICCV DeepFake Analysis and Detection Workshop*, 2023.