Internship Report

Lucas SALAND

July 17, 2024

Contents

Introduction

This M1 internship was carried out at CRIStAL in the SIGMA team under the supervision of Patrick Bas. The internship lasted for three months during which we worked on a challenge from the AID on the detection of images generated by AI.

Work environment : CRIStAL and SIGMA

CRIStAL is a laboratory which research focus on computer science, signal and automatic control. It is under the supervision of the University of Lille, CNRS and Centrale Lille. The laboratory is divided in 34 research teams grouped in 9 Thematic Groups. SIGMA team is part of DatInG : Data Intelligence Group. SIGMA is a team of 15 permanent staff which focuses on machine learning, statistics and signal processing. Some of the research topics are Monte-Carlo methods, signal processing with tensorial approaches and information security.

The challenge from AID

The Agence de l'innovation et de défense (AID) launched a challenge on détecting modified or generated images. This challenge aimed at detecting three types of images :

- fully AI-generated images;
- images partially modified by AI;
- images partially modified with more usual image processing tools such as photoshop.

This challenge was divided in two tasks : A and B. Task A focused on images fully generated by AI. AID would provide 10000 images. The goal was to identify which images were real and which one were generated. On top of this, we could provide which generator was used to generate images. Task A could be treated as binary classification problem with the two classes being real and generated images. It could also be treated as a multi-class classification problem where the classes would be the real images and all the generator used. The main difficulty was that the generators used were kept secret until the last day of the challenge.
Task B focused on the detection of partially modified images. The objectives were :

- detecting real images and modified images;
- identifying the tool used for modification;
- localisation of modification on images.

I worked on the challenge in a team with 3 other interns and 5 permanent staff from SIGMA. During the internship, I worked on task A.

AI image generation

We should now go over an overview of image generation with AI. In recent years, the quality of images generated with AI models skyrocketed. These models appear as promising new tools for art generation and data augmentation for machine learning. In 2014, Generative Adversarial Networks were introduced by Ian J. Goodfellow and his colleagues in [2]. A GAN is composed of two main components : a generator and a discriminator. The generator take random vector as input and tries to generate images that are indistinguishable from real images. It tries to fool the discriminator. On the other hand, the discriminator's goal is to differentiate between real data from the training set and fake data produced by the generator.

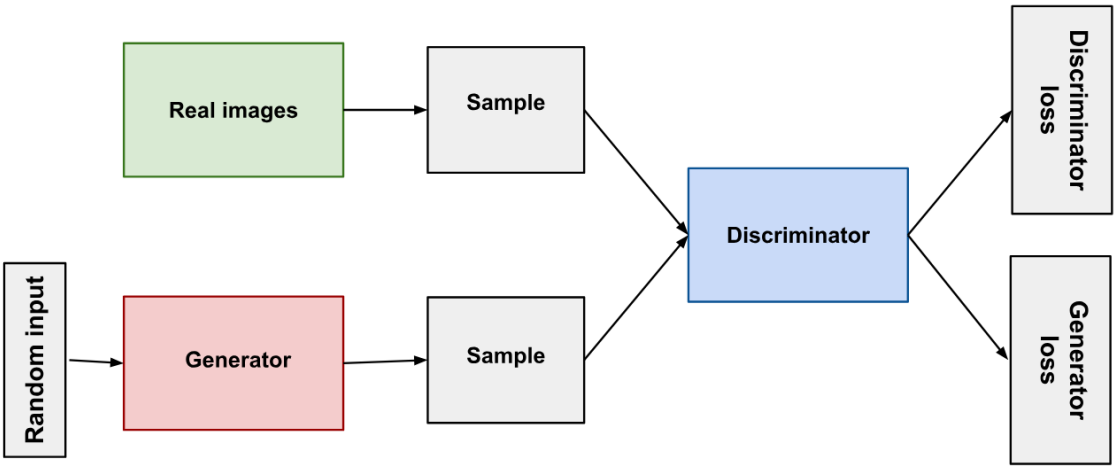

Figure 1: GAN architecture

Diffusion models use a different approach to generate images.

More on diffusion model

1 AI-generated images detection

As improvements happen at an impressive rate in the field of generative models, concerns about security issues rose as well. These tools could be used to manipulate information and combined with social media they would be weapon of massive disinformation. As generative model keep improving, we need to develop new tools to detect generated images. But what should we look for in generated images in order to differentiate them

from real images? To develop an efficient detector, we decided to explore a semantic approach. Let's first give a definition of semantic. For an image, the semantic refers to the content of the image. The objects and things that are perceived by the humans. Images generated by AI tend to have a poor semantic content in comparison to real images. But how should we process images to extract their semantic content? That's where image encoders come into play. More specifically CLIP's image encoder.

1.1 CLIP

[1]

1.2 Impact of JPEG compression

1.3 Adding diversity to the data

1.4 Bigger datasets and neural network

1.5 Pair training

1.6 Filling the holes with fine tuning

2 Additional explorations

2.1 Color features

2.2 DINO as an alternative to CLIP

[4]

2.3 Tip-Adapter

[3] [5]

2.4 Understanding CLIP features

AID real img are bad crops with poor semantic content so why does CLIP detector performed well ? Adversarial attack.

Conclusion

References

[1] Davide Cozzolino et al. *Raising the Bar of AI-generated Image Detection with CLIP*. Apr. 29, 2024. arXiv: 2312.00195 [cs]. URL: http://arxiv.org/abs/2312.00195. Pre-published.

[2] Ian J. Goodfellow et al. *Generative Adversarial Networks*. June 10, 2014. arXiv: 1406.2661 [cs, stat]. URL: http://arxiv.org/abs/1406.2661. Pre-published.

[3] Sohail Ahmed Khan and Duc-Tien Dang-Nguyen. *CLIPping the Deception: Adapting Vision-Language Models for Universal Deepfake Detection*. Feb. 20, 2024. arXiv: 2402.12927 [cs]. URL: http://arxiv.org/abs/2402.12927. Pre-published.

[4] Maxime Oquab et al. *DINOv2: Learning Robust Visual Features without Supervision*. Feb. 2, 2024. arXiv: 2304.07193 [cs]. URL: http://arxiv.org/abs/2304.07193. Pre-published.

[5] Renrui Zhang et al. *Tip-Adapter: Training-free Adaption of CLIP for Few-shot Classification*. July 19, 2022. arXiv: 2207.09519 [cs]. URL: http://arxiv.org/abs/2207.09519. Pre-published.