CS7NS5: Security & Privacy Assignment 1

Cormac Sharkey
20333458

March 2025

MAI Dissertation: Security & Privacy Considerations

The project I have undertaken for my dissertation is titled "Day-to-Night Colour Transfer Using Stable Diffusion". The purpose of this project is to further develop Deep Learning models that will contribute to the ongoing research topic of day-to-night image translation. More broadly, the research falls into the field of unpaired image translation, a field that looks to use Deep Learning techniques to create synthetic images between unnatural domain pairs. A classic example of an unnatural domain pair is horse and zebra, in which no two photos of a horse and a zebra exactly align with each other. The day-to-night aspect is further inspired by cinema post-production, as many nighttime shots in film and television are captured during the day for convenience and edited back into a nighttime setting. This is a costly technical process, thus developing an effective model for automation would be ideal.

Stable Diffusion plays a vital role in the project, as it is a subset Deep Learning model that excels at synthetic image generation. The most convincing AI generated imagery around today, regardless of content or style, is created with Stable Diffusion or other Diffusion-based models. Up until now, no researcher has fully explored using Stable Diffusion in unpaired image translation, and there is potential for it to make a significant impact. However, despite the project's usefulness and its potential for practical application, there are things to consider. There are numerous security and privacy considerations that may pose issues to the ongoing research and final results, but also to any other researchers looking to follow a similar path on this topic.

Google Colab

The first security and privacy consideration of the project revolves around the use of an online tool for model training and gathering results. Google Colab [1] is an online service that allows any Google account holder to connect to a virtual machine hosting a Jupiter Notebook Python environment, allowing them to avail of computing resources and GPUs. The service is provided by Google and has become a popular tool for both full-time researchers and hobbyists to train and experiment with AI models. The service offers a free tier, with limited GPU access and a cap on runtimes, which is most useful for small, simple models or for one-time uses. However, there are various paid tiers, allowing access to better GPUs, longer runtimes and even a terminal interface into the virtual machine. As well as that, paid tiers work through a compute credits system, which require a set number of credits be consumed every hour of connection. More credits can be purchased at any time. With all of this functionality, Google Colab becomes the perfect tool for anyone who cannot afford a high-end GPU or computer. As part of this project, Google Colab was used extensively for the above discussed reasons.

Despite its usefulness for Deep Learning model training, it suffers from a significant security flaw that leaves many users open to data breaches and cyberattacks. As part of Colab, Google have implemented functionality to mount a user's Google Drive on the virtual computer, meaning any scripts or data you have stored there can be executed or used on the virtual machine. It is a clever addition to the service, as it is the easiest way to run pre-made code on Colab without copy-pasting into notebooks or cloning a Git repo. However, to mount the drive, users must grant the Colab notebook extreme privileges. These include create, delete, edit, and view access on all files in their Google Drive, and view access on all contact information stored in their Google account. Once these permissions have been granted, the drive is mounted. It is a necessary step for mounting the drive, and most users click through it to continue working, but it presents enormous potential for malicious attack and data theft.

Colab notebooks can be made publicly available, for both viewing and code execution. This means that any bad actor can create a notebook with malicious code, which can be easily accessed and used by any Google account holder. A data breach then becomes as simple as getting any unsuspecting user to run the notebook's

code when signed in, and the attacker has access to all data in their Google Drive. Even worse, tricking someone into using the code blindly is not hard, as it is easy to hide small amounts of malicious code in a notebook filled with a few hundred lines. Distribution of the notebook can be through scam emails, online forums, or message board posts. They can even be through official webpages, if the notebook has valid and functional code, with the malicious code hidden deep within it.

To take an example, ngrok [2] is a tool used for making applications or APIs available online, such as setting up online endpoints. Normally, the tool is not malicious and has genuine practical applications. However, it has been shown [3] that, by using ngrok in a Colab environment, a user can set up a Python server within the virtual machine and link its access to a URL. Through that URL, the user can access the directories of the computer, which could have a user's mounted Google Drive, if that code was added before the ngrok executions. Using ngrok, a bad actor can gain access to a user's Google Drive in a few lines of code, all because Google Colab's drive mounting puts them at risk.

Fortunately, any code used in this project that was external came from research papers and their respective Github repositories, and I was careful to read the code to understand it, but also to make sure it was fit for purpose. The code itself was also never designed to be used in Google Colab, drastically reducing the likelihood of it containing malicious code meant to abuse the Colab environment. But, regardless, it was never checked for malicious code that could cause a data breach. It is a significant security and privacy issue for the project, as well as to the other researchers and casual users of the service.

Membership Inference Attacks

The second security and privacy consideration of the project focuses on the Deep Learning models themselves. When training an AI model, data is the most important part, as every model becomes incredibly effective at its purpose, by training for prolonged periods of time on substantial amounts of relevant data. In some cases, this may be sensitive data, such as medical imagery or addresses. For this project, this data consists of cityscape images during both nighttime and daytime. As with most cities, there are people, cars, signs, and more interspersed within the images, which may be considered personal data. This is because not all information in these images is redacted or blurred out to prevent models from learning and copying this censoring as a behavioural trait. As a result, the models may be considered as being trained on sensitive data, but other researchers in the field of Deep Learning and AI use much more sensitive data, like medical images.

Despite containing sensitive information, any user of the model post-training will not be able to access this data due to the nature of the models and their training. The user will only be able to observe the model's behaviour, which, when combined with the fact that researchers do not leak any sensitive information their model is trained on, makes them perfectly safe for anyone to use. Unfortunately, recent work has rendered this claim false, as that safety has been shattered with the threat of Membership Inference Attacks, or MIA [4]. These attacks, performed usually on tabular data, involve a bad actor reverse engineering the training data based on the model itself. This is a result of the models performing better or showing unique behaviour when exposed to data it has been trained on, because it has seen it before, numerous times. In MIAs, a bad actor can make educated guesses on some general data that was used, which, when applied to the model and its output observed, allows them to narrow their guesses until they are confident, and they have the training data.

Research has shown that even Generative Adversarial Networks, or GANs, are susceptible this type of attack [5]. This type of network is a core part of the project, potentially putting the models at risk. While the data used in the project is not incredibly sensitive, the underlying model architecture can be exploited this way in other applications. The models in the project are designed with unpaired image translation as their application, not only day-to-night. This means that another researcher with access to the model architecture may use it with data of other domains, such as medical imaging. For example, two unpaired domains may be satellite imagery of properties, and map-style images of properties. The application here would be to automate the process of generating map layouts for housing estates and land based on satellite imagery, useful for Google Maps and other technologies. However, non-obfuscated satellite images are sensitive data. If bad actors could reverse engineer the data out of this model, they would have private property data for potentially thousands of houses and estate, allowing them to plan robberies or conduct other malicious actions.

Speaking on this security risk beyond the scope of the project, this is a real threat to many AI models currently circulating around the Internet, which are freely available for users to access and use, or to download and run locally. The chances of a successful MIA hinge on two key factors: the number of susceptible models being created every day and circulating the Internet, and the availability of heavy-duty computing resources. Unfortunately, based on the widespread use of AI today and the accelerated growth of computing hardware companies, MIAs are likely to increase in both occurrence and success. Even worse, there is no clear approach

that can be taken for this project to avoid MIAs, as the architecture of the model is unchangeable in a way that can prevent them.

DeepFakes

The third and final security and privacy consideration of the project involves the creation of DeepFake images. These are fake images generated by high-performance gen AI models that look incredibly convincing and passable as real photos. Their creation and distribution have seen a rise in recent times, with increasingly explicit DeepFakes appearing across the Internet. As a result, ordinary people with photos of themselves online are becoming victims of DeepFake creation, as the models used primarily edit existing images to present a different context. This allows for explicit and/or false image creation without the photo owner's consent and has become a major issue for celebrities with many people stealing their likeness illegitimately.

The project is centred on day-to-night images, but the underlying architecture, as mentioned previously, focuses on image translation of unpaired image domains. A bad actor with access to this architecture and training data may be able to create a model for generating DeepFakes from existing photos. If the model were trained on the right type of data, it would be possible to create an image generator that took any photo of a person and turned it into explicit imagery, or changed the photo's context to whatever they would want it to be. For example, making a politician appear as though they have ties to a terrorist organisation or are corrupt.

This would all be without the photo owner's consent, as millions of images can be easily scraped from the Internet. DeepFakes present a huge threat to an individual's privacy, as, despite the fact the owner may have willingly posted images of themselves online, the resultant application of these images is malicious and harmful. Unfortunately, the only confirmed way to mitigate this issue would be on the photo owner's end, as they would need to avoid posting more photos of themselves or other people. As for the generative models used to synthesize DeepFakes, there aren't really any ways to mitigate the creation of these images, as many generative AI architectures are available online, some pre-trained and with no safeguarding to disable malicious functionality. Unfortunately, the models developed in this project fall under this category of generative AI.

References

[1] Ekaba Bisong. Google Colaboratory. In *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*, pages 59–64. Apress, Berkeley, CA, 2019. ISBN 978-1-4842-4470-8. doi: 10.1007/978-1-4842-4470-8_7. URL https://doi.org/10.1007/978-1-4842-4470-8_7.

[2] ngrok. ngrok | API Gateway, Kubernetes Networking + Secure Tunnels. URL https://ngrok.com/.

[3] 4n7m4n. Careful Who You Colab With:, July 2022. URL https://antman1p-30185.medium.com/careful-who-you-colab-with-fa8001f933e7.

[4] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017. URL https://ieeexplore.ieee.org/abstract/document/7958568/.

[5] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. LOGAN: Membership Inference Attacks Against Generative Models, August 2018. URL http://arxiv.org/abs/1705.07663. arXiv:1705.07663 [cs].