WASP Software Engineering Course Module 2024

Asma Raeisi

August 27, 2024

1 Introduction

I am working on the development of an image synthesis framework to generate realistic synthetic images. We focus on generating such kinds of synthetic data that resemble real-world data in order to train machine learning models for object detection, classification, and instance segmentation tasks. So our aim is to enhance the accuracy and performance of deep learning models in a wide scope of real-world applications. In our project, we will use augmented reality (AR) and neural rendering for the generation of highly realistic images. AR techniques will be used to place the images on top of the real data and merge the virtual with physical environments. Neural rendering is also a recent and advanced technique for generating images through 3D models. So, the use of neural rendering and AR techniques makes it possible to generate this novel type of data as training data for the deep learning models, while this type of data did not exist before. We can then train our models on either synthetic data alone or integrated with real data, thereby increasing the size of the training set. Finally, we will evaluate the performance of models trained on synthetic or integrated data to assess the possibility of performance improvement.

2 Concepts from Robert's Lectures

2.1 Fraction of ML code in Systems

When we take a closer look at any real-world ML systems, we realize that the ML code is just a small fraction of each complex system while the majority of the system is composed of other infrastructure components such as data pipelines, feature extraction processes, monitoring, and process management mechanisms. In our work, we need strong data pre-processing, augmentation, and validation procedures to create a robust synthetic image generation model and ensure the quality and diversity of the generated images. These images must be monitored and tested continuously to ensure that they meet the realistic standards required for training models. Thus, we have to understand the complexity of these surrounding components in ML applications.

2.2 Behavioral Software Engineering

Behavioral software engineering (BSE) focuses on understanding humans' behaviors to design software in a way that meets their needs effectively. When it comes to my research topic on synthetic image generation, by applying the principles of psychology, BSE might help make images look more real; synthetic images that actually look like images that humans expect to see. For instance, collecting and analyzing the feedback from users will further help improve our model's performance based on real-world use cases. The training set becomes as diverse as real-world images, making our model generalize in dealing with real-world scenarios.

2.3 Cognitive Bias

Cognitive bias is the systematic pattern of deviation from norm or rationality in judgment, where inferences about other people and situations may be drawn illogically. This may impact my research in several ways. For example, if we believe our image synthesis model is already perfect, then we may not test it thoroughly for changes. This could make the model have some flaws that we don't see, because we are just looking for results to confirm our belief. Then, we might ignore suggestions or new techniques that improve a model because they do not fit with our existing beliefs. Accordingly, it might not allow the model to reach its full potential.

3 Concepts from the Guest Lectures

3.1 AI Role in Four Key Steps to Release a Product

As it was mentioned in Per's presentation, we have four necessary steps to create a product that can be released to the public: define, design, develop, and test. These steps can be applied to most fields of research when addressing a problem through the development of an algorithm is necessary. In that case, we need to test and verify the algorithm and make sure it works. Utilizing AI to address problems in each of these steps is one of the biggest advantages of modern technology that can make our lives easier. In our research, we can also apply it at the design, development, and testing stages in order to expedite the process of model development and to guarantee its robustness in different scenarios. Using such advancements is very important, and we will fall behind if we don't use these technologies.

3.2 Uncertainties in AI/ML algorithms

One important takeaway from Dhasarathy's lecture is also the use of LLMs to validate models that have been created. Using an LLM approach can indeed help us explore some intricate scenarios that are difficult for humans to check due to their complexity. However, I think in some cases a human needs to be in the loop to ensure that the LLMs are performing accurately on their own. Based on the discussion we had, we all know that the integration of AI/ML will create uncertainties. Suppose we have an autonomous car detection model and train it on our synthetic images. Since AI algorithms have uncertainties, synthetic images may not accurately reflect the full diversity of real-world conditions such as lighting variations, weather effects, and unexpected obstacles. Although test cases run by LLMs might show a good performance, this reality gap can restrict the adaption of the model to real applications. So, the model might not work reliably in all kinds of environments and it will affect the safety of automated driving systems. Accordingly, we need to recognize these potential weaknesses and work to fix them.

4 Paper [1]

4.1 Core Idea

I selected this paper titled "Privacy and Copyright Protection in Generative AI: A Lifecycle Perspective" [1]. The core idea of this paper is about the growing concerns of copyright and privacy in the domain of Generative AI which stems from the reliance of models on large datasets. When we train a model on a large dataset it might cause some moral and legal problems. They discuss one way to address this issue by using a lifecycle-centric strategy. This means taking copyright and privacy protection into account at every stage of the data life-cycle, from training and collecting to distribution and deployment. This step is

crucial in developing AI systems. Addressing privacy and copyright from a life-cycle perspective ensures that AI systems can handle data responsibly. Thus, public trust will be gained, and legal penalties will be avoided.

4.2 How the Paper Relates to My Research

As I mentioned earlier, I am working on developing a synthetic data generation framework. In my research, privacy and copyright protection come into play. We have to be careful about training datasets that contain the faces of individuals. We must ensure that the generated images are not of copyrighted content. So, we can use licensed training data and incorporate mechanisms to avoid replicating copyrighted content.

4.3 Project Fit

If we have a large AI-intensive software project whose goal is to train models on synthetic data, then my research would fit into that project. By applying the approach outlined in the paper [1] and using a lifecycle-centric strategy when working with data, we can make any AI-intensive model more reliable. So, we can use a lifecycle-centric strategy and consider privacy and copyright at every stage of synthetic data generation process. Therefore, we can ensure that data is protected at every stage, from creation to deployment. So, each AI-intensive model trained on the synthetic data will not violate privacy or copyright laws.

4.4 Project Adaptations

Three techniques were proposed in this paper [1] to protect privacy and copyright:

- 1. Consent tagging which is one method to ensure privacy and facilitates tracking and verification of authorship.
- 2. AI bill of materials which maintains transparency and records the origin and ownership of the data
- 3. Machine forgetting which helps the model to effectively enable unlearning.

So, using consent tagging and AI bill of materials in our data generation process might help to ensure all synthetic data is traceable and compliant with data ownership and privacy laws. The next approach we can use in our research to protect copyright and privacy can be machine forgetting. Using this in our research might allow us to remove specific data points that violate copyright and privacy from our synthetic datasets. So, we might be able to correct or delete some parts of the dataset that might cause privacy concerns.

5 Paper [2]

5.1 Core Idea

In this section, I selected a paper titled "What About the Data? A Mapping Study on Data Engineering for AI Systems" [2] which discusses the critical role of data engineering in the development of AI systems. They discuss some of the existing literature on data engineering for AI systems from an AI engineering perspective. Moreover, they show how important it is to collect, prepare, and manage data properly to effectively build and deploy AI models. In addition, it highlights the concept of data-centric AI, which

means focusing on carefully preparing and managing data before using it to build AI systems. This includes activities such as data ingestion, transformation, serving, and storage. This work is important because it addresses a critical gap in the AI and data science fields by shifting the focus from predominantly modeling and algorithmic discussions to the essential role of data engineering.

5.2 How the Paper Relates to My Research

We focus on synthetic data generation using AR and neural rendering for AI applications. This paper provides a broader context for how the data engineering processes might support such synthetic data integration into AI workflows. Data engineering will lead us to have a consistent data generation pipeline, which is very important for ensuring the quality and reliability of the synthetic data.

5.3 Project Fit

In this paper [2], the authors discuss several AI data engineering architectures that facilitate the management of data, leading to the production of high-quality and consistent information to support downstream use cases. Moreover, the paper shows the importance of a comprehensive data ecosystem and robust data validation practices which lead to high standards of data quality and integrity. Any AI-based project can implement these practices and ensure the data used during training and inference is reliable and accurate.

We can use the AI-based models for downstream tasks and train them on our generated images. Since the models have robust data pipelines and architectures, we may conclude that any suboptimal results can be attributed to synthetic data issues, such as lack of generalization or diversity.

5.4 Project Adaptations

Since AI models will not work properly without high-quality data, we might pay more attention to the practices of data engineering to ensure an efficient and reliable data flow. For example, we can meticulously collect diverse image datasets. Then, we have to preprocess the data by normalizing image data and applying data augmentation techniques in order to increase data variability and volume. Data validation can be the next step in the process, which checks the quality of generated images. Taking these steps into account might allow us to generate high-quality data that might enhance the performance of any AI models we train.

References

[1] Dawen Zhang, Boming Xia, Yue Liu, Xiwei Xu, Thong Hoang, Zhenchang Xing, Mark Staples, Qinghua Lu, and Liming Zhu. Privacy and copyright protection in generative ai: A lifecycle perspective. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI*, pages 92–97, 2024.

[2] Petra Heck. What about the data? a mapping study on data engineering for ai systems. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI*, pages 43–52, 2024.