ROBOT PERCEPTION AND VISION

FINAL PROJECT

Muslim Alaran 5/4/23 ROBT 502

ABSTRACT

In recent times, we have witnessed the huge success of deep learning and artificial intelligence, most notable and popular of them all being GPT-3 (ChatGPT) – a large language model whose success is highly attributed to the transformer model. After the success of these deep learning models, the new paradigm is shifting towards achieving general artificial intelligence and the application of these models, libraries, and tools to increase efficiency and effectiveness in our daily lives. Therefore, in this work, I focus on creating a conceptual plan on how two of the AI tools, Nvidia Picasso and Omniverse, courtesy of Nvidia can be used to replace the traditional process/method of creating datasets. I focus on how the two tools can be used in generating synthetic visual datasets, thereby leading to an increased efficiency in model development and elimination of complexities associated with the manual creation and collection of visual datasets.

INTRODUCTION

Studies in Artificial Intelligence (AI) has increased significantly in various fields as the overall quantity of AI publications have doubled since 2010. Of these studies in AI, the main subjects of machine learning, computer vision, and pattern recognition have remained prevalent. However, after 2014, the industry has taken the lead in the creation and release of significant machine learning models with different applications[1]. The most commonly known application being the intelligent chatbot, ChatGPT which is powered by the large language model, *GPT-3* [2]. The reason for industry taking this lead is not far-fetched as industries have access to the computational and financial resources required to create these models [1]. These developments in AI have led to the release of state-of-the-art results and performance saturation on traditional benchmarks [1]. Not only are these saturations being achieved at incredibly fast speed, efforts in AI are also becoming focused on achieving general artificial intelligence, and an example similar to this is seen in the *Prismer model* [3] which integrates multiple models such as object detection model, OCR detection model amongst others to achieve image captioning and visual question and answering (VQA).

Large-scale datasets have become the core of model development and evaluation in AI. The developments and success witnessed in the field of AI can be attributed to datasets as they are foundational and fundamental to the training, benchmarking, and evaluation of AI models. Their collection and creation shape the problems and techniques that are investigated during model development as they serve as the foundation for this process [4]. Benchmark datasets such as MS-COCO [5], ImageNet [6] and OpenImage [7] datasets in the visual object recognition field and GLUE [8] benchmark dataset for English textual understanding have aided significant developments in the field of AI. Therefore, the importance of datasets in machine learning cannot be overemphasized.

However, various studies have highlighted the limitations of certain machine learning datasets in assessing human-like reasoning skills, as well as the concerning societal implications associated with the development and utilization of these datasets [4]. Although benchmark datasets are widely believed to be beneficial in contributing to the generality of trained systems [9], they are impacted by some limitations and biases [10] Generally, current data practices tend to overlook the subjective judgments, biases, and contextual factors inherent in dataset creation. Nevertheless, such details are crucial for evaluating the suitability of a dataset for a specific application, enabling thorough error analysis, and recognizing the substantial challenges involved in constructing effective datasets [4].

For example, a noticeable lack of subjects with darker skin tones in contrast to those with lighter skin tones has been detected in significant facial analysis datasets and image datasets utilized to teach self-driving cars how to identify pedestrians [11], [12]. In a similar manner, [13] discovered that the ImageNet dataset contained millions of person-related images categorized with offensive labels, such as racial slurs and derogatory phrases although such labels were removed afterwards. Also, most of the object recognition dataset have been collected mostly from Western countries. All of these can result in the decreased effectiveness of AI models trained with such datasets. There is also the method of collecting huge data from the internet to form datasets, this method is not robust due to the limitations previously highlighted. One might therefore, resort to collecting and creating datasets manually, this method however, is expensive, time-consuming and can lead to a reduction in the efficiency of model development. I recently experienced this in a project involving the creation of a dataset containing 1000 images. It is therefore imperative, to develop a conceptual plan to generate synthetic datasets (focus of this plan is on visual datasets) with the emerging AI tools rather than continue with the traditional process of collecting and creating datasets. This method would lead to increased efficiency in model deployment and increased effectiveness in the capabilities of AI models.

Existing Methods of Visual Data (Images and Videos) Collection

The following are some of the existing methods of visual data collection

- Collection of large number of images and videos from the web
- Increased reliance on crowd-workers to collect datasets and perform annotations

Methodology / Conceptual Plan

The plan is to use Nvidia Picasso Tool and Nvidia's Omniverse to generate synthetic visual datasets with the aim of removing time-lags, increasing efficiency, effectiveness, and making models robust to shortcomings previously mentioned.

Nvidia's Picasso: Picasso is a visual language model making service for clients who intend to build custom models trained with licensed or proprietary content. This means that users have complete control over the content of the dataset (images and videos) generated. Users can also be intentional about the type of images generated. For example, a user can intentionally prompt Picasso for an image of a busy shop or canteen containing different people of different colors. With Picasso, users can generate a photorealistic image, high-resolution video, or a detailed 3D geometry by giving as input, a text or sentence of what type of image or asset is to be generated. The use of a generative AI-powered tool like Picasso can definitely replace the traditional process of manual collection of data by humans thereby leading to increased efficiency as time is conserved with such AI tool. Also, problem of biases, subjective factors can also be eliminated since users and developers can be intentional about the type of dataset used to train AI models. On the other hand, Picasso when used together with Nvidia's Omniverse Replicator (this platform allows industries to design, build, operate, and optimize physical products and factories digitally before making a physical replica) can find application in industry such as generation of synthetic video dataset for self-driving cars and for robotics systems being trained on reinforcement learning models. This can greatly reduce the expenses and time needed to create the required dataset for such complex systems.

Asides from the fact that photorealistic images and videos can be generated quickly, by giving text input to Picasso, users also have complete control and can create their own unique dataset by integrating other 3D modeling software with Picasso. The 3D models generated can be rendered and used to generate a large quantity of labeled data. This method can also be implemented by developers to replace the traditional process of generating visual datasets. Figure 1.0 is a block diagram showing the plan/workflow for a team intending to use Picasso to generate large quantity of visual datasets.

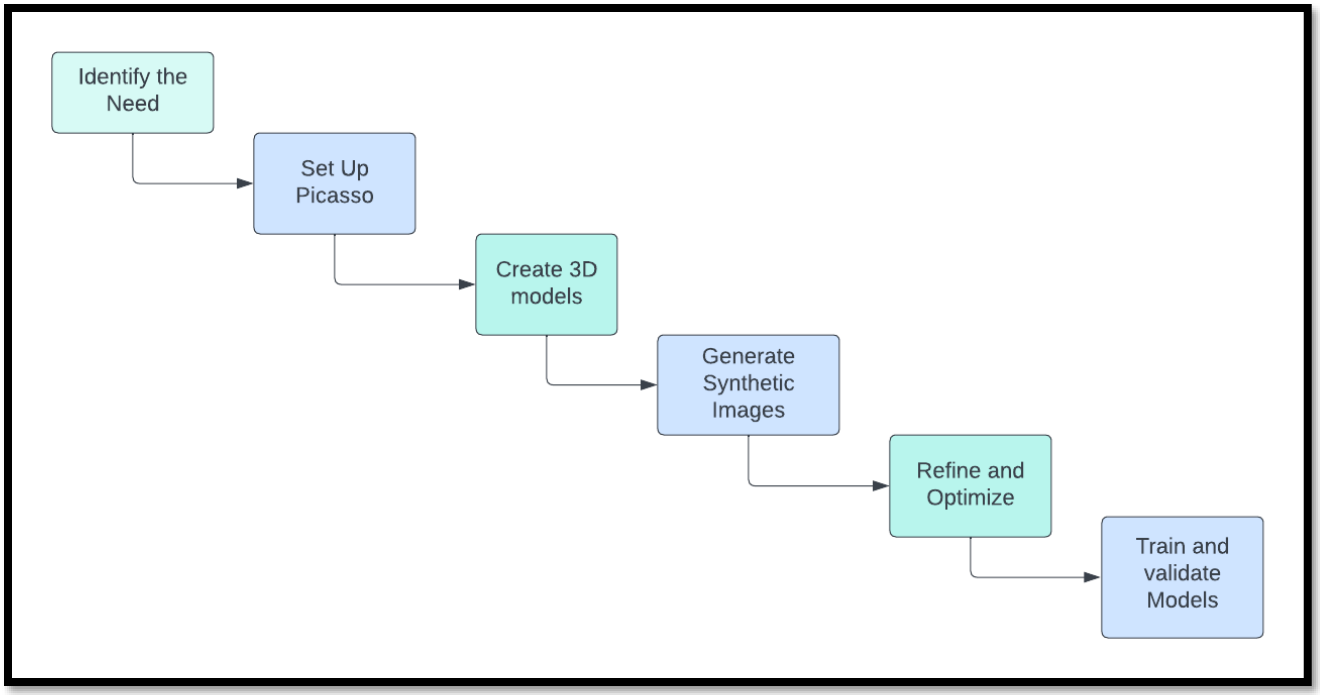

Figure 1.0: Block Diagram of a conceptual plan of generating datasets using Nvidia's Picasso.

Figure 1.0 shows the block diagram of the method of generating datasets using Nvidia's Picasso.

- **Identify the Need**: Firstly, the type of categories of dataset needed is defined and this depends on the task of the model. Datasets required for models powering intelligence in self-driving cars will be different from datasets required for models performing VQA. Once the types of datasets needed are identified, then Picasso can be set up.
- **Set up Picasso**: The tools needed to generate the image datasets such as 3D modeling software tools and texture mapping tools together with Picasso software are launched.
- **Create 3D models**: The 3D models of the objects and scenes that are needed in the datasets are created. The image and video datasets can be made more realistic by adding realistic textures using Picasso.
- **Generate Synthetic Images:** The images in the dataset are now generated by rendering the images of the 3D models. This can result in the fast generation of large quantities of labeled data.
- **Refine and Optimize**: The 3D models and images are further refined and adjusted to meet the requirements of the dataset.
- **Train and Validate Models:** The generated datasets are then split into test and validation dataset and applied to models.

CONCLUSION

AI has come to stay and will continue to evolve. It is only a matter of time before it finds its way to our everyday lives. Therefore, many conventional ways/methods will be replaced by AI due to its ability to increase efficiency and effectiveness. A conceptual plan to replace one of such methods has been discussed. The traditional way of creating visual dataset which include manual methods, sourcing from internet (which might sometimes contain noisy data, biases, offensive categories, racial slurs, harmful and problematic representations, non-consensual images) can be replaced with Nvidia's Picasso tool. With the right team, unique visual datasets of all types can be generated quickly with Picasso thereby reducing the time gap in model creation and development.

Bibliography

[1] Nestor Maslej *et al.*, "Artificial Intelligence Index Report 2023 Introduction to the AI Index Report 2023," 2023.

[2] T. B. Brown *et al.*, "Language Models are Few-Shot Learners," May 2020, [Online]. Available: http://arxiv.org/abs/2005.14165

[3] S. Liu, L. Fan, E. Johns, Z. Yu, C. Xiao, and A. Anandkumar, "Prismer: A Vision-Language Model with An Ensemble of Experts," Mar. 2023, [Online]. Available: http://arxiv.org/abs/2303.02506

[4] A. Paullada, I. D. Raji, E. M. Bender, E. Denton, and A. Hanna, "Data and its (dis)contents: A survey of dataset development and use in machine learning research," *Patterns*, vol. 2, no. 11. Cell Press, Nov. 12, 2021. doi: 10.1016/j.patter.2021.100336.

[5] T.-Y. Lin *et al.*, "LNCS 8693 - Microsoft COCO: Common Objects in Context," 2014.

[6] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *Int J Comput Vis*, vol. 115, no. 3, pp. 211–252, Dec. 2015, doi: 10.1007/s11263-015-0816-y.

[7] A. Kuznetsova *et al.*, "The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale," Nov. 2018, doi: 10.1007/s11263-020-01316-z.

[8] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding," 2018.

[9] M. K. Scheuerman, A. Hanna, and E. Denton, "Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development," *Proc ACM Hum Comput Interact*, vol. 5, no. CSCW2, Oct. 2021, doi: 10.1145/3476058.

[10] D. Boyd and K. Crawford, "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon," *Inf Commun Soc*, vol. 15, no. 5, pp. 662–679, Jun. 2012, doi: 10.1080/1369118X.2012.678878.

[11] B. Wilson, J. Hoffman, and J. Morgenstern, "Predictive Inequity in Object Detection," Feb. 2019, [Online]. Available: http://arxiv.org/abs/1902.11097

[12] I. D. Raji and G. Fried, "About Face: A Survey of Facial Recognition Evaluation," 2021. [Online]. Available: www.aaai.org

[13] K. 'Crawford and "Paglen Trevor," "Excavating AI," *Excavating AI: The politics of Training Sets for Machine Learning*, Sep. 19, 2019.