Synthetic Data Generation for Insect Detection

Roan Meijer, Minor Computer Vision & Data Science
Supervisor: Lucas Ramos

Winter 2024

Introduction

- Collecting, annotating and cleaning data is a task that can be labor intensive, expensive and slow.
- Current data collection methods for insects include placing yellow sticky plates at farms [1], which must be manually annotated by professionals.
- Class imbalances can pose a problem when collecting insect data [2], due to the uncertainty regarding which insects will appear.

Materials and Methods

- An alternative method collection data, can be by generating images using AI. This is cheaper, faster and requires less work.
- Stable Diffusion XL is a customizable text-to-image model that creates realistic looking images, while Dall·E 2, created by OpenAI, creates more abstract looking images, using its powerful prompt analysis.
- Without proper prompt engineering and settings, unexpected results will appear as shown in figure 5.

Abstract

Collecting images for the training of neural networks is a time consuming and difficult process. To try and remedy this, an alternative method of data collection is explored, namely data generation, using generative models DALL-E and Stable Diffusion XL. The research has proven it is possible to substitute up to 75% of images used to train a model with AI generated images without losing accuracy. DALL-E has proven to be the superior model in this use case. The accuracy can be further improved applying data augmentation.

Figure 1. Realistic looking images of crickets generated by Stable Diffusion XL

Figure 2. Abstract looking images of crickets generated by DALL·E 2

Figure 3. An example from each class used in the dataset

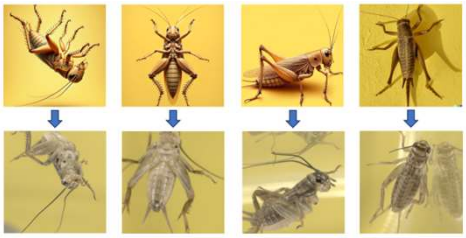

Figure 4. AI generated crickets & real crickets. This image demonstrates how AI generated crickets and real crickets can look alike when they are in the same pose

Figure 5. When an insect image is generated with improper settings, weird mutations can occur.

Experiments

- For our experiment, we focus on replacing real images with as much synthetic data as possible, without hurting performance.
- Both the models, Dall-E and Stable Diffusion XL will be tested in this experiments, with and without data augmentation. An example of data augmentation can be found in Figure 7.

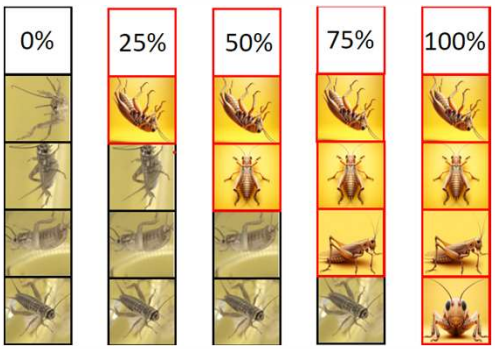

Figure 6. Visualization of the data substitution experiment.

Figure 7A. Original Image

Figure 7B. Mirrored Image

Results

- As seen in Figure 8 and 9, the experiments show that under the right circumstances, up to 75% of the data can be substituted with AI generated data without losing accuracy.
- DALL-E has proven to be the superior model in the experiment, overall achieving an accuracy of 10-20 percent higher then Stable Diffusion XL.
- Data augmentation has proven to be useful when training a model on AI generated data, boosting the accuracy with an average of 10 percent.

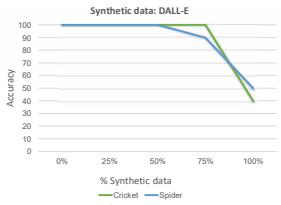

Figure 8. Accuracy of the model trained on DALL-E generated images.

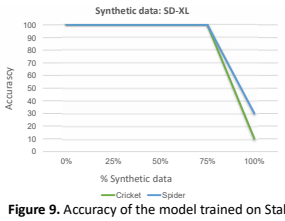

Figure 9. Accuracy of the model trained on Stable Diffusion XL generated images.

Conclusions

- It is possible to use AI generated data to enhance the performance of a classification algorithm.
- 75% of data in our dataset can be replaced by AI generated data without losing accuracy, significantly lowering the required amount of data.
- Replacing all the data in a dataset with AI generated data, provides results substantially better than random predictions.
- To demonstrate the recognition abilities of the model, a demo utilizing bounding boxes and prediction has been made, as seen in Figure 10.

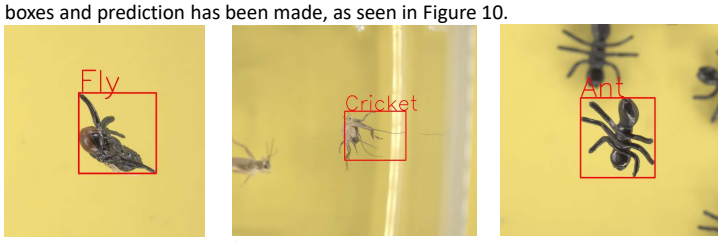

Figure 10. Bounding boxes around classified insects

Acknowledgements

- This project is financially supported by ELFPO and performed within the POP3+ Fryslân project Innovatie luizendetectie

References

- [1] Heinz, K. M., Parrella, M. P., & Newman, J. P. Time-Efficient Use of Yellow Sticky Traps in Monitoring Insect Populations. Journal of Economic Entomology (1992)
- [2] Johnson, J.M., Khoshgoftaar, T.M. Survey on deep learning with class imbala
- nce. J Big Data 6, 27 (2019)