Data centric approach for oil palm trees detection

Issouf TOURE*
data354
Abidjan, Cote d'ivoire
issouf.toure@data354.co

Monsia DOUGBAN
NLP researcher at data354
Abidjan, Cote d'ivoire
monsia.dougban@data354.co

Fabrice ZAPFACK
CTO data354
Abidjan, Cote d'ivoire
fabrice.zapfack@data354.co

Abstract

To enable African agritech startups to keep up with the progress made in recent years in the field of artificial intelligence, we decided to set up a tool for automatically counting oil palm trees in drone images. This tool is a continuation of digital africa's commitment through its data for digital africa (D4DA)[2] program, which aims to promote the use of data by African agritech startups, and is also an improvement on a first solution obtained following a challenge organized on the zindi platform[3]. This first solution was based on a regression approach to predict the number of palm trees in images, without necessarily saying where these palm trees are actually located in the image a sort of black box, as we like to say. The new solution proposed consists in detecting the palm trees in the images for counting purposes, with a view to making the model more explicable. This solution was implemented using a data-centric AI approach with the faster r-cnn pre-trained model. Indeed, the faster r-cnn model was fine-tuned, adapting the output of the last layer to the number of objects we wished to detect in the images, in this case a single object (the oil palm). It should also be noted that with faster r-cnn, we add +1 to the number of objects to define the total number of classes, this +1 corresponding to the background class. The results obtained in this study are quite encouraging: with 312 images annotated on the principle of data centric AI, which we divided into train-test (90%,10%) for training, we obtained an average precision (AP) of 77% at the threshold of 0.75 IoU on the test data. These results can be considerably improved on the basis of data-centric AI principles. Our main aim in this study was to demonstrate the interest of this new approach in the field of data science in general and computer vision in particular.

***Keywords*—** data-centric AI, model-centric AI, objects detection, deep learning, Artificial Intelligence

1 Context and Motivation

Major advances in artificial intelligence in recent years have given rise to new AI models such as chatgpt, mid journey, Segment Anything Model (SAM) and new approaches to creating these AI models such as few short learning, zero short learning and data centric AI. To enable African countries to integrate these new technologies into their technological development process, we decided to implement the data centric AI approach in the

*Computer Vision researcher

[2]https://digital-africa.co/en/

[3]https://zindi.africa/competitions/digital-africa-plantation-counting-challenge

5th Deep Learning Indaba Conference (DLI 2023).

creation of AI models for oil palm counting in drone images.This solution will be used in several African countries, in particular Côte d'Ivoire.Indeed, Côte d'Ivoire, the world's fifth-largest oil palm producer, and Africa's second-largest producer behind Nigeria here, is faced with a significant lack of technological tools in the palm sector in particular, and in the agricultural sector in general, due in part to the considerable illiteracy rate among agricultural players. With a view to addressing this problem, our solution aims to provide African AgriTech startups with a fairly reliable tool for estimating harvests in palm-growing areas, by means of a complete count of the palm trees in these areas.

2 Data centric vs Model centric

Model-centric and data-centric AI: what do we need to understand about these two approaches? [3] . In model-centric AI, the focus is on the model with a fixed dataset, whereas in data-centric AI, the focus is on the data. Indeed, model-centric AI is based on the objective of producing the best model for a given dataset, while data-centric AI is based on the objective of producing the best dataset to feed a given ML model. In the latter approach, the emphasis is on continuous data improvement, whereas in model-centric AI, the focus is on models with a fixed dataset. A data-centric AI pipeline might look something like this:

1. Explore the data, solve the fundamental problems and transform it to make it suitable for ML.
2. Train a basic ML model on the correctly formatted dataset.
3. Use this model to help you improve the dataset (by pre-annotating new data)
4. Try out different modeling techniques to improve the model on the enhanced dataset and get the best model

To deploy the best ML systems, steps 3 and 4 can be repeated several times.

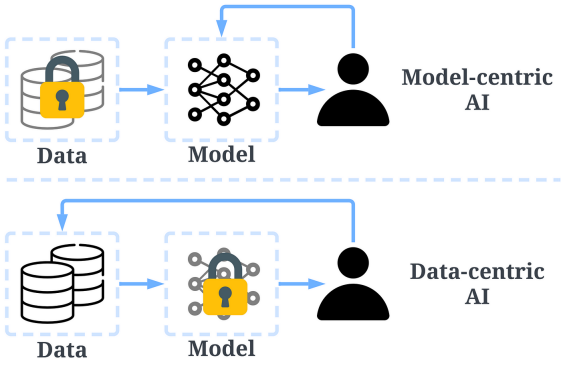

Figure 1: data centric vs model centric [1]

3 Our approach

The general approach of this study is based on data centric AI for the detection of oil palm trees in images collected by drones. To implement this approach, we used the following components:

- Kili Technology
  Kili[4] is an annotation platform with a very large community of annotators and online resources for learning how to use it, as well as a very practical and intuitive annotation interface. Kili's api makes it possible to implement the data-centric AI approach. So, once the model has made predictions on new data, we use the Kili api to push these predictions onto the platform from our work environment (Google Colab, Anaconda), which serve as pre-annotations for the data concerned, so we become reviewers instead of annotators in the process of annotating this data.
- Google Collaboratory
  We used Google Colab as our working environment because of the availability of GPUs (T4, V100, A100). For this study, the T4 GPU was used.
- PYTORCH
  The pytorch 2.0.1 framework was used to develop our artificial intelligence model. The advantage of

[4]https://kili-technology.com/

pytorch is that it enables us to work on large volumes of data, loading data only in batches using the dataloader, thus avoiding RAM saturation.

- FASTER R-CNN
  We opted for transfer learning in this study, for which we fine-tuned the faster r-cnn model to achieve the baseline, which we then improved using the data-centric approach explained in section 2. In fact, to fine-tune the faster r-cnn model, we adapted the output of the last layer to the number of objects we wished to detect in our images, in this case a single object (the oil palm). Note also that with faster r-cnn, we add +1 to the number of objects to define the total number of classes, this +1 corresponding to the background class. In the end, we ended up with a total number of classes equal to 2.

Our overall approach is summarized in the graph below.

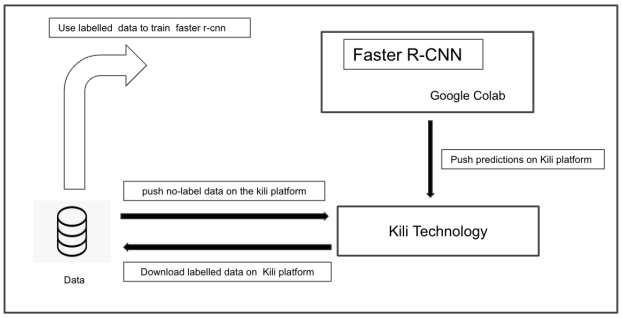

Figure 2: data centric pipeline

Once we had defined our approach, we began with the first iteration, which involved manually annotating 70 images in PASCAL VOC format on the KILI platform we used to create our baseline. Then, in the second iteration, we used the model created to pre-annotate a further 120 images, which saved us a considerable amount of time in the annotation process. These new annotated images were added to the first 70 manually annotated images to re-train the model. This further improved model performance. This process was used to re-label a further 122 images in the third iteration. Finally, we annotated 312 images. To create the final model, this data was divided into train-test. 90% were used for the train and 10% for the test, i.e. 281 train images and 31 test images.

4 Results

For each iteration we trained the model on 20 epochs with a batch size of 4 using Google Collaboratory's T4 GPU. The results obtained in Average Precision (AP) at a threshold of IoU=0.75 on the test data are shown in the below table.

Table 1: Results of training

Iterations	Number of new annotated data	Labeling time	Training data	AP(%)
1	70	3H	70	37,14
2	120	1H30mins	190	55,83
3	122	35mins	312	77,6

Figure 3: image before prediction vs image after prediction

5 Discussion

To address the problem of lack of labeled data in AI model creation in this study, we implemented the new data-centric approach to AI model creation. It should be noted that this approach makes it possible to use an infinite source of unlabeled data available on the web. The results achieved in this study showed that as the number of labeled data increases, so does the model's performance, enabling new data to be labeled much more quickly, reducing labeling time by up to 70% or even 80%. The uniqueness of this study lies in the use of a data-centric approach in the creation of an AI model for an agritech problem (oil palm trees counting).

6 Conclusion

To implement our automatic oil palm trees detection system in drone images, we propose a new data-centric approach, in which we use Kili Technology platform as the labeling platform to implement our approach and faster r-cnn as the pre-trained model. This approach in data science opens up a new era in the implementation of AI models without having to worry about the absence of labeled data, since it enables continuous data labeling. This study has shown that AI model creation is not necessarily linked to the use of the best DL architecture or ML algorithm or to the best pre-trained model on a fixed dataset, but can also depend on a continuous improvement of the data on which a selected model is trained. As future work, we plan to implement new AI models for the agritech sector (cocoa disease detection [2] , cocoa crop harvest forecasting) again using this new data-centric approach.

References

[1] Daochen et al. Data-centric ai: Perspectives and challenges. *arXiv:2301.04819v3*, 2023.

[2] Dong et al. Crop disease diagnosis with deep learning-based image captioning and object detection. *Applied Sciences*, 13, 2023.

[3] Oussama H. Hamid. Data-centric and model-centric ai: Twin drivers of compact and robust industry 4.0 solutions. *Applied Sciences*, 13, 2023.