Help Protect the Great Barrier Reef in detecting crown-of-thorns starfish in underwater image data

Project Proposal

Clément Apavou
MVA Master's student
École Normale Supérieure
clement.apavou@ens-paris-saclay.fr

Guillaume Serieys
MVA Master's student
École Normale Supérieure
serieysguillaume@gmail.com

Abstract

The project proposed is a Kaggle research code competition which aims to accurately identify a starfish species in real-time by building an object detection model trained on underwater videos of coral reefs.

1. Problem

1.1. Context

The Australia's Great Barrier Reef is the world's largest coral reef and home to 1,500 species of fish, 400 species of corals, 130 species of sharks, rays, and a massive variety of other sea life. Unfortunately, this environment of magnificent beauty is threatened by the overpopulation of a coral-eating starfish species named the crown-of-thorns starfish (COTS).

Scientists, tourism operators and reef managers established a large-scale intervention program to control COTS outbreaks to ecologically sustainable levels. To know where the COTS are, a traditional reef survey method, called "Manta Tow", is performed by a snorkel diver. While towed by a boat, they visually assess the reef, stopping to record variables observed every 200m. While generally effective, this method faces clear limitations, including operational scalability, data resolution, reliability, and traceability.

1.2. Application

In order to develop new survey and intervention methods in COTS control, the Great Barrier Reef Foundation established an innovation program. The new survey and intervention methods will consist of collecting thousands of reef images using underwater cameras and then apply AI technology with a near real-time object detection model trained on underwater videos of coral reefs to improve the efficiency and scale at which reef managers detect and control COTS outbreaks. Thus, our work will help researchers identify species that are threatening Australia's Great Barrier Reef and take well-informed action to protect the reef for future generations.

2. Dataset

2.1. Description

In this competition, we will predict the presence and position of crown-of-thorns starfish in sequences of underwater images taken at various times and locations around the Great Barrier Reef. Predictions take the form of a bounding box together with a confidence score for each identified starfish. An image may contain zero or more starfish. The dataset contains a training set of 23,501 annotated images and a test set of 13,000 images whose annotations are not accessible. Each training image have the following metadata:

- video_id: ID number of the video the image was part of.
- video_frame: The frame number of the image within the video.
- sequence: ID of a gap-free subset of a given video.
- sequence_frame : The frame number within a given sequence.
- image_id: ID code for the image, in the format 'video_id-video_frame'.
- annotations: The bounding boxes of any starfish detections. A bounding box is described by the pixel coordinate (x_min, y_min) of its upper left corner within the image together with its width and height in pixels.

2.2. Difficulties observed

First, the crown-of-thorns starfish are difficult to distinguish because of their coral-like appearance. Also, images are taken from a distance which makes it even more difficult. An image of the training set showing the difficulty of the task is in the figure 1.

Second, the training set contains images from only three videos. This details, can be a very challenging detail to train a model so that it generalizes well on data that it has never seen or with contexts different from those seen during the training.

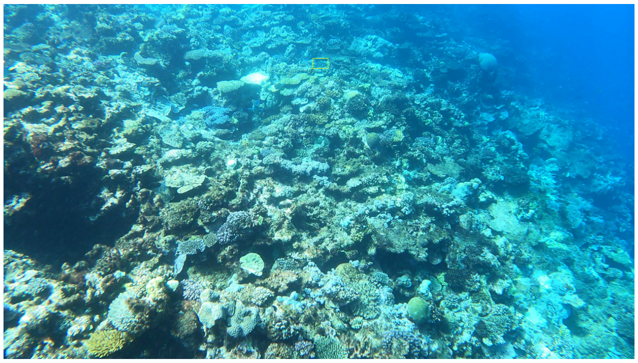

Figure 1. Image of the training set with a COTS not very visible with a bounding box annotation.

3. Approaches

3.1. Handling of the dataset

The first major step to adress the problem is to take charge of the dataset and manipulate annotations. Indeed, we must first analyse it and then distribute images into the training and validation sets with an approximately equal proportion of starfish in both and that there is no overlap between our sets. Furthermore, if we want to use the fact that images are video frames and use the timeline context, the split must take into consideration the images order in a video.

3.2. Object detectors architectures and features extractors

In the second part, we plan to try state-of-the-art object detectors particularly Mask R-CNN [5] (slower than YOLO but more accurate) with Dynamic Head [2] as object detection head to boost its performance, and as backbone we will use state-of-the-art features extractors based on transformers architecture such as Swin Transformers V_1 [8] or the last release version V_2 [7]. The authors of the articles [8, 7] provide an efficient implementation of their proposed architectures which save GPU memory consumption and make it feasible to train large vision models with regular GPUs [6]. We will do experiments with different object detectors and backbones to keep the best model for our problem.

3.3. Training and hyperparameter tuning

We plan to adapt their implementation for our problem and train it by ajusting hyperparameters and obtain a baseline. Then, we have the intention to apply efficient data augmentation and also try data auto augmentation proposed in [1] to increase the learning scope of the model and thus improve its generalisation capability.

3.4. External underwater dataset, unsupervised pre-training methods and camouflaged object detection

In the third part, we have thought to use open source external underwater datasets and apply methods to pretrain the backbone with unsupervised techniques such as proposed in [3]. Furthermore, as explain in 2.2, the COTS has an appearance very similar to that of corals, so it is difficult to detect them. After some research, we found that the camouflaged object detection is a poor topic in work, and to our best knowledge, the published works are for semantic segmentation tasks. So, we will investigate camouflaged object semantic segmentation methods proposed in the articles [9] and SINet [4] to get inspiration and develop a method for the COTS detection more generally for the detection of camouflaged objects.

4. Evaluation

4.1. Metrics

The competition is evaluated on the F2-Score at different intersection over union (IoU) thresholds. The F2 metric weights recall more heavily than precision, as in this case it makes sense to tolerate some false positives in order to ensure very few starfish are missed.

The metric sweeps over IoU thresholds in the range of 0.3 to 0.8 with a step size of 0.05, calculating an F2-score at each threshold. For example, at a threshold of 0.5, a predicted object is considered a ”hit” if its IoU with a ground truth object is at least 0.5.

A true positive is the first (in confidence order) submission box in a sample with an IoU greater than the threshold against an unmatched solution box.

The final F2-Score is calculated as the mean of the F2-scores at each IoU threshold. Within each IoU threshold the competition metric uses micro averaging; every true positive, false positive, and false negative has equal weight compared to each other true positive, false positive, and false negative.

4.2. Kaggle submission

In our submission, we are also asked to provide a confidence level for each bounding box. Bounding boxes are evaluated in order of their confidence levels. This means that bounding boxes with higher confidence will be checked first for matches against solutions, which determines what boxes are considered true and false positives. Furthermore, the competition uses a hidden test set that will be served by an API to ensure you evaluate the images in the same order they were recorded within each video.

References

[1] Ekin Dogus Cubuk, Barret Zoph, Dandelion Mané, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation policies from data. *CoRR*, abs/1805.09501, 2018. 2

[2] Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic head: Unifying object detection heads with attentions. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7369–7378, 2021. 2

[3] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. Up-detr: Unsupervised pre-training for object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1601–1610, June 2021. 2

[4] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[5] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 2

[6] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer object detection code github. `https://github.com/SwinTransformer/Swin-Transformer-Object-Detection`, 2021. 2

[7] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution, 2021. 2

[8] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, October 2021. 2

[9] Jing Zhang, Yunqiu Lv, Mochu Xiang, Aixuan Li, Yuchao Dai, and Yiran Zhong. Depth-guided camouflaged object detection. *CoRR*, abs/2106.13217, 2021. 2