GREAT BARRIER REEF: STARFISH DETECTOR

Nick Wagner and David Unger

Institute of Signal Processing and System Theory, University of Stuttgart, Germany

1. Problem Statement

Abstract

The Great Barrier Reef in Australia is known world-wide for its diverse animal and coral species. Recently, the overpopulation of the coral-eating crown-of-thorns starfish starts threatening the existance of many corals. To allow divers to efficiently remove these star fishes from the corals, an object detection algorithm is used.
The algorithm is based on the ”YOLO”–framework and takes video sequences as input, detects the starfishes and draws bounding boxes around them.

Dataset

- ~23000 images from 3 video sequences
- bounding box labels from csv
- single class object detection problem

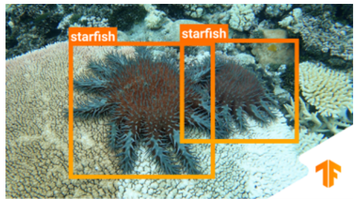

Figure 1: example bounding boxes [1]

2. Architecture

The basis for the following customized architecture is the YOLO (”You Only Look Once”) paper [2]. The most significant changes are the removal of a dense layer, the addition of a dropout layer and dimension changes.

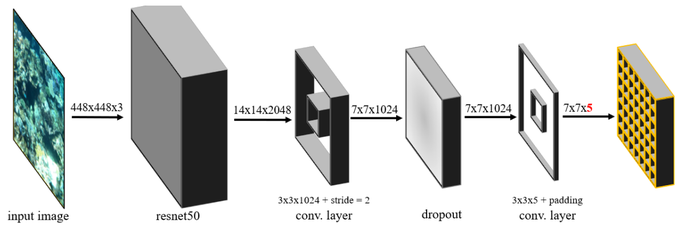

Figure 2: customized YOLO architecture

- **input layer**: image size of 448x448 with three color channels (R,G,B)
- **resnet50**: transfer model that is pretrained on ImageNet data
- **conv1**: reduction of channels and grid size to 7x7 due to stride
- **dropout**: generalization technique to prevent overfitting
- **conv2**: final reduction of channels to 5 with padding

3. Network Output

The final output of the network is a 7x7 grid. Each grid cell predicts one bounding box, which results in 49 predicted bounding boxes.

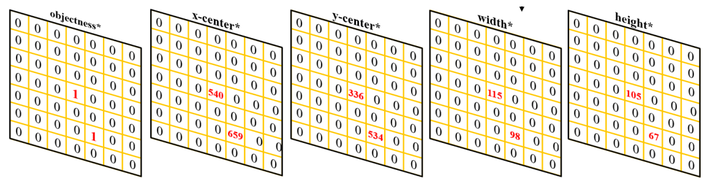

Figure 3: customized YOLO architecture

- **objectness**: confidence of the prediction
- **x-center**: x-coordinate in relation to its cell center
- **y-center**: y-coordinate in relation to its cell center
- **width**: box-width measured in relation to the cell size
- **height**: box-height measured in relation to the cell size

* The values are shown in pixel for understanding purposes. In reality, smaller values relative to the cell size are easier for the model to predict.

Hence, the implementation is done that way.

4. Input Pipeline

Compared to a classification network, a detection network requires a more sophisticated input pipeline. This mainly comes from the labels which are more complex and in addition dependent on the image content and scale, rotation, crop and flip variant. The input pipeline consists of the following steps:

- csv loading
- bounding box text to grid conversion and back
- image reading and resizing
- image augmentation

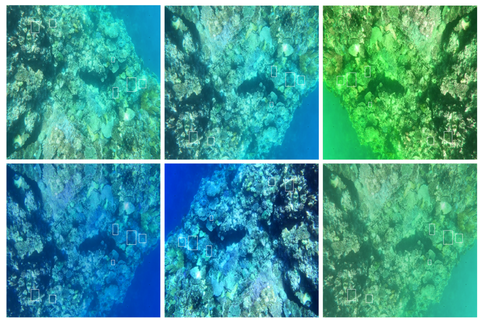

Figure 4: input image augmentated differently

5. Loss Function

To ensure that the network learns the bounding boxes as represented by the ground truth labels, using a standard loss function like MSE is not sufficient. Therefore, the loss function from the YOLO paper is used and adapted that it fits the described problem setup. The loss function consists of four parts, that are summed up during training:

- objectness loss: $l_{obj} = \lambda_{obj} \sum_{i=0}^{(S-1)^2} \mathbb{1}_i^{obj} (1 - \hat{p}_i)^2$
- no object loss: $l_{noobj} = \lambda_{noobj} \sum_{i=0}^{(S-1)^2} \mathbb{1}_i^{noobj} (0 - \hat{p}_i)^2$
- box center loss: $l_{center} = \lambda_{center} \sum_{i=0}^{(S-1)^2} \mathbb{1}_i^{obj} [(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2]$
- box size loss: $l_{size} = \lambda_{size} \sum_{i=0}^{(S-1)^2} \mathbb{1}_i^{obj} [(\sqrt{w_i} - \sqrt{\hat{w}_i})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_i})^2]$

The loss term weightings λ are determined in a way, that all loss terms are in a similar range and therefore are optimized in a similar strength.

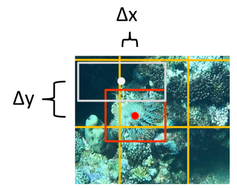

Figure 6: bounding box center loss visualized

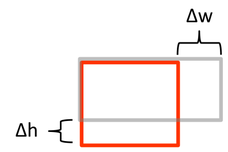

Figure 7: bounding box size loss visualized

6. Challenges

The proposed method works well on the training dataset and shows that the architecture is in general capable of detecting starfishes. Due to the challenging dataset, the YOLO network is not able to detect the starfishes in the test dataset. The task is also for humans very challenging, since the starfishes are hard to detect in the original resolution and even harder to detect in the downscaled version with 448x448 pixel, as one can see in figure 9.

Figure 9: top: original resolution, bottom: downscaled resolution

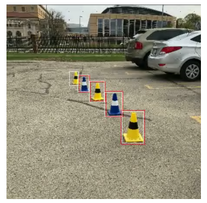

Figure 10: traffic cone detection performance

To endorse this hypothesis of the too challenging dataset for the standard YOLO network, the almost same network that failed to generalize on the starfish dataset was trained to detect traffic cones. The network is able to achieve over 90% precision and an IOU >0.4 on the test data, which are decent results that show the performance of the architecture.

Deep Learning Lab 2022, February 8, Stuttgart, Germany

[1] *https://www.kaggle.com/c/tensorflow-great-barrier-reef* [2] *Redmon, J., et al.: You only look once: Unified, real-time object detection. In: CVPR (2016)*