Focus-CNN for object detection

Paweł Wawrzyński, 2022.06.04

Abstract

The proposed network Focus Convolutional Neural Network is an architecture for object detection in images. It is based on two neural component: (1) the focus network that indicates in the input image a potential location where the object could be and (2) the classifier that verifies if the object is there.

Direct competitor: Faster R-CNN.

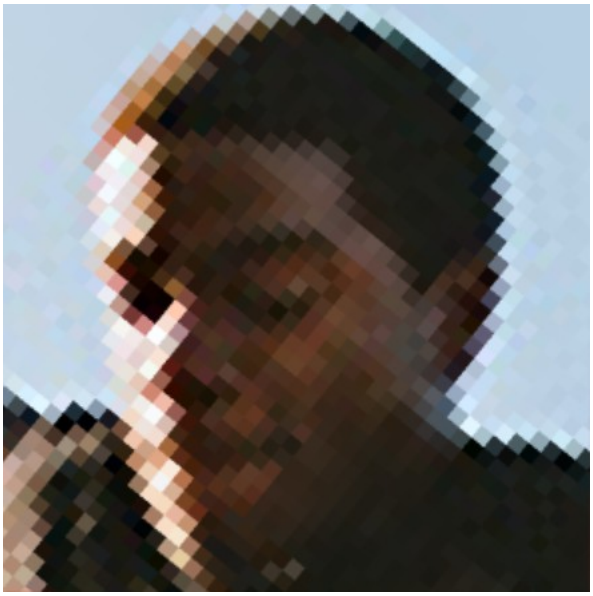
Architecture

Focus network

The focus network is fed with an image. It outputs 5 numbers:

1. x coordinate of the location in the image where the object is likely to be
2. y coordinate of the location in the image where the object is likely to be
3. log(scale), where scale says how much the part of the object needs to be zoomed to the predefined resolution
4. angle at which the indicated part needs to be rotated to its normal view
5. likelihood at which the located part contains the wanted object

Classifier



The classifier is fed with the zoomed and rotated part of the image indicated by the focus network. It output one scalar that says if the image contains an object of the given class or not.

Training

The training can be based on a dataset of images with objects indicated by bounding boxes.

Focus network pretraining

The network is fed with original and rotated images from the dataset. Its job is to learn to indicate the smallest squares that contain the bounding boxes of the objects. The network is also fed with the images that do not contain the object; its job then is to produce the likelihood value equal to zero (the rest of the outputs do not matter).

Classifier pretraining

The network is pretraining with zoomed parts of the images from the dataset. These parts are either defined by the bounding boxes (then the output should be “yes”) or random parts (then the output should be “no”).

Fine tuning

The architecture is fed with the images that either contain or not contain the required objects. The classifier is to output a correct “yes” or “no”. The gradient flows backward through both the networks.

Open question

What happens if there are many object of the given class in the image?

Temporary answer: Then the focus network should indicate the object that is the closest to the image center. The focus network can in principle be applied recursively to parts of the input images to indicate

all the objects.