Deep Learning Project

Alfonso Sanchez

April 17, 2018

1 OBJECTIVE

The objective of this project is the implementation of a fully convolutional neural network (fcn) for the control of a drone, which has the objective of follow a specific target. The fcn training and validation will be based on data provided by Udacity, having as objective an accuracy of .40, in terms of the Intersection over Union (IoU) metric. In the next figure is presented the objective of the project.

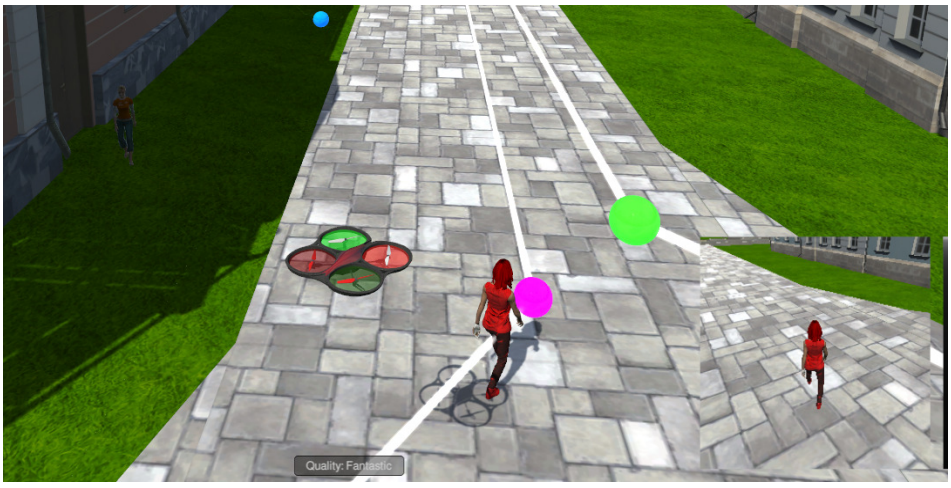

Figure 1: Drone objective.

2 NETWORK ARCHITECTURE

A Fully Convolutional Neural Network was implemented for this project, because of its robustness and success when implemented to find important characteristic in images. A fcn is used instead of a convolutional neural network because we are interested in obtaining spatial properties of a predefined target, with respect to its environment.

Then, a convolutional neural network is used to get important characteristics of images that would allow the network to identify a specific target, by adapting their filters weights. However, as we are interested in knowing the location of the target with respect to its environment, the convolutional neural network is transformed into a fully convolutional neural network, to get the spatial information.

For the convolutional neural network, it was used an encoder function to generate each of the convolutional neural network layers, function that reduces the depth for each generated layer relying on the *SeparableConv2DKeras* function, which acts as pooling layers. Furthermore, the *BatchNormalization* function is used to normalize the inputs for each layer when training them, in order to normalize their means and variances.

Hence, the last encoder layer is connected to a $1x1$ convolutional layer, which encodes the content of the images to let us know if something specific is in it, preserving spatial characteristics. This is used to determine if a person of interest, or target, is present in the images.

Lastly, in order to know the specific location for the target with respect to the images, decoder layers are used, transforming the convolutional nueral networks into fully convolutional neural networks. This is possible using an *upsampling* function. The output of the decoder layers give us enough information to know the location of the target.

2.1 NETWORK ARCHITECTURE CHARACTERISTICS

In the following table are presented the characteristics of the network used, their number of layers, the number of filters used, the strides and the kernels sizes. Additionally, in figure 2 is presented the network architecture corresponding to the parameters presented.

Lalyer	**size**	**Filters**	**Strides**	**KernelSize**
Encoder 1	80 x 80	32	2	3 (default)
Encoder 2	40 x 40	64	2	3 (default)
Encoder 3	20 x 20	128	2	3 (default)
1x1 Convolution	20 x 20	256	1	3
Decoder 1	40 x 40	128	2	3 (default)
Decoder 2	80 x 80	64	2	3 (default)
Decoder 3	160 x 160	32	2	3 (default)

The number of filters were used to acquired specific characteristics of the images, while the strides were used to determine the scan of the images using kernels of specific size. In figure 2 are also presented a brief idea of how the network is used to determine the position of the target. Furthermore, the fully covolutional neural network was implemented *skipping connections* for the decoder parts.

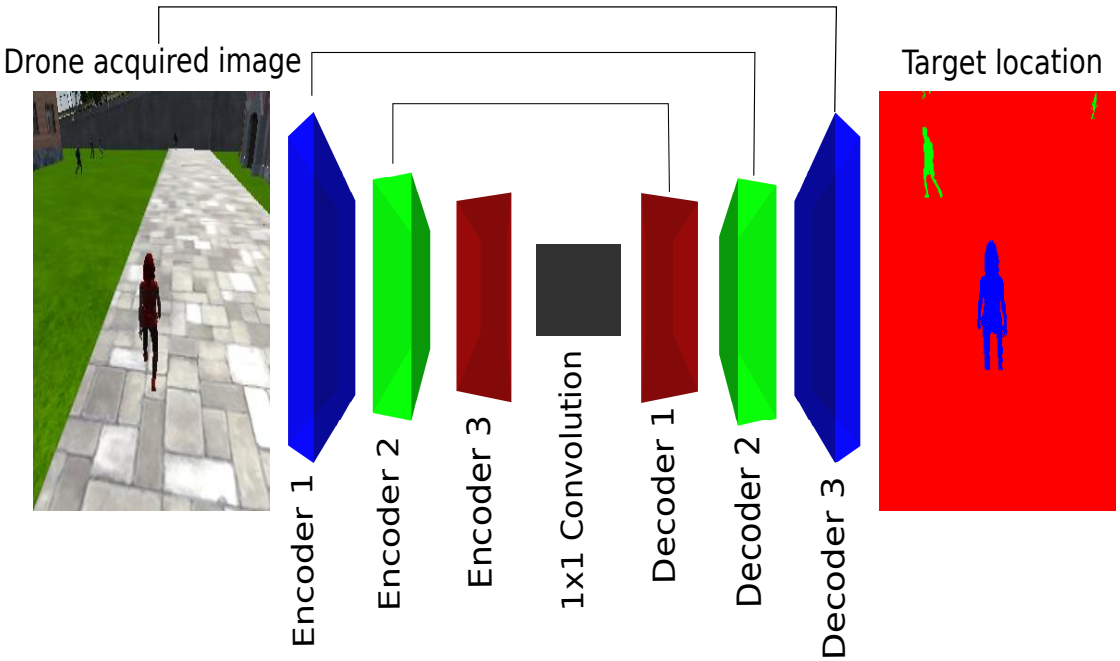

Figure 2: Fcn architecture, implemented for the problem.

The selected network allow us to identify the target and to know its location with respect to the envormonet. Different configurations of newtworks were used, using less encoders and decores, decrease the filters for each layer and changing the by increasing the stride values, however the presented networls gave the best peformance in terms of the accuracy metric when compared with the rest of the implemented networks.

3 HYPERPARAMETERS SELECTION

The selection of hyperparameters was based on the amount of images available for the training and validation of the previously presented network, considering the network characteristics. In the following table can be observed the hyperparameters used for the training and validation of the network:

Parameters	Value
learning rate	.002
batch size	50
num epochs	50
steps per epoch	Number of Training Images / batch size
validation steps	Number of Validation Images / batch size
workers	10

The learning rate was set to .002, after tryed different values from .02 to .001. The size of the data batches used to train the network, instead of using all available data at one time, was set to 50. The batch_size was to 50 because the trained network got its best accuracy with this value, when compared against batches of sizes equals to 30 and 40. Additionally, the best number of epochs to train the networks was 50, when compared against values of 30 and 200, in terms of accuracy and training and validation times.

Furthermore, the steps per epochs and validation steps were set to:

$$steps\_per\_epoch = \frac{Number\_of\_Training\_Images}{batch\_size} \tag{1}$$

$$validation\_steps = \frac{Number\_of\_Validation\_Images}{batch\_size} \tag{2}$$

The steps per epochs and validation steps values were selected in order to tried to divide the data into batches of exactly equal size, as mentioned in the *TensorFlow for Deep learning module*. Finally, the hyperparameter workers value was set to 10, after tried values of 4 and 8.

4 FULLY CONNECTED AND 1X1 CONVOLUTIONAL LAYERS

A fully connected layer is normally used to identify specific characteristic from a specific input. This is possible by adapting their layers weights accordingly to what we would like to find, were all layers nodes are fully connected among them. However, when compared with more robust networks, or layers, it has less information since it represents data in 2D.

On the other hand, a $1x1$ convolution layer preserve all possible information, when compared with a fully connected layer, since it represent data in 4D. In our problem, this allow us to preserve spatial information, that can be expanded using decoder layers and upsampling techniques to relate them with the image obtained by the drone.

Nonetheless, both layers, or networks, can be used to obtain information that allow us to identify specific patterns from a predefined input, with the difference of the amount of information that each of them can provide. However, more information is not always the best, that will depend on the problem that is trying to be solved.

5 IMAGES PROCESSING

The encoding of images is important to get specific features that can be used to identify objects of interest, for example, the following filters are used to obtain the silhouettes of specific profiles of a car, depending on the filter. Then, more filters can be used on the filtered images in order to get more specific patterns, which can be used to identify objects of interest.

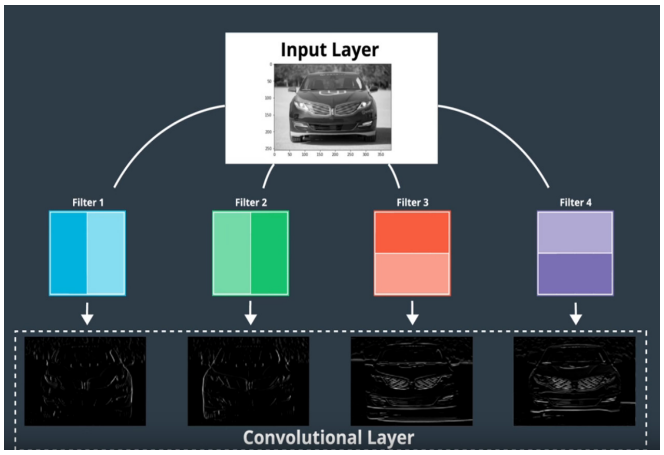

Figure 3: Effects of encoder layers.

The use of $1x1$ convolutional layers, instead of fully connected layers, retains spatial information of the input to the network, which can be used in combination with upsamples techniques to generate decoder layers to reconstruct the position of the object of interest with respect to its environment. This can be observed better in the following figure, where are presented the performance of the network identifying the object of interest.

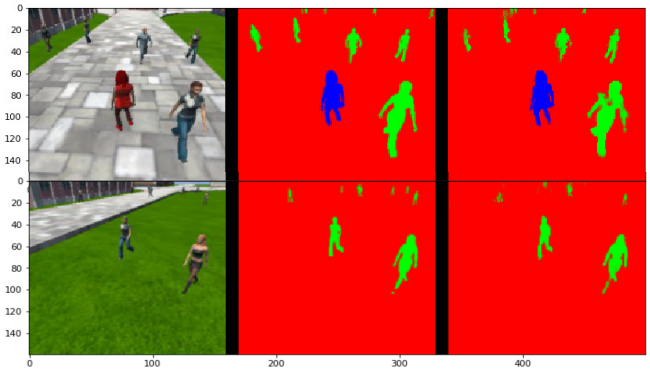

Figure 4: Top, target identified by the FCN. Bottom, the performance of the network when the target is not present in the image acquired by the drone.

6 DATA LIMITATIONS

It is important to remark that the trained network only works for the specific target used to train it. If a different target is required to be followed, or to be recognized, then the network should be trained using images when the desired target is present. For example, if a dog is used as target, its breed, color, and any distinguishable feature needs to be considered into the trained network, the same for a car and a cat, where their size and colors need to be considered. This is necessary for the network in order to adapt their weights to identify the object of interest.