Scope

The scope of this project is to train a fully convolutional neural network so that a drone can recognize a specific person and follow them in simulation.

Intro

In previous lessons, neural networks were trained to recognize images of letters. Is this a picture of an A? B? etc. Next the idea of a convolutional neural network (CNN) was introduced. Rather than applying the model to the entire image at once, a CNN will apply the model to a small window of the full image, sliding this window across the entire image. The advantage of a CNN over a regular deep neural network is that a CNN can be trained to recognize ‘Is this a picture of an A’ with the A appearing anywhere in the picture. To train a deep neural network to do the same would require a much larger data set. Fully convolutional networks (FCN) are an improvement over CNNs. In a FCN the fully connected layer is replaced with a 1x1 convolutional layer, allowing spatial information to be retained. The benefit of a FCN is that it will say not just ‘Is this a picture of an A?’ but ‘where in this picture is the A?’

Network Structure

Encoder

The encoder portion of the network is a series of convolution steps. The encoder gradually squeezes spatial dimensions while increasing the depth. More depth provides more feature maps for semantic segmentation.

1x1 Convolution

The use of a fully connected layer would change the output dimensions of the convolution tensors from 4D to 2D, causing spatial information to be lost. Replacing the fully connected layer with a 1x1 convolution maintains the spatial information.

Decoder

The decoder upscales the output of the encoder such that it is the same size as the original image. This allows the FCN to be applied to images of any size.

Skip Connections

Skip connections are connections of one layer’s output to a non-adjacent layer’s input. As convolutions are applied to an image the network is effectively looking closer and closer at the image. By including information from a different level of “zoom,” bigger picture information is retained.

Gradually squeeze spatial dimensions while increasing the depth

Depth corresponds roughly to semantic complexity