Udacity RoboND Deep Learning Project: Follow Me

Brett Gleason

# 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.

## Separable Convolutions

The encoder and decoder blocks of the FCN were constructed using separable convolutions. In a separable convolution a convolution is performed on each channel of the input layer individually, and then a 1x1 convolution is performed on the output. Using separable convolutions reduces the number of parameters that are needed in the convolution step, decreasing the computational burden.

## Batch Normalization

Another technique used to increase the training speed of the network is batch normalization. To do batch normalization the inputs to each layer of the neural network are normalized, instead of only the inputs to the first layer. The encoder, 1x1 convolution, and decoder layers are all implemented using batch normalization. In addition to increasing the overall training speed of the network, networks using batch normalization can use higher learning rates than other networks.

## Bilinear Upsampling

The decoder must create a layer with more spatial information than the previous layer. In order to do this the decoder must use a technique to interpolate between points in the input layer. The method used for this project is called bilinear upsampling. For this process the weighted average value of the four nearest known pixels are used to estimate the value for the new pixel.

# Code

The model for this project was built using Keras, an open source library containing abstractions for the constituent pieces of neural networks.

## Encoder Block

The encoder block accepts three arguments: the input layer, the number of feature maps, and the stride length of the convolution. A batch-normalized separable convolution is applied to the input layer to create the output layer.

## 1x1 Convolution

The 1x1 convolution accepts two arguments: the input layer and the number of feature maps. A batch-normalized 1x1 convolution is applied to the input layer to create the output layer. A separable convolution is not used because it would not reduce the number of parameters for a 1x1 convolution.

## Decoder Block

The decoder block accepts three arguments: the input layer, the skip connection layer, and the number of feature maps. The input layer is first upsampled using bilinear upsampling. Next the upsampled input layer is concatenated with the skip connection layer. A batch-normalized separable convolution is applied to the concatenated input layers to create the output layer.

# Training

## Run 1

## Run 2

## Run 3

## Run 4

## Run 5

Depth corresponds roughly to semantic complexity