13 - Fully convolutional networks

Juan Felipe Cerón
Universidad de los Andes
Cra 1 Nº 18A - 12, Bogotá - Colombia
jf.ceron1010@uniandes.edu.co

Daniel Rodriguez
Universidad de los Andes
Cra 1 Nº 18A - 12, Bogotá - Colombia
da.rodriguez1253@uniandes.edu.co

1. VERY Abstract

Deep neural networks have great success classifying images or in other words they are very good at attaching labels to images. Here, we have less success in the task of segmenting an image or in other words attaching labels to pixels. We use an implementation of the Fully Convolutional Network published in [**?**].

2. The problem

Semantic segmentation is the task of grouping the pixels in an image so that the resulting groups correspond to the different objects in it, similarly to the way a person would identify them.

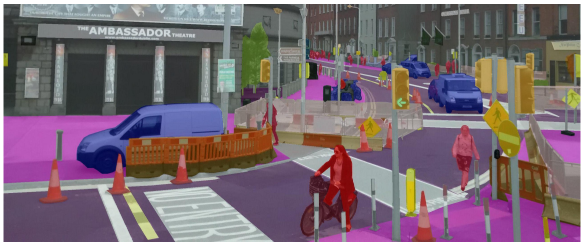

Figure 1. We want something like this: (Source: https://sthalles.github.io/deep_segmentation_network/)

Not long ago, this task was carried out in unsupervised procedures due to the lack of labelled datasets and faith in neural networks. FCN was one of the first successful supervised algorithms applied in this context. In a supervised schema, we can treat this problem as a pixel-wise classification: We attempt to assign the correct class label to each pixel, out of a set of known classes specific to our training set.

3. Measuring success

There are two popular metrics for this:

3.1. Class-wise

For each image and each class, we want to convey how well the pixels from the class were labelled in the image. One way to do this, which balances precision and recall, is the Jaccard coefficient. Let G_c be the set of ground truth pixels of class c in the image and E_c the set of pixels estimated by our algorithm to be of class c. Then the Jaccard coefficient for class c can be defined as

$$J_c = J(G_c, E_c) = \frac{G_c \cap E_c}{G_c \cup E_c}. \tag{1}$$

We can then average J_c over all of the classes (this can be a weighted average) to obtain a goodness measure for the segmentation of a particular image.

3.2. Pixel-wise

Let C be the number of classes in the dataset. Then we can measure success in the classification of an image just as we would in a classification problem with C classes (by mean accuracy, for example).

4. The method

Suppose we start with an image of height h and width w, and with h number of channels. Fully convolutional networks have three basic ingredients.

- Convolution
- Pooling
- Activation

In this method, those ingredients are arranged in the following manner:

- Add two convolutional layers.
- Mix ReLU between the convolutional layers.
- Maxpool the output to reduce model's complexity.
- Salt (Repeat) to taste.

Since we are dealing with segmentation tasks we must return an image. In order to do so, the authors deconvolute the image or in less fancy words, they reexpand the image. They use local information previously extracted from the image to smoothen out the reexpansion.

Here, we trained the 32s network from scratch, the 32s network pretrained with VGG weights, the 16s network and the 16S network pretrained with the 32s weights. The first 2 networks were trained for 48 hours and the latter 2 for 36 hours. The network was implemented such that the chekpoint was updated every time a better model was built.

The main advantage of Fully Convolutional Layers over Fully Connected Layers is that they need not to have a fixed input size.

5. The problem with the method (us) ((CUDA)).

The checkpoint was so heavy that CUDA ran out of memory.

6. BATMAN OF THE FUTURE

This method is so popular that it already has 9000 citations. A famous spin off is called U-Net, a network which parallers its encoding and decoding structure. It was initally