Implement SegNet Encoder-Eecoder Architecture.

Arik Md. Isthiaque
1103096
Section B
Level 4, Term 2
Department of Computer Science and Engineering
BAIUST

November 4, 2019

1 Topic Characteristics

SegNet, is an encoder decoder architecture built with deep convolutional neural network for pixel-wise semantic segmentation for images and videos. It is generally built for roads, buildings, cars, pedestrians and for differentiated the context between roads and side-walks [1]. SegNet extracts low resolution pixel information from input image and classify them as per pixel. This extraction must generate some patterns which can be used in boundary localization.

Our main contribution on this paper is to study and analysis the details of SegNet encoder decoder architecture and reduce the parameters of convolutional networks and perform Canny edge detector algorithm for more boundary localized feature maps, so that it can be more memory efficient and produce more robust sematic segmentation.

Image segmentation is a process where the objects of an image is classified by their pixel information. In an image certain pixel share some common characteristics. By identifying them we can label the pixels, so they can be detected from the image. We use image segmentation to find objects in images like text (OCR), cars (automated vehicle), cancer, tumor (medical image). In recent days, deep learning algorithms got a success in handwritten OCR, NLP [3], [4]. It has also a huge active interest for pixel-wise semantic segmentation. Recently there are some approaches are being carried out to adopt deep learning algorithm for pixel-wise semantic segmentation [5]. The results of these approaches are very encouraging [6]. But we still need better architecture to run the segmentation process. So, improvise the current architectures is needed for more robust segmentation.

2 Working Hypothesis

The architecture will take RGB image as input and the input will be directly fed into the encoder network. The encoder network is topologically identical to the 13 convolutional layers in the VGG16 network [1]. The network will perform convolution with a filter bank to extract a set of feature maps. Then they will go for a batch normalization process [9]. Then an element-wise ReLU will be applied. After each convolutional operation a max-pooling will be performed. Max-pooling is used to achieve translation invariance over small spatial shifts in the input image. The most important phase of this architecture is the decoder network. Here, there is no learning involved in this network. First upsampling will be performed in the feature maps. Here the upsampling process will be performed by inverse convolution using a fixed or trainable multi-channel upsampling kernel. The final decoder output feature maps will be fed to a soft-max classifier for pixel-wise classification. After that we will run Canny edge detector for more localized boundary detection.

3 Methodology

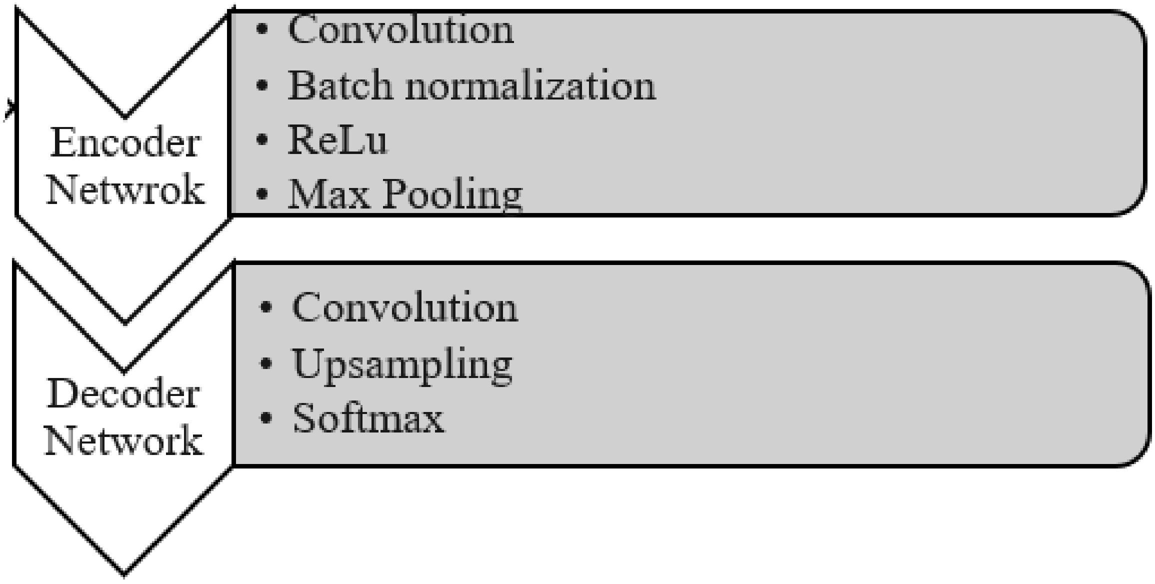

Figure 1: The proposed methodology