KiU-Net

Astarag Mohapatra

October 2, 2022

1 Motivation behind the paper

1. U-Net is a popular segmentation mask. It has an encoder and decoder layer. In the encoder layer we convolutional and max pooling layers and in the decoder layer we have upsampling, convolutional (3x3 and 1x1 filter size).

   - The encoder and decoder layer are concatenated by skip connections.
   - This helps the model to learn both high level and low level features.
   - Max-pooling layers destroys locality of objects (as it reduces the dimensions). But due to the concatenation, locality is conserved and it leads to better segmentation

2. But U-Net performance suffers in segmenting smaller objects. But in medical images smaller objects play an major role in diagnostics.

   - Like in the paper it is pointed out that absence of septum pellucidum which is 6mm in length, can lead to diagnosis of brain dis-orders.

3. Also accuracy is affected by improper boundary edges of ventricles. So due to all these problems we U-Net is considered as under-complete architectures.

2 What problems the paper focuses on?

1. The main problem we talked about U-Net is problem in detection of smaller objects. Smaller objects require smaller receptive fields.

   - Receptive fields can be thought of as the area that the convolutional filter size concentrates on. So receptive field is the size of the input which produces one node in the feature map.
   - Suppose we have an MNIST data set of $28x28$ images and filter size of $3x3$ then the filter concentrates on 9 pixels(3x3x1, 1 for gray scale color channel) at a time which has 9 weight parameters and 1 bias parameter.

   - But after a convolution followed by max pooling operation we get dimension of the image as 13x13. We know that max pooling discards the unimportant information and retains the highlighting pixel or average of the pixels.
   - So now the same filter size 3x3 concentrates on 9 pixels on the 13x13 image but implicitly it is concentrating on 18 pixels in the original image (28x28 image). So the receptive field size is increasing due to max pooling layers

2. In U-Net encoder layer we have max pooling layers which leads to increase in receptive field size but for smaller objects we need smaller receptive field size.

3. So this problem is tackled here, Ki-Net or Kite-Net is an over-complete architecture

3 Ki-Net or Kite-Net

1. This network is over-complete in spatial sense as it does not decrease the dimension of the image.It has an encoder and decoder layer
   - In encoder layer we have an convolutional operation followed by bilinear interpolation up sampling. This increases the dimension. This leads to decrease in receptive field size and it will help the network to capture smaller and finer details.So it restricts the size of the receptive field by up sampling layers
   - To compensate for the increase in dimension, in decoder layer we have convolution followed by max-pooling operation.

2. But alone this layer cannot achieve satisfactory results. We need to augment this architecture with U-Net to get better results. Ki-Net helps to learn small level features and when both results are combined we get superior segmentation. The combined architecture is called as **KiU-Net**
   - Ki-Net helps to capture low-level features and U-Net captures high level features and both of them learn together.
   - We have both the networks running parallel and instead of concatenating their outputs in each convolutional layer we combine the feature maps. This is called as CRFB (Cross-Residual Fusion block).
   - The feature map (output of each layer) from U-Net is passed through Conv2D layer and then up sampled. Then this is added to the feature map corresponding to the Ki-Net.
   - The feature map (output of each layer) from Ki-Net is passed through Conv2D layer and then max pooling. Then this is added to the feature map corresponding to the U-Net.

- So in short, Ki-Net learns lower features and U-Net learns high level features and both of them share their knowledge with each other through CRFB and finally outputs from both layer is concatenated and passed through 1x1 convolution to get the number of required channels.

4 Data sets and Implementation details

1. Cross-entropy loss was used here with Adam optimizer and batch size of 1. The learning rate was set to 0.001 and was trained for 100 epochs.

2. Data sets included

    - GLAS or Gland Segmentation dataset
    - RITE (Retinal image vessel tree extraction).
    - Three models were compared, KiU-Net, U-Net and Seg-Net and KiU-Net gave superior dice score and Jaccard Index (IoU).

5 Experimentation results and Inferences

1. KiU-Net is able to perform so effectively as it consumes less time because it has less number of parameters. It does not want the network to be deep rather it tackles the problem of smaller object segmentation by restricting the receptive field size. So it has specialized architectures to solves issues of smaller and higher level features, so it doesn't need to be deep. Less parameters hence less time for convergence.

2. Only Ki-Net will give poor results as it is unable to capture high level features, but combined with U-net it gives better results.

3. It gives nearly 2% increase in dice score and 4% increase in Jaccard index compared to state-of-the art model.

4. Also the CRFB gave better performance than if we have combined both layers by skip connections. So combining feature maps is a better idea than skip connections.

6 TL;DR

1. Max pooling leads to increase in receptive field size and it misses smaller objects. So to tackle it we need to restrict the size of receptive field.

2. Thus we tackled this problem by introducing an architectural change that is Ki-Net. It has encoder and decoder layer just like U-Net but in encoder layer we have Conv2d and upsampling followed by ReLU activation. In

decoder we have Conv2d and max pooling to compensate for the increase in dimension

3. It can detect small features and it needs to be augmented with U-Net which will detect high level features. They are combined by CRFB instead of skip connections which gave better results.

4. Ki-Net learns lower features and U-Net learns high level features and both of them share their knowledge with each other through CRFB (combining the goods of both world) and finally outputs from both layer is concatenated and passed through 1x1 convolution to get the number of required channels. This gave superior results compared to state-of-the art models.