Attention to Scale: Scale-aware Semantic Image Segmentation

(2016)

Liang-Chieh Chen et al.

Resume

July 14, 2019

---

1 Introduction

In this work, the authors propose an attention mechanism that learns to softly weight the multi-scale features at each pixel location, an seen in the figure below, we pass the image at two different scales through the network, and pass the activation before the softmax to the attention model, the attention model will give us weigts for each scale, that we'll use to do a weighted sum of the two score maps predicted by each scale for the final predictions.

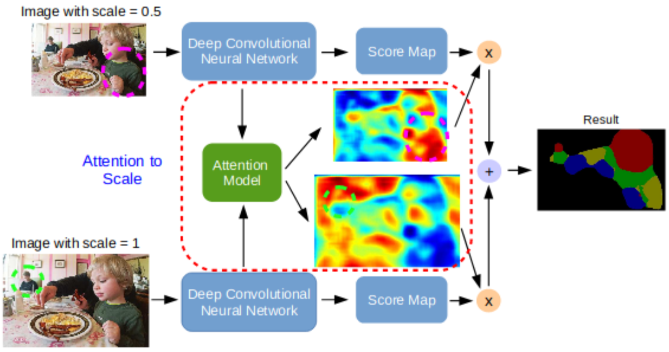


Figure 1. Model illustration. The attention model learns to put different weights on objects of different scales. For example, our model learns to put large weights on the small-scale person (green dashed circle) for features from scale = 1, and large weights on the large-scale child (magenta dashed circle) for features from scale = 0.5. We jointly train the network component and the attention model.

2 Attention mecanism

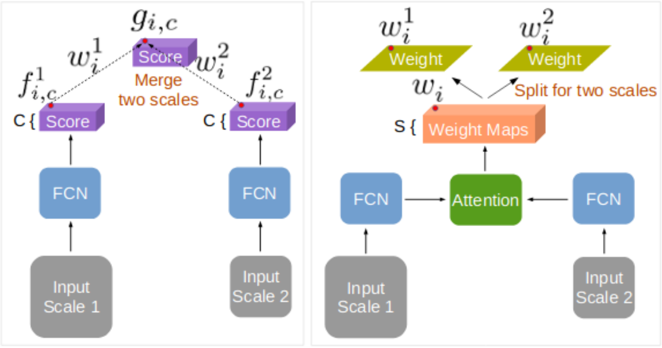
The attention mecanism works by first feeding the features maps of the FC7 pf VGG to the attention model, this is then passed through two convolutionnal layers, the first outputs 512 channels, and the last one outputs S channels, each one of these chanles are the weights for the corresponding scale, from the paper, even if it is note stated clearly, the attention model, obtains S feature maps of FC7, that are resized to the smaller size of all of them, and them each volume will be responsible to produce one channel of the S channels in the output, and after this we'll also resize the score maps (the un-normalized propabilities, before being passed to the softmax layer) we'll weight and add, first we resize them, and multiply each volume with its respective weights and add the results and pass them through the softmax to fet the results.

Each scale is passed through the DeepLab and produces a score map for scales, denoted as $f^S_{i,c}$, where i ranges over all the spatial positions and $c \in \{1, \ldots, C\}$ where C is the number of classes of interest. The score maps $f^s_{i,c}$, are resized to have the same resolution with respect to the finest scale by bi-linear interpolation. And finally we compute the weighted sum of score maps at (i, c) for all scales:

$$g_{i,c} = \sum_{s=1}^{S} w_i^s \cdot f^s_{i,c}$$

And the weight are computed by a softmax of computed feature maps by the last layer of the attention model:

$$w_i^s = \frac{\exp\left(h_i^s\right)}{\sum_{t=1}^{S} \exp\left(h_i^t\right)}$$



(a) (b)

Figure 3. (a) Merging score maps (*i.e.*, last layer output before SoftMax) for two scales. (b) Our proposed attention model makes use of features from FCNs and produces weight maps, reflecting how to do a weighted merge of the FCN-produced score maps at different scales and at different positions.