Attention to Scale: Scale-aware Semantic Image Segmentation

Liang-Chieh Chen*
lcchen@cs.ucla.edu

Yi Yang, Jiang Wang, Wei Xu
{yangyi05, wangjiang03, wei.xu}@baidu.com

Alan L. Yuille
yuille@stat.ucla.edu
alan.yuille@jhu.edu

Abstract

Incorporating multi-scale features in fully convolutional neural networks (FCNs) has been a key element to achieving state-of-the-art performance on semantic image segmentation. One common way to extract multi-scale features is to feed multiple resized input images to a shared deep network and then merge the resulting features for pixel-wise classification. In this work, we propose an attention mechanism that learns to softly weight the multi-scale features at each pixel location. We adapt a state-of-the-art semantic image segmentation model, which we jointly train with multi-scale input images and the attention model. The proposed attention model not only outperforms average- and max-pooling, but allows us to diagnostically visualize the importance of features at different positions and scales. Moreover, we show that adding extra supervision to the output at each scale is essential to achieving excellent performance when merging multi-scale features. We demonstrate the effectiveness of our model with extensive experiments on three challenging datasets, including PASCAL-Person-Part, PASCAL VOC 2012 and a subset of MS-COCO 2014.

1. Introduction

Semantic image segmentation, also known as image labeling or scene parsing, relates to the problem of assigning semantic labels (*e.g.*, "person" or "dog") to every pixel in the image. It is a very challenging task in computer vision and one of the most crucial steps towards scene understanding [18]. Successful image segmentation techniques could facilitate a large group of applications such as image editing [17], augmented reality [3] and self-driving vehicles [22].

Recently, various methods [11, 15, 37, 42, 58, 34] based on *Fully Convolutional Networks* (FCNs) [38] demonstrate astonishing results on several semantic segmentation benchmarks. Among these models, one of the key elements to successful semantic segmentation is the use of multi-scale features [19, 45, 27, 38, 41, 34]. In the FCNs setting, there are mainly two types of network structures that exploit multi-scale features [54].

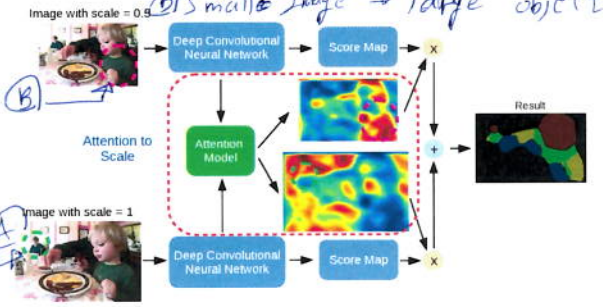

Figure 1. Model illustration. The attention model learns to put different weights on objects of different scales. For example, our model learns to put large weights on the small-scale person (green dashed circle) for features from scale = 1, and large weights on the large-scale child (magenta dashed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

The first type, which we refer to as *skip-net*, combines features from the intermediate layers of FCNs [27, 38, 41, 11]. Features within a skip-net are multi-scale in nature due to the increasingly large receptive field sizes. During training, a skip-net usually employs a two-step process [27, 38, 41, 11], where it first trains the deep network backbone and then fixes or slightly fine-tunes during multi-scale feature extraction. The problem with this strategy is that the training process is not ideal (*i.e.*, classifier training and feature-extraction are separate) and the training time is usually long (*e.g.*, three to five days [38]).

The second type, which we refer to as *share-net*, resizes the input image to several scales and passes each through a shared deep network. It then computes the final prediction based on the fusion of the resulting multi-scale features [19, 34]. A share-net does not need the two-step training process mentioned above. It usually employs average- or max-pooling over scales [20, 14, 44, 15]. Features at each scale are either equally important or sparsely selected.

Recently, attention models have shown great success in several computer vision and natural language processing

*Work done in part during an internship at Baidu USA.