Bag of Surrogate Parts: one inherent feature of deep CNNs

Yanming Guo
y.guo@liacs.leidenuniv.nl
Michael S. Lew
mlew@liacs.nl

LIACS Media Lab
Leiden University
Leiden, the Netherlands

In this paper, we first develop a new feature from the last pooling layer (i.e. $pool_5$) of VGG, called Bag of Surrogate Parts (BoSP), and its spatial variant, Spatial BoSP (S-BoSP). Next, we propose a scale pooling scheme for better handling the objects that may appear in different shape, positions and scales. Finally, aiming that the traditional data augmentation focuses more on part the original image, we further raise a global constrained augmentation method to make a more comprehensive prediction. The details of our contributions are described below:

Bag of Surrogate Parts (BoSP)

We take the feature maps as surrogate parts and assume that the activation values represent the assignment strengths for these parts. Therefore, given the architecture, the number of the surrogate parts is inherently determined, as is the same with the number of feature maps. For each spatial unit, we calculate its assignment strengths for the surrogate parts by observing its activation values. The one-by-one processing of these spatial units can be viewed as densely sampling and assigning regions of the input image. Finally, we sum the assignment strengths for the surrogate parts and form a vector accordingly, i.e. BoSP, whose length is the same with the number of the feature maps. The framework of the proposed BoSP feature is shown in Figure 1.

Specifically, the BoSP for this image can be represented as Eq.(1):

$$BoSP = \sum_{i=1}^{n} [P_1^i, P_2^i, \cdots, P_j^i, \cdots, P_M^i] \tag{1}$$

P_j^i is the assignment strength of region i on surrogate part j.

$$P_j^i = \begin{cases} 0 & \text{if } A_j^i < mean(A^i) \\ A_j^i / max(A^i) & \text{if } A_j^i \geq mean(A^i) \end{cases} \tag{2}$$

On top of BoSP, we further propose its spatial variant, called S-BoSP, by dividing the image equally into multiple sub-regions (9 regions in 3 rows and 3 columns), and concentrating the BoSP inside each sub-region.

Scale Pooling

The BoSP/S-BoSP described above only concern the spatial units at the finest level, and handle them in a disjoint way, which means to sample and assign regions in input images with fixed size and position. However, the objects may appear in different shapes, positions and scales, the independent processing of the spatial units may capture different parts of the same object and is inferior to assign the objects of different scales, thus makes the resulting feature less discriminative. To address this problem, we propose a scale pooling technique, which improves the assignment of objects with different scales and deformations by handling regions of different sizes and positions, together with max pooling operations inside each region. The procedure of scale pooling is illustrated in Figure 2.

The scale pooling scheme can handle the image regions of different sizes and different positions, making the assignment be more relaxable. Besides, benefited from the max pooling operation, the scheme is robust to object deformation inside each coarse spatial unit. Furthermore, the introduction of scale pooling would not enlarge the feature dimension and does not affect the efficiency greatly.

Global Constrained Augmentation

It is mostly beneficial to utilize data augmentation technique. Without extra data, one common approach of data augmentation is to generate numerous sub-images from the input image, and average the sub-image features as the augmented image feature. Although this approach could extract more information from one image, it only considers individual parts of the input image, and fails to handle the input image entirely. To make a more comprehensive prediction, we add a global constraint term upon the prediction of augmented features.

The specific procedure is: given an input image, we first resize it to 224×224, and extract the global feature. This feature focuses more on the entire image, and we can achieve the global prediction based on it,

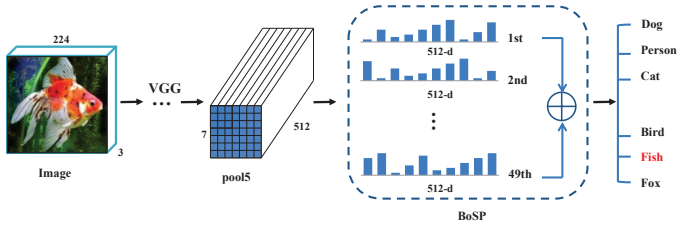

Figure 1: The framework to extract BoSP from the $pool_5$ layer of VGG. We can calculate the statistical histogram of the surrogate parts by observing the activation values.

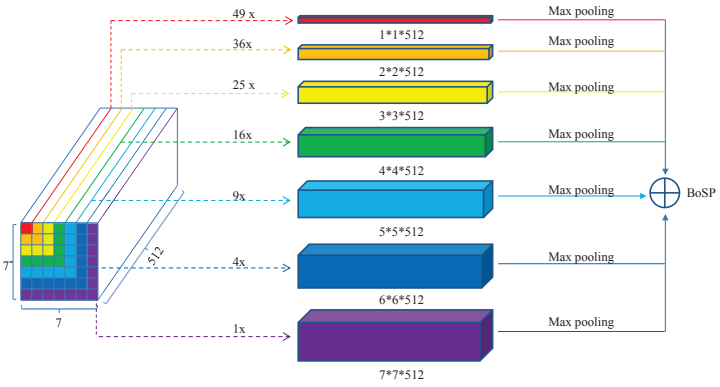

Figure 2: The illustration of scale pooling technique for BoSP (We can extract different number of features from 7 scales. For example, there are 49 red strips for the smallest scale, and only 1 purple strip for the largest scale. Then we max pooling the features inside each scale and add up to form the final feature).

denoted as Pre_{global}. Next, we resize the image to make the smallest side equal S while keeping its ratio, and crop regions of 224×224 with stride of 32 pixels. Thereby, we formulate several sub-images from the input image, each sub-image may only contain part of the original object. The image feature is the average of the sub-image features, and this feature concerns more about parts of the image, and based on it, we make the part prediction, denoted as Pre_{part}. The final prediction is the product of the global prediction and part prediction:

$$Pre_{fusion} = Pre_{global} \times Pre_{part} \tag{3}$$

As our feature is derived from the convolutional layers, the input image could be of any size, and we do not need to explicitly crop sub-images. In practice, we only need to input the resized image once to extract the augmented BoSP/S-BoSP.

Experiment

There are some valuable findings through our experiment:

1) In terms of efficiency and accuracy, it is more beneficial to derive the BoSP feature from higher layer, i.e. $pool_5$, than lower layers.

2) Scale pooling method could improve the performance of BoSP/S-BoSP without enlarging the feature dimension, and the improvement can be very large. For example, scale pooling increases the BoSP features of Caltech101 and Oxford102 by more than 3%.

3) Regardless of the differences of the global-based prediction and part-based prediction, it is always beneficial to incorporate them by utilizing the global constrained augmentation scheme.

4) Our proposed approach improves the state-of-the-art of Caltech101, Oxford102, SUN397 considerably, and achieves comparable result with previous best performance on Indoor67 dataset, while comes in lower dimension.