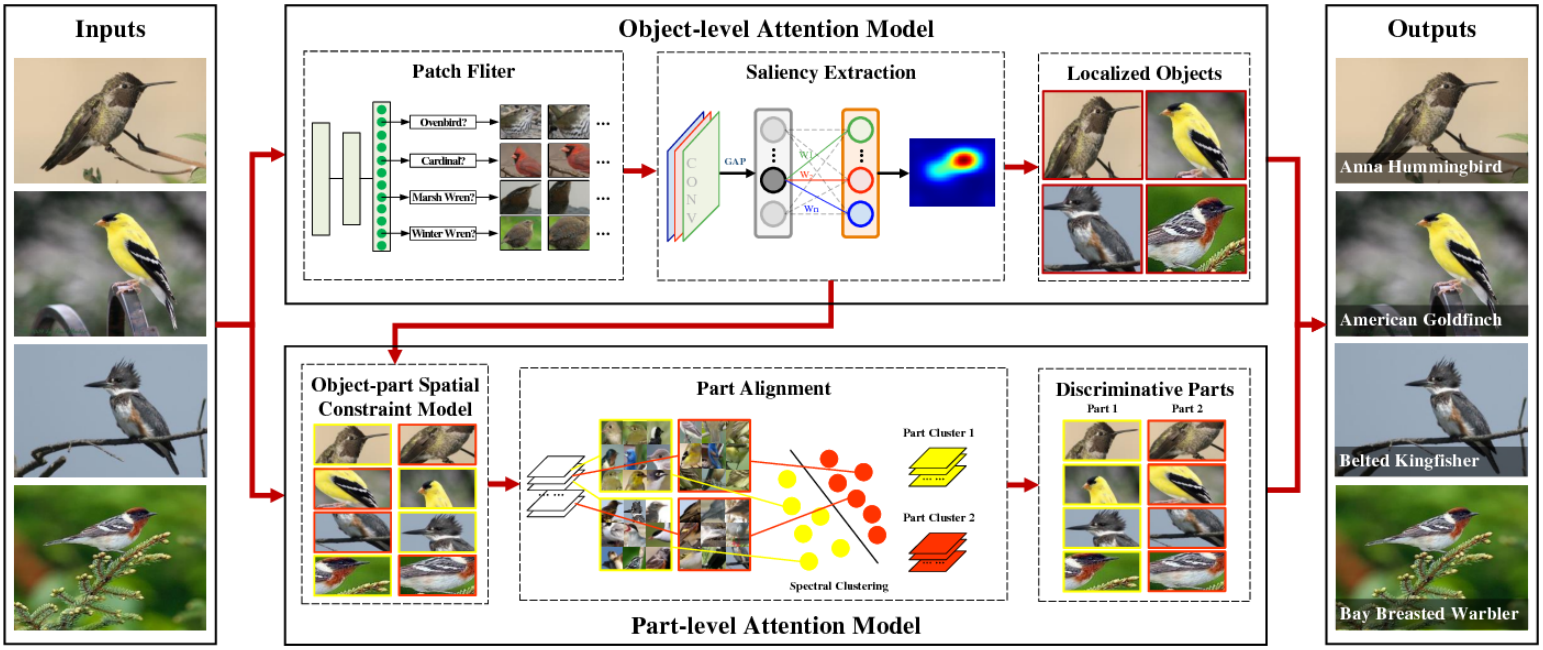
Rishabh Ramteke

November 5, 2019

Discriminative Localization for Fine-grained image recognition



Introduction & Motivation of the Problem

In this blog, I will explain about Fine-grained recognition and the approach used by Yuxin Peng [1] . It is the task of distinguishing between visually very similar objects such as identifying the species of a bird, the breed of a dog or the model of an aircraft. The apparent differences between such categories are only very subtle and can be easily overwhelmed by those caused by factors such as post, viewpoint, or location of the object in the image. In terms of visual categories, such classes have high inter-class and low inter-class variance. This makes the fine-grained recognition task extremely challenging. However, they mainly have two limitations: (1) Relying on object or parts annotations which are heavily labor consuming. (2) Ignoring the spatial relationship between the object and its parts as well as among these parts, both of which are significantly helpful for finding discriminative parts.

Distinguishing between objects in this scenario often implies focusing on details from coarser to finer levels such as a beak of a bird (as shown in figure) because this work as discriminating features for recognition. Semantic part localization can facilitate fine-grained categorization by explicitly isolating subtle differences in appearance associated with specific object parts such as the beak of the bird. Localizing the parts in an object is therefore important for establishing correspondence between object instances and discounting object post variations and camera view position changes.

Therefore, the paper by Yuxin Peng, proposes the object-part attention driven discriminative localization (**OPADDL**) approach for weakly supervised fine-grained image classification, and the main novelty is that the Object-part attention model integrates two level attentions: object-level attention localizes objects of images, and part-level attention selects discriminative parts of object. Both are jointly employed to learn multi-view and multi-scale features to enhance their mutual promotion.

Brief description:

Object-level attention selects image patches relevant to the object and part-level attention selects discriminative parts, which is the first work to classify fine-grained images without using object and parts annotations in both training and testing phase. **Object-level attention** model utilizes the global average pooling in CNN to extract the saliency map for localizing objects of images, which is to learn object features. **Part-level attention** model first selects the discriminative parts and then aligns the parts based on the cluster pattern of neural network, which is to learn the subtle and local features. The object-level attention model focuses on the representative object appearance, and the part-level attention model focuses on the distinguishing specific differences of parts among subcategories. Both of them are jointly employed.

Object spatial constraint enforces that the selected parts are located in the object region to ensure their high representativeness. **Part spatial constraint** reduces the overlap among parts and highlights the saliency of parts to eliminate the redundancy and enhance the discrimination of selected parts. Combining these two spatial constraints not only exploits the subtle and local discrimination for

promoting parts selection significantly, but also achieves a notable improvement on fine-grained image classification.

The Object-level attention model consists of two components: **patch filter** and **saliency extraction**. The first component is to filter out the noisy image patches and retain these ones relevant to the object for training a CNN called ClassNet, to learn multi-view and multi-scale features for the specific subcategory. Then the second component is to extract the saliency map via global average pooling in CNN for localizing the object of image.

Part-level attention model consists of two components: object-part spatial constraint model and part alignment. The first is to select the discriminative parts, and the second is to align the selected parts into clusters by the semantic meaning.

Mathematical background of the solution:

Patch Filter : We remove the noisy patches and select relevant patches through a CNN, called FilterNet, which is pre-trained on the ImageNet dataset, and then fine-tuned on the training data. The patch filter is performed only in the training phase and only uses image-level subcategory label.

Saliency Extraction: Given an image **I**, the activation of neuron **u** in the last convolutional layer at spatial location (x, y) is defined as $\mathbf{f_u(x, y)}$, and $\mathbf{w^c_u}$ defines the weight corresponding to subcategory **c** for neuron **u**. The saliency value at spatial location (x, y) for subcategory c is computed as follows:

$\mathbf{M_c(x, y) = \sum_u w^c_u f_u(x, y)}$ where $M_c(x, y)$ directly indicates the importance of activation at spatial location (x, y) leading to the classification of an image to subcategory c. Through object-level attention model, we localize objects in images to train a CNN called ObjectNet for obtaining the prediction of object-level attention.

Object-Part Spatial Constraint Model: Let $\mathbb{p}$ denotes all the candidate image patches and P = {p1, p2, ..., pn} denotes the n parts we selected from $\mathbb{p}$ as the discriminative parts for each given image. The object part spatial constraint model considers the combination of two spatial constraints by solving the following optimization problem:

$$P^* = arg\max_{\mathbb{P}} \Delta(P) \qquad \Delta(P) = \Delta_{box}(P)\Delta_{parts}(P)$$

Where **Δ(P)** is a scoring function over two spatial constraints

$\mathbf{\Delta_{box}(P)}$ denotes the object spatial constraint and $\mathbf{\Delta_{parts}(P)}$ denotes the part spatial constraint

$$\Delta_{box}(P) = \prod_{i=1}^{n} f_b(p_i)$$

where

$$f_b(p_i) = \begin{cases} 1 & IoU(p_i) > threshold \\ 0 & otherwise \end{cases}$$

and

$$\Delta_{parts}(P) = log(A_U - A_I - A_O) + log(Mean(M_{A_U}))$$

where A_u is the union area of n parts, A_I is the intersection area of n parts, A_o is the area outside the object region and Mean(M_{AU}) is defined as follows: $\textbf{Mean}(M_{AU}) = \frac{1}{|Au|}\sum_{i,j} M_{ij}$

Algorithm 1 Part Alignment

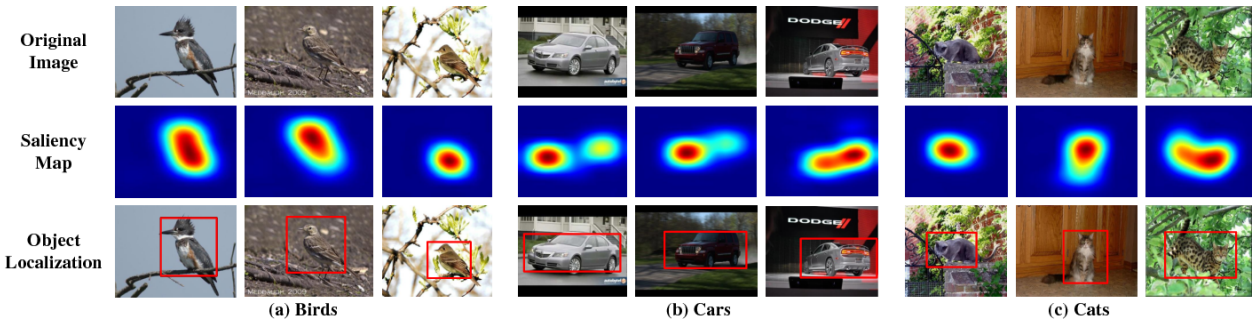
Input: The ith selected part p_i; The part clusters $L = \{l_1, l_2, ..., l_m\}$; And the number of neurons in penultimate convolutional layer d.
Output: The cluster that p_i is aligned into l_c.

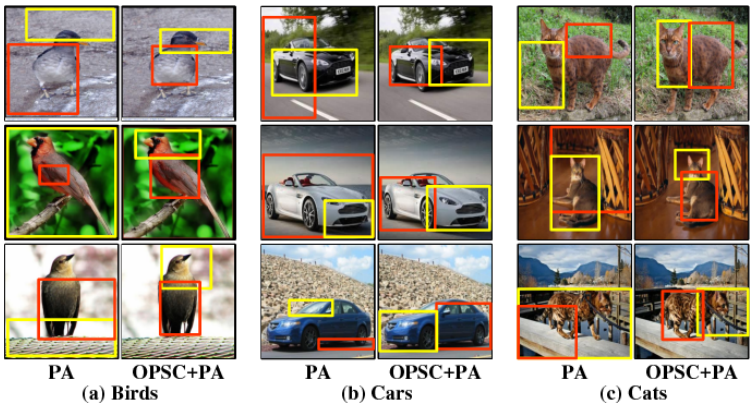
Set $score_k = 0; k = 1, ..., m$.
Warp p_i to the size of receptive field on input image of neuron in penultimate convolutional layer.
Perform a feed-forward pass to compute p_i's activations $F_i = \{f_{i1}, f_{i2}, ..., f_{id}\}$.
for $k = 1, ..., m; j = 1, ..., d$ **do**
  if jth neuron belongs to cluster l_k **then**
    $score_k = score_k + f_{ij}$.
  end if
end for
$c = \arg\max_k score_k$.
return l_c.

We follow the following algorithm for part alignment.

Visual results or tables for comparisons:

Here, we will see the visual results of this approach and compare it with other benchmark algorithms.





OPSC refers to object-part spatial constraint model, and OPSC+PA refers to combining the above two approaches, which is adopted in our OPADDL approach. The yellow and orange rectangles denote the selected discriminative parts via the two approach, which respond to the heads and bodies of objects.

Method	Train Annotation		Test Annotation		Accuracy (%)	Net
	Object	Parts	Object	Parts		
Our OPADDL Approach					**85.83**	VGGNet
FOAF [7]					84.63	VGGNet
PD [6]					84.54	VGGNet
STN [21]					84.10	GoogleNet
Bilinear-CNN [25]					84.10	VGGNet&VGG-M
Multi-grained [24]					81.70	VGGNet
NAC [20]					81.01	VGGNet
PIR [13]					79.34	VGGNet
TL Atten [15]					77.90	VGGNet
MIL [40]					77.40	VGGNet
VGG-BGLm [12]					75.90	VGGNet
Dense Graph Mining [35]					60.19	
Coarse-to-Fine [39]	√				82.50	VGGNet
Coarse-to-Fine [39]	√		√		82.90	VGGNet
PG Alignment [11]	√		√		82.80	VGGNet
VGG-BGLm [12]	√		√		80.40	VGGNet
Triplet-A (64) [41]	√		√		80.70	GoogleNet
Triplet-M (64) [41]	√		√		79.30	GoogleNet
Webly-supervised [42]	√	√			78.60	AlexNet
PN-CNN [10]	√	√			75.70	AlexNet
Part-based R-CNN [5]	√	√			73.50	AlexNet
SPDA-CNN [23]	√	√	√		85.14	VGGNet
Deep LAC [43]	√	√	√		84.10	AlexNet
SPDA-CNN [23]	√	√	√		81.01	AlexNet
PS-CNN [22]	√	√	√		76.20	AlexNet
PN-CNN [10]	√	√	√	√	85.40	AlexNet
Part-based R-CNN [5]	√	√	√	√	76.37	AlexNet
POOF [34]	√	√	√	√	73.30	
GPP [14]	√	√	√	√	66.35	

From the results, it's clear that this new approach works better compared to other benchmark algorithms.

References:

1. 'Object-Part Attention Driven Discriminative Localization for Fine-grained Image Classification' paper by Yuxin Peng
   https://pdfs.semanticscholar.org/560e/a5ae5ee59dae108a8c7adbda6acdfc5e4cf6.pdf?_ga=2.239126410.882889251.1572941678-762048108.1561077286

2. Coursera Deep Learning in Computer Vision course

   https://www.coursera.org/lecture/deep-learning-in-computer-vision/fine-grained-image-recognition-lBgu5