# The Application of Two-level Attention Models in Deep Convolutional Neural Network for Fine-grained Image Classification

Tianjun Xiao[1], Yichong Xu[2], Kuiyuan Yang[2], Jiaxing Zhang[2], Yuxin Peng[1], Zheng Zhang[3]
[1]Institute of Computer Science and Technology, Peking University. [2]Microsoft Research, Beijing. [3]New York University Shanghai.

Fine-grained classification is to recognize subordinate-level categories under some basic-level category, e.g., classifying different bird types [1], dog breeds, flower species, etc. Counter intuitively, intra-class variance can be larger than inter-class in this task. Consequently, it is technically challenging. Specifically, the difficulty of fine-grained classification comes from the fact that discriminative features are localized not just on foreground object, but more importantly on object parts(e.g. the head of a bird). Therefore, most fine-grained classification systems follow the pipeline: finding foreground object or object parts (*where*) to extract discriminative features (*what*). As fine-grained classification datasets often provide detailed annotations of bounding box and part landmarks, most methods rely on some of these annotations to achieve better accuracy [2, 3].

In this paper, we propose to apply visual attention to fine-grained classification task using deep neural network. Our pipeline integrates three types of attention: the bottom-up attention that propose candidate patches, the object-level top-down attention that selects relevant patches to a certain object, and the part-level top-down attention that localizes discriminative parts. We combine these attentions to train domain-specific deep nets, then use it to improve both the *what* and *where* aspects. Importantly, we avoid using expensive annotations like bounding box or part information from end-to-end. The weak supervision constraint makes our work easier to generalize.

The motivation of object-level attention is to learn better feature and introduce more variance by training and testing on the image patches containing features belong to the basic-level object. To achieve that, we turn a Convolutional Neural Net (CNN) pre-trained on ILSVRC2012 1K category into a *FilterNet*. *FilterNet* selects patches relevant to the basic-level category from bottom-up region proposals. The selected patches drive the training of another CNN into a domain classifier, called *DomainNet*.

The aim of part-level attention is to focus on the certain object parts to achieve detailed description and pose normalization. Our method achieves this without using additional annotations like bounding box and part landmarks Empirically, we observe clustering pattern in the internal hidden representations inside the *DomainNet*. Groups of neurons exhibit high sensitivity to discriminating parts. Thus, we choose the corresponding filters as *part-detector* to implement part-level attention.

Both levels of the attention model can achieve better accuracy compared with the one processing original image. However, their functionality and strength differ, primarily because they admit patches of different nature. Finally, we merge the prediction results of the two level attention methods to utilize the advantage of the two. Figure 1 shows the complete pipeline.

We have verified the effectiveness of the method on the subsets of ILSVRC2012 dataset and CUB200_2011 dataset. Table 1 shows part of the experiment results. Our pipeline delivered significant improvements and achieved the best accuracy under the weakest supervision condition. The performance is competitive against other methods that rely on additional annotations.

Table 1: Accuracy and Annotation Comparison between methods

| | Training | | Testing | | |
|---|---|---|---|---|---|
| Method | BBox | Part | BBox | Part | Accuracy (%) |
| Object-level | | | | | 67.6 |
| Part-level | | | | | 64.9 |
| Two-level | | | | | **69.7** |
| No attention | | | | | 58.8 |
| BBox | ✓ | | ✓ | | 68.4 |
| DPD [2] | ✓ | | ✓ | | 70.5 |
| Part RCNN [3] | ✓ | ✓ | | | 73.5 |

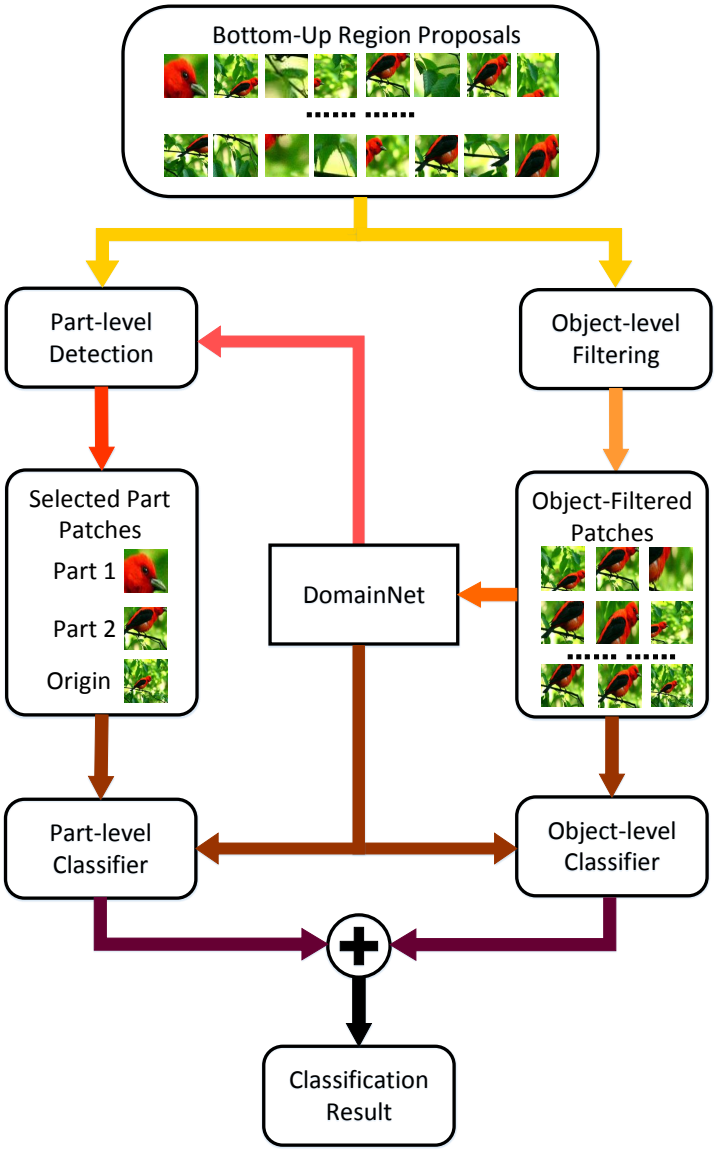


Figure 1: The complete classification pipeline. The darker the arrow is, the later this operation will be executed. Two levels of top-down attentions are applied on the bottom-up proposals. One conducts object-level filtering to select relevant patches. The other conducts detection to detect parts. DomainNet can provide the part detectors for part-level method and also the feature extractor for both of the two level classifiers. The prediction results of the two classifiers are merged in later phase to combine the advantages of the two level attentions.

[1] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-UCSD birds 200. 2010.

[2] Ning Zhang, Ryan Farrell, Forrest Iandola, and Trevor Darrell. Deformable part descriptors for fine-grained recognition and attribute prediction. In *ICCV*, 2013.

[3] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based r-cnns for fine-grained category detection. In *ECCV*. 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.