WEAKLY SUPERVISED DEEP CONVOLUTIONAL NETWORKS FOR FINE-GRAINED OBJECT RECOGNITION IN MULTISPECTRAL IMAGES

Bulut Aygüneş, Selim Aksoy

Department of Computer Engineering
Bilkent University
06800, Ankara, Turkey
{bulut.aygunes,saksoy}@bilkent.edu.tr

Ramazan Gökberk Cinbiş

Department of Computer Engineering
Middle East Technical University
06800, Ankara, Turkey
gcinbis@ceng.metu.edu.tr

ABSTRACT

The challenging task of training object detectors for fine-grained classification faces additional difficulties when there are registration errors between the image data and the ground truth. We propose a weakly supervised learning methodology for the classification of 40 types of trees by using fixed-sized multispectral images with a class label but with no exact knowledge of the object location. Our approach consists of an end-to-end trainable convolutional neural network with separate branches for learning class-specific and location-specific scoring of image regions. Comparative experiments show that the proposed method simultaneously learns to detect and classify the objects of interest with high accuracy.

Index Terms— Weakly supervised learning, object recognition, multispectral image analysis

1. INTRODUCTION

Increasing spatial resolution and richer spectral information have led to new challenges regarding the increasing detail in the appearance of objects in remotely sensed images. Fine-grained object recognition is one of such challenges, which differs from traditional object recognition and classification problems with respect to the low between-class variance among a large number of closely related categories.

In addition to the fine-grained nature of the objects of interest, the sizes of the objects can also introduce new challenges. For example, registration errors can cause an offset between the pixels in the data source and the ground truth locations. Furthermore, precise pixel-level labeling can often be difficult to obtain, especially when resolution is too low for annotating objects directly on the imagery. Thus, a shift of a few pixels can introduce a significant uncertainty to the data set, if the objects of interest themselves, such as trees, cover an area of a few pixels.

This work was supported in part by the TUBITAK Grant 116E445 and BAGEP Award of the Science Academy.

The problem that we consider is fine-grained classification of fixed-sized images cropped from multispectral (MS) scenes into one of 40 types of trees. Although the trees in our data set have different sizes, most trees can fit inside a region of 4×4 pixels at $2m$ spatial resolution. However, larger images need to be cropped to account for the aforementioned ground truth uncertainties. Such settings cause a sample to have a single image-level label without information about the exact location of the object instance inside the image.

This problem can be studied from a weakly supervised learning (WSL) perspective. In the WSL setting, every sample image has a label which indicates that there is a certain object somewhere within the image, but does not provide any location information. Similar WSL problems have been studied in the remote sensing literature. For example, Han *et al.* [1] presented a binary object detection method where positive instances for the object of interest were sampled based on saliency scores while trying to keep inter-class separability and intra-class compactness high. New positive training samples were selected using a classifier trained on the current training set, and a new classifier was trained on the updated training set in an iterative fashion. Zhang *et al.* [2] followed a similar iterative instance mining procedure by updating the negative instance set and the classifier in each iteration.

These WSL approaches differ from ours in the sense that they use iterative instance mining to improve the classifier which is then used to detect objects in a sliding window fashion. Our method, instead, focuses on directly training a model that detects and classifies the object of interest given the whole input image, by using end-to-end learning from weakly labeled training instances. Sumbul *et al.* [3] presented an alternative approach in the same problem setting by using an attention mechanism that learns a weighted combination of features extracted from fixed-sized regions obtained at each possible position in the image for classification.

Our contribution in this paper is the development of a weakly supervised deep detection network (WSDDN) model [4] for fine-grained object recognition with registration uncertainty between the MS images and the ground truth data. We

show that the proposed model achieves higher classification accuracy while being able to localize the objects of interest in test images. In the following, Section 2 summarizes the data set, Section 3 describes the methodology, Section 4 presents the experiments, and Section 5 provides the conclusions.

2. DATA SET

The data set consists of 8-band MS WorldView-2 imagery at $2m$ spatial resolution [3]. There are 48,063 images containing street trees of 40 different types, where each image is centered at a coordinate provided in the point GIS data [5]. Even though most trees fit within a 4×4 pixel window, we choose to use a neighborhood of 12×12 pixels around each ground truth location to account for the registration errors.

3. METHODOLOGY

In this section, we first outline the problem and our weakly supervised learning (WSL) approach. Then, we explain the way we realize this WSL approach as an end-to-end trainable convolutional neural network.

Problem definition. The goal is to learn an image classification model over a set of training images containing objects with unknown positions: each training image $\mathbf{x}$ is annotated with some class label $\mathbf{y}_{gt}$ denoting one of the C object classes, and, the image is presumed to contain an instance of the corresponding class but the exact position of the object instance within the image is unknown. Using these training images, we aim to learn an accurate classification function $f(\mathbf{x})$ that maps a given image $\mathbf{x}$ to one of the C classes.

For simplicity, we assume that each image is $N \times N$ pixels, and, each object instance corresponds to a smaller $W \times W$ image region. While objects naturally vary in size, such fixed-sized regions can provide an effective representation for the instances of compact objects in most cases, as long as the region size roughly corresponds to the typical object size. In our experiments, the size of encapsulating but unaligned images is $N = 12$ (Section 2), and the region size W is a hyper-parameter to be tuned, as it poses a trade-off between granularity versus correctness of localization, and also, inclusion of context versus elimination of background clutter.

Framework. The core challenge in WSL is to develop a model that can (implicitly or explicitly) localize the true sub-images both at train time and test time so that the image representation can be obtained from the actual object content rather than the background clutter. To tackle this problem, we develop a WSL approach inspired from the Weakly Supervised Deep Detection Networks (WSDDN) [4].

In the original WSDDN approach, an external algorithm is used to extract a set of *candidate regions* at each image, and, the classification score for an input image is formulated

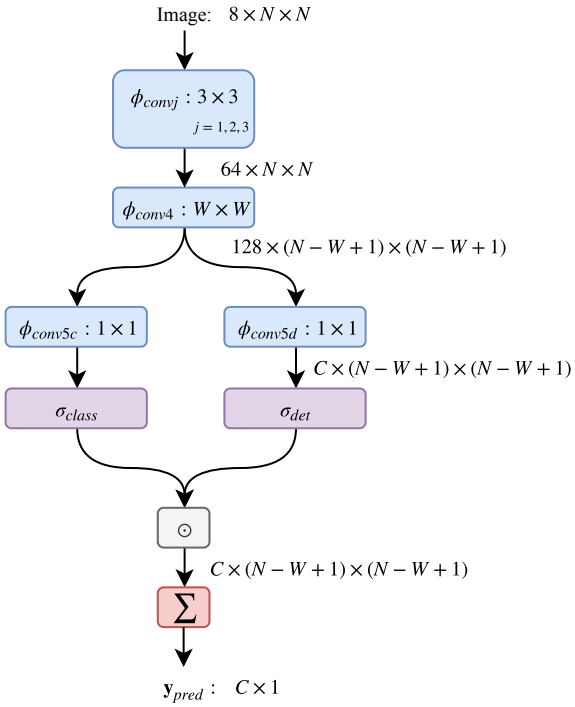

Fig. 1. Our weakly supervised fine-grained classification model. The layers ϕ_{conv1}, ϕ_{conv2}, ϕ_{conv3} use zero-padding to preserve spatial dimensions. No padding is used in ϕ_{conv4}.

in terms of the estimated *region classification* (σ_{class}) and *region detection* scores (σ_{det}). WSDDN builds upon the idea that a region should positively contribute to the image-wide score of a class only if the region is assigned to that class, and, the region is (one of) the top-scoring regions among other ones in terms of the corresponding detection scores. Following this idea, the final image-wide class scores are obtained by summing over the final per-region scores, formulated as the product of σ_{class} and σ_{det}.

We build our model on the WSDDN approach with two important differences: (i) we utilize a different deep network architecture tailored for WSL on multispectral remote sensing imagery; (ii) we side-step the external candidate region extraction algorithm, and, efficiently implement dense-sliding window candidate region extraction mechanism in terms of a convolutional layer. As a result, we do not need a region-level feature pooling mechanism either, and, are able to simply train the deep network in an end-to-end manner. The details of our WSL model are provided in the following part.

Our multispectral WSL network. Our proposed network consists of the following parts: (i) convolutional feature extraction, (ii) candidate region extraction, (iii) region scoring, and (iv) region-based image scoring. The feature extraction part is adapted from [3] and consists of three convolutional layers, ϕ_{conv1}, ϕ_{conv2}, and ϕ_{conv3}. Each layer utilizes input

zero-padding, convolves with 64 filters of spatial size 3×3, and yields a $64 \times N \times N$ feature tensor.

The resulting feature tensor is then fed into the ϕ_{conv4} layer, which convolves with 128 filters of spatial size $W \times W$, with no zero-padding. This layer effectively implements the combined candidate region generation and representation extraction steps of WSDDN in an efficient manner. More specifically, this layer corresponds to cropping candidate regions of size $W \times W$ in a sliding window fashion, and, extracting a 128-dimensional feature vector from each one by a linear transform. In total, it provides $(N - W + 1)^2$ windows.

The output of the candidate region extraction step is fed into two separate branches for region classification and region detection. We efficiently implement both branches by two parallel convolutional layers with C kernels of spatial size 1×1. Therefore, each layer corresponds to applying a C-class linear classifier to the candidate regions.

In the region classification branch, the resulting classification scores are transformed by a softmax *over the classes*:

$$[\sigma_{class}]^c_{i,j} = \frac{\exp\{[z_{class}]^c_{i,j}\}}{\sum_{k=1}^{C} \exp\{[z_{class}]^k_{i,j}\}} \quad (1)$$

where $[z_{class}]^c_{i,j}$ is class-c classification score of the region at position (i, j). Similarly, in the detection branch, the resulting scores are transformed by a softmax *over the regions*:

$$[\sigma_{det}]^c_{i,j} = \frac{\exp\{[z_{det}]^c_{i,j}\}}{\sum_{u=1}^{N-W+1} \sum_{v=1}^{N-W+1} \exp\{[z_{det}]^c_{u,v}\}} \quad (2)$$

where $[z_{det}]^c_{i,j}$ is class-c detection score of the region at (i, j).

The region classification and detection branches can be interpreted as soft-assigning regions to classes and soft-selecting the top-scoring regions within each class, respectively. The final region-based image scores are obtained by summing over the element-wise products of the per-region classification and detection scores:

$$[\mathbf{y}_{pred}]^c = \sum_{i=1}^{N-W+1} \sum_{j=1}^{N-W+1} [\sigma_{class}]^c_{i,j} \odot [\sigma_{det}]^c_{i,j}. \quad (3)$$

This can be interpreted as focusing on top-scoring regions of each class, where the regions are also consistently soft-assigned to the same class.

The resulting C scores are finally passed through softmax to obtain image-level class probability scores. The loss for a single image is defined as the cross-entropy loss between the obtained probabilities and the ground truth class label.

We note that the ReLU function is applied to the outputs of the first four convolutional layers. The overall multispectral WSL network architecture is summarized in Figure 1.

4. EXPERIMENTS

In this section, we present the experimental setup, the details of the baselines used for comparison, and our results.

Experimental setup. We conduct our experiments on two versions of the data set: the original set with 40 classes and the subset with 18 classes as in [5]. We split 60% of our data sets as the training sets, 20% as the validation sets, and 20% as the test sets. We train our models on the training sets using the Adam optimizer with learning rate 10^{-3}. We use batch normalization and dropout regularization after all convolutional layers before the classification and detection branches. Drop probability is chosen as 0.25 for ϕ_{conv1}, ϕ_{conv2}, and ϕ_{conv3}, and 0.5 for ϕ_{conv4}. ℓ_2-regularization with weight 10^{-5} is used for all parameters. Mini-batch size is chosen as 100. All of these hyper-parameters are same as in [3].

Due to significant imbalance in the data set, we oversample the minority classes in the training sets, so that each class has an equal number of instances. For unbiased evaluation in an imbalanced setting, we use normalized accuracy (i.e., the average of per-class accuracies) as the evaluation metric.

We initialize ϕ_{conv1}, ϕ_{conv2}, and ϕ_{conv3} from the basic CNN model pretrained with 12×12 inputs as explained below. We use early-stopping to terminate the training process. If normalized validation accuracy does not increase for over 200 epochs, we decrease the learning rate by a factor of 10, and continue training from the checkpoint with the highest validation accuracy. We stop the training if no increase is observed for another 200 epochs, and use the model with the highest normalized validation accuracy for testing.

Baselines. We compare our results with four other methods. The first method classifies 4×4 pixel regions centered at the point GIS data with a network of three convolutional and two fully-connected layers [3]. This corresponds to the ideal setting when there is no registration error between the image data and the point GIS reference. The network architecture consists of the same ϕ_{conv1}, ϕ_{conv2}, ϕ_{conv3} layers described in Section 3, followed by two fully-connected layers that output a C-dimensional vector of class scores. The second and third methods use the same architecture as the first one with input regions of size 7×7 and 12×12 pixels centered at the same locations, respectively. These three methods are referred as basic CNN models in Table 1.

The fourth baseline is called recurrent attention model, which learns to attend discriminative regions in input images for fine-grained object classification [6]. The model works in a multi-scale fashion, where each scale attends a more localized region inside the attended region of the previous scale. A module named attention proposal network learns where to attend, and a classification network learns to classify the attended region with a score higher than the classification score of the previous scale. To achieve this, an inter-scale ranking loss is defined between the consecutive scales, and is used for training the attention proposal network. Additionally, an intra-scale classification loss is used to train the classification networks. For comparison, we use the classification results of the second scale network of a two-scale recurrent attention model that is based on the code provided as part of [6].

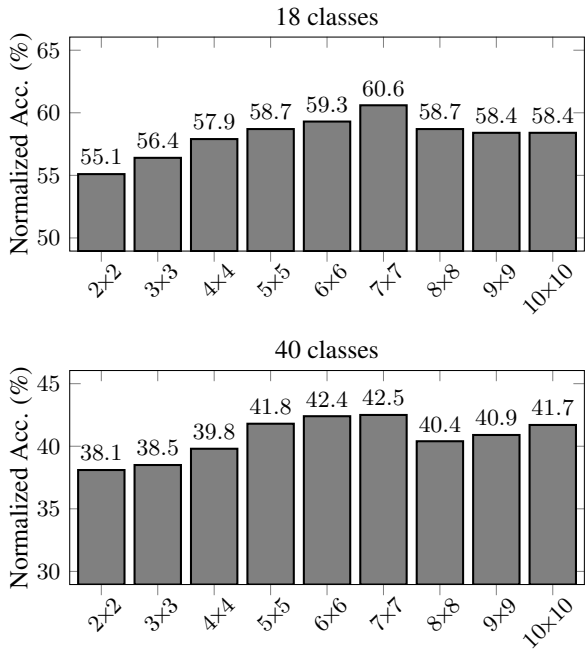

Fig. 2. Impact of region size (W) on normalized test accuracy.

Table 1. Normalized test accuracy (%) of different methods.

	18 classes	40 classes
Basic CNN model (4×4) [3]	39.0	25.1
Basic CNN model (7×7) [3]	44.6	30.6
Basic CNN model (12×12) [3]	47.7	34.6
Recurrent attention model [6]	51.6	36.6
Proposed framework (7×7)	60.6	42.5

Results. We experimented with different region sizes $W \in \{2, \ldots, 10\}$ to assess the impact of the region size on the performance of the model. The highest scores are obtained when 7×7 regions are used both in 18-class and 40-class settings as seen in Figure 2. Therefore, the rest of the results in this section are presented using the model trained with $W = 7$.

As seen in Table 1, the basic CNN performs better when trained using the whole 12×12 images. This confirms the need for using larger neighborhoods when there is uncertainty regarding the locations of the objects, but with an increasing risk of introducing additional background clutter. On the other hand, the proposed method performs better than all others for both data sets by using a WSL model that incorporates a class-specific and location-specific scoring scheme.

Qualitative evaluation of the localization capability of our approach is done by visualizing the highest scoring regions on test images as in Figure 3. The region scores are obtained from the per-region classification and detection scores in (3) by selecting the scores that correspond to the predicted class.

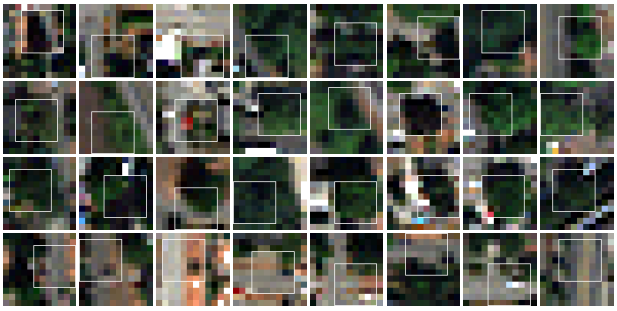

Fig. 3. Example test images from the 40-class data set with the highest scoring regions marked in white. The first two rows are correctly classified samples. The ones in the third row are misclassified although the highest scoring regions correspond to trees. The fourth row contains misclassified examples with erroneous detection outputs.

5. CONCLUSIONS

We studied the problem of fine-grained recognition of small objects, where registration errors between the image source and the ground truth introduce a significant uncertainty in the training data. We showed that it is possible to overcome this uncertainty by approaching the problem from a weakly supervised learning perspective by using a deep network that simultaneously learns to detect and classify the objects of interest with a higher performance than several baselines.

6. REFERENCES

[1] J. Han et al., "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3325–3337, 2015.

[2] F. Zhang, B. Du, L. Zhang, and M. Xu, "Weakly supervised learning based on coupled convolutional neural networks for aircraft detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 9, pp. 5553–5563, 2016.

[3] G. Sumbul, R. G. Cinbis, and S. Aksoy, "Multisource region attention network for fine-grained object recognition in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, 2019, to appear.

[4] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *IEEE CVPR*, June 2016.

[5] G. Sumbul, R. G. Cinbis, and S. Aksoy, "Fine-grained object recognition and zero-shot learning in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 770–779, 2018.

[6] J. Fu, H. Zheng, and T. Mei, "Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition," in *IEEE CVPR*, July 2017.