Object Detection in Satellite Images Based on Active Learning Utilizing Visual Explanation

Kazuki Uehara, Hirokazu Nosato, Masahiro Murakawa, and Hidenori Sakanashi
Artificial Intelligence Research Center (AIRC)
National Institute of Advanced Industrial Science and Technology (AIST)
Tsukuba, Ibaraki, Japan
{k-uehara, h.nosato, m.murakawa, h.sakanashi}@aist.go.jp

***Abstract*—Convolutional neural networks (CNNs) have attracted much attention for object detection in satellite images. However, creating an annotated dataset requires a lot of time and user workload due to which the remote sensing domain has insufficient labeled datasets for training CNNs. We exploit an active learning (AL) framework for training CNNs with a small labeled dataset. AL obtains the training dataset by asking a human user to label the samples. For efficient AL, an intelligent query strategy is essential because the performance of a CNN depends on the collected dataset. Thus, in this study, we propose a query strategy to train CNNs effectively; this is done by choosing effective samples for training both the classifier and feature extractor. The strategy selects samples according to the gap between the classifier's prediction and visual explanation, which is the class discriminative part of an image derived from the extracted feature maps. Experimental result shows that a CNN trained with samples queried by our strategy had a 95% reduction in training samples requirement while maintaining 94% detection performance compared to a CNN trained with a complete dataset. Furthermore, the proposed strategy also reduced the required training samples by 30% compared to the conventional strategy to yield the same performance.**

***Index Terms*—Active learning, Visual explanation, Annotation, Object detection, Satellite images**

I. Introduction

Automatic object detection in satellite images is essential for various applications such as change detection [1] and land use monitoring [2]. In recent years, convolutional neural networks (CNNs), which are a type of deep neural networks, have shown excellent performance in many visual tasks; thus, they are also being studied for remote sensing [3], [4]. Their high performance is because they can learn effective feature extraction according to the data. Such a deep model generally requires numerous labeled samples to train the model, i.e., both feature extractor and classifier, for yielding robust performance. However, in the remote sensing domain, there are insufficient labeled samples because annotation of samples consumes a lot of time and effort. Thus, an effective training method for CNNs that requires a small dataset is necessary.

Active learning (AL) is an excellent solution to address this issue because it can reduce user workload for creating labels by asking the user to label useful samples from an unlabeled sample pool. AL methods are attracting increasing interest in the remote sensing domain [5], [6]. Moreover, several studies that incorporate CNNs into the AL framework have emerged because combining these methods is advantageous and practical [7]–[9]. AL is an iterative training method in which there is interaction between a user and a classifier. For each iteration in AL, the user is asked to label informative samples, which is expected to be the most effective for training the classifier; the classifier is then retrained with newly added labeled samples [10].

However, conventional query strategies only suggest samples based on the classifier's prediction, which makes them inadequate for training CNNs [5], [9], [10]. As mentioned above, CNN needs to be trained not only the classifier but also the feature extractor.

Therefore, we propose a query strategy to effectively train both the feature extractor and the classifier for CNNs. The proposed strategy utilizes not only the classifier's prediction but also the activation in the extracted features to determine the priority of query samples. Specifically, the query strategy selects samples for which there is a large gap between the classifier's prediction and the activation in the extracted feature maps for the concerned class. We define the magnitude of the gap as the *degree of disagreement.* This strategy is based on the assumption that if samples have a higher degree of disagreement, training of either the classifier or the feature extractor will be insufficient for such samples. To obtain the activation in a feature, we adopt visual explanation methods [11], [12], which can highlight class discriminative pixels based on the extracted feature.

An additional advantage of using the visual explanation method is that it can add visual aids to the queried samples because the method can highlight informative pixels. In many cases, labeling uncertain samples for the classifier requires considerable concentration because such samples are ambiguous for humans as well. Therefore, highlighting pixels can provide more information to human users as visual assistance.

The main contribution of our work is two-fold. First, we propose a query strategy to increase the training efficiency by utilizing the degree of disagreement between the activation in feature representation and classifier's prediction for the concerned class. Second, we introduce a visual explanation method in the AL framework to facilitate the labeling task for the users.

978-1-7281-3140-5/19/$31.00

II. Active Learning Through Visual Explanation

This section describes our proposed query strategy and the AL framework. The proposed strategy is based on the assumption that if there is a discrepancy between visual explanation and classifier results for samples, training either the feature extractor or classifier for the samples is insufficient. For example, a sample with lower prediction probability highlighted by the visual explanation can be easily overlooked by the classifier. Therefore, querying such samples for labeling and adding them as training data can be useful for training a CNN.

To select such samples, the proposed strategy determines query priority based on the degree of disagreement between classifier's prediction (probability) and the activation in feature maps for the concerned class. In other words, the strategy assigns a higher priority to samples with a larger degree.

To calculate the degree, we adopt a gradient-weighted class activation map (Grad-CAM) [12] because it can localize pixels that have an impact on the concerned class based on the extracted feature. We summarize the activation of the Grad-CAM by a method explained in section II-B to calculate the degree.

Section II-A explains an overview of the AL framework utilizing our query strategy, and Section II-B presents the proposed query strategy in detail.

A. Active Learning

Let $D = \{x_i\}_{i=1}^{l+u}$ be a dataset consisting of a labeled set $D^L = \{x_i, y_i\}_{i=1}^{l}$ and unlabeled samples $D^U = \{x_i\}_{i=l+1}^{l+u}$. In this study, a sample x_i represents a satellite image patch, and a label y_i indicates whether the sample x_i is the target object. In the iterative process, the framework picks up q samples $Q_t = \{x_i\}_{i=1+l}^{q+l}$ from unlabeled sample pool D^U based on the degree of disagreement, explained in Section II-B; it asks the user to assign their labels and suggests the samples to the user through visual explanations. The samples labeled by the user are added to the dataset D^L and removed from the unlabeled sample pool. Algorithm 1 presents an overview of our AL framework.

Fig. 1 illustrates a golf course in a satellite image and also shows examples of an AL query. In the figure, (a) and (b) respectively show the original satellite image and ground truth of the detection target, which is a golf course in this example. Cyan rectangles in Fig. 1 (c) depicts target objects detected by the classifier. Bold green rectangles in Fig. 1 (d) show samples, which are preferentially queried by our proposed strategy.

In addition, as shown in the figure, the heat-map makes the target object easy to find. Thus, querying highlighted samples is more helpful to the users than obtaining original samples alone.

Algorithm 1 :AL with visual explanation

Input:

Initial training set $D_t^L = \{x_i, y_i\}_{i=1}^{l}$.
Unlabeled samples $D_t^U = \{x_i\}_{i=l+1}^{l+u}$.
Number of samples q queried in each iteration.
A classifier f_θ.
Initialization : $t = 1$

1: **repeat**
2: Train a classifier using current labeled dataset D_t^L.
3: **for** $(i = 1\ to\ u)$ **do**
4: Rank unlabeled sample $x_i \in D^U$ based on the degree of disagreement. (explained in Section II-B.)
5: **end for**
6: Select q samples to be labeled, and display the samples to the user with visual explanation.
7: The user assigns a label to the queried samples. $Q_t = \{x_j, y_j\}_{j=l+1}^{q+l}$
8: Add the labeled samples to the labeled set. $D_{t+1}^L = D_t^L \cup Q_t$
9: Remove the labeled samples from the unlabeled pool. $D_{t+1}^U = D_t^U \backslash S$
10: $t = t + 1$
11: **until** Stopping criterion is met.
12: **return** Trained classifier f_θ

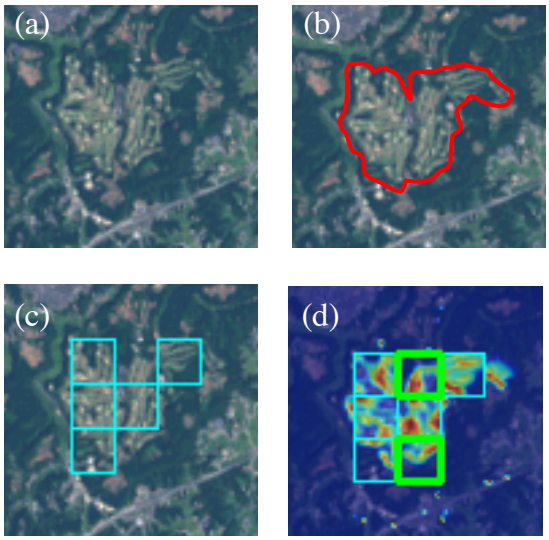

Fig. 1. Illustration of visual explanation with AL framework (an example of a golf course detection task). (a) Original satellite image of the golf course. (b) Ground truth of the golf course. (c) Candidates of the image patches to be queried. (d) Candidates with visual explanation, shown as a heat-map. Red region show important parts for detecting target objects.

B. Proposed Query Strategy

To determine the query priority, we need to calculate the degree of disagreement between classifier's prediction and the activation in a feature map. Figure 2 shows an overview of the procedure to calculate the degree of disagreement. Because the activation of a Grad-CAM is represented as an image (matrix), we summarize the activation into a scalar value. Then, samples are arranged in descending order of the degree and are queried from the top for the user.

We define the degree of disagreement $d(x)^c$ for class c of

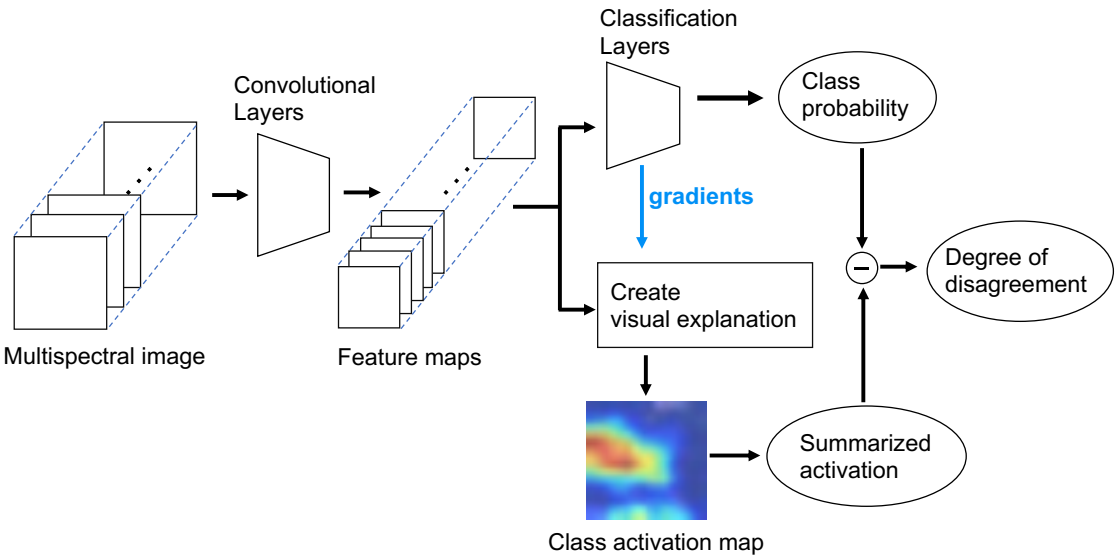

Fig. 2. Overview of the calculation procedure for the degree of disagreement between the classifier prediction and activation in an extracted feature. We transform a satellite image to feature maps through convolutional layers and then calculate the class probability through the classification layers. To obtain a class discriminative feature map, the gradients for the target class are back-propagated and combined with the feature maps. After creating the class activation map, we summarize the map and calculate the magnitude of disagreement between the class probability and summarized activation map.

sample x as follows.

$$d(x)^c = |p_\theta(y|x)^c - a(x)^c|, \tag{1}$$

where $p_\theta(y|x)^c$ denotes the probability of class c in probability vector y, predicted by the classifier for parameter θ; $a(x)^c$ is the summarized activation calculated by the Grad-CAM for class c. We calculate $a(x)^c$ by averaging the top $m\%$ of the sub-pixels of class activation map M^c, which is calculated by the following equation.

$$M^c = ReLU(\sum_k w_k^c A^k), \tag{2}$$

where $ReLU$ is a rectified linear unit function, which is a transfer function; and A^k represents the feature map of the k-th convolutional layer. The importance weights w_k^c are defined as follows:

$$w_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}. \tag{3}$$

To calculate the importance weights, we compute the gradient of the prediction score y^c (before the softmax function), which corresponds to feature maps A^k of the convolutional layer. Then, the gradients are globally averaged.

III. Experiments

To evaluate the proposed strategy, we performed an object detection task starting with a limited number of datasets. The objective of this experiment was to compare query strategies to determine which one can be effectively used for training with fewer labeled samples to reduce user workload. In this study, we chose golf courses as the target object to be detected. Detecting golf courses is a challenging task because they include many variations, e.g., variations in the course arrangement (shape), scale, and vegetation.

A. Dataset

We used images from Landsat 8 because the satellite has been accumulating data for a long time (since 1972), which makes the satellite images valuable from the perspective of practical applications. Moreover, Landsat images are freely available. Landsat images are captured across different wavelengths at a spatial resolution of 15 m (panchromatic), 30 m (ultraviolet, visible, near-infrared, and shortwave infrared), and 100 m (thermal). To exploit spectral information, we used the first seven spectral bands of the satellite images having same spatial resolution (30 m).

We conducted patch-based object detection because the size of the target object was very small relative to the size of the entire image. To construct the dataset for object detection, we divided the entire satellite images into a grid of image patches. The size of an image patch was 16 $\times$ 16 pixels, corresponding to an area of 480 $\times$ 480 m^2. Therefore, the image patches used in this experiment had bands $\times$ height $\times$ width of 7 $\times$ 16 $\times$ 16. For creating the ground truth, the image patches with more than 20% of their pixels on the annotated target object (golf course) were defined as positive samples, whereas the others were defined as negative samples.

We selected two images for the experiment, as shown in Fig 3. Image (a) was used for interactive training and labeling and (b) was used for evaluating the trained model. Image (a) included 3,418 positive image patches and 158,056 negative image patches. Image (b) included 570 positive image patches and 161,188 negative image patches.

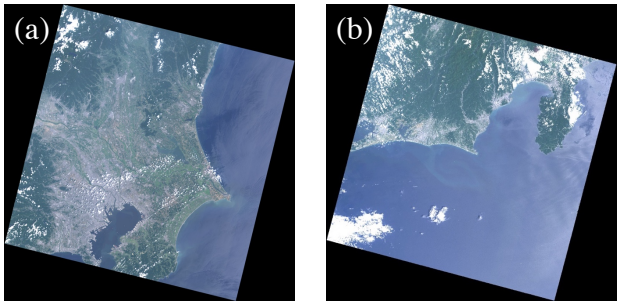

Fig. 3. Images used for training and testing. (a): LC81070352015218LGN00 (Kanto) used for collecting samples and training the model. (b): LC81080362016196LGN00 (Tokai) used for testing the model trained on the image in (a).

TABLE I
NETWORK ARCHITECTURE OF THE CNN, EMPLOYED IN AL.

layer	kernel size	output
input	-	$7 \times 16 \times 16$
convolution	3×3	$32 \times 14 \times 14$
convolution	3×3	$32 \times 12 \times 12$
convolution	3×3	$32 \times 10 \times 10$
fully connected	-	1024
fully connected	-	2

B. Network Architecture

We adopted the CNN architecture which was successfully applied for object detection in Landsat 8 images [3] and added one fully connected layer to the network to improve detection performance. Thus, the network had three convolutional layers and two fully connected layers (Table I). Each convolutional layer transformed its output through a rectified linear unit (ReLU) function.

To train the model effectively, we applied batch normalization [13] before the ReLU function. We also added a dropout layer with 50% probability after the third convolutional layer to avoid overfitting.

C. Experimental Conditions

We compared three strategies, namely, the proposed query strategy, the least confidence query strategy, and the random query strategy for evaluation. The least confidence query strategy, which is popularly used in AL, preferentially selects samples with a confidence level close to 50% [10], as assigned by the CNN. The random query strategy randomly selects samples. In this experiment, the queried samples were labeled according to the ground truth instead of the user to make the experiments fair.

First, a CNN model was trained from the labeled samples. The numbers of initial positive samples and initial negative samples were 20 and 400, respectively. In each iteration, 20 samples were selected to be labeled by using one of the above three query strategies, and the process was repeated until the number of labeled samples reached 500.

Because the dataset was highly imbalanced, we evaluated the performance by an F-measure, which is a harmonic mean of precision (correctness) and recall (completeness). The performance of each strategy was calculated as an average of 10 trials.

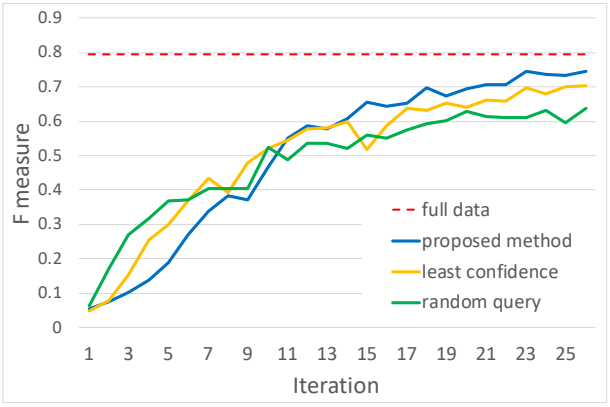

Fig. 4. Transition of detection performance of three strategies.

D. Results

Figure 4 shows the performance of the models trained by the three sampling strategies for each iteration. The solid lines show the performance of the models trained in AL, whereas the dashed red line shows the performance of the model trained on the complete dataset with 3,418 positive samples and 16,000 negative samples. Therefore, the dashed line depicts the upper limit of the detection performance for this experiment. The strategy yielding a performance such that its solid line is closest to the dashed line in fewer iterations is considered to be better than the others.

Until the fifth iteration, the random query strategy showed better performance than the other strategies. This result shows that labeling samples to include a variety of random samples is effective in the early stages of training models. However, after a certain amount of samples were accumulated (after the tenth iteration), the improvement in performance of the random query strategy was slower than that of the other strategies. After the tenth iteration, the model trained with our proposed strategy outperformed the other two strategies. This result indicates that the proposed strategy could more effectively query samples than the other two strategies for training the CNN after a certain number of samples have been collected.

As a result, the model trained by using our strategy with 500 training samples[1](approximately 5% of positive samples) achieved a performance of 94% relative to the upper limit in this experiment, whereas the least confidence strategy yielded a performance of 88%. Focusing on the number of iterations, the model trained by the proposed strategy yielded a 0.7 F-measure at the 18th iteration (labeled 360 samples). In contrast, the model trained by the least confidence strategy took 23 iterations (labeled 460 samples) to yield the same F-measure. The proposed strategy reduced approximately 30%

[1]The average number of additional data for ten trials included 143 positive samples and 357 negative samples

of labeled samples compared to the least confidence strategy. Thus, using the proposed strategy for AL can reduce user workload for labeling.

IV. Conclusion

To improve the effectiveness of AL, we proposed a query strategy that can effectively train CNN models by using a degree of disagreement between classifier's prediction and activation in an extracted feature. To calculate this degree, we introduced activation in the extracted feature as a visual explanation method. Introducing the visual explanation method into AL had two advantages: (1) allowing calculation of the degree for effective query strategy and (2) providing visual aids for humans by highlighting target objects.

We compared our strategy with conventional query strategies, i.e., the random query and least confidence query. The result showed that the proposed strategy outperforms conventional strategies after a certain number of samples are collected. Therefore, the strategy based on the degree of disagreement provides more effective samples for training a CNN.

V. Acknowledgement

The authors thank Dr. Ryosuke Nakamura from AIST for his valuable comments. This paper is based on results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

References

[1] Gang Chen, Geoffrey J. Hay, Luis M. T. Carvalho, and Michael A. Wulder, "Object-based change detection," *International Journal of Remote Sensing*, vol. 33, pp. 4434–4457, 2012, doi:10.1080/01431161.2011.648285.

[2] Laurent Durieux, Erwann Lagabrielle, and Andrew Nelson, "A method for monitoring building construction in urban sprawl areas using object-based analysis of spot 5 images and existing gis data," *Journal of Photogrammetry and Remote Sensing*, vol. 63, no. 4, pp. 399–408, 2008, doi:10.1016/j.isprsjprs.2008.01.005.

[3] Tomohiro Ishii, Edgar Simo-Serra, Satoshi Iizuka, Yoshihiko Mochizuki, Akihiro Sugimoto, Hiroshi Ishikawa, and Ryosuke Nakamura, "Detection by classification of buildings in multispectral satellite imagery," in *International Conference on Pattern Recognition (ICPR)*, 2016, doi:10.1109/ICPR.2016.7900150.

[4] Yanfei Zhong, Feng Fei, Yanfei Liu, Bei Zhao, Hongzan Jiao, and Liangpei Zhang, "Satcnn: satellite image dataset classification using agile convolutional neural networks," *Remote Sensing Letters*, vol. 8, no. 2, pp. 136–145, 2017, doi:10.1080/2150704X.2016.1235299.

[5] Devis Tuia, Michele Volpi, Loris Copa, Mikhail Kanevski, and Jordi Muñoz-Marí, "A survey of active learning algorithms for supervised remote sensing image classification," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 3, pp. 606–617, 2011, doi:10.1109/JSTSP.2011.2139193.

[6] Jun Xu and Renlong Hang, "A new committee-based active learning (cbal) approach to hyperspectral remote sensing data classification," *Remote Sensing Letters*, vol. 5, no. 6, pp. 511–520, 2014, 10.1080/2150704X.2014.928423.

[7] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin, "Cost-effective active learning for deep image classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 12, pp. 2591–2600, 2017, doi:10.1109/TCSVT.2016.2589879.

[8] Ozan Sener and Silvio Savarese, "Active learning for convolutional neural networks: A core-set approach," in *International Conference on Learning Representations*, 2018.

[9] Juan Mario Haut, Mercedes E. Paoletti, Javier Plaza, Jun Li, and Antonio Plaza, "Active learning with convolutional neural networks for hyperspectral image classification using a new bayesian approach," *IEEE Transactions on Geoscience and Remote Sensing*, vol. abs/1312.6034, 2013, doi:10.1109/TGRS.2018.2838665.

[10] Burr Settles, "Active learning literature survey," Tech. Rep. 1648, University of Wisconsin-Madison, Jan. 2010.

[11] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliba, and Antonio Torralba, "Learning deep features for discriminative localization," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, doi:10.1109/CVPR.2016.319.

[12] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *International Conference on Computer Vision (ICCV)*. IEEE, 2017, doi:10.1109/ICCV.2017.74.

[13] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, 2015, vol. PMLR37, pp. 448–456.