Environmental Modelling & Software 91 (2017) 127–134

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Automatic land cover classification of geo-tagged field photos by deep learning

Guang Xu [a,*], Xuan Zhu [a], Dongjie Fu [b], Jinwei Dong [c,d], Xiangming Xiao [d]

[a] School of Earth, Atmosphere and Environment, Monash University, Clayton Campus, Clayton, VIC 3800, Australia
[b] State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, 100101, China
[c] Key Laboratory of Land Surface Pattern and Simulation, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
[d] Department of Microbiology and Plant Biology, and Center for Spatial Analysis, University of Oklahoma, Norman, OK 73019, USA

ARTICLE INFO

Article history:
Received 4 July 2016
Received in revised form
19 January 2017
Accepted 2 February 2017

Keywords:
Deep learning
Convolutional neural network
Transfer learning
Multinomial logistic regression
Land cover
Crowdsourced photos

ABSTRACT

With more and more crowdsourcing geo-tagged field photos available online, they are becoming a potentially valuable source of information for environmental studies. However, the labelling and recognition of these photos are time-consuming. To utilise such information, a land cover type recognition model for field photos was proposed based on the deep learning technique. This model combines a pre-trained convolutional neural network (CNN) as the image feature extractor and the multinomial logistic regression model as the feature classifier. The pre-trained CNN model Inception-v3 was used in this study. The labelled field photos from the Global Geo-Referenced Field Photo Library (http://eomf.ou.edu/photos) were chosen for model training and validation. The results indicated that our recognition model achieved an acceptable accuracy (48.40% for top-1 prediction and 76.24% for top-3 prediction) of land cover classification. With accurate self-assessment of confidence, the model can be applied to classify numerous online geo-tagged field photos for environmental information extraction.

1. Introduction

Global land cover mapping is a fundamental method to monitor and evaluate global changes for environmental research and policy making. Remote sensing based classification is considered as the most efficient way for land cover mapping, but it always requires ground referencing data for training (calibration) and validation. Field survey is the general approach to acquiring the ground referencing data. During field surveys, photographs are often used to record detailed information of particular types of land cover at specific locations. Information provided by these photos can be used to help classify and validate land cover maps derived from analyses of aerial or satellite imagery. Lots of efforts have been made to archive these field photos. For example, from 1999 to 2011, the United States Geological Survey has conducted a project named “Land Cover Trends” (Gallant et al., 2004). During the project, 13,000 field photos were collected with ecoregion labels, as a nation-wide, geo-referenced dataset for land cover change mapping and as training or test site data for remote sensing image classification (Soulard and Sleeter, 2012).

However, field photo collecting by experts at a large scale is always labour-intensive and time-consuming. The crowd-sourced field photos have become a useful source employed by researchers. Since 2011, the University of Oklahoma has set up a Global Geo-Referenced Field Photo Library (Xiao et al., 2011), and also released the mobile app “Field Photo” (freely available in Google Play store and Apple Store for public use) to collect geo-referenced field photos from other researchers; and the library now contains more than 150,000 field photos (in public mode) with manually labelled land cover types. Furthermore, in 2013, the Geo-Wiki (Fritz et al., 2012) project released its mobile app “Geo-Wiki Pictures” which enables the public to share landscape photographs with detailed land cover types and other environmental information. This platform has accumulated more than 17,800 pictures so far.

These pictures can be used for validation of land cover maps at local to global scales (Fritz et al., 2012; Dong et al., 2013). However, with the aid of non-professional volunteers, crowdsourced field

* Corresponding author.
E-mail address: xg1990@gmail.com (G. Xu).

http://dx.doi.org/10.1016/j.envsoft.2017.02.004
1364-8152/

photos are sometimes misclassified with low accuracy. Average producer's accuracy of volunteers ranges from 52% to 62% (Foody et al., 2013). The experiment by Sparks et al., (2015) showed that the overall accuracy of volunteer-based Earth observation is around 70%, which is comparable to the result of GEO-Wiki. The accuracy of crowdsourced photo interpretation is becoming a bottleneck for its development.

Additionally, the volume of unlabelled online photos has been increasing at a rapid speed. In 2015, Yahoo released the YFCC (Yahoo Flickr Creative Commons) dataset (Thomee et al., 2015) containing 100 million online photos. Panoramio (http://www.panoramio.com/) from Google has also collected countless photos of the world, which remain to be utilised. However, the processing speed of public participated photo recognition is limited by the number of volunteers. Alternative efficient techniques should be developed.

With the fast development of deep learning technology, it becomes much more likely to make artificial intelligence to aid field work and help with the interpretation of ground referencing data for land cover types. In the image recognition field of deep learning, convolutional neural network (CNN) (Fukushima, 1980) is becoming the most promising algorithm, which incorporates convolutional and max-pooling layers into traditional neural networks for image feature extraction. It has already demonstrated satisfactory results for digit number recognition (LeCun et al., 1998), face detection (Garcia and Delakis, 2002; Osadchy et al., 2007; Strigl et al., 2010), pedestrian detection (Sermanet et al., 2013) and object detection (Long et al., 2015). However, these technologies have not been used specifically for the identification of land cover types. Therefore, the exploration of the state-of-the-art deep learning technology on photo recognition for land cover classification is needed to promote the automatic generation of the training and validation samples for large scale land cover mapping.

In this study, the classification model is built and tested for land cover type recognition by using the field photos from the global geo-referenced field photo library (http://eomf.ou.edu/photos), based on the CNN. Manually tagged pictures of land cover are used for model training and validation. The model performance and credibility are also assessed.

2. Methodology

2.1. Transfer learning

Training a complex neural network from scratch is always very slow on a large training set. Thus, transfer learning was proposed to apply a pre-trained neural network to another related problem (Caruana, 1995; Bengio et al., 2011; Bengio, 2012; Donahue et al., 2013). The idea of transfer learning is based on the fact that the knowledge learned from one task could be applied to solve other similar problems (Pan and Yang, 2010). In this way, the researchers could save much time for model training. A pre-trained neural network will include both its model structure and the network weights trained with large datasets. The pre-trained CNN models can always capture important features from common photos. Thus, they can be widely used for different applications.

There are mainly two kinds of strategies to take advantage of pre-trained models: feature extraction and fine-tuning. Fine-tuning means continuing training the pre-trained CNN model with another new dataset, according to the task of interest. This process will adjust the network weights of the pre-trained model to fit its outputs as close to new labels as possible, which has been proven to be effective by Yosinski et al., (2014). The benefit of fine-tuning is less time consuming because the training starts from pre-trained models. This technique has already been used for image style recognition (Karayev et al., 2013).

Unlike the fine-tune, feature extraction works by removing the last layer of a CNN model (output layer) and treating the output data of the second last layer as extracted features (also called CNN codes), which are always high dimensional vectors and implicitly represent characteristics of input images. The extracted features then can be analysed by other classifying models, such as logistic regression, multinomial logistic regression or support vector machine. In feature extraction, the pre-trained CNN model acts as the image feature extractor in the whole workflow. This framework has also shown competitive performance compared with other sophisticated models (Razavian et al., 2014).

Feature extraction is suitable when the research dataset is not similar to the original training dataset of the pre-trained CNN model in terms of sample size or sample content when it may take too much time to fine-tuning a CNN model. In this study, the Inception-v3 model was pre-trained by the ImageNet dataset (Russakovsky et al., 2015), which contains more than 10,000,000 labelled images depicting over 10,000 object categories. However, the Global Geo-referenced Field Photo Library has only nearly 30,000 training samples for landscape classification, which may cause overfitting if the CNN model is fine-tuned until the first layer. Thus feature extraction was chosen in this research.

The overall model framework for our study is shown in Fig. 1. The pre-trained CNN model Inception-v3 from Google (Szegedy et al., 2015) was chosen, because of its excellent performance on image recognition. The Inception-v3 model reached a top-5 error rate of 3.46%, which is even better than the error rate 5.1% of human (Karpathy, 2016) at the same image recognition challenge. By removing the last output layer of the pre-trained Inception-v3 model, the image feature extractor was then acquired with the output of 2048 CNN codes (image features). The CNN codes were then classified by a weighted multinomial logistic regression model for land cover type recognition.

2.2. Weighted multinomial logistic regression

In neural networks, multinomial logistic regression (Arbib, 2003) is the most widely used classification model as the last layer of a network, because it is straightforward and efficient. Multinomial logistic regression, also called softmax regression, is the generalised form of logistic regression, which can be used to model and predict probabilities that samples belong to more than two independent types. Its mathematic form is:

$$h_\theta(x) = \begin{bmatrix} P(y=1|x,\theta) \\ \dots \\ P(y=k|x,\theta) \end{bmatrix} = \frac{1}{\sum_{j=1}^{k} e^{\left(\theta^T x\right)}} \begin{bmatrix} e^{\left(\theta^{(1)T} x\right)} \\ \dots \\ e^{\left(\theta^{(k)T} x\right)} \end{bmatrix}, \tag{1}$$

where x represents an input variable, or a single sample, which is a $m \times 1$ dimensional vector, where m is the number of features of the input variable; k is the number of categories, into which the input variable to be classified; $h_\theta(x)$ represents the predicted probabilities that x belongs to each of k classes; θ is the multinomial logistic regression model parameter, an $m \times k$ matrix. In this research, the input variable x for each sample contains the CNN codes extracted by Inception-v3 model, which is a 2048 dimensional vector.

The process of model training is to find out the best model parameter θ that minimises the difference between the predicted and the actual probability that samples belong to each category. The gradient descent method is used to search for optimum parameter θ, which is an iterative algorithm that updates parameter θ step by step: