Landcover classification using texture-encoded convolutional neural networks: peeking inside the black box

MALIK KARIM
Department of Geography and Environmental Studies, Wilfrid Laurier University
malikkarim360@gmail.com

COLIN ROBERTSON
Department of Geography and Environmental Studies, Wilfrid Laurier University
crobertson@wlu.ca

ABSTRACT

Pattern recognition, object detection, and image classification are typical areas in which contemporary computer vision algorithms are being deployed. In high-resolution remotely sensed image classification problems, texture features can be crucial to the recognition of different landcover classes. Using a classical convolutional neural network (CNN) and texture-encoded CNN variant, early concatenation CNN (EC-CNN), we explore the relevance of texture-based features in landcover classification. In this paper, we demonstrate the utility of using shallow layers of a CNN for learning discriminative local texture features in very high-resolution images. We apply these models to a case study problem involving ground lichen classification in a tundra ecosystem and found that the texture EC-CNN out-performed the non-texture based classical CNN. Given that deep learning models are often perceived as "black-boxes", in order to illustrate how effective texture models represent landcover features, we extract feature maps from each model to provide a visual interpretation of the texture patterns learned by the various models. The CNN model saliency maps contain more localized patterns which are not easily interpretable, visually. The EC-CNN model on the other hand contain patterns that are more intuitive and representative of fine-grained textures. Furthermore, almost all filters in the model detected locally significant patterns in the landscape. This finding suggests the potential generalizability of texture-based CNNs and that classification errors associated with such models might be lower than that of traditional CNNs.

1. Introduction

Landcover classification is becoming increasingly vital and more sophisticated as remotely sensed data are now available at high temporal and spatial resolutions.
Landcover information is crucial for monitoring, and reporting on transitions in vegetation types across time and space (Albert and Gonz, 2017). Remotely sensed image classification represents a domain in which computer vision methods are now widely applied. Landcover type discrimination is a typical task in which machine learning algorithms are frequently deployed. The recent successes of artificial neural networks in object detection, pattern recognition and scene classification tasks have resulted in growing interest in exploring the capabilities of these models (Hinton et al., 2012). While traditional landcover mapping methods focus on spectral homogeneity of classes as the basis for discrimination, there are several cases where the spatial arrangement of features on the landscape is a key discriminant feature. Forest fragmentation, thermokarst, and vegetation patterns in arid ecosystems are all typically recognized by their texture (i.e., spacing and arrangement) rather than their spectral characteristics. Most traditional classification techniques and

classifiers are unable to achieve high accuracy in these settings due to high data dimensionality and scale dependencies (Basu et al., 2018). On recognizing this limitation, there has been increasing interest in approaching texture mapping using ANN techniques (Lloyd et al., 2004). Andrearczyk and Whelan (2016) found that a texture explicit CNN model out-performed the state-of-the art models in texture datasets. In a related study, Cimpoi and Vedaldi (2015) demonstrated CNN models capability to detect textural classes in scenes characterized by diverse texture categories .

Research in landcover classification and landscape comparison has demonstrated the potential of CNNs, especially in determining similarities in urban land use patterns (Albert and Gonz, 2017). Despite the level of improvements achieved using CNNs, the fact that higher hierarchical features are often used to perform classifications in classical CNNs raises questions pertaining to their generalizability, transferability and error rate across domains, locations, and datasets. Higher-level CNN features or activations lack geometric invariance, thus diminishing their robustness and generalization power for classification and mapping across variable scene configurations (Gong et al., 2014). Lower-layer feature maps however contain significant local information that capture texture and hence extracting and concatenating such features could build texture explicit models which are robust and less prone to classification errors. Research on the use of dense lower CNN features information in classification task is limited. This necessitates the specification of explicit texture models capable of learning and accurately representing textured landcovers.

We specify a texture-encoded CNN for feature extraction and classification of lichen for a landscape in a low-tundra ecosystem in Northwest Territories, Canada. We further compare a classical CNN with that of our texture-based model. Given that CNNs are to some degree, considered “black boxes”, we adapt existing approaches to compare activation feature maps from the CNN models (Jacobs and Goldman, 2010; Selvaraju et al., 2017). The contributions of this paper are therefore two-fold: (a) design of a simple texture-encoded CNN for landcover classification, and (b) use of computer vision techniques to visualize and gain insight into how well the models learn texture patterns.

2. Experimental methods

2.1 Texture-encoded CNN model design

CNNs consist of filter banks capable of extracting hierarchical spatial features using a weight-sharing framework (Cimpoi et al., 2015). Texture analysis and synthesis has been implemented in CNNs and proven to represent discriminative local features (Gatys et al., 2015; Ustyuzhaninov et al., 2016). Our design of a texture-based CNN is motivated by previous findings that lower-layers pool dense orderless features and capture relevant local patterns as compared to higher-layers which contain global shape information (Cimpoi et al., 2015). Our approach is also inspired by Andrearczyk and Whelan (2016) technique that derives an energy feature vector from the penultimate pooling layer and concatenates them with the first fully connected layer.

In the context of landscape or landcover classification and similarity search, global shape information present in fully connected layers is of little relevance as spatial patterns often bear no uniquely defined geometry across space and time. Our method encompasses the concatenation of multi-layer features and learning a representation of the data generating process concurrently. In the feature fusion framework, feature maps from three hierarchical layers are concatenated; these are subsequently flattened to feature vectors to construct the first fully connected (FC) layer. Here, we attempt to derive a non-existing fully connected layer FC1 via merging features from all preceding layers (i.e., pool1, pool2, and pool3). The classical CNN is also designed in parallel to enable performance comparison with the texture CNN. Figure 1 depicts the architecture of a classical CNN and a texture-encoded CNN. The architecture of both CNNs (i.e., the size of hidden layers) is adapted from the VGG model (Simonyan and Zisserman, 2014), with slight modifications in the first and second hidden layers to learn 96 features. Table 1.0 provides further information regarding the model architecture.

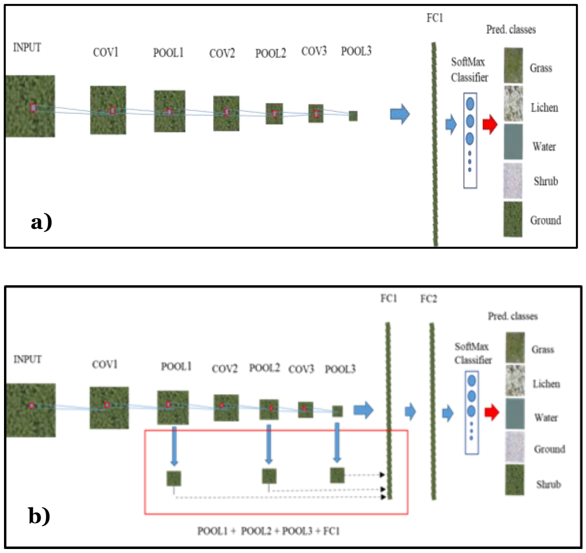

Figure 1.0 A classical CNN (a) and a texture-based EC-CNN (c)

Table 1.0 Models depth and layer structure

	cov1	cov2	cov3	FC1	FC2
ECCNN	64×5×5 pl2×2, str 2	96×5×5 pl2×2, str 2	96×5×5 pl2×2, str 2	128 ×1	128 ×1
CNN	64×5×5 pl2×2, str 2	96×5×5 pl2×2, str 2	96×5×5 pl2×2, str 2	128 ×1	

2.2 Sample preparation and model training

Imagery from an unmanned aerial vehicle (UAV) as well as georeferenced ground photographs from a study site at Daring Lake, Northwestern Territories, Canada are used in this study. Spatial resolutions of the sample imagery are 0.05 cm and 1 cm for the ground photos and UAV, respectively. To generate the training samples, high resolution ground-truth photos from the study site were tiled into 225 × 225 patches of homogenous landcover categories. We also generated 225 × 225 non-homogeneous tiles comprising all the three vegetation categories. These are reserved for probing into CNN layers to unravel what the models have learned. The tile dimensions used are significant enough to contain relevant discriminative features in the 5 × 5 receptive field of the CNN filters (Basu et al., 2015). The lichen class contains 1300 samples while the green and colored vegetation categories comprise 1000 images in each class, resulting in 3300 patches. The green vegetation class mainly comprise sedges, birch and alder shrubs and grasses, while the colored vegetation category represents primarily dwarf shrubs in the genus *Arctostaphylos* (i.e., bearberry). Figure 2.0 depicts samples of the landcover classes. The ground photos are used exclusively for model training, and classification is implemented on 450 × 450 tiles of UAV imagery. The rectified linear unit (ReLU) method is employed in the convolutional layers and fully connected layers to effect non-linear input transformation. All images are standardized to zero mean and unit variance. Image standardization is essential to mitigating the ReLU signal saturation and ensuring convergence of the gradient descent algorithm (Nair and Hinton, 2010). In training, the stochastic gradient algorithm with cross-entropy cost function is used. The learning rate, momentum and number of iterations are set to 0.01, 0.9 and 50, respectively. In order reduce overfitting, a 50% drop-out is applied to both convolutional and fully connected layers. Shutting 50% hidden units (drop-out) is an effective regularization technique for minimizing potential overfitting. The learning rate, and momentum were held unchanged. An epoch of 50 was selected after training the models for incremental iterations of 20, 30, 40, and 50. The models performance did not improve appreciably after 40 epochs, thus the selection of 50 epochs for training both models. The training, validation and test data comprise 60%, 20% and 20%, respectively. The experiment is conducted with CPU implementation using Keras + Tensorflow in Python. Training duration is 9 hours on i7-3930K CPU @ 3.20GHz, and 52.0GB windows computer.

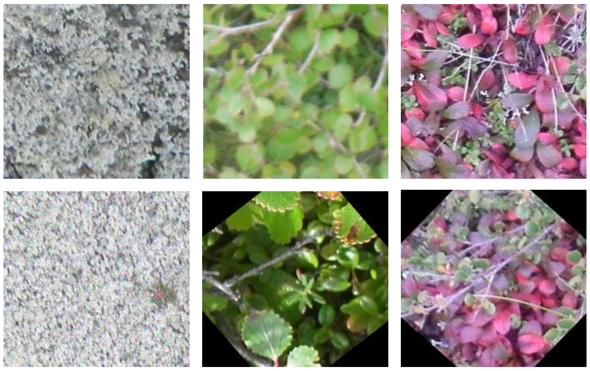

Figure 2.0 Samples of landcover types in the study site; lichen (first column), green vegetation (second column), and colored vegetation (third column)

3. Results and discussion

The model accuracy reports on the test sample are 97.50% and 96.52%, respectively, for EC-CNN and CNN models; this suggests a marginal improvement in performance of the texture-based model compared to the classical CNN, and probably accounts for the close similarity in the classification results outputted by the models. A plausible cause of this slight difference in test accuracy can be attributed to the relatively small training sample size used. With sample size of 2640 (60% of 3300) used in training and validation (20% of 3300), the traditional CNN is likely to demonstrate competitive performance. Figure 3.0 shows training/validation losses and training/ validation accuracies for both models. It can be observed that the EC-CNN losses and accuracies peak and stabilize after 25 epochs. The CNN on other hand has its losses and accuracies peaking after 40 epochs but do not exhibit clear stability. It is possible that the CNN requires additional training time to learn. However, it should be noted that CNNs are susceptible to over training, resulting in poor performance on unseen data.

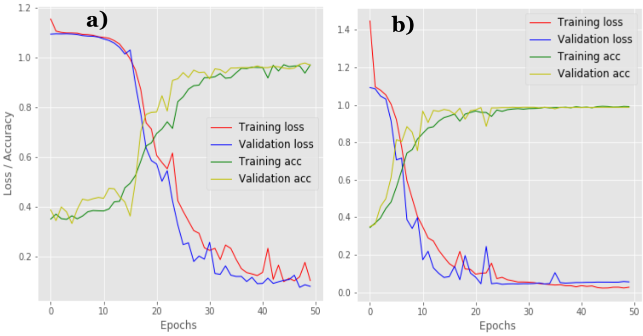

Figure 3.0 Training and validation history; classical CNN (a) and a texture-based EC-CNN (b)

Figure 4.0(b-c) illustrates tundra vegetation classification using the models. The maps appear to be difficult to distinguish visually, though close examination reveals some shift in spatial distribution of the landcover types, notably, the colored vegetation class. In order to enable direct class comparison, a difference map (Figure 4.0d) between 4.0b and 4.0c is generated. It can be noticed that the models show large discrepancy in classifying the color vegetation class, while agreeing closely in lichen and green vegetation prediction. It appears that the observed disagreement stems from shift in class allocation of pixels between the green and color vegetation types. This may be explained by the fact that the two categories contain predominantly similar textures. Hence the models are relying on spectral signatures to discriminate. It must be emphasized that ground photos were mostly samples from the similar locations as the UAV images. Though the imagery come in different spatial resolutions, training on ground photos and predicting vegetation classes on UAV mitigates the effects of potential overfitting in training data being transferred to the classification stage. This also tests as well as exposes the potential generalizability of the models on an unseen dataset.

Our approach to probing the performance of the models in the classification of the various landcover categories surprisingly reveals that the texture-based model prediction of the landcover types is visually intuitive and interpretive; in fact, visual inspection shows that the model has learned relevant feature maps about lichen composition and configuration, and that it is easier for users to link, for example, what the model thinks is lichen in the original map to the class activation maps extracted from layer three (cov3) of EC-CNN model. This implies the model predictions may be more robust and accurate despite the competing performance seen in the CNN. In Figure 5.0, class activation maps (CAM) extracted from lichen filters found in the models (cov3 layers) shed light on the patterns learned by each model to discriminate lichen. It can be observed that the EC-CNN features are denser than the CNN features which exhibit sparsity and smoothness.

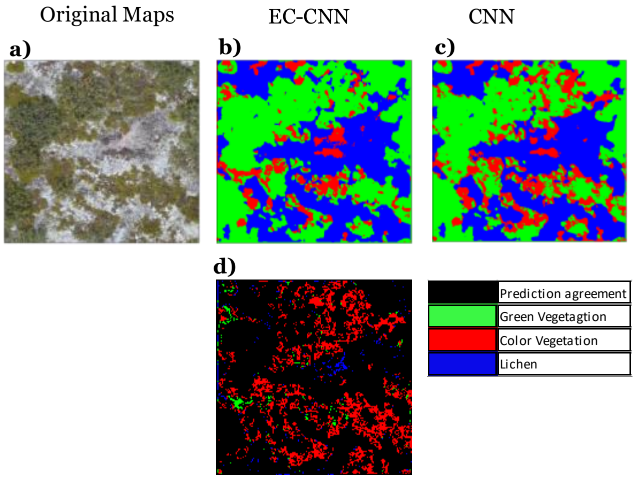

Figure 4.0. Classification of UAV map (**a**), using EC-CNN (**b**), CNN (**c**), and difference map (**d**) derived from (**a)** and (**b)**

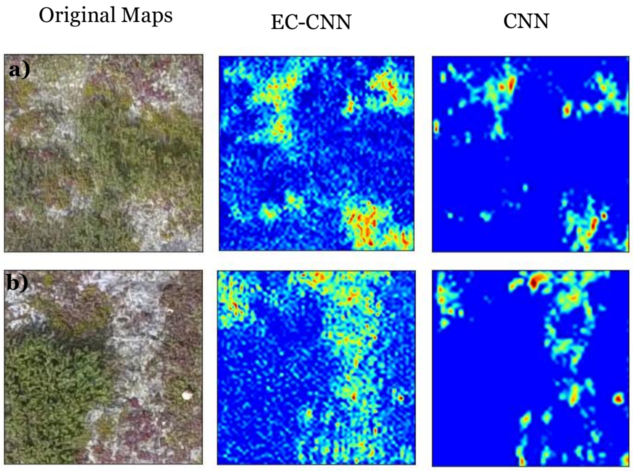

Figure 5.0. Class activation maps for Lichen. First column shows sample UAV images, second column (EC-CCN model CAM), and third column (C-CNN model CAM)

4. Conclusion

In this paper, we demonstrate the robustness of texture-based CNN in a lichen dominated tundra ecosystem classification task. Using gradient-based class activation maps as in (Selvaraju et al., 2017), we show that our texture-encoded CNN can be more effective and less error prone in classification problems as its CAMs are more intuitive and visually depict a particular landcover class the model predicts. Further research is essential to testing the accuracy of the texture model on benchmark texture datasets. It is also worth exploring the performance of different architectures of the texture model. The effectiveness of texture-based features in change detection and landscape similarity search is another research area worth investigating.

Acknowledgements

The authors acknowledge NSERC for supporting this research.

References

Albert, A., & Gonz, M. C. (2017). Using Convolutional Networks and Satellite Imagery to Identify Pa erns in Urban Environments at a Large Scale, 1357–1366.

Andrearczyk, V., & Whelan, P. F. (2016). Using filter banks in Convolutional Neural Networks for texture classification R. *Pattern Recognition Letters*, *84*, 63–69.

Basu, S., Ganguly, S., Mukhopadhyay, S., DiBiano, R., Karki, M., & Nemani, R. (2015). DeepSat-A Learning framework for Satellite Imagery.

Basu, S., Mukhopadhyay, S., Karki, M., DiBiano, R., Ganguly, S., Nemani, R., & Gayaka, S. (2018). Deep neural networks for texture classification—A theoretical analysis. *Neural Networks*, *97*, 173–182.

Cimpoi, M., Maji, S., & Vedaldi, A. (2015). Deep filter banks for texture recognition and segmentation. *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference On*, 3828–3836.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2015). Texture Synthesis Using Convolutional Neural Networks, 1–9.

Gong, Y., Wang, L., Guo, R., & Lazebnik, S. (2014). Multi-Scale Orderless Pooling of Deep Convolutional Activation Features Yunchao. *In European Conference on Computer Vision*, 392–407.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R. R., Yosinski, J., Lipson, H. (2012). Improving neural networks by preventing co-adaptation of feature detectors, 1–18.

Jacobs, D. E., & Goldman, D. B. (2010). Cosaliency : Where People Look When Comparing Images.

Lloyd, C. D., Berberoglu, S., Curran, P. J., & Atkinson, P. M. (2004). A comparison of texture measures for the per-field classification of Mediterranean land cover. *International Journal of Remote Sensing*, *25*(19), 3943–3965.

Nair, V., & Hinton, G. E. (2010). Rectified Linear Units Improve Restricted Boltzmann Machines. *Proceedings of the 27th International Conference on Machine Learning*, (3), 807–814.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Proceedings of the IEEE International Conference on Computer Vision*, *2017–Octob*, 618–626.

Ustyuzhaninov, I., Brendel, W., Gatys, L. A., & Bethge, M. (2016). Texture Synthesis Using Shallow Convolutional Networks with Random Filters, 1–9.