A COMPARATIVE ANALYSIS OF DEEP LEARNING TECHNIQUES FOR CROP TYPE RECOGNITION IN TEMPERATE AND TROPICAL REGIONS FROM MULTITEMPORAL SAR IMAGE SEQUENCES

L. E. Cué[1], J. D. Bermudez[1], P. Achanccaray[1], I. D. Sanches[3], P. N. Happ[1], R. Q. Feitosa [1,2]

[1] Pontifical Catholic University of Rio de Janeiro, Brazil
[2] Rio de Janeiro State University, Brazil
[3] National Institute for Space Research, Brazil

Comissão IV - Sensoriamento Remoto, Fotogrametria e Interpretação de Imagens

ABSTRACT

This work presents a comparative analysis of Deep Learning (DL) approaches for crop recognition from multitemporal sequences of SAR images. Convolutional Neural Networks (CNN) and Autoencoders (AE) are compared with a Random Forest (RF) classifier, all of them running on a feature space formed by *image staking*. Hand-crafted texture features were used for the RF approach. The DL approaches learned how to extract features for each pixel as part of training, whereby neighboring pixels were also considered. To evaluate the techniques under different crop dynamics, two Sentinel-1A datasets were used as testbed. The first one, from Germany, represents a comparatively simple dynamics of temperate regions, where typically there is a single harvest per year. The second dataset is from a tropical region in Brazil, which is characterized by a complex crop dynamics due to the diversity of crops with different phenology, and multiple harvests per year. In our experiments the DL approaches outperformed the RF baseline in almost all experiments, with a consistent superiority of CNN over AE.

Keywords: crop type classification, multitemporal SAR images, Deep Learning, tropical region

1- INTRODUCTION

Agricultural monitoring and crop yield estimation are important and challenging tasks to enhance food security. Remote sensing data provides a cost-effective way for agricultural monitoring and management. With the launch of more Synthetic Aperture Radar (SAR) satellites in the recent years, such as Sentinel 1-A, high spatial resolution images with low revisit time became affordable and allows to capturing changes over time related to different phenological stages of different crops.

Support Vector Machines and Random Forest classifiers have been applied for crop classification in pixel-wise approaches (Mosleh et al., 2015, Sonobe et al., 2015). However, these methods disregard the spatial context. Object-based classification, extracting quantitative attributes from segments, has also been applied (Jiao et al., 2014). Nevertheless, the usage of segments might not exploit spatial context appropriately because most segmentation procedures ignore semantic.

As the appearance of a parcel changes over time, the temporal context must also be considered for crop classification. Spatio-temporal Markov Random Fields (Liu et al., 2008) and Conditional Random Fields (Kenduiywo et. al., 2016) have successfully integrated both spatial and temporal information. These approaches achieved high accuracies, but also require expert knowledge about the problem.

Deep Learning (DL) techniques have recently gained broad interest in the remote sensing (RS) community. Such techniques contain specific supervised and unsupervised feature-learning algorithms, which learn features from labeled and non-labeled data. In fact, the state-of-the-art of RS image classification is based on DL techniques, such as Autoencoders (AEs) (Romero et. al, 2016), and Convolutional Neural Networks (CNNs) (Kussul et. al., 2017), which integrate both spatial and temporal context in unsupervised and supervised ways, respectively.

Most of crop type recognition researches have been conducted upon databases from temperate regions, where crop dynamics is comparatively simple because there is usually just a single crop per parcel during the whole season. On the other hand, crop dynamics in tropical areas is more complex due to multiple agricultural practices such as irrigation, non-tillage, crop rotation and multiple harvests per year, which make the traditional methods not suitable for the aforementioned approaches. However, most publications about crop recognition from multitemporal remote sensing images rely on datasets from temperate regions.

In this scenario, the present work presents a comparative analysis among supervised and unsupervised DL based approaches for crop classification from sequences of multitemporal SAR images. Two Sentinel-1A datasets were used, the first one from a temperate and the second one from a tropical region.

The rest of this paper is organized as follows. Section 2 introduces the basic architecture of AE and CNN. Section 3 presents the classification strategies compared in this work. Section 4 describes the datasets as well as the experimental protocol. Section 5 presents and discusses the experimental results. Section 6 closes the paper with a summary of the main conclusions and points to future directions.

2- FUNDAMENTALS

A. *Autoencoders (AEs)*

An Autoencoder is a Neural Network that uses an unsupervised learning algorithm to reproduce at its output the pattern presented at its input (Goodfellow et. al., 2016a). It has a hidden layer, whose outcome is used as a representation of the input. The network works in two phases: the encoder and the decoder. The encoder projects a d-dimensional input data onto an internal k-dimensional (k is the number of nodes in the hidden layer) representation space. The decoder projects it back to the original d-dimensional input space.

The learning process searches the parameter space for the set of values that minimizes the reconstruction error, a measure of the average discrepancy between the input and the corresponding output of the AE. Once the parameters have been learned, the encoder is used to generate for any input the corresponding internal representation, which is expected to be more discriminative than the original one.

B. *Convolutional Neural Networks (CNNs)*

CNNs (Goodfellow et. al., 2016b) are a Neural Networks in which an input image is convolved with multiple kernel matrices to extract specific features. The amount of spatial context taken into account relates to the size of the kernels. In the basic CNN architecture, after the image has been processed by a convolutional layer, a pooling layer reduces the data resolution.

During the convolution the kernels slide over the input image. Each of these kernels can be seen as a feature identifier. The pooling is a form of non-linear down-sampling that reduces the amount of data in the spatial domain, and consequently the number of model parameters and the computational load. It also helps to avoid overfitting.

The fully connected layer connects all neurons at the three dimensional output of the pooling layer to a one dimensional layer, which performs a nonlinear function. The final layer of a CNN comprises a single node for each class and an activation function (typically a *softmax*), which delivers class posterior probabilities.

3- METHODS

This section describes the three crop recognition methods evaluated in this work. We chose as baseline the standard image stacking approach (*RF-stack*), the most widely used approach for multitemporal remote sensing image analysis (Schneider, 2012). The other two methods are based on DL techniques: *Autoencoders* (*AE-stack*) and *Convolutional Neural Networks* (*CNN-stack*). An explanation of each of these methods is given in the following subsections.

A. *RF-stack*

This method consists in stacking the pixel wise features over all epochs. So, there is a single feature space for all epochs. Pixels at the same spatial coordinates share the same representation in this space. A Random Forest (RF) classifier assigns a class label to each pixel along the sequence. In our experiments the spatial context was captured by texture features.

B. *AE-stack*

In this approach, temporal and spatial context are exploited as part of the AE training. An AE is trained for each epoch separately extracting new feature representations. Similar to *RF-stack*, a classifier operates in the resulting feature space.

The descriptor x of a pixel in each image is a vector that comprises a $w \times w \times d$ patch centered at that pixel, where d is the depth of each input image. The training/inference procedure consists of the following steps: 1) select randomly M patches for each epoch in the sequence, 2) train an AE for each epoch using the corresponding set of M patches, 3) compute the representation of each pixel in each image using the encoded mapping functions learned by each AE, 4) take as descriptor the concatenation of those representations over the whole sequence and 5) as in *RF-stack*, apply a RF classifier using the aforementioned descriptors.

C. *CNN-stack*

The CNN architecture tested in this study consists of four layers: convolutional, max-pooling, fully connected and *softmax* layer. Similarly to (Kussul et. al., 2017), we train a CNN to describe a pixel location taking into account its neighborhood. In this method we use the concatenated original bands of all epochs in a given sequence. The descriptor x of a pixel in each image sequence is a dnw^2 dimensional vector that comprises the $w \times w \times d$ patch/sub-image centered at the same position in all n images in the sequence.

4- EXPERIMENTS

A. *Study areas and data*

Two datasets were used in our experiments.

The *Hannover* dataset covers the surroundings of Hannover city, in Northern Germany, an extent of 1728 km^2. It consists of a sequence of 45 dual polarized (VH and VV) Sentinel-1A images captured from October 2014 to September 2015, with three to five images per month. Crops found in this area are barley, rye, wheat, canola, grassland, maize, potato and sugar beets. These crops go through different phenological stages within a season (Figure 1). Typical of temperate regions, in this dataset each parcel belongs to the same class over the whole season.

Campo Verde is a municipality in the state of Mato Grosso, Brazil, with an extension of 4782 km^2. The dataset consists of a sequence of 14 dual polarized (VH and VV) Sentinel-1A images captured from October 2015 to July 2016, with one or two images per month. The main crops found in this area are soybean, maize and cotton. Some minor crops such as beans and sorghum are also present. We joined in the class *non-commercial crops* (NCC), millet, brachiaria and crotalaria. Other classes present in the dataset are pasture, eucalyptus, uncultivated soil (e.g. bare soil, soil with weeds, soil with crop residues), turf grass and cerrado (Brazilian savanna). Figure 2 shows the class occurrences per month in the dataset.

B. *Feature Extraction*

To test the *RF-stack* method we used hand-crafted features. As in (Kenduiywo, 2016), four features were computed for both image bands from the GLCM matrix (correlation, homogeneity, mean and variance) in four directions (0, 45, 90 and 135 degrees) using a 3×3 window. Then, each pixel was represented by a feature vector of dimensionality 32.

Patches from the original bands of each epoch were selected as input features for DL techniques, each patch with sizes of 3×3×2 for *AE-stack*, and 5×5×2 and 7×7×2 for *CNN-stack* for *Hannover* and *Campo Verde*, respectively.

C. *Experimental Protocol*

In all experiments we only classified the latest image of each data set. Sequences of different lengths were considered by adding earlier images consecutively. In the experiments on *Hannover* dataset pixels within a parcel carried the same crop label along all sequence. For *Campo Verde,* up to two crops may come about in a parcel. We also considered the class uncultivated soil. Recall that in all experiments only the final image of the dataset was classified.

The proportion of reference data for training and testing for both datasets was about 20% and 80%, respectively. In order to balance the number of training samples for all classes we replicated samples of less abundant classes in *Campo Verde*. Since *Hannover* has comparatively few samples, we applied a data augmentation procedure. We selected 7,000 and 50,000 training samples per class for *Hannover* and *Campo Verde* datasets respectively.

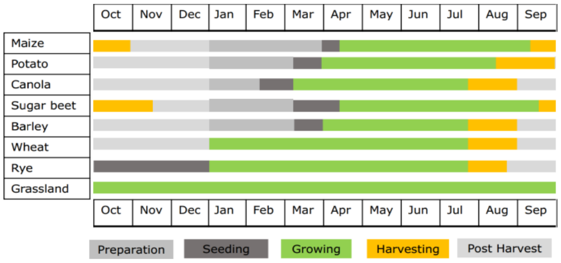

Fig. 1 – *Hannover*'s crop phenology stages. Adapted from (Kenduiywo et. al., 2016).

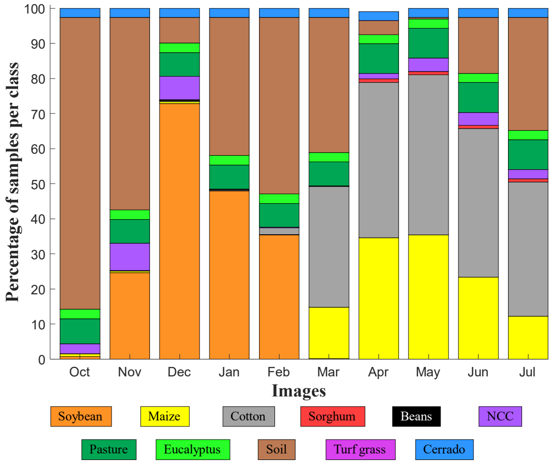

Fig. 2 – Campo Verde's class occurrences per image.

5- RESULTS

Results for *Hannover* and *Campo Verde* datasets are shown in Figure 3 and Figure 4, respectively. The Figures summarize the results in terms of accumulated *F1-score* (bars) values per class and *Overall Accuracy* (*OA*) (curves) obtained for each method described in Section 3. Each group of bars corresponds to a sequence length. Within a group, the bar represents from left to right, the results for *RF-stack, AE-stack and CNN-stack.* Notice that the maximum possible value of the accumulated *F1-score* is equal to the number of classes × 100%, i.e., 800% for *Hannover* and 900% for *Campo Verde.*

For both datasets the accumulated *F1-score* and *OA* improved as more images were added to the sequence, but only until a certain length. From then on the performance stabilized and no significant improvement could be obtained by lengthen the sequence. This is consistent with the intuition that data from an increasingly distant epoch contribute less and less to discriminate the crop class in the current epoch.

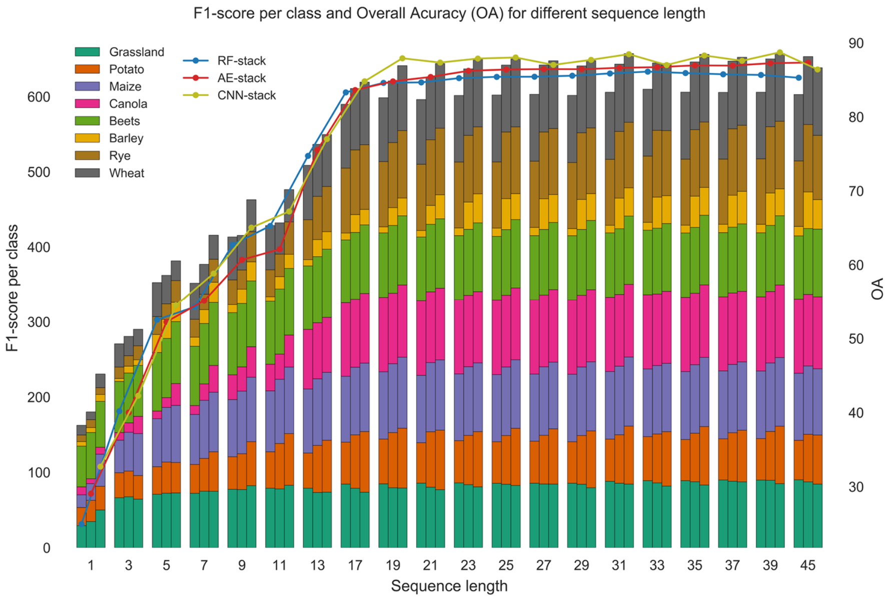

Fig. 3 – Accumulated F1-score per class (stacked bar plot) and Overall Accuracy (line plot) for different sequence lengths for *Hannover* dataset. From left to right in each bar group: *RF-stack*, *AE-stack* and *CNN-stack*.

For the *Hannover* dataset (Figure 3) we achieved improvements on *OA* from 25% to 86%, from 29% to 87.3% and from 32% to 88.7% for *RF-stack, AE-stack* and *CNN-stack,* respectively, as more images were considered. In terms of *OA* all three methods delivered similar results, being *CNN-stack* consistently the best performing approach. *RF-stack* and *AE-stack* alternated as second/third best method for different sequence lengths. For the first 19 epochs (from September to February) the *OA* had a considerable improvement as more images were added to the sequence. For sequences between 19 and 45 images the performance remained almost constant for all methods. This behavior can be explained by the crop phenology stages shown in Figure 1. From February to September there was more diversity among crops. Before February some crops are undistinguishable, as they are in the *post-harvest* and *preparation* stage and cannot be easily discriminated upon SAR data.

The accumulated *F1-scores* reveal how each method performs on a class by class basis. Major improvements were achieved by DL techniques for increasing sequence lengths, being more evident for summer barley and potato.

The results for *Campo Verde* (Figure 4) showed improvements on *OA* from 18.6% to 65.7%, from 43% to 66% and from 43.8% to 69% for *RF-stack, AE-stack* and *CNN-stack* respectively. Similar to *Hannover* for sequence up to 7 epochs (between March and July) the *OA* improved consistently for increasing sequence lengths. For sequences comprising 8 to 14 epochs the performance remained almost constant for all methods. This behavior can be explained by considering Figure 2. Prior to March, soybean is the dominant crop, which is replaced in March by maize and cotton. So, the information before March is less correlated to the crops prevailing in the end of the sequence. In terms of both accumulated *F1-score* and *OA* the DL based techniques improved considerably for increasing sequence length. Noticeably, *CNN-stack* outperformed all other methods for all sequence lengths, with a gain of up to 16% in *F1-score* for sorghum, non-commercial crops, pasture, eucalyptus and cerrado.

Finally, the best *OA* for *Campo Verde* was about 69%, whereas for *Hannover* it came close to 89%. We claim that the superior results on the *Hannover* dataset are due to the comparatively simple crop dynamics.

6- CONCLUSION

In this work, we reported the results of a comparative analysis of supervised and unsupervised Deep Learning based techniques for crop type recognition in a tropical and in a temperate region on sequences of multitemporal Sentinel-1A images. The results confirmed that the accuracy of the recognition improves as multitemporal data are used. However, the gain tends to vanish as data from increasingly remote epochs are added to the sequence.

Deep Learning techniques outperformed the standard Random Forest approach in almost all experiments. By and large, the CNN approach was the best performing among all evaluated methods, mainly for the tropical dataset.

It is further worth noticing that the accuracies were better for the dataset from a temperate than from a

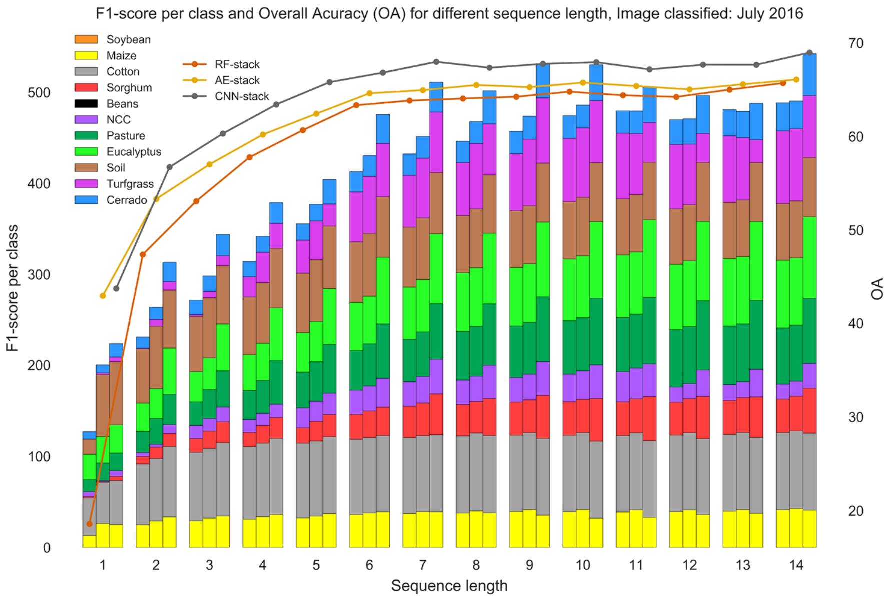

Fig. 4 – Accumulated F1-score per class (stacked bar plot) and Overall Accuracy (line plot) for different sequence lengths for *Campo Verde* dataset. From left to right in each bar group: *RF-stack*, *AE-stack* and *CNN-stack*.

tropical region. Even though there were more available training samples for *Campo Verde* than for *Hannover,* the classification performance was remarkably higher for the *Hannover.* This is most probably due to the comparatively more complex crop dynamics. Future works will focus on incorporating prior knowledge about crop phenology into the DL based classification models.

ACKNOWLEDGEMENTS

The authors acknowledge the support provided by CNPq and CAPES, and to Damian Bargiel from Technische Universität Darmstadt to provide the reference to the Hannover dataset.

REFERENCES

Mosleh, M. K.; Q. K. Hassan and E. H. Chowdhury, 2015. Application of remote sensors in mapping rice area and forecasting its production: A review. Sensors, Vol. 15, Nº 1, pp.769-791.

Sonobe, R.; H. Tani; X. Wang; N. Kobayashi and H. Shimamura, 2015. Discrimination of crop types with terrasar-x-derived information. Physics and Chemistry of the Earth, Parts A/B/C, Vol. 83-84, pp.2-13.

Jiao, X.; J. M. Kovacs; J. Shang; H. McNairn, D. Walters; B. Ma and X. Geng, 2014. Object-oriented crop mapping and monitoring using multitemporal polarimetric radarsat-2 data. ISPRS Journal of Photogrammetry and Remote Sensing, Vol. 96, pp.38-46.

Liu, D.; K. Song; J. R. Townshend and P. Gong, 2008. Using local transition probability models in markov random fields for forest change detection. Remote Sensing of Environment, Vol. 112, Nº 5, pp.2222-2231.

Romero, A.; C. Gatta and G. Camps-Valls, 2016. Unsupervised deep feature extraction for remote sensing image classification. IEEE Transactions on Geoscience and Remote Sensing, Vol. 54, Nº 3, pp.1349-1362.

Kussul, N.; M. Lavreniuk; S. Skakun and A. Shelestov, 2017. Deep learning classification of land cover and crop types using remote sensing data. IEEE Geoscience and Remote Sensing Letters, Vol. 14, Nº 5, pp.778-782.

Goodfellow, I.; Y. Bengio and A. Courville, 2016a. Autoencoders, in Deep Learning. MIT press, pp.499-523.

Goodfellow, I.; Y. Bengio and A. Courville, 2016b. Convolutional Networks, in Deep Learning. MIT press, pp.326-366.

Schneider, A., 2012. Monitoring land cover change in urban and peri-urban areas using dense time stacks of landsat satellite data and a data mining approach. Remote Sensing of Environment, Vol. 124, pp.689-704.

Kenduiywo, B. K.; D. Bargiel1 and U. Soergel, 2016. Crop type mapping from a sequence of terrasar-x images with dynamic conditional random fields. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. III-7, pp.59-66.