End-to-End CNN-CRFs for Multi-date Crop Classification **Using Multitemporal Remote Sensing Image Sequences**

Laura E. Cué La Rosa¹, Dário A. Borges Oliveira² and Raul Queiroz Feitosa¹

¹Pontifical Catholic University of Rio de Janeiro (PUC-Rio), Brazil ²IBM Research, Brazil

Abstract

Accurate crop type identification from remote sensing data in tropical regions is a challenging task. The more favorable weather conditions permit more flexible agricultural practices, which creates complex crop dynamics. Synthetic Aperture Radar (SAR) systems provide high-resolution images independent of daylight, cloud coverage, and climate conditions, turning into a cost-effective tool for crop mapping in tropical regions. Conditional random fields (CRFs) models have been used with great success to exploit spatial and temporal contexts in crop type classification. Even if such approaches deliver high accuracy, they often rely on manually designed features requiring domain-specific knowledge. In this context, deep learning methods such as convolutional neural networks (CNNs) proved to be an appealing alternative for feature learning in the context of remote sensing image classification. This work introduces an end-to-end model that combines CNNs and CRFs for crop recognition in areas characterized by complex spatio-temporal dynamics, typical of tropical regions. The proposed framework consists of two modules: the first implements a CNN that models spatial and temporal contexts from the input data, and the second implements a CRF module that models temporal dynamics considering label dependencies. Experiments are presented for an agricultural region in Brazil using multi-temporal SAR images sequences. The experiments showed significant improvements in the F1 score against a baseline model that does not include temporal dependencies in the learning process.

Keywords

Crop mapping, SAR images sequences, Convolutional neural networks, Conditional random fields.

1. Introduction

The expected increase in human population from 7.7 billion in 2019 to 8.5 billion in 2030 [1], coupled with the predicted worldwide growth of per capita income, pressures the demand for food in the future [2, 3] and reinforces the need for efficient and sustainable agricultural policies that ensure food security. In this context, accurate estimation of crop area extents is indispensable for farm monitoring and yield prediction. The Synthetic Aperture Radar (SAR) is a crucial cost-effective tool that generates high-resolution imagery independent from daylight, cloud coverage, and climate conditions.

In temperate regions, agricultural practices are characterized by a single crop type per parcel during the productive season, simplifying crop dynamics analysis. Conversely, in tropical regions, crop dynamics are more complex due to the multiple harvests per year, crop rotation, and other agricultural practices [4]. The diverse crop dynamics observed in tropical regions often require multitemporal approaches for efficient crop discrimination.

Probabilistic Graphical Models, such as Hidden Markov Models (HMM) and Conditional Random Fields (CRFs), have been successfully used to exploit both spatial and temporal contexts in crop type classification [5, 6, 7]. These approaches deliver high accuracies but often rely on feature engineering and require expert knowledge.

Alternatively, convolutional neural networks (CNNs) proved to be a robust option for remote sensing image classification, as they can learn optimal features directly from raw data. In [8], the authors used CNNs for multi-date crop recognition considering a SAR multi-temporal image sequence in a tropical region and applied a post-processing algorithm based on an HMM. The so-called Most-Likely Class Sequence (MLCS) enforces prior knowledge about crop occurrence over time in the target region. 3D CNN have been also employed for crop mapping from multi-temporal remote sensing images considering the temporal and spatial context [9]. In [10], the authors proposed a bidirectional convolutional recurrent neural network (ConvLSTM) that takes into account the spatio-temporal context and delivers a prediction for each date of interest. The authors also applied the MLCS algorithm that further improved the per-date accuracy by obtaining the most-likely class sequences based on temporal constraints. Besides achieving competitive performance, the ConvLSTM classifier does not consider these

CDCEO 2021: 1st Workshop on Complex Data Challenges in Earth Observation, November 1, 2021, Virtual Event, QLD, Australia. ☆ lauracue@aluno.puc-rio.br (L. E. C. L. Rosa); dariobo@br.ibm.com (D. A. B. Oliveira); raul@ele.puc-rio.br

⁽R.O. Feitosa) © 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CORDING) Commons License Attribution 4.0 International (CC BY 4.0).



Figure 1: General framework of the proposed CNN-CRF method. The network is trained end-to-end using the CRF loss function.

constraints during training.

Recently, CNNs and RNNs have been combined with CRFs in an end-to-end framework to model the intra-class temporal progression. These models were first proposed for sequence tagging for natural language processing and then extended for video action segmentation [11, 12]. However, they are yet to be explored for crop mapping from multi-temporal remote sensing images sequences to the best of our knowledge. This work seizes that opportunity and proposes a novel method that combines deep learning and graphical models in an end-to-end framework from crop mapping in tropical regions from multi-temporal image sequences.

2. METHOD

Our work presents a method for making dynamic decisions for multi-date crop recognition combining linear-chain CRFs and CNNs in an end-to-end formulation. The proposed framework named *CNN-CRF* is presented in Figure 1 and consists of three modules: a CNN module, a linear-chain CRF module, and Viterbi decoding. It takes as inputs a multi-temporal remote sensing image sequence $\mathbf{x} \in R^{T \times H \times W \times K}$ where each image at each time-step covers the same region; *H*, and *W* represent the height and width, respectively; and *T* and *K* are the length of the sequence and the number of

bands, respectively.

2.1. 3D CNN module

The sequence of images is fed to the CNN module that learns spatio-temporal features from the input sequence and produces a score for each label at each time step. We denote the output of the CNN at each time-step as $U(\mathbf{x}, \mathbf{y}_t; \omega)$, where ω is the network parameters. These are the emission scores (also known as unaries) that serve as input to the CRF module and are related to the input sequence's posterior probability at time-step *t*. The CNN module can consist of any CNN capable of modeling spatial and temporal context, such as 3D CNN and Convolutional LSTM networks that have been used with success for crop mapping [9, 10].

2.2. CRF module

Using the unaries from the CNN module, we applied a sequential CRF to decode labels for the whole multi-temporal input sequence jointly. The linear-chain CRF uses features that depend on the input, i.e. the predicted unaries, and on a transition matrix that models the temporal relation among pairs of adjacent labels. We define the transition matrix **Tr** with size $(T - 1 \times |S| \times |S|)$, where each **Tr**_t is the transition score matrix considering the two adjacent epochs t - 1

and t. For each \mathbf{Tr}_t , elements in row *i* and column *j* relates to the probability of the input sequence being the *i*-th crop type at time-step t, considering the *j*-th crop type in the previous time-step t - 1. The CRF probabilistic model defines the family of conditional probability $p(\mathbf{y}|\mathbf{x})$ as

$$p(\mathbf{y}|\mathbf{x}) = \frac{\exp\left\{\sum_{t} U(\mathbf{x}, y_t; \omega) + \sum_{t} \mathbf{Tr}[t, y_t, y_{t-1}]\right\}}{Z(\mathbf{x})}$$

where $Z(\mathbf{x})$ is the partition function (a normalization constant) that ensures the normalization between 0 and 1. We proposed to train the CNN and CRFs modules end-to-end by minimizing the negative log-likelihood (NLL), which is equivalent to maximizing our data's log-likelihood. Given a set of training samples pair $\mathbf{A} = \{\mathbf{x}^i, \mathbf{y}^i\}_{i=1}^A : \mathbf{y} \in \mathcal{Y}$, where \mathcal{Y} is the set of all label sequences, the NLL loss function reads as:

$$\mathcal{L}(\omega, \mathbf{Tr}) = -\sum_{i} \log p(\mathbf{y}^{i} | \mathbf{x}^{i}).$$

2.3. Data-driven transition scores

We explored two variants for learning the transitions scores automatically from the training data. We first tested a variant that has been widely used in natural language processing [11], that we called $CNN-CRF_G$ hereafter. $CNN-CRF_G$ assumes that the transition matrix is shared over time, i.e. a global matrix whose transitions scores are independent of the time-step. Hence, the $CNN-CRF_G$ model considers $\mathbf{Tr}[1] = \mathbf{Tr}[2] = ... = \mathbf{Tr}[T]$, and the transition scores are estimated directly from training data.

Despite the success of similar models for sequence tagging in natural language problems, learning a global transition matrix can be too restrictive for crop phenology changes, principally in tropical regions where crop transitions are subject to agricultural practices that may vary according to the productive season. In this sense, we propose a second variant called $CNN-CRF_A$, that learns a transition matrix for each pair of *adjacent* epochs conditioned to the observed transitions in the training set.

2.4. Viterbi decoding

During inference, the objective is finding the most likely sequence given an unseen input sequence, i.e., the maximum scoring sequence according to our model. That corresponds to solving the following equation: $\hat{\mathbf{y}} = \arg \max_{\mathbf{y} \in \mathcal{Y}} p(\mathbf{y}|\mathbf{x})$. The naive solution would require computing the probability for all possible sequences, which is computationally infeasible. Instead, we can solve the above equation more efficiently using the Viterbi algorithm [13].



Figure 2: Percentage of samples per class in each epoch of the Campo Verde dataset.

3. Experiments

3.1. Study area and Data

We evaluate the proposed models in Campo Verde, an agricultural region located in Brazil. Campo Verde [4] is a ublic dataset that provides the land-use classes by month (between October 2015 and July 2016) and 14 pre-processed SAR Sentinel-1 with VV and VH polarization. The major crops found in the region are soybean, maize and cotton. Others minor crops are beans and sorghum. The rest of the land-use classes are composed by Non-Commercial Crops (NCC), pasture, eucalyptus, turfgrass, cerrado and soil. The agricultural practices in the region consist of two seeding periods for the major crops, soybeans span from October to February, and maize and cotton from Mars to July. Figure 2 shows how the area is distributed among different crops along the months which allow to observed the complex crops' dynamics characteristic of the region. The reference data consisted of 608 crop fields split into two disjoint sets, 50% of each class selected for training and the other 50% for testing, using stratified random sampling.

3.2. Experimental Protocol

The CNN-CRF model takes as input an image patch and computes the class score for the central pixel of the patch. From the SAR images sequences, in each epoch, we randomly cropped 50,000 image patches of $14 \times 15 \times 15 \times 2$ on the fly to train the network. To alleviate the class imbalance problem, we replicated samples from minority classes and employed rotations and horizontal and vertical flips during training.

Our work primarily focused on how the inclusion of a



Figure 3: Metric performance for each one of the variants and the baseline model.

CRF module affects the classification performance; therefore, we do not evaluate various architectures for the CNN module and experiment with a simple 3D-CNN only. As fully-connected layers often contain large numbers of parameters, we only employed convolutional layers. The network consists of three processing blocks that comprehend two successive 3D convolutions + batch normalization + LeakyReLU operations, followed by an average pooling that reduces the spatial dimension by 2. The 3D convolutions consider a spatial context of 3×3 and a temporal context of 5 epochs. Our dataset has several images per epoch/month, but we are interested in a unique prediction for each month. Hence we implemented the last processing block to map the temporal dimension to the desired length. Finally, a last convolutional layer with linear activation with |S| kernels of size $1 \times 1 \times 1$ delivers the emission scores. The second term of the CRF model is implemented by defining the transition matrix of size $(T-1) \times |S| \times |S|$. That matrix is estimated during training from the training data. At inference time, the high dimensional remote sensing input images were cropped in densely overlapping images patches and the final crop map was constructed by predicting each patch associated with each image coordinate. As a baseline, we trained a 3D-CNN that delivers per epoch class probability, called CNN hereafter. The architecture is the same as the CNN-CRF variants without the transition matrix. We trained the model using a per-date categorical cross-entropy loss function. Note that the classification decision at each time-step is conditionally independent of its neighbors in such schema. We trained all models with Adam optimizer for 50 epochs with 64 image patches per batch.

3.3. Accuracy Assessment

The performance of the evaluated methods was expressed in terms of Overall Accuracy (OA), producer's accuracy (PA), user's accuracy (UA), and F1 score (F1). The OA represents the proportion of correctly classified samples; hence it is a global metric that depends on larger classes. The PA (also known as recall) represents the probability that a particular class on the reference is correctly classified. The UA (also known as precision) is the probability that a pixel classified into a given class actually represents that class on the reference. Finally, the F1 score is the harmonic mean of UA and PA. The F1 score is a more suitable metric in scenarios where class distribution is uneven, as in our dataset.

3.4. Results

Figure 3 summarizes the per-month results obtained for CNN- CRF_G and CNN- CRF_A , in terms of OA, average F1 score (avgF1), average producer's accuracy (avgPA) and average user's accuracy (avgUA), for a sequence comprising the 14 SAR images from October 2015 to July 2016. The horizontal axis contains the month being classified. In this figure, we report just one result per



Figure 4 Maps of the class output for each method for a selected area for the nine annotated months. GT stands for ground truth. Same color legend as in Figure 2.

ore per class for July (best values are highlighted ir			
	CNN	CNN-CRF _G	CNN-CRF _A
maize	46.3	50.0	48.9
cotton	89.1	89.9	89.9
sorghum	61.4	52.6	62.3
NCC	35.0	32.2	47.9
pasture	74.9	71.3	78.9
eucalyptus	96.3	95.2	93.9
soil	76.4	75.4	74.4
turfgrass	74.0	71.9	85.5
cerrado	73.0	67.2	74.2

 Table 2

 F1 score per class for July (best values are highlighted in bold).

month, considering the nine annotated months. In addition, the figure presents the results for the baseline model where we report the performance for the *CNN* output (first bar in the figure).

67.3

72.9

69.6

Average

The results revealed that CNN-CRFA consistently outperformed CNN-CRFG method in terms of OA and avgUA. In contrast, CNN-CRFG outperforms CNN-CRFA in terms of avgPA. Nonetheless, looking at the F1 score, $CNN-CRF_A$ overcomes $CNN-CRF_G$ in 5 out of the 9 months by a high margin, and for the other 4 months, we observed a drop in performance by a shallow margin. We also observed that CNN-CRF_A model significantly outperforms the baseline model in 8 out of the 9 months considering the F1 score, achieving up to 4.4% of improvement. Similar behavior was observed for avgUA, with CNN- CRF_A reporting the best results for all months. CNN-CRF_G reported higher values in terms of PA. In contrast, CNN-CRF_G presented worse results compared to CNN model in terms of OA for almost all months. Considering F1 score, CNN-CRF_G also presented low performance in 4 months compared to CNN.

Table 2 reports the F1 score for all crop types for July, when the higher drop in performance was observed for CNN- CRF_G model compared with the baseline. Notice that CNN- CRF_G reported a significant drop for *sorghum* and *cerrado*. Figure 2 shows that these are two minority classes with very different crop's dynamics. That result indicates that global transitions could potentially favor more abundant classes. In contrast, CNN- CRF_A had higher robustness, achieving better results for 7 out of the 9 classes when compared with the baseline method.

Figure 4 presents the classification maps for each method and month in a selected area. Notice that for all methods, the prediction maps were more accurate for those classes with a higher number of samples: *soil* for October, November, and March; *soybeans* from December to February; and *maize* from May to June. Looking at the field located on top-right in the selected area, we observed that the *CNN* method frequently

misclassify *pasture* as others classes such as *soybeans* and *soil* in the sequence. These changes are inconsistent with the temporal dynamic observed in the region. In contrast, the *CNN-CRF* variants correctly identify the class *pasture* for almost all pixels within this field along the whole sequence. In addition we observed that for some fields all models erroneously classify *pasture* as *cerrado*. These errors were more frequently observed for the *CNN-CRF*_G model.

Our results indicate that learning a transition matrix for each pair of adjacent months enables training a more robust model for modeling the complex crop dynamics observed in our tested region. Our results also revealed that jointly decoding a chain of labels ensures temporal consistency and improves the per-month classification. Finally, we observed that the inclusion of the temporal context for adjacent dates brings more benefits for less abundant classes, in which any error in classification may cause a significant variation in the analyzed metrics.

4. Conclusions and future work

This work introduced a hybrid deep learning architecture for multi-temporal crop recognition, which combines the spatio-temporal context encoding of 3D CNNs and the temporal modeling capabilities of CRFs, in an end-to-end framework. In contrast to existing similar approaches, which learn a global transition matrix that models the temporal dynamics, our method proposes to learn different transitions matrices, one for each pair of adjacent dates. We tested the models using a publicly available multi-temporal SAR dataset from a tropical region with highly complex spatio-temporal crop dynamics.

The experiments indicated that the proposed end-toend framework consistently outperformed a baseline model that disregards temporal modeling. As further research, we intend to test an approach that penalizes the transitions that may not occur according to the expert knowledge in the target region. In addition, further detailed studies will be conducted to assess the influence of the temporal context modeled by the CNN and the CRF modules in the classification accuracy. In a subsequent step, we will test other CNN architectures (e.g., Fully Convolutional LSTM networks) for learning the emission scores.

Acknowledgments

We gratefully acknowledge the financial support offered by the Brazilian National Council for Scientific and Technological Development (CNPq) and the Foundation for Support of Research and Innovation, Rio de Janeiro State (FAPERJ).

References

- U. Nations, World population prospects 2019: Highlights., Department of Economic and Social Affairs, Population Division ST/ESA/SER.A/423 (2019).
- [2] N. Ramankutty, Z. Mehrabi, K. Waha, L. Jarvis, C. Kremen, M. Herrero, L. H. Rieseberg, Trends in global agricultural land use: implications for environmental health and food security, Annual review of plant biology 69 (2018) 789–815.
- [3] B. L. Bodirsky, J. P. Dietrich, E. Martinelli, A. Stenstad, P. Pradhan, S. Gabrysch, A. Mishra, I. Weindl, C. Le Mouël, S. Rolinski, et al., The ongoing nutrition transition thwarts longterm targets for food security, public health and environmental protection, Scientific reports 10 (2020) 1–14.
- [4] I. Del'Arco Sanches, R. Q. Feitosa, P. M. Achanccaray Diaz, M. Dias Soares, A. J. Barreto Luiz, B. Schultz, L. E. Pinheiro Maurano, Campo verde database: Seeking to improve agricultural remote sensing of tropical areas, IEEE Geoscience and Remote Sensing Letters 15 (2018) 369–373. doi:10.1109/LGRS.2017.2789120.
- [5] P. B. C. Leite, R. Q. Feitosa, A. R. Formaggio, G. A. O. P. da Costa, K. Pakzad, I. D. Sanches, Hidden markov models for crop recognition in remote sensing image sequences, Pattern Recognition Letters 32 (2011) 19–26.
- [6] S. Siachalou, G. Mallinis, M. Tsakiri-Strati, A hidden markov models approach for crop classification: Linking crop phenology to time series of multisensor remote sensing data, Remote Sensing 7 (2015) 3633–3650.
- [7] P. Achanccaray, R. Q. Feitosa, F. Rottensteiner, I. Sanches, C. Heipke, Spatial-temporal conditional random field based model for crop recognition in tropical regions, in: 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), IEEE, 2017, pp. 3007–3010.
- [8] L. E. Cué La Rosa, R. Queiroz Feitosa, P. Nigri Happ, I. Del'Arco Sanches, G. A. Ostwald Pedro da Costa, Combining deep learning and prior knowledge for crop mapping in tropical regions from multitemporal sar image sequences, Remote Sensing 11 (2019) 2029.
- [9] S. Ji, C. Zhang, A. Xu, Y. Shi, Y. Duan, 3d convolutional neural networks for crop classification with multi-temporal remote sensing images, Remote Sensing 10 (2018) 75.
- [10] J. A. C. Martinez, L. E. C. La Rosa, R. Q. Feitosa, I. D. Sanches, P. N. Happ, Fully convolutional recurrent networks for multidate crop recognition from multitemporal image sequences, ISPRS

Journal of Photogrammetry and Remote Sensing 171 (2021) 188–201.

- [11] L. Liu, J. Shang, X. Ren, F. Xu, H. Gui, J. Peng, J. Han, Empower sequence labeling with task-aware neural language model, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018.
- [12] J. Li, S. Todorovic, Set-constrained viterbi for setsupervised action segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10820–10829.
- [13] A. Viterbi, Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, IEEE transactions on Information Theory 13 (1967) 260–269.