A recurrent deep Learning model for large scale crop type mapping using time series of Sentinel 2 images

Candidate: Giulio Weikmann
Supervisor: Lorenzo Bruzzone
Assistant Supervisor: Claudia Paris

1 Introduction and Motivation

Precision agriculture is a strategy that integrates Internet of Things (IoT) instruments in crop management in order to improve the efficiency, the productivity, the quality and the sustainability of agricultural production. In particular, food production systems need to optimize the usage of water, energy and fertilizers, reduce pollution and soil degradation while maximizing the agricultural yields. The impact of farming is extremely huge and agriculture faces several challenges. Natural resources have to be monitored, due to their limits, farmers have to operate in sustainable boundaries, in order to reduce the impact on the environment, the agriculture must be intensified to achieve food security but has to remain sustainable. The first step to perform precision agriculture is the correct crop type mapping of the different coltures.

In this framework, Earth Observation (EO) data represent a continuous and reliable source of information to perform accurate crop type mapping and to objectively monitor agricultural areas at large scale. Moreover, with the advent of European programs such as Copernicus, completely full, open and free EO data are now available at global level. The data acquired can be used to train deep learning classifiers, able to extract important statistics and analytic needed for the creation of productivity maps, fertilizer application maps, and the definition of the optimal sewing date. Through EO data, large areas can be constantly monitored, collecting a time series of multispectral images at high spatial resolution, i.e. 10 m. In particular, Sentinel 2 allows the characterization of such phenological parameters of different crop types due to their spectral properties and several multi-temporal radiometric indices typically used to perform crop type classification. However, the analysis of such data requires the definition of suitable architectures and processing techniques, able to exploit all the available temporal information at a reasonable complexity. In this regard, deep learning becomes a powerful tool for remote sensing imagery classification. Moreover, the classification of crop types is a complex scenario, that requires the definition of a multi-temporal approach since cultivation change their spectral and textural appearance according to their crop type growth cycle. Hence, each type of crop has its own phenological characteristics and development time. However, the classification problem taken into consideration brings several different challenges due to: (i) the cloud coverage which hampers the use of several images in the Time Series (TS), (ii) the irregular temporal sampling at large scale since the sensor acquires images at different dates over different regions, (iii) the spatial variability of the spectral signature of the same crop, which may be very different in different locations, and (iv) the imbalanced classification task since some crop types are cultivated extensively while many are poorly represented from the statistical view point. Moreover, most of the pretrained deep learning model are made up for the classification of single images or hyperspectral data. Only few deep learning methods addressed the classication of long TSs of images, focusing on the temporal component. In this context, a large reliable training set is required for the successful training of a multi-temporal deep learning model from scratch.

2 Methodology

In order to deal with multitemporal agricultural datasets, all the images in the TS are required to accurately represent the temporal trends of the crop types. The longer is the TS, the bigger is the

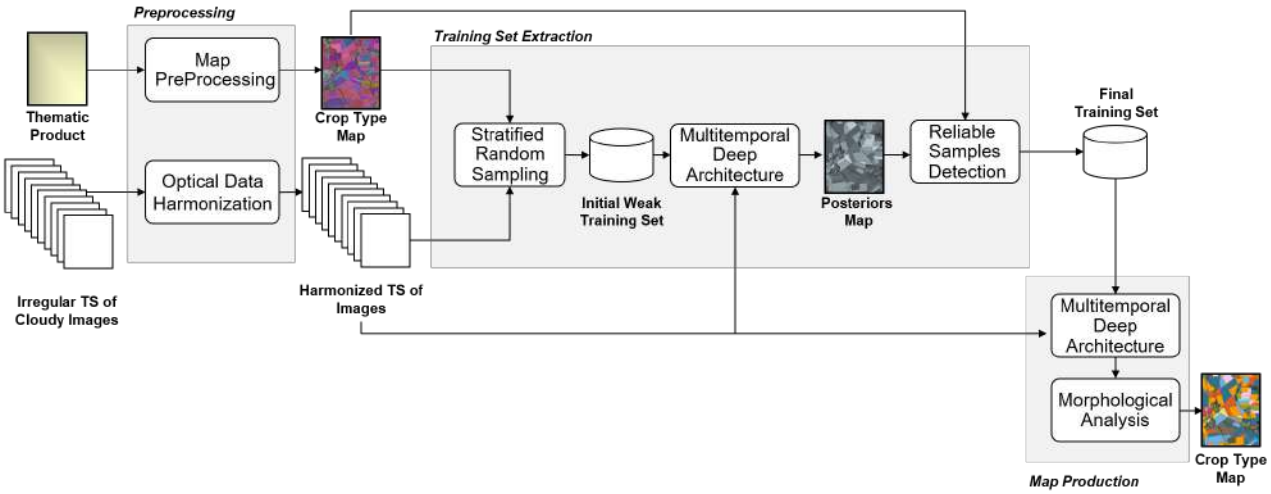

Figure 1: Architecture of the proposed method for crop type mapping by using TSs of Sentinel 2 images and LSTM.

capability of the network to model the temporal dynamic [1]. However, while this is feasible when working at a single tile-level, scaling to larger areas introduce several problems. The different TSs may differ in terms of acquisition dates, acquisition length and satellite orbit. Moreover, the cloud coverage severely hampers the TSs acquired, thus leading to irregular sampling of the study area. This requires: (i) the harmonization of the TSs from the temporal view point, and (ii) the removal of the cloud coverage in the scene. In particular, we generate monthly composite images, where each pixel is generated by computing the band-wise median of the cloud-free multispectral pixels of the images present in the short TS of images. The cloudy pixels present in the masks are removed from the median calculation to strongly reduce the possibility of having cloud coverage present in the monthly composite, thus removing high reflectance value that corrupts the spectral signature of the crops.

Then, the pre-processing of the considered thematic product is performed. The initial map legend has been revised in order to define a classification scheme made of classes that can be discriminated using the spectral and temporal information. Then, to increase the probability of selecting informative and representative pixels, the border of the crops are eroded. This condition allows us to extract clean pixels, avoiding those between neighboring crops characterized by a mixed spectral signature.

In order to successfully train a network from scratch, a large training set representative of the considered study area is required. Nowadays, several crop type maps are available at country scale in several European states due to the Common Agricultural Policy of the European Union. However, these maps are based on farmers declarations and, although they provide an extreme amount of information, they may contain (i) annual crop type declarations that do not match with the real cultivation due to the rotation practice, and (ii) polygons which do not match exactly with purely spectral parcels. Given the complexity of the problem, an automatic method able to extract reliable training samples in an unsupervised way has been defined and the main workflow can be seen in Fig. 1. The methodology adopted is based on three steps: (i) a stratified random sampling to generate a statistically balanced initial weak training set, (ii) a LSTM trained with the weak training set to generate a preliminary crop type map, and (iii) a reliable sample selection analysis to generate the final training set.

During the first stage, the prior probability of each crop type is extracted from the thematic product, in order to obtain a statistically balanced initial training set. The number of samples extracted is directly proportional to the total number of pixels associated to each crop type present in the analyzed area. Due to the highly imbalanced prior probabilities in the problem considered, the classes are initially split into two main categories (most-present and least present classes) defined by a threshold computed automatically considering the statistical distribution of the crop types. To this end, the considered threshold is computed at the 65^{th} quantile of the prior probabilities. The computed prior probabilities are then multiplied and summed by two user-defined constants to (i) guarantee a minimum amount of samples per crop type, and (ii) mitigate the imbalanced classification task.

The initial weak training set is fed to the LSTM adopted in the proposed method. The model generates a preliminary crop type map and a pixel-wise posterior map. The preliminary map obtained

is compared with the original thematic products to detect crops associated to the same class in both the maps. The pixel-wise posterior probability is used to quantify the confidence of the classificator for each pixel. In particular, the last layer of the LSTM outputs the confidence level of the architecture according to the softmax function, where the output is converted into a probability distribution representing the predicted classes [2]. For each class, the samples located where both maps agree, i.e. $\mathbb{M}^{pred} \cap \mathbb{M}^{er}$, and having high confidence values are considered as possible candidates for the final training set. In particular, for the uth class ω_u, only the samples having confidence value higher than a threshold computed as the 25^{th} percentile of the posterior probabilities associated to pixels belonging to ω_u are considered. This criteria allows us to sharply increase the probability of selecting correctly associated to their labels and with a pure spectral signature.

To fully exploit the seasonality of the target, we considered a multitemporal deep learning model consisting in a multi-layer LSTM which provides classification result at pixel level. Since the classification results benefit from an initialization of each layer with a different number of cells, the proposed LSTM is composed by three layers of 200, 125 and 100 hidden units respectively, connected to a fully connected layer whose output is fed to a softmax function. The architecture considers a cross-entropy loss function between the predicted and the ground truth class, whose output value is a probability value between 0 and 1. The loss is backpropagated at each iterations through the network as gradients, which are then used by RMSprop optimizer (a mini-batch version of rprop [3]). Taking into account the imbalanced classification task which characterizes agricultural areas, the cross entropy loss function is changed according to what presented in [4]. In particular, the final loss function can be rewritten as:

$$H' = \sum_{u=1}^{U} H'_u = \sum_{u=1}^{U} \frac{n_{max}}{n_u} H_u \tag{1}$$

where U is the total number of classes, n_{max} the number of samples associated to the dominant class having the higher number of training samples, and n_u the number of samples associated to the uth class ω_u. However, after modifying the cost function, the output represent only an approximation of the a posteriori probability. For this reason, the network's weights obtained by training the network with the proposed loss function are reused as initial weights for the final training of the LSTM using the standard multiclass entropy loss.

3 Experimental results and Conclusion

Different experiments have been carried out to (i) assess the effectiveness of the pre-processing step, (ii) evaluate the need of training the deep learning model with a large training database extracted from the whole study area, and (iii) compare the classification accuracy of the weighted LSTM with respect to the state-of-the-art models. To assess the effectiveness of the proposed architecture, experiments have been carried out in Austria in a study area having spatial extent of $3600km^2$. In order to accurately model the phenological trends of the crops, the Sentinel 2 imagery employed range from September 2017 to August 2018, to represent an agronomic year. The data has been preprocessed in terms of (i) atmospheric correction, and (ii) spatial interpolation of the 20m channels at 10m resolution. The experiments have been carried out in three neighboring tiles, namely T33UUP, T33UVP, and T33UWP. As expected, the different tiles selected have different TSs in order of length and acquisition dates (T33UUP, T33UVP, and T33UWP have respectively 40,70, and 40 images).

To extract the training set, the publicly available 2018 Austrian crop type map [5] has been considered. The defined legend of the Austrian crop type map has been deeply analyzed and aggregated to select a set of crop types suitable for the task proposed. The final classification scheme is made up of 15 classes, namely: "Legumes", "Grassland", "Maize", "Potato", "Sunflower", "Soybean", "Winter Barley", "Winter Caraway", "Rapeseed", "Winter Sugar Beet", "Beet", "Spring Barley", "Winter Wheat", "Winter Triticale" and "Permanent Plantations".

To perform the qualitative and the quantitative evaluation of the results obtained, we focus on tile T33UVP, which is particularly interesting due to the heterogeneity of crops present in the scene and the position on two overlapping orbits, allowing the acquisition of more images than the two

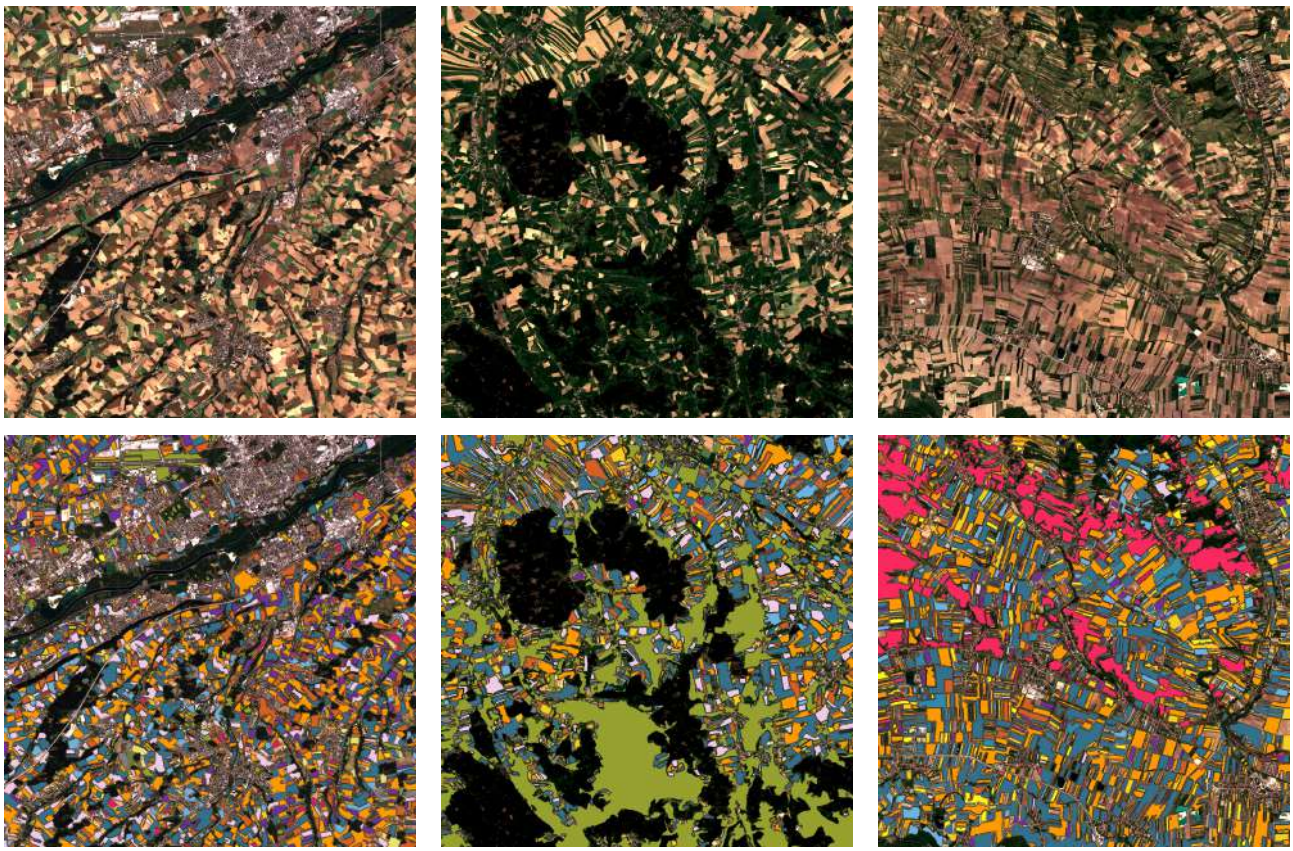

Figure 2: Qualitative example of crop maps obtained for the tiles (i) 33UVP, (ii) 33UUP, and (iii) 33UWP The upper images represents the original Sentinel 2 images in true colors, while the bottom images show the predicted maps.

neighboring tiles. To generate spatially and statistically independent training, the tiles have been split into 100 non-overlapping patches, each assigned randomly to training set, test set, and validation set. The choice of performing patch division extraction allow us to evaluate the capability of scaling at large scale, using area spatially disjointed. The hyper-parameters of the network are selected according to a standard grid search approach.

The first experiment aims to test the impact of the harmonization step on the classification results by comparing the classification results obtained at local level, i.e., using only tile T33UVP. The deep learning model adopted has been trained on (i) the whole TS of images, and (ii) the 12 monthly composites. As expected, the training set obtained considering the monthly composite achieves a lower classification accuracy compared to the one considering the whole TS of images. However, the OA and the F1 are still comparable with the performance obtained using the whole TS. Hence, the monthly composite approach is able to harmonize the different TSs of the neighboring tiles, allowing the (i) extraction of training samples from the whole study area, (ii) usage of a single deep learning model for the whole study area, and (iii) production of crop type maps at large scale.

After assessing the effectiveness of the harmonization step, in the second experiment we evaluated the need of training the deep learning model using a large training database. We compare the classification results on the T33UVP test set using a training set: (i) extracted only from the tile T33UVP, (ii) extracted from T33UUP and T33UWP, and (iii) extracted from the whole study area. Using the training set extracted from the whole area we achieve the highest accuracy. Table 1 shows the results of each experiment conducted. In particular, the OA% and the F1% of the network trained on the monthly composites considering the neighboring tiles are higher than the ones obtained using the whole TS trained with the local training set. Moreover, the minor classes present in the scene are better classified due to the introduction of training samples from the neighboring tiles. The qualitative results of the weighted LSTM over the three tiles analyzed can be seen in Fig. 2.

In order to evaluate the performance of the weighted LSTM, the model has been compared with recent multi- temporal state-of-the-art deep architecture used for crop type mapping. In particular, we considered MSResNet [6], InceptionTime [7] and StarRNN [8]. Table 2 reports the accuracy of

Table 1: Crop type classification results obtained on the test set of tile T33UVP. The Overall Accuracy (OA%), Producer Accuracy (PA%), User Accuracy(UA%) and the Fscore (F1%) are reported for the LSTM trained on (i) the whole T33UVP TS of 70 images, (ii) the 12 monthly composites TS of tile T33UVP, (iii) the 12 monthly composites of tiles T33UUP and T33UWP, and (iv) considering the training set extracted using the monthly composites from each tile.

	Training Set											
	TS of 70 images (UVP)			**TS of 12 composites (UVP)**			**TS of 12 composites (UUP UWP)**			**TS of 12 composites (UVP UUP UWP)**		
	PA%	**UA%**	**F1%**	**PA%**	**UA%**	**F1%**	**PA%**	**UA%**	**F1%**	**PA%**	**UA%**	**F1%**
Legumes	66.96	81.79	73.64	58.09	61.35	59.68	53.77	52.98	53.37	66.07	76.59	70.94
Grassland	87.01	89.07	88.03	87.82	89.38	88.60	73.24	91.59	81.40	90.00	90.50	90.25
Maize	94.07	97.4	95.71	93.85	92.18	93.00	95.55	87.35	91.27	94.34	93.12	93.72
Potato	90.34	87.91	89.11	77.90	84.99	81.29	85.6	66.47	74.83	88.45	85.56	86.98
Sunflower	87.01	38.85	53.71	73.20	26.50	38.91	42.79	43.95	43.36	76.20	58.30	66.06
Soy	61.74	91.10	73.60	50.26	86.05	63.45	48.69	74.85	59.00	58.47	90.40	71.01
Winter Barley	80.54	91.17	85.76	80.43	91.50	85.61	70.02	85.85	77.13	80.87	93.85	86.87
Winter Caraway	90.92	75.65	82.58	93.96	76.35	84.24	89.82	27.80	42.45	94.01	86.40	90.04
Rapeseed	64.54	57.16	60.63	60.38	57.80	59.06	58.69	38.53	46.52	61.39	63.07	62.22
Winter Sugar Beet	88.69	96.50	92.43	93.18	97.80	95.43	71.25	97.90	82.47	91.92	99.00	95.32
Beet	83.98	97.67	90.30	80.04	96.90	87.67	80.71	95.52	87.50	88.57	96.17	92.21
Spring Cereals	82.60	76.94	79.67	80.71	72.49	76.38	73.65	66.09	69.66	84.13	77.34	80.59
Winter Wheat	83.08	80.32	81.68	81.19	84.50	82.81	90.75	69.72	78.86	87.30	83.75	85.49
Triticale	47.10	67.39	55.44	58.39	56.23	57.29	51.22	48.56	49.85	63.01	60.48	61.72
Perm. Plantations	92.19	56.74	70.25	81.39	61.04	69.76	61.57	49.65	54.97	82.01	70.47	75.80
OA%		83.23			80.89			66.18			**84.54**	

Table 2: Crop type classication results obtained on tile T33UVP. The Overall Accuracy (OA%), and the Fscore (F1%) score are reported for the baseline methods InceptionTime, MSResNet, StarRNN and the proposed weighted LSTM.

		Baselines			**Proposed**
		Inc. Time	**MSResNet**	**StarRNN**	**LSTM**
	Prior%	**F1%**			
Legumes	3.1	58.66	**75.62**	71.76	70.94
Grassland	30.9	**90.55**	87.56	89.93	90.25
Maize	12.1	**95.95**	94.33	95.12	93.72
Potato	6.6	83.25	**91.99**	87.99	86.98
Sunflower	2.3	67.11	**71.80**	67.83	66.06
Soy	3.7	**85.14**	79.89	74.97	71.01
Winter Barley	3.9	74.47	90.92	**86.95**	86.87
Winter Caraway	1.6	88.46	**92.09**	87.64	90.04
Rapeseed	3.7	44.00	**62.43**	47.10	62.22
Winter Sugar Beet	3.7	**97.52**	93.82	97.48	95.32
Beet	3.6	89.10	**96.53**	90.59	92.21
Spring Cereals	7.7	73.45	77.84	80.19	**80.59**
Winter Wheat	5.1	76.44	84.75	82.25	**85.49**
Triticale	4.1	55.71	58.23	51.69	**61.72**
Perm. Plantations	7.4	**80.87**	59.98	78.78	75.80
OA%		82.35	84.48	83.97	**84.54**

the different architectures. The architectures have been trained on the three tiles analyzed and the performance are evaluated on tile T33UVP. The LSTM outperformed the three baselines achieving an OA% of 84.54. Moreover, the proposed LSTM achieved the best minimum F1% score (61.72%).

The results showed that the proposed method is able to: (i) harmonize optical data across different tiles and mitigate the cloud coverage problem, (ii) extract automatically a large training database representative of the different crops in the area, (iii) train a LSTM from scratch, and (iv) exploit temporal and spectral properties of the TSs analyzed. We showed that the monthly composite approach does not heavily affect the classification accuracy that can be obtained by using the whole TS of images. Moreover, the loss of accuracy is balanced by the possibility of using a much larger training set acquired over neighboring regions. In fact, using the whole training set, the model obtained an OA% of 84.54%, outperforming the accuracy obtained with the whole TS of 70 images with a local training set. Moreover, higher classification accuracy are achieved also on the minor classes compared to the ones obtained on the whole TS of images with the local training set.

References

[1] Y. T. Solano-Correa, F. Bovolo, L. Bruzzone, and D. Fernández-Prieto, "A method for the analysis of small crop fields in sentinel-2 dense time series," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 3, pp. 2150–2164, 2020.

[2] B. Gao and L. Pavel, "On the properties of the softmax function with application in game theory and reinforcement learning," 2017. [Online]. Available: https://arxiv.org/abs/1704.00805

[3] M. Riedmiller and I. Rprop, "Rprop - description and implementation details," 04 2004.

[4] L. Bruzzone and S. Serpico, "Classification of imbalanced remote-sensing data by neural networks," *Pattern Recognition Letters*, vol. 18, no. 11, pp. 1323–1328, 1997. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167865597001098

[5] AgrarMarkt Austria. (2018) Invekos schläge Österreich 2018. Accessed 15 December 2021. [Online]. Available: https://www.data.gv.at/katalog/dataset/f7691988-e57c-4ee9-bbd0-e361d3811641

[6] F. Wang, J. Han, S. Zhang, X. He, and D. Huang, "Csi-net: Unified human body characterization and pose recognition," 2018. [Online]. Available: https://arxiv.org/abs/1810.03064

[7] H. Ismail Fawaz, B. Lucas, G. Forestier, C. Pelletier, D. Schmidt, J. Weber, G. Webb, L. Idoumghar, P.-A. Muller, and F. Petitjean, "Inceptiontime: Finding alexnet for time series classification," *Data Mining and Knowledge Discovery*, vol. 34, pp. 1–27, 11 2020.

[8] M. O. Turkoglu, S. D'Aronco, and J. Wegner, "Gating revisited: Deep multi-layer rnns that can be trained," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, pp. 1–1, 03 2021.