DuPLO: A DUal view Point deep Learning architecture for time series classification

Roberto Interdonato, Dino Ienco, Raffaele Gaetano, Kenji Ose

To cite this version:

HAL Id: lirmm-02011262
https://hal-lirmm.ccsd.cnrs.fr/lirmm-02011262
Submitted on 20 Mar 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
DuPLO: A DUal view Point deep Learning architecture for time series classification

Roberto Interdonato∗
CIRAD, UMR TETIS, Montpellier

Dino Ienco
IRSTEA, UMR TETIS
LIRMM, University of Montpellier, Montpellier

Raffaele Gaetano
CIRAD, UMR TETIS, Montpellier

Kenji Ose
IRSTEA, UMR TETIS, Montpellier

Abstract

Nowadays, modern Earth Observation systems continuously generate huge amounts of data. A notable example is represented by the Sentinel-2 mission, which provides images at high spatial resolution (up to 10m) with high temporal revisit period (every 5 days), which can be organized in Satellite Image Time Series (SITS). While the use of SITS has been proved to be beneficial in the context of Land Use/Land Cover (LULC) map generation, unfortunately, machine learning approaches commonly leveraged in remote sensing field fail to take advantage of spatio-temporal dependencies present in such data.

Recently, new generation deep learning methods allowed to significantly advance research in this field. These approaches have generally focused on a single type of neural network, i.e., Convolutional Neural Networks (CNNs) or Recur-

∗Corresponding author
Email addresses: roberto.interdonato@cirad.fr (Roberto Interdonato),
dino.ienco@irstea.fr (Dino Ienco), raffaele.gaetano@cirad.fr (Raffaele Gaetano),
kenji.ose@irstea.fr (Kenji Ose)
rent Neural Networks (RNNs), which model different but complementary information: spatial autocorrelation (CNNs) and temporal dependencies (RNNs). In this work, we propose the first deep learning architecture for the analysis of SITS data, namely DuPLO (DUal view Point deep Learning architecture for time series classificatiOn), that combines Convolutional and Recurrent neural networks to exploit their complementarity. Our hypothesis is that, since CNNs and RNNs capture different aspects of the data, a combination of both models would produce a more diverse and complete representation of the information for the underlying land cover classification task. Experiments carried out on two study sites characterized by different land cover characteristics (i.e., the Gard site in France and the Reunion Island in the Indian Ocean), demonstrate the significance of our proposal.

Keywords: Satellite Image Time Series, Deep Learning, Land Cover Classification, Sentinel-2

1. Introduction

Modern Earth Observation (EO) systems produce huge volumes of data every day, involving programs that provide satellite images at high spatial resolution with high temporal revisit period. High-resolution Satellite Image Time Series (SITS) represent a practical way to organize this information, which is particularly useful for area monitoring tasks. A notable example is the Sentinel-2 mission, which is part of the Copernicus Programme, i.e., a programme developed by the European Space Agency (ESA) that involves a constellation of satellites monitoring different aspects of the Earth surface. The Sentinel-2 mission allows to monitor the entire Earth Surface at 10m of spatial resolution with a revisit period between 5 and 10 days, supplying optical information ranging from visible to near and medium infrared. One of the main advantages of this mission is that the produced data are publicly available.

For these reasons, the use of SITS is gaining increasing success in a plethora of different domains, such as ecology, agriculture, mobility, health, risk moni-
In the context of Land Use/Land Cover (LULC) classification, exploiting SITS can be fruitful to discriminate among classes that exhibit different temporal behaviors [10], i.e., with the respect to the results that can be obtained using a single image. In [7], the authors propose to exploit SITS data to extract homogeneous land units in terms of phenological patterns and, later, for the automatic classification of land units according to their land-cover. The effectiveness of Sentinel-2 SITS to produce land cover maps at country scale has been showed in [8], demonstrating the practical interest of such data source. In [9], the authors combine multi-source optical (Landsat-8) and radar (Sentinel-1) SITS in order to improve land cover maps on the agricultural domain. Another example is supplied in [3] where optical SITS are leveraged to characterize grassland area as proxy indicator for biodiversity, food production, and global carbon cycle.

Despite the usefulness of temporal trends that can be derived from remote sensing time series, most of the strategies proposed for SITS analysis tasks [11] [12] [8] [7], directly apply standard machine learning approaches (i.e. Random Forest, SVM) on the stacked images, thus ignoring any temporal dependencies that may be discovered in the data. Indeed, such algorithms make the assumption that the information (spectral bands and timestamps) are independent from each other.

Recently, the deep learning revolution [13] has shown that neural network models are well adapted tools for automatically managing and classifying remote sensing data. The main characteristic of these models is the ability to simultaneously extract features optimized for image classification as well as the associated classifier. Moreover, unlike standard machine learning approaches, they can be used to discover spatial and temporal dependencies in SITS data. Deep learning methods can be roughly categorized in two families of techniques: convolutional neural networks [13] (CNNs) and recurrent neural networks [14] (RNNs). CNNs are well suited to model the spatial autocorrelation available in an image and they are already a well-known tool in the field of Remote Sensing [13] [9] [15]. Conversely, RNNs are specifically tailored to manage time dependencies [14].
from multidimensional time series and they are recently starting to get attention in the Remote Sensing community [16] [17] [18]. Such models explicitly capture temporal correlations by recursion and they have already proved to be effective in different domains such as speech recognition [19], natural language processing [20] and image completion [21].

Considering the analysis of SITS data, few works already exist which exploit deep learning to analyze such kind of data. A CNN based strategy is employed in [9] to deal with land cover classification in the agricultural domain. The main idea is to obtain a single image by stacking all the images in an input SITS, then using it to fed a CNN-based model. The results demonstrate the quality of the proposed approach w.r.t. a standard machine learning algorithm (i.e., Random Forest classifier). In [17] a binary change detection classification task (i.e., change vs. no-change) has been addressed using a RNN model on a small time series of two dates. The authors in [10] propose a RNN-based approach for land cover classification on optical SITS data. The work evaluates deep learning methods on very high resolution and high resolution data on two different study sites. These preliminary results have paved the way to the use of such models for the analysis of Satellite image time series data. Always in the context of land cover classification tasks, RNN have also been used for the analysis of radar SITS data [22] [23].

Even if we acknowledge the existence of a significant body of work dealing with the use of deep learning for the analysis of SITS data, at the same time it should be observed how previous works were tied to the choice of a specific network model, i.e., focusing either on Recurrent or Convolutional Neural Networks. Even though both approaches have been shown to be effective on the analysis of SITS data, our hypothesis is that, since they capture different knowledge aspects, a combination of both would produce a more diverse and complete representation of the information. For this reason, we propose a deep learning architecture for the analysis of SITS data, namely DuPLO (DUal view Point deep Learning architecture for time series classificationOn), based on the combination of Convolutional and Recurrent neural networks. To the best of our
knowledge, *DuPLO* is the first example of deep learning architecture combining RNN and CNN approaches for the analysis of satellite image time series.

The idea behind *DuPLO* is to take advantage of the fact that the two strategies (i.e., CNN and RNN) focus on different characteristics of the data, so that addressing the problem from a dual view point will allow to exploit as much as possible the complementary information produced by such models. Experiments carried out on two study sites characterized by different land cover characteristics (i.e., the *Gard* site in France and the *Reunion Island* in the Indian Ocean), demonstrate the effectiveness of our proposal compared to state of the art approaches for the characterization of land cover mapping on SITS data. Furthermore, the quantitative and qualitative results emphasize how the combination of CNN and RNN is beneficial for the classification task compared to the use of a single neural network model.

The rest of the article is structured as follows: Section 2 introduces the proposed deep learning architecture for land cover classification from SITS data. The study sites and the associated data are presented in Section 3 while, the experimental setting and the evaluations are carried out and discussed in Section 4. Finally, Section 5 draws conclusions.

2. *DuPLO*: A DUal view Point deep Learning architecture for time series classification

Figure 1 shows a visual representation of the proposed *DuPLO* deep learning architecture for the satellite image time series classification process. The optical time series of Sentinel-2 (S2) satellite images in input is first processed independently by the two branches consisting of two different neural networks: a CNN (cf. Section 2.1) and a RNN (cf. Section 2.2). Each branch supplies complementary information for the discriminative process since they look at the information from different points of view. The result of each branch is a feature vector summarizing the extracted knowledge. Each vector is used independently to train an auxiliary classifier (cf. Section 2.3), while the concatenation of both
feature vectors (i.e., in a single feature vector of 2048 descriptors) is used to feed the final classifier that produces the land cover decision.

2.1. CNN Branch

Figure 2 depicts the network architecture associated to the CNN branch of DuPLO. For this branch, we took inspiration from the VGG model [24], a well-known network architecture usually adopted to tackle with standard Computer Vision tasks. The basic idea behind this model is to constantly increase the number of filters along the network, as long as a reasonable size of the feature maps has been reached.

Inspecting this architecture, we can observe that the first convolution has a kernel of 3×3 and it produces 256 feature maps. The second convolution still applies a kernel of size 3×3 outputting 512 feature maps while the last convolution, with kernel size 1×1, is only devoted to increase the final number of extracted features to 1024. All the convolutions are associated with a linear filter, followed by a Rectifier Linear Unit (ReLU) activation function [25] to induce non-linearity and a batch normalization step [26] that accelerates deep network training convergence by reducing the internal covariate shift.
The ReLU activation function is defined as follows:

$$ReLU(x) = \text{Max}(0, W \cdot x + b)$$

(1)

This activation function is defined on the positive part of the linear transformation of its argument \((W \cdot x + b)\) where \(x\) is the input information and \(W\) and \(b\) are parameters learned by the neural network model. The choice of ReLU nonlinearities is motivated by two factors: i) the good convergence properties it guarantees and ii) the low computational complexity it provides \[25\].

Even though the proposed architecture shows a reasonable number of parameters, the training of such models may be difficult and the final model can suffer by overfitting \[27\]. In order to avoid such phenomena, following a common practice for the training of deep learning architectures, we add Dropout \[27\] after each batch normalization step. We set the drop rate equal to 0.4 meaning that 40% of the neurons are randomly deactivated at each propagation step, at training time.
2.2. RNN Branch

Recurrent Neural Networks are well established machine learning techniques that demonstrate their quality in different domains such as speech recognition, signal and natural language processing \cite{28, 20}. Unlike standard feed forward networks (e.g., Convolutional Neural Networks – CNNs), RNNs explicitly manage temporal data dependencies since the output of the neuron at time t-1 is used, together with the next input, to feed the neuron itself at time t.

Recently, recurrent neural network (RNN) approaches have been successfully applied to tasks in the remote sensing field, e.g., to produce land use mappings from time series of optical images \cite{16} and to recognize vegetation cover status using Sentinel-1 radar time series \cite{29}. Motivated by these recent research results, we introduce a RNN module to manage information from the Sentinel-2 time series with the aim to extract an alternative representation from the data. In our model, we choose the GRU unit (Gated Recurrent Unit) introduced in \cite{30} since it has a moderate number of parameters to learn and its effectiveness in the field of remote sensing has already been proved \cite{16, 31}. Due to the fact that we consider patches of satellite images, centered around a central pixel, we do not use the GRU unit directly on the radiometric information. First, we use a shallow CNN, we name \textit{SCNN}, to process the patches at each timestamp and, subsequently, we feed the RNN model with the information extracted by the convolutional models. Finally, we couple the Gated Recurrent Unit with an attention mechanism \cite{32}.

The structure of the \textit{SCNN} is reported in Figure \ref{fig:3}. This shallow network is composed only by two convolutional layers, with 32 and 64 filters respectively, producing an output vector composed by 64 features. This step allows to extract the information carried out by the spatial neighborhood of the considered pixel before considering the temporal behavior of the satellite image time series. Also in this case, each convolution is associated with a linear filter, followed by a Rectifier Linear Unit (ReLU) activation function \cite{25} to induce non-linearity and a batch normalization step \cite{26}.

Once the \textit{SCNN} is applied on the patches describing the time series, the
input of a RNN unit is a sequence \((x_{t_1}, \ldots, x_{t_T})\) where a generic element \(x_{t_i}\) is a feature vector of cardinality equals to 64 (extracted by \(SCNN\)) and \(t_i\) refers to the corresponding time stamp.

Equations 2, 3 and 4 formally describe the \(GRU\) neuron.

\[
\begin{align*}
    z_{t_i} &= \sigma(W_{zx}x_{t_i} + W_{zh}h_{t_{i-1}} + b_z) \quad (2) \\
    r_{t_i} &= \sigma(W_{rx}x_{t_i} + W_{rh}h_{t_{i-1}} + b_r) \quad (3) \\
    h_{t_i} &= z_{t_i} \odot h_{t_{i-1}} + \\ &\quad (1 - z_{t_i}) \odot \tanh(W_{hx}x_{t_i} + W_{hr}(r_{t_i} \odot h_{t_{i-1}}) + b_h) \quad (4)
\end{align*}
\]

The \(\odot\) symbol indicates an element-wise multiplication while \(\sigma\) and \(\tanh\) represent Sigmoid and Hyperbolic Tangent function, respectively.

The \(GRU\) unit has two gates, update \((z_t)\) and reset \((r_t)\), and one cell state, i.e., the hidden state \((h_t)\). Moreover, the two gates combine the current input \((x_t)\) with the information coming from the previous time stamp \((h_{t-1})\). The update gate effectively controls the trade off between how much information from the previous hidden state will carry over to the current hidden state and
how much information of the current time stamp needs to be kept. On the other hand, the reset gate monitors how much information from the previous timestamps needs to be integrated with current information. As all hidden units have separated reset and update gates, they are able to capture dependencies over different time scales. Units more prone to capture short-term dependencies will tend to have a frequently activated reset gate, but those that capture longer-term dependencies will have update gates that remain mostly active [30]. This behavior enables the GRU unit to remember long-term information.

Attention mechanisms [32] are widely used in automatic signal processing (1D signal or language) and they allow to join together the information extracted by the GRU model at different time stamps. The output returned by the GRU model is a sequence of learned feature vectors for each time stamp: \((h_{t_1}, ..., h_{t_N})\) where each \(h_{t_i}\) has the same dimension \(d\). Their matrix representation \(H \in \mathbb{R}^{T \times d}\) is obtained vertically stacking the set of vectors. The attention mechanism allows us to combine together these different vectors \(h_{t_i}\), in a single \(rnn\_feat\), to attentively combine the information returned by the GRU unit at each of the different time stamps. The attention formulation we use, starting from a sequence of vectors encoding the learned descriptors \((h_{t_1}, ..., h_{t_T})\), is the following one:

\[
\begin{align*}
  v_a &= \tanh(H \cdot W_a + b_a) \\
  \omega &= \text{SoftMax}(v_a \cdot u_a) \\
  rnn\_feat &= \sum_{i=1}^{T} \lambda_i \cdot h_{t_i}
\end{align*}
\]

where matrix \(W_a \in \mathbb{R}^{d \times d}\) and vectors \(b_a, u_a \in \mathbb{R}^d\) are parameters learned during the process. These parameters allow to combine the vectors contained in matrix \(H\). The purpose of this procedure is to learn a set of weights \((\omega_{t_1}, ..., \omega_{t_T})\) that allows the contribution of each time stamp to be weighted by \(h_{t_i}\) through a linear combination. The \(\text{SoftMax}(\cdot)\) [16] function is used to normalize weights \(\omega\) so that their sum is equal to 1. The output of the RNN module is the feature vector \(rnn\_feat\): it encodes temporal information related to \(ts_i\) for the pixel \(i\).
With the aim to supply an equal amount of information from each branch of analysis of the DuPLO model, we set the value of \( d \) (the size of the learned feature vector) equal to 1024.

The features extracted by each branch of the architecture are combined by concatenation, resulting in a feature vector of 2048 descriptors, i.e., 1024 from the CNN branch and 1024 from the RNN branch. In this way we assure an equal contribution from each branch to the final decision. Many other combination techniques are possible (e.g., sum, gating, and so on) but we rely on standard concatenation following recent practices introduced by working on multi-source remote sensing analysis \([18, 33]\).

2.3. Training of DuPLO model

One of the advantages of deep learning approaches, compared to standard machine learning methods, is the ability to link, in a single pipeline, the feature extraction step as well as the associated classifier \([13]\). This ability is particularly important when different flows of information need to be combined together, such as in our scenario where we need to couple different representations of the same data source. In addition, the different features learned via multiple non-linear combination of the radiometric information are optimized for the specific task at hand, i.e., land cover mapping.

To further strengthen the complementarity as well as the discriminative power of the learned features for each branch, we adapt the technique proposed in \([34]\) to our problem. In \([34]\), the authors propose to learn two complementary representations (using two convolutional networks) from the same image. The discriminative power is enhanced by two auxiliary classifiers, linked to each group of features, in addition to the classifier that uses the merged information. The complementarity is enforced by alternating the optimization of the parameters of the two branches. In our case, we still have a unique source of information (an optical Sentinel-2 time series of satellite images) but we manage it via two processing branches that differ from each other regarding the particular deep learning strategy we employ.
In detail, the classifier that exploits the full set of features is fed by concatenating the output features of both CNN \((\text{cnn}_\text{feat})\) and RNN \((\text{rnn}_\text{feat})\) modules together. Empirically, we have observed that the RNN module overfits the data. To alleviate this problem, we add a Dropout layer \([27]\) on \(\text{rnn}_\text{feat}\) with a drop-rate equals to 0.4. The learning process involves the optimization of three classifiers at the same time, one specific to \(\text{rnn}_\text{feat}\), a second one related to \(\text{cnn}_\text{feat}\) and the third one that considers \([\text{rnn}_\text{feat}, \text{cnn}_\text{feat}]\).

The cost function associated to our model is:

\[
L_{\text{total}} = \alpha_1 * L_1(\text{rnn}_\text{feat}) + \\
= \alpha_2 * L_2(\text{cnn}_\text{feat}) + \\
= L_{\text{fus}}([\text{cnn}_\text{feat}, \text{rnn}_\text{feat}]) \tag{8}
\]

where

\(L_i(\text{feat})\) is the loss function associated to the classifier inputed with the features \(\text{feat}\). In our case, for all the classifiers (the auxiliary and the main ones) we adopt two fully connected layers of 1024 neurons with ReLU activation function plus a final output layer with as many neurons as the number of land cover classes to predict. The SoftMax activation function is finally applied \([13]\) on the output layer with the aim to produce a kind of probability distribution over the class labels.

Each cost function is modeled through categorical cross entropy, a typical choice for multi-class supervised classification tasks \([10]\).

\(L_{\text{total}}\) is optimized end-to-end. Once the network has been trained, the prediction is carried out only by means of the classifier involving the concatenated features \([\text{cnn}_\text{feat}, \text{rnn}_\text{feat}]\). The cost functions \(L_1\) et \(L_2\), as highlighted in \([34]\), operate a kind of regularization that forces, within the network, the features extracted to be discriminative independently.
3. Data

The analysis was carried out on the Reunion Island, a French overseas department located in the Indian Ocean and on part of the Gard department in the South of France. The Reunion Island dataset consists of a time series of 34 Sentinel-2 (S2) images acquired between April 2016 and May 2017 while the Gard dataset consists of a time series of 37 S2 images acquired between December 2015 and January 2017. All the S2 images used are those provided at level 2A by the THEIA pole. We only use bands at 10m Blue, Green, Red and Near Infrared (resp. B2, B3, B4 and B8). A preprocessing was performed to fill cloudy observations through a linear multi-temporal interpolation over each band (cf. Temporal Gapfilling [8]), and the NDVI radiometric indice was calculated for each date [8, 36]. A total of 5 variables (4 surface reflectances plus 1 indice) is considered for each pixel of each image in the time series. The spatial extent of the Reunion Island site is 6 656 × 5 913 pixels while the extent for the Gard site is 4 822 × 6 748 pixels. Considering the former benchmark, the field database was built from various sources: (i) the Registre parcellaire graphique (RPG) reference data of 2014, (ii) GPS records from June 2017 and (iii) photo interpretation of the VHSR image conducted by an expert, with knowledge of the territory, for distinguishing between natural and urban spaces. Also for the latter benchmark, the field database was built from various sources: (i) the Registre parcellaire graphique (RPG) reference data of 2016 and (ii) photo interpretation of the VHSR image conducted by an expert, with knowledge of the territory, for distinguishing between natural and urban areas.

1Data are available via http://theia.cnes.fr preprocessed in surface reflectance via the MACCS-ATCOR Joint Algorithm [35] developed by the National Centre for Space Studies (CNES).

2RPG is part of the European Land Parcel Identification System (LPIS), provided by the French Agency for services and payment.
<table>
<thead>
<tr>
<th>Class</th>
<th>Label</th>
<th># Objects</th>
<th># Pixels</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Crop Cultivations</td>
<td>380</td>
<td>12090</td>
</tr>
<tr>
<td>2</td>
<td>Sugar cane</td>
<td>496</td>
<td>84136</td>
</tr>
<tr>
<td>3</td>
<td>Orchards</td>
<td>299</td>
<td>15477</td>
</tr>
<tr>
<td>4</td>
<td>Forest plantations</td>
<td>67</td>
<td>9783</td>
</tr>
<tr>
<td>5</td>
<td>Meadow</td>
<td>257</td>
<td>50596</td>
</tr>
<tr>
<td>6</td>
<td>Forest</td>
<td>292</td>
<td>55108</td>
</tr>
<tr>
<td>7</td>
<td>Shrubby savannah</td>
<td>371</td>
<td>20287</td>
</tr>
<tr>
<td>8</td>
<td>Herbaceous savannah</td>
<td>78</td>
<td>5978</td>
</tr>
<tr>
<td>9</td>
<td>Bare rocks</td>
<td>107</td>
<td>18659</td>
</tr>
<tr>
<td>10</td>
<td>Urbanized areas</td>
<td>125</td>
<td>36178</td>
</tr>
<tr>
<td>11</td>
<td>Greenhouse crops</td>
<td>50</td>
<td>1877</td>
</tr>
<tr>
<td>12</td>
<td>Water Surfaces</td>
<td>96</td>
<td>7349</td>
</tr>
<tr>
<td>13</td>
<td>Shadows</td>
<td>38</td>
<td>5230</td>
</tr>
</tbody>
</table>

Table 1: Characteristics of the Reunion Island Dataset

3.1. Ground Truth Statistics

Considering both datasets, ground truth comes in GIS vector file format containing a collection of polygons each attributed with a unique land cover class label. To ensure a precise spatial matching with image data, all geometries have been suitably corrected by hand using the corresponding Sentinel-2 images as reference. Successively, the GIS vector file containing the polygon information has been converted in raster format at the Sentinel-2 spatial resolution (10m).

The final ground truths are constituted of 322 748 pixels (resp. 2 656 objects) distributed over 13 classes for the Reunion Island dataset (Table 1) and 1 157 260 pixels (resp. 2 538 objects) distributed over 8 classes for the Gard benchmark (Table 2). We remind that the ground truth, in both cases, was collected over large areas.
<table>
<thead>
<tr>
<th>Class</th>
<th>Label</th>
<th># Objects</th>
<th># Pixels</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Corn Crop</td>
<td>260</td>
<td>115264</td>
</tr>
<tr>
<td>2</td>
<td>Barley Crop</td>
<td>76</td>
<td>12279</td>
</tr>
<tr>
<td>3</td>
<td>Other Crops</td>
<td>430</td>
<td>127236</td>
</tr>
<tr>
<td>4</td>
<td>Rice</td>
<td>208</td>
<td>124943</td>
</tr>
<tr>
<td>5</td>
<td>Orchards</td>
<td>242</td>
<td>26142</td>
</tr>
<tr>
<td>6</td>
<td>Olive Trees</td>
<td>203</td>
<td>9367</td>
</tr>
<tr>
<td>7</td>
<td>Meadow</td>
<td>230</td>
<td>59018</td>
</tr>
<tr>
<td>8</td>
<td>Vineyard</td>
<td>259</td>
<td>76815</td>
</tr>
<tr>
<td>9</td>
<td>Forest</td>
<td>171</td>
<td>73915</td>
</tr>
<tr>
<td>10</td>
<td>Urbanized areas</td>
<td>266</td>
<td>10584</td>
</tr>
<tr>
<td>11</td>
<td>Water Surfaces</td>
<td>193</td>
<td>521873</td>
</tr>
</tbody>
</table>

Table 2: Characteristics of the Gard Dataset

4. Experiments

In this section, we present and discuss the experimental results obtained on the study sites introduced in Section 3. We carried out several experimental stages, in order to provide a complete analysis of the behavior of DuPLO:

- we provide an ablation study in which we evaluate the importance of the different components of DuPLO (Section 4.2);
- we perform an extensive quantitative analysis, comparing the global classification performances and the per-class results obtained by DuPLO w.r.t. competing methods and baseline approaches (Section 4.3);
- we provide a qualitative discussion considering the land cover maps produced by our framework compared to those produced by competing methods (Section 4.4).
4.1. Experimental Settings

For our analysis, we selected as competing methods two state of the art techniques commonly employed for the classification of SITS. As state of the art standard machine learning tool, we selected a Random Forest (RF) classifier \cite{RF, RF2}, while as regards deep learning techniques we selected the Recurrent Neural Network approach (LSTM) recently proposed in \cite{LSTM}, since it demonstrates interesting performances in the classification of SITS. Furthermore, similarly to what proposed in \cite{LSTM}, we also investigate the possibility to use the deep learning architecture to obtain a new data representation for the classification task. To this end, we feed a Random Forest classifier with the features extracted by DuPLO, naming this approach RF(DuPLO).

All the Deep Learning methods (including DuPLO) are implemented using the Python Tensorflow library. During the learning phase, considering both DuPLO and LSTM, we use the Adam method \cite{Adam} to learn the model parameters with a learning rate equal to $2 \cdot 10^{-4}$. The training process is conducted over 300 epochs with a batch size equals to 128. The number of hidden units for the RNN module is fixed to 1024.

As concerns DuPLO, we perform a preprocessing phase in order to associate each pixel to its surrounding area (i.e., to force the learning process to take into account the spatial context). We consider patches with a spatial extent equals to $5 \times 5$, where each patch represents the spatial context of the pixel in position (2,2). This means that for each timestamp we have a cube of information of size $(5 \times 5 \times 5)$, since 5 is the number of raw bands and indices involved in the analysis. Finally, all the patches are associated to a land cover label that corresponds to the label of the central pixel. The values are normalized, per band (resp. indices) considering the time series, in the interval $[0, 1]$. Regarding the LSTM and RF approaches, according to standard literature \cite{LSTM}, the input is represented by the time series information associated to each pixel.

We divide the dataset into three parts: training, validation and test set. Training data are used to learn the model while validation data are exploited for model selection varying the parameters of each method. Finally, the model
that achieves the best accuracy on the validation set is successively employed to perform the classification on the test set. More in detail, we use 30% of the objects for the training phase, 20% of the objects for the validation set while the remaining 50% are employed for the test phase. We impose that all the pixels of the same object belong exclusively to one of the splits (training, validation or test) to avoid spatial bias in the evaluation procedure [8]. Considering the RF models, we optimize the model via two parameters: the maximum depth of each tree and the number of trees in the forest. For the former parameter, we vary it in the range \{20,40,60,80,100\} while for the latter one we take values in the set \{100, 200, 300, 400, 500\}.

Experiments are carried out on a workstation with an Intel (R) Xeon (R) CPU E5-2667 v4@3.20Ghz with 256 GB of RAM and four TITAN X GPU. The assessment of the classification performances is done considering global precision (Accuracy), F-Measure [10] and Kappa measures.

It is known that, depending on the split of the data, the performances of the different methods may vary as simpler or more difficult examples are involved in the training or test set. To alleviate this issue, for each dataset and for each evaluation metric, we report results averaged over ten different random splits performed with the previously presented strategy.

4.2. Ablation Analysis

In this set of experiments we investigate the interplay among the different components of DuPLO, setting up an ablation analysis in which we disentangle the benefits of the different parts of our framework. To this end, we compare DuPLO with three variants of the original model: i) excluding the use of the auxiliary classifiers (DuPLO\textit{noAux}), ii) considering only the Convolutional Neural Network branch (Cbranch) and iii) considering only the Recurrent Neural Network branch (Rbranch). The results are reported in Table 3 (resp. Table 4) for the Reunion Island (reps. Gard) study site. In both cases we can note that DuPLO outperforms the other variants. This fact underlines that: all the different components play a role in the classification process and support our
intuition that combining different models would produce a more diverse and complete representation of the information. The second finding we can underline is that $C_{branch}$ and $R_{branch}$ behave similarly, showing no real difference in terms of obtained $Accuracy$, $F$-Measure and Kappa measure. Even tough the two variants transform the original information differently, the extracted multifaceted knowledge lets them achieve similar performances. Finally, the use of auxiliary classifiers to boost the discriminative power of each branch independently is confirmed to be effective. It can be observed how for both datasets $DuPLO$ outperforms the variant without auxiliary classifiers ($DuPLO_{noAux}$) as well as the variants involving only one branch of the proposed architecture (i.e., $C_{branch}$ and $R_{branch}$).

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>F-Measure</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>$DuPLO$</td>
<td>83.72% ± 1.08%</td>
<td>83.73% ± 1.03%</td>
<td>0.8089 ± 0.0122</td>
</tr>
<tr>
<td>$DuPLO_{noAux}$</td>
<td>80.28% ± 1.68%</td>
<td>80.25% ± 0.68%</td>
<td>0.7685 ± 0.0075</td>
</tr>
<tr>
<td>$C_{branch}$</td>
<td>79.50% ± 1.00%</td>
<td>79.48% ± 1.11%</td>
<td>0.7594 ± 0.0118</td>
</tr>
<tr>
<td>$R_{branch}$</td>
<td>79.11% ± 1.66%</td>
<td>78.97% ± 1.71%</td>
<td>0.7547 ± 0.0191</td>
</tr>
</tbody>
</table>

Table 3: Accuracy, F-Measure, Kappa considering different ablation of $DuPLO$ on the Reunion dataset

<table>
<thead>
<tr>
<th></th>
<th>Accuracy</th>
<th>F-Measure</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>$DuPLO$</td>
<td>96.36% ± 0.52%</td>
<td>96.28% ± 0.55%</td>
<td>0.9513 ± 0.0067</td>
</tr>
<tr>
<td>$DuPLO_{noAux}$</td>
<td>95.86% ± 0.34%</td>
<td>95.78% ± 0.33%</td>
<td>0.9446 ± 0.0041</td>
</tr>
<tr>
<td>$C_{branch}$</td>
<td>96.01% ± 0.42%</td>
<td>95.91% ± 0.40%</td>
<td>0.9466 ± 0.0056</td>
</tr>
<tr>
<td>$R_{branch}$</td>
<td>96.02% ± 0.43%</td>
<td>95.91% ± 0.40%</td>
<td>0.9466 ± 0.0056</td>
</tr>
</tbody>
</table>

Table 4: Accuracy, F-Measure, Kappa considering different ablation of $DuPLO$ on the Gard dataset

4.3. Comparative Analysis

Table 5 and Table 6 report the results obtained by $RF$, $LSTM$, $DuPLO$ and $RF(DuPLO)$ on the Reunion Island and Gard study sites, respectively. We can observe how on both benchmarks $DuPLO$ outperforms both state of the art competing methods. However, we can also note that using $DuPLO$ as
feature extractor \((RF(DuPLO))\) provides always the best average performances in terms of \textit{Accuracy}, \textit{F-Measure} and Kappa measure. This is in line with recent work in remote sensing \cite{29, 38} and it is due to the fact that, once the newly extracted features are optimized for a particular task, classical machine learning techniques are able to efficiently leverage the information richness carried out by such features.

<table>
<thead>
<tr>
<th></th>
<th>\textit{Accuracy}</th>
<th>\textit{F-Measure}</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>82.99% ± 1.04%</td>
<td>82.40% ± 1.09%</td>
<td>0.7989 ± 0.0119</td>
</tr>
<tr>
<td>LSTM</td>
<td>76.66% ± 1.21%</td>
<td>76.57% ± 1.11%</td>
<td>0.7260 ± 0.0140</td>
</tr>
<tr>
<td>DuPLO</td>
<td>83.72% ± 1.08%</td>
<td>83.73% ± 1.03%</td>
<td>0.8089 ± 0.0122</td>
</tr>
<tr>
<td>(RF(DuPLO))</td>
<td>\textbf{86.12% ± 1.21%}</td>
<td>\textbf{86.00% ± 1.24%}</td>
<td>\textbf{0.8366 ± 0.0143}</td>
</tr>
</tbody>
</table>

Table 5: \textit{REUNION}

<table>
<thead>
<tr>
<th></th>
<th>\textit{Accuracy}</th>
<th>\textit{F-Measure}</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>96.04% ± 0.40%</td>
<td>95.71% ± 0.44%</td>
<td>0.9469 ± 0.0046</td>
</tr>
<tr>
<td>LSTM</td>
<td>95.05% ± 0.55%</td>
<td>94.81% ± 0.59%</td>
<td>0.9338 ± 0.0066</td>
</tr>
<tr>
<td>DuPLO</td>
<td>96.36% ± 0.52%</td>
<td>96.28% ± 0.55%</td>
<td>0.9513 ± 0.0067</td>
</tr>
<tr>
<td>(RF(DuPLO))</td>
<td>\textbf{96.78% ± 0.50%}</td>
<td>\textbf{96.70% ± 0.51%}</td>
<td>\textbf{0.9569 ± 0.0061}</td>
</tr>
</tbody>
</table>

Table 6: \textit{GARD}

4.3.1. \textit{Per-Class Analysis on the Reunion Island benchmark}

Table 7 and Table 8 summarize the per class \textit{F-Measure} performances of the different methods for the \textit{Reunion Island} and \textit{Gard} study site, respectively.

Considering the \textit{Reunion Island} benchmark (Table 7), we can observe that, considering the main competing approaches \((RF, LSTM \text{ and } DuPLO)\), our framework supplies the best classification results on nine over thirteen land cover classes. These classes are: (0), (1), (2), (3), (8), (9), (10), (11) and (12) (resp. \textit{Crop Cultivations}, \textit{Sugar cane}, \textit{Orchards}, \textit{Forest plantations}, \textit{Herbaceous Savannah}, \textit{Bare rocks}, \textit{Urbanized areas}, \textit{Greenhouse crops}, \textit{Water surfaces} and \textit{Shadows}). The highest gains are obtained in correspondence of \textit{Bare rocks}, \textit{Urbanized areas} and \textit{Greenhouse crops} classes with an improvement of 8, 9 and 20
point of F-Measure, respectively. Considering the characteristics of such classes, the significant gains obtained by DuPLO are the results of the effectiveness of our approach to exploit the temporal behavior supplied by the time series as well as the fact that our approach integrates a small amount of spatial context via the patch (a $5 \times 5$ image grid) centered around the analyzed pixel. Regarding all the other land cover classes (Forest Plantations, Meadow, Forest and Shrubby savannah) the performances of DuPLO are comparable with the results obtained by the other approaches. Looking at the RF(DuPLO) method, we can observe that this solution provides the best performances over all the land cover categories and, in particular, we can note that the RF algorithm largely benefits from the feature extracted by the deep learning architecture increasing its average results of more than 3 points. We also investigate the confusion between each pair of classes and we report, for each competing method, the confusion matrix in Figure 4. The visual results support the previous discussion, since a closer look at the heat maps representing the confusion matrix points out that DuPLO and RF(DuPLO) are more precise than the competitors. This consideration emerges from the fact that the corresponding heat maps (Figure 4c and Figure 4d) have a more visible diagonal structure (the dark red blocks concentrated on the diagonal). This is not the case for Random Forest (Figure 4a) and LSTM (Figure 4b) where the distinction between different classes is less sharp.

<table>
<thead>
<tr>
<th>Method</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>61.67%</td>
<td>91.94%</td>
<td>70.12%</td>
<td>66.63%</td>
<td>85.45%</td>
<td>85.45%</td>
<td>73.61%</td>
<td>67.64%</td>
<td>73.96%</td>
<td>82.08%</td>
<td>84.87%</td>
<td>86.41%</td>
<td></td>
</tr>
<tr>
<td>LSTM</td>
<td>42.68%</td>
<td>88.33%</td>
<td>64.85%</td>
<td>52.46%</td>
<td>79.34%</td>
<td>79.34%</td>
<td>60.01%</td>
<td>60.01%</td>
<td>70.82%</td>
<td>81.63%</td>
<td>84.25%</td>
<td>86.51%</td>
<td></td>
</tr>
<tr>
<td>DuPLO</td>
<td>62.36%</td>
<td>92.09%</td>
<td>73.24%</td>
<td>70.61%</td>
<td>82.65%</td>
<td>86.51%</td>
<td>72.12%</td>
<td>71.63%</td>
<td>85.46%</td>
<td>87.47%</td>
<td>85.46%</td>
<td>87.47%</td>
<td>88.81%</td>
</tr>
<tr>
<td>RF(DuPLO)</td>
<td>65.72%</td>
<td>92.98%</td>
<td>75.39%</td>
<td>73.22%</td>
<td>85.46%</td>
<td>87.36%</td>
<td>75.78%</td>
<td>87.36%</td>
<td>86.32%</td>
<td>92.05%</td>
<td>85.46%</td>
<td>94.87%</td>
<td>89.29%</td>
</tr>
</tbody>
</table>

Table 7: Average F-Measure per class for the Reunion Island Dataset

4.3.2. Per-Class Analysis on the Gard benchmark

Considering the Gard benchmark (Table 8), we can observe that all the main competing methods achieve similar performances on all the land cover classes with the exception of the Barley Crop land cover category. While the results on all other land cover classes are satisfactory regarding all the approaches, con-
versely, on this class we can observe a clear different behavior between DuPLO and the other algorithms. RF and LSTM have a poor behavior on Barley Crop with a maximum (average) F-Measure of 31.55%. On the other hand, we can underline that our strategy is capable to reach an F-Measure of 59.30% with an increase of more than 27 percentage points compared to the best state of the art method. Inspecting deeply the results, we have seen that confusion arises between the Corn Crop and Barley Crop classes. This is due to a very similar behavior between such classes, which makes it difficult to discriminate between
them. The ability of DuPLO to disentangle and characterize the temporal behavior of each class efficiently supports the classification of such classes. We can also observe that the RF(DuPLO) strategy provides the best performances over all the land cover categories with the exception of the Forest class, where its performance is similar to the one of the best approach, namely DuPLO (97.60% vs 97.71%).

Figure 5 shows the heat maps representing the confusion matrices on the Gard study site. Also in this case, the visual results support the previous discussion. The heat maps representing the confusion matrices pinpoint that DuPLO and RF(DuPLO) avoid confusion on some particular classes on which the competitors fail. More specifically, we can observe this phenomenon in two different scenarios: the first one is related to the confusion between the Corn Crop and the Barley Crop land cover classes while the second case involves the Olive Trees class. In the former case, DuPLO and RF(DuPLO) are more precise on the Barley Crop while in the latter case, our approaches make some confusion between the Olive Trees and the Meadow, while the other approaches confuse the Olive Trees class not only with Meadow, but also with Vineyard and Forest.

### Table 8: Average F-Measure per class for the Gard Dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>RF</td>
<td>93.40%</td>
<td>31.95%</td>
<td>96.33%</td>
<td>98.69%</td>
<td>91.16%</td>
<td>72.70%</td>
<td>82.64%</td>
<td>92.53%</td>
<td>96.71%</td>
<td>97.15%</td>
<td>99.80%</td>
</tr>
<tr>
<td>LSTM</td>
<td>94.30%</td>
<td>27.58%</td>
<td>95.63%</td>
<td>98.48%</td>
<td>73.37%</td>
<td>69.95%</td>
<td>76.95%</td>
<td>89.48%</td>
<td>96.72%</td>
<td>90.60%</td>
<td>99.82%</td>
</tr>
<tr>
<td>DuPLO</td>
<td>94.92%</td>
<td>59.30%</td>
<td>96.90%</td>
<td>98.99%</td>
<td>90.47%</td>
<td>70.90%</td>
<td>83.28%</td>
<td>91.69%</td>
<td>97.71%</td>
<td>96.77%</td>
<td>99.81%</td>
</tr>
<tr>
<td>RF(DuPLO)</td>
<td>95.26%</td>
<td>62.65%</td>
<td>97.43%</td>
<td>99.10%</td>
<td>82.18%</td>
<td>74.64%</td>
<td>84.66%</td>
<td>93.39%</td>
<td>97.60%</td>
<td>98.63%</td>
<td>99.86%</td>
</tr>
</tbody>
</table>

4.4. Qualitative Inspection of Land Cover Maps

In Figure 6 and 7, we report some representative map classification details on the Gard and Reunion Island datasets considering the RF, LSTM and DuPLO, respectively. With the purpose to supply a reference image with natural colors for the map classification details, we have used the multispectral information obtained from a Very High Spatial Resolution image acquired on the same area.
in the interval spanned by the time series. More in detail, for each study site, we used the multispectral bands of a SPOT7 image at a spatial resolution of 6m.

Regarding the Gard study site, the first example (Figures 6a, 6b, 6c and 6d) depicts an area mainly characterized by forest, meadows and olive trees. On this area, we can observe that both RF and LSTM present confusion between these three classes and do not preserve the geometry of the scene. This is underlined by the salt and pepper error presents in their land cover maps. Conversely, we
can observe that DuPLO supplies a sharper (i.e., more homogeneous) characterization of the three land cover classes, especially concerning the forest class.

The second example (Figures 6e, 6f, 6g and 6h) represents a rural area principally characterized by different crop types. In this case we highlight a barley crop, in order to confirm the strong improvement in performance of DuPLO upon RF and LSTM on this class, already emerged from the quantitative analysis (cf. Table 8). It is evident how both RF and LSTM fail to correctly classifying the Barley Crop, mistakenly identifying it as Corn Crop (interleaved
by some Meadow. Conversely, it can be seen how DuPLO correctly identifies the correct extent of the Barley Crop.

The third example (Figures 6i, 6j, 6k and 6l) involves an urban area, mixed with rural areas and forest. We highlight a large asphalted street which diagonally cuts the selected area. It can be noted how both RF and LSTM classify the street as water while, DuPLO correctly classifies it as Urban Areas. We remind that in our nomenclature we do not have a land cover class for street and, for this reason, detect a street as Urban Areas is a reasonable compromise in our scenario.

Summarizing, on this study area, the qualitative analysis of the land cover maps demonstrates the effectiveness of DuPLO compared to the other approaches on some specific classes like Barley Crop and Corn Crop, confirming the quantitative results reported in Table 8. The analysis also shows how DuPLO provides sharper and spatially coherent classification in mixed areas with respect to competitors which provide land cover maps affected by evident salt and pepper errors.

As concerns the Reunion Island study site, the first example (Figures 7a, 7b, 7c and 7d) depicts a forest area. It can be noted how both RF and LSTM, similarly to what observed for the first Gard example (Figures 6a–6d), are not able to preserve the geometry of the scene, erroneously placing sugar cane and orchards areas among the forest ones. Conversely, DuPLO confirms its ability to produce a sharper and homogeneous demarcation of the forest.

The second example (Figures 7e, 7f, 7g and 7h) shows a coastal area, where we highlight two urban settlements. The salt and pepper error that has been found to characterize the land cover maps produced by RF and LSTM can be clearly observed, with spots of meadow, orchards, greenhouse crops and sugar cane in the middle of urban areas, while the land cover map produced by DuPLO depicts far more clear urban areas.

In the third example (Figures 7i, 7j, 7k and 7l) a bare rocks area can be observed, produced by a lava flow of the Piton de la Fournaise volcano on the eastern side of the island. It can be seen how RF and LSTM place urban areas
and water in the middle of the bare rocks while, DuPLO correctly identifies all the rocky area. It is interesting to observe how DuPLO correctly identifies the border between the ocean and the mainland, while both competitors fail.

To sum up, the qualitative inspection of the land cover maps produced for the Reunion Island study site confirms the quantitative results discussed in Section 4.3 and, consolidates the observations drawn when discussing the land cover maps produced for Gard, especially for what concerns the ability of DuPLO to produce sharper and more spatially coherent land cover maps with respect to both competitors.

In order to have a wider example of how the land cover classification produced by DuPLO differs from the ones provided by competing methods, the
land cover maps produced by DuPLO, RF and LSTM on the Reunion Island study site can be explored on our website.

5. Conclusion

In this paper, a novel Deep Learning architecture to deal with optical Satellite Image Time Series classification has been proposed. The approach, named DuPLO, leverages complementary representation of the remote sensing data to obtain a set of descriptors able to well discriminate the different land cover classes. The two-branch architecture involves a CNN and a RNN branch that process the same stream of information and, due to the difference in their structures, produces a more diverse and complete representation of the data. The final land cover classification is achieved by concatenating the features extracted by each branch. The framework is learned end-to-end from scratch.

The evaluation on two real-world study sites has shown that DuPLO achieves better quantitative and qualitative results than state of the art classification methods for optical SITS data. In addition, the visual inspection of the land cover maps has advocated the effectiveness of our strategy.

As future work, we plan to extend the proposed approach towards the integration of other type of remote sensing data considering a multi-source scenario. For instance, our Deep Learning strategy can be extended to combine optical and radar SITS (i.e. Sentinel-2 and Sentinel-1 data) for land cover classification.

6. Acknowledgements

This work was supported by the French National Research Agency under the Investments for the Future Program, referred as ANR-16-CONV-0004 (DigitAg), the GEOSUD project with reference ANR-10-EQPX-20, the Programme

\[https://193.48.189.134/index.php/view/map/?repository=duplo&project=duplo\]

Be aware that predictions are not available on some areas of the map (i.e. on the right side of the volcano) where there were not enough data available to perform the temporal gapfilling preprocessing on the SITS due to cloud issues (cf. Section 3).

References


URL https://doi.org/10.1016/j.ecoinf.2014.06.007


28


[33] R. Gaetano, D. Ienco, K. Ose, R. Cresson, Mrfusion: A deep learning architecture to fuse pan and ms imagery for land cover mapping, CoRR abs/1806.11452.


[38] R. Gaetano, D. Ienco, K. Ose, R. Cresson, Mrfusion: A deep learning architecture to fuse PAN and MS imagery for land cover mapping, CoRR abs/1806.11452. arXiv:1806.11452
URL http://arxiv.org/abs/1806.11452