MULTI-TEMPORAL LAND COVER CLASSIFICATION WITH LONG SHORT-TERM MEMORY NEURAL NETWORKS

M. Rußwurm,* M. Körner

Technical University of Munich, Chair of Remote Sensing Technology, Computer Vision Research Group
Arcisstraße 21, 80333 Munich, Germany – {marc.russwurm,marco.koerner}@tum.de

Commission III, WG III/7

KEY WORDS: Long Short-Term Memory, Recurrent Neural Networks, Sentinel 2, Crop Identification, Deep Learning, Land Cover Classification

ABSTRACT:

Land cover classification (LCC) is a central and wide field of research in earth observation and has already put forth a variety of classification techniques. Many approaches are based on classification techniques considering observation at certain points in time. However, some land cover classes, such as crops, change their spectral characteristics due to environmental influences and can thus not be monitored effectively with classical mono-temporal approaches. Nevertheless, these temporal observations should be utilized to benefit the classification process. After extensive research has been conducted on modeling temporal dynamics by spectro-temporal profiles using vegetation indices, we propose a deep learning approach to utilize these temporal characteristics for classification tasks. In this work, we show how *long short-term memory* (LSTM) neural networks can be employed for crop identification purposes with SENTINEL 2A observations from large study areas and label information provided by local authorities. We compare these temporal neural network models, *i.e., LSTM* and *recurrent neural network* (RNN), with a classical non-temporal *convolutional neural network* (CNN) model and an additional *support vector machine* (SVM) baseline. With our rather straightforward LSTM variant, we exceeded state-of-the-art classification performance, thus opening promising potential for further research.

1. INTRODUCTION

In earth observation, the problem domain of *land cover classification* (LCC) has educed a variety of techniques until today. Many approaches rely on mono-temporal observations and concentrate on *spectral* or *textural* features describing observations acquired at one specific point in time. However, some land cover classes—such as, *e.g.,* vegetation and especially crops—are difficult to classify by mono-temporal approaches (Foerster et al., 2012), as vegetation changes its spectral and textural appearance within its species-dependent growth cycle. Especially crops develop these temporal dynamics in a systematic and thus predictable manner, dependent on phenology and the applied crop calendar (Valero et al., 2016; Whitcraft et al., 2014). These *temporal* features can be utilized for classification by suitable techniques.

In the recent past, the *deep learning* community has developed a variety of architectures producing impressive results for a wide range of applications. Among these applications, *long short-term memory* (LSTM) neural networks are commonly utilized to handle sequential information in various problem domains, such as natural language processing and text or voice generation. In contrast to mono-temporal models, these LSTM networks can store a theoretically unlimited amount of evidence and make decisions in that actual temporal context. In text generation, for instance, the subsequent word is chosen from the vocabulary body wrt. to a sequence of preceding words. These generated texts imitate the language, grammar, and word choice of the training data.

In this work, we propose to use LSTM networks for the purpose of crop classification from earth observation data. In experiments carried out on a series of SENTINEL 2A images collected over the entire growth season of the year 2016, the effect of multi-temporal features has been evaluated by comparing the performance of multi-temporal models, namely LSTM networks and RNNs, with mono-temporal *convolutional neural network* (CNN) models and a *support vector machine* (SVM) baseline.

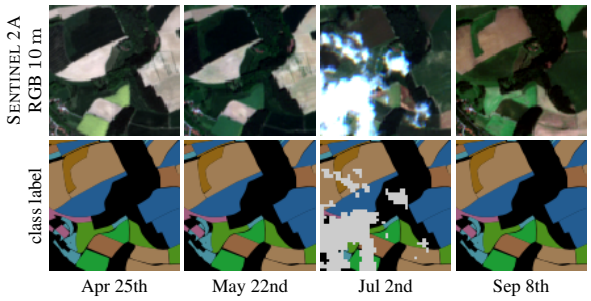

Figure 1. Sequence of observations during the growth season of the year 2016. SENTINEL 2A RGB bands (top row) illustrate the systematic and characteristic change of crops over the season. Class labels (bottom row) show the ground truth labels, as used for the training process. Coverage of the ground is considered by additional *covered* classes, as shown as *clouds* at July 2nd. The systematic temporal changes of spectral reflectances can benefit identification of crops, as exploited by our proposed multi-temporal land cover classification network.

1.1 Remote Sensing of Phenology

Vegetation follows specific periodic growth cycles determined by the plant's biology. The study of these cycles is known as *phenology* and describes characteristic events such as *germination*, *flowering*, or *senescence*. Along with these phenological events,

*Corresponding author

plants change their reflective spectral characteristics which can be observed via remote sensing technologies. Phenological characteristics are assumed to change in a predictive manner and can thus be utilized for identification, as long as farming practices and environmental conditions remain unchanged or are considered in the model (Odenweller and Johnson, 1984; Foerster et al., 2012). Figure 1 illustrates reflective RGB responses of different crops in SENTINEL 2A observations along the growth season. Fields containing the same types of cultivated crops change their spectral appearance uniformly within the series of observation. This is due to a combination of the crops phenological cycles and farming practices, such as the date of seeding and or harvesting.

Commonly, vegetation remote sensing uses *vegetation indices*—*e.g.,* the *normalized difference vegetation index* (NDVI) or *enhanced vegetation index* (EVI)—to observe these changes over a temporal series of observations (Reed et al., 1994). However, these indices usually consider only a limited number of bands which are predominantly sensitive to chlorophyll and water content, *i.e.,* red and near infrared wavelengths. Further spectral bands are often discarded, even though that information is perceived by the satellite and may also contribute to the classification procedure. Additionally, these approaches need to filter high-frequent coverages, such as clouds, as preprocessing routines or by applying *upper envelope filters* (Bradley et al., 2007) to remove negative outliers from the temporal profiles. Overall, these manually-crafted functional models might not be able to represent the complex effects of various influencing factors—such as, for instance, weather conditions, sunlight exposure, or farming practices—which are encoded in the reflectance signal. For these reasons, very recent research has started to employ *deep learning* techniques to overcome these limitations for crop yield prediction (You et al., 2017) and classification of phenological events (Ikasari et al., 2016).

2. RELATED WORK

Even though vegetation analysis with continuous monitoring over the growth season dates back many decades (Reed et al., 1994), only recently space-born sensors—namely the LANDSAT and ESA SENTINEL series—provide sufficient *ground sampling distance* (GSD) and temporal resolution for single-plot field classification. Thus, classical land-cover classification approaches have concentrated on multi- or hyperspectral sensors at one single observation time. Matton et al. (2015) propose a generic methodology for global cropland mapping and statistical temporal features derived from LANDSAT-7 and SPOT images for *K-means* and *maximum likelihood* classifiers on eight test regions on the entire world. Following this approach, Valero et al. (2016) use SENTINEL 2A images to create a binary cropland/non-cropland mask by using *randomized decision forests* (RDF) classifiers on statistical temporal features extracted from spectro-temporal profiles. Foerster et al. (2012) make first attempts to utilize temporal information for per-plot identification by extracting spectro-temporal NDVI profiles and adjusting these profiles by additional agro-meteorological information to account for seasonal variations in phenology. They use LANDSAT-7 images aggregated over several years from a large study area in north-east Germany and classify twelve crop classes in total. While these approaches follow a generic feature extraction and classification pipeline, Siachalou et al. (2015) utilize a *hidden markov model* (HMM) approach which retains sequential consistency of multi-temporal observations on four LANDSAT-7 and one RAPIDEYE observation of Thessaloníki, Greece in 2010. Methodically similar to ours, Lyu et al. (2016) utilize RNNs and LSTM architectures to multispectral LANDSAT-7 and hyperspectral EO-1 HYPERION images, but—in contrast to our approach—for the purpose of change detection.

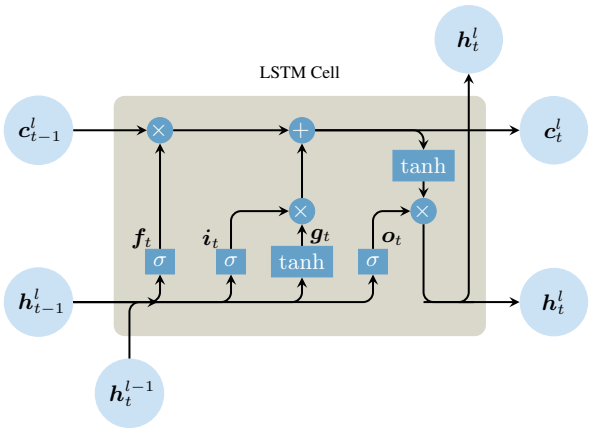

Figure 2. LSTM cells in layer l recieve information of previous observations at time step $t-1$ by means of two vectors: The hidden layer $\boldsymbol{h}_{t-1}^l$ represents the output information at previous time and provides *short-term memory*, similar to RNN cells. The cell state vector $\boldsymbol{c}_{t-1}^l$ carries *long-term* information. Based on the forget gate $\boldsymbol{f}_t$, information in $\boldsymbol{c}_t$ is discarded, input $\boldsymbol{i}_t$ und modulation $\boldsymbol{g}_t$ gates write new information to $\boldsymbol{c}_t$. The updated cell state $\boldsymbol{c}_t$ and the output gate $\boldsymbol{o}_t$ evaluate the hidden output layer $\boldsymbol{h}_t^l$. (Figure adapted from `colah.github.io/posts/2015-08-Understanding-LSTMs`)

3. METHODOLOGY

3.1 Neural Network Architectures

Traditional classification systems are assembled from sequential building blocks, *e.g.,* feature extraction, classification, and post processing, as summarized by Ünsalan and Boyer (2011). Features, which are expected to be significant for classification, are extracted from available observations, *e.g.,* via estimating the parameters of functional models (Bradley et al., 2007). These features are further passed as inputs to classifiers like, for instance, *maximum likelihood* classifiers, SVMs, or RDFs. The optimal choice of feature extraction methodology and classifiers depends on the actual classification task and available data.

In contrast to these approaches, *artificial neural network*s (NNs) are trained in an *end-to-end* manner, solely based on raw *input data* $\boldsymbol{x} \in \mathbb{R}^n$ and *output labels* $\hat{\boldsymbol{y}} \in \mathbb{R}^c$. NNs are usually used for supervised learning to approximate non-linear response functions, *e.g.,* class probabilities, by a sequence of affine transformations $\boldsymbol{W}_{\text{data}} \cdot \boldsymbol{x} + \boldsymbol{b}$ passed to (usually non-linear) activation functions $\sigma : \mathbb{R}^m \mapsto \mathbb{R}^m$, *e.g.,* sigmoid or tangent functions. A *loss function* quantifying the divergence between predicted and actual class probabilities is minimized at each training step by back-propagating residuals and adjusting the network weights $\boldsymbol{W}_{\text{data}} \in \mathbb{R}^{n \times m}$ and biases $\boldsymbol{b} \in \mathbb{R}^m$. Neural networks are commonly arranged in multiple stacked *layers* with the hidden output of one layer forming the input of the consecutive layer. The *number of neurons*, expressed as dimensions of hidden vectors m and *number of layers* l, are common hyper-parameters of which the optimal combination is determined based on classification performance on an `validation` dataset distinct from the training data corpus.

Feature extraction and classification are performed in a joint manner, as the network can inherently select which parts of the input is

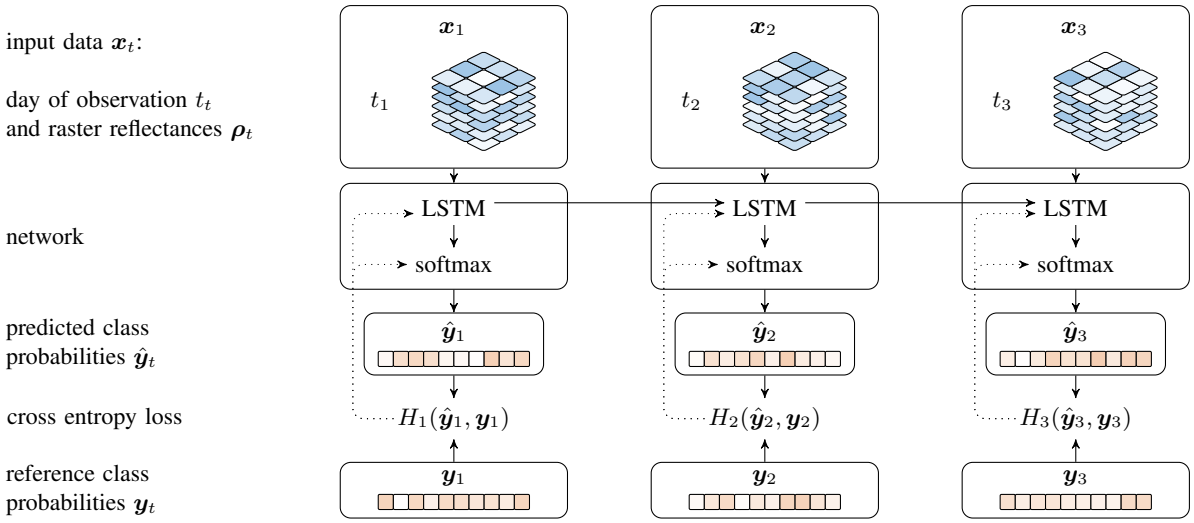

Figure 3. Unrolled flow of data for classification and training. Class probabilities $\hat{\boldsymbol{y}}_t$ are calculated by multiple stacked LSTM layers and one softmax layer. LSTM layers have additional access to cell state vectors $\boldsymbol{c}_{t-1}$ and $\boldsymbol{h}_{t-1}$ originating from previous observations, indicated by an arrow between processing columns, which benefits extraction of temporal features. At each training step, the *cross entropy loss* $H_t(\hat{\boldsymbol{y}}_t, \boldsymbol{y}_t)$ between predicted and reference class probabilites is calculated. To minimize the loss function, gradients (dotted) are calculated by Adam optimizer and adjust the network weights in involved in LSTM and softmax layers.

important for classification based on available data. This scheme ensures that neural network architectures can be applied to a variety of tasks and scenarios, as long as a sufficient quantity of training data is available. Additionally, information which is not important for classification can be gradually ignored by the network. Hence, all available information can safely be provided to the deep learning model. Nevertheless, various neural network architectures have been developed for application to certain fields which—by design—excel at some types of features.

Mono-temporal models *Feed-forward neural networks* process *input data* in a one-directional pipeline. Image processing and segmentation feed forward neural networks incorporate additional *convolutional* layers to account for local neighborhoods and thus are well suited for recognition of shapes and textural patterns. Due to these properties, *convolutional neural networks* (CNNs) are already applied in earth observation for high resolution satellite imagery (Hu et al., 2015) or semantic segmentation (Castelluccio et al., 2015).

Multi-temporal models RNNs (Werbos, 1990) are potentially well suited for processing sequential data, such as temporal sequences of observations, as the network has access to information of the previous observation for the classification of current observation. At each network layer l and observation t, a hidden output vector $\boldsymbol{h}_t^l$ is derived from the output of the previous observation $\boldsymbol{h}_{t-1}^l$ and the input of the current observation $\boldsymbol{h}_t^{l-1}$. Hence, decisions can be based on the *context* of previous observations, thus making RNNs a useful architecture for language processing, text generation, or voice recognition.

Inducing a further level of complexity, *long short-term memory* (LSTM) networks (Hochreiter and Schmidhuber, 1997) introduce an additional *cell state* vector $\boldsymbol{c}_t^l \in \mathbb{R}^m$ providing long-term memory capabilities, as at each observation t information can be stored or retrieved to varying extents. Figure 2 illustrates the flow of vectorized information within one LSTM cell layer l. At each observation t, the previous cell state vector $\boldsymbol{c}_{t-1}^l$ is manipulated by a set of gates, *i.e.,* the *forget gate* $\boldsymbol{f}_t^l$ for discarding information and the *input gate* $\boldsymbol{i}_t^l$ combined with the *modulation gate* $\boldsymbol{g}_t^l$ for writing additional information to $\boldsymbol{c}_t^l$. The output gate $\boldsymbol{o}_t^l$ is derived based on $\boldsymbol{h}_{t-1}^l$ and $\boldsymbol{h}_t^{l-1}$ and provides the same functionality as a RNN layer. The new output vector $\boldsymbol{h}_t^l$ is then obtained by an element-wise multiplication of the output gate $\boldsymbol{o}_t^l$ and the cell state $\boldsymbol{c}_t^l$. While our approach is based on these LSTM networks (Zaremba et al., 2014), a variety of variations have been presented in the past (Gers et al., 2002; Graves and Schmidhuber, 2005; Graves et al., 2013; Kalchbrenner et al., 2015).

3.2 Approach

We employ LSTM neural networks for the purpose of crop classification on a per-plot scale from medium resolution satellite imagery. Figure 3 illustrates the classification and training scheme of our approach for at multiple consecutive observations. At each observation t, n_s spectral *bottom-of-atmosphere* reflectance measurements $\boldsymbol{\rho}_i \in \mathbb{R}^{(k \times k) \cdot n_s}$, along with the day of observation t_i are fed to the network as input vector $\boldsymbol{x}$. A series of l LSTM layers process the data with additional information of the previous observation $t-1$. A *softmax* layer produces probabilities for each class $\hat{\boldsymbol{y}}_t$. At each training step, the *cross-entropy loss* wrt. predicted and actual class probabilities is calculated. Gradients are calculated by *Adam* optimizer (Kingma and Ba, 2014) and back-propagated in order to adjust the network weights.

Crops have been chosen as subject of classification since these land cover classes are expected to change in a characteristic manner, as explained in Section 1.1, thus making these classes ideal subjects to demonstrate the capabilities of temporal modeling by LSTM networks.

4. EXPERIMENTS

In order to evaluate the classification performance of the LSTM architecture and to investigate the effects of temporal features on the classification results, we trained multiple *multi-temporal* models, *i.e.,* based on LSTM networks and RNNs, and *mono-temporal* alternatives, *i.e.,* a CNN model and a baseline SVM, on the dataset described in Section 4.1. In the following, we describe the body of

Figure 4. In our experiments, we used an *area of interest* of 102 km × 42 km (black rectangle) in the north of Munich, Germany, as study area containing 137 k field plots.

obtained data in Section 4.1 and the performed data aggregation to obtain input and label vectors in Section 4.2. Section 4.3 explains the intuitions behind the different dataset partitioning regimes to obtain `training`, `validation`, and `evaluation` subsets in the context of phenological temporal features and regional environmental influences. The training and evaluation process is explained in Section 4.4 and results are shown in Section 4.5.

4.1 Data Material

To train the large number of neural network weights, a feasible body of raster and label data is necessary. For this reason a large *area of interest* (AOI) of 102 km × 42 km in the north of Munich, Germany, has been selected (*cf.* Figure 4) due to its homogeneous geography, climate conditions, and farming practices which suggest similar environmental conditions. A `raster` dataset of 26 SENTINEL 2A images, acquired between 31st December, 2015 and 30th September, 2016, has been retrieved from ESA SCIHUB and atmospherically corrected by SEN2COR software. For consistency reasons with the LANDSAT series, *blue*, *green*, *red*, *near-infrared* and *shortwave infrared 1* and *2* bands were selected for this evaluation. Field geometries and cultivated crop labels of 137 k fields in the AOI have been provided by the *Bavarian Ministry of Agriculture* (*Bayrische Staatsministerium für Ernährung, Landwirtschaft und Forsten*).

The distribution of fields per crop class is not uniform with common cultivated crops, *e.g., corn* or *wheat*, dominating the dataset, while other crops, *e.g., sugar beet* or *asparagus*, are less represented (*cf.* Figure 5). Nevertheless, from 172 unique field crops, 19 *field* classes have been selected with at least 400 field-plots in the AOI.

4.2 Data Extraction

The field geometries of the `field` and reflection values of the `raster` dataset have been discretized to a 100 m × 100 m grid of *points of interest* (POIs). Each POI contains information of network input $\boldsymbol{x}$ and classification ground truth labels $\boldsymbol{y}$ in a 30 m × 30 m neighborhood. The network input vector $\boldsymbol{x}$ incorporated *bottom-of-atmosphere* reflection values directly derived from the `raster` dataset combined with the day of observation

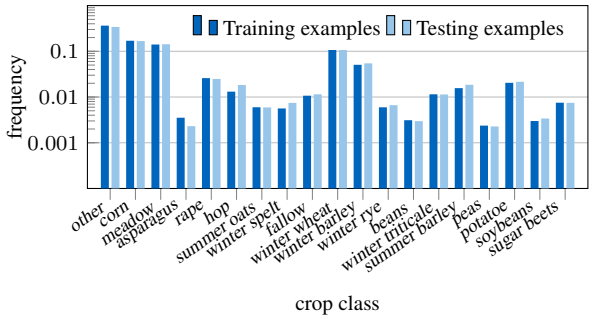

Figure 5. Distribution of classes by fields in the AOI. Common field crops, such as *corn* or *wheat*, dominate the dataset with 28k and 22k fields, respectively, while, *e.g., asparagus* or *peas* are cultivated in less than 600 fields. Only crops with at least 400 occurances have been included in the dataset.

normalized as *fraction of year*. The network labels $\boldsymbol{y}$ have been formed by two types of classes.

1. *Field* classes were derived from the `field` dataset, namely *corn*, *meadow*, *asparagus*, *rape*, *hop*, *summer oats*, *winter spelt*, *fallow*, *winter wheat*, *winter barley*, *winter rye*, *beans*, *winter triticale*, *summer barley*, *peas*, *potatoe*, *sugar beets*, *soybeans*, and the default class *other*.

2. *Covered* classes *cloud*, *water*, *snow*, and *cloud shadow*, account for high frequent coverages and are provided by the *scene classification mask* of SEN2COR extracted from the `raster` dataset.

If POIs were located at the border of multiple classes, class labels have been weighted with respect to a local 30 m neighborhood.

4.3 Dataset Partitioning

Commonly, two sets of parameters need to be determined when selecting and training the neural network architecture. *Weights* $\boldsymbol{W} \in \mathbb{R}^{n \times m}$ are adjusted during the training process by back-propagation of residuals and *hyper-parameters* $\boldsymbol{\theta}$ are chosen following a grid search regime in order to find the optimal network structure for the classification task. To ensure that these parameters are chosen independently, training of network weights and evaluation of hyper-parameters was performed on `training` and `validation` datasets, respectively. A third `evaluation` dataset is used for to calculate accuracy measures of neural network independently from network weights and parameters. While the `evaluation` dataset remained unchanged, `training` and `validation` were redistributed in multiple *folds*. This practice maximizes the number of training samples and avoids misrepresentations of classes containing small numbers of features in the respective dataset. Hence, the body of POIs was divided in the three respective datasets.

As discussed in Section 1.1, the dates of phenological events are influenced by environmental conditions which vary at large spatial distances. To average these environmental conditions, the body of data is divided randomly with respect to the extent of the AOI. A pure random assignment of individual POIs would ensure an even distribution of POIs but may assign POIs of the same field to the different datasets and thus cause dependencies between datasets.

For these reasons, the POIs were not assigned completely randomly to the respective datasets, but have been first partitioned in

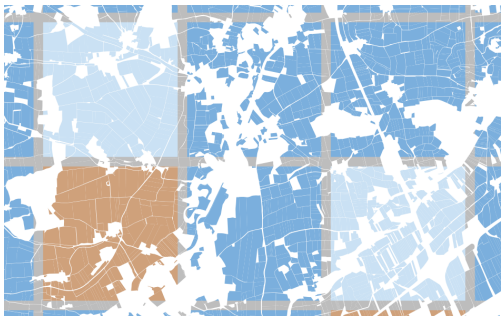

Figure 6. Illustration of the `field` dataset overlayed with 3 km × 3 km blocks dedicated for `training` (blue), `evaluation` (lightblue), and `validation` (orange). A circumferential margin of 200 m ensures that field plots are not located in two distinct datasets. `Training` and `evaluation` datasets are randomly reassigned at each cross-validation *fold*.

blocks. These 476 blocks of 3 km × 3 km were in turn randomly assigned to `training`, `validation`, and `evaluation` in a 4:1:1 ratio (*cf.* Figure 6). A circumferential margin of 200 m ensures that POIs located on the same field were not assigned to different datasets. At each *fold*, `training` and `evaluation` blocks got reassigned randomly while the `validation` dataset remained unchanged.

4.4 Experimental Setup

In total, 135 networks of each architecture have been trained the body of `training` data over 30 epochs with varying hyper-parameters $l \in \{2, 3, 4\}$ and $r \in \{110, 220, 330, 440, 880\}$. This process has been repeated in 9 folds of randomly reassigned `training` and `validation` datasets. Dropout with probability $p_{\text{dropout}} = 0.5$ was used for regularization. Additional to the investigated neural network architectures, a baseline SVM with *radial basis function* (RBF) kernel was trained on a balanced dataset of 3,000 samples per class extracted from the `training` dataset. The optimal hyper-parameters, *i.e.,* slack penalty C and RBF scaling factor γ, have been chosen based on a grid search over $C \in \{10^{-2}, ..., 10^{6}\}$ and $\gamma \in \{10^{-2}, ..., 10^{3}\}$. All networks have been trained within 8 hours on a NVIDIA DGX-1 server equipped with eight NVIDIA TESLA P100 GPUs and 16 GB VRAM each. Five networks have been able to be trained on each GPU in parallel, making the large grid search of parameters possible. While neural networks were implemented in TENSORFLOW, the SVM baseline based on the SCIKIT-LEARN framework.

The best network performances have been achieved by networks with hyper-parameter settings $\boldsymbol{\theta}_{\text{LSTM}} = (l = 4, r = 440)$, $\boldsymbol{\theta}_{\text{RNN}} = (4, 880)$, and $\boldsymbol{\theta}_{\text{CNN}} = (3, 880)$. The SVM baseline performed best with $\boldsymbol{\theta}_{\text{SVM}} = (C = 10, \gamma = 10)$.

4.5 Results

In this section, we evaluate the performance of the trained networks at multiple scales from general performance of each neural network architecture to the performance of best networks on individual classes.

4.5.1 Training performance Figure 7 shows the overall accuracy of each architecture on the `validation` dataset within the training process by means of the average overall accuracy as indication of general performance of the respective architecture. Variations of observed accuracy were presumably caused by the

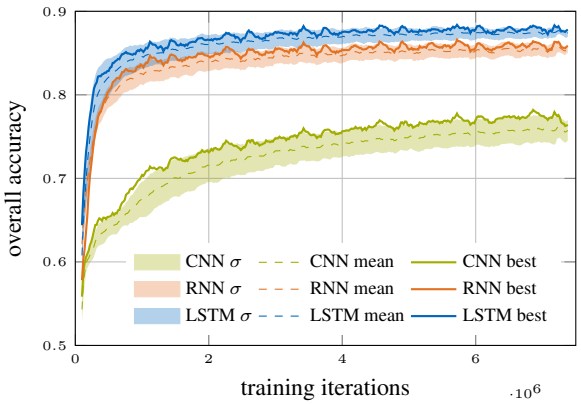

Figure 7. Evolution of overall validation accuracy performance during the training process of 135 networks of each CNN, RNN, and LSTM architecture with different hyperparameter settings. Means (dashed lines) and standard deviation intervals (shaded areas) indicate the general performance of each architecture, the most performand network is superimposed separately (solid line).

chosen hyper-parameter sets and are indicated by their standard deviation intervals. These standard deviations turned out to be relatively small compared to the performances of LSTM, RNN, and CNN models and suggest that, on this dataset, the choice of the actual architecture has larger influence on classification performance than the choice of the involved hyper-parameters. Overall, LSTM networks and RNNs achieved significantly better accuracies over the entire training process with LSTM models performing slightly better than their RNN competitors. The networks which achieved best accuracies are plotted separately as solid lines, as these will be evaluated in detail in the following.

4.5.2 Accuracy measures per best network The best performing networks, opposed to the SVM baseline, have been tested in detail on the `evaluation` dataset. Results of these experiments are compiled in Table 1. Additionally, *covered* and *field* classes have been separated from each other. Similarly to the previous figure, multi-temporal models achieved better accuracies compared to mono-temporal competitors. This performance gain is supposed to be largely caused by the *field* classes which—in contrast to *covered* classes—contain temporal phenological characteristics likely to be utilized by LSTMs and RNNs. *Covered* classes achieved similar accuracies in all of the models with also the baseline SVM achieving good classification accuracies. Apparently, the characteristics of these classes are more distinctive and can be utilized by all models.

4.5.3 Class confusions Similar results can be observed from the confusion matrices shown in Figure 9 based on the best performing networks and the SVM baseline shown in Figure 8. The class frequencies are normalized by the sum of ground truth classes to obtain the precision measure. While some classes represent distinct cultivated crop, other classes—such as *meadow*, *fallow*, or *other*—can not be defined precisely. Consequently, these classes performed worse during our experiments and were more likely confused with a variety of other classes, as becoming apparent in in the confusion matrix. Further chance for confusion was observed in the case of classes with are botanically related to each other and thus share similar spectral and temporal features. For instance, the classes *triticale*, *wheat*, and *rye* have been commonly confused, as *triticale* is a hybrid of the latter two classes. The CNN model, in general, performed worse compared to LSTM and RNN. Some classes (*e.g., sugar beets*, *wheat*) achieved good accuracies,

Table 1. Performance evaluation of our proposed LSTM-based method in comparison to standard RNNs and mono-temporal baselines based on CNNs and SVMs. As *cover* classes (*i.e.*,, *cloud*, *cloud shadow*, *water*, and *snow*) are usually comparatively easy to recognize, we restrict our evaluation to unbiased performance measures with respect to the remaining *field* classes,

Measure	Multi-temporal models						Mono-temporal models					
	LSTM (ours)			RNN			CNN			SVM (baseline)		
	all	*cover*	*field*	*all*	*cover*	*field*	*all*	*cover*	*field*	*all*	*cover*	*field*
ov. accuracy	84.4	98.5	**76.2**	83.4	98.4	74.8	76.8	98.4	59.9	40.9	90.4	31.7
AUC	97.2	99.4	**96.8**	96.5	99.2	95.9	91.7	99.3	90.2	87.1	98.9	84.8
kappa	56.7	65.8	**54.9**	54.5	64.1	52.7	28.6	61.9	22.2	34.3	83.2	24.9
f-score	57.7	67.5	**55.8**	55.6	66.0	53.6	30.1	64.3	23.6	40.3	85.0	31.7
precision	63.8	74.8	**61.8**	59.2	72.3	56.7	47.3	69.1	42.2	40.3	83.1	32.2
recall	56.0	62.8	**54.7**	55.0	62.1	53.6	29.1	61.4	22.9	40.6	87.4	31.7

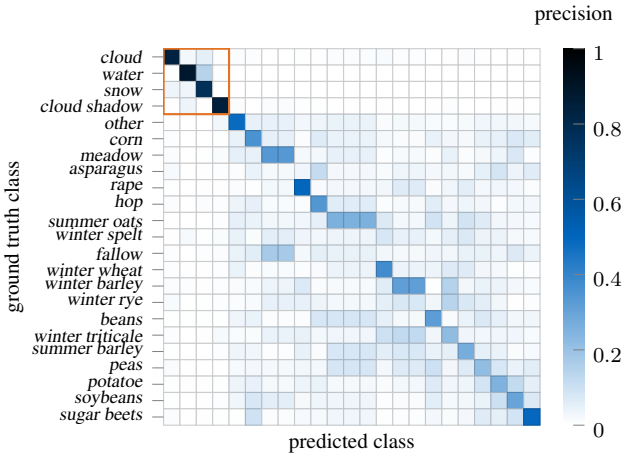

Figure 8. Confusion matrix of SVM classification. While *covered* classes, indicated by the orange-framed submatrix, have been classified well, *field* classes were confused broadly.

while others have not been classified at all or got interchanged with classes showing particular spectral similarities. Presumably the spectral characteristic of some crops are distinct enough for classification without temporal information, while other crops are not distinguishable solely by spectral features. The SVM model showed broad confusion for all *field* classes which indicates that this classifier does not provide a sufficient amount of capacity to encode detailed crop-specific characteristics. SVM, though, performed especially well on the *covered* classes, of which the spectral characteristics are more distinctive.

4.5.4 Accuracy over sequence of observation As LSTM and RNN networks utilize additional information of previous observations, accuracy is expected to improve with increasing sequence length. Figure 10 illustrates this relationship, where the recall values on *field* classes of the three respective networks were calculated by day of observation. While all models, in general, performed equal during the first couple of observations, the accuracy of LSTM and RNN models increased with the sequence of observations. The later the observation is registered, the more context information is available to the temporal models to evaluate the classification decision. The performance of temporal models increases especially at the beginning of vegetation period between March and April (day of year 100), likely due to characteristic phenological events. This trend continues to late summer, up to which crops are harvested, and fields are prepared for the next season, which may cause the slight decrease by the end of growth season.

5. DISCUSSION

In this work, we have shown how to employ LSTM and RNN models for land cover classification. Large-scale experiments have been conducted on a real-world dataset acquired from open-access satellite data together with *in-situ* annotations provided by local authorities. These experiments have shown that LSTM and RNN networks are able to directly utilize temporal information, namely phenological characteristics of crops, for classification and achieve superior results compared to models which—by design—can not benefit from these features but solely rely only on spectral and textural characteristics. All models performed well on classes which do not incorporate characteristic temporal information, such as clouds, while crops can be reliably better classified by LSTM networks and RNNs utilizing distinct phenological events.

Our LSTM model achieved good classification accuracies compared state-of-the art, while considering a notably larger number of crop classes (Foerster et al., 2012; Siachalou et al., 2015). While the *hidden markov model* approach of Siachalou et al. (2015) is methodically closest to our deep learning strategy, their relatively small study area together with their small number of classes impede direct comparison. Our deep learning approach achieved better accuracy performance than the approach of Foerster et al. (2012) using spectro-temporal NDVI profiles and adjusting these by additional agro-meteorological information (*cf.* Figure 11). Their considerably large test area is located in north-east Germany and fields are comparable with ours in terms of cultivated crops and farming practices. Hence, their work is most comparable in terms of data.

In this work, LSTM and RNN architectures have been shown to perform similar, with the LSTM model achieving slightly better accuracies consistently over all evaluation schemes. With increasing observation length, LSTM models may be able to exhaust their full long-term memory capabilities which may be advantageous for monitoring multiple years or data with a higher temporal frequency. As a side effect, the LSTM model have learned *cloud* and *cloud shadow* detection along with field classifications with good overall accuracy of 98.5% by providing additional *covered* classes to the network. Hence, no preprocessing or cloud filtering is necessary, as the network learns to distinguish these coverages in the training process based on provided class labels. This argues for the flexibility of deep learning approaches allowing large amounts of data to be processed without the need of manual data preprocessing and feature selection.

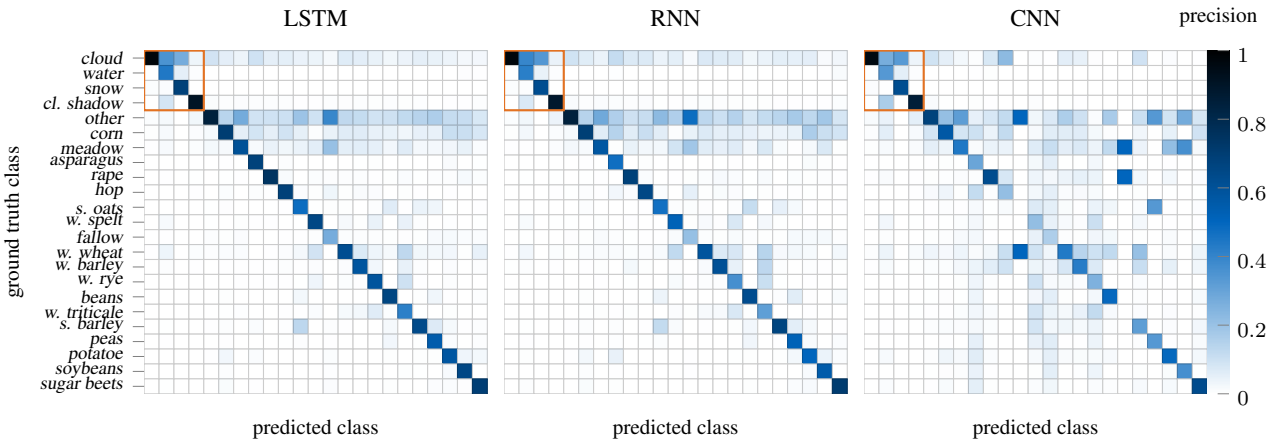

Figure 9. Confusion matrices reporting class-wise average accuracy values of LSTM, RNN, and CNN architectures. The orange-framed submatrix comprises *covered* classes. While LSTM and RNN showed similarily good performances, as these networks utilize temporal features, the CNN network performed generally worse, as temporal characteristic features are not accessible to the CNN architecture.

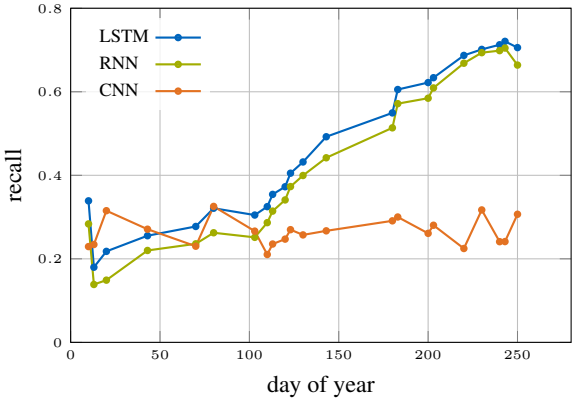

Figure 10. Evolution of performance of LSTM, RNN, and CNN networks with increasing number of observations. The more observations have been registered, the more context information was available to multi-temporal LSTM and RNN networks to assist the classification decision. CNN models can not utilize temporal information, thus performance remained at a nearly constant level with increasing observation length.

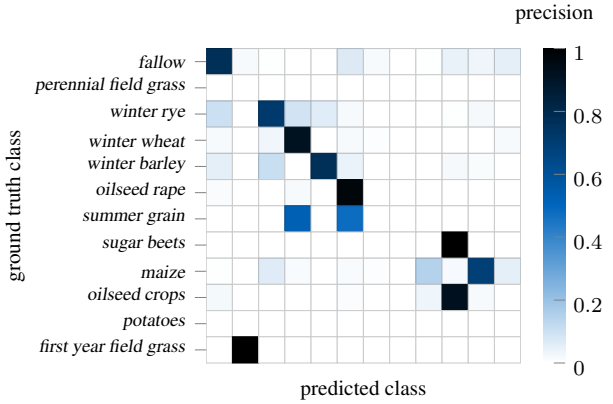

Figure 11. Confusion matrix reporting class-wise average accuracy values reported by Foerster et al. (2012).

6. CONCLUSIONS

Many applications and architectures of the *deep learning* community can be applied to the domain of earth observation for efficient, large scale data processing. This work has demonstrated the applicability of *long short-term memory* (LSTM) networks, originating from speech and text generation, for earth observation. Earth observation, in particular, has to face increasing amounts of data from a variety of multi-modal sensors, such as the SENTINEL, LANDSAT, or MODIS satellite series. The acquired information needs to be processed on a large scale and in an efficient manner exploiting all available information. Considering these requirements, neural networks provide flexibility in terms of preprocessing and provided data, as, *e.g.,* no cloud filtering is necessary if the network has been trained on additional *cloud* classes. Additionally, data which is not significant for the given task will be ignored. We believe that a holistic data approach—comprising temporal, spectral, and textural information—has the potential to yield superior results in future applications. Our presented approach limits textural and spatial features by only observing a small neighborhood of 30 m around each POI to concentrate on available temporal information. In future work, a CNN encoder prepended to the LSTM network could additionally benefit the classification accuracy, as richer textural features would be extracted in a perceptive field optimally chosen by the network.

7. ACKNOWLEDGEMENTS

We would like to thank the Bavarian Ministry of Agriculture for providing information of cultivated crops in high geometrical and semantic accuracy. The TITAN X PASCAL used for this research was donated by the NVIDIA CORPORATION.

References

Bradley, B. A., Jacob, R. W., Hermance, J. F. and Mustard, J. F., 2007. A curve fitting procedure to derive inter-annual phenologies from time series of noisy satellite ndvi data. *Remote Sensing of Environment* 106(2), pp. 137–145.

Castelluccio, M., Poggi, G., Sansone, C. and Verdoliva, L., 2015. Land Use Classification in Remote Sensing Images by Convolutional Neural Networks. *arXiv preprint arXiv:1508.00092* pp. 1–11.

Foerster, S., Kaden, K., Foerster, M. and Itzerott, S., 2012. Crop type mapping using spectral-temporal profiles and phenological information. *Computers and Electronics in Agriculture* 89, pp. 30–40.

Gers, F. A., Schraudolph, N. N. and Schmidhuber, J., 2002. Learning precise timing with lstm recurrent networks. *Journal of machine learning research* 3(Aug), pp. 115–143.

Graves, A. and Schmidhuber, J., 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks* 18(5), pp. 602–610.

Graves, A., Mohamed, A.-r. and Hinton, G., 2013. Speech recognition with deep recurrent neural networks. In: *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, IEEE, pp. 6645–6649.

Hochreiter, S. and Schmidhuber, J., 1997. Long Short-Term Memory. *Neural Computation* 9(8), pp. 1735–1780.

Hu, F., Xia, G.-S., Hu, J. and Zhang, L., 2015. Transferring Deep Convolutional Neural Networks for the Scene Classification of High-Resolution Remote Sensing Imagery. *Remote Sensing* 7(11), pp. 14680–14707.

Ikasari, I. H., Ayumi, V., Fanany, M. I. and Mulyono, S., 2016. Multiple regularizations deep learning for paddy growth stages classification from landsat-8. *arXiv preprint arXiv:1610.01795*.

Kalchbrenner, N., Danihelka, I. and Graves, A., 2015. Grid long short-term memory. *arXiv preprint arXiv:1507.01526*.

Kingma, D. and Ba, J., 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lyu, H., Lu, H. and Mou, L., 2016. Learning a Transferable Change Rule from a Recurrent Neural Network for Land Cover Change Detection. *Remote Sensing* 8(6), pp. 1–22.

Matton, N., Canto, G. S., Waldner, F., Valero, S., Morin, D., Inglada, J., Arias, M., Bontemps, S., Koetz, B. and Defourny, P., 2015. An Automated Method for Annual Cropland Mapping along the Season for Various Globally-Distributed Agrosystems Using High Spatial and Temporal Resolution Time Series. *Remote Sensing* 7(10), pp. 13208–13232.

Odenweller, J. B. and Johnson, K. I., 1984. Crop identification using landsat temporal-spectral profiles. *Remote Sensing of Environment* 14(1-3), pp. 39–54.

Reed, B. C., Brown, J. F., VanderZee, D., Loveland, T. R., Merchant, J. W. and Ohlen, D. O., 1994. Measuring Phenological Variability from Satellite Imagery. Journal of Vegetation Science, 5(5), 703–714. http://doi.org/10.2307/3235884suring Phenologica. *Journal of Vegetation Science* 5(5), pp. 703–714.

Siachalou, S., Mallinis, G. and Tsakiri-Strati, M., 2015. A Hidden Markov Models Approach for Crop Classification: Linking Crop Phenology to Time Series of Multi-Sensor Remote Sensing Data. *Remote Sensing* 7(4), pp. 3633–3650.

Ünsalan, C. and Boyer, K. L., 2011. Review on Land Use Classification. In: *Multispectral Satellite Image Understanding: From Land Classification to Building and Road Detection*, Springer, pp. 49–64.

Valero, S., Morin, D., Inglada, J., Sepulcre, G., Arias, M., Hagolle, O., Dedieu, G., Bontemps, S., Defourny, P. and Koetz, B., 2016. Production of a Dynamic Cropland Mask by Processing Remote Sensing Image Series at High Temporal and Spatial Resolutions. *Remote Sensing* 8(1), pp. 1–21.

Werbos, P. J., 1990. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE* 78(10), pp. 1550–1560.

Whitcraft, A. K., Becker-Reshef, I. and Justice, C. O., 2014. Agricultural growing season calendars derived from MODIS surface reflectance. *International Journal of Digital Earth* 8(3), pp. 173–197.

You, J., Li, X., Low, M., Lobell, D. and Ermon, S., 2017. Deep gaussian process for crop yield prediction based on remote sensing data.

Zaremba, W., Sutskever, I. and Vinyals, O., 2014. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*.