Crop Type Mapping with Multi-Temporal and Multi-Spatial Satellite Imagery

Rose Rustowicz, Robin Cheong, and Lijing Wang
roserust@stanford.edu, robin20@stanford, lijing52@stanford.edu
Stanford University

Abstract

We investigate crop type classification for small holder farms in both Ghana and South Sudan, Africa, using temporal remote sensing imagery and deep learning methods. Our model accommodates both multi-resolution spatial and temporal inputs to produce predictions. We encounter challenges with sparse data labels, class imbalance, and high cloud cover, and achieve an average F1 score and overall accuracy of 0.571 and 58.9% in Ghana, 0.774 and 84.2% in South Sudan, and 0.892 and 94.9% in Germany, surpassing state-of-the-art performance on this separate baseline data set.

1 Introduction

According to the UN, approximately 815 million people around the world are undernourished (2). In particular, countries in East Africa suffer more severely in facing the hardships of food insecurity and malnutrition. Within our study region of Ghana, 33-40% of people face chronic malnutrition in some districts (9). 74-82% of children in northern Ghana suffer from anaemia (9), and economically, Ghana loses 6.4% of it's GDP to child under-nutrition (1). Ghana's employment is dominated by agriculture, where 90% of households in northern Ghana depend on agricultural livelihoods (9). Three of the UN's Sustainability Goals in particular relate to these statistics and motivate this work – *zero hunger*, *good health and well being*, and *decent work and economic growth*, which may all be improved with better understanding of food systems and food security. Accurate crop type maps may be useful for mapping cropland for food yield estimation, understanding farmer crop choice and other growing decisions, gaining insight into interactions of crop types with environmental factors, and information on crop diversity and nutrition outcomes.

We leverage a combination of remotely sensed data with deep learning algorithms to address our motivation of improving food security by mapping crop type from space. We explore crop type classification in Ghana and South Sudan specifically, where this problem is particularly relevant (9), (1), (2). Given a temporal stack of satellite imagery over an agricultural area, we classify scene pixels into their corresponding crop types via semantic segmentation.

2 Related work

Within this study, we explore supervised machine learning techniques for land cover classification of agricultural crop types.

In recent years, deep learning methods have been increasingly used in crop classification over more traditional methods (6), (7). Marc Rußwurm and Marco Körner (10) demonstrated how long short term memory (LSTM) cells, a type of recurrent neural network made to process sequences, could be applied to temporal crop signatures and outperform a convolutional neural network (CNN) for the same task. They report a 76.2% overall accuracy and 55.8% f1-score for 19 crop type classes. They later extend the work with a bidirectional convolutional LSTM network to categorize 17 crops in Munich, Germany. They achieve an overall accuracy of 89.7% and show their model learns to detect and ignore clouds without the need of significant cloud pre-processing (11).

CS230: Deep Learning, Winter 2019, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

(3) uses a simple deep fully connected neural network to aggregate information across different timestamps and applies this method to multi-temporal Landsat data to predict maize versus soybean in Illinois. They report an overall accuracy of 96%. (4) apply a 3D UNet for crop type mapping and achieve 95% accuracy over four crop types. In video segmentation, (13) use a fully convolutional network as a feature extractor which is fed into a recurrent unit cell for semantic segmentation in video, making use of both spatial and temporal features to improve performance. Our approach has unique challenges in that smallholder farms such as those in Africa tend to have smaller fields and sparser ground truth labels as compared to larger studies conducted in places such as the United States and Germany. Smaller fields lead to less pixels of information, while sparse labels introduce missing data gaps. Additionally, the growing season in our study area is dominated by rain and cloud cover, leading to low visibility in optical imagery. Among the works that study Africa, we note that incorporating both radar and optical information often improves performance, but that data is limited (5).

3 Dataset and Features

Through the Lobell Lab at Stanford, we were provided with sparse ground truth data for our study regions in South Sudan and northern Ghana. Ground truth labels consist of geo-referenced polygons, where each polygon represents an agricultural field boundary with a crop label. Labels are created as raster masks where pixels contained within polygons have a class label that corresponds to crop type, while all other pixels outside the labeled fields are set to zero. We use Sentinel-1 and Sentinel-2 satellite data collected over the study region extents to relate spectral data to these label masks. Images are exported to correspond with the 2017 ground truth data, and the number of time stamps for a scene varies from less than 16 to greater than 100 observations. Both satellites have a 10m spatial resolution and a temporal revisit rate of 6 - 12 days. In addition to this data, we incorporate higher spatial and temporal resolution satellite imagery from Planet Labs. Planet's Dove Satellites have a spatial resolution of approximately 3m, which image the entire land mass of the earth each day. With high cloud cover and small field sizes, we believe incorporating this new satellite source will be beneficial.

We subdivide our area of interest into 32 x 32 pixel grids and split according to a 80 / 10 / 10 split for train, validation, and test. Our splitting algorithm ensures that there is no overlap of fields in splits, and also attempts to best preserve the relative percentages of all crops, allowing for consistent class balances in all splits. The right-most plot in Figure 1 shows the data split that was used for Ghana.

In terms of pre-processing, we normalize all input bands to zero mean and unit variance based on statistics from our training set. We build random data augmentation into our data loader to include both rotation and flips. As input features, we use all ten Sentinel-2 bands (blue, green, red, NIR, four red edge bands, and two SWIR bands), both Sentinel-1 bands (VV and VH polarizations), and all four Planet bands (blue, green, red, NIR). We also include day of year as an input band, and construct additional bands commonly used in remote sensing. For example, for Planet and Sentinel-2, we use NDVI and GCVI vegetation indices. For Sentinel-1, we use a ratio of the two bands, VH/VV, as an additional input.

4 Methods

4.1 Random Forest Baseline

To measure baseline performance, we use a random forest classifier applied to each pixel location, where a pixel location has both spectral (the input bands from all satellites) and temporal (observations of the same plot n the ground in time) features. Random forest is a standard method in many land cover classification studies in remote sensing. It is an ensemble machine learning algorithm that creates a forest of thousands of sub-optimal decision trees that we use for classification.

4.1.1 Model Architecture

Because our model input is a sequence of images, we choose to include both CNNs and RNNs within our model architecture in an attempt to capture both spatial and temporal information. Figure 2 shows our model architecture.

To begin, each image in a time series is put through an encoder network that has shared weights between all timestamps. The encoder uses a series of convolution and down sampling layers, with batch normalization and a Leaky ReLU non-linearity after each convolution layer. High resolution Planet imagery is input at the top of the network, which flows through two U-Net "tiers" before being concatenated with features from the input lower resolution imagery from Sentinel-1 and Sentinel-2.

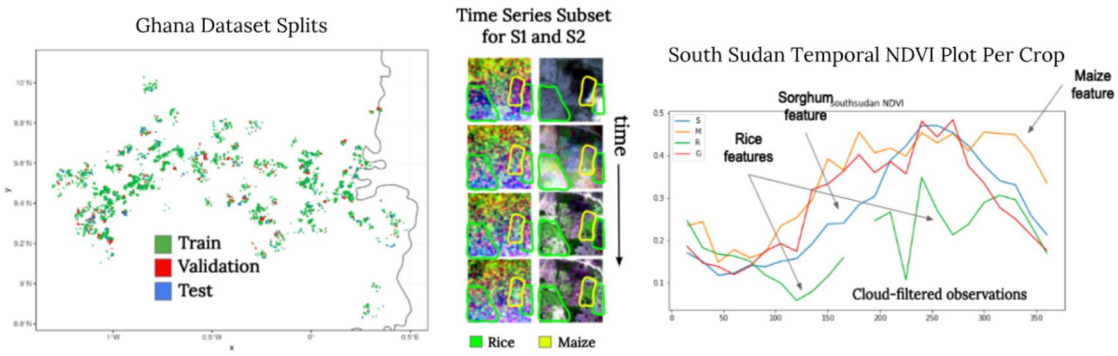

Figure 1: Data set Visualizations. (Left) Data set splits for the Ghana data set. Each dot represents a field location overlaid on a map of Ghana. Notice that splits are well distributed. (Center) A visual plot of input data. The left column shows a time series of Sentinel-1, while the right column shows a time series of Sentinel-2 images. Rice and maize fields are outlined in green and yellow, respectively. Note that crops are not easily differentiated by eye. Also note that we generally use at least twenty time stamps in our models, and that only a small subset (four) is shown here. (Right) Average NDVI of cloud-filtered pixels across time for each crop in South Sudan: maize (yellow), rice (green), sorghum (blue), and groundnut (red). In this case, we can see visual differences between crops, but we note that this is a smoothed out signal from all of the input data, and that the standard deviations for each crop are very large.

.

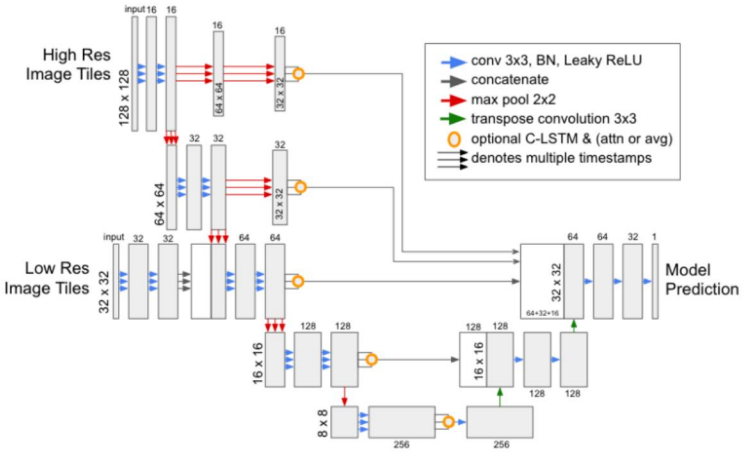

Figure 2: The model architecture used in this study.

.

Before features are input into the decoder network, we experiment with using both a Convolutional LSTM (C-LSTM) (12) and attention to aggregate the relevant information within the time series.

When using the C-LSTM, the convolution features from each time step of the encoder are input to the C-LSTM recurrent cell, a variation of the traditional LSTM, in order to capture temporal information. LSTMs take an input x_t in a vectorized form. CLSTMs, as introduced by Shi (12), convert the matrix vector multiplication in the gates of the LSTM to a convolution operation, allowing the input x_t to be a matrix instead of a vector. Specifically, the gates become:

$$\begin{aligned} i_t &= \sigma(W_{xi} * x_t + W_{hi} * h_{t-1} + W_{ci} \circ c_{t-1} + b_i) \\ f_t &= \sigma(W_{xf} * x_t + W_{hf} * h_{t-1} + W_{cf} \circ c_{t-1} + b_f) \\ c_t &= f_t \circ c_{t-1} + i_t \circ tanh(W_{xc} * x_t + W_{hc} * h_{t-1} + b_c) \\ o_t &= \sigma(W_{xo} * x_t + W_{ho} * h_{t-1} + W_{co} \circ c_t + b_o) \end{aligned} \tag{1}$$

where $*$ is the convolution operator.

At first, we only put the bottom most encoded features through the C-LSTM. At the output of the RNN cell, we experiment with taking both the last hidden state as well as the average of the hidden states. Upon finding that averaging hidden states is helpful, we also investigate the use of attention. We implement two types of self-attention to use rather than averaging, with the thought that the attention mechanism would give a more intelligent weighted average result for further improvement.

The first type of attention we implement is based on (8), in which two matrix multiplications and intermediate non-linearities yield the desired weights for the weighted attention summation. Within this method, attention weights are defined as:

$$A = \sigma(W_{s2} tanh(W_{s1} H^T)) \tag{2}$$

where σ denotes the softmax function, which forces the weights to sum to one. In this expression, W_{s1} is a weight matrix of dimensions $d \times u$, where u is the number of hidden states in the sequence. W_{s2} is a weight matrix of dimensions $d \times r$ where r is the number of attention heads we wish to use to represent the weighted signal. Since H^T has shape $n \times u$, the resulting matrix A has shape $n \times r$, which is used to weight each of the n hidden states, which are them summed to yield the final representation. We implement this attention function using two fully connected layers with weights W_{s1} and W_{s2} and no bias.

We also investigate self-attention as described in (14). In this case, we start with our encoded features H, which we project into matrices of queries, keys, and values via three matrix multiplications. To get a score for the first hidden state, we take an inner product of the first query vector with the key vectors for every sequence element. These inner products give a weight that is used to weight the values in the value matrix V for each time stamp.

Once we have processed all of the encoded features, the decoder uses transpose convolutions to transform the results back up to the size of our predictions. Our model is trained end-to-end with a weighted cross entropy loss function, where weights are chosen as a function of class balance. To handle sparse labels, we only calculate loss on valid image regions that have labels. We mask all others and set the loss at these locations to zero.

5 Experiments/Results/Discussion

We experiment with several features and tweaks without our model architecture. After investigation, we find that using the C-LSTM at the bottom of the "U" and averaging across all hidden states from the C-LSTM and encoding layers produces the best results. We find that incorporation of Planet data also helps performance. Unfortunately, neither version of attention type seems to improve results. We may need to do more hyperparameter tuning to fully investigate if attention is useful for our model.

In terms of hyperparameter search, we search across optimization method (Adam vs. SGD), learning rate, weight decay, number of timestamps, using loss weight, and attention type vs averaging. We find that Adam with a learning rate of 0.001, weight decay of 0.1, 25 timestamps, use of weighted loss, and averaging produces reasonable results across all countries with the exception of Germany where we only have the ability to use 14 timestamps due to data limitations. We may require further tuning to reach optimal performance for each country. Since we have class imbalance, accuracy results are biased toward the dominating crops. To account for this and to give an equal treatment to classification importance across all classes, we decide to compute the F1 score for each class and then average to give an unweighted average F1 score.

Table 3 gives a quantitative overview of our model results compared with the random forest baseline as well as the state-of-the-art reported metrics on the Germany data set (11). Our model performs best in both F1 and accuracy on the large data sets in Ghana and Germany, but is outperformed by random forest on the smaller data set in South Sudan. We hypothesize that the model was not able to perform as well due to lack of data and ability to generalize in this case, although performance across both methods is still generally much higher than in Ghana. We also notice a high performance across both African countries with rice. If we look back on the plots visualizing the data in time in Figure 1 on the right-most plot, we notice that rice seems to differentiate itself the most, which may explain why the model is able to easily differentiate the crop compared to the others.

Code is available here: `https://github.com/roserustowicz/crop-type-mapping`

6 Conclusion/Future Work

In conclusion, we construct a model that incorporates both CNNs and RNNs for semantic segmentation of multi-temporal, multi-spatial satellite images. We predict crop type with reasonable performance in Ghana and

Country	Num Pixels	Random Forest F1 / Accuracy	Our Model F1 / Accuracy	Rußwurm & Körner (2018) F1 / Accuracy
South Sudan	~65k	0.837 / 92.3	0.774 / 84.2	--
Ghana	~575k	0.408 / 59.0	0.571 / 58.9	--
Germany	~1330k	0.251 / 67.7	0.892 / 94.9	0.831 / 89.7

Figure 3: Quantitative results across countries and models.

.

Crop type	Our Model TestF1	Rußwurm & Körner (2018) Test F1
Sugar Beet	0.951	0.853
Summer Oat	**0.806**	**0.758**
Meadow	0.908	0.882
Rapeseed	**0.977**	**0.926**
Hop	0.942	0.917
Winter Spelt	**0.719**	**0.656**
Winter Triticale	0.714	0.618
Beans	**0.950**	**0.896**
Peas	0.885	0.804
Potato	**0.944**	**0.895**
Soybeans	0.895	0.877
Asparagus	**0.897**	**0.837**
Winter Wheat	0.962	0.903
Winter Barley	**0.961**	**0.910**
Winter Rye	0.732	0.607
Summer Barley	**0.934**	**0.854**
Maize	0.979	0.939
Macro Avg. f1-score	**0.892**	**0.831**
Overall Accuracy	0.949	0.897
Kappa Coefficient	**0.937**	**0.870**

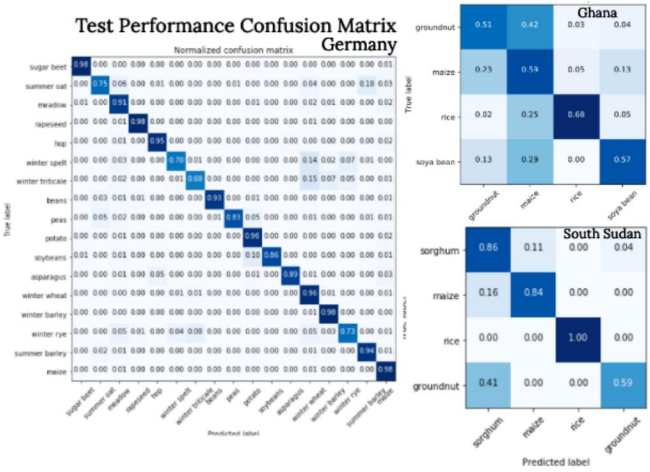

Figure 4: Quantitative performance of our model across all countries. We surpass state-of-the-art performance on the Germany data set and show reasonable results for both Ghana and South Sudan

.

South Sudan where data is limited and of poor quality due to high cloud cover, class imbalance, and lack of labels. When applied on a large data set in Germany, we surpass state-of-the-art performance on this task. We assumed that attention would allow our model to improve based on selective averaging of features across time, but did not find this to be the case. We will need to run more experiments to tune possibilities of using attention within our model. Given more time, we hope to investigate ways to improving performance in crop type mapping for small holder farms where signal is low and labels are lacking. We would consider semi-supervised learning methods with deep generative models, or transfer learning from larger data sets in the United States or Europe.

7 Contributions

Rose curated the Planet data set, worked on modifications to the U-Net to incorporate multi-temporal inputs, implemented the random forest baseline, and organized and processed all necessary data for running the comparison with the Germany data. Robin added many features to the model such as recurrent dropout, recurrent batch norm, and variable length sequences (not shown here). Both Rose and Robin worked on adding attention to the model. Lijing worked with baseline methods last quarter as well as work in comparing our model with a 3D U-Net

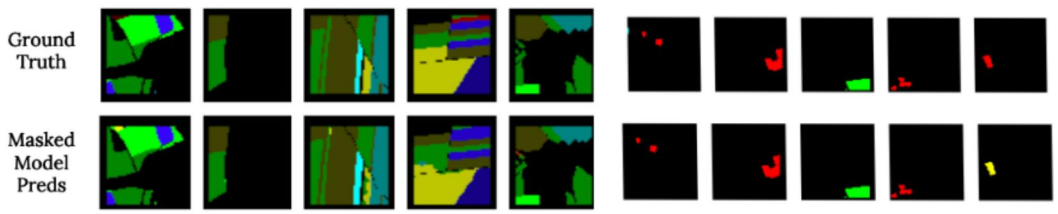

Figure 5: Qualitative Performance in Germany (left) and South Sudan (right). Ground truth labels are in the top row while model predictions are in the bottom row. Each color corresponds to a different crop type.

.

(not shown here). She also worked on incorporating the original U-Net code. Zhongyi Tang exported data for Sentinel-1 and Sentinel-2, while all other authors contributed to other pre-processing steps. This project was worked on as a part of the Lobell Lab and the AI and Sustainability Lab, with advisement from Dr. David Lobell, Dr. Stefano Ermon, and Dr. Mashall Burke. We would also like to thank Burak Uzkent, and other members of the AI and Sustainability Lab and Lobell Lab for useful discussions around this project.

References

[1] The cost of hunger in africa. Technical report, UN Economic Commission for Africa, 2014.
[2] The sustainable development goals report 2018. Technical report, United Nations Department of Economic and Social Affairs, New York, June 2018.
[3] Y. Cai, K. Guan, J. Peng, S. Wang, C. Seifert, B. Wardlow, and Z. Li. A high-performance and in-season classification system of field-level crop types using time-series landsat data and a machine learning approach. *Remote Sensing of Environment*, 210:35 – 47, 2018.
[4] S. J. et al. 3d convolutional neural networks for crop classification with multi-temporal remote sensing images. 2018.
[5] G. Forkuor, C. Conrad, M. Thiel, T. Ullmann, and E. Zoungrana. Integration of optical and synthetic aperture radar imagery for improving crop mapping in northwestern benin, west africa. *Remote Sensing*, 6(7):6472–6499, 2014.
[6] K. Kamilaris and F. X. Prenafeta-Boldu. A review of the use of convolutional neural networks in agriculture. *The Journal of Agricultural Science*, pages 1–11, 2018.
[7] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov. Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geoscience and Remote Sensing*, 14, 2018.
[8] Z. Lin, M. Feng, C. N. dos Santos, M. Yu, B. Ziang, B. Zhou, and Y. Bengio. A structured self-attentive sentence embedding. *ICLR*, 2017.
[9] W. F. Programme. Ghana, 2018.
[10] M. RuÃwurm and M. KÃ¶rner. Multi-temporal land cover classification with long short-term memory neural networks. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-1/W1:551–558, 2017.
[11] M. RuÃwurm and M. KÃ¶rner. Multi-temporal land cover classification with sequential recurrent encoders. *ISPRS International Journal of Geo-Information*, 7(4), 2018.
[12] X. Shi, Z. Chen, H. Wang, D. Yeung, W. Wong, and W. Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *CoRR*, abs/1506.04214, 2015.
[13] S. Valipour, M. Siam, M. Jagersand, and N. Ray. Recurrent fully convolutional networks for video segmentation. *arXiV*, 2016.
[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *31st Conference on Neural Information Processing Systems (NIPS)*, 2017.