Understanding Satellite-Imagery-Based Crop Yield Predictions

Mark Sabini, Gili Rusak and Brad Ross
Stanford University
{msabini, gili, bross35}@stanford.edu

Abstract

Researchers like [26] have already trained Convolutional Neural Networks to predict crop yields by county in the US using satellite images. We aim to improve upon and better understand [26]'s methodology and results. In line with their work, we use nine spectral and temperature bands from relatively low resolution satellite images as our features for predicting county-level corn and soybean yields. To ease training, we reduce the dimensionality of our data by assuming that the position of pixels doesn't impact the average yield (the permutation invariance assumption), which allows us to use pixel intensity histograms as features.

By making [26]'s model deeper, we achieve better prediction accuracy, showing that there is still signal to extract from the data. To better understand whether our models can distinguish between crops, we compute saliency maps for each image/crop pair and compare maps for various crops. We find that our model distinguishes between crops, and that, in line with previous yield prediction research, the infrared and temperature bands of images taken during peak growing season contribute the most to discrimination ability.

1. Introduction

Being able to predict crop yields accurately allows governments to plan the production, distribution, and consumption of food more effectively, combat food insecurity, and prepare for shortages and supply shocks well in advance. Historically, prediction of localized crop yields has been impossible on a large scale due to a dearth of sufficiently granular yet globally available predictive data. As a result, entities interested in predicting crop yield have relied on expensive and noisy agricultural censuses. However, the advent of remote sensing data collected regularly by satellites orbiting the globe and powerful image-processing techniques like Convolutional Neural Networks (CNNs) promises increased prediction coverage and even accuracy.

Remote sensing-based crop yield prediction at the US county level using CNNs has already been demonstrated by papers like "Deep Gaussian Process for Crop Yield Prediction Based on Remote Sensing Data" by You et. al. [26]. In the paper, You et. al. [26] propose several novel techniques to make prediction possible [26]. First, they reduce the dimensionality of their relatively small imagery dataset by using histograms of pixel values as features rather than raw pixel values. Second, the output of their CNN is fed into a deep gaussian process to account for spatial correlation of yield between counties [26]. With these techniques, the authors are able to drastically outperform existing prediction techniques based on remote sensing data. By adding layers to [26]'s vanilla CNN model, we were able to increase accuracy further even without including their Gaussian Process layer. Thus, we were able to show that there was still signal to be extracted from the satellite imagery dataset used by [26].

In addition to being a powerful prediction technology, CNNs have the potential to provide insight about the underlying mechanisms that drive real-world phenomena such as growth of different crops. By analyzing yield prediction models trained on historical yields of several different crops and their interactions with input data, we were able to show that such models are able to distinguish between pixels important for predicting specific types of crop. We corroborated both natural intuition and existing research about which factors are most important for predicting the yields of differing crops. In conducting the experiments described in this paper, we were able to demonstrate that analyzing trained CNN models could be used to understand the world, not just predict the future.

2. Related Work

2.1. Remote-Sensing-Based Crop Yield Prediction

While the paper by You et. al. [26] uses CNNs for crop prediction and forms the basis for our work, it is far from the first to attempt to predict crop yield via an easily-measurable proxy. Some of the most popular proxies are normalized-difference vegetation indices (NDVIs), which are positively correlated with crop yield [16]. As far back

as 1983, researchers such as J.L. Hatfield used vegetation indices from infrared and red wavelengths to predict potential yield (as given by genetics) and actual crop yield (as measured in a harvest) [20, 5].

In 2013, David Lobell published a paper in which he utilized MODIS satellite images to measure and analyze crop yield gaps (the difference between the potential yield and the actual yield). A key technique used by Lobell involved determining relationships between crop yields and vegetation indices computed from "remote sensing measurements of light at red and near-infrared (NIR) wavelengths" [15], indicating that some wavelengths of light (and thus bands) may be more significant for producing accurate yield predictions than others.

The simplest approach given by Lobell uses only vegetation indices, but these are not the only factors correlated with crop yield. A 2014 paper by David Johnson discovered that in addition to vegetation indices, "MODIS daytime land surface temperature was negatively correlated" to crop growth mid-summer [9], suggesting that prediction accuracy can be improved by incorporating other remotely-sensed measurements.

2.2. Visualizing Predictions

Once one obtains a model that attains high crop yield prediction accuracy, a commonly-created visualization is a nationwide map of predicted county-level crop yields [2]. However, such images provide little insight into how the model interacts with the raw satellite data to produce its predictions. As a result, we visualize data not on the country scale, but rather on the *county* scale.

One visualization for understanding which pixels of a raw image contribute most to a complex model's output is the saliency map, as introduced by Karen Simonyan et al. [21]. Here, a forward pass is performed through the model, and then the gradients of the output with respect to the *input data* (rather than the weights) are computed and plotted as an image. While saliency maps are relatively simple to generate, other such "importance" visualizations exist as well, such as occlusion maps (which measure error relative to the position of an occluded region of the image) [27]. However, saliency maps are relatively inexpensive to generate compared to some of these other visualizations, which is why we focus on them in our paper.

2.3. Other Applications of Remote-Sensing Data

Crop yield prediction is only one of many applications that were improved by the combination of satellite imagery with machine learning models. In more environmental applications, reseachers have been able to use remote-sensing data to detect oil spills [11], locate and assess the severity of forest fires [18], and survey penguin populations from space [3] even without deep convolutional neural networks. Satellite imaging and machine learning have inspired numerous social and economic applications as well, such as poverty detection [8], GDP estimation at sub-national scales [24], racial makeup of city populations [17], and urban sprawl measurement [23]. Given the effectiveness of such methods, predictions based on remote-sensing data will only become even more ubiquitous and profitable in the future.

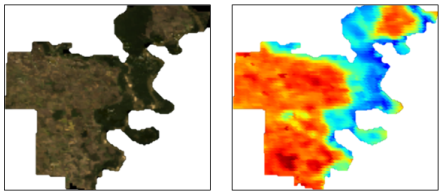

Figure 1. Shown above are raw visualizations of the input data corresponding to Desha County, AR. On the left is an RGB visualization of spectral data using the first, third, and fourth bands. On the right is a heatmap visualization of daytime temperature data.

3. Dataset and Features

3.1. Raw Data

The input to our training pipeline consisted of raw satellite images; we used remote sensing data collected by NASA's MODIS instrument, which is made publically available through Google Earch Engine [4]. MODIS images are collected 46 times per year at 500 meter resolution (i.e. each pixel in a MODIS image represents a region of 500 meters by 500 meters). Of these 46 times per year, we used a subset of 32 times that occur during the growing seasons for corn and soybeans.

The images we used had seven spectral bands (ranging from 459-2155 nm) [14] and two temperature bands (daytime and nighttime) [13] collected by the MODIS instrument. In addition, we used landcover masks to ignore regions within a county not corresponding to cropland [12]. Our final dataset included images from counties located in 11 different agriculturally-important states in the United States.

While MODIS's 500 meter resolution may initially seem quite coarse, the average size of a corn farm in Iowa is 349 acres [19], which equates to approximately 1.5×10^6 square meters. At 500 meter resolution, this means that each farm will actually consist of approximately four pixels.

As we were predicting crop yields at the county level, we used data from the National Agricultural Statistics Service (NASS) for the ground truth county-level crop yield corresponding to the various crops that we studied. In order to enable reuse of raw satellite image data between crops, we restricted ourselves to using yields of soybeans and corn, since they are grown in similar regions of the US. We used yield data and images from 2003 through 2012 as our train-

ing set, and data from 2013 as our validation set.

3.2. Features

As explained by You et al. [26], there were only approximately 7,500 usable county/year pairs in the training dataset we assembled. As a result, training on the raw images taken of each county over the course of each year (roughly 100 pixels tall $\times$ 100 pixels wide $\times$ 9 bands $\times$ 32 images/year $= 2,880,000$ pixels per county/year pair) would be somewhat infeasible, since the dimension of the raw features would be much larger than the size of our training set.

To avoid this dimensionality problem, we employed You et al.'s *permutation invariance* assumption, which states that each pixel's *value* determines its contribution to the county's yield, not its *location*. Intuitively, this assumption means that a farm is the same farm no matter where it is in relation to other farms, and thus produces the same yield regardless of where it is. Of course, it is possible that a model trained on raw images could pool information from neighboring counties to make more accurate predictions, but by invoking permutation invariance we assume there is little predictive signal to be extracted by doing so.

Leveraging this assumption, we a constructed 32-bucket histogram of pixel values for each band of every multispectral input image, thus allowing us to represent each image instead as a 32×9 histogram matrix. Since there were 32 images taken of each county every year, for a given county/year pair, we stacked the histogram matrices constructed from the 32 images taken over the course of the year to create a $32 \times 32 \times 9$ tensor representing the pixel values across all times and bands for that county and year. We then used these histogram tensors as our inputs to our models. In short, by assuming permutation invariance as You et al. [26] did, we were able to reduce the dimension of the inputs to our model from approximately 3 million features to only 9216 features, rendering model training much more feasible on a smaller dataset.

4. Methods

4.1. Crop Yield Prediction

For this project, we focused on predicting soybean and corn yield in 2013 by training on satellite images and yield data collected for counties in our 11 states of interest between 2003 and 2012.

4.2. Model Architectures

Figure 2 describes various architectures that we implemented for this project. The *reference* architecture is the one described in [26]. We also explored building deeper networks to determine if we could achieve higher accuracies than the original reference network, thus indicating that there was still signal to be extracted from the data. In the table, each layer CONV(c, f, s) represents a convoutional layer with c filters of size $f \times f$ with stride s, followed by a ReLU nonlinearity, a batch normalization layer, and a dropout layer with keep probability p. We denote the number of consecutive layers of a given type (i.e., CONV(c, f, s) by the respective number in the table. For example, in the reference architecture, the network has one CONV$(128, 3, 1)$ layer, followed by a CONV$(128, 3, 2)$ layer, etc. The final layer for all architectures that we experimented with was a fully connected layer of size 2048. We used batch normalization [7] and dropout layers [22] after each convolutional layer in order to reduce overfitting. For training, we used the Adam optimizer [10] to minimize the training loss. We used

$$L_2 = \frac{1}{2}\sum_{i=1}^{N}(\text{pred}_i - \text{real}_i)^2$$

as our training loss between our predicted yield and our real yield labels. We used

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\text{pred}_i - \text{real}_i)^2}$$

as our error metric on our validation set.

4.3. Crop Differentiation

In their paper, You et al. focused exclusively on crop yield prediction for soybeans, mostly "since it [was already] widely investigated in prior work" [26]. Given the relatively low resolution of the input images however, we wanted to determine whether models trained on these images were actually able to distinguish between farms that grow different types of crops, or whether the models were only able to identify which pixels were most likely to be farms producing any sort of crop at all. If these models were in fact able to meaningfully discriminate between crop types when making yield predictions, then we could be more confident that the model was not just taking advantage of circumstantial correlation between yields of common crops in the same counties and instead using meaningful differences in the characteristics of multispectral images that were most predictive of different crops' yields to make its predictions. As a result, we focused the rest of our work on characterizing the discriminability of corn and soybeans using the models above.

We chose corn as our alternative crop because it is an extremely common agricultural product in much of the US. In particular, its ubiquity means it is grown in many of the same counties as soybeans, allowing us to compare predictions of each crop for the same counties. All of the experiments described below were conducted by training two

	Reference [26]	Deep1	Deep2	Deep3	Deep4
CONV(128, 3, 1)	1	1	1	2	2
CONV(128, 3, 2)	1	1	1	1	1
CONV(256, 3, 1)	1	1	2	2	2
CONV(256, 3, 2)	1	1	1	1	1
CONV(512, 3, 1)	2	3	3	3	3
CONV(512, 3, 2)	1	1	1	1	1
CONV(1024, 3, 1)	1	1	1	1	1
FC(2048)	1	1	1	1	1

Figure 2. The CNN model architectures that we implemented to predict yield

models based on the reference model architecture described by You et al., a *soybean model* trained using the soybean yield for each county as histogram tensor labels and a *corn model* trained using the corn yield for each county as histogram tensor labels. We then compared attributes of the models themselves as well as their predictions to try to determine the degree to which the architecture proposed by You et al. is able to discriminate between crops.

4.3.1 Output Rescaling

As a first test of the differentiation ability of the model proposed by You et al., we attempted to predict corn yield in 2013 for each county by first standardizing the distribution of soybean yield predictions made using the soybean model to zero mean and unit variance using the mean and standard deviation of soybean yield in our validation set and then rescaling the distribution to the mean and variance of corn yield in our validation set, and vice versa. More formally, if $\hat{\mu}_{soy}$ and $\hat{\sigma}_{soy}$ are the estimated mean and variance of soybean yield in our validation set, $\hat{\mu}_{corn}$ and $\hat{\sigma}_{corn}$ are the mean and variance of corn yield in our validation set, and $\hat{y}_{soy}$ and $\hat{y}_{corn}$ are the predicted yields for corn and soybeans using their respective models, then the predictions made using this rescaling technique can be expressed as follows:

$$\tilde{y}_{corn} = \frac{\hat{y}_{soy} - \hat{\mu}_{soy}}{\hat{\sigma}_{soy}} \cdot \hat{\sigma}_{corn} + \hat{\mu}_{corn}$$

$$\tilde{y}_{soy} = \frac{\hat{y}_{corn} - \hat{\mu}_{corn}}{\hat{\sigma}_{corn}} \cdot \hat{\sigma}_{soy} + \hat{\mu}_{soy}$$

In some sense, this technique can be thought of as an extremely crude form of transfer learning. Instead of the final layer of the network being a fully-connected layer with bias of $0 \in \mathbb{R}^n$ and weight matrix I_n that acts as the identity transformation, it is instead "retrained" manually to have weight matrix $\frac{\hat{\sigma}_{soy}}{\hat{\sigma}_{corn}} I_n$ and bias $\left(\frac{\hat{\sigma}_{corn}\hat{\mu}_{soy}}{\hat{\sigma}_{soy}} + \hat{\mu}_{corn}\right) \cdot 1 \in \mathbb{R}^n$.

Having rescaled the outputs of the original models to the scale of the other crop, we then computed the RMSE of these predictions on the validation sets for the new crops and compared them to the RMSE of the original models. If there was a meaningful drop in performance between the original and rescaled models, then we could conlude that at least to some degree, the original architecture could learn to predict yield based on characteristics of the data unique to the crop whose yield it was trained to predict. If there was little decrease in performance, then there might be a high degree of similarity between characteristics of the two crops in the input data that confounded the architecture's ability to distinguish between them. However, similar prediction accuracies could also result from a particularly high degree of correlation between the yields of the two crops between counties, rather than a meaningful lack of discriminative ability, meaning other tests would be needed to confirm indistinguishability.

4.3.2 Saliency Maps

In addition to examining the transferrability of predictions from one crop to another, we wanted to determine which attributes of the data were the most predictive of soybean or corn yield specifically. In particular, we wanted to ask whether or not our models could identify which pixels from each histogram tensor were the most informative for predicting the yield of one crop *but not the other crop*. If the architecture learned meaningful differences in the importance of various pixels to yield prediction accuracy between crops, then we could say with some confidence that the architecture was able to distinguish between crops when learning to make predictions. To determine whether or not the model did in fact learn different representations for different crops, we computed saliency maps inspired by Zeiler et al.[27] for each image and yield prediction model. We then scaled the entries of the saliency maps for each model appropriately to account for different yield magnitudes between crops and compared the maps using several different distance metrics. Intuitively, if the differences between saliency maps were large, then it would be clear that the pixels that needed to change the least to positively or negatively impact loss ereifferent from crop to crop. If they were

small, then the pixels that needed to change the least to positively or negatively impact loss would not be meaningfully different from crop to crop.

More formally, if L_c is the L_2 loss function for crop c and $I_{iy} \in \mathbb{R}^{32\times32\times9}$ is the set of pixel histograms for county i in year y, the saliency map for featureset I_{iy} and crop c is defined as

$$S_{ciy} = \frac{\partial L_c}{\partial I_{iy}}$$

We then normalize each entry of S_{ciy} by its maximum magnitude entry as follows:

$$\tilde{S}_{ciy} = \frac{S_{ciy}}{\max_{jk\ell} \mid S_{ciy,\,jk\ell} \mid}$$

To compare the impacts of different pixels on the losses corresponding to two crops c_1 and c_2, we used three different distance metrics. First, we computed the RMSE of saliency map entries for crops c_1 and c_2 across counties $i = 1, \ldots, N$ in year y:

$$D_s(c_1, c_2, y) = \sqrt{\frac{1}{N} \sum_{i,j,k,\ell} \left(\tilde{S}_{c_1 iy,\,jk\ell} - \tilde{S}_{c_2 iy,\,jk\ell} \right)^2}$$

Next, we computed the L_1 norm of the difference between the saliency maps for crops c_1 and c_2 across counties $i = 1, \ldots, N$ in year y:

$$D_1(c_1, c_2, y) = \frac{1}{N} \sum_{i,j,k,\ell} \mid \left| S_{c_1 iy,\,jk\ell} - \tilde{S}_{c_2 iy,\,jk\ell} \right|$$

Finally, we computed the average percent difference between the saliency maps for crops c_1 and c_2 across counties $i = 1, \ldots, N$ in year y:

$$D_p(c_1, c_2, y) = \frac{1}{N} \sum_{i,j,k,\ell} \frac{\left| \tilde{S}_{c_1 iy,\,jk\ell} - \tilde{S}_{c_2 iy,\,jk\ell} \right|}{\left| \tilde{S}_{c_1 iy,\,jk\ell} \right|}$$

5. Results

5.1. Accuracy Improvements

You et. al. trained their data on various architectures including a CNN, a CNN with Gaussian Processes, an LSTM, and an LSTM with Gaussian Processes [26]. They found that the best results came from the CNN and the CNN with Gaussian Processes. For our project we used You et. al.'s base CNN architecture as our reference architecture and experimented with improving accuracy using deeper models.

In figure 3, we summarize the training and validation loss over the $25,000$ iterations for the reference architecture with different dropout keep probabilities. Note that we plotted the moving average of these losses as opposed to the actual data points in order to capture the general trend of the loss and eliminate some noise. We found that increasing our dropout keep probability to 0.5 caused overfitting. This is reflected by the divergence of the training and validation loss; training loss continues to decrease while validation loss flattens. On the other hand, we observed that the simple model ($p = 0.1$) didn't overfit since the training and validation loss are quite similar, plus or minus some noise.

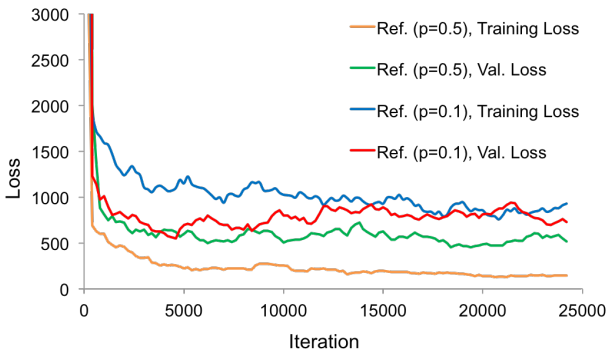

Figure 3. CNN model training and validation loss for different dropout parameter values

In figure 4, we summarize the validation accuracies for the year 2013 of the models we trained. For each of our experiments, we list the minimum achieved RMSE on the validation error during the $25,000$ training iterations. We also provide a unitless measurement of loss called "% Mean Yield," which measures what percent the RMSE for a given model and crop is of the true mean yield for that crop in 2013.

The Deep3 architecture performed the best of all models that we experimented with based on the 2013 validation set. Further, several of the deeper architectures performed better than the baseline from the original paper on the same training and validation sets. This suggested that there was still signal to be extracted from the data by making the architectures more complex.

5.2. The Linear Model and Histogram Sums

As one of our experiments, we trained a linear model that predicted yield for a given histogram tensor I_{iy} for crop c, county i, and year y as $\hat{y}_{ciy} = W_c I_{iy} + b_c$. The reason we tested such a simple model was to see whether or not predictions were summable, i.e. $\sum_{i,y} \hat{y}_{ciy} = W_c \sum_{i,y} I_{iy} + b_c$. Intuitively, if this relationship were true, then we could segment each raw satellite image into many small "patches," generate a histogram tensor of the same dimensions for each patch, and then predict yield for each patch separately, providing predictions at much finer geographies that would still be accurate when summed back up to the county level. However, the superior performance of deep models indicates that yield is likely a highly nonlinear function of his-

Architecture	Crop	Min RMSE	% Mean Yield
Baseline [26]	Soybean	5.50	12.10%
Ref. $p = 0.25$	Soybean	5.82	12.80%
Ref. $p = 0.5$	Soybean	5.37	11.81%
Ref. $p = 0.1$	Soybean	5.79	12.74%
Deep 1	Soybean	5.74	12.63%
Deep 2	Soybean	5.57	12.25%
Deep 3	Soybean	**5.24**	**11.53%**
Deep 4	Soybean	5.28	11.61%
Linear	Soybean	7.37	16.21%
Ref. $p = 0.25$	Corn	18.49	11.73%
Ref. $p = 0.50$	Corn	18.67	11.84%

Figure 4. 2013 validation set min RMSE and percent error of mean yield for various architectures. The dropout keep probability (p) is 0.25 unless otherwise specified.

togram counts, meaning it would be difficult to achieve high prediction accuracy on finer geographies without retraining a model specifically for those geographies.

5.3. Output Rescaling

After making predictions using the rescaled outputs, we found that the accuracy of the rescaled predictions was roughly comparable to the predictive performance of the original models. As shown in 5, the RMSE computed from predictions made by rescaling soybean yield predictions on 2013 data to corn yield scale was only 2.825% percent of average corn yield larger than the RMSE of the original model. More surprisingly, the predictions made by rescaling corn yield predictions to soybean yield scale marginally outperformed the original model's predictions; the RMSE for the rescaled predictions was 0.638% of average soy yield smaller than the RMSE for the predictions produced by the original model. The similarity of predicted yield distributions to the true yield distributions can be seen in 6; the real distributions of corn and soybean yield in the validation set (shown in green and blue, respectively), differ only slightly from the predicted distributions for those crops using the rescaling technique (shown in red and yellow), demonstrating the effectiveness of the rescaling technique.

Model	RMSE	% Mean Yield
Original Corn	19.90	12.62%
Rescaled Corn	24.35	15.44%
Original Soybean	5.94	13.07%
Rescaled Soybean	5.65	12.43%

Figure 5. The results of our output rescaling experiment. % Mean Yield is computed by dividing RMSE by the mean yield of the relevant crop in the validation set.

While we were tempted to conclude that the results of

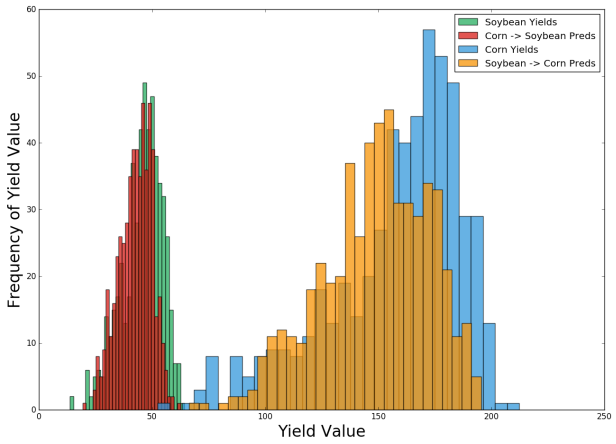

Figure 6. The real distributions of crop yields compared with the predicted yield distributions computed by rescaling predictions for one crop to predictions for the other crop. The rescaled predictions are quite similar to the true yield distributions.

this test point to a lack of discriminatory ability on the part of the reference architecture, we thought there were other explanations for why performance seemed unreasonably high. In particular, the correlation between corn and soybean yields (as measured by Pearson's correlation coefficient) is $\rho = 0.818$, meaning the two quantities are extremely collinear. It would make sense then that since counties that produce lots of corn also tend to produce lots of soybeans, simply changing the magnitude of a yield prediction for a different crop to match the scale of the crop yield actually being predicted would provide a prediction quite similar to the desired quantity.

Of course, this explanation doesn't completely explain why corn predictions rescaled to soybean yield scale outperformed the original soybean model. While we can't be sure that this explanation is sufficient, we hypothesize that rescaling the outputs acted as an extremely crude form of post-training regularization. Using the corn model to predict soybean yield introduced some random noise into predictions since corn and soy predictions are not perfectly correlated. Thus, if the soybean model overfit the training data at all, using this rescaling process instead would likely result in a more generalizable model. However, it is also possible that the validation set we selected (namely 2013 yield data) happened to yield corn predictions particularly similar to the set of true soybean yields.

5.4. Saliency Maps

5.4.1 Individual Maps

First, we computed the distances between saliency maps using the distance metrics defined above. The results are displayed in 7.

Distance Metric	Distance
D_s	0.141%
D_1	0.092
D_p	1151.93%

Figure 7. The distances between saliency maps computed using different distance metrics

While the D_s and D_1 distances appeared small, it turned out to be an illusion due to the fact that the entries of the normalized saliency maps were bounded between -1 and 1. In fact, the average percent distance D_p was over 1100%, indicating that there were actually very large differences between saliency maps on average. Thus, our results demonstrate that the model architecture was able to distinguish between pixels important for corn yield prediction and pixels important for soybean yield prediction.

Once a scaled saliency map $\tilde{S}_{ciy}$ had been computed for a given crop, county, and year, we found it useful to visually and qualitatively analyze it in the context of the original raw image $R_{iy} \in \mathbb{R}^{H \times W \times (32 \times 9)}$, rather than the histogram tensor I_{iy}. Letting $I_{iytb} \in \mathbb{R}^{32}$ be the slice of histogram I_{iy} corresponding to time t and band b and $R_{iytb} \in \mathbb{R}^{H \times W}$ be the slice of histogram R_{iy} corresponding to time t and band b, we constructed the saliency map visualization V_{ciytb} of the yield of crop c for county i in year y at time t and band b like so:

$$V_{ciytb} = \left[\left(\tilde{S}_{ciy,\,tb} \right)_{b(m,n)} \right]_{m,n=1,1}^{H,W}$$

where $b(m,n)$, the bucket corresponding to the pixel $(V_{ciytb})_{m,n}$ is given by:

$$b(m,n) = \left\lfloor \left((R_{iytb})_{m,n} - 1 \right) \cdot 32/4999 \right\rfloor$$

To aid with interpretation, we created visualizations that corresponded to the maps that minimized and maximized the distance D_s between $\tilde{S}_{1iy,\,tb}$ and $\tilde{S}_{2iy,\,tb}$) computed over i, y, t, b. Large negative gradients (which decrease loss and improve accuracy) are shown in white, while large positive gradients (which increase loss and hurt accuracy) are shown in black.

We found that while counties with the smallest saliency map differences had nearly identical visualizations for both crops (as expected), the counties with the largest saliency map differences displayed visible inversion, in which white pixels in the soybean saliency map visualization were black in the corn visualization, and vice-versa. This pattern indicated to us that the model was able to distinguish between soybeans and corn.

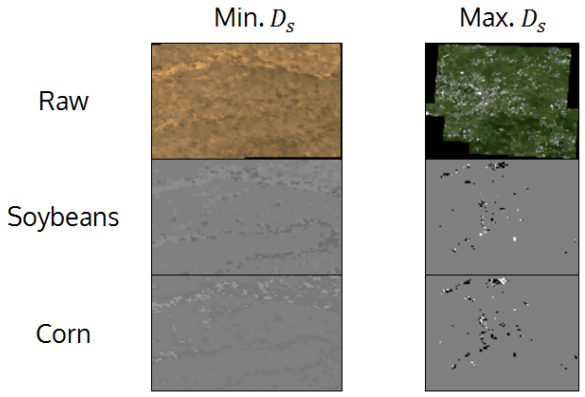

Figure 8. This figure visualizes the saliency maps with the smallest and largest differences. The images on the left depict Furnas County, NE, in 2013, at time slice 0 and band 3, while the images on the left depict Muskingum County, OH, in 2013, at time slice 20 and band 2.

5.4.2 Physical Factors Determining Distinguishability

Having discovered that the reference architecture was in fact capable of distinguishing between pixels important for corn yield vs. soybean yield with widely varying confidence (as demonstrated by the two example saliency maps provided above), we wanted to understand what physical phenomena might contribute to such differences in the distinguishability of the two crops. To do so, we hypothesized that distinguishability would vary across times of year and bands. To test whether or not there were differences in the informative power of time and bands in distinguishing between crops, we computed the average distance between saliency maps for each band and time of the year separately. Note that we omitted D_p distances from our plots below because they were not any more informative than the D_s and D_1 distances, and being orders of magnitude larger than the other distance metrics made them difficult to include on the same plots.

As can be seen in 9, images taken between May and October exhibited the largest differences in saliency maps between corn and soybeans, thus indicating that, somewhat intuitively, images taken from the middle to the end of the growing seasons for corn and soybeans were the most informative for distinguishing between different crops when predicting yield. This trend was qualitatively corroborated by the minimally and maximally different saliency maps in 8; the minimally different saliency maps were computed for an image taken around March, and the maximally different saliency maps were computed for an image taken around July. These results were also consistent with You et al.'s experiments [26], which show that including images taken later in the growing season in the set of features improves predictive accuracy in general; if images taken

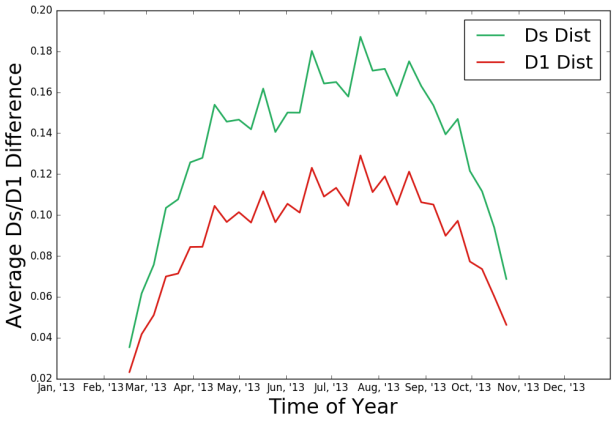

Figure 9. The D_s and D_1 distances between every county's saliency map computed for each time of the year separately

later in the growing season improve predictive power, they should likely also help in discriminating between crops.

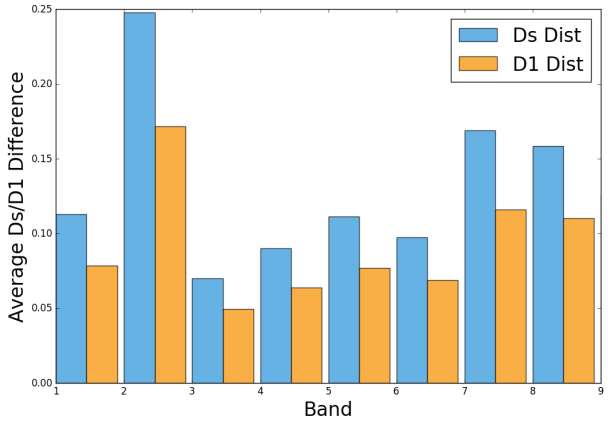

Figure 10. The D_s and D_1 distances between every county's saliency map computed for each band separately

We also investigated the relative impact of different bands on the architecture's ability to discriminate between crops by determining which bands had the largest differences in saliency maps across counties according to our distance metrics. As demonstrated in 10, bands 2, 8, and 9–the infrared band and the two temperature bands–had the largest differences in saliency maps across distance metrics. These results are in line with the findings of [15] and [9] which state that infrared and temperature are the most informative types of data for yield prediction models.

6. Future Work

In the future, the most important work we could do would be to test the efficacy of the permutation invariance assumption. In principle, one could build a convolutional model that was trained on the raw images themselves but had a similar number of parameters to the models we trained on histogram tensors. If we were to train a model on raw images and achieve higher accuracy than the models examined in this paper, we would be able to demonstrate that there was important signal to be extracted from the location of pixels within each image. Unfortunately, the images in our dataset were of varying sizes and aspect ratios, making it more difficult to construct a model since any architecture we used would have to be able to handle variable image sizes. Several ideas we had to circumvent such difficulties included making use of the Spatial Pyramidal Pooling layer developed by [6] or standardizing image sizes by padding them to be the size of the largest image and then passing in the original width and height of each image as additional features. Of course, if we were not able to achieve better results by training on raw images, we would not be able to rule out the importance of pixels' locations in providing accurate crop yield predictions. However, we would have demonstrated that avoiding the permutation invariance assumption does not provide a de facto advantage over models that do leverage the assumption.

7. Conclusion

In this paper, we demonstrated that it is possible to achieve better crop yield prediction accuracy using MODIS satellite imagery by employing more complex models. This finding indicates that there was still meaningful signal to be extracted from the data we collected. As was made clear by the experiments we conducted, the model architectures we studied were able to distinguish between pixels important for predicting corn yield and pixels important for predicting soybean yield. In particular, the infrared and temperature bands of images taken between May and October were the most informative features for differentiating between corn and soybean yield. Further, these findings matched what one might expect based on understanding the physical phenomena that underly crop growth. In some sense, we discovered that the architectures we studied were able to learn something about how interactions between real-world phenomena impact the growth of different crops. While these experiments were not empirical enough to determine whether the representations the network learns were actually based on physical processes as opposed to pure correlation, they point to a future in which neural networks are used not just for prediction, but also for building models that help us better understand important physical and social processes at the structural level.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia,

R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] D. K. Bolton and M. A. Friedl. Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. *Agricultural and Forest Meteorology*, 173:74–84, 2013. 2

[3] P. T. Fretwell, M. A. LaRue, P. Morin, G. L. Kooyman, B. Wienecke, N. Ratcliffe, A. J. Fox, A. H. Fleming, C. Porter, and P. N. Trathan. An emperor penguin population estimate: the first global, synoptic survey of a species from space. *PLoS One*, 7(4):e33751, 2012. 2

[4] Google Earth Engine Team. Google earth engine: A planetary-scale geo-spatial analysis platform. https://earthengine.google.com, 12 2015. 2

[5] J. Hatfield. Remote sensing estimators of potential and actual crop yield. *Remote Sensing of Environment*, 13(4):301–311, 1983. 2

[6] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *CoRR*, abs/1406.4729, 2014. 8

[7] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015. 3

[8] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016. 2

[9] D. M. Johnson. An assessment of pre-and within-season remotely sensed variables for forecasting corn and soybean yields in the united states. *Remote Sensing of Environment*, 141:116–128, 2014. 2, 8

[10] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[11] M. Kubat, R. C. Holte, and S. Matwin. Machine learning for the detection of oil spills in satellite radar images. *Machine learning*, 30(2-3):195–215, 1998. 2

[12] Land Processes Distributed Active Archive Center. Land cover type yearly l3 global 500 m sin grid. 2014. 2

[13] Land Processes Distributed Active Archive Center. Land surface temperature emissivity 8-day l3 global 1km. 2014. 2

[14] Land Processes Distributed Active Archive Center. Surface reflectance 8-day l3 global 500m. 2014. 2

[15] D. B. Lobell. The use of satellite data for crop yield gap analysis. *Field Crops Research*, 143:56–64, 2013. 2, 8

[16] B. Ma, L. M. Dwyer, C. Costa, E. R. Cober, and M. J. Morrison. Early prediction of soybean yield from canopy reflectance measurements. *Agronomy Journal*, 93(6):1227–1234, 2001. 1

[17] J. Maantay and A. Maroko. Mapping urban risk: Flood hazards, race, & environmental justice in new york. *Applied Geography*, 29(1):111–124, 2009. 2

[18] D. Mazzoni, L. Tong, D. Diner, Q. Li, and J. Logan. Using misr and modis data for detection and analysis of smoke plume injection heights over north america during summer 2004. In *AGU Fall Meeting Abstracts*, volume 1, page 0853, 2005. 2

[19] National Agriculture in the Classroom. A look at iowa agriculture. 2016. 2

[20] P. Schreinemachers. The (ir-) relevance of the crop yield gap concept to food security in developing countries. *With an application of Multi Agent Modeling to Farming Systems in Uganda, University of Bonn (forthcoming)*, 2005. 2

[21] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 2

[22] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014. 3

[23] P. C. Sutton. A scale-adjusted measure of urban sprawl using nighttime satellite imagery. *Remote sensing of environment*, 86(3):353–369, 2003. 2

[24] P. C. Sutton, C. D. Elvidge, and T. Ghosh. Estimation of gross domestic product at sub-national scales using nighttime satellite imagery. *International Journal of Ecological Economics & Statistics*, 8(S07):5–21, 2007. 2

[25] J. You. crop_yield_prediction. https://github.com/JiaxuanYou/crop_yield_prediction. 9

[26] J. You, X. Li, M. Low, D. Lobell, and S. Ermon. Deep gaussian process for crop yield prediction based on remote sensing data. *Association for the Advancement of Artificial Intelligence*, 2017. 1, 3, 4, 5, 6, 7

[27] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. *European Conference on Computer Vision*, 2014. 2, 4

8. Supplementary Material

The code that we wrote ourselves and repurposed from [25] can be found at https://github.com/brad-ross-35/crop-yield-prediction-project