Image Classification for Poverty Estimation

A case study on Rwanda

ST449 Artificial Intelligence and Deep Learning
LT 2019
Instructor: Milan Vojnovic
TA: Tianlin Xu

Author Name and Candidate Number:

Hongjin (Nicole) Lin
22454

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Introduction

Reliable data on local socio-economic characteristics is essential for policy-makers to design effective development programs and interventions. However, such data requires costly and logistically difficult surveys in terms of time and human resources. Developing countries that need such data the most tend to be those least capable of collecting them, such as countries in the African continent. According to the World Bank, 23 out of Sub-Saharan Africa's 48 countries conducted less than two household surveys from 2000 to 2010 (World Bank, 2016). With the rapid development of sensor technology, consistent and high-resolution satellite imagery increasingly becomes an alternative source of data for accurate estimation of geo-spatial and socio-economic characteristics of almost every part of the world.

Stanford's sustainability and artificial intelligence lab is one of the forerunners in applying image classification techniques on satellite imagery for better poverty prediction. Jean et al (2016) from the lab proposes a novel deep learning approach using Multi-Step Transfer Learning (MSTL) in Convolutional Neural Networks (CNNs) to extract socio-economic data from daytime satellite imagery. The rationale behind using MSTL rather than directly using CNNs on daytime satellite images is that labels for these images rely on existing surveys, which typically only contain information in the cluster level, resulting in training data size much smaller than those typically used in CNN applications. The MSTL approach uses nightlight luminosity, a noisy but readily available and sizable proxy for poverty, to train a fully convolutional CNN and use the features from the last layer to estimate wealth.

The MSTL approach is detailed in an earlier paper by Xie et al (2016), where the authors apply the method to Uganda. Evaluating the performance metrics of the approach against different image classification approaches, they conclude that the MSTL model outperforms the rest with an accuracy rate of 0.716.

	Survey	ImgNet	Lights	ImgNet +Lights	Transfer
Accuracy	0.754	0.686	0.526	0.683	**0.716**
F1 Score	0.552	0.398	0.448	0.400	**0.489**
Precision	0.450	0.340	0.298	0.338	**0.394**
Recall	0.722	0.492	0.914	0.506	0.658
AUC	0.776	0.690	0.719	0.700	**0.761**

Table 1: Performance of approaches in Xie et al (2016). The survey model uses logistic classifier with features provided in surveys directly. The ImgNet model uses the VGG F model trained on ImageNet directly. The Lights model uses logistic classifier with average light intensity, summary statistics, and histogram-based features for each area. The ImgNet+Lights model uses a concatenation of ImageNet features and nightlight features. The Transfer model is the multi-step transfer learning approach, which will be detailed later.

This report aims to evaluate the MSTL approach against three other classification approaches for a smaller country—Rwanda, a landlocked East African country with an area of 26,338 km² (Uganda has an area of 241,037 km²) as of 2010 (World Bank data). It explores four solutions: 1) estimating wealth with a CNN using daytime satellite images directly, 2) using VGG16 net trained on ImageNet as the base CNN model for daytime satellite images to estimate wealth (direct transfer learning), 3) using a ridge logistic model to estimate wealth using matched nightlight luminosity, and 4) the MSTL approach proposed by Xie et al (2016).

The goal of this report is to implement the approaches correctly and see if the same comparative performance holds for data of a different country, rather than improving the accuracy rate in Xie et al (2016).

Problem Setup

This report introduces the problem of poverty estimation as a binary classification problem.

Wealth labels are grouped into two classes: 0 for "poor" if below the national median and 1 for "rich" if above the national median.

Transfer Learning

Like in Xie et al (2016), this report follows the standard mathematical definition of Transfer Learning in Pan and Yang (2010). A *domain* $D = \{X, P(X)\}$ consists of a feature space X and a probability distribution $P(X)$. Given a domain, a *task* $T = \{Y, f(.)\}$ consists of a label space Y and a predictive function $f(.)$ which models $P(y|x)$ for $y \in Y$ and $x \in X$. Given a source domain D_S and learning task T_S, and a target domain D_T and learning task T_T, *transfer learning* seeks to improve the learning of the target predictive function $f_T(.)$ in T_T using the knowledge from D_S and T_S, where $D_S \neq D_T$ or $T_S \neq T_T$ or both.

The multi-step transfer learning we adopt here consists of two transfer learning problems. In the first transfer learning step, we have an object classification learning task $T_{ImageNet}$ as the source learning task on the source domain data $D_{ImageNet}$ consisted of images from ImageNet. The target learning task $T_{nightlight}$ is an image classification task to predict nightlight from source domain data $D_{satellite}$ consisted of daytime satellite images. Both learning tasks use CNNs. In the second transfer learning step, we have the $T_{nightlight}$ as the source learning task and using knowledge (features from the last convolutional layer) to classify wealth as the target task T_{wealth}. The first learning task uses CNN and the second learning task uses ridge classifier.

Data Description

There are three data sources: 1) Rwanda's poverty and GPS data, 2) nightlight luminosity, and 3) daytime satellite images, all of the same year 2010.

Both poverty and GPS data is collected from the Demographic and Health Surveys (DHS) conducted by USAID. Poverty data is contained in a STAT data file and GPS data is contained in shape files. Standard DHS provides household-level surveys about every 5 years. The surveys encompass many topics, including gender, education, and HIV. This report will focus on wealth and calculate wealth by taking the averages of assets of households (Demographic and Health Surveys, 2012). Although the survey data does not provide coordinate location of each household, it does group households into 492 clusters based on locations. Poverty data is therefore at the cluster level; images from the same cluster share the same wealth class.

Nightlight luminosity is collected from the National Oceanic and Atmospheric Administration (NOAA), which provides annual night images of the world with 1 km² resolution. The light intensity values are averaged for each year, ranging from 0 to 62 (as integers). The nightlight data is in raster format, presented as a dot matrix representing grids of pixels. Using GPS data from the DHS, we can match a location within the country with its corresponding nightlight intensity. In the MSTL approach, we use daytime satellite images to estimate nightlight intensity. Since there exists high class imbalance, this report groups images into two classes: 0 as "dark" if nightlight intensity is 0 and 1 as "bright" if nightlight intensity is from 1 to 62. However, the resulting classes remain highly imbalanced; the number of images in the dark class is about 19 times that in the bright class. To solve the

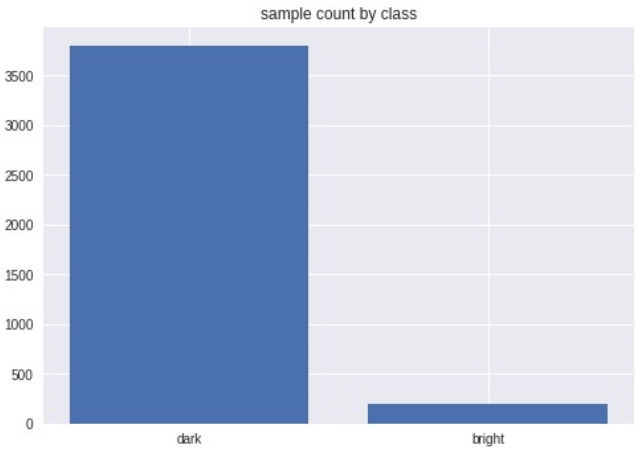

Figure 1: Class histogram based of nightlight intensity.

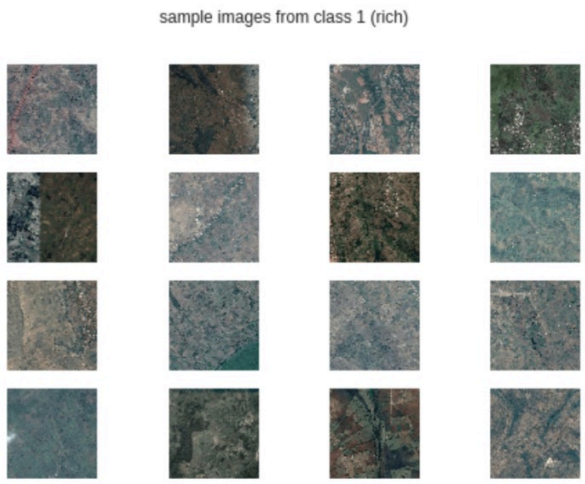

Figure 2: Sample daytime satellite images from class 1 (rich). The rich class seems to contain more roads, houses, and farmlands.

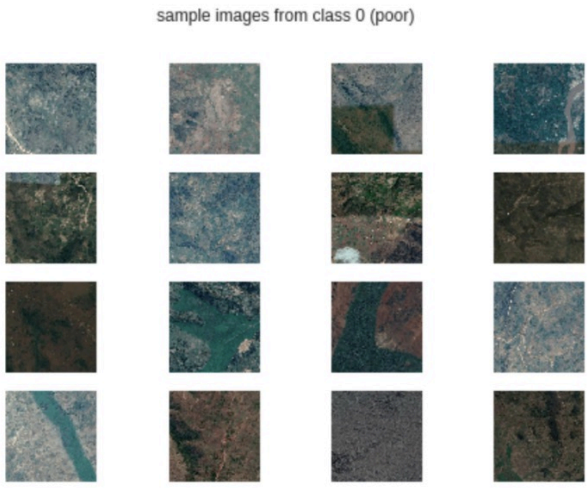

Figure 3: Sample daytime satellite images from class 0 (poor). The poor class seems to contain more forests, rivers, and mountains.

class imbalance problem, this report assigns 19 times larger weights to training images from the bright class in the training process.

Daytime satellite images are collected from Google Maps Static API and stored in files depending on their corresponding nightlight luminosity. Images are downloaded at a zoom level 16 (pixel resolution about 2.5m) of size 400 x 400, each covering 1 km² of area, similar in size to pixels in the nightlight data. There are then approximately 26,338 images in total (because Rwanda has an area of 26,338 km²). However, not all images have nightlight luminosity data available, resulting 22,204 images in total. After matching each image with nightlight luminosity and cluster-level wealth data based on its GPS information, we can see that we have more images for the poor class than the rich class. This report randomly select the same number of images from each class for further analysis. Furthermore, due to the limitation of Google drive storage space and Colab's computational power, this report randomly sample 2,000 images from each class, resulting in 4,000 images in total—2,800 for training and 1,200 for validation. This is a notable reduction of sample size. However, later results show that the relative performance of different solutions still holds with a smaller dataset. Since the goal of this report is to compare the different solutions proposed by Xie et al (2016), rather than outperforming their results, our dataset is sufficiently large. However, to compensate the sample loss, we will augment our data and generate more images during training.

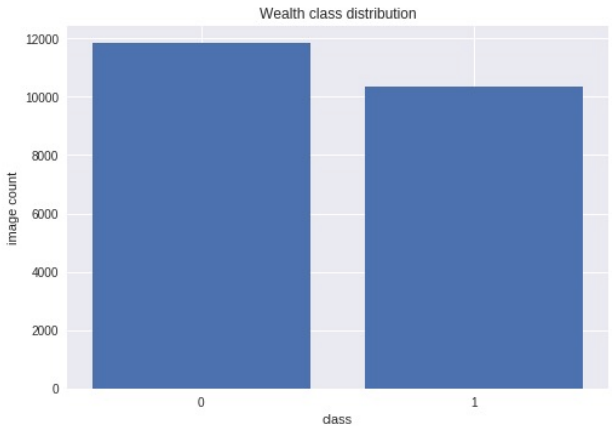

Figure 4: Class histogram based of wealth.

Approaches and Evaluation

CNN with daytime satellite images

We start with a direct implementation of CNN on daytime satellite images to predict wealth class as a baseline. The CNN model are trained using mini-batched gradient descent with a learning rate of 0.001 on 100 epochs. It consists of three convolutional layers, each with kernel size (3,3) and ReLU as the activation function and each followed by a

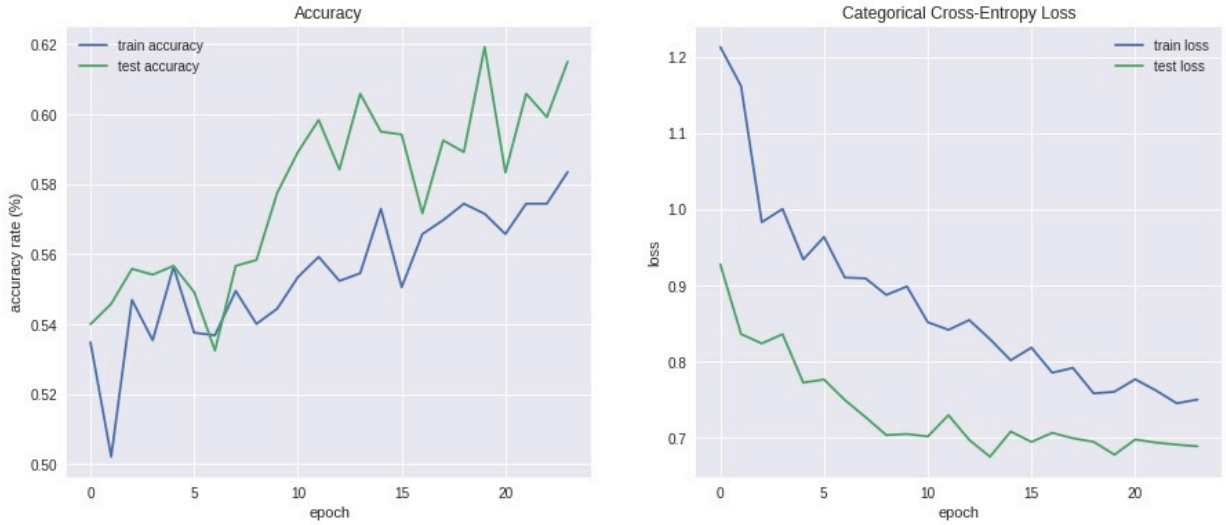

Figure 5: Training and validation accuracy and loss plots for baseline CNN model.

batch normalization step, average pooling layer with pooling size (2,2), and 25% dropout. Following the convolutional layers is a fully connected layer before outputting the two-dimensional prediction (one-hot representation of two wealth classes).

The benefit of using a stack of three 3x3 kernels is that it has the same receptive field as one 7x7 layer, but with more depth and non-linearities and fewer parameters per layer. Batch normalization has been proven to boost good validation accuracy, especially by ResNet (He et al. 2015). As the parameters of the previous layers change, the distribution of each layer's inputs also changes during training. This might cause vanishing gradients as it requires lower learning rates. Just as we want to normalize our data before training, normalization for each training mini-batch inputs can solve this problem. Moreover, both the pooling layers and 25% dropouts help avoid overfitting.

To adopt a similar architecture like the VGG16 net, which we will soon use for transfer learning. We randomly crop the 400 x 400 images to 224 x 224 (input size for VGG16 net) before training. To compensate sample loss in data preprocessing, we augment training images at each batch. More training images are generated from rotating, flipping, shifting and zooming existing data. We finally monitor the training process using an early stopping criteria. The model would stop training if validation loss hasn't decreased for 10 epochs. There are many other potential ways to improve the performance of the model. We are only concerned with the performance of a basic CNN model on the daytime satellite images in this approach.

Training stopped at around 30 epochs and produced a final validation prediction rate of 0.615. Note here that the training accuracy is lower than validation accuracy (training loss is higher than validation loss). This is because we have implemented data augmentation at each batch, adding variance to each training step while increasing validation accuracy. The dropout regularization layers in our model might also contribute to this pattern, because dropout is activated during training but deactivated during validation, generating higher variance in the training set.

Direct Transfer Leanring

We then apply transfer learning directly to daytime satellite images to predict wealth class. The source learning task $T_{ImageNet}$ is to classify objects on the source domain data $D_{ImageNet}$ consisted of images from ImageNet. The target task T_{wealth} is to classify wealth on the target domain data $D_{satellite}$ consisted of the

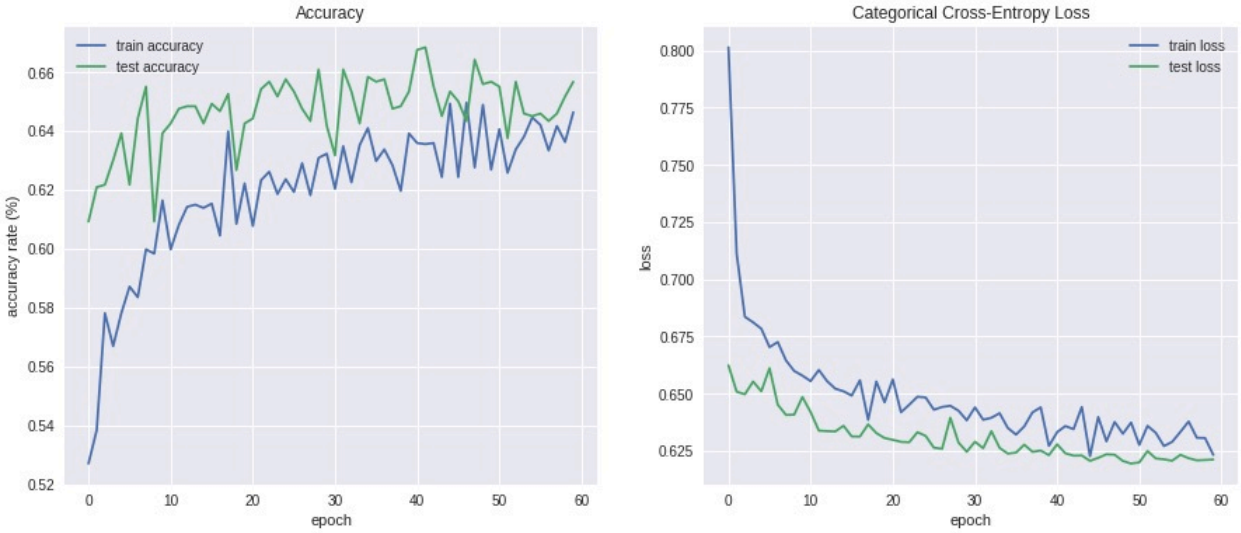

Figure 6: Training and validation accuracy and loss plots for the direct transfer learning model using VGG16 net.

Figure 7: Architecture of VGG16 net. Diagram adopted from lecture notes 04 (Vojnovic, 2019).

daytime satellite images. Many CNN architectures have achieved high performance in the first learning task, such as ResNet, VGG net, and GoogLeNet, all are recent winners of the annual large scale visual recognition challenge (LSVRC). This report chooses VGG16 (16 layers CNNs) for the task. VGG16 has 13 convolutional layers and three fully connected layers with five maximum pooling steps (Simonyan and Zisserman, 2014). It has less number of parameters than previous models with its use of three-layer (3,3) kernels, less complicated than GoogLeNet but outperforming it in many single-network classification tasks.

In this approach, we do not modify the network nor train any parameters in the convolutional layers, but to use the stored weights directly to estimate wealth class. Since the final layer of the pre-trained VGG16 makes prediction of 1000 dimensions, while our target is two dimensions (0 for poor and 1 for rich), we will only import the convolutional layers of VGG16 and stack a fully connected layer on top of it and make prediction.

Like in the first approach, this report uses data augmentation at every batch, monitors the training process and stops training if validation loss has not decreased for 10 epochs. The model was trained for about 60 epochs and produced an accuracy rate of 0.657, a significant improvement from the base line CNN model without using transfer learning. Note that the pattern of validation accuracy higher than training accuracy persists here because of data augmentation every batch and dropout steps in training.

Nightlight

To see how well nightlight intensity can estimate wealth class directly, we used a ridge logistic classifier with nightlight intensity as the only predictor ranging from 0 to 62 (as integers). The best model is found by tuning the parameter α, coefficient of the penalty term, using 10-fold cross validation. The best accuracy rate is 0.545, a very poor predictive performance. This shows that highlight luminosity is a noisy predictor of wealth. Because wealth data is at a cluster level, while high nightlight intensity tends to appear in a small area within a cluster, nightlight intensity is not the best proxy for local wealth. However,

it might provide some important features of urban area, as shown in the next approach.

Multi-Step Transfer Learning
Like we have formalized the problem in the problem setup section, MSTL is a two-step transfer learning problem. The first step is to use pre-train model on images from ImageNet to use satellite images to estimate highlight intensity class (0 for dark and 1 for bright). The second step is to use the features trained on the first step to estimate wealth class using a ridge logistic classifier.

In the first step, as we have seen from the direct transfer learning approach using satellite images, we can randomly crop the images of original size 400 x 400 to match the input size of VGG16 224 x 224. Although this is a reasonable approach, as the original model was trained by cropping a larger 256 x 256 image to 224 x 224 (Simonyan and Zisserman, 2014), Xie et al (2016) proposed a more compelling approach—a fully convolutional CNN (FCNN). FCNN allows multiple overlapping crops of the 400 x 400 image and compute each crop simultaneously. The outputs of each layer could be reused as the convolution slides the network over a larger input. Instead of a scalar output, the new output is a two-dimensional map of filter activations. In our case, the output is of size (2,2,4096).

This report uses the same training techniques as before, except that we weigh our training images by the class they belong to, as we have a significant class imbalance problem in this approach. The number of images from the dark class is 19 times that from the bright class. We therefore weigh images from the bright class 19 times higher than the majority dark class. After about 30 epochs of training, the model produced an accuracy rate of 0.944. The high accuracy rate is not surprising, as the two classes are imbalanced (F1 score is 0.471). Since this is just the intermediate step, we don't care too much about the accuracy; we want features that can serve as a good proxy for wealth for the next classification problem.

The next step is to use the features obtained from the last convolutional layer as variables to estimate wealth class using a ridge classifier. Since we have a large amount of features ($2 * 2 * 4096$), to avoid overfitting, we use PCA to select the top 100 most important features before training the final ridge model. Like we did in the model using nightlight directly, we use 10-fold cross validation to choose the best α, the coefficient of the penalty term. The best model outputs an accuracy rate of 0.673, F1 score of 0.652, and AUC of 0.708, all are the highest among the four different approaches we have explored.

	CNN	Direct	Nightlight	MSTL
Accuracy	0.615	0.657	0.545	**0.673**
F1	0.525	0.617	0.173	**0.652**
Recall	0.427	0.556	0.095	**0.615**
Precision	0.680	0.693	**0.905**	0.694
AUC	0.658	0.702	0.543	**0.708**

Table 2: Evaluation of approaches explored in this paper. Note that in Xie et al (2016), "poor" is interpreted as the "positive" class (1), while this report has done the opposite. This is why our precision and recall scores are flipped. High precision in the nightlight model can be explained by the large amount of samples with 0 nightlight intensity. More images are classified as "poor." In other words, we have more true positives.

The table shows that the multi-step transfer learning approach outperforms the rest approaches, which matches the results obtained by Xie et al (2016) using data for a different country.

Conclusion
This report has successfully replicated the results of Xie et al (2016) for Rwanda. Although the accuracy of the best model is still quite low, the prospects of image classification for development mapping is limitless as we continuously witness the incredible breakthroughs in computer vision. The recent

emergence of the powerful Generative Adversary Networks might be able to solve the problem of lack of labeled images in our task (Goodfellow et al. 2014). If we can generate realistic labeled data, it would boost our training performance significantly. Furthermore, as we can see from the ridge classifier using nightlight data, nightlight luminosity might not be the best proxy for wealth. Other alternative candidates, such as the number of roads or houses in each area, might be a better proxy that can capture the socio-economic features of a region for poverty estimation.

References

Demographic and Health Surveys (2012). Description of the Democratic and Health Surveys Data File, V. Accessed April 8th. https://dhsprogram.com/pubs/pdf/DHSG4/Recode5DHS_23August2012.pdf

Google Maps Statics API. Accessed April 12th. https://developers.google.com/maps/documentation/maps-static/intro

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., et al. (2014). Generative Adversarial Nets. *NIPS'14 Proceedings of the 27th International Conference on Neural Information Processing Systems*. V 2, Pages 2672-2680.

He, K., Zhang, X., Ren, Z., Sun, J. (2015) Deep Residual Learning for Image Recognition. arXiv:1512.03385[cs.CV].

Jean, N., Burke, M., Xie, M., et al. (2016). Combining satellite imagery and machine learning to predict poverty. *Science* 353 (6301), 790-794.

National Centers For Environmental Information (NOAA). Version 4 DMSP-OLS Nighttime Lights Time Series. Accessed April 12th. https://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html

Pan, S. and Yang, Q. (2010) A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*. Volume 22 Issue 10.

Sustainability and artificial intelligence lab. Stanford University. Accessed April 10th. http://sustain.stanford.edu/predicting-poverty

Simonyan, K. and Zisserman, A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs.CV].

World Bank (2016). While Poverty in Africa Has Declined, Number of Poor Has Increased. Online blog post. Accessed April 12th. http://www.worldbank.org/en/region/afr/publication/poverty-rising-africa-poverty-report

World Bank Data. Rwanda and Uganda profile. Accessed April 12th. https://databank.worldbank.org/data/views/reports/reportwidget.aspx?Report_Name=CountryProfile&Id=b450fd57&tbar=y&dd=y&inf=n&zm=n&country=RWA

Xie, M., Jean, N., Burke, M., et al. (2016). Transfer Learning from Deep Features for Remote Sensing and Poverty Mapping. *Association for the Advancement of Artificial Intelligence*. arXiv:1510.00098v2.

Appendix

Refer to the accompanying Jupiter Notebook for codes that generate the findings from this written report.