Temporal Poverty Prediction using Satellite Imagery

Derek Chen
Stanford University
Department of Computer Science
derekchen14@cs.stanford.edu

Abstract

In order to effectively alleviate poverty, the measurement and tracking of support initiatives over time is a necessary step for targeting aid efforts and guiding policy decisions. However, obtaining such data is time and labor intensive, so coverage of poverty stricken areas is often sparse or non-existent. Previous research has shown that a viable alternative for measuring poverty levels can be achieved through remote sensing methods. Specifically, satellite images processed through convolutional neural networks have shown promise in predicting the intensity of nighttime lights, which can then be used to gauge the underlying poverty level. This paper attempts to extend on past work by finding methods for measuring the change in poverty levels across different years using the same type of publicly available data. We are able to verify the original results of predicting poverty at a single point in time. More work is still needed to produce meaningful results in predicting temporal poverty.

1. Introduction

In the effort of poverty alleviation, enacting policies and sending supplies are necessary, but insufficient steps since poverty is a complex, multifaceted issue that requires sustained levels of long-term support to address. Thus, among many supplementary activities, measuring the effect of aid efforts over time is critical for determining whether interventions had their intended impact. Studying past events, researchers could pinpoint the moment a certain policy was introduced and compare the change in state of that location to determine whether resources were allocated efficiently. In turn, this knowledge could inform government officials who are interested in identifying the ideas which made the biggest contribution when deciding on new policies.

However, while poverty data is most needed in the developing world, it is also where the data scarcity issue is most pressing because of the high cost associated with conducting on the ground surveys. Therefore, this data gap is one of the crucial challenges to overcome in order to alleviate poverty in areas where help is needed most.

In recent years, deep learning approaches applied to large-scale datasets have revolutionized the field of computer vision, leading to substantial improvements in numerous tasks such as image classification [14], localization [16] and object detection [13]. In theory, modern approaches including Convolutional Neural Networks (CNNs) can also be applied to remote sensing imagery to extract socioeconomic factors that directly measure policy impact. On the one hand, tremendous amounts of satellite data is being captured every year that can serve as inputs into training such a network. On the other hand, a major obstacle still exists in that the target outputs needed to train the networks is precisely the information missing in the first place.

Fortunately, the emergence of novel data sources stemming from the explosion of information technology could help to close this gap [8]. In particular, maps of nighttime light intensity may serve as a proxy for measuring the poverty level of a town or village. Previous work has hinted at the possibility of using the combination of satellite imagery data and nighttime lights to produce features that are subsequently useful for predicting poverty [19]. This project extends on such work by adding a temporal factor, such that while the original task was to predict an asset index score given a satellite image of a village from a certain year, the updated task is to predict the change in asset scores between two time periods given two satellite images of the same village across different years.

2. Related Work

In order to circumvent the lack of ground truth economic measurements, a key insight to making progress is the use of transfer learning to pre-train the model to perform well on another task before incorporating survey results containing the final poverty scores. Pan and Yang establish a foundation for transfer learning and explore applications between different machine learning domains [12]. Following the explosion of deep learning, Oquab et al. apply transfer learning with CNNs to classify images in the PASCAL VOC by reusing layers trained on ImageNet [11].

While satellite images have been used in the past for predicting nighttime lights as a proxy for tracking growth and economic levels [17], only relatively recently have such images been analyzed with modern deep learning techniques. Using convolutional neural networks, Jean et al. are able to create models that capture 55 to 75% of the variation in average household asset wealth in the countries they examined [6]. Since then, other research has emerged which has evaluated the best strategy for exploiting the power of ConvNets in the context of remote sensing [10].

3. Data Sources

In order to gain a good understanding of the complete approach to evaluating a temporal shift in poverty, it helps to first examine a simplified pipeline and the data fed in at each stage. In the first phase of training, our network initializes its weights in the RGB channels using a model originally trained on ImageNet, which allows its first layers to immediately have strong capabilities in edge detection and similar coarse tasks [3]. In the second phase, the network is trained to predict nighttime light intensities from daytime imagery, simultaneously learning features that are useful for poverty prediction. In the last phase, an intermediate output of the network is transferred over to a separate regression model to predict the final wealth score. Even before entering the prediction pipeline, each data source must already undergo a series of pre-processing steps, which this section will now describe, as well as other attributes of interest.

3.1. Survey Data

The entire process starts and ends with survey data collected by the World Bank in the Living Standard Measurement Study (LSMS). This information resides as a number of CSV files for each country, where each file contains columns for location (latitude, longitude), years, and wealth scores. The years covered vary based on country, but the majority of data comes from panels conducted between 2009 and 2014. For each of those years, survey participants are asked a number of questions, with the queries of interest for us being those around assets (e.g. Do you own a motorcycle? Do you own a washing machine?). This is used as a way of measuring poverty by translating the responses into a normalized wealth score.

Roughly 3000 households are surveyed per country, but the usable information is more limited because the location of each household is listed as the latitude and longitude of its village. In all likelihood, we would have averaged the household data together anyway to reduce the noise coming from any single family, and because the resolution of satellite images fit multiple houses into a single pixel, but as it stands we are forced to reduce our count of examples by a factor of $\sim$6 due to an element beyond our control.

Band	Description	Wavelength (micrometers)	Spatial Resolution
B1	Blue	0.45-0.52	30m
B2	Green	0.52-0.60	30m
B3	Red	0.63-0.69	30m
B4	Near Infrared (NIR)	0.77-0.90	30m
B5	Shortwave Infrared 1	1.55-1.75	30m
B6	Thermal	10.4-12.5	1800m
B7	Shortwave Infrared 1	2.09-2.35	30m
B8	Panchromatic	0.52-0.90	15m

Table 1. Landsat-7 ETM+ Spectral Bands and attributes

3.2. Training Data

Landsat-7 is managed by the U.S. Geological Survey (USGS) [18], and provides images for every year dating from 1999 to 2016. The satellite circles the globe and complete a full orbit in roughly 16 days, which implies the existence of a rich and plentiful data source that can be used for historical analysis and future research in bridging the poverty data gap 1. Since Landsat-7 revisits the same area multiple times a year, a composite image is created by averaging the numerous snapshots in order to get a single representation of the time period. Furthermore, masking is applied to minimize the impact of clouds and tiling is performed on overlapping portions of the image in order to arrive at the final raw satellite mosaic, which is giant TIFF file covering 10 bands of information and whose view spans hundreds of kilometers.

Out of the ten bands, one band in particular specifies pixels containing NaNs where information is missing, possibly because the location experienced cloud coverage the entire year. Therefore, we impute those pixels with 0 values to prevent the network from breaking when encountering those areas. After accounting for a second thermal band, this leaves the eight major bands used for analysis, including but not limited to RGB channels, as described in more detail in the bands table 1. As another pre-processing step, we pan-sharpen the multi-spectral bands with the panchromatic band to create hyperspectral images that ideally contain the high spatial resolution of the panchromatic image while while maintaining the high spectral resolution of the multi-spectral image.

3.3. Intermediate Proxy

Generally speaking, nightlights can be used a rough proxy for wealth in that brighter cities tend to be wealthier. Thus, a CNN trained on Landsat-7 images to predict nightlight values should theoretically by primed to predict wealth scores using the same inputs. Our nightlights data comes from the NGDC Earth Observation Group (EOG) [9] and serve as a measure of the brightness of a city

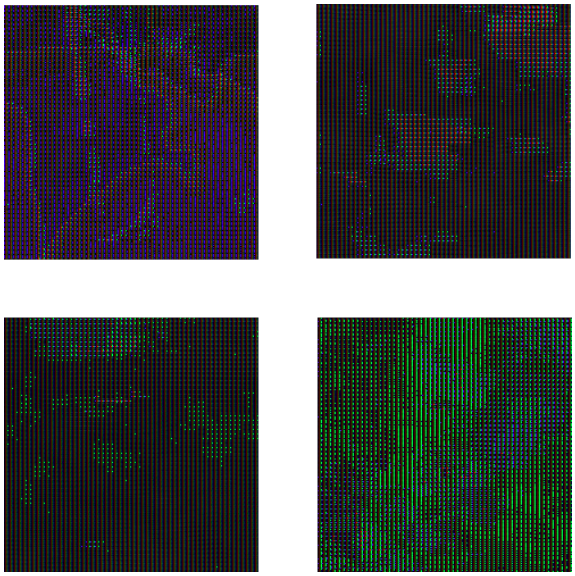

Figure 1. Samples of random locations chosen from Tanzania (upper left), Nigeria (upper right) and Malawi (lower right) in 2013 for RBG-bands only. Sample of location for Uganda (lower left) chosen from 2012.

as night. Specifically, We use data from Defense Meteorological Satellite Program (DMSP), which provides images at a 30 arc-second resolution ($\sim$1.0 km).

Similar to the Landsat-7 data, we start with a data source that has averaged the nighttime light intensity values across the 12-month period. The resulting values range from 0-63 where higher values represent brighter areas. Following the observations from [19], we discretize the values into three buckets to simplify training. Values from 0 to 8 are considered "dark", values of 8 to 35 are considered "medium", and the remaining values from 35 to 63 are considered "bright". Examples of each category can be seen in Figure 4. While there exist cities in Africa that have high luminosity, these areas are exceedingly rare. As seen in 2 and 3, Mwanza and Gulu represent two of the largest cities in Tanzania and Uganda, respectively, yet have nightlight values that would only place them in the "medium" category. Out of 848 samples in Uganda, only 235 have values that go beyond the "dark" rating. In fact, roughly two-thirds of cities have nightlight values that are actually 0, causing the signal in the data to be quite sparse.

3.4. Alternate Sources

Since we are studying the change in poverty over time, we necessarily had to use partially outdated training data in order to line up with the available data in the original LSMS survey. For example, suppose the goal was to predict the poverty level for a single year for village in Uganda. Then, one could use LSMS survey data from 2014, Landsat-8

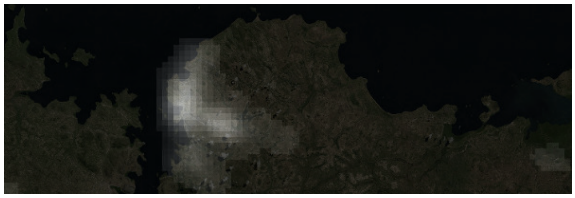

Figure 2. Visualization of nightlight values in Mwanza, Tanzania in 2013 overlaid on top of city map. *Avg pixel: 28*

Figure 3. Visualization of nightlight values in Gulu, Uganda in 2013 overlaid on top of city map. *Avg pixel: 19*

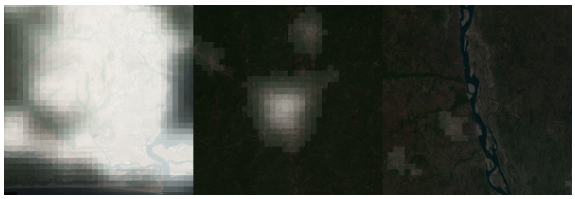

Figure 4. Nightlight images of (left) Lagos, (center) Ondo and (right) Adaigba, all cities in Nigeria. *Avg pixel: 63, 21, 7*

satellite data and Visible Infrared Imaging Radiometer Suite (VIIRs) nightlight scores, both of which have 2x higher resolution over their older counterparts. In contrast, in order to compare to a previous period, our research also requires data from the most recent survey of Uganda before 2014, which happens to come from 2012. However, as alluded to earlier, Landsat-8 satellite was launched in 2013, so it is unable to provide any data for 2012. Thus, a major challenge in reproducing the results of previous research is the noticeably lower resolution of the viable datasets.

With that said, new LSMS surveys are actively being conducted that could allow for the comparison across time periods using more modern data sources. Additionally, alternate data sources exist which may serve as ground truth labels for representing poverty. The Rural Economic and Demographic Survey (REDS) from the National Council of Applied Economic Research covers hundreds of households in India across a number of years. Additionally, the Indonesian Family Life Survey (IFLS) conducted by RAND includes five waves of panels ranging between 1993 and 2016. As such, any of the first three components of the prediction pipeline can and should be swapped out for newer data whenever any becomes available 5.

4. Approach

Continuing the high level example described in the previous section, suppose a wealth score has already been generated for a certain location. Going beyond past work, this project also adds a time component such that there are now two images, perhaps one from 2011 and one from 2013, which have ground truth wealth scores of 2.4 and 2.8, respectively. Then, the model would ideally output a prediction of 0.4 representing a growth in wealth and drop in poverty. This section will describe the end-to-end pipeline for producing this final difference in wealth scores.

4.1. Data Augmentation

Based on the village (lat,lon) coordinates provided by the survey, one key method of data augmentation is sampling around those areas to generate additional data for feeding into the network. Training data is sampled from a 7x7 grid around the main location, yielding roughly 10,000 examples per country per year. Validation data is sampled from a 3x3 surrounding grid, and yields about 500 examples. Finally, testing data is also sampled from a 3x3 grid around the main location, and results in about 200 examples per country. Recall that the raw satellite image lives in one large file, so the next step is extracting 224x224 pixel images based on the locations determined above. Given those images, mirroring is performed in both horizontal and vertical directions, which is allowed because unlike flipping a picture of a cat upside down, the aerial view of a city lacks such a sense of direction. A subset of images are also generated by randomly jiggling the brightness and/or contrast of the satellite images, resulting in a five fold increase over the original amount.

With satellite images at hand, corresponding nightlight values are retrieved by looking up the coordinates in the nightlight image and translating the corresponding pixel value into the appropriate brightness class. All of this is packaged into a tf-record file, the preferred data format for the Tensorflow framework, ready to be fed into the two networks for training. Note that at this point, there are two groups of train/val/test datasets – one for training a network representing the "before" time period and another dataset for training a separate network representing the "after" time period.

4.2. Architecture

While previous research utilized a pre-trained VGG-net for predicting nightlights, since the time of publication, open-source versions of ResNet have been released along with weights that have also been pre-trained on ImageNet. While many earlier convolutional networks used a simple linear structure of layers, Residual Networks (ResNet) take advantage of shortcut connections between blocks in the sequential layers [4]. These skip connections allow the loss gradient to flow through deep networks largely unimpeded during backpropagation, improving the viability of very deep networks. Residual networks also make liberal use of (spatial) batch normalization layers [5], which accelerate training by addressing shifting input distributions to layers by forcing features to conform to a zero-mean, unit variance distribution before being passed into the next layer. Means and variances are calculated across batches while training, and a running average is kept for later use during the inference phase.

Balancing the desire for more depth against the practical concerns of training time, a ResNet 50-layer model was chosen over because it was the smallest network size that still contained bottleneck connections. Most residual blocks followed a structure of [1x1 Conv-BN-ReLU, 3x3 Conv-BN-ReLU, 1x1 Conv-BN] at which point the data is merged with the identity connection. With three brightness classes to predict, the model was trained using a traditional softmax with cross-entropy loss function. Initial experiments tried a small handful of optimization schemes, with Adam always performing near the top, so it was chosen as the optimizer for all future trials.

4.3. Feature Extraction

At this point, it would be reasonable to assume that the next step is to save the weights and the replace the final prediction layers with a different loss to directly predict wealth scores, but there exist two main reasons to avoid going down this path. First, because the final wealth scores are continuous in nature, a classification model expecting discrete values would not be able to be directly applied. More importantly, the limited number of wealth scores means the network would quickly overfit when properly tuned.

Thus, given two trained and fine-tuned networks, we instead perform feature extraction by first passing in the test data into the network for inference. Once a full session has run, the activations from the penultimate fully-connected layer are compacted so all the rows referencing the same village are averaged together, resulting in a $(N, 4096)$ numpy vector where N equals the number of villages in the original testing data. This process is performed twice such that there is one vector encoding the knowledge of "before" and another vector representing the "after" time period.

4.4. Generating Scores

Each vector representation is now fit onto a ridge regression model to predict the final wealth scores for each time period. Compared to ordinary least squares, ridge regression adds a degree of bias to reduce the standard errors.

$$\underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i^\top \theta)^2 + \lambda \|\theta\|_2^2$$

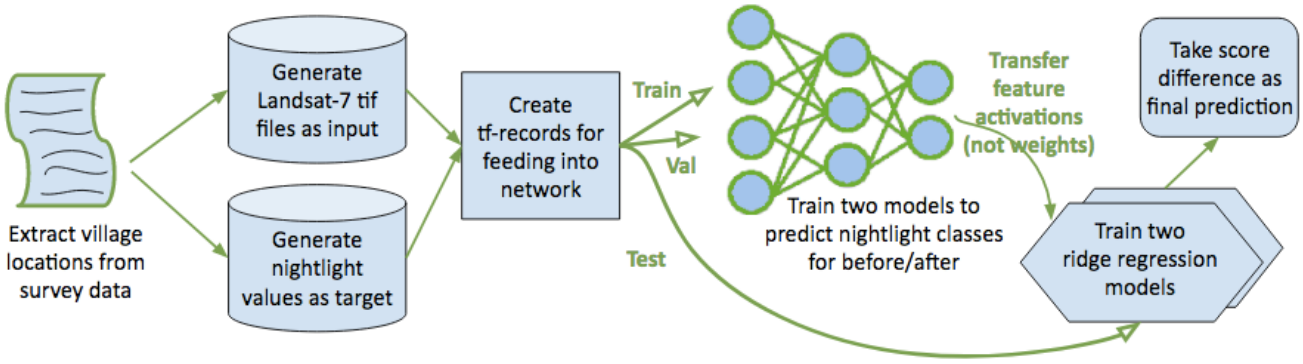

Figure 5. End-to-end pipeline for making predictions of wealth score differences.

L2 penalty was chosen over L1 as found in LASSO since there was no reason to believe sparsity of weights would lead to improvement in prediction. This belief held up empirically as both forms were tested and ridge regression performed consistently better. Hyperparameter λ was chosen through 5-fold cross validation. With two wealth scores at hand, the finishing step is taking the difference between the two to get the final prediction of the change in poverty. For evaluation, a Pearson R^2 correlation value was calculated comparing the predicted difference in wealth scores against the actual difference in wealth scores.

5. Experiment

In order to achieve our highest accuracy, the pipeline described above was tweaked such that data was fed in through a number of methods. Namely, data was grouped together in different ways to match the "before" and "after" periods from which we can take a change in poverty scores. As it turned out, the major bottleneck in the process was creating tf-records, clocking in at an average of 8 hours to generate the training file and 3 hours each to generate the validation and testing files. In comparison, running on a Tesla K40c GPU, training took roughly 4 hours to reach convergence, running for 30 epochs (roughly 5000 timesteps) per network.

5.1. Baseline

While the proposed pipeline follows many of the same steps as the original research [6], the data used throughout training comes from slightly different sources. Thus, as an initial step to verify the validity of the transfer learning stage, preliminary work was done to measure the correlation between LSMS wealth scores and nightlight values for Tanzania, Malawi, Nigeria, and Uganda 6. We found that most countries had a R^2 value between 0.45 and 0.65, which were close to, but not quite matching the original results of 0.55 to 0.75. Thus, given that our data was also

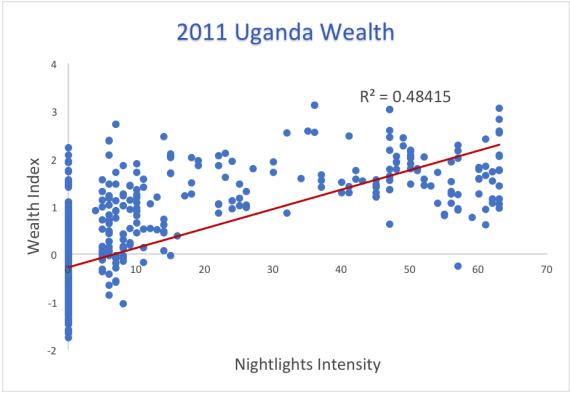

Figure 6. Nightlights correlation with wealth index score for Uganda in 2011.

strictly lower in resolution, a reasonable goal would be to come close to reproducing the original results for single year predictions.

However, when evaluating the correlation of difference in wealth scores against the difference in nightlight values, we found most countries had a R^2 value between -0.03 and 0.02, which is to say that we found no correlation whatsoever. In an attempt to ameliorate the issue, we then tried correlation with just an indicator variable, where the value is considered 1 when the difference in nightlight values is greater than 10 and 0 otherwise. We found that this improved R^2 scores to roughly the 0.35 range, but also skewed the data since very few villages experienced such a large jump in nightlight intensity during the two year span from 2011 to 2013. With our foundation set, we now move onto training methods involving deep learning techniques.

5.2. Single Year, Single State

The most straightforward idea for creating predictions reflects the approach outlined in earlier sections whereby a model is trained for each year to get two wealth scores,

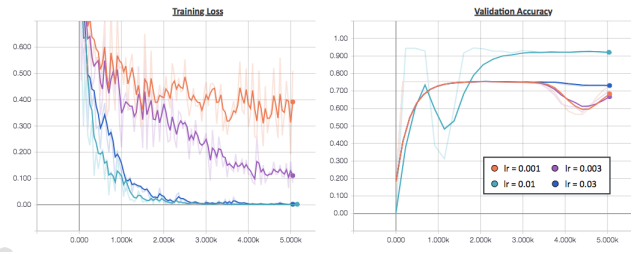

Figure 7. Training loss and validation accuracy at various learning rates for Tanzania 2011.

Learning Rate			
Tanzania - 0.001	68.4%	Nigeria - 0.001	62.0%
Tanzania - 0.003	66.8%	Nigeria - 0.003	65.2%
Tanzania - 0.01	91.7%	Nigeria - 0.01	78.4%
Tanzania - 0.03	73.1%	Nigeria - 0.03	76.4%
Multi-spectral vs. RGB-Only			
Tanzania - All bands	91.7%	Tanzania - RGB only	88.7%
Uganda - All bands	74.6%	Uganda - RGB only	70.2%
Batch Size			
Tanzania - 75	82.5%	Tanzania - 115	90.2%
Tanzania - 85	85.6%	Nigeria - 95	78.4%
Tanzania - 95	91.7%	Nigeria - 115	78.9%

Figure 8. Validation accuracy for specified parameters at the end of training. Most were trained to 30 epochs, although some were stopped early at 15-20 epochs if it did not appear able to catch up to previous trials.

and then a difference is taken between the two scores as the final prediction for the change in poverty. The first set of trials involves experimenting with the observed country. Concretely, Tanzania, Uganda, and Nigeria were chosen because their nightlight scores had the highest correlation with the underlying wealth data. The specific time periods were chosen because those were the most recent years LSMS survey data was readily available.

After resolving numerous training bugs, we arrive at our results which show Uganda with the highest reported R^2 value for a single year at 0.852 and a Tanzania holding the highest R^2 value for a difference score at 0.172 2. We currently hypothesize that this is due to the fact that Tanzania nightlights had the highest correlation to their wealth scores of the three chosen countries. While Uganda did offer the most data available for training, the best score was somewhat of an anomaly and perhaps the second trial of 0.513 is a more accurate measure, which still leaves Uganda as the best single year performer.

5.3. Parameter Tuning

Before moving onto mixing data sources, various hyperparameters and options were tuned on single year, single state models to maximize performance gains and minimize time needed to train more complicated models. The initial learning rate of 0.00001 was far too low, and it was not until we raised this up to 0.001 until we saw converging results. At this point, the learning rate was tripled each time to find the optimal setting for Tanzania, which ended at $\alpha = 0.01$. We repeated this process with Nigeria and found similar results, so this learning rate was used for all future runs.

Generally speaking, larger data means longer training time, so we also wanted to validate the need for the extra five multi-spectral bands beyond the RGB channels. Processing the tf-records already required splitting the channels into two parts since loading the ImageNet weights only applied to RGB bands. After running tests on Tanzania and Uganda, the validation accuracy was higher by roughly 5% on the full spectrum tests, so all bands were used moving forward. Next, having larger batch sizes typically improves the stability of training, which occasionally leads to faster training time. With that said, the default batch size of 128 seemed to cause memory overflow issues. This was lowered to 64 and then up by increments of ten to 75, 85, 95, 105 and 115. A batch size of 95 seemed to give the best performance without leading to any memory warnings being ejected.

Given time constraints, not all possible combinations of parameters were tested to completion. Ideally, we would have also tuned the decays rates and regularization strengths. Results for experiments with at least 15 epochs are included in Table 8 below.

5.4. Aggregated Methods

Note that in the outlined approach, the before and after time periods were left purposely unspecified because the major experiments involves trying different aggregation methods in an attempt to increase the amount of data used for training. Rather than using wealth scores from a single year and single country, we now group together the data in multiple ways to simulate different effects.

5.4.1 Aggregate by Time

Grouping by location involves putting together all time periods shared the same location. Concretely, the ”before” time period includes 2009 and 2011 data from Tanzania grouped together into one dataset to train a single ResNet. Separately, the ”after” time period includes 2011 and 2013 data from Tanzania to train a second ResNet. The same process was repeated for Uganda with results shown in the table above 2. The results are generally mixed in this category with Tanzania performing generally better, but not by any large margins. Tuning the model increased validation accuracy, but unfortunately did not do much to change the correlation between the predicted vs. actual poverty scores. Additionally, we note that there is a drop-off in correlation as we go from single year predictions to making a prediction of the the difference in poverty.

Trial	Before	After	Difference
Single year, Single Model			
2.1	0.434	0.143	0.172
2.2	0.404	0.513	0.170
2.3	0.208	0.174	0.038
Aggregate by Time			
3.1	0.342	0.489	0.218
3.2	0.236	0.350	0.236
Aggregate by Location			
4.1	0.170	0.114	0.135
Combine Everything			
5.1	0.086	0.024	0.089
Stacked Image			
6.1	0.002	0.003	< 0.001

Table 2. Major results for R-squared values. Before and After represents the value gained from individual models. Difference is the score we are attempting to ultimately predict.

5.4.2 Aggregate By Location

Grouping by time involves aggregating all the different locations that have data from the same years. In this case, the "before" time period includes 2011 data from Tanzania, Nigeria, and Malawi, and the "after" time period includes the 2013 data from those same countries. Uganda is unable to participate because its panel surveys were conducted in 2012 and 2014. Compared to previous runs, we notice that results are quite mixed since the difference prediction ends up doing better than the single year prediction for the "after" period of 2013 when grouping the countries together. The performance has actually slightly dropped off, which means that we are actually seeing worse results as we aggregate more data together. We suspect that the information from different countries is pulling the network into unfriendly optimization terrain, and that the network performs well in validation only by overfitting on limited training data.

5.4.3 Combine All

The final aggregation idea is to throw every data source we have at the problem to see if that will move the needle. This effectively uses data from all years and all countries, but unfortunately continued the trend that more data led to worse results. We suspect this occurred due to clashes in data signals. Observing the graphs showed choppy behavior where validation accuracy would jump around in an unsteady manner. This partially may have been the way we were feeding the data into the network, and that the data from each country/year combination was being grouped together, rather than properly shuffled. Overall, this method still requires more tuning to find the root underlying cause of underperformance.

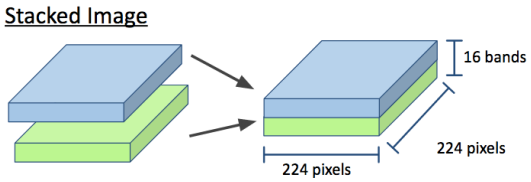

Figure 9. Representation of how the images are stacked together.

5.5. Stacked Images

The previously outlined methods basically repeat the process described in the original paper twice with slight tweaks in data sources to arrive at poverty scores. In that sense, none of these methods seriously tackle the temporal nature of the problem since they are calculating the difference in wealth scores only by subtracting one from the other at the very end. One idea that does take time into account is to train a single network on a paired image input and paired value output, which is what we try for our final experiment.

More specifically, the network consumes a 224x224x16 stacked satellite image composed of two 224x224x8 images to predict a difference in nightlight values. Since the nightlights are now calculated as a difference, there is the possibility of negative scores, so the original method of three brightness classes do not apply. Additionally, the change in nightlight intensities were much smaller than the original values we were dealing with before. After analysis of range of values available, we decided that creating bins where values from the initial time period to the next either (a) dropped in value (b) had a value from 0 to 3 or (c) gained in value by greater than 3. Since most cities did not experience a change, keeping the range in Bucket B limited was needed to get enough examples for the other categories.

Based on the success of the "Aggregation by Time" method, we trained our Stacked Images network using the same data sources. However, the performance of the stacked images ended up doing quite poorly with final correlation for a difference in scores registering at below 0.001 R^2. Part of the reason is the lack of time for tuning hyperparameters beyond what worked well for the individual models, which would probably be our highest value task given more time.

6. Conclusion

We found that the "Aggregation by Time" trials generally performed the best out of all the various methods tested. Additionally, our best parameters included a learning rate at 0.01 and batch size at 95. Overall, the major surprise was that while more data didn't really hurt validation loss or accuracy, it also did not magically make the correlation improve. In fact, we witnessed a drop in the actual measurement of poverty scores when we combined all the data

available together to train a single pair of models. We speculate that the diversity of the data, especially in the change in satellite imagery from one year to the next, caused issues that the network could not overcome.

Although, we were able to produce similar cross-sectional results, the ability to predict a difference in poverty levels across time is still lacking. We suspect this is largely due to the fact that the proxy for measuring a change in poverty does not correlate well with the final task, which is a critical flaw since luminosity has shown strong evidence for economic statistics [1], but at varying degrees depending on the task at hand [7]. Thus, alternate proxies for poverty, such as measures of agricultural productivity through crop yields or output per worker [15], could hold the key to finding effective methods for calculating changes in wealth over time. As briefly mentioned, having more time to tune hyperparameters would probably be our first task moving forward. Given the time to find newer options, using a different data source that is either higher resolution would also be a obvious move.

Another direction to take involves implementing models that replace the linear regression portion with a component that directly takes into account difference in wealth as the target. For example, at the risk of overfitting, rather than feature extraction, we could swap out the last layer for a linear SVM and continue to fine-tune the network [10]. Alternatively, a Siamese network taking in "before" and "after" images as its two halves could effectively enhance the aggregation activities performed in our experiments while simultaneously removing the need for an extra processing step [2]. Finally, along the lines of image captioning models, we could feed in features generated by a CNN as inputs into a RNN [20], with the belief that survey data spanning multiple time periods are more likely to take advantage of the time-series nature of recurrent neural networks.

Once an end-to-end pipeline can reliably measure a change in economic activity, we will apply the network to filling in gaps in knowledge where survey data is missing completely. Furthermore, as modern deep learning matures as a discipline, it can be worthwhile to consider how such sophisticated computational approaches can be applied outside of academic settings. To that extent, we are documenting progress so that others who may lack advanced technological mastery can still take advantage of the similar methods. We believe would be a meaningful step towards the overarching goal of helping policymakers gain a better understanding of their real-world problems in order make more informed decisions.

7. Acknowledgements

This project would not have been possible without the support and resources of the Sustain AI Lab at Stanford University. In particular, Stefano Ermon and Marshall Burke offered amazing guidance, and Anthony Perez was instrumental in moving the project forward. Thanks also goes out to George Azzari and Matt Davis for providing access to clean data.

References

[1] X. Chen and W. D. Nordhaus. Using luminosity data as a proxy for economic statistics. *Proceedings of the National Academy of Sciences*, 108(21):8589–8594, 2011.

[2] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 539–546. IEEE, 2005.

[3] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *Icml*, volume 32, pages 647–655, 2014.

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[5] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[6] 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.

[7] C. Mellander, J. Lobo, K. Stolarick, and Z. Matheson. Nighttime light data: A good proxy measure for economic activity? *PloS one*, 10(10):e0139779, 2015.

[8] K. Murthy, M. Shearn, B. D. Smiley, A. H. Chau, J. Levine, and D. Robinson. Skysat-1: very high-resolution imagery from a small satellite. In *SPIE Remote Sensing*, pages 92411E–92411E. International Society for Optics and Photonics, 2014.

[9] NOAA. National geophysical data center: F18 and f17 nighttime lights composites, 2014. Data for all years ranging from 2009 to 2013 are explored.

[10] K. Nogueira, O. A. Penatti, and J. A. dos Santos. Towards better exploiting convolutional neural networks for remote sensing scene classification. *Pattern Recognition*, 61:539–556, 2017.

[11] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724, 2014.

[12] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[13] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

[14] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[15] K. Schneider and M. K. Gugerty. Agricultural productivity and poverty reduction: Linkages and pathways. *Libraries Test Journal*, 1(1):56–74, 2011.

[16] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[17] 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.

[18] USGS. Remote sensing missions component of the us geological survey land remote sensing program, 2014. Data for all years ranging from 2009 to 2013 are explored.

[19] M. Xie, N. Jean, M. Burke, D. Lobell, and S. Ermon. Transfer learning from deep features for remote sensing and poverty mapping. *arXiv preprint arXiv:1510.00098*, 2015.

[20] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015.