Utilizing Latent Embeddings of Wikipedia Articles to Predict Poverty

Evan Sheehan
CS Department
Stanford University
esheehan@stanford.edu

Zaid Nabulsi
CS Department
Stanford University
znabulsi@stanford.edu

Chenlin Meng
CS Department
Stanford University
chenlin@stanford.edu

Abstract

In this paper, we propose a novel method for the task of poverty prediction, utilizing natural language processing on latent embeddings of Wikipedia articles, as well as satellite imagery, to predict poverty for geographical regions, providing an alternative to on-the-ground surveys and nightlights estimation. We demonstrate there are latent traits in the articles which correlate strongly with poverty. This framework can be deployed across the globe and provides a successful and intriguing link between latent textual embeddings and socioeconomic applications.

1 Introduction

Elimination of extreme poverty is one of the foremost UN Sustainable Development Goals [1]. However, in order to assess progress made on this front, traditional methods of poverty prediction and estimation revolve around the utilization of laborious and expensive on-the-ground surveys. While these surveys are often quite accurate and precise, it is highly impractical and costly to increase their temporal frequency, and many regions of the globe lack the infrastructure to even conduct them. As such, estimating the level of need for desperate regions of the globe and distributing resources accordingly requires the ability to assess extreme poverty and other lifestyle metrics without the use of these on-the-ground surveys. Although this is conventionally done using the nightlight satellite imagery of a region, we utilize Wikipedia [2] articles as research into the corpus and distribution of these articles has suggested they could possibly contribute to socioeconomic factor prediction [16]. By providing detailed textual information about locations and entities in a region deemed important enough by the open-source crowds to document, we view the articles as data rich proxies representing the area around them. As such, though the article types are diverse (eg. dam, town, company) and often do not contain explicit information on wealth, in aggregate we demonstrate they possess a confluence of latent features which are robust predictors of poverty.

In this paper, we utilize latent embeddings of Wikipedia articles, as well as satellite imagery, to predict poverty levels of regions. In short, the input to our model consists of the embeddings of the k-closest geolocated Wikipedia articles to the coordinate of interest (and for our final model, a nightlight satellite image histogram of the region). The output is a poverty prediction (a continuous number between -2 and 2). We use a variety of different methods to accomplish this task, before arriving at a model that outperforms the current state-of-the-art results. Lastly, we provide a study of article embedding activations and what features the model is learning in order to better understand what is happening.

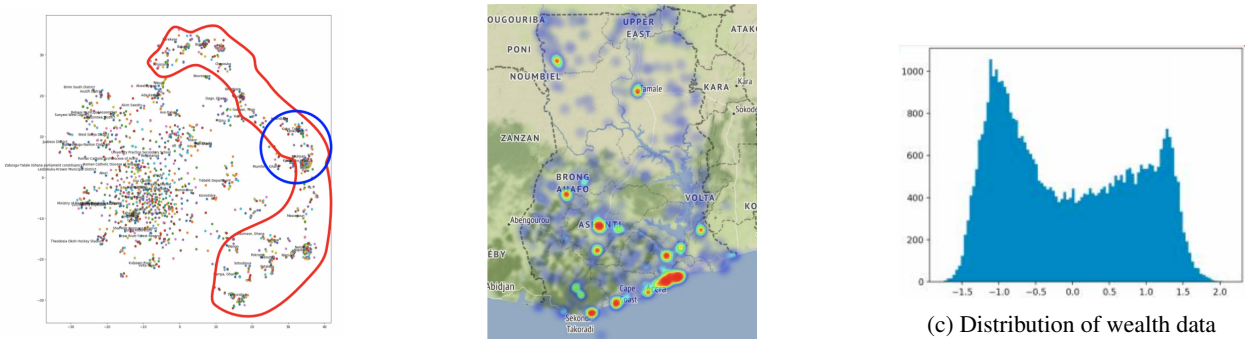
2 Related Work

Previous attempts to predict poverty in geographical regions have been limited to the use of satellite imagery. [7] uses nightlight satellite imagery to predict wealth levels, and [13] utilizes multispectral satellite imagery with an increased number of channels to improve upon [7]'s results. [19] also utilizes features derived from satellite imagery to extract features for poverty prediction. To the best of our knowledge, there hasn't heretofore been any work done on the extraction and utilization of textual features for poverty prediction purposes, though [16] proposes that geolocated Wikipedia articles could prove a useful source for such features. Representation learning is often used to learn mappings to latent domains, and [8] engages in this for satellite imagery, providing a framework for understanding traditional approaches to the task, while [10] and [12] validate this from a textual perspective.

Preprint. In review, CS229.

3 Problem Definition and Dataset

In order to utilize latent Wikipedia articles to predict poverty levels around the world, we first obtained a corpus of Wikipedia articles from the June 2018 dump, parsed it into its constituent articles, and extracted all those which were geolocated. This process netted over 1 million geolocated articles. We then trained a Doc2Vec model [11] on this text corpus to create a system for obtaining standardized article embeddings, which embedded each article into a vector $V \in R^{300}$. As a sanity check, we examined the Tsne embeddings of articles from various regions to verify that the model was learning useful similarities (Figure 1). For our groundtruth wealth data, we obtained data regarding the wealth index, or poverty level, at over 44,000 coordinates



(a) Tsne of Ghanaian articles.

(b) Wealth coordinate map for Ghana

(c) Distribution of wealth data

Figure 1: Left: Tsne of a subset of Ghanaian articles, with a cluster of rural village articles noted; Tsne clusters similar embeddings together; Middle: Heatmap of wealth coordinates for Ghana, with red being wealthier and blue being poorer. Right: Distribution of groundtruth wealth data after scaling (range between -2 and 2.)

across Africa from the Stanford sustainability laboratory, with each point's value scaled between -2 and 2, higher numbers indicating greater wealth. For our Multi-Modal model, we also utilized nightlight VIIRS images [5], obtained from [13] and covering 5km by 5km regions with 224 x 224 pixels.

To train our models, we specifically focused on five countries: Ghana, Malawi, Nigeria, Tanzania, and Uganda. These are five of the most common countries for evaluating new model performance on and thus allow us to compare our model to benchmarks in the literature. For each of these countries, we trained a model on 80% of the points within it, used the other 20% as our dev set, and then evaluated the model on the other four countries. This was repeated for each of the countries. This approach is motivated by the fact that our nearest article and nightlights images have high spatial dimensions, so primarily evaluating across national boundaries guarantees no train and test contamination. Also, this cross-boundary evaluation strategy and train/dev/test split is common practice for the space and is observed in much of the literature in this task [7] [13].

4 Approaches

4.1 Baselines

Doc2Vec SVM Regression: For a baseline, we created a simple support vector machine regression [3]. Specifically, for each coordinate we make a prediction for, we obtained the ten (hyperparameter) closest geolocated Wikipedia articles to the point. Next, we obtained a 300-dimensional vector embedding of each of the articles through Doc2Vec [17] (see Dataset for more information), and averaged all the vectors to get one, 300-dimensional vector. We then passed that vector through a support vector regression, using the following loss:

$$L = \begin{cases} 0 & |y - \hat{y}| < \epsilon \\ |y - \hat{y}| - \epsilon & otherwise \end{cases} \tag{1}$$

Support vector regression uses the same principles as SVM classification, where input spaces are mapped into higher dimensional spaces using kernels, essentially converting non-separable problems to separable problems. [15]. However, in the regression problem, a margin of tolerance ϵ, as in Equation 1, is set, since there are infinite possible outputs. This is called a "soft margin", while $\hat{y}$ is the kernel output of a weight vector and an input. [18]

Averaging Doc2Vec Neural Network: For our second baseline, we designed a simple, feed-forward neural network, where our input is the same as that for our SVM baseline, but now, we pass the embedding through a multi-layer perceptron to get a poverty prediction. We train the network with mean squared error for our loss, given by:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{2}$$

Neural Networks work by passing the input through a series of matrix multiplications, in order to learn a complex relationship between the input features and the output. These networks minimize the loss, and update weights through a process called gradient descent, finding the optimal set of weights for the task by backpropagating the derivative with respect to each of the weights. [14]

4.2 Models

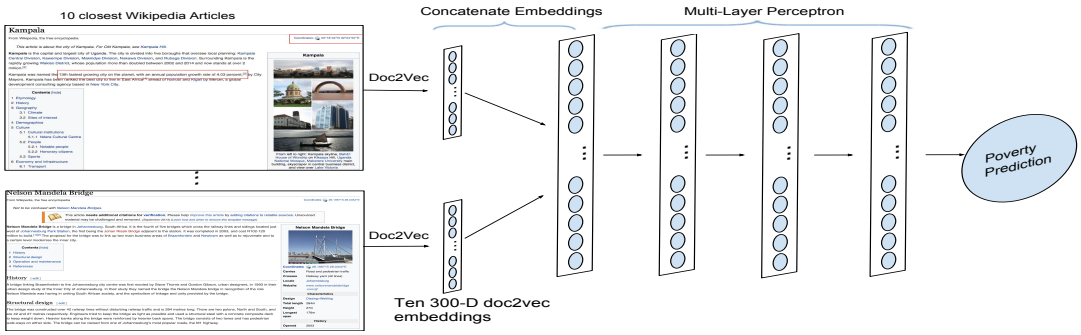


Figure 2: Diagram of our Wikipedia embedding model. The ten (hyperparameter) closest geolocated Wikipedia articles are obtained and the Doc2Vec embeddings are generated and concatenated into one large 3000-dimensional vector. We then append 10 nodes to that vector, representing the distance between each doc and the point of interest.

Wikipedia Embedding MLP: Following our baselines, we proceed with a similar, slightly more advanced model to help improve on our results. From our baselines, we observe that averaging the document embeddings results in a non-trivial loss of information as ten vectors are cut down to a single one of the same size. Thus, we decided to concatenate the document embeddings instead of averaging them so as to allow our model to utilize all the information, resulting in one large, 3000-dimensional vector. Furthermore, the distance from the articles to the point of interest is an important factor to consider, as the further away the article is, the less influential it should be. Thus, we also append the distance from each article to the point of interest to our vector, resulting in our final, 3010-dimensional vector that is input to our neural network. Note that since this neural network has more inputs than our second baseline, it required a more complicated architecture, as discussed under the Experiments section. We train this network similarly to our second baseline, with mean squared error loss (Equation 2).

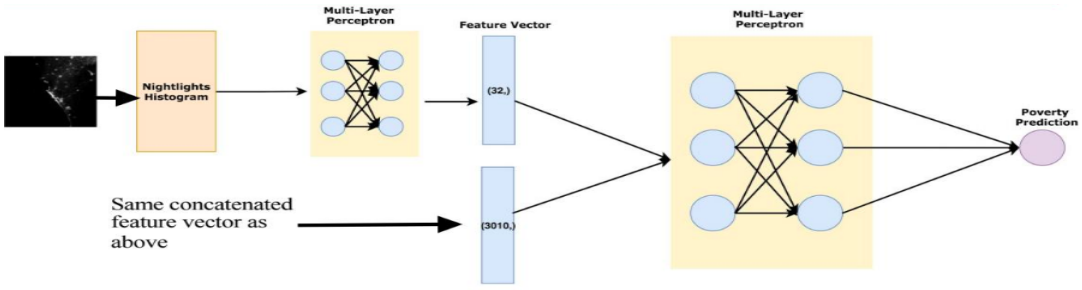
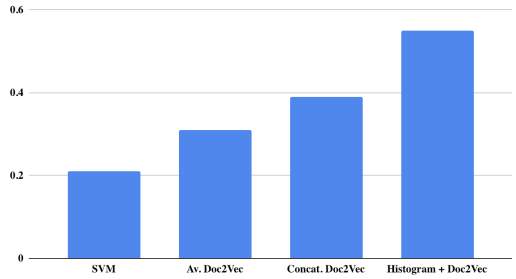


Figure 3: Diagram of our Multi-Modal Model, which utilizes the same idea as our Wikipedia Embedding MLP, but also concatenates in the nighlights histogram as input to a more complicated network.

Multi-Modal Model: Following our previous model, we decided that in addition to using Wikipedia articles for the task of poverty prediction, we could also use the satellite nightlights imagery of the region of interest. It has been shown in current state-of-the-art methods for poverty prediction, as discussed previously, that nighttime satellite images of the region of interest are quintessential. Thus, we decided to extend our model to use the image as an input in addition to our Wikipedia documents as this will provide our model with more information, and thus, potential for better results.



	UGANDA	TANZANIA	NIGERIA	GHANA	MALAWI
UGANDA	-	0.725	0.704	0.636	0.722
TANZANIA	0.554	-	0.569	0.513	0.558
NIGERIA	0.421	0.539	-	0.457	0.573
GHANA	0.505	0.521	0.501	-	0.544
MALAWI	0.413	0.549	0.475	0.392	-

Figure 4: Left: Comparison of average cross-boundary performance for all models; Right: Multi-Modal Model Results - Trained on column country, tested on row; metric - pearson's r^2.

We utilize the same input as our Wikipedia Embedding MLP, for the bottom part of our network, as depicted in Figure 3. In the top part of our network, we first obtain a histogram of pixel values from the nightlights image $V \in R^{256}$, and we pass it through several dense layers with ReLU activations, and obtain a 32-dimensional tensor. We then concatenate this tensor with our original 3010-dimensional Wikipedia Embedding feature, and feed it through a series of fully connected layers before a regression prediction is made. We train using mean squared error loss (Equation 2).

5 Experiments and Results

To evaluate our models, the primary metric we use is the squared Person correlation coefficient [6], which is a measure of correlation between predicted and observed values. We use this as it tells us how well what we are predicting matches groundtruth, and this metric is widely used in literature for this task.

We observed a steady improvement in r^2 values as we progressed from our baselines to our final architectures (Figure 4 left, Figure 5), as expected, since our models became more complicated. When transitioning from the Averaged MLP to the Concatenated MLP, we altered the number of units in each layer from 256 to 512, and increased the number of hidden layers from 3 to 5, to take into account the more complicated input set. Additionally, we iterated through different types of activation functions before settling on Leaky ReLU [20], which performed best. This is because it avoided the dying neuron problem frequently caused by the ReLU [4], where neurons are never activated due to the gradient of the ReLU being 0 at every negative number. We found that the Adam optimizer [9] with learning rate 0.001 worked best, as it combines the advantages of momentum and RMSprop. Finally, we settled on batch size of 32, which is standard for this task, and cleanly fit in memory.

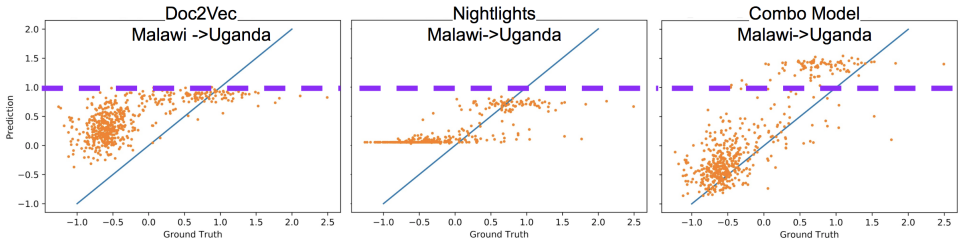
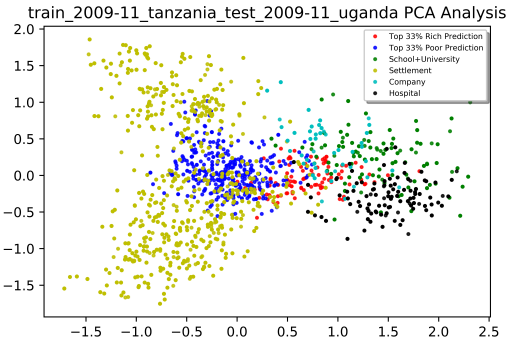


Figure 5: Model results trained on Malawi and tested on Uganda. The leftmost graph shows results for a model trained only on Doc2Vec, the centermost shows results for a model trained only on the nighlights imagery, and the rightmost shows results for our Multi-Modal Model, which combines both inputs. As is clear, the combination of inputs yields the best results.

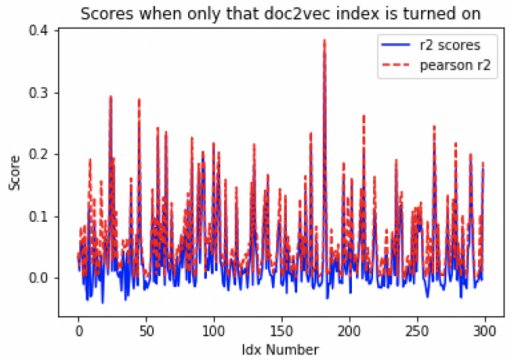
These changes, combined with a more detailed input, boosted results and began to yield r^2's challenging the state-of-the-art without using any visual input. Finally, after combining this with nightlights imagery in the Multi-Modal Model, we attained state-of-the-art results, beating [7] by over 10% for nearly all models. These results suggest that the latent article features are indeed robust predictors of poverty which allow networks to perform the task better than previously possible. In our final model, based on our results as shown on the right side of Figure 4, we see that training on certain countries yields better results than training on certain other countries, indicating a slight overfitting problem, as the model is not able to generalize to all countries with the same outcome. However, in the future, this can be mitigated with more data and training on multiple countries at once.

For error analysis, we observe that often the points which the model receives the highest MSE for are ones which come from areas where points with disparate ground truth values are tightly packed together, such as urban areas where there are slums and wealthy places in close proximity to each other. This makes sense, since the spatial resolution of our input (5km x 5km image)

does not allow us to discern features at a higher resolution than its dimensions (ie. any two points within 5km of each other often share much of the same image and thus receive similar predictions; this is a common issue with the nightlights approach).



(a) Wikipedia Embeddings PCA Analysis.



(b) Masked Activation Results (all other indices set to 0), with pearson's r^2 for each index shown.

Figure 6

To help us better understand Wikipedia article embeddings, especially the components that play key roles in poverty prediction, we performed PCA analysis on the first two principle components of Wikipedia article embeddings. More specifically, we picked the article embeddings of the richest 33% and the poorest 33% points in the Ugandan test set and compared their PCA projections to the projections of 'school', 'university', 'hospital', 'company', and 'settlement' articles that appeared in the test set (see Figure 6a). We observed that the average Doc2Vec embeddings for wealthier places tend to cluster with 'school', 'university', 'hospital', and 'company' article embeddings, while poorer places were more related to 'settlement' articles.

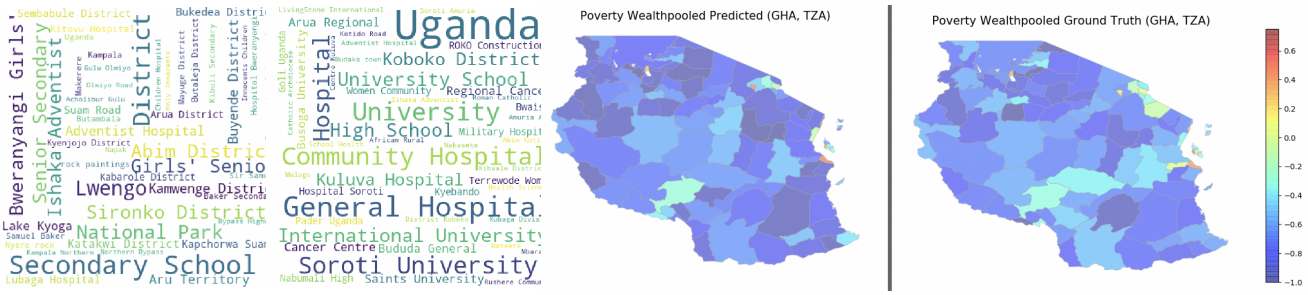


Figure 7: Left: Titles with largest values at indices in embedding most predictive of wealth level (24, 182; larger words indicate larger values); Right: Admin 2 level prediction vs ground truth for model trained on Ghana, evaluated on Tanzania.

Additionally, in our attempt to better understand the embeddings and their significance, we evaluated all our models on Uganda, utilizing our Average MLP model, to analyze R^{300} inputs to see which embedding indices were most important for the model's predictions. We did this by evaluating the model 300 times, each time masking all but one of the input dimensions, and observing the resulting r^2 values. We found several indices, most notably indices 24 and 182, that were highly predictive (obtaining .3 and .4 r^2 by themselves, respectively; see Figure 6b). Searching our inputs, we found that Ugandan articles which possess the highest values at these indices are often those related to healthcare and education (ie. hospitals and schools; see Figure 7). This analysis is very informative, as it reveals that our model is learning to look at the healthcare and education of a region (via Wikipedia articles) in order to predict the poverty of the region. In other words, our model is learning that the healthcare and educational status of a region is very informative for predicting poverty, which, in reality, is true.

6 Conclusions

In this paper, we present preliminary yet state-of-the-art results for poverty prediction, using Wikipedia article embeddings and nightlights histograms. Before this becomes usable in real-world applications, we plan to acquire Wikipedia datasets with article creation dates in order to match them with appropriate nightlights imagery and explore using Wikipedia articles written in different languages, such as French, as they may provide us with more Africa-related articles. Overall, we feel our approach is a novel method that holds promise for future exploration and application.

Code

The code for our models and error analysis can be found at the following link:
https://drive.google.com/open?id=1ZmOzz7UXOE2eSesRWLSdSAK5X13lvznb

Contributions

All members of this team contributed equally to the research presented in this paper. Specifically, Evan worked on obtaining and pre-processing the dataset, as well as developing and running our two main models. Zaid worked on developing and running the baselines, as well as conducted error analysis. Chenlin worked on developing the second model and running experiments with all of our models, as well as conducting error analysis.

We would like to acknowledge the help of Stefano Ermon, Marshall Burke, and David Lobell in brainstorming ideas to tackling this problem, as will as their assistance in finding datasets and exploring previous research efforts. We'd also like to thank Christopher Yeh for his feedback and guidance.

References

[1] Un sustainable development goals. `https://www.un.org/sustainabledevelopment/sustainable-development-goals/`. Accessed: 2018-12-07.

[2] Wikipedia online encyclopedia. `https://www.wikipedia.org/`. Accessed: 2018-12-07.

[3] D. Basak, S. P. Pal, and D. Patranabis. Support vector regression. *Neural Information Processing – Letters and Reviews*, 2007.

[4] J. Chen, S. Sathe, C. Aggarwal, and D. Turaga. Outlier detection with autoencoder ensembles. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pages 90–98. SIAM, 2017.

[5] C. D. Elvidge, K. Baugh, M. Zhizhin, F. C. Hsu, and T. Ghosh. Viirs night-time lights. *Int. J. Remote Sens.*, 38(21):5860–5879, Nov. 2017.

[6] J. Hauke and T. Kossowski. Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quaestiones geographicae*, 30(2):87–93, 2011.

[7] 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.

[8] N. Jean, S. Wang, G. Azzari, D. Lobell, and S. Ermon. Tile2vec: Unsupervised representation learning for remote sensing data. *arXiv preprint arXiv:1805.02855*, 2018.

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

[10] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196, 2014.

[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[12] G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[13] A. Perez, C. Yeh, G. Azzari, M. Burke, D. Lobell, and S. Ermon. Poverty prediction with public landsat 7 satellite imagery and machine learning. 11 2017.

[14] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[15] B. Scholkopf and A. J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[16] E. Sheehan, B. Uzkent, C. Meng, Z. Tang, M. Burke, D. Lobell, and S. Ermon. Learning to interpret satellite images using wikipedia. *arXiv preprint arXiv:1809.10236*, 2018.

[17] G. Shperber. A gentle introduction to doc2vec. *Medium*.

[18] A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[19] S. M. Xie, N. Jean, M. Burke, D. B. Lobell, and S. Ermon. Transfer learning from deep features for remote sensing and poverty mapping. In *AAAI*, 2016.

[20] B. Xu, N. Wang, T. Chen, and M. Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.