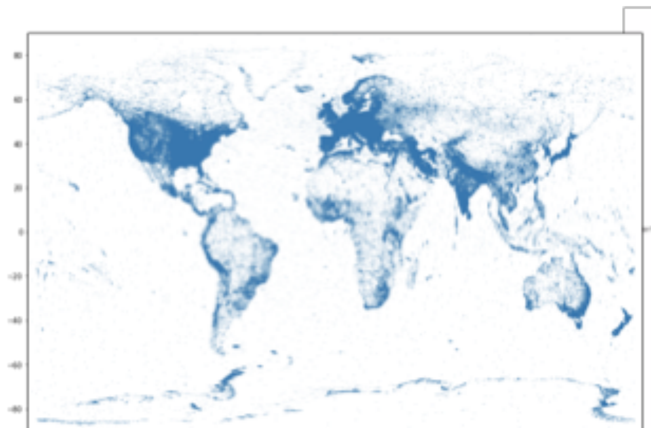
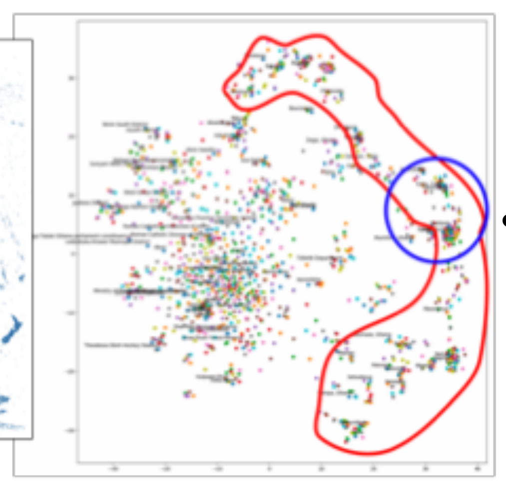
Using Latent Embeddings of Wikipedia Articles to Predict Poverty

Evan Sheehan, Chenlin Meng, Zaid Nabulsi

Background & Summary

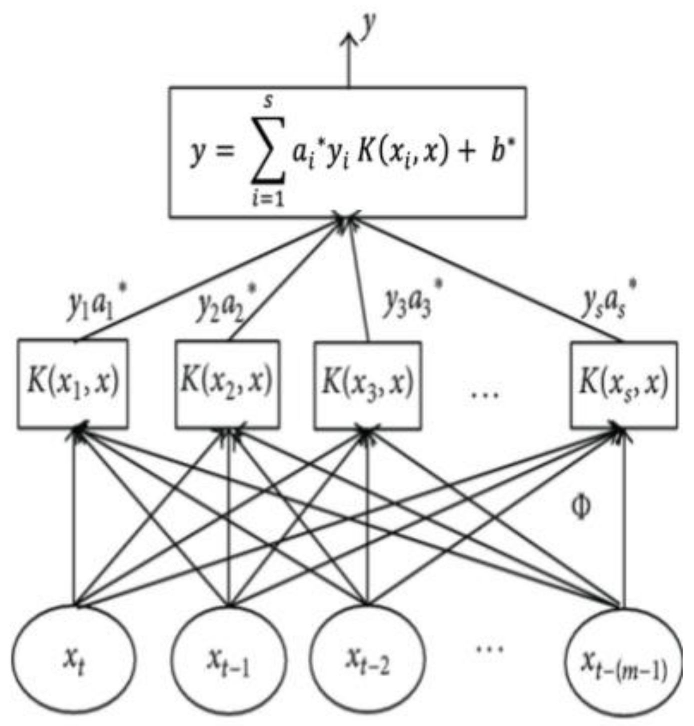
In this project, we propose a novel method for the task of poverty prediction through the use of geolocated Wikipedia articles. Traditional state-of-the-art models rely on nightlights images to regress on the problem. We explore the utilization of the latent embeddings of these articles (Sheehan et. al. suggest geolocated Wikipedia articles can be used as socioeconomic proxies for their surrounding regions) for wealth index prediction. These articles contain almost no information about poverty or wealth at face-value. However, we obtain results suggesting that latent features within these articles strongly correlate with poverty, allowing us to perform regression on points throughout Africa and challenge the state-of-the-art results.





Approach & Methods

Baseline Models



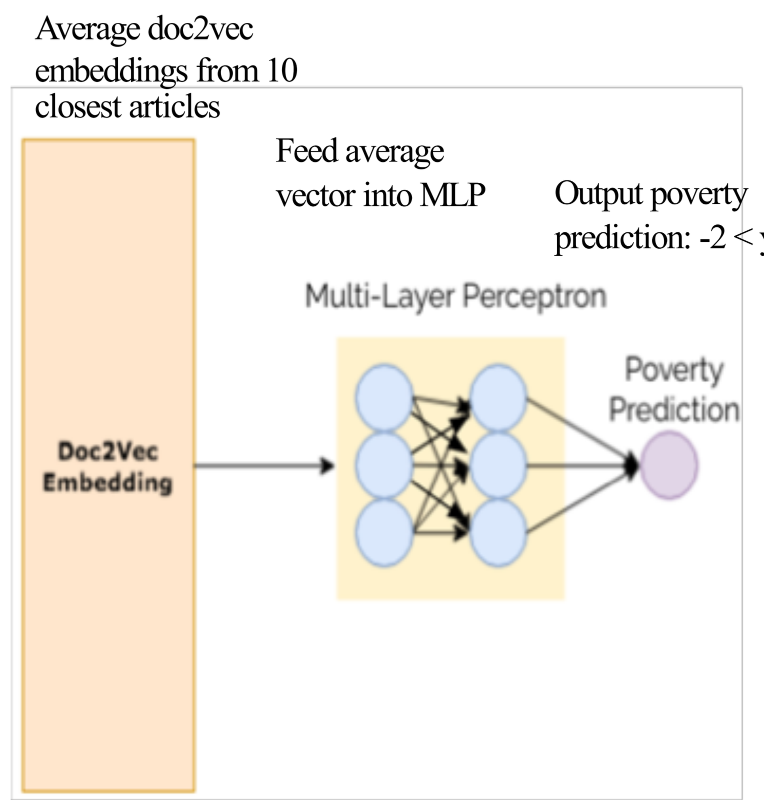
Doc2Vec SVM Regression

- *Goal*: Create a simple model to sanity check data and get a sense of the difficulty of the task.
- *Approach*: Use support vector machine regression to predict the poverty index from Doc2Vec embeddings of 10 closest articles. Use loss function:

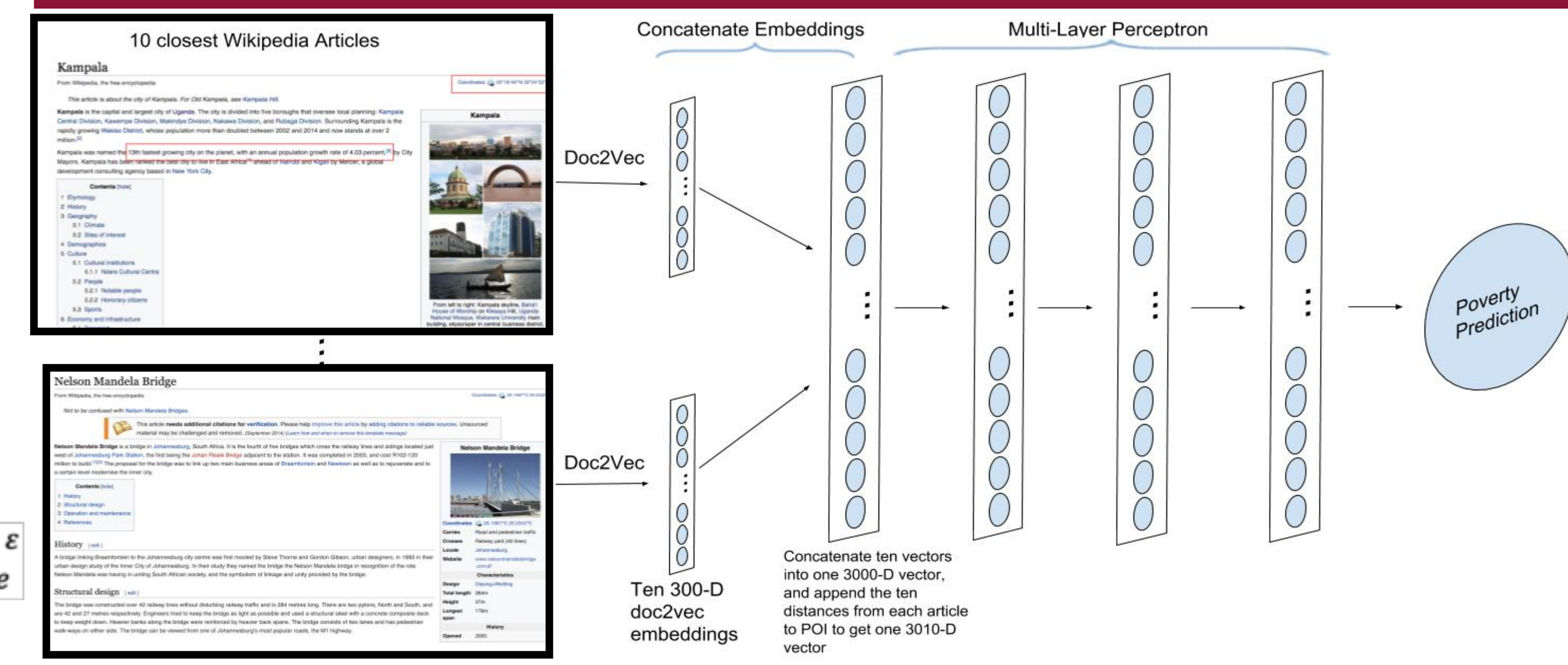
$$L = (y, \hat{y}) = f(x) = \begin{cases} 0, & |y - \hat{y}| < \varepsilon \\ |y - \hat{y}| - \varepsilon, & otherwise \end{cases}$$

Doc2Vec Neural Network

- *Goal*: Design a fully connected neural network as a second baseline.
- *Approach:* Train an MLP from scratch with a regression output corresponding to the poverty value.
- Take 10 closest articles to coordinate of interest, get Doc2Vec embedding of each, average the 10 feature vectors to get input for MLP.



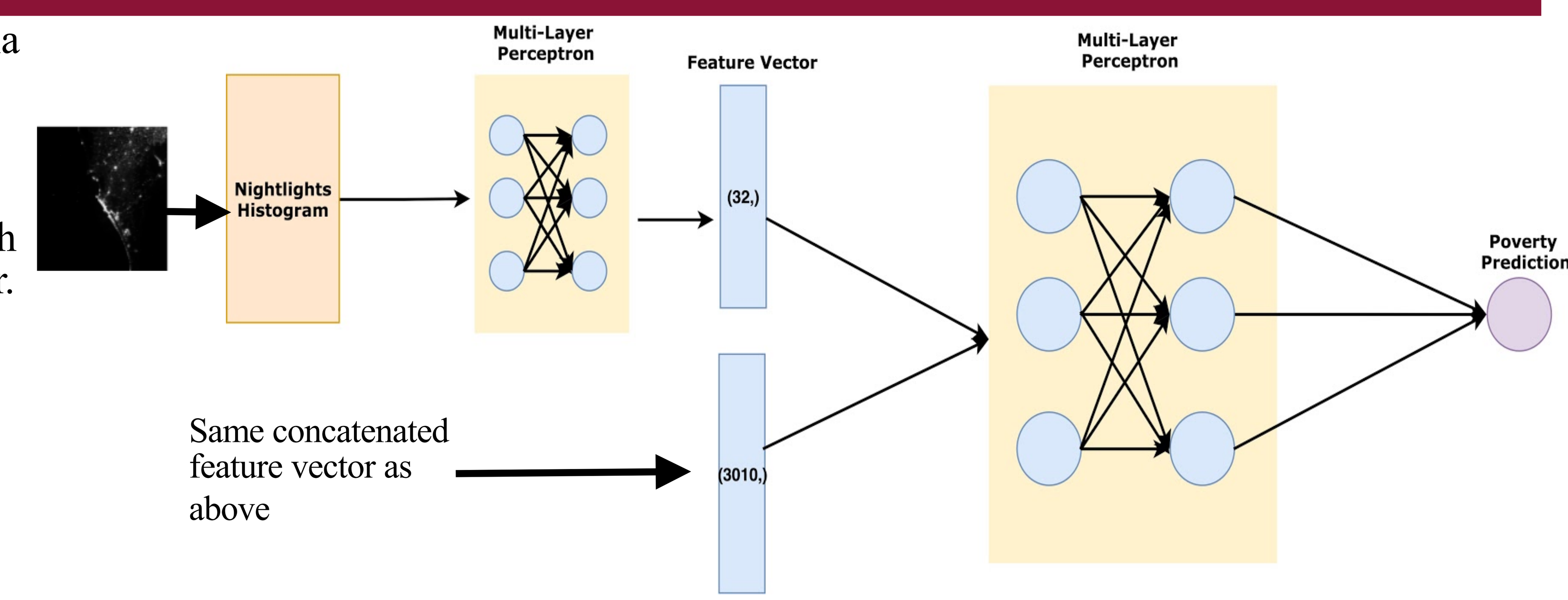
Wikipedia Embedding MLP



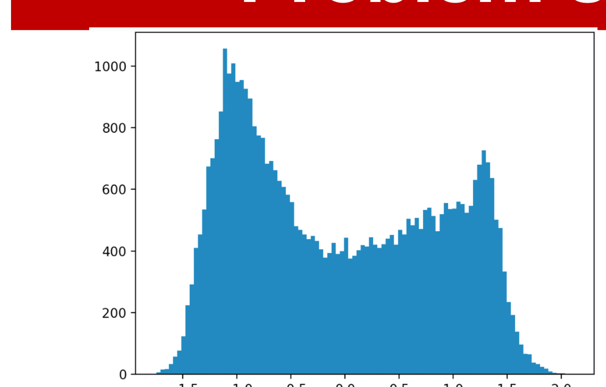
- In this model, similar to our second baseline, we take the ten closest geolocated Wikipedia articles to the point of interest, and pass each through a Doc2Vec model, to get ten 300-dimensional vectors
- We then concatenate the ten vectors, and for each vector, we also append the distance of that Wikipedia article from our point of interest, to get a 3010-dimensional vector.
- We pass the 3010-dimensional vector through an MLP to get a poverty prediction.

Multi-Modal Model

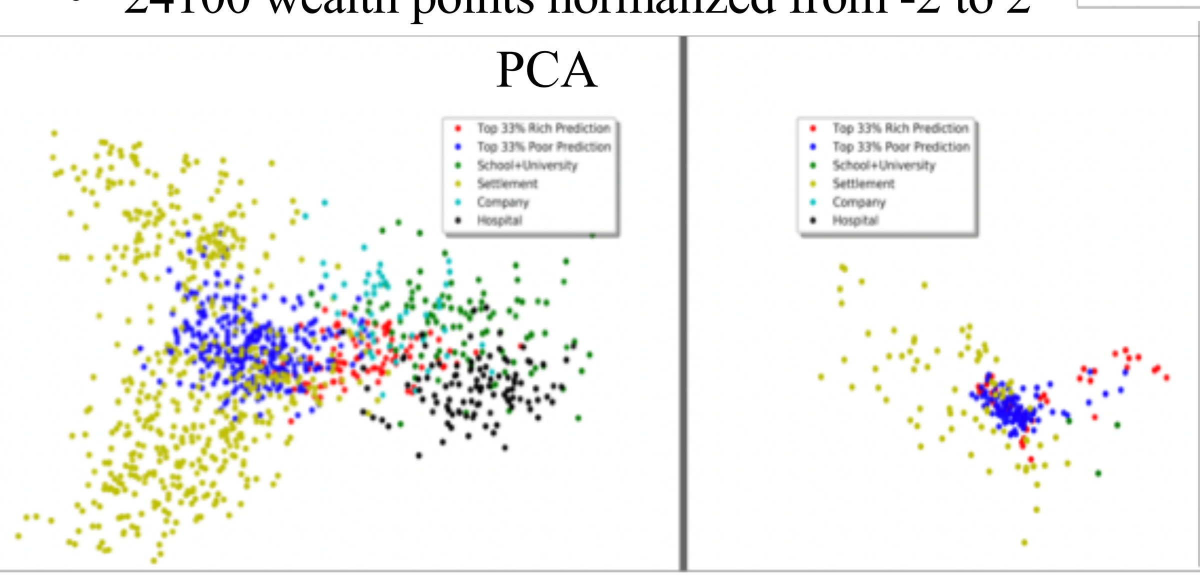
- In this model, we utilize both Wikipedia embeddings as well as the nighttime image of the region of interest.
- We generate a histogram from the nightlights image, and feed that through an MLP to obtain a 32-D feature vector.
- We also generate the same 3010-D vector described above through Doc2Vec embeddings.
- We concatenate both inputs and pass them through an MLP to get our final poverty prediction.



Problem & Data

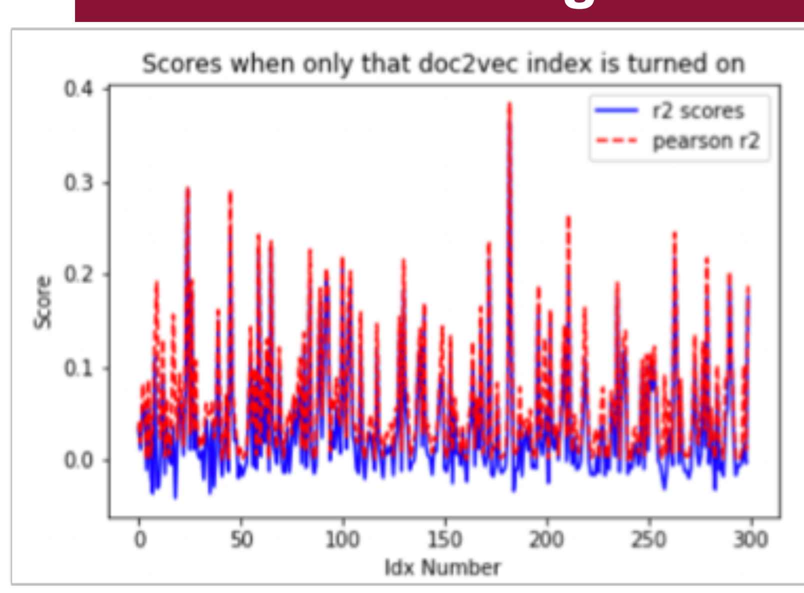


- Goal - predict poverty level given geolocated Wikipedia articles (1 mil. articles scraped).
- Data from Stanford Sustain Lab, UN World Bank, and DHS
- 24100 wealth points normalized from -2 to 2

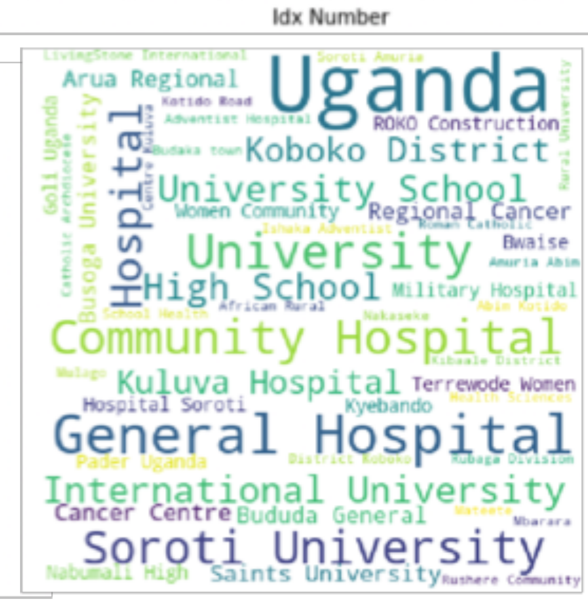


Results

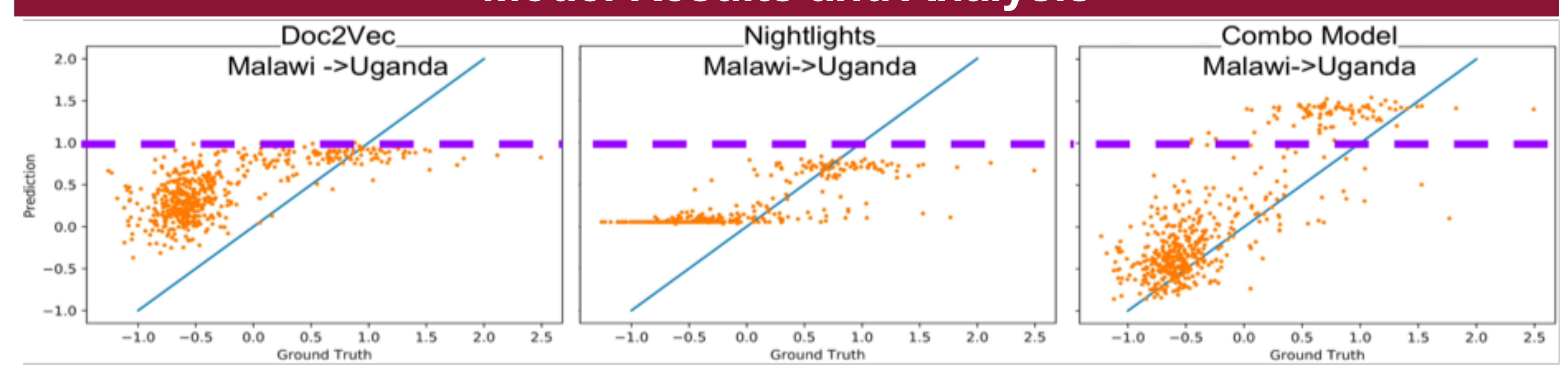
Embedding Activation Analysis



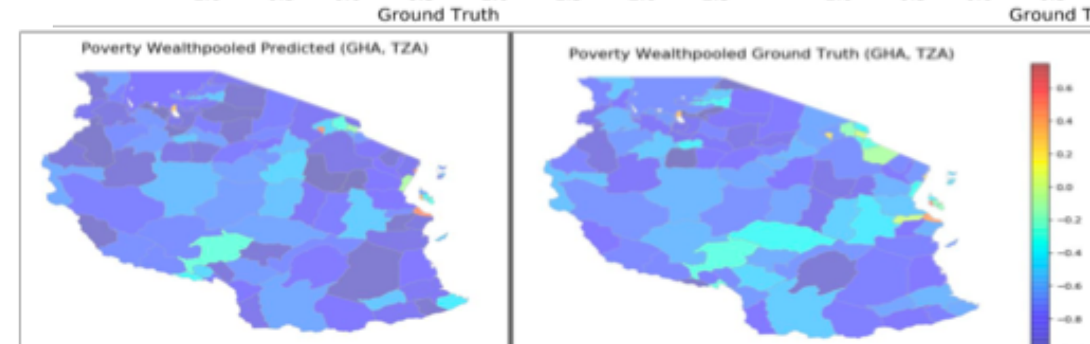
On the left, masked activation of each embedding index is shown (all other indices are set to 0), along with its corresponding r^2 value. We see that indices 24 and 182 yield the highest r^2. Below, we see the article titles which posses the highest values in those indices. On the left titles for index 24 are shown, while on the right, titles for index 182 are shown. Healthcare and education are important factors.



Model Results and Analysis

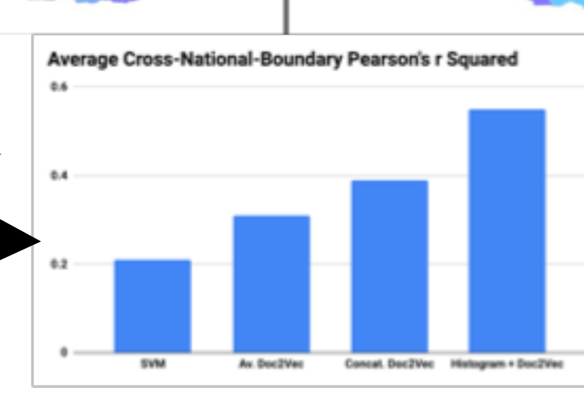


Observed vs. predicted wealth values for models trained on Malawi and tested on Uganda. Leftmost graph shows only Doc2Vec input, center graph shows only nightlights histogram input, and rightmost graph shows model with both inputs



Predicted vs. ground truth value for concatenated Doc2Vec model trained on Ghana and tested on Tanzania.

Average r^2 value for each of our 4 models trained and tested on Uganda, Tanzania, Nigeria, Ghana, and Malawi (excluding testing on trained country).



Cross-national boundary r^2 results for our multi-modal model. Trained on column, and tested on row countries. Outperforms current state-of-the-art models

	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	-

Further Work

So far, we have detailed a novel comparative approach for the task of poverty prediction, in particular, using latent Wikipedia embeddings to predict wealth levels with r^2's that outperform state-of-the-art models. Our results suggest that combining nightlights imagery with Doc2Vec embeddings creates large improvements. In the future, we plan to experiment with more multi-modal architectures that show promise, such as the use of convolutional neural networks for the imagery.

References

[1]N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon. Combining satellite imagery an dmachine learning to predict poverty.Science, 353(6301):790–794, 2016.[2]T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space.arXiv preprint arXiv:1301.3781, 2013.[3] G. A. Miller. Wordnet: a lexical database for english.Communications of the ACM, 38(11):39–41, 1995.[4]A. Perez, C. Yeh, G. Azzari, M. Burke, D. Lobell, and S. Ermon. Poverty prediction with public landsat 7satellite imagery and machine learning. 11 2017.[5]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.[6]S. M. Xie, N. Jean, M. Burke, D. B. Lobell, and S. Ermon. Transfer learning from deep features for remotes ensing and poverty mapping. InAAAI, 2016.[7]S. K. Yarlagadda, D. Güera, P. Bestagini, F. M. Zhu, S. Tubaro, and E. J. Delp. Satellite image forgery detection and localization using gan and one-class classifier.CoRR, abs/1802.04881, 2018.[8]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.[9]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.