Predicting Educational Opportunity using Satellite Imagery within the United States

Greg DePaul & Hugo Valdivia

Stanford University - CS 230: Deep Learning

Objective

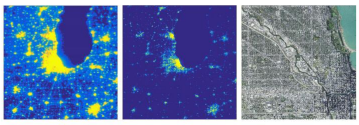
We set out to build a model that, given satellite imagery and structured covariate data, is able to provide a reasonable estimate of the performace of a school district, as measured by the National Assessment of Educational Progress (NAEP) scale. This is done in the hopes of discovering critical features in satellite imagery, that have been used to predict poverty as in [1], which can serve as determinants of educational opportunity.

Description of Dataset

Our structured data is provided by the Stanford Education Data Archive (SEDA), which compiles a wide range of data describing educational performance of over 12,000 school districts within the United States. Some of the information this data set provides us with is:

- School District ID
- Socioeconomic Status Composite Index
- Racial Diversity
- Mean Score of Student Body [Score]

For our baseline model we used Night-time Light Intensity Data, specifically the datasets available from the Defense Meteorological Satellite Program Operational Linescan System (DMSP-OLS) and the Visible Infrared Imaging Radiometer Suite (VIIRS). Our more generalized GatedCNN model makes use of the Daytime Imaging available from the LandSat-8 dataset. This allows the model to make a more accurate measure based on the discernible pieces of infrastructure.



(a) DMSP-OLS (b) VIIRS (c) LandSat-8

Figure 1: Satellite Imagery of the Greater Chicago Area

Model Architecture - GatedCNN

GatedCNN Architecture: Drawing from [2], for our model, we fine-tune three CNN's separately on our three different types of satellite input data and then we have introduced a small dense 'Gate' to find a weighted function based on the recommendation of the three networks. Two small convolutional networks are used for the small pixelled DMSP and VIIRS images. The ResNet-50 architecture pre-trained on ImageNet is used for the LANDSAT Data.

Loss Function: We equipped this architecture with the *Huber Loss Function*:

$$\mathcal{L} = \begin{cases} \frac{1}{2}(\hat{Y}_i - Y)^2, & |\hat{Y}_i - Y_i| < \delta \\ \delta(|\hat{Y}_i - Y_i| - \frac{1}{2}\delta) & otherwise \end{cases}$$

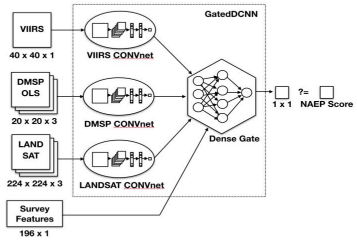


Figure 2: GatedCNN architecture

Data Augmentation

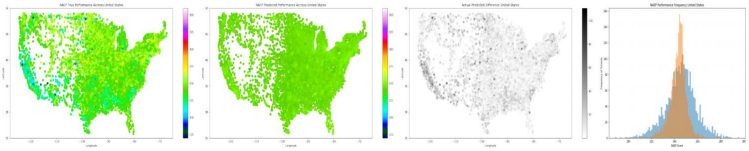
We perform augmentation of the images via the Keras built-in *ImageDataGenerator* class; applied to randomly rotate our train images between epochs. This rotates the square image between 0 and 360 degrees and then interpolates the remainder of the image to maintain its dimensions.

Results

While we can achieve R^2 values quite close to 1 on the train set for individual models, the best results on the validation set show that it is difficult to generalize. The Day-time ResNet-50 model is the best individual model, and our current best model on the validation set is the gated model; on the other hand, we see that the general mass distribution is well-maintained.

Model	Epochs	R^2
Night-lights DMSP CONV	670	-.012
Night-lights VIIRS CONV	120	.010
Day-Time LANDSAT CONV	40	.060
GatedCNN Model	1000	.160

Table 1: Prediction R^2 Accuracy



(a) Actual Performance (b) Predicted Performance (c) Absolute Differences (d) Overlay of Distributions

Figure 3: Comparison of Predicted Performance versus Actual Performance

Discussion and Conclusion

By our results, we see that it is possible to generalize upon a local infrastructure to gain some insight into education opportunity. By making use of a GatedCNN, we were able to generalize over a diverse set of satellite input features.

For the most part, it appeared DMSP was the least likely to contribute to the Gate Model Prediction. When we investigated it's ability to predict, it fell incredibly short compared to that of VIIRS and LANDSAT. This could be due to interference to its measurements, possibly the result of cloud coverage. Conversely, the LANDSAT model which ran through ResNet-50 appeared to perform the best and would thus earn a greater weight in the Gate Prediction.

Future Work

We recommend the following directions for future work on this topic:

- Better method for extrapolating over missing / sparse data features. This would allow for using the more structured data within SEDA to better predict model performance.
- Alternative methods of interpreting performance by bucketing schools and then classifying local infrastructure into those buckets.

References

[1] Ishfaq et al. Duan, Chartock.
Predicting poverty with satellite imagery in bangladesh and india.
2018.

[2] Alex; McCool Christopher; Upcroft Ben; Corke Peter; Sanderson Conrad Ge, ZongYuan; Bewley.
Fine-grained classification via mixture of deep convolutional neural networks.

[3] S.F. Reardon.
Educational opportunity in early and middle childhood: Variation by place and age.
CEPA Working Paper, 17(12), 2018.