Predicting Educational Opportunity within the United States using Satellite Imagery

Greg DePaul & Hugo Valdivia

***Abstract*—It is a sizable challenge to collect education data from school districts in the United States. The Stanford Education Data Archive (SEDA) stores a wide range of covariate data for over 12,000 school districts all over the US; however, due to the difficulty of gathering such information, the data is incomplete. Our project uses Satellite Imagery in order to construct accurate measures of education opportunity all over the United States. We create a supervised model that takes in satellite images and incomplete structured data for a particular school district and produces the performance of that school district, as measured by the National Assessment of Educational Progress scale. Our best model achieves a 0.629 R^2 on our test set, indicating that we have successfully developed a regression model.**

I. Introduction and Related Works

We set out to build a model that, given satellite imagery and structured covariate data, is able to provide a reasonable estimate of the performance 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.

In [1], researchers working with Stefano Ermon, of Stanford University, applied deep learning methods on Satellite imagery in order to develop a regression model of income for the countries of India and Bangladesh. We viewed this work very critically when it came to applying our own methods. For their results, Ermon et al. were able to achieve R^2 values of 0.3251 and 0.1080 for India and Bangladesh, respectively, which gives us an expectation for our own model's achievable performance.

While Dr. Ermon's paper considers the component networks for learning each of the satellite images separately with no attempt to combine these results into a composite estimate for poverty. In order to incorporate such a method to composite multiple networks that have the same output but different input features, we sought a network model like an ensemble network, mixture network, as well as a gated network [2]. We eventually decided, based on this paper, that a GatedCNN was the best choice to base our model.

II. Description of Data Set

A. Structured Data

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

- School District ID
- Socioeconomic Status Composite Index
- Racial Diversity
- Mean Score of School District

These, however, are not the only attributes that are available; SEDA also provides over 150 additional features (e.g., pupil-teacher ratio, percent of households with 5-17 year olds living in poverty, percent of adults with a BA or higher, et cetera). We merged their main dataset with their covariate dataset to arrive at data for 12,139 school districts all over the United States.

The covariate dataset, however, was not perfect; around 12.3% of the entries were missing. To tackle this problem, we tried a number of approaches. First, since all but 5 of the attributes were non-negative, we substituted all of the missing entries with the value -1 in hopes that our network would learn to discern the missing values. Second, we tried to substitute the missing entries with the mean value of the present entries. Thirdly, we tried to delete all attributes which were missing any values; in practice, this resulted in around 2,000 school districts. Lastly, we set thresholds and deleted attributes that had more missing entries than the threshold and then filled in the then fewer missing values. Empirically, we found that the best results came from substituting the mean.

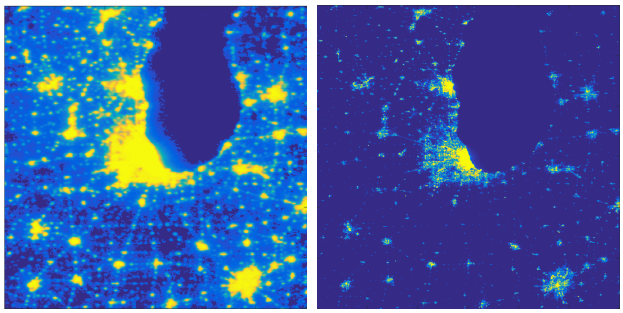
B. Satellite Imagery

For our baseline model and eventually GatedCNN model, we used Night-time Light Intensity Data 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. The daytime data allows the model to make a more accurate measure based on the discernible pieces of infrastructure while the night-time data provides inference on the relative level of wealth locally. These three sources of satellite image data can be seen in Figure 1.

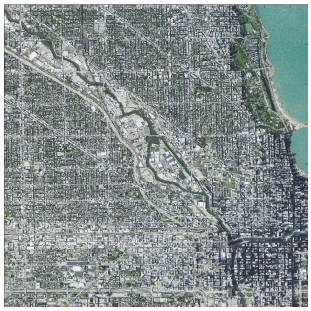
III. Initial Approach to Solution

A. Data Input:

For each school district, we identify its respective Latitude / Longitude from a simple dictionary lookup. We



(a) DMSP-OLS (b) VIIRS



(c) LandSat-8

Fig. 1. Satellite images selected from our four datasets, depicting the greater Chicago area

then use this Lat / Long pair to identify a square image for our DMSP-OLS, VIIRS and LANDSAT datasets. We establish a pixel radius to be 10, which we see that since VIIRS has twice the resolution as DMSP, we know we should return 20x20 and 10x10 pictures available for each Lat / Long pair from each dataset. On the other hand, LANDSAT, with it's high pixel density, yields the largest image of size 600x600 of the local surrounding area. To start our analysis for our model, we extracted a single mean performance value for close to 12,000 school districts. Given this number of data points, we used a 60-20-20 splits for our train-dev-test sets.

B. Baseline Architecture:

For our baseline model, we employed a shallow feed forward network, given that at the time, we had access only to VIIRS and DMSP, so the images being processed over were relatively small enough to be handled efficiently in such a way. This architecture is composed of four hidden layers, each with a ReLU activation function, and predicts the mean NAEP score of the school district located at the center Latitude / Longitude Pairing. This model doesn't include the categorical data for our data set, but instead serves to give motivation as to whether we can reliably draw education performance from such imagery. To address a variance problem, we added L2-regularization to our model.

C. Loss Function:

Drawing from the use of this particular loss function in [1], we use Tensorflow's preprogrammed version of the Huber Loss with a delta of 1.0, which we found helped to deal with outliers in our data set; intuitively, the Huber

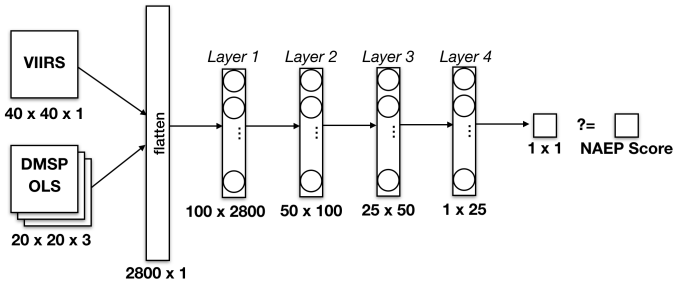


Fig. 2. Baseline architecture

loss acts as squared error when the loss is small and as absolute error when the loss is large. We began by using a mean-squared error, but this proved to give us worse performance than the Huber loss.

IV. Initial Results

We found that a small feed-forward neural network with 4 hidden layers could overfit the train data quite well; however, this simply did not generalize to the validation set. We did not have the ability to create more data in this setting, so we tried regularization, to decrease our variance. While we could achieve R^2 values quite close to 1 on the train set, we could not achieve a positive R^2 value on the dev set; with our best model, the R^2 value on the dev set was approximately -0.65. Since we used a feed-forward network that simply flattens an image and then connects each value of the image to every neuron in the hidden layers, this architecture is not ideal for capturing the spatial relationships of images, so these results were to be expected; however, they serve as a lower threshold for our further analysis. We move to convolutional networks in our final model, as these architectures are specifically tailored to handle proximal relationships within an image.

V. Introducing Our Architecture

A. 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. A GatedCNN appealed to us because it's effective for learning to form an opinion on the recommendations of smaller components, and then weighs them based on which component it feels would best predict based on the image input.

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. We experimented with the smaller DenseNet121 architecture, but we found that the performance was worse than the ResNet-50. We chose a linear weighting network for the dense Gate, which we feel

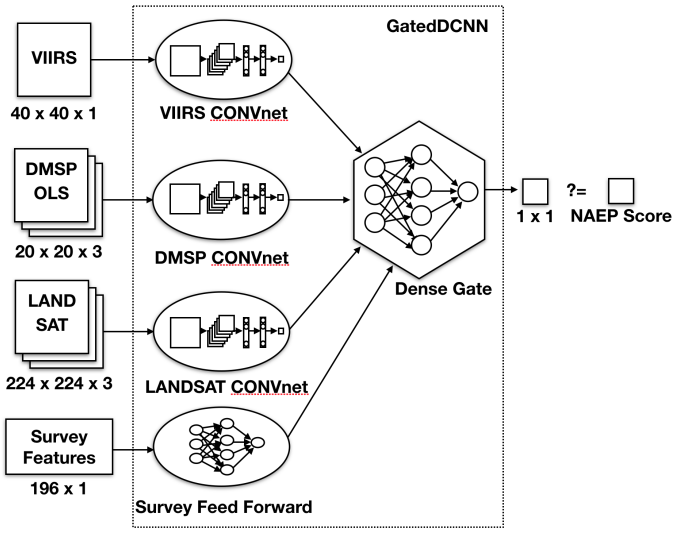


Fig. 3. GatedCNN architecture

is adequate to handle the recommendations of our three Satellite inputs as well as the our survey features.

B. 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. Given that we were training two convolutional networks from scratch and fine-tuning a very deep one, this step gave a huge boost to our performance. We first tried vertical and horizontal flips, but we found that the rotation setting allowed for the most generalization to our validation set.

C. Hyperparameter Tuning

For our network, we had the typical parameters of learning rate, batch size, and depth of our networks. On the other hand, our network also introduced new hyperparameters such as picture size (radius) and zoom level, which then dictates the size of our input set. We experimented with hyperparameters, especially at the mercy of AWS, which had a tendency to crash when we increased batch sizes past a certain threshold. Similarly, we want hyperparameters like picture radius and zoom level to be large enough to capture a lot of information, but not too large to capture multiple schools as well as making the network perform longer searches. For our model, we ended choosing much of our hyperparameters to maximize processing power on AWS. The concluding hyperparameters are:

- Batch Size: 16 due to memory for the LANDSAT model. VIIRS and DMSP had a size of 1024.
- Small Convnet Depth: 1 Conv Layer, 1 Pool, 2 Dense, which is necessary because the images are small and also the input set is small.
- Learning Decay: Set to zero because this tended to slow the network learning
- Image Radius: for DMSP and VIIRS, we chose an approximate 10 mile radius, which equates to a total radius of 10 pixels for VIIRS and 20 pixels for DMSP. For LANDSAT, we simply used the maximum radius possible, which is 600x600 pixel.
- Zoom Level: We played around with zoom levels from 15 to 18. We found that zooming too far out led to an image in which buildings and other features became hard to distinguish, while a zoom level of 18 or more just left you with an image of the school itself. We found that a zoom level of 17 led to the best results for discerning features.

VI. Final Results

To measure the effectiveness of our model, we used R^2 metric of accuracy to see how closely related our regression model was to the actual performance of the school districts. 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. As indicated by the Table in (4a), the Daytime ResNet-50 model is the best individual model under this metric, and our current best model on the validation set is the gated model.

To gain a better understanding of our results, we sought visualization to understand this high dimensional problem. Figures (4b) and (4c) displays actual and the predicted NAEP scores respectively in color for each plottable school district. The better the performance, the higher on the color band towards red, while lower scores tend to lean towards blue. Both models display a scale like appearance, caused by a high density of scores that rank yellow-green on the colormap, indicated most schools perform within this margin.

We attempt to visualize the overall difference in model performance, as seen in figure (4d). Immediately, we see that our model scores differ from the true scores mostly within the Californian basin as well as the North-Midwest. We attribute this to schools in close proximity to one another, while having vastly different performance. This can be due to a variety of reasons, such as population density.

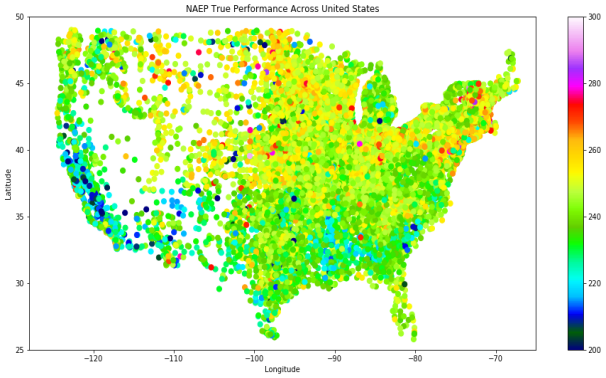
A better way to see model performance, we found, is plotting the frequency of school performances by histogram (Figure 4e). Our model, in yellow, is able to non-parametrically identify the mean of the actual score performance well. However, we see evidence of bias, which we attribute to the Gate estimating the schools scores pessimistically.

VII. Conclusion

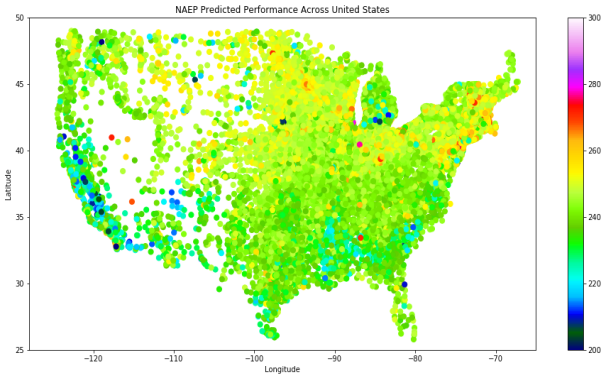
By our results, we see that it is possible to generalize upon a local infrastructure to gain some insight into education opportunity. By making use of a GatedCNN, we were able to generalize over a diverse set of satellite

Model	Val	Test
Night-lights DMSP CONV	.052	.028
Night-lights VIIRS CONV	.083	.060
Day-Time LANDSAT CONV	.056	.055
GatedCNN Model w/o Struct Data	.134	.134
Structured Data	.613	.618
GatedCNN Model w/ Struct Data	.632	.629

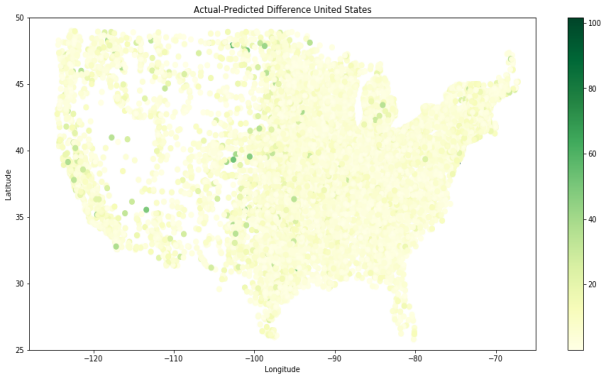
(a) Prediction R^2 Accuracy



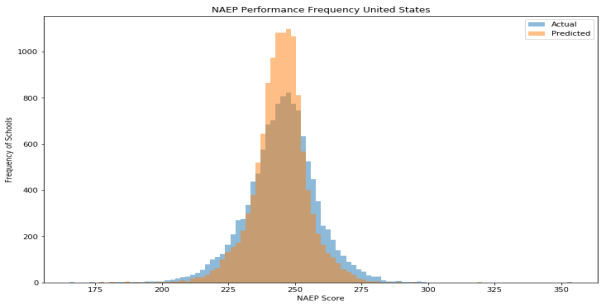
(b) Actual Performance



(c) Predicted Performance



(d) Absolute Differences



(e) Overlay of Distributions

Fig. 4. Comparison of Predicted Performance versus Actual Performance

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.

VIII. 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.

IX. Inspiration and Contribution

Greg and Hugo met in an education seminar at Stanford on "Project Based Learning." Their mutual dedication to bettering education lead them to pursue this project when they both teamed up for CS 230 : Deep Learning. Greg was able to develop much of the input pipeline for retrieving the satelite imaging as well as the visualizations of the imagining. Hugo dedicated his time to focus on developing the ResNet-50 model to discern the information-rich LANDSAT. Together, they constructed the GatedCNN which lead to their results.

The code for this project can be accessed at

https://github.com/valdivia4/cs230project.git

References

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

[2] Christopher McCool Ben Upcroft Peter Corke Conrad Sanderson ZongYuan Ge, Alex 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.

[4] Benjamin R. Shear Erin M. Fahle Demetra Kalogrides Sean F. Reardon, Andrew D. Ho. and Richard DiSalvo. Stanford education data archive (version 2.0), 2017. data retrieved from http://purl.stanford.edu/db586ns4974.

[5] Benjamin R. Shear Erin M. Fahle Demetra Kalogrides Sean F. Reardon, Andrew D. Ho. and Richard DiSalvo. Version 4 dmsp-ols nighttime lights time series, 2013. data retrieved from https://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html.

[6] National Geophysical Data Center. Version 1 viirs day/night band nighttime lights, 2018. data retrieved from https://ngdc.noaa.gov/eog/viirs/download_dnb_composites.html.