CS230

Using Satellite Imagery, Street-Level Imagery and Multi-Modal CNNs to Predict Women's Educational Attainment

Disa Alda Naomi
Stanford University
disaalda@stanford.edu

Haya Hidayatullah
Stanford University
hayah@stanford.edu

Abstract

Women's educational attainment is a key indicator that differentiates developed countries from developing countries and hence, the UN uses it to measure a country's progress. However, it is difficult to track this metric on a local level, especially in developing countries and rural areas. On the other hand, with the development of imaging technology, there is an abundance of satellite images and street-level images that have been found to have success with predicting developmental indicators. In our project, we attempt to predict women's educational attainment using both street and satellite-level imagery. In this task, we use transfer learning with pretrained ResNets trained on our images for feature extraction. We extract a feature vector for each satellite image and an average feature vector for its associated street images which we concatenate into a combined feature vector. This feature vector is then fed into a CNN to predict women's educational attainment. Our results suggest that (lowest test MSE of 13.3) our combined multi-modal CNN does not improve predictions above a baseline using satellite-imagery alone. This also suggests that street-level and satellite imagery together may not be informative for our prediction task.

1 Introduction

While data is key in measuring progress in achieving various Sustainable Development Goals ("SDG"), often there is only few data available at local levels. On the other hand, with the increase in the imaging technology, satellite and street-level imagery are becoming increasingly available at local levels, and related works [4, 5] have shown that such data can be predictive of SDG-relevant metrics.

In this project we are particularly interested in predicting women's educational attainment using satellite imagery and street-level imagery. There is evidence that women's educational attainment is influenced by their environmental factors, such as access to educational facilities, housing conditions, public infrastructure and public transportation among other things.

The input to our network is a set of images, where each example in the dataset consists of one satellite image and between 1 - 5 street-level images [1]. We plan to first use the satellite images to output a predicted value for the women educational attainment for the region. Then, we plan to augment our input with street level of images in order to experiment with improving the model's performance. Thus, we will produce a baseline model using just satellite images, and compare it against a multi-modal network that incorporates both satellite and street-level images.

CS230: Deep Learning, Winter 2018, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

2 Related work

A research problem can be characterized as multimodal when the process in which something happens or is experienced involves multiple distinct signals. Multimodal machine learning aims to build models that can process and relate information from multiple modalities. [5]

Using multi-modal approaches in machine learning is usually aimed improving model performance by utilizing different kinds of information that individual data sources can provide when combined. However, the method of combination/fusion still remains as experimental challenges in the field, e.g. early and late fusion, hybrid fusion or just simple feature vectors concatenation. [2] In this project we aim to experiment with multimodal models, by combining potentially informative distinct signals from the satellite and street images of the same localities in predicting women's educational attainment of the particular location.

3 Dataset and Features

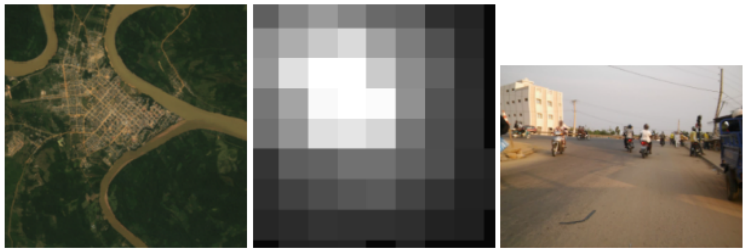

Figure 1: Satellite Imagery (Left), Nightlights (Middle), and Street-Level Mapillary Images (Right)

Our predicted label, the women educational attainment metric is derived from survey data from the Demographic and Health Surveys ("DHS") program. SustainBench summarized the survey data into geographical IDs unique for each satellite image and also "cluster-level" labels, where each "cluster" corresponds to a village or local community. The women's education metric is created by taking the cluster level mean of "education in single years" among women between the ages of 15 and 49 and is a continuous variable thresholded at 18 years.

The original dataset from SustainBench consists of 117,062 examples from 76 countries, where each example in the dataset consists of one 255x255x8px satellite image and 1 - 300 street-level RGB images with 1024px shortest length. [1]

For each dataset of satellite images and street-level images, we created a dataset that maps the label to an image, including metadata of the images such as the geographical and cluster IDs, longitude and latitude coordinates, as well as its SDG metrics.

We made use of images from a list of 10 countries ('AM', 'BJ', 'CD', 'CM, 'GH', 'KY', 'MD', 'NM', 'NP', 'ZW'). We only made use of the RGB channels of the satellite as ResNet models only take RGB images. We used min-max scaling to normalize the values of each image. To ensure that the street-level images are of the same input size, we resized all of the street-level images used to 224x224 RGB images per SustainBench's guidance on its GitHub page.

Our final dataset contains 1,891 satellite images with 8,961 associated street images. We limited the number of associated street images to a maximum of 5 street images for each satellite image for our first model. For our second model, in an attempt to improve performance, we limited the number of associated street images to 1. Finally, we split up our working dataset into train, validation, and test set by time to give a roughly 60/20/20 split, randomly shuffling images of different countries to each set. We trained the model using the training set, tuned the hyperparameters with the validation set, and analyzed the model's performance on the unseen test set.

4 Methods

4.1 Baseline Model

Our baseline model is the pre-trained ResNet-18 CNN-based model, with satellite images as inputs and women educational attainment as outputs. We chose the ResNet model as it has performed well for many computer vision tasks, including those with satellite imagery as inputs, as cited by the SustainBench paper. We compared performances of three types of ResNet architectures ('ResNet-18', 'ResNet-34', and "ResNet-50') and found that the ResNet-18 performs the best on our dataset.

4.2 Multi-Modal CNN

We implemented a novel multi-modal network, where we have separate convolutional neural networks ("CNNs") to learn embeddings of satellite and street-level images of a certain region.We incorporated pre-trained weights from ResNet as building blocks of our multi-modal network. We experimented with various CNN architectures, including different types of ResNet architectures and using various number of number of layers and activation layers, and tuning hyperparameters to achieve the best performance on our dataset.

The model takes in one satellite image and its associated street images (upto 5) as input. We train ResNet on the images and extract the second last layer to map the image to its feature vector in R^k where k = 512. For the street images, we averaged the feature vectors to one (512 x 1) vector and concatenate both satellite and street mappings into one vector of (1028 x 1). We use ResNet 50 for street-level images as it improved model performance. Using the final concatenated feature vector, we train with MSE as our loss function to predict the regression output for each example. Our best performing model architecture is shown below.

We hypothesized that this novel, comprehensive approach to predicting women's educational level will improve performance, and that the additional data. The model optimizes MSE in training and we will analyze the same regression evaluation metrics.

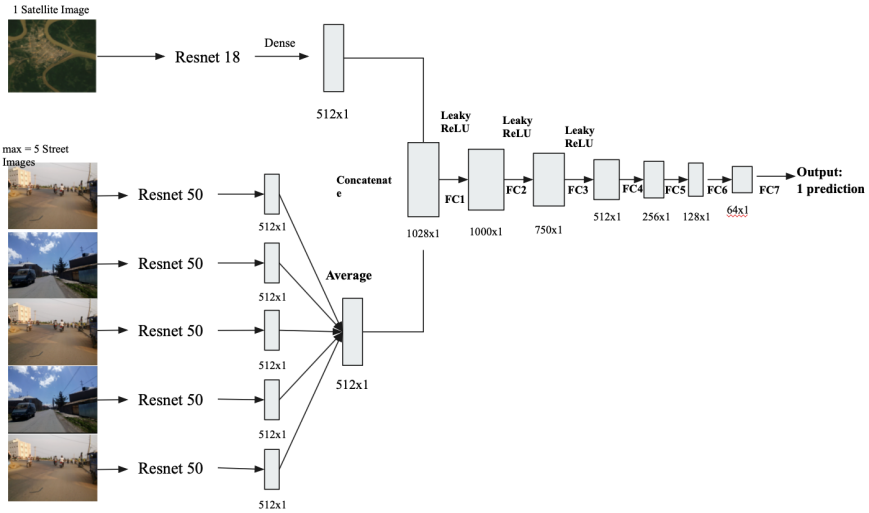

Figure 2: Final Multi-Modal Network Architecture.

5 Experiments/Results/Discussion

5.1 Baseline Model

Our baseline model achieved a RMSE of 1.831. This is comparable to the standard deviation of the level of women's educational attainment of 2.834.

We conducted experiment of several hyperparameters, such as the batch size, learning rate, number of epochs, and weight decay used in training. Finally, we chose the final hyperparameters: batch size = 128, LR = 0.0001, number of epochs = 10, and weight decay = 0.0001.

Baseline Model (only on Satellite Images)

Model	R-Squared	MSE	RMSE	MAE	MAPE
ResNet-18	0.400	3.353	1.831	1.455	0.341
ResNet-34	0.309	3.857	1.964	1.542	0.355
ResNet-50	0.281	4.017	2.004	1.592	0.370

Table 1: Performance of ResNet Models.

5.2 Multi-Modal CNN: Results

Our first model consisting of 1 Fully-Connected layer had an extremely high test MSE of 354. Thus, we experimented with several hyperparameters and 3 different architectures to see if we could improve the model. We tested different number of fully-connected layers (1,3,7) and the addition of Leaky ReLU activations (between 3 of the 7 FC layers). We experimented with Dropout but it increased training time exponentially and produced worse test MSE. Thus, we decided not to use dropout layers. Our experimental results are listed in detail below. Additionally, we implemented 2 different models which can be seen below, one with 5 street images and one with 1 street image, test if street-level imagery is adding pure noise.

Model	R-Squared	MSE	RMSE	MAE	MAPE
5 Street	-0.483	13.26	3.64	2.98	0.42
1 Street	-2.78	38.86	6.23	4.72	0.73

Table 2: Performance of Multi-Modal Models.

Hyperparameter	Value
Batch Size	64
Learning Rate	0.0001
Number of Epochs	50
Weight Decay	0.001

Table 3: Hyperparameters of Best Multi-Modal Model.

Model 1: 5 Street-Level images with 1 Satellite Image (WD = 0.001 except for 1FC where WD = 0.0001)

Architecture	Batch Size	Epochs	LR	Train MSE	Dev MSE	Test MSE	Test RMSE	Test R2
7FC-3ReLU	64	50	0.0001	1.669	6.891	13.255	3.641	-0.482
7FC-3ReLU	64	50	0.00005	1.467	7.348	13.891	3.727	-0.553
7FC-3ReLU	64	50	0.00001	2.029	8.074	23.750	4.873	-1.294
7FC	64	50	0.0001	1.208	8.294	27.136	5.209	-1.515
7FC	32	50	0.0001	3.712	8.865	29.812	5.460	-1.763
3FC	32	30	0.0001	1.660	7.730	30.746	5.545	-1.850
3FC	64	30	0.01	11.782	16.420	63.154	7.947	-
3FC-1ReLU	64	30	0.001	10.354	11.456	10.412	3.227	-0.029
1FC	128	50	0.003	-	-	354.830	18.837	-41.073

Model 2: 1 Street-Level Image with 1 Satellite Image (weight decay = 0.001)

Architecture	Batch Size	Epochs	LR	Test MSE	Test RMSE	Test R2
7FC-3ReLU	64	50	0.0001	38.861	6.234	-2.781
3FC	64	50	0.0001	23.180	4.815	-1.745
3FC	32	25	0.0001	76.215	8.115	-8.037

6 Conclusion/Future Work

We found that 7 FC layers with 3 Leaky ReLU activations model was the highest performing out of all of our experiments. We think that this architecture with non-linear activations works better because the underlying relationship between inputs and outputs is non-linear. Furthermore, our best model performs worse than our baseline with satellite imagery performs better with RMSE of 1.831 which is within 1 standard deviation (2.834) of women's education attainment. Based on these initial results, street-level images do not seem to be correlated with women's educational attainment and we hypothesize that street-level imagery may be adding noise to the model.However, when we tested this hypothesis further (1 street image vs. 5) our model with 5 street images outperforms the one with 1 street image. Thus, we cannot conclusively say that the addition of street-level images is adding pure noise.

For future work, if we have more computational resources we would benefit from using a larger training data set, and possibly incorporating more than five satellite images for each example to conclusively test our hypothesis about street images. In the same vein, we could run a baseline model with only street-level images. Additionally, our best model suggests a non-linear relationship so a deeper network with more non-linear activations could potentially improve results as well. We may also benefit from using data from other countries to make our model more robust against different types of satellites and street images that may be collected from diverse countries and environments. We could also potentially improve our model by limiting the outputs to be between 0 and 18 (the range of women's educational attainment).

If we have more time, we may also look into defining our own loss function instead of using MSE as our objective loss function. This might be helpful for our task since we are combining two different types of images as inputs, and each input type could contribute differently to the difficulty of the learning task. If, for instance, satellite images prove to be more informative in predicting women's educational attainment, then we might want to put larger weight on the loss term that minimizes the error from satellite images inputs.

7 Contributions

Both members contributed to the project. Disa Alda Naomi implemented the data pre-processing pipeline, coded for baseline satellite models, and wrote the base code for the multi-modal network. Haya Hidayatullah imported data for data processing, built data set for combined images, focused on coding and debugging for the multi-modal model, ran hyperparameter tuning experiments and experimented with different architectures to improve models. Both members worked on the final project report and the final project presentation.

References

[1] "DHS Survey-Based Datasets." SustainBench, https://sustainlab-group.github.io/sustainbench/docs/datasets/dhs.html#references.

[2] K. Gadzicki, R. Khamsehashari and C. Zetzsche, "Early vs Late Fusion in Multimodal Convolutional Neural Networks," 2020 IEEE 23rd International Conference on Information Fusion (FUSION), 2020, pp. 1-6, doi: 10.23919/FUSION45008.2020.9190246.

[3] Lee, J, et al. "Predicting Livelihood Indicators from Community-Generated Street-Level Imagery." Proceedings of the AAAI Conference on Artificial Intelligence, 35(1):268–276, 5 2021. ISSN 2374-3468. URL https://ojs.aaai.org/index.php/AAAI/article/view/16101.

[4] "Models and Pre-Trained Weights¶." Models and Pre-Trained Weights - Torchvision 0.14 Documentation, https://pytorch.org/vision/stable/models.html.

[5] T. Baltrušaitis, C. Ahuja and L. -P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423-443, 1 Feb. 2019, doi: 10.1109/TPAMI.2018.2798607.

[6] Yeh, Christopher, et al. "SustainBench: Benchmarks for Monitoring the Sustainable Development Goals with Machine Learning." arXiv preprint arXiv:2111.04724 (2021).

[7] Yeh, Christopher, et al. "Using publicly available satellite imagery and deep learning to understand economic well-being in Africa." Nature Communications, 11(1), 5 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-58916185-w. https://www.nature.com/articles/s41467-020-16185-w.

[8] Nguyen, A, et al. "Predicting Water Quality Index from Urban Satellite and Street Level Imagery using a Multi-Modal CNN." CS 230 Fall 2021 Project http://cs230.stanford.edu/projects_fall_2021/reports/103169540.pdf