CS230

Predicting Water Quality Index from Urban Satellite and Street Level Imagery using a Multi-Modal CNN

Anh Nguyen
Department of Computer Science
anhng@stanford.edu

Xinqi Wang
Department of Computer Science
xinqiw@stanford.edu

Selena Sun
Department of Computer Science
selenas@stanford.edu

Abstract

Access to clean water is a crucial metric to keep track of human development in many countries. However, collecting data for water quality has been labor-intensive and expensive. On the other hand, with the ever growing number of publicly available datasets for geographical images, and the rapid development of machine learning/deep learning techniques, application of these innovations to prediction of water conditions presents an interesting approach. In this project, we attempt to explore the correlation between street-level imagery combined with satellite imagery and water quality. We architecture a novel multi-modal model that takes in the average features learned from CNN model trained on the street-level dataset concatenated with the features learned from a CNN model trained on satellite dataset. The result is a combined effect of each street and satellite imagery trained separately on a CNN model. Our test MSE of 1.238 indicates that although street-level images are better in predicting water quality, these images alone are not enough to make informative predictions.

1 Introduction

There has been a lack of progress in the United Nations Sustainable Development Goals (SDGs) due to a lack of data on key environmental and socioeconomic indicators. The indicators are calculated through the analysis of ground survey data, which is limited in quantity [1]. Machine learning (ML) and deep learning (DL) have made it possible to use globally available data to make progress towards the SDGs [2-3].

We work with the SustainLab at Stanford to create a model to monitor progress for the SDG of ensuring access to safe drinking water for all by 2030 [1]. We explore and evaluate satellite and street level images as sources to predict Water Index (1-5 continuous scale, where 5 is the "highest quality") of local clusters using a novel multi-modal CNN architecture. Although models have been built to use satellite images to predict Asset Wealth Indices, none have incorporated street-level images, which can add more granularity to the model's predictions. Our multi-modal approach can be utilized to inform policy, financial allocation, and urban planning decisions to more effectively make advances towards the SDG of equitable access to drinking water.

CS230: Deep Learning, Fall 2021, Stanford University, CA.

2 Related work

The approaches to measuring water quality index range from Convolutional Neural Networks to Long Short-Term Memory Networks (LSTMs) to Graph Convolutional Networks (GCNs). Some literature is surrounding a time series-based estimation of water quality, given a series of input measurements from the water [9][10]. In [9], Liu et al. used a LSTM deep neural network to measure water quality indicators in the Yangtze River basin (pH, dissolved oxygen, chemical oxygen demand, etc.). However, these methods are sensor-intensive and unable to generalize to an entire region. They necessitate the reading of constant sensors located in the water, and require an on-the-ground team to ensure the hardware is deployed correctly.

Other methods have used satellite images to predict certain indicators over a region of land. One model deploys a deep learning CNN on publicly available satellite imagery, and is able to explain 70% of variation in wealth in most countries it was tested on[2]. However, it seems that the fidelity of the measurement could further be improved by adding street-level images, as suggested by [11]. The street-level image approach was particularly clever, since it used images closer to the ground to predict livelihood indicators, a metric that intuitively seems more accurately predicted from closer-up images. This paper also deployed three learning methods (image-wise learning, cluster-wise learning, and cluster-wise GCN learning) on the street-level images, where they saw varying degrees of success between all three methods in different countries (Kenya vs. India).

The benchmark model for the water quality index has an accuracy of 0.4 using a k-nearest neighbors (kNN) model trained on only satellite images[4]. Inspired by the success of using street-level images to predict livelihood indicators, we design a model that separately trains a CNN model on street-level images and satellite images, then aggregates the learned features from these CNN models. The aggregated features are then fed into a NN model to predict water quality index of a cluster.

3 Dataset and Features

We used both satellite and street-level images to train our model. The labels were sourced from the DHS survey. All of this data was aggregated by the SustainLab at Stanford: www.github.com/sustainlab-group/sustainbench.

3.1 Water Quality Index Survey Data

The DHS surveys provide the water quality for each household surveyed (water quality is ranked on a 1-5 continuous scale, where 5 is the "highest quality"). The Water Quality Index is the average score of households in a cluster. Sustainlab has summarized the household-level data into "cluster-level" labels, where a "cluster" roughly corresponds to a village or local community [4]. 179 DHS surveys from 56 countries spanning 1996-2019 were used to create labels.

3.2 Satellite and Street-level Imagery

The SustainBench dataset provides collections of images under different country codes, ranging from approximately 100-400 images per country per year. The total number of satellite images in the dataset having a valid water quality index label is 87,938. Around 20% of the satellite images have at least one corresponding street image. For each of these satellite images, there are between 1 to 100 street-level images (~800,000 total images).

Satellite Images:

The satellite imagery consists of both daytime images (multispectral - MS) from the Landsat 5/7/8 satellites and nightlights (NL) images from the DMSP and VIIRS satellites [4]. For each cluster from a DHS survey in a country by year, a 255x255x8 image (7 MS bands, 1 NL band) is provided.

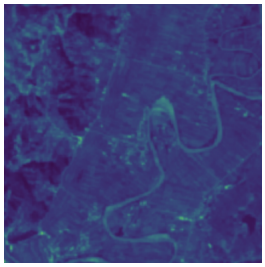
We sampled a total of 4,500 satellite images. We collected these images by randomly sampling 100 images from each set of images grouped by the DHS survey in a country by year.
We split the dataset into train/validation/test buckets with the ratio 80/11/8, roughly abiding by

standard practices. Specifically, we split the countries into the following groups:

- Train set = 12 countries: TZ, BF, CM, GH, IA, KM, LS, ML, MW, NG, PH, TG
- Validation set = 4 countries: BJ, BO, CO, DR
- Test set = 2 countries: AM, AO

Note: There are 7 countries with missing data that we excluded: HN, ID, JO, KH, MA, MB, NI.

Figure 1.: An example of a satellite image



Street-Level Images:

A maximum of 100 images within 0.01 degrees lat/long of a DHS cluster and were captured within 1 year of a DHS cluster datapoint were retrieved and labelled with their corresponding cluster [4]. The raw images have 3 channels (RGB) and were preprocessed into 236 x 256 x 3 vectors. The resolution of the original images is approximately (250-1000)x(250-1000) pixels.

- Train set = 6 countries: CD, MD, ZW, CM, GH, NP
- Validation set = 2 countries: BJ, BO
- Test set = 2 countries: AM, AO

Figure 2.: Street-level images

3.3 Summary of Data Used in Separate Models

Dataset	Resolution	Train Examples	Dev Examples	Test Examples
Satellite	255*255*8	3600	500	400
Street	256*256*3	10279	1152	1000

4 Methods

We modify the pretrained ResNet50 model from tensorflow to train satellite imagery and street-level imagery separately, and extract the second last layer from each model to concatenate them and input into a regression model. These images were trained separately since nature of the features for satellite imagery and street-level imagery are different. ResNet50 in Tensorflow is pretrained on ImageNet dataset. Transfer learning are used for both satellite model and street model because lower level

features in general images learned by ResNet50 are useful for our data. ResNet50 in Tensorflow only accept data with 3 channels, we modify it to accept 8-channel data.

Water quality index data is used as labels input into our model. Each label is a numeric value ranges from 1 to 5. So the output of our model is bounded data in [1, 5]. We perform normalization on input labels to be ranged in [0, 1], use Sigmoid activation as the final layer and scale the output back from [0, 1] to [1, 5].

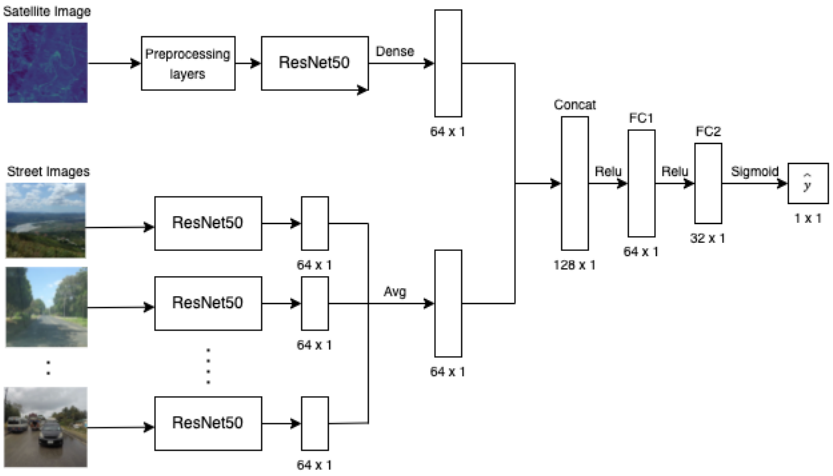
After training each CNN model separately for satellite and street-level imagery to predict the water quality index for each cluster, we extract 64 features for each satellite image and street image. For each satellite image, the average of the features for its corresponding street-level images is calculated and concatenated to the feature for this satellite image. This creates a feature vector of 128 features combining the satellite image and street-level images for a cluster.

In other words, our aggregate feature vector $h(f(x), g(r_1), ..., g(r_{100}))$ is

$$h(f(x), g(r_1), ..., g(r_k)) = [f(x), (\frac{1}{k}) * (g(r_1) + ... + g(r_k))]$$

We then train a neural network (NN) on these aggregated feature vectors make predictions for water quality index.

Figure 3.: Model Architecture



5 Experiments/Results/Discussion

We use mean squared error (MSE) loss when training and validating, and also use it to measure the performance of predictions.

5.1 Architecture Choices

We experiment with two CNN models during training. One is a 7-layer small CNN defined by ourselves, and one is ResNet50 with transfer learning. The small CNN gives lower test MSE loss than our modified ResNet50 model when trained on a small dataset (around 400 examples). However, on our sampled dataset (more than 3000 examples in train set), our modified ResNet50 gives better result. A probable cause is that the large architecture of ResNet50 requires an equally large dataset to learn meaningful gradient information. Since the ResNet50 performs better on large datasets, we decide to use our modified ResNet50 as the model to train the satellite and street-level imagery.

5.2 Hyperparameter Tuning Experiments and MSE Results

Due to the computation limitation of our EC2 instance, we are not able to set a batch size larger than 32 and an epoch number larger than 15. We choose Optimizer Adam and use the default learning rate 0.001.

Test MSE is calculated based on predictions and normalized input labels. Normalized labels ranges in [0, 1]. The true lables ranges in [1, 5]. Test MSE on Scaled Output is calculated by true labels and prediction values scaled back from [0, 1] to [1, 5]. Please refer to Appendix to see Hyperparameter Tuning Experiments and MSE Results Table.

5.3 Best Training Results

Model	Batch Size	Epoch	Frozen Layers Num	Dropout	Train MSE	Dev MSE	Test MSE	Test MSE on Scaled Output
Satellite	32	10	35	None	0.045	0.051	0.101	1.624
Street	32	5	45	None	0.147	0.075	0.060	0.965
Aggregate (2 FC)	16	10	N/A	None	0.052	0.040	0.077	1.238

5.4 Discussion

The overall high test MSEs suggest that satellite imagery and street-level imagery have weak correlation with water quality index. The best MSE result on scaled output for satellite model, street model, and aggregation model is higher than 0.9 while water quality index ranges from 1 to 5.

The average test MSE for the street imagery is lower than that for the satellite imagery could be attributed to the fact that satellite images are zoomed out representation of a location and thus cannot capture geographical nuances relating to water quality as well as the ground-level features learned from street imagery. The average test MSE for the aggregation model is higher than that for street-level imagery but lower than the test MSE for satellite imagery, demonstrating a blended effect of the two types of datasets.

To avoid overfit, we experiment with adding 1-layer Dropout for both satellite model and street model. As for satellite model, there should be no overfit issue because after we add 0.2 dropout, the result is much worse. The reason could be that satellite model cannot learn enough information from the weak correlation between input imagery and labels. Additionally, adding dropout will make the model learn even less. As for street model, 0.2 dropout does not result in a significant difference. Further, 0.5 dropout encounter a worse Test MSE. Based on these results,we decide not to add dropout to our final models.

6 Conclusion/Future Work

In this academic project from using data from the SustainBench project, we investigated the correlation between satellite/street-level imagery and water quality index using a novel multimodal NN model. Our results suggest that datasets for these images, alone or combined, are not significant indicators for water quality prediction even though street data performs slightly better than satellite imagery alone and both of them combined.

There are several improvements on the datasets that can be promising. Currently train/dev/test sets are from different countries. Different countries may have different types of terrains, weather and populations. Data learned from some countries may not be applied to others. One can consider sampling data from geographically similar or neighboring countries to achieve a relatively same distribution on dev and test sets. Specifically, one can experiment with putting countries from the same continent or latitude in dev and set. In addition, blurry and general satellite imagery may not be able to provide clear enough information on water quality of a location. Unmasking techniques or data filtering efforts to generate satellite images with unhindered view on reservoirs might be helpful.

More explorations on the model are also worth to try. With more powerful computational resources, we could expand our search for hyperparamters on the whole dataset provided by SustainBench. Different ways to aggregate satellite and street features can also be explored. For example, NN-based ideas such as attention-based transformer network, GCNs or "deep sets" models might prove to be more powerful in leveraging satellite and street image information for water quality prediction.

7 Contributions

All members contributed plenty of effort to this project. Each member had specific aspects to focus on and there were also collaboration works. Anh focused on importing data, preprocessing street images, extracting features, and coding up and tuning the aggregation model. Xinqi focused on the satellite model, hyperparameter tuning experiments for both satellite model and street model, and providing code skeletons for all models. Selena focused on coding up the street model and report/poster write-ups. All team members contributed to the final report and the video.

Acknowledgments

We would like to thank the Stanford Sustain Lab and Christopher Yeh, for their support and suggestions.

References

[1]"The Sustainable Development Goals Report 2021." The Sustainable Development Goals Report, 2021. https://doi.org/10.18356/9789210056083.

[2] C. Yeh, A. Perez, A. Driscoll, G. Azzari, Z. Tang, D. Lobell, S. Ermon, and M. Burke. "Us ing 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.

[3] J. Lee, D. Grosz, B. Uzkent, S. Zeng, M. Burke, D. Lobell, and S. Ermon. "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. https://ojs.aaai.org/index.php/AAAI/article/view/16101.

[4] C. Yeh, C. Meng, S. Wang, A. Driscoll, E. Rozi, P. Liu, J. Lee, M. Burke, D. Lobell, and S. Ermon. "SustainBench: Benchmarks for Monitoring the Sustainable Development Goals with Machine Learning." 2021. https://openreview.net/forum?id=5HR3vCylqDnoteId=FL6Sr6Ks0J.

[5] Theyazn H. H Aldhyani, Mohammed Al-Yaari, Hasan Alkahtani, and Mashael Maashi4. "Water Quality Prediction Using Artificial Intelligence Algorithms." Applied Bionics and Biomechanics, 2020. https://doi.org/10.1155/2020/6659314.

[6] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention Is All You Need." arXiv.org, December 6, 2017. https://arxiv.org/abs/1706.03762.

[7] Sanchez-Lengeling, Benjamin, Emily Reif, Adam Pearce, and Alexander B. Wiltschko. "A Gentle Introduc tion to Graph Neural Networks." Distill, September 8, 2021. https://distill.pub/2021/gnn-intro/.

[8] Zaheer, Manzil, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. "Deep Sets." arXiv.org, April 14, 2018. https://arxiv.org/abs/1703.06114.

[9] Liu, Ping, et al. "Analysis and Prediction of Water Quality Using LSTM Deep Neural Networks in IOT Environment." MDPI, Multidisciplinary Digital Publishing Institute, 7 Apr. 2019, https://www.mdpi.com/2071-1050/11/7/2058.

[10] Baek, Sang-Soo, et al. "Prediction of Water Level and Water Quality Using a CNN-LSTM Combined Deep Learning Approach." MDPI, Multidisciplinary Digital Publishing Institute, 3 Dec. 2020, https://www.mdpi.com/2073-4441/12/12/3399.

[11] Lee, Jihyeon, et al. Predicting Livelihood Indicators from Community-Generated Street-Level Imagery. 26 Feb. 2021, https://arxiv.org/pdf/2006.08661v6.pdf.

Appendix

7.1 Hyperparameter Tuning Experiments and MSE Results

Exp No.	Model	Batch Size	Epoch	Frozen Layers Num	Dropout	Train MSE	Dev MSE	Test MSE	Test MSE on Scaled Output
1	Satellite	16	10	45	None	0.176	0.105	0.182	2.915
2	Satellite	32	10	45	None	0.064	0.076	0.128	2.056
3	Satellite	32	15	45	None	0.175	0.102	0.175	2.805
4	Satellite	32	10	35	None	0.045	0.051	0.101	1.624
5	Satellite	32	10	25	None	0.176	0.104	0.181	2.896
6	Satellite	32	10	35	0.2	0.170	0.101	0.179	2.864
7	Street	16	5	45	None	0.045	0.043	0.075	1.213
8	Street	32	5	45	None	0.147	0.075	0.060	0.965
9	Street	32	5	35	None	0.151	0.081	0.072	1.152
10	Street	32	5	45	0.2	0.147	0.075	0.060	0.965
11	Street	32	5	45	0.5	0.034	0.040	0.068	1.099
12	Aggregate (1 FC)	16	10	N/A	None	0.053	0.040	0.080	1.285
13	Aggregate (1 FC)	32	10	N/A	None	0.054	0.040	0.078	1.245
14	Aggregate (2 FC)	16	10	N/A	None	0.052	0.040	0.077	1.238
15	Aggregate (2 FC)	32	10	N/A	None	0.053	0.040	0.078	1.254