Abstract

In this project, the researchers develop several models to classify whether images of houses contain a solar panel or do not contain a solar panel. Two models were determined as effectively classifying the images into the correct categories, as compared to a basic logistic regression model. The first was a Histogram of Oriented Gradients Support Vector Machine, with an accuracy on the test set of 83%. The second model was a Computational Neural Network, with an accuracy of correctly classifying the test images of 96%. Together, these models provide an effective tool to allow for the classification of residential homes as either having or not having solar panels. These tools can be used to further understand the distribution of solar panels in residential areas, allowing governments or other agencies to examine the effectiveness of their policy proposals as related to alternative energy and climate change.

Introduction

Global climate change is a serious problem with far reaching, long term consequences to all nations and economies [1](https://www.zotero.org/google-docs/?26hLyC). However, these technologies still continue to be very expensive compared to more traditional means of energy generation, such as coal plants. To try and counteract this, many governments, including the government of the United States, have funnelled money into both alternative energy research and alternative energy production. In the US, many states also have programs designed to increase solar PV energy, such as the Duke-Energy Solar Rebate Program [2](https://www.zotero.org/google-docs/?alTmN1). Despite numerous government funds, these technologies still only constitute a minority of energy generation in the US; the US Energy Information Administration states that renewables only accounted for 17.5 percent of energy generation in 2019 [3](https://www.zotero.org/google-docs/?D50rCg). The ultimate goal of this project will be to help the government inform best policies to encourage a broader use of renewable energy sources, particularly as applied to residential consumers. To that end, this project will focus on residential consumer’s use of solar PV arrays on their houses. This is an attractive research group because it can measure both popular support for renewable energy sources (as opposed to industries complying with federal law) and the effect of government policies to encourage the use of renewable energy sources.

Specifically, this project seeks to use satellite imagery to computationally determine if a house has a solar PV array. This can help governments to determine if their policies of financially supporting solar PV arrays are yielding more solar panels on private homes. This process can then be used in conjunction with other studies to determine if government policies concerning residential solar PV arrays are effective at increasing the use of solar panels.

Background

Computer Vision has been a growing field in the past several decades. Recently, with the combination of feature extraction and machine learning methods, the accuracy of the models that allow computers to be able to “see” and identify specific objects from pictures has risen dramatically. This project seeks to use these methods to recognize solar panels from satellite images.

1.1 Solar panel identification from satellite images

Solar panel identification is beneficial for both government and private organizations since they can make energy policies or marketing plans by analyzing the solar development patterns. There are already some previous studies applied to machine learning methods to identify solar panels from satellite images. For example, “The DeepSolar Project” conducted by Stanford University, “constructed a comprehensive high-fidelity solar deployment database for the contiguous U.S.” 4 The purpose of the project is to create a database for people who are interested in how the solar panels are installed in the different area in the U.S. to conduct further analysis or research. The team applied the classification method based on Google Inception V3 to identify whether there is any solar panel in the piece of the image.4 If an image is classified as having a solar PV array, segmentation using the CAMs method would be conducted to estimate the size of the solar panels.5 Though the purpose of the project is slightly different from this current project, the basic steps are the same as this project- detect solar panels from aerial images. Image processing and image classification are other important parts of this project to identify solar panels from the images.

1.2 Image Processing

Image processing is a subfield of signal processing, which uses computers to process digital images. This has been studied for decades since the 1950s.6 SSince the digital images are represented using matrices, this allows scaling, color conversion, image enhancement, and other useful methods to be applied to the image by adjusting the value in the matrices. Moreover, image processing can be used to filter out the information from the high dimensional features of the images. In a study of detecting plant diseases, image processing was applied to the image, for example, transforming the color into greyscale in order to filter out extraneous information then conduct image classification.7 Similar image processing methods are conducted in this project to prepare training data for the chosen image classification models.

1.3 Image Classification

Image classification “refers to a process in computer vision that can classify an image according to its visual content”.8 However, because of the high dimensions of features of the images, it is computationally infeasible to use all these features to train models. Moreover, “look[ing] at high-resolution images it is very likely that a neighboring pixel belongs to the same land cover class as the pixel under consideration.”7 Therefore, “implementing feature extraction, and selecting suitable variables for input into a classification procedure are all important” 9 since we can make full use of the features and also reduce the dimension of the data. Some image processing methods such as Histogram of Oriented Gradients (HOG) 10or Scale-Invariant Feature Transform (SIFT)11 are famous as tools to extract informative features from the original image. With these filtered features, researchers can apply supervised machine learning methods to conduct image classification with high accuracy compared to using original images. One of the frequently used methods that can be applied to image classification are Support Vector Machines (SVM).12 In the study of image classification, it is stated that SVM “can generalize well on difficult image classification problems where the only features are high dimensional histograms,” by comparing the performance with KNN-based models, tuning the SVM model itself, remapping the input data.13 The analysis process of this project would be similar to the previous work since the goal of this project is to find a good combination of transformed original data and a well-tuned model to identify solar panels in the image with high accuracy.

Data

The data for this project was received as part of a kaggle competition. The data consists of a labeled test and training set of residential housing images. The images are a three channel visible spectrum color aerial imagery, from several locations in California [8](https://www.zotero.org/google-docs/?7mTHZu). The labels state whether the image contains a solar panel (labeled as 1) or does not contain a solar panel (labeled as 0). The data was previously manually processed by Bradbury et al. to only have a single house per image, and the sets were manually labeled. The training data is slightly skewed towards images which do not contain a label:

(Image here showing the skewness of the data)

The images are all of residential homes; below is a representative example of a house without solar panels:

This is an image of a representative example of a house with solar panels:

Conclusions

Out of the models created, the two best models, which achieved a significantly higher accuracy than a basic logistic regression, were the HOG SVM and the CNN models. The HOG SVM performed acceptably, achieving an accuracy of 83% on the test dataset. This model is much faster than the CNN, and could have applications where speed is necessary, but accuracy is not as much of a concern. In terms of strict accuracy, the CNN performed much better, with an accuracy of 96% at classifying images as correctly having or not having a solar panel on the house. The CNN clearly has an accuracy advantage over all other models at correctly classifying the images, however, it was much slower than any other model. This model would be preferred when accuracy is required, but speed is not a concern. As described in the introduction above, the solar panel image classification goal of this project could help inform policymakers or other stakeholders as they seek to experiment with and determine the best method to increase solar panel use. With an accuracy of 96%, the CNN model could be used to effectively determine the efficacy of policies targeted at increasing solar panel usage. This can be done by feeding satellite images into the model, allowing any level of government to analyze their policies effectiveness. Using this information, the government can then tailor the policies to further experiment with policies and legislation, with the ultimate goal of increasing solar panel usage, reducing reliance on fossil fuels, and combating climate change. Further research could include ways to boost the accuracy of this model even further, and to apply these image classification techniques to further questions of interest.

Roles

For all the group assignments Team 5 uses a rotating leadership (a new leader per assignment). This gives each member of the Team Project Management Experience, creative control, and helps distribute workload. For this assignment Derek was the team leader and he decided to split the project into two separate teams, one that was Derek and Andrew Team One, and Team Two (Akshay, Melody, and Tzu-Chun). Team One focused on building the baseline/non-CNN models and writing and editing the report. Team One built several different models to test performance, the first being the group baseline of logistic regression as well as several other methods such as Naïve Bayes and Random Forest. These ultimately did not perform as well as the HOG SVM and CNN.

Derek focused on building the baseline models, and worked on the roles section of the report. Andrew ran several test models, and supported Derek in his work on the baseline models. Andrew also wrote the abstract, introduction, data, and conclusion sections, and edited the rest of the report.

Team Two focused on building more sophisticated models because both Akshay and Melody had prior experience with machine learning. Akshay focused on constructing the SVM HOG model, and wrote the methods and results section for that part of the project. Melody focused on constructing the CNN, and wrote the methods and results section for that model. Tzu-Chun ran several models, and supported Akshay and Melody. Tzu-Chun also wrote the background section of the report.

References

1. Mora, C. et al. Broad threat to humanity from cumulative climate hazards intensified by greenhouse gas emissions. Nat. Clim. Change 8, 1062–1071 (2018).

2. DSIRE. https://programs.dsireusa.org/system/program?zipcode=27517.

3. What is U.S. electricity generation by energy source? - FAQ - U.S. Energy Information Administration (EIA). https://www.eia.gov/tools/faqs/faq.php?id=427&t=3.

4. Yu, J., Wang, Z., Majumdar, A. & Rajagopal, R. DeepSolar: A Machine Learning Framework to Efficiently Construct a Solar Deployment Database in the United States. Joule 2, 2605–2617 (2018).

5. Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, Antonio Torralba 2016. Learning Deep Features for Discriminative Localization. Retrived from: ”http://cnnlocalization.csail.mit.edu/Zhou\_Learning\_Deep\_Features\_CVPR\_2016\_paper.pdf”.

6. Lipkin, L. <strong>Picture Processing by Computer.</strong> <span class="sc">Azriel Rosenfeld.</span> Academic Press, New York, 1969. x + 198 pp., illus. $11.50. Computer Science and Applied Mathematics. Science 169, 166 (1970).

7. Jayme Garcia Arnal Barbedo 2013. Digital image processing techniques for detecting, quantifying and classifying plant diseases. Retrived from: “https://link.springer.com/article/10.1186/2193-1801-2-660”.

8. David Kaeli, Perhaad Mistry, Dana Schaa, Dong Ping Zhang 2015. Heterogeneous Computing with OpenCL 2.0- Chapter 9.

9. Lu, D. & Weng, Q. A survey of image classification methods and techniques for improving classification performance. Int. J. Remote Sens. 28, 823–870 (2007).

10. Navneet Dalal and Bill Triggs 2005. Histograms of Oriented Gradients for Human Detection. Retrived from: “https://lear.inrialpes.fr/people/triggs/pubs/Dalal-cvpr05.pdf”.

11. Lowe, D. G. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 60, 91–110 (2004).

12. Corinna Cortes and Vladimir Vapnik 1995. Support-Vector Networks. Retrived from: “http://image.diku.dk/imagecanon/material/cortes\_vapnik95.pdf”.

13. Olivier Chapelle, Patrick Haffner, and Vladimir N. Vapnik 1999. Support Vector Machines for Histogram-Based Image Classification. Retrived from: “http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.459.9821&rep=rep1&type=pdf”.

14. Bradbury, Kyle; Saboo, Raghav; Johnson, Timothy; Malof, Jordan; Devarajan, Arjun; Zhang, Wuming; et al. (2016): Modesto Aerial USGS Imagery from the Distributed Solar Photovoltaic Array Location and Extent Data Set. figshare. Dataset. https://doi.org/10.6084/m9.figshare.3385789.v1.