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. Most energy generation in the world comes from fossil fuels, the byproducts of which are implicated in causing the current phase of global warming. Much research has gone into looking for alternative sources of energy, including wind and solar power. 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. Despite numerous government funds, these technologies still only constitute a minority of energy generation in the US. 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 panels 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. There have been numerous programs in several US states to try and encourage the use of solar panels, including <https://www.energysage.com/solar/cost-benefit/solar-incentives-and-rebates/>

This project seeks to use satellite imagery to computationally determine if a house has a solar panel. This can help governments to determine if their policies of financially supporting solar panels 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 panels are effective at increasing the use of solar panels.

Background

Image classification

This project builds on a st examining image classification. In many ways, this work is a continuation of early contextual pattern recognitions from the 1960’s and 1970’s (look at toissant’s paper, find more papers to quote) <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.72.5050&rep=rep1&type=pdf>

Further developments in computer technology and machine learning have made image classification a thriving field.

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 labels state whether the image contains a solar panel (labeled as 1) or does not contain a solar panel (labeled as 0). The training data is slightly skewed towards images which do not contain a label:

(Image here showing data)

(For the data section, I am not sure what else to put).

Methods

Three models were used which were eventually decided on as the final three models. The three final models were a logistic regression model (to set a baseline to test the other models against), a HOG SVM, and a CNN.

(I am not sure who’s logistic regression we are using as the baseline, but I wrote up a bit)

To create the logistic regression model, the images first underwent a principle component analysis process which reduced the dimensions from the RGB color scheme into grayscale gradient. The training grayscale images were then used to fit a logistic regression model,

(HOG SVM Akshay will most likely write it up)

(CNN is Melody’s job)

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

Akshay used a Histogram of Oriented Gradients PCA, combined with a support vector machine model, to prepare the second model submitted for the final part of the competition. Akshay also wrote (parts of the report). Tzu Chun made several models, including (add more). Derek Wales made several models, including (add more). Melody Li created a Computational Neural Network model for the images, which was then submitted as part of the final part of the competition. Andrew Patterson created and tested several models while using the grayscale component reduction, including a test of a basic logistic regression and an experimental QDA test, both of which were submitted as part of the competition. Andrew also wrote the introduction, background, conclusion, references, and other parts of the rest of the report.

References