Improving a Rooftop Solar Panel Segmentation Model's Robustness to Domain Shifts

Stanford CS231N Project

Kerrie Wu
Stanford University
kerriewu@stanford.edu

Julian Cooper
Stanford University
jelc@stanford.edu

Andrea van den Haak
Stanford University
vandenhaak@stanford.edu

Problem & Motivation. We are interested in segmenting and classifying rooftop solar panels from satellite images, with a focus on residential areas. In many countries (including the United States), the rollout of rooftop solar has been ad hoc and largely untracked. Besides industrial sites and city buildings, electricity utilities are largely blind to where and how much solar neighborhoods have accumulated. Being able to estimate this (and its rate of increase) allows utilities to predict power output reasonably well (from meteorology forecasts) and adjust grid management and investment.

While Stanford's DeepSolar team have largely solved this challenge for the United States **? ?**, their model is not robust to "distribution shifts". This means that the model accuracy drops when tested on datasets from other regions with different satellite imagery to its USA-based training data. With the help of the DeepSolar team, we want to investigate different techniques for modifying the DeepSolar segmentation model in order to improve its robustness to distribution shifts. This will improve the model's generalizability for use in countries other than the United States while requiring less (or ideally none) labeled data from each other country for finetuning the model.

Dataset & Evaluation. In the interest of helping to solve this specific problem for rooftop solar segmentation, Kasmi et al **?** recently published a labelled dataset for France (released January 2023). The dataset provides ground truth segmentation masks for 13303 images from Google Earth25 and 7686 images from the French national institute of geographical and forestry information (IGN). We will use this to test our model's ability to handle distribution shifts by training our model on United States data, but validating and testing our model on France data. We will use quantitative metrics such as effective robustness and relative robustness (as defined by **?**) to measure the effectiveness of our approaches, and DICE/F1 scores to measure segmentation model performance irrespective of robustness.

Literature Review. In addition to reviewing the France dataset and DeepSolar codebase, we have also read several papers on techniques for measuring robustness to and handling distribution shift.

- Taori, 2020 **?**: Major contributions included defining effective robustness and relative robustness metrics. The authors also concluded that robustness to synthetic data distribution shifts do not imply robustness to natural distribution shifts, and that a more diverse training set improved robustness.
- Yao, 2022 **?**: Describes a data augmentation method called LISA adds interpolations between original input-output training example pairs to the training data. The interpolation pairs can be selected in a targeted manner to improve robustness to domain shifts.
- Volpi, 2018 **?**: Describes how to adversarially augment the training data to achieve a more robust model. At each training iteration, the training examples fed to the model are augmented with examples that are considered difficult for the current model.
- Huang, 2022 **?**: Describes RefSeq, a method that involves training a separate proxy segmentation-generating model in addition to the main segmentation model, and asking the main segmentation model to "reflect" and refine on its originally predicted segmentation given the proxy segmentation, to achieve a better result at test time.

Proposed Modeling Approach. Based on the survey of the above papers, some of the approaches that we plan to try are interpolation-based data augmentation through LISA **?**, regularized fine-tuning **?**, adversarial data augmentation **?**, and test-time/training adaptation methods involving auxiliary models such as RefSeq **?**. We expect that our experiments will yield plots comparing robustness metrics and DICE/F1 scores across different approaches.