Generalizability of Neural Network-based Identification of PV in Aerial Images

PennState Hazleton

Joseph Ranalli[1*] and Matthias Zech[2]
[1]Penn State Hazleton, Hazleton, PA, USA
[2]DLR Institute of Networked Energy Systems, Oldenburg, Germany
*Corresponding Author: jar339@psu.edu

DLR Deutsches Zentrum für Luft- und Raumfahrt
German Aerospace Center
Institute of Networked Energy Systems

Introduction: Automating PV Detection

Deep learning approaches utilizing Neural Networks (NN) have been applied by investigators to efficiently identify photovoltaic (PV) installations from aerial and satellite photos. Broad application of the technique is limited by access to training data, but labeling training data is very time intensive. Previous studies have observed challenges when applying models trained on one data set to a different test data set, limiting their generalizability. We investigated generalizability of trained networks, by training separate models on six data sets. Each of the models was cross-tested using test data from each other data set to evaluate model performance. In addition to the listed data sets, a synthetic combination was generated as an even sample of images from all six sets.

ID	Location	Source	Tot. Tiles	Tile Size	Resolution	Scaled Res.
CA-F	Fresno, California	USGS Orthoimagery	1,044	625x625 px	0.3 m/px	0.32 m/px
CA-S	Stockton, California	USGS Orthoimagery	4,192	625x625 px	0.3 m/px	0.32 m/px
FR-G	France	Google Earth	13,303	400x400 px	0.1 m/px	0.07 m/px
FR-I	France	French IGN	7,865	400x400 px	0.2 m/px	0.14 m/px
DE-G	Oldenburg, Germany	Google Earth	1,325	639x640 px	0.18 m/px	0.2 m/px
NY-Q	Queens, New York	NYS Orthoimagery	1,007	625x625 px	0.15 m/px	0.16 m/px

Data sets were prepared for NN model use

Several processing steps were required to create consistent data sets. CA-F, CA-S and NY-Q had large images (5000 x 5000 px) that were sliced into smaller tiles. A subset of 1,000 images was randomly selected from each data set. Only images that contained PV were selected for the subsets. When read from disk, images were scaled to 576 x 576 px to match the input size of the NN model. This led to different effective zoom levels across data sets. Data sets reflected geographic variety, including Europe and different areas of the United States. Images were manually categorized based on qualitative inspection to provide a description of different features between the data sets.

CA-F featured more rural and agricultural areas than others. NY-Q had a significant number of flat-roof or commercial systems and had a more urban character.

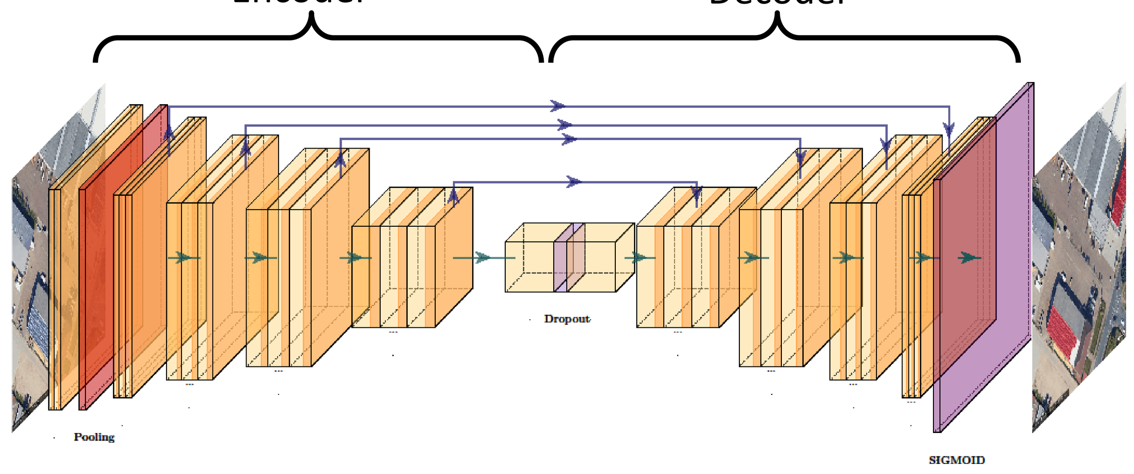
ID	Large/Flat	Open Spaces	Ag.	Water	Util. PV	# Bldg/Tile
CA-F	71	144	38	7	1	20-40
CA-S	74	75	14	39	0	20-40
FR-G	10	23	0	0	0	2-5
FR-I	16	90	15	0	0	5-10
DE-G	55	83	6	7	5	10-20
NY-Q	130	12	0	6	0	10-20

Training used existing libraries and NN architectures

Data sets were randomly split into 200 test, 720 training and 80 validation images. U-net models with a ResNet-34 backbone were initialized with pretrained weights from the ImageNet competition and were trained for up to 200 epochs. Data augmentation was used to virtually expand the data sets, using the parameters listed. The implementation of u-net used was from the *segmentation_models* python library. Weights from the epoch with the lowest validation loss were used for model evaluation.

Method	Value
Rotation	± 30°
Width Shift	± 10%
Height Shift	± 10%
Zoom	± 20%

Data augmentation parameters



U-net's use a coupled encoder-decoder architecture. The encoder transforms the image into a filtered representation. The decoder reverses the process producing an image. This study used u-net with a ResNet-34 backbone.

Models did not generalize well

Performance of each model was evaluated in terms of Intersection over Union (IoU), Precision and Recall. Each model performed best on its own test data. Outside of a few examples of generalization for one other test set, models performed poorly on unseen test data. The combination model performed adequately for most test data, reflecting the fact that it was jointly trained on each of the test data sets.

IoU Score		Test Data					
		CA-F	CA-S	FR-G	FR-I	DE-G	NY-Q
Model	CA-F	0.71	0.35	0.11	0.36	0.06	0.16
	CA-S	0.55	0.61	0.11	0.22	0.17	0.19
	FR-G	0.03	0.00	0.81	0.45	0.13	0.26
	FR-I	0.13	0.19	0.35	0.69	0.31	0.56
	DE-G	0.18	0.29	0.11	0.29	0.63	0.44
	NY-Q	0.07	0.22	0.15	0.47	0.40	0.81
	CMB-6	0.35	0.45	0.67	0.61	0.49	0.68

Precision		Test Data					
		CA-F	CA-S	FR-G	FR-I	DE-G	NY-Q
Model	CA-F	0.87	0.46	0.36	0.48	0.07	0.25
	CA-S	0.82	0.79	0.51	0.31	0.22	0.24
	FR-G	0.10	0.03	0.91	0.76	0.41	0.52
	FR-I	0.63	0.64	0.95	0.79	0.67	0.77
	DE-G	0.70	0.65	0.83	0.91	0.77	0.82
	NY-Q	0.59	0.66	0.90	0.87	0.75	0.90
	CMB-6	0.79	0.80	0.75	0.70	0.72	0.77

Recall		Test Data					
		CA-F	CA-S	FR-G	FR-I	DE-G	NY-Q
Model	CA-F	0.79	0.59	0.15	0.58	0.59	0.35
	CA-S	0.62	0.72	0.13	0.47	0.59	0.37
	FR-G	0.06	0.01	0.88	0.52	0.15	0.29
	FR-I	0.15	0.23	0.36	0.84	0.37	0.67
	DE-G	0.19	0.33	0.11	0.30	0.79	0.48
	NY-Q	0.07	0.24	0.15	0.50	0.47	0.89
	CMB-6	0.38	0.49	0.86	0.82	0.61	0.85

Some inferences can be made by inspecting image results

Examples of all models making predictions on one of the best performance test images from each data set are show below. All models performed relatively well on commercial rooftop systems (e.g. NY-Q picture below). The FR-G data set was particularly hard for models to predict, which we partially attribute to its uniquely high zoom level. A few models seemed to be reliant on the presence of module frames to identify the presence of panels.

Conclusion

We have described some of the strengths and weaknesses of generalization across six separate aerial imagery datasets applied for identification of PV. This study found that generalization of models trained on a single dataset is relatively challenging when applied to other datasets. Models generally did well on large commercial-scale systems, but experienced poor generalization that was not tied to a specific image context. Further work is needed to test methodologies that may improve the generalizability of the trained models and address differences within source data.

Acknowledgments

The authors acknowledge contributions by Richard Ray and Brian Tylutke in labeling the NY-Q data set. J. Ranalli acknowledges financial support by Penn State School of Engineering, Design and Innovation and Penn State Hazleton.