# Fine-tuning transformers for PV detection in varying resolutions

Rajanie Prabha[1,3], Iván Higuera-Mendieta[1,2]
Stanford Doerr School of Sustainability[1], Environmental Change and Human Outcomes (ECHOLab)[2], Sustainable Systems Lab (S3L)[3]

## Introduction

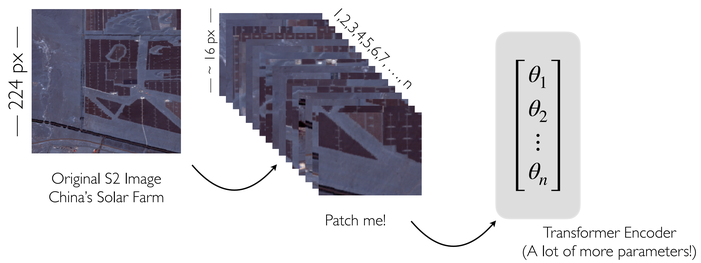
Lack of accurate labels and high-resolution data pose an obstacle to object identification in remote sensing. This problem is evermore present in data-scarce contexts in the global south, where labels and imagery are wanting. We propose a set of experiments using a ViT transformer and a set of different fine-tuning experiments using very-high resolution imagery, and publicly available medium-resolution imagery to explore the classification performance leverage from different fine-tuning experiments. We chose Solar PV detection as our downstream binary classification task.

## Data and architecture

*High-resolution [GoogleMaps]*

*Medium-resolution [Sentinel-2]*

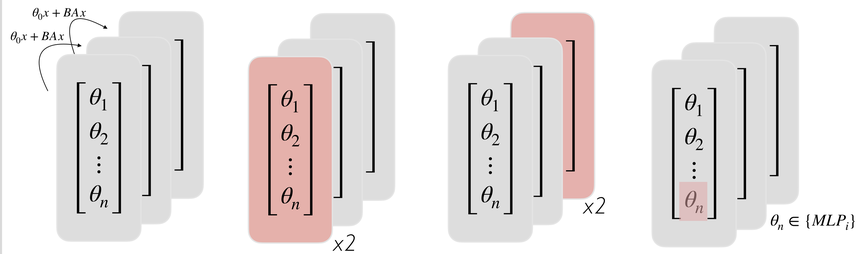

*Visual Transformer (ViT)*




Why ViT? Everyone is paying attention! During initial experiments with ResNet18, we found that the ViT architecture provided significant performance leverage and more stability across epochs without overfitting.

## Methods and experiments

We fine-tune our datasets using five different fine-tuning strategies and a 80-20 split for training and validation.




*LoRA finetuning:*

$$Wx + BA^T x$$

Where $A$ and $B$ are matrices with lower rank than $W$

*For evaluation:*

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

## Results

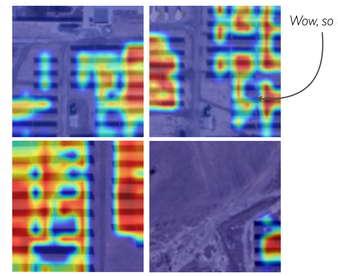
To compare the different FT strategies we compare the **F1** scores for each of our AOI and product combinations. For our model/sample of interest, LoRA has the better performance.

| | All parameters | First layer | Last layer | LoRA | Linear |
|---|---|---|---|---|---|
| Cali [HR] | 0.909 | 0.846 | 0.896 | 0.892 | **0.915*** |
| China [HR] | 0.964 | 0.959 | 0.961 | **0.969** | 0.957* |
| China [S2] | 0.833 | 0.846 | 0.792 | **0.853** | 0.751* |

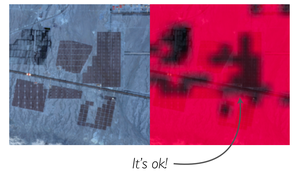
Observations:
1. LoRA overall performs better with the least overfitting and lowest computations.
2. For medium resolution Sentinel, fine-tuning the first two blocks gives better performance than others. Our intuition is that the data domain shifts between two datasets is significant [shown in discussion section].
3. Out of sample dataset evaluation shows poor generalizability across domains.

## Activation Maps

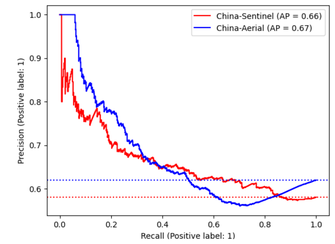
*High-resolution activation layers for China*

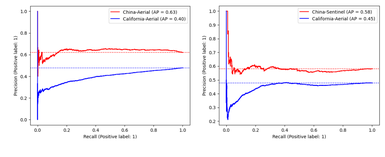



*Low-resolution activation layers for China*



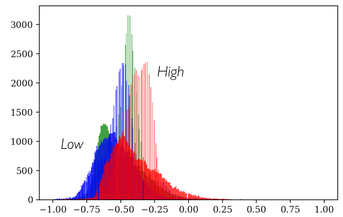

Out-of-sample dataset evaluation shows low generalization abilities of our fine-tuned models




## Discussion and future work

Most of pre-trained vision models (i.e. ResNet, CIFAR) used for down streaming tasks are trained on RGB imagery with certain color distributions. For remote sensing tasks, the changes in resolution and color generate a distributional-shift that might explain an optimization upper-bound during the fine-tuning:




- Exploring more fine-tuning alternatives and ViT regularizations ($\theta$ decay and layer drop).
- Implement architectures that can parses different wavelengths, as these provide *new feature spaces.*