PyViT-FUSE: A Foundation Model for Multi-Sensor Earth Observation Data

Manuel Weber, Carly Beneke - EarthDaily Analytics

Introduction

We introduce PyViT-FUSE, a model designed to process multi-source, multi-modal images at their native resolutions. It enables flexible band selection through a channel fusion approach leveraging attention mechanisms. Unlike many EO foundation models that employ masking for self-supervised learning [1-3], we adapt SwAV (Swapping Assignments between Views) [4], eliminating the need for a decoder and pixel-space reconstruction which enables learning of embeddings independent of input band combinations, allowing the same model to be applied across diverse downstream tasks. Our unique model architecture compresses rich input information through three distinct modules: Input module, Fusion module and Pyramidal Vision Transformer.

Methodology

Architecture

Input module

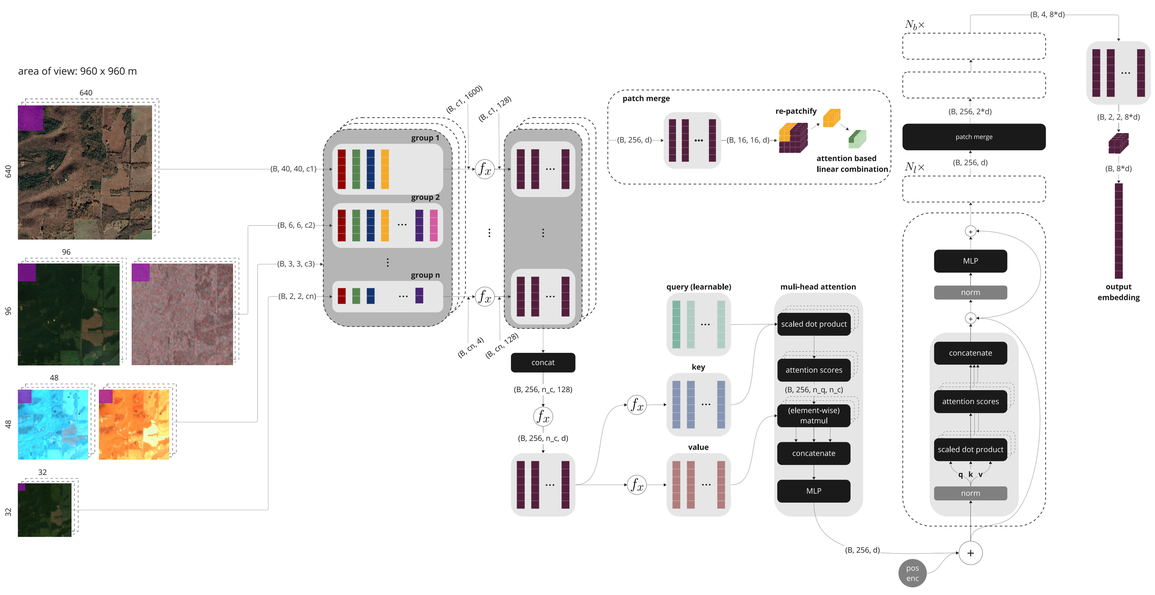
- Patchify input modalities
- Projection onto common feature space

Fusion module

- Fuse bands via attention into single token per patch
- Learnable query vector, key and value vectors from projection

Pyramid ViT

- Process token sequence
- Spatial merging with fusion module



Each ViT pyramid block follows the transformer implementation of [5]. Spatial merging reduces the sequence length by factor 4 at each stage. 2D position encoding is added at each block input. Each pyramid block consists of N_l multi-head attention + MLP layer groups while the entire pyramid architecture contains N_b blocks. These two parameters, as well as the internal feature dimension, enable scaling the number of model parameters.

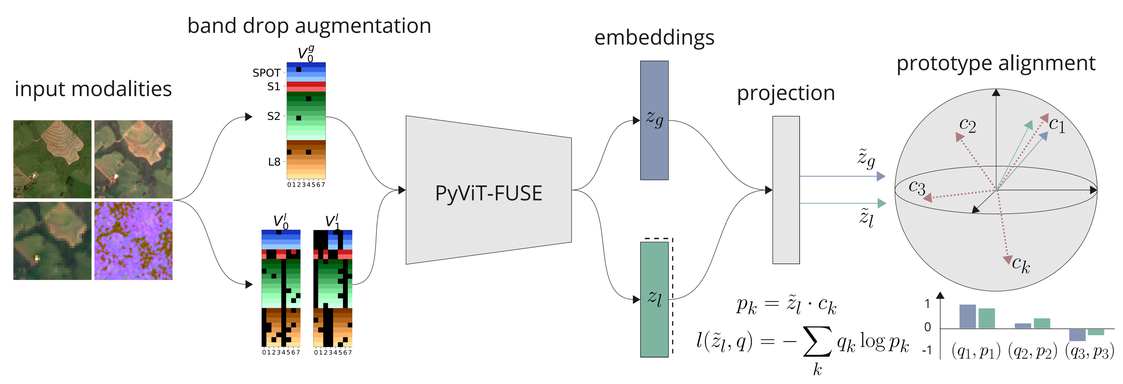
The model is transferable to a different set of input modalities by re-training the input and fusion module while keeping the pyramid ViT fixed, allowing versatile use across many data sources without re-training from scratch.

Training

Global view: All modalities with light augmentations applied (random noise add/mult, single channel drop)

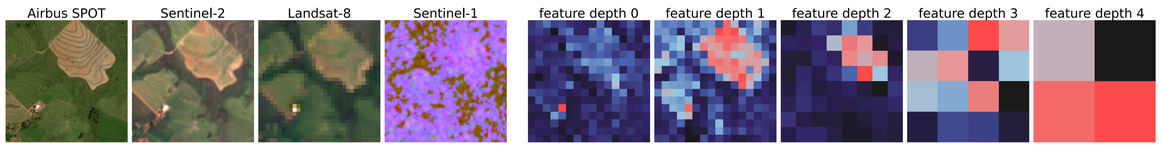
Local views: Hard augmentation applied by dropping modalities at random + weak augmentations

Prototype alignments of projected global view embeddings (q) provide the target while the alignments of local views (p) correspond to the predictions.
Loss function is defined as the cross-entropy of p and q.



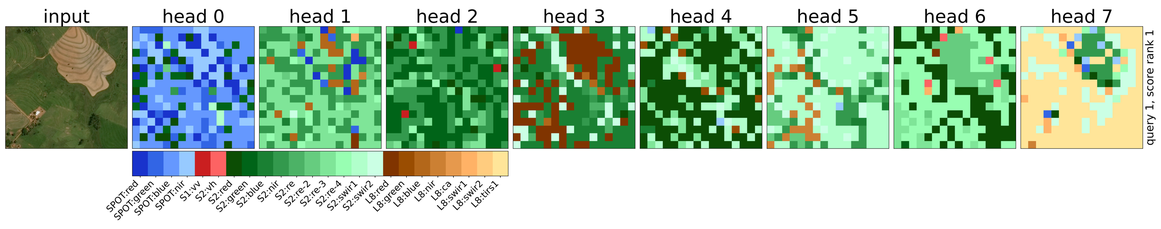
Results

Learned feature maps



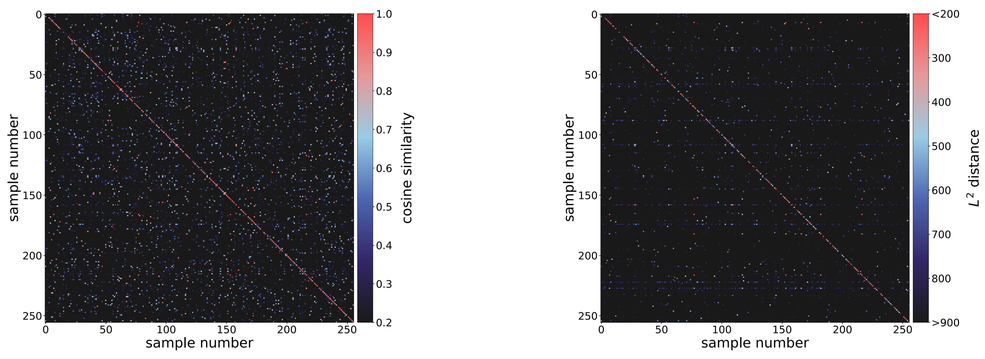
Sample input and corresponding feature maps averaged along the channel dimension at multiple depths of the ViT pyramid

Band importance



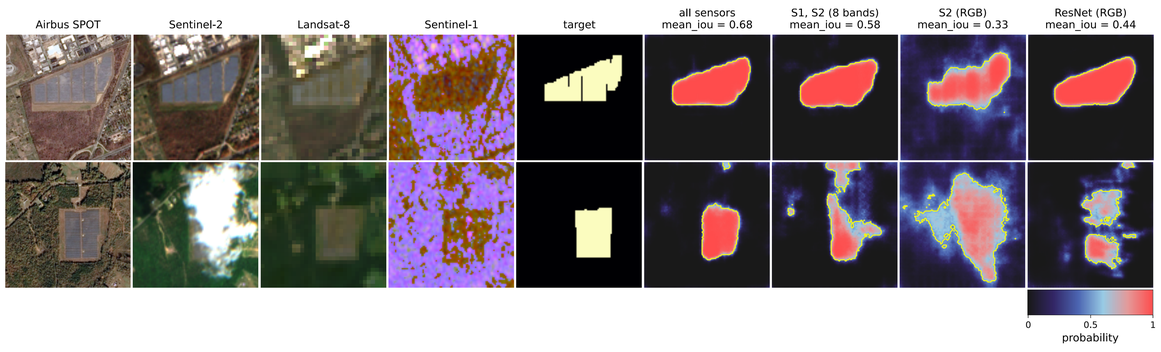
Visualization of band importance through highlighting the band with highest attention score in the fusion module

Feature similarity

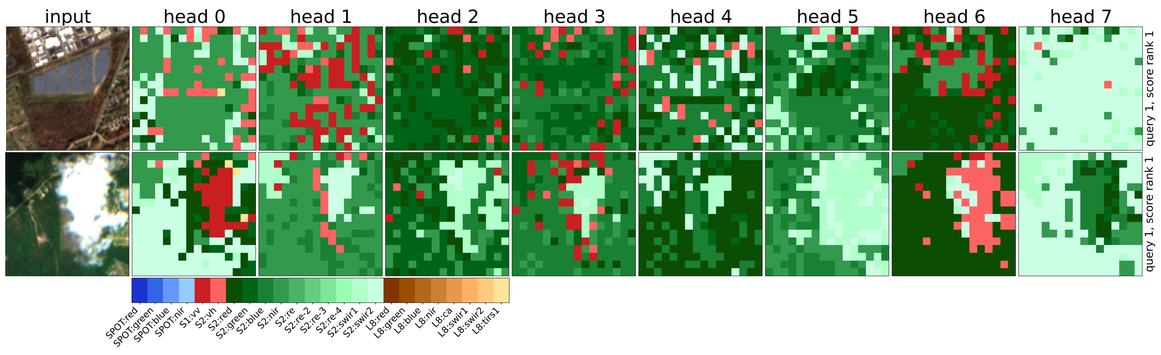


Measurement of cosine similarity and L2 distance between the embeddings of a global view of a given sample against a local view of all other samples in the batch

Experiments



Example downstream task: Fine-tune segmentation decoder from feature maps on Solar PV dataset, demonstrating the benefit of multi-modal input for more accurate predictions



Visualization of attention scores for an input consisting of only Sentinel-1 and Sentinel-2 bands indicates location of cloud obstruction, validating the function of the fusion module to optimally combine bands at the given location

Conclusions

We demonstrate a novel approach for the fusion of mixed-resolution, multi-modal satellite imagery, its interpretability due to attention mechanism and leverage an innovative concept for self-supervised training of a large model on vast amounts of data. As the next step, we will run a number of benchmark tests to quantify the model capability.

[1] Reed et al., Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning, 2023
[2] Astruc et al., Omnisat: Self-supervised modality fusion for earth observation, 2024
[3] Jakubik et al., Foundation models for generalist geospatial artificial intelligence, 2023
[4] Caron et al., Unsupervised learning of visual features by contrasting cluster assignments, 2021
[5] Vaswani et al., Attention is all you need, 2023

Corresponding author: manuel.wbr@gmail.com