SatMAE

Introduction

Satellite images often contain more than just RGB—they include additional spectral bands like infrared (IR) and ultraviolet (UV), and can also have a temporal dimension (images of the same location at different times). SatMAE is a method for self-supervised pretraining on such data to improve downstream tasks like classification and segmentation.

Method

SatMAE uses a Masked AutoEncoder (MAE) to pretrain a Vision Transformer (ViT) on unlabeled satellite data. Each input image is split into 16×16 patches, which are then linearly embedded and sent to the transformer encoder. A large fraction of these patches is masked randomly.

To handle multi-spectral inputs, SatMAE groups spectral bands into sets (RGB with one dimension, and SWIR and IR in another dimension since they are "close" and might contain similar information) and stacks them along the channel dimension. Temporal data is incorporated by treating multiple time frames as additional channels or as separate grouped inputs depending on the configuration.

Only the visible patches are encoded, and the decoder reconstructs the full image using both position embeddings and the encoded tokens. The model learns useful feature representations by minimizing the reconstruction loss between the original and predicted pixel values.

This pretraining strategy allows SatMAE to extract transferable features from complex, unlabeled satellite imagery.