Rethinking Transformers Pre-training for Multi-Spectral Satellite Imagery

DISCLAIMER: Summarized by AI

Problem they are trying to solve / Purpose of method

What are the previous problems that need to be solved?

- Satellite imagery varies in *scale* (due to differences in ground sampling distance - GSD) and in *modality* (e.g., RGB, multi-spectral).
- Existing pre-training methods for satellite imagery (e.g., SatMAE, Scale-MAE) are limited:
    - **SatMAE** does not capture multi-scale information effectively.
    - **ScaleMAE** uses GSD-based positional encoding, which is not compatible with multi-spectral imagery (because different bands have different resolutions).
- These limitations hinder the performance of models on downstream remote sensing tasks.

Why is the method introduced/needed?

- There's a need for a *simple, scalable*, and *modality-agnostic* pre-training method that can leverage the *multi-scale nature* of satellite data across both optical and multi-spectral domains.
- The goal is to improve feature representation learning by integrating multi-scale reconstruction into the masked autoencoder (MAE) framework, leading to better generalization and performance.

How does it differ from other methods?

Compare what is different from other methods.

- Unlike **SatMAE**, which reconstructs only at a single scale, **SatMAE++** performs *multi-scale reconstruction*, enabling the model to learn representations that are robust to scale variation.
- Unlike **ScaleMAE**, SatMAE++:
    - Does *not* rely on GSD-based positional encodings (which are incompatible with multi-spectral data).
    - Uses standard sinusoidal positional encodings for *simplicity and generality.*
    - Introduces a *lightweight convolution-based upsampling block* instead of a complex Laplacian decoder for multi-scale image reconstruction.

What makes this method unique?

- Simplicity: Uses standard positional encodings and avoids modality-specific designs.
- Generalizability: Works with both optical and multi-spectral imagery.

- Effectiveness: Achieves state-of-the-art results on several benchmarks with faster convergence during finetuning.

How the method works

Simple overview:

1. Builds on the MAE framework but extends it to perform *multi-scale image reconstruction.*
2. Uses a transformer encoder-decoder to reconstruct an image at a base resolution.
3. Adds convolutional upsampling blocks to reconstruct higher-resolution versions of the input.
4. Combines losses from all scales to guide the training process.

More details:

- **Input:** Multi-spectral or RGB image at high resolution.
- **Downsampling:** Input is downsampled to multiple resolutions (e.g., (H, W), (2H, 2W), (4H, 4W)).
- **Encoding:** Only the lowest resolution (H, W) is input to the MAE.
- **Decoding:** The decoder reconstructs the image at (H, W) using MSE loss.
- **Upsampling blocks:** Use transpose convolution followed by residual conv layers to upsample features to (2H, 2W) and (4H, 4W), reconstructing the image at these scales using L1 loss.
- **Loss Function:** Weighted sum of losses from all scales:

$Loss = \alpha_1 * L_1 + \alpha_2 * L_2 + (\alpha_3 * L_3)$, if using 3-scale setup