Enchancing the processing of Hyper Spectral satellite image data with specialized vision transformers

Andras Bodrogai

2575706B

Halfway Progress Report

This report is intended to give a brief summary for the project for optimizing vision transformers for Hyper Spectral Image (HSI) Processing.

Proposal

Motivation

With the continuous improvements to mobile edge devices and with the increasing amount of data being processed taken by satellites, the need for an optimized and better performing deep neural network is increasing. Current state of the art methods based on Convolotional Neural Networks (CNN) are starting to find their way into multiple feature extraction pipelines for hyper spectral image processing, such as land coverage extraction, shore detection, etc. However, with the original introduction of the Vision Transformer (ViT) architecture, a new door opened, in terms of approaching the problem of image processing, prioritizing global feature extraction. However many of the architectures proposed based on the original ViT are not optimized for the task of processing HSI data, such that to optimize the performance of such a networks, both architecturally, and during pretrainig.

Aims

The goal of the project is to optimize the vision transformer architecture proposed originally by the paper An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale for HSI processing. The paper proposed a new type of network based on the transformer architecture, and built around global attention giving it a significant edge over traditional convolutional models. This project aims to propose new pipeline for training HSI processing networks. This is achieved by combining, both spatial and spectral analysis of the data in a one network, providing better awareness of the features present in the data. Combined with enchanced pretraining techniques

Progress

So far the progress has been made in the following areas:

- implemented training pipeline
    - data loading
    - trainers
    - visualizations
    - loggers
    - callbacks
    - custom layers
- Collected dataset
- Implemented the Masked Autoencoder (MAE) transformer architecture proposed in the paper Masked Autoencoders Are Scalable Vision Learners
- Performed self-supervised pre-training on the MAE architecture

Problems and risks

Problems

Pretraining the networks proved to be a challange, as the training of the network requires both a lot of computational power, time and data.

The later the I used the WHU-OHS Dataset providing 10000 images of size 512x512 image patches each with 32 bands split between train, validation and test sets. The images were split further to 64x64 patches to provide more data for training, resulting in 300000 images for training.

For training the network originally I utalized a personal machine, However as transformer networks are very data hungry, it is very time consuming to train them.

Risks

The main risk of the project as it stands now, is the time remaining which requires the planned steps to be prioritized accordingly. As mentioned above the computational requirements of the networks are very high such that the number of tries for a full pre training is limited.

Plan

As the foundation is in place and we already collected a baseline performance for the model, the goal now is it improve uppon the architecture. Which will be done in the following order:

- Changing the current pretraining pipeline to a more advanced one, such as the one proposed in the paper Emerging Properties in Self-Supervised Vision Transformers.
- Changing the current MAE architecture to a more specialized model for HSI processing such as the Masked SST, providing a better understanding of the spectral features present in the data.
- Changing the current Naive tokanizer from patches to a more advanced spatial tokanizer, likely mitigating the problem of the square patches being present in the predicted image

The time plan for the project is as follows:

- Implementing the new pretraining pipeline based on either DINO, DINOv2, iBOT including adding Registers, for stabilizing the training as it was proposed in the paper Vision transformers need registers
  Planned time: present - Jan 8
  Model Evalution: Jan 8 - Jan 15

- Implementing a hyper spectral optimized architecture, such as the Masked SST
  Planned time: Jan 15 - Jan 22
  Model evaluation: Jan 22 - Jan 29

    - Implementing a spatial tokanizer
    - Implementing a spectral tokanizer

- Updating the previously implemented spatial tokinezer to a more advanced one.
  Planned time: Jan 29 - Feb 5
  Model evaluation: Feb 5 - Feb 12

- Comparing performances of the models on downstream tasks
  Planned time: Feb 12 - Feb 19

Note: the above information is subject to change.