Does spatial resolution (i.e. meters per pixel) matter for self-supervised learning (SSL) methods for multispectral imagery?

NYU Center for Data Science

Fiona Chow, Jack Savage, Oliver Xu, Pallabi Chandra

LINDSAY LAB

Problem Description

Background → Applications → Problem → Proposal

Background

Technological advancements have led to a significant increase in the velocity and volume of aerial multispectral imagery data.

Multispectral images are images captured in multiple bands of the electromagnetic spectrum, providing data beyond what is visible to the human eye.

Applications

1. **Climate change monitoring through land cover analysis**
2. **Optimizing resource management in agriculture**
3. **Planning rescue operations in disaster management**

Problem

Expert annotations of such images are expensive and time consuming.

While self-supervised learning methods are a great way to annotate such data without human annotation by learning representations without labels using deep neural nets, notably, **these models often overlook the control of resolution (i.e. meters per pixel).**

Proposal

Hence, in this capstone project, we will aim **to test whether models pre-trained on one resolution transfer well or poorly to tasks that use images of different resolutions.**

Summary of Related Work

Among the first to propose a deep learning based mechanism for hyperspectral data classification in remote sensing were Chen et al. in 2014[1], who suggested the usage of stacked autoencoders for extracting potentially useful high level features from images. While many other authors have investigated similar combinations of a self-supervised learning method, frequently an autoencoder, paired with a classifier, such as Othman et al.[2] in 2016, **a September 2022 literature review by Wang et al.[3] did not report any relevant works explicitly investigating the effect of varying image resolution on model performance,** which is a very common scenario given the many different resolutions that appear in the various remote sensing datasets in use for model training **and hence an important one to investigate.**

Dataset

The National Agriculture Imagery Program (NAIP)[4]

- **405,000 image patches**
- **64/16/20 train/validation/test split** ratio of non-overlapping tiles
- Image patches are sourced from diverse California landscapes
- **Six landcover/landuse categories**
- Categories adhere to the National Land Cover Data algorithm
- Each image patch measures **28x28 pixels**
- **Each patch** corresponds to a **single landcover/landuse class**
- Data is captured in **four spectral bands:**
  - Red
  - Green
  - Blue
  - Near Infrared
- Maintains a **ground sample distance of 1 meter**
- Horizontal accuracy is within six meters of identifiable ground control points

Approach and Evaluation Protocol/Metrics

Data Preprocessing

1) Resize images to required image resolution
2) Data augmentation

E.g. Resizing images from 28 by 28 by 4 (W, H, C) to 14 by 14 by 4 and applying 'RandomVerticalFlip'.

Pre-training:

Learn useful representations from unlabeled data. This process involves training a unified encoder structure on two different self-supervised learning models: SimSiam and an autoencoder. We assess binary cross-entropy loss for the autoencoder and negative cosine similarity loss for SimSiam.

E.g. Training models on data with a spatial resolution of 14 by 14 by 4.

Transfer Learning:

Train a **classifier head** using labeled images of varying resolutions while **initializing the encoder weights from** the **previous** step. We assess cross-entropy loss for the classifier.

E.g. Training classifier on data with a spatial resolution of 28 by 28 by 4.

Evaluation:

Evaluate classification test accuracy with labeled data.

To investigate the impact of resolution, we **compare test classification accuracy across three scenarios**: **constant** encoder and classifier **resolutions**, **increasing resolutions,** and **decreasing resolutions**. Further details are provided in the results section.

E.g. Evaluating test accuracy between class labels from test data with spatial resolution of 28 by 28 by 4 and predicted class labels of similar resolution.

Traditionally in computer vision, resolution is a measure of the pixel count in our subject image. Within the context of aerial imagery, **spatial resolution refers to the number of ground level meters represented by each image pixel.** By reducing pixel resolution in our images, we are simultaneously decreasing spatial resolution due to the information loss from compression.

In varying spatial resolution, we wanted simulate a likely real world use case where a pre-trained self-supervised network is trained at a different spatial resolution than our classification dataset. Following this, **we established three possible scenarios: an increase, a decrease, or an equivalency in resolutions between pre-trained encoder and classifier.** We chose to then evaluate these cases across three different resolutions: 28 x 28 (1 meter per pixel), 14 x 14 (2 meters per pixel), and 2 x 2 (14 meters per pixel).

To determine the effects of differing resolution, we utilized the self-supervised encoder to classifier pipeline featured in our literature review.[3] We implemented two self-supervised architectures common to the field: SimSiam and a convolutional autoencoder. For classification, we used a simple fully connected feedforward neural network.

We used a unified encoder structure for both models. The structure is made of convolutional layers of increasing depth that are distilled into a 1D vector using 2D pooling on feature map width and height. With global average pooling, we ensured a constant encoder output shape regardless of input shape. Hyperparameters were also fixed across all models to ensure valid comparison of results.

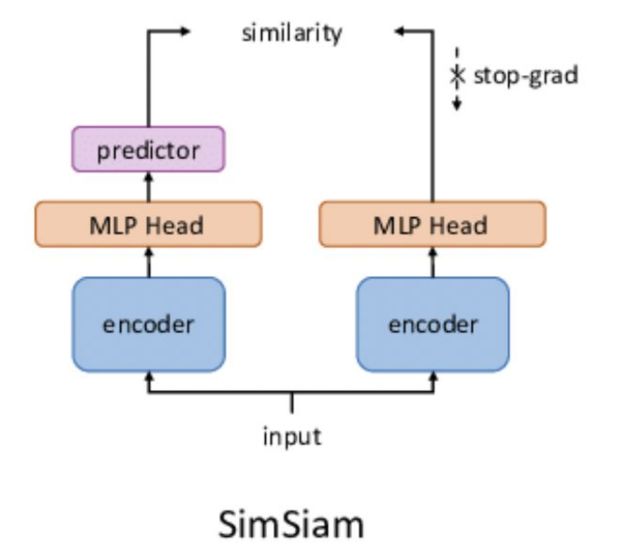

Figure 1. Simsiam structure outline

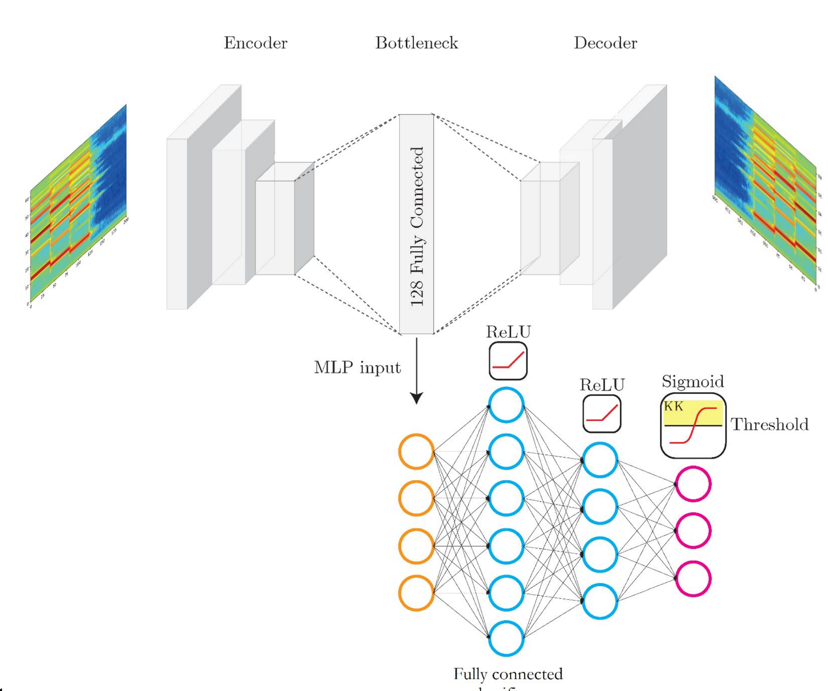

Figure 2. Convolutional autoencoder and classifier
These models are trained in parallel

Results

Model	Latent Dimension	Encoder Dimensions	Classifier Dimensions	Loss Function	Classifier Test Accuracy [%]
I. Constant resolution					
Autoencoder	128	[28, 28]	[28, 28]	Binary Cross-entropy	98.14
		[14, 14]	[14, 14]		97.8
		[2, 2]	[2, 2]		96.82
SimSiam		[2, 2]	[2, 2]	Negative Cosine Similarity	52.01
		[14, 14]	[14, 14]		51.73
		[28, 28]	[28, 28]		48.59
II. Increasing resolution					
Autoencoder	128	[14, 14]	[28, 28]	Binary Cross-entropy	98.19
		[2, 2]	[28, 28]		96.81
		[2, 2]	[14, 14]		95.56
SimSiam		[2, 2]	[14, 14]	Negative Cosine Similarity	67.25
		[2, 2]	[28, 28]		64.83
		[14, 14]	[28, 28]		61.73
III. Decreasing resolution					
Autoencoder	128	[28, 28]	[14, 14]	Binary Cross-entropy	95.48
		[14, 14]	[2, 2]		91.4
		[28, 28]	[2, 2]		86.3
SimSiam		[14, 14]	[2, 2]	Negative Cosine Similarity	62.73
		[28, 28]	[2, 2]		58.44
		[28, 28]	[14, 14]		45.98

Table 1. Classifier performance on (**I**) pipeline where self-supervised model was trained at the same resolution as classifier; (**II**) pipeline where self-supervised model was trained at a lower resolution than classifier; (**III**) pipeline where self-supervised model was trained at a higher resolution than classifier.

From Table 1, the **autoencoder shows that classifier performance is best with constant resolution.** We see worse results when varying resolution, scaling with the gap between resolutions, and better results when training with more complex (higher) resolutions in general. **Increasing resolution produced better results than decreasing resolution.** From Figure 3, autoencoder reconstructions look very similar to original image patches.

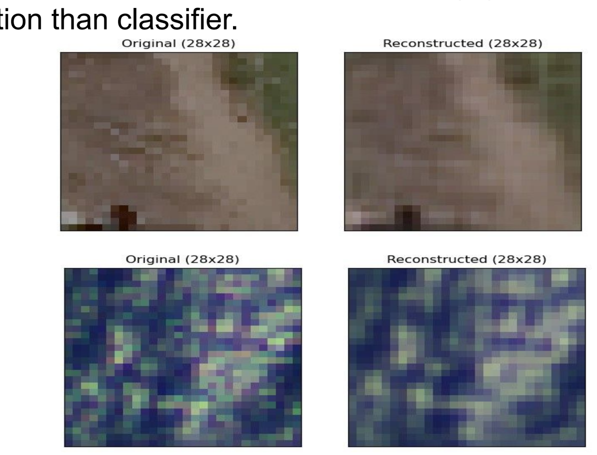

Figure 3. Autoencoder reconstructions

Overall, SimSiam classifier performance is lower than that of the autoencoder. This could be the result of the nature of training of SimSiam being unstable.[3] From Table 1, **SimSiam demonstrates stronger performance in scenarios involving increasing or decreasing resolutions.** Similar to autoencoder, it achieves higher accuracy in enhancing resolution rather than reducing it. **Surprisingly, SimSiam delivers better results with 2x2 dimensions compared to 28x28 dimensions in the mix,** underscoring the variability in self-supervised models' responses to different situations.**This implies that higher resolution does not always equate to superior accuracy, highlighting the complexity of self-supervised models' behavior. Finally, different self-supervised methods may prioritize varying resolutions, emphasizing that the choice of spatial resolution is method-dependent in multispectral imagery.**

Future Work

- Explore mixed resolutions pre-trained models
- Benchmark against TorchGeo (widely used deep learning library with pretrained models for multispectral imagery
- Explore other SSL models like MoCo
- Evaluate on a more complex dataset like Satlas that contains smaller objects

Acknowledgements

We would like to thank Grace Lindsay and Brian McFee for their guidance and support throughout the project.

References

[1] Y. Chen, Z. Lin, X. Zhao, G. Wang and Y. Gu, "Deep Learning-Based Classification of Hyperspectral Data," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 7, no. 6, pp.
2094-2107, June 2014, doi: 10.1109/JSTARS.2014.2329330.
[2] Esam Othman, Yakoub Bazi, Naif Alajlan, Haikel Alhichri & Farid Melgani (2016) Using convolutional features and a sparse autoencoder for land-use scene classification, International Journal of Remote Sensing, 37:10, 2149-2167, DOI: 10.1080/01431161.2016.1171928
[3] Yi Wang, Conrad M Albrecht, Nassim Ait Ali Braham, Lichao Mou, and Xiao Xiang Zhu. 2022. Self-supervised learning in remote sensing: A review. arXiv preprint arXiv:2206.13188 (2022)
[4] Saikat Basu, Sangram Ganguly, Supratik Mukhopadhyay, Robert Dibiano, Manohar Karki and Ramakrishna Nemani, DeepSat - A Learning framework for Satellite Imagery, ACM SIGSPATIAL 2015