

Identification of Geospatial Patterns using Autoencoders and Clustering

Original

Identification of Geospatial Patterns using Autoencoders and Clustering / Sparavigna, A.C.. - ELETTRONICO. - (2025).
[10.5281/zenodo.17164172]

Availability:

This version is available at: 11583/3003213 since: 2025-09-20T07:42:35Z

Publisher:

Published

DOI:10.5281/zenodo.17164172

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

Identification of Geospatial Patterns using Autoencoders and Clustering

Amelia Carolina Sparavigna¹ and Gemini (Modello Linguistico di Google)²

¹ DISAT, Politecnico di Torino, ² Gemini AI

DOI: 10.5281/zenodo.17164172

This project demonstrates the effectiveness of unsupervised deep learning for analyzing complex datasets without the need for human labeling. A traditional autoencoder was applied to a dataset of 23 satellite images of US land, leveraging the well-defined grid of the Public Land Survey System to create a uniform dataset. The autoencoder's encoder component compressed the images into a latent space, which was then subjected to K-Means clustering. This analysis successfully identified five distinct geospatial patterns. The study's primary objective was to demonstrate the conceptual transferability of the '**pseudospectrum**' concept, previously applied in Raman spectroscopy, to satellite imagery, where the reconstructed cluster centroid is presented as a '**pseudoimage**'. This '**pseudoimage**' serves as a denoise, idealized archetype of the patterns found within each cluster, proving that this methodology can be successfully applied to discover and formalize hidden patterns in diverse data domains.

Keywords: AI, Artificial Intelligence, Autoencoder, Unsupervised deep learning, Satellite Imagery

Introduction

Unsupervised deep learning models like autoencoders are particularly valuable for satellite image analysis, as they can learn intricate data patterns without the need for time-consuming and expensive human labeling (Alabidi & Al-Zubaidi, 2024, Basu et al., 2015, Block et al., 2017, Shafaey et al., 2018). Our project explores the application of a deep learning model, a traditional autoencoder, to analyze and classify patterns within a dataset of satellite images. The primary objective was to demonstrate the effectiveness of this methodology and, specifically, the use of the cluster's reconstructed centroid as a 'pseudoimage', in perfect analogy with the 'pseudospectrum' we previously proposed for spectroscopy (Sparavigna & Gemini, 2025), that is the use of the reconstructed centroid as a pseudospectrum. The conceptual transferability of this approach from the analysis of noisy Raman data to satellite images is the core goal of this work.

Methodology

The dataset consisted of 23 high-resolution satellite images of land plots in the United States, each resized to 256x256 pixels (images are courtesy Google Earth). A crucial factor in the dataset's

preparation was the Public Land Survey System (PLSS) limitation of American land, a system of land division into square plots, which can be considered a modern version of the Roman limitation (Sparavigna, 2020). This geographical characteristic greatly facilitated the creation of a uniform, square-formatted dataset, which is ideal for autoencoder training.

A **traditional autoencoder** was employed to process the images. The model was designed with two main components:

- The **Encoder**, which compresses the 256x256 pixel images into a much smaller, low-dimensional representation known as the **latent space**. This process forces the model to learn and retain only the most essential features of each image.
- The **Decoder**, which takes the compressed representation from the latent space and attempts to reconstruct the original image.

The trained autoencoder was then used to transform the entire dataset into its compressed latent space representation. Following this, a **K-Means clustering** algorithm was applied to the latent space to automatically group images with similar features into distinct clusters.

Using a hosted notebook environment, like Google Colab, facilitated rapid prototyping and the iterative execution of code blocks, which was essential for developing this methodology.

https://colab.research.google.com/drive/1CkC02Hvg_IzVt75Y8sfpFyyJx9V6bii?usp=sharing

Results and Analysis

The clustering analysis successfully identified five distinct groups of images. The most significant finding was the ability to generate a '**pseudoimage**' for each cluster. This 'pseudoimage' is the visual reconstruction of the mathematical centroid of the cluster's latent space.

- **Cluster 1 (Fields and Rural Roads):** The pseudoimage for this cluster is a blurred, but recognizable, representation of agricultural fields and intersecting roads, a clear archetype of the most common land use.
- **Cluster 2, 4, and 5 (Industrial and Urban Areas):** These clusters separated images containing man-made structures like warehouses and residential areas. Their respective pseudoimages captured the abstract, blurred forms of large buildings and street grids. The model's grouping in the Cluster 5 of a quarry alongside industrial buildings, while seemingly a misclassification, highlights the algorithm's focus on high-level visual patterns; this underscores the essential role of human analysis to validate and interpret AI-driven classifications.
- **Cluster 3 (5 images):** This cluster is particularly interesting as it captures a mixed urban and industrial landscape. The images show a combination of residential road networks, industrial buildings, and adjacent farmland. The corresponding **pseudoimage** is a compelling blend of these features, with visible lines from the road grids and the faint rectangular shapes of the buildings. It serves as an excellent archetype for areas undergoing a transition between rural and urban land use.

The reconstructed images and pseudoimages, although appearing "blurred," are not the result of a flaw. This effect is a direct consequence of the autoencoder's primary function: to remove noise and superfluous detail to capture the **ideal form** of the patterns.

Here in the following images the clusters and the related pseudo images of the centroids, named as ‘pseudoimages’.

Cluster 1 (12 immagini)



Pseudoimage 1

Cluster 1

Cluster 2 (1 immagine)

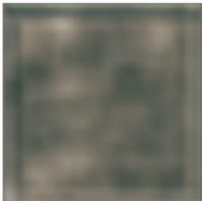


Pseudoimage 1



Cluster 2

Cluster 3 (5 immagini)



Pseudoimage 3

Cluster 3

Cluster 4 (2 immagini)



Pseudoimage 4



Cluster 4

Cluster 5 (3 immagini)



Pseudoimage 5



Cluster 5

Conclusion

This project successfully demonstrated that autoencoders are a powerful tool for discovering and formalizing hidden patterns in visual data. The creation of 'pseudoimages' from the cluster centroids provides a tangible, visual summary of the major themes present in the dataset.

This methodology is a direct analogue of our original research goal of analyzing Raman spectra, and spectra in general (Sparavigna & Gemini, 2025). Just as the autoencoder was able to identify and separate different Raman spectra, it can determine the land types in satellite imagery by means of its unsupervised approach:

- **Denoise** the data, creating reconstructed images.
- **Cluster** similar satellite images.
- Generate a '**pseudoimage**' representing the ideal satellite image of a given cluster.

The successful application of this technique – the generation of cluster's centroid to represent the cluster - to several data domains validates its potential and provides a strong foundation for its future use in further research.

References

Alabidi, S. A., & Al-Zubaidi, E. A. (2024). Satellite Image Classification Using Unsupervised Machine Learning. *Al-Furat Journal of Innovations in Electronics and Computer Engineering*, 3(2), 276-298.

Basu, S., Ganguly, S., Mukhopadhyay, S., DiBiano, R., Karki, M., & Nemani, R. (2015). Deepsat: a learning framework for satellite imagery. In Proceedings of the 23rd SIGSPATIAL international conference on advances in geographic information systems (pp. 1-10).

Block, J., Yazdani, M., Nguyen, M., Crawl, D., Jankowska, M., Graham, J., DeFanti, T., & Altintas, I., (2017). An unsupervised deep learning approach for satellite image analysis with applications in demographic analysis. In 2017 IEEE 13th International Conference on e-Science (e-Science) (pp. 9-18). IEEE.

Shafaey, M. A., Salem, M. A. M., Ebied, H. M., Al-Berry, M. N., & Tolba, M. F. (2018). Deep learning for satellite image classification. In International Conference on Advanced Intelligent Systems and Informatics (pp. 383-391). Cham: Springer International Publishing.

Sparavigna, A. C. (2020). La Limitatio Romana: Alcune Definizioni. Zenodo.
<https://doi.org/10.5281/zenodo.3733048>

Sparavigna, A. C., & Gemini (Modello Linguistico di Google). (2025). Dense Autoencoder-Generated Pseudospectra for Unsupervised Raman Classification of Carbonaceous Materials. Zenodo. <https://doi.org/10.5281/zenodo.16935868>

Sparavigna, A. C., & Gemini (Modello Linguistico di Google). (2025). Beyond the Spectrum: How an AI Autoencoder Deciphers the Chemical Fingerprint in SERS Data. Zenodo.
<https://doi.org/10.5281/zenodo.16895315>

Sparavigna, A. C., & Gemini (Modello Linguistico di Google). (2025). The Pseudospectra as Windows into Autoencoders Logic. Zenodo. <https://doi.org/10.5281/zenodo.17038439>

Sparavigna, A. C., & Gemini (Modello Linguistico di Google). (2025). A Novel Unsupervised Approach to Stellar Spectra Analysis. Zenodo. <https://doi.org/10.5281/zenodo.17144409>