Assignment 12: topoDL VAE

40 Points

Deliverable: Notebook (.ipynb file) with all required code to complete the stated tasks. Answer all questions in Markdown cells.

Overview: Build an anomaly detection algorithm using a variational autoencoder (VAE) to detect sections of topographic maps that include surface mine disturbance as anomalies.

Background Questions

B1: Explain the difference between an autoencoder and a variational autoencoder. (5 Points)

B2: Explain the concept of a latent space. (5 Points)

B3: Explain the reparameterization trick required to implement VAEs and why it is necessary. (5 Points)

B4: Explain the loss function for a VAE. What is the purpose of the reconstruction loss and the KL-divergence loss? (5 Points)

Tasks

Edit the anomaly detection PyTorch example to detect chips in the topoDLMini dataset that have mine disturbance as anomalies. This will involve training a model using only the background chips. The code should include the following components. (20 Points)

1. Read in the *trainDF.csv* and *testDF.csv* files from the topoDLMini folder as Pandas DataFrames. You will not need the validation set data.
2. Subset out only “background” training samples.
3. Split the test set samples into separate “background” and “positive” sets.
4. Use the DataFrames to create a DataSet subclass to read in the topo images and their associated label: “positive” or “background”. You will not need the masks. Make sure the topo map chips are rescaled from 0 to 1 and converted to a 32-bit float data type. The “background” and “positive” cases should be assigned unique numeric codes with a data type of long integer.
5. Instantiate DataSets and DataLoaders for the “background”-only training set, “background”-only test set, and “positive”-only test set.
6. Define a VAE architecture. You can use the architecture used in the example module.
7. Define a VAE loss. You can implement the loss from the example module that sums MSE and KL-divergence loss.
8. Instantiate the model.

9. Train the model for at least 30 epochs.
10. Use the trained model to predict to the “background”-only test set and “positive”-only test set separately. Create a grouped kernel density plot showing the distribution of the reconstruction loss for the “background”-only test set and “positive”-only test set (like the one in the example).
11. Using the graph, discuss the results. Could the model be used to detect topo chips that include surface mining? If so, what reconstruction threshold would you suggest using?