Assignment 7: DataSet and DataLoader for CNN-Based Scene Labeling Task

40 Points

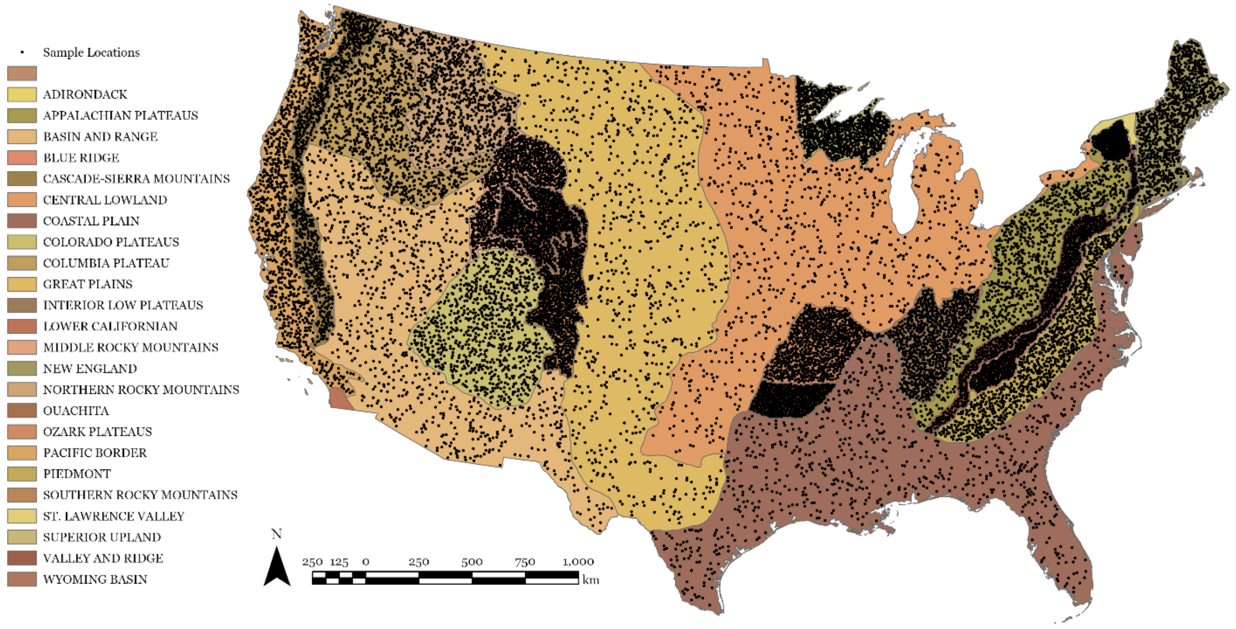
Deliverable: Notebook (.ipynb file) with all required code to complete the stated tasks.

Overview: The goal of this assignment is to build a DataSet subclass, check that it is working properly, and define a DataLoader for a CNN-based scene classification problem. You will use the defined DataSet and DataLoader in the following two assignments.

Data:

This assignment makes use of the *physioDL* dataset, which is available on FigShare:

Maxwell, A.E., 2024. physioDL: A dataset for geomorphic deep learning representing a scene classification task (predict physiographic region in which a hillshade occurs). https://doi.org/10.6084/m9.figshare.26363824.



The task presented in this dataset is to predict the physiographic province of an area based on a hillshade image. Terrain data were derived from the 30 m (1 arc-second) 3DEP product across the entirety of CONUS. Each chip has a spatial resolution of 30 m and 256 rows and columns of pixels. As a result, each chip measures 7,680 meters-by-7,680 meters. Two datasets are provided. Chips in the *hs* folder represent a multidirectional hillshade while chips in the *ths* folder represent a tinted multidirectional hillshade. You will use the data in the *hs* folder. Data are represented in 8-bit (0 to 255 scale, integer values). Data are projected to the Web Mercator projection relative to the WGS84 datum. Data were split into training, test, and validation partitions using stratified random sampling by physiographic province. 70% of the samples per region were selected for training, 15% for testing, and 15% for validation. There are a total of 16,325 chips. The following 22 physiographic regions are represented: "ADIRONDACK" , "APPALACHIAN PLATEAUS", "BASIN AND RANGE", "BLUE RIDGE", "CASCADE-SIERRA MOUNTAINS", "CENTRAL LOWLAND", "COASTAL PLAIN", "COLORADO PLATEAUS", "COLUMBIA PLATEAU", "GREAT PLAINS", "INTERIOR LOW PLATEAUS", "MIDDLE ROCKY MOUNTAINS", "NEW ENGLAND", "NORTHERN ROCKY MOUNTAINS", "OUACHITA", "OZARK PLATEAUS", "PACIFIC BORDER", and "PIEDMONT", "SOUTHERN ROCKY MOUNTAINS", "SUPERIOR UPLAND", "VALLEY AND RIDGE", and "WYOMING BASIN".

physioDL.csv: Table listing all image chips and associated physiographic province (id = unique ID for each chip; region = physiographic province; fnameHS = file name of associated chip in *hs* folder; fnameTHS = file name of associated chip in *ths* folder; set = data split (train, test, or validation).

chipCounts.csv: Number of chips in each data partition per physiographic province.

Tasks:

T1: Read in the *physioDL.csv* file as a Pandas DataFrame. (5 Points)

T2: Generate a new column in the table where each physiographic region is represented using a numeric code. (5 Points)

T3: Use the split column in the *physioDL.csv* file to partition the data into separate training, validation, and test Pandas DataFrames. (5 Points)

T4: Build a DataSet subclass that meets the following criteria: (10 Points)

1. Accepts an input Pandas DataFrame.
2. Returns a hillshade chip from the *hs* folder as a PyTorch tensor of shape (1, 256, 256) with a 32-bit float data type and the numeric index representing the associated class as a long integer.
3. Hillshade chips should be read using Rasterio.
4. The hillshade chips are 8-bit and scaled from 0 to 255. They should be rescaled to 0 to 1 by dividing by 255.

T5: Instantiate training, validation, and testing DataSets using your new subclass. (5 Points)

T6: Instantiate training, validation, and testing DataLoaders using your instantiated Datasets. Use a mini-batch size of 32. (5 Points)

T7: Print summary information for a mini-batch of chips. Confirm that the chips have the correct shape/dimensionality, the correct data types (32-bit float for images and long integer for class indices), and range of cell values and class indices. (5 Points)