A Comparative Study on Traditional Machine Learning and a Simple CNN with Land Use Satellite Image Classification

September 2024

1 Introduction

With the growing number of satellite launches with modern sensor packages for image capture and land surveying, a wealth of publicly available datasets have emerged to encourage both commercial innovation and academic research with wide applications from agriculture, urban planning, crisis response, climate modeling, defense, and so on [4, 5]. One specific domain of satellite imagery is land use image classification: the process of identifying and categorizing regions of the Earth by how they are utilized (agricultural, urban, etc.) or by their physical characteristics (rivers, forests, water bodies, etc.) [4, 5]. To classify images of the Earth at scale, machine learning (ML) approaches provide a means to classify land use with granularity and accuracy. In this paper, we explore and compare a simple convolutional neural network (CNN) with a more traditional random forest.

The structure of the report is as follows: (1) an introduction to the task; (2) the problem formulation; and, (3) the methodology involving an exploration of the EuroSAT dataset, the feature selection process, the hypothesis space, loss functions, and model validation. In stage two of the report, we will discuss (4) the results of our comparison and (5) a conclusion with discussions on where to proceed next.

2 Problem Formulation

We will use the EuroSAT dataset [4, 5] to perform supervised multi-class image classification in the following form: given a satellite image patch, find the correct land use category from a predefined list. Sentinel-2 images are freely and openly accessible as part of the Copernicus program, and the EuroSAT dataset with its labeled curation can be accessed through https://github.com/phelber/EuroSAT. The dataset consists of 27,000 labeled satellite image patches with 2,000 to 3,000 images per class across 10 land use categories. Each image patch is 64x64 pixels with a feature vector containing 13 spectral bands.

The dataset comes labeled into 10 categories: (1) Annual Crop, (2) Forest, (3) Herbaceous Vegetation, (4) Highway, (5) Industrial, (6) Pasture, (7) Permanent Crop, (8) Residential, (9) River, and (10) Sea/Lake. Each image patch is associated with one of the above class labels denoted by $y \in \{1, \dots, 10\}$. The categories are nominal with no inherent order or rank.

3 Methodology

3.1 Data Points

The EuroSAT dataset comprises of 27,000 geo-referenced images of Europe captured by the Sentinel-2A satellite's Multispectral Imager. The image patches in the dataset are uniformly 64x64 pixels with 13 spectral bands and an associated class label. As each image comprises of pixels of multispectral bands, each pixel is represented by 13 intensity values, one for each spectral band. Though the intensity values are stored to a resolution of 12 bits as integers–from 0 to 4096–, these values represent a continuous spectrum of light intensity and are treated as continuous numerical data. The feature vector $\mathbf{x}$ for a single image patch can be represented as $\mathbf{x} = [b_1, \dots, b_{13}]$ where b_i represents the values for spectral band i for all pixels in the patch.

3.2 Feature Selection and Engineering

Before training, each image patch will undergo the following processing protocol. First, each spectral band will be normalized to ensure consistent scale across bands using min-max normalization: $X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}}$. Next, we apply PCA to reduce the dimensionality of the 13 spectral bands while retaining essential variance and reducing noise [6]. We keep the top 6 principle components to retain 95% of the variance in the data. Then, we calculate three common spectral indices widely used for land cover classification: Normalized Difference Vegetation Index (NDVI) [9], Normalized Difference Water Index (NDWI) [8], and Normalized Difference Built-up Index (NDBI) [10].

$$\text{NDVI} = \frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red}}, \quad \text{NDWI} = \frac{\text{Green} - \text{NIR}}{\text{Green} + \text{NIR}}, \quad \text{NDBI} = \frac{\text{SWIR} - \text{NIR}}{\text{SWIR} + \text{NIR}} \tag{1}$$

We then compute the Gray Level Co-occurrence Matrix (GLCM) features on the first principle component to capture spatial patterns and extract contrast, correlation, energy, and homogeneity. These features help to capture spatial relationships between pixels which can help distinguish between different land cover types.

The result of the feature engineering protocol results in a feature vector consisting of 6 PCA components, 3 spectral indices, and 4 texture features. To assess the importance of each feature, we analyze the explained variance ratio of the PCA components, the distribution of spectral indices across different classes, the correlation matrix of all engineered features, and the PCA loadings to understand the contribution of each original spectral band.

3.3 Model Selection

In our land use classification task, we compare a Convolutional Neural Network with a Random Forest approach. Here, we compare the strength of the more modern approach against a robust classical machine learning method. We aim to evaluate the trade-offs between automatic feature learning (CNN) and manual feature engineering (Random Forests).

Convolutional Neural Networks We propose using CNN as a model for this classification task. Specifically, we will use a ResNet-50 architecture [3] pre-trained on ImageNet [1] and fine-tuned on the EuroSAT dataset. The motivation behind using a CNN is that they are well-suited for image data and capable of learning hierarchical features directly from raw pixel values. ResNet-50 in particular has shown excellent performance on various image classification tasks, including remote sensing applications [7]. The architecture consists of 50 layers, including convolutional layers, batch normalization, ReLU activation functions, and skip connections. The residual blocks mitigate vanishing gradients to improve training stability. The general structure of ResNet-50 is an initial convolutional layer, max pooling layer, 4 residual block stages, global average pooling, and a fully connected layer with softmax activation. Lastly, the choice of using a transfer learning setup is to speed up the training process and save compute time by starting with a model pre-trained on the large ImageNet dataset and having learned general image features. The hypothesis space for our ResNet-50 can be represented as:

$$H = h : X \to Y | h(x) = \text{softmax}(W_L \cdot f_{L-1}(...f_1(x))) \tag{2}$$

where f_i represents convolutional and pooling layers, W_L represents the weights of the final fully connected layer, and softmax as the activation function for multi-class classification.

Random Forests To provide a point of comparison and validate the CNN approach, we also implement a Random Forest model. A Random Forest is an ensemble of decision trees in which each tree in the forest is built from a bootstrap sample of training data, and each node is a random subset of features for splitting. The hypothesis space can be represented as:

$$H_{RF} = h : X \to Y | h(x) = \text{mode}(h_1(x), h_2(x), ..., h_T(x)) \tag{3}$$

where h_i represents individual decision trees, and T is the number of trees in the forest. For our implementation, we use 100 tree with no maximum depth. We apply a minimum samples split of 2 and a minimum samples leaf of 1. The maximum features shall be $\sqrt{n}$ where n is the number of features.

Random Forest is a strong choice as our traditional machine learning model as we have obtained a rich set of features comprised of PCA components, spectral indices, and textures. Furthermore, it is less robust to overfitting due to their ensemble and can allow us to assess the importance of the engineered features to the classification decision.

3.4 Loss Function

In this report, we employ two models that utilize the loss function in different ways. We utilize categorical cross-entropy loss for the CNN and the Gini impurity for the Random Forest [2].

3.4.1 Categorical Cross-Entropy

The categorical cross-entropy loss function is well-suited for multi-class classification and provides a measure of dissimilarity between predicted and true probability distributions. The loss heavily penalizes incorrect predictions while pushing the network into more decisive precisions when paired with softmax activation. Cross-entropy is also differentiable, allowing us to leverage gradient-based convex optimization of our CNN. The loss function is defined as:

$$L(y, \hat{y}) = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{4}$$

where C is the number of classes, y_i is the true probability of the sample belonging to class i, and $\hat{y}_i$ is the predicted probability of the sample class belonging to class i.

3.4.2 Gini Impurity

For the Random Forest model, we use Gini impurity as the criterion for splitting nodes during the tree construction process. The Gini impurity measures how often a randomly chosen element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset. The metric reaches its minimum when all cases in a node fall into a single target category. The Gini impurity is defined as:

$$G = \sum_{i=1}^{C} p_i(1 - p_i) \tag{5}$$

where C is the number of classes and p_i is the proportion of samples of class i at a given node.

3.5 Model Validation

To ensure robust evaluation of our models, the dataset is split into training, validation, and test sets in a stratified manner to ensure each set maintains the overall class distribution of the original dataset. The split ratios are as follows: training set: 70% (18,900 images), validation set: 15% (4,050 images), and test set: 15% (4,050 images). The 70/15/15 gives us plenty of data for training and enough for robust testing and validation. We apply a random state seed of 7 to maintain deterministic behavior.

To evaluate and compare our models, we employ the following metrics: accuracy, precision, recall, and F1 score. As a harmonic mean of precision and recall, the F1 score gives us a better overall performance measure accounting for both precision and recall of each class. Below are the validation metric expressions where TP, TN, FP, FN are True Positives, True Negatives, False Positives, and False Negatives respectively.

$$\begin{aligned} \text{Accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} & \text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}} \\ \text{F1} &= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} & \text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \end{aligned} \tag{6}$$

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: a large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 06 2009.

[2] Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016. http://www.deeplearningbook.org.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[4] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Introducing eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 204–207. IEEE, 2018.

[5] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019.

[6] Ian T. Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Phil. Trans. R. Soc. A.*, 374, 2016.

[7] Ying Li, Haokui Zhang, Xizhe Xue, Yenan Jiang, and Qiang Shen. Deep learning for remote sensing image classification: A survey. *WIREs Data Mining and Knowledge Discovery*, 8(6):e1264, 2018.

[8] S. K. McFEETERS. The use of the normalized difference water index (ndwi) in the delineation of open water features. *International Journal of Remote Sensing*, 17(7):1425–1432, 1996.

[9] Compton J. Tucker. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, 8(2):127–150, 1979.

[10] Yong Zha, Jingqing Gao, and S. Ni. Use of normalized difference built-up index in automatically mapping urban areas from tm imagery. *International Journal of Remote Sensing - INT J REMOTE SENS*, 24:583–594, 02 2003.