Visual Recognition by Satellite Imagery

Pratham Singhal (2021082), Yash Yadav (2021117), Manshaa Kapoor (2021540)

April 25, 2024

1 Introduction

Satellite picture categorization is a complex task involving remote sensing, computer vision, and machine learning, with existing techniques ineffective due to the unpredictability of satellite data and the absence of a single labeled high-resolution dataset. Satellite picture scene categorization is a challenging task involving terabytes of data with large variances. Detecting multiple land cover classes, such as trees, meadows, barren areas, and water bodies, is challenging due to their higher intra-class variability compared to highways. Satellite photography data's unpredictable nature has historically hindered deep neural network-based classification algorithms from delivering human-like performance. However, recent studies, particularly in the deep learning community, aim to retrofit deep learning approaches for high-resolution satellite data categorization. Proper land cover forecasting is crucial for monitoring environmental changes and human settlement expansion, and experts often spend significant time viewing satellite photographs. This study aims to analyze and evaluate two automated algorithms for land cover categorization using deep learning techniques, focusing on CNN architecture for multi-class semantic segmentation on satellite data. The goal is to improve visual recognition and evaluate the efficiency and accuracy of land cover categorization systems, contributing to ongoing research in deep learning for automated satellite image processing.

2 Problem Statement

Recent advancements in deep learning have revolutionized the field of land cover classification using satellite imagery. These advancements have enabled the development of automated approaches that can accurately and efficiently classify land cover types, offering a valuable tool for environmental monitoring and urban planning. This project focuses on evaluating two recent approaches for land cover classification using deep learning, with a specific emphasis on Convolutional Neural Network (CNN) architecture. By assessing the accuracy and efficiency of these approaches, this project aims to contribute to the ongoing research in automated analysis of satellite imagery. The evaluation will be conducted using a new dataset that represents a diverse range of land cover types and environmental conditions, ensuring that the models are tested under realistic scenarios. Through this evaluation, we seek to identify the strengths and weaknesses of these approaches and provide insights that can help improve future land cover classification models.

3 Research Paper 1

The paper "Land Cover Classification with U-Net: Satellite Image Multi-Class Semantic Segmentation Task with PyTorch Implementation of U-Net" by Srimannarayana Baratam introduces a methodology for land cover classification using the U-Net model, originally designed for biomedical image segmentation. The article discusses U-Net's implementation using PyTorch and its application in multi-class semantic segmentation of satellite pictures, highlighting its significant capacity for local and global information collection.

3.1 Model Architecture

• U-net is a prominent fully-convolutional architecture used for semantic picture segmentation. It is divided into two primary sections: the contractive (left) and expansive paths (right). The contracting and expanding paths are linked using skip connections.
• The contractive route involves repeating two 3x3 convolutions. Each 3x3 convolution is followed by a ReLU function and a 2x2 Max-pooling operation.
•Every downsampling of the feature map in the contractive route doubles the size of the channels. On the expanding route, however, the feature map is up-sampled before being subjected to a 2x2 convolution that divides the number of channels by two.
• The feature map is created by concatenating the contracting path and 2x2 convolution results, which are then up-sampled further.
• The architecture's final layer uses 1x1 convolution to reduce 64 components to the necessary 7 classes.

3.2 Dataset

• Source: Kaggle competition on Land Cover Classification (DeepGlobe 2018)
• Type: Satellite imagery with pixel-wise land cover annotations (segmentation)
• Region: Covers various locations worldwide
• Number of Classes: 21 land cover classes (e.g., buildings, forest, water, etc.)

3.3 Loss Functions

3.3.1 Simple Categorical Cross Entropy

While developing and training the network using a basic cross-entropy, it was discovered that the network predictions frequently landed on a single color for all pixels in every picture input. Because of the previously described substantial class imbalance, it is possible to conclude that this hue correlates to the over-represented class — yellow/agriculture land in our example.

Class	Mask Colour	Pixel Count (in millions)	Proportion
Urban Land	Cyan	461.19	11.27%
Agricultural Land	Yellow	2379.65	58.14%
Range Land	Magenta	343.12	8.38%
Forest	Green	444.92	10.87%
Water	Blue	138.39	3.38%
Barren Land	White	323.39	7.90%
Unknown	Black	2.35	0.06%

Table 1: The distribution of classes in our training+validation set

$$\text{loss}(x, class) = -\log\left(\frac{\exp(x[class])}{\sum_j \exp(x[j])}\right) = -x[class] + \log\left(\sum_j \exp(x[j])\right)$$

Figure 1: Simple Categorical Cross Entropy Loss Function

3.3.2 Dice Loss

The authors used Dice Loss for multi-class segmentation. However, the gradients appeared to be "freezed" during training the model, and it was deduced that specific actions on torch variables will detach the result.

$$L_{Dice} = 1 - \frac{1}{N_{classes}} \sum_{i=1}^{N_{classes}} \frac{2p_{y_{true}} p_{y_{pred}}}{p_{y_{true}} + p_{y_{pred}} + \varepsilon}$$

Figure 2: Dice Loss Function

3.3.3 Weighted Cross Entropy

Given an acceptable learning rate, this technique provides loss convergence over time. The weight for each class was determined by dividing the total occurrences of the class with the lowest presence by the total instances of each class in the dataset (training).

$$\text{loss}(x, class) = weight[class]\left(-x[class] + \log\left(\sum_j \exp(x[j])\right)\right)$$

Figure 3: Dice Loss Function

3.4 Results

3.4.1 Training Curves

The training loss is a weighted cross-entropy loss. The validation loss, however, is the basic cross-entropy.

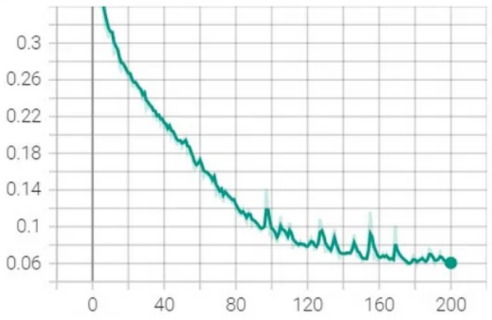

Figure 4: Training Curve showing the loss convergence over the training

3.4.2 Metrics

The authors evaluated the model using the Intersection of Union (IoU) measure and compared the results to the land cover segmentation task. IoU is a statistic that measures the accuracy of a segmentation method by overlapping the model's predictions with the ground truth.

3.4.3 Performance

The model achieves an IoU score of 0.608 at 30 epochs using the same picture scaling factor on the test set that is different from the training data. However, the model was trained on fewer photos, with lower resolution and batch size, and with half-precision due to memory/computing constraints. Thus, it is claimed that performance would undoubtedly increase in the absence of such bottlenecks.

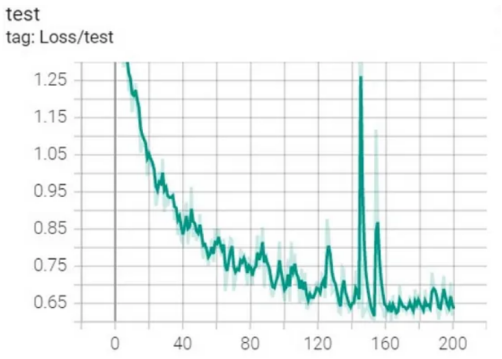

Figure 5: Training Curve showing the loss convergence over the validation

$$IoU = \frac{1}{N_{classes}} \sum_{i=1}^{N_{classes}} \frac{predictions_{c_i} \cap ground\_truth_{c_i}}{predictions_{c_i} \cup ground\_truth_{c_i}}$$

Figure 6: IoU Metric Function

4 Research Paper 2

4.1 Model Architecture

• Our proposed approach improves distribution separability for satellite picture classification by augmenting a current CNN architecture with handmade texture characteristics.
•The model comprises of two convolutional layers with 32 and 64 feature maps, each with a 3×3 kernel and a Rectified Linear Unit (ReLU) layer.
• The next layer is a max-pooling layer with a 2x2 kernel.
• The max pooling layer is followed by a 0.25 dropout layer. This is followed by a feature fusion layer, which combines the handmade features with the CNN bottleneck representations.
• The fused features are fed into a fully connected dense layer with 32 neurons, followed by a ReLU layer and a fully connected dense layer with 128 neurons. Batch normalization is also used. After this layer, there is a ReLU layer followed by a dropout layer with a rate of 0.2.
• The last layer is a Softmax layer using the crossentropy loss function. The Adadelta optimizer (Zeiler

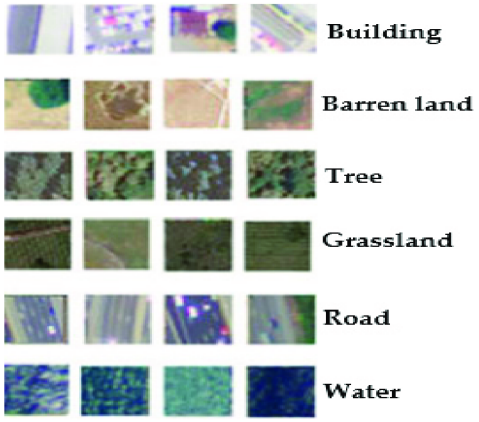

Figure 7: Sample image patches of SAT-6 Dataset

2012) was adopted in the framework.
• The feature extraction step generates 150 features from the input picture. The primary characteristics used for categorization are mean, standard deviation, variance, second moment, direct cosine transformations, and so on.

4.2 Dataset

SAT-6 Dataset:
• Type: High-resolution aerial imagery
• Classes: 21 land cover categories including agricultural fields, airplanes, baseball diamonds, beaches, and buildings.
• Resolution: Typically 256x256 pixels, with minor variations.
• Format: RGB (Red, Green, Blue) channels.
• Access: Tensorflow

4.3 Performance Analysis

• visualizes the map responses learned from the first fully connected dense layer, those learned from the second fully connected dense layer, and the decision boundaries, respectively, for a CNN augmented with handcrafted features while the top row shows the same for the same CNN without the handcrafted features.
• It can be seen from Figure 8 that fusing handcrafted features helped improve discriminative feature learning (see Figure 8(B), bottom row, where the others class is already more compactly clustered than in the top) providing robust separation of the decision boundaries (see Figure 3(C) where the bottom row shows clearer separation of the classes than the top where the classes trees, grassland, and others are not robustly separable and the intra-class distances are more). This is corroborated by the higher distances between means and the lower standard deviations for the handcrafted features as shown in Table 9.
• The authors compares the performance of the model with related models referred to in different research paper (mentioned in Figure 8).

Methods	SAT-4 Accuracy (%)	SAT-6 Accuracy (%)
DBN (Basu et al. 2015a)	81.78	76.47
SDAE (Basu et al. 2015a)	79.98	78.43
CNN (Basu et al. 2015a)	86.83	79.10
DeepSat (Basu et al. 2015a)	97.95	93.92
Contrastive loss (Simo-Serra et al. 2015)	98.74	98.55
MLP (Z-score) (Zhong et al. 2017)	94.76	97.46
DCNN (Ma et al. 2016)	98.41	96.04
TradCNN (Z-score) (Zhong et al. 2017)	98.43	98.34
D-DSML-CaffeNet (Gong et al. 2018)	99.51	99.42
SatCNN (linear) (Zhong et al. 2017)	99.55	99.58
SatCNN (Z-score) (Zhong et al. 2017)	99.69	99.61
Triplet networks (Liu and Huang 2018)	99.76	99.71
DeepSat V2 (The proposed method)	99.90	99.84

Figure 8: Comparison of classification accuracy percentage of various methods on and SAT-6 datasets.

5 Dataset

We would be analysing the results of both the models on a dataset the models didn't while training. For this purpose, we will be using the UC Merced Land Use Dataset.
• Type: The dataset consists of high-resolution aerial imagery captured from above, providing a detailed view of the Earth's surface.
• Number of Classes: There are 21 distinct land cover classes represented in the dataset. These classes include agricultural areas, airplanes, baseball diamonds, beaches, buildings, chaparral, dense resi-

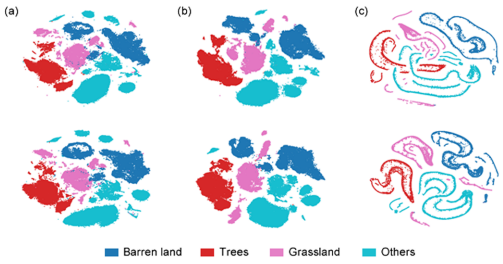

Figure 9: Visualization of learned representations and decision boundaries for SAT-4 dataset. Top row, regular CNN model which has no handcrafted features fused. Bottom, proposed framework which has handcrafted features fused. (a) Feature maps learned from the first dense layer. (b) Feature maps learned from the second dense layer. (c) Decision Boundaries.

Dataset	Type	Distance between Means	Mean of Standard Deviations
SAT-4	Raw Images	0.1994	0.1166
	Handcrafted DeepSat Features	0.8454	0.0435
SAT-6	Raw Images	0.3247	0.1273
	Handcrafted DeepSat Features	0.9726	0.0491

Figure 10: Distance between Means and Means of Standard Deviations for raw image values and DeepSat feature vectors for SAT-4 and SAT-6.

dential, etc. This variety allows for a comprehensive analysis of land cover types commonly encountered in urban and natural environments.

• Image Resolution: The images in the dataset typically have a resolution of 256x256 pixels, providing sufficient detail for accurate classification and analysis. However, there may be slight variations in resolution across the dataset.

• Format: The images are stored in RGB format, where each pixel is represented by three color channels: Red, Green, and Blue. This format enables the visualization of the images as color images, allowing for the detection of patterns and features based on color information.

6 Analysis and Inference

• In the predicted images for the U-Net and DeepSat Models for the pasture lands, we could identify how differently both of the models have predicted the images. U-Net has identified more of the 90 percent image as pasture land, and color coded similarly. Whereas the DeepSat V2, identified the whole image as a 97percent probability of having a pasture land.

• A visible difference here is that UNet can segment a larger scope of the satellite image whereas the DeepSat V2 need more accurate and zoomed satelite image to give a proper probability of the class.

• The predicted River images by both of the models, are identified correctly. Similarly as the previous trend we can see both of the models results are baised over the magnification of the satellite image, whereas the Unet model is more accurate pixel wise.

• Image magnification while clicking the picture from the satellite can be decieving while predicting the type of geography.

• Coexistence of different Geography in one image is again a problem for DeepSAT V2, since it can give baised results.

• U-Net can miscalssify various highly correlated types like agriculture land and pasture land, since the pixel resolution can be baised around the type of geography.

• River and Highways can be misclassified if they happen to exist coherently in the same picture by both of the models, since the decideing factor is way to noisy for the model, given if the training sample is small for both of the types.

• pixel wise upsampled segmentation is good for real time analysis whereas the DeepSAT V2 classification can be really usefull in static image analysis where the magnification task is simplified.

• both of the models can be used simultaneously in any of the image analysis tool to simplify the complexity at varied steps.

7 Individual Analysis - Pratham Singhal

In this report we recreated the two existing CNN models for Semantic Image segmentation and Classification of the Satellite Images of different geographic

terrains. Comparing different CNN models like DeepSAT, DEEPSAT V2, StrNet, UNET, etc for the classification of the images, we found that satellite images can be classified to different Geographic Classes, depending upon its features. But it is not a pixel upsampled semantic segmentation for the different classes, hence limiting its usage to only accurately identifying the image type instead of differently identifying the pixels of different classes within the image. For this we take up an another simple U-NET model which does the semantic segmentation of pixel wise different classes of geographic terrains. This model helps us identify the presence of each type of geographies present in the view point of the satellite image. This type of model is very useful in real time imagery, where we can identify the geography of the satellite image simultaneously, and identify the live location of the specified type of the terrain. Whereas the later one can be useful in passive identification of the geography type. Whereas the future prospects of a new model could be devised where we have the chance to merge both of them at one place and use image classification first and then based on the type of the image we recreate the pixel wise location of the predicted type into the image. This gives us an accurate location and identification of the terrain and could be computationally feasible.

8 Individual Analysis - Yash Yadav

In this report, we recreated two CNN models for segmentation and classification for high resolution Satellite Image Dateset. We identified two different tasks on an image data which is segmentation and classification and tried to provide a better analysis by segmenting the data and then classifying the segmented result. Comparing different CNN models like DeepSAT, DEEPSAT V2, StrNet, UNET, etc for the classification of the images, we found that satellite images can be classified to different Geographic Classes, depending upon its features. This model helps us identify the presence of each type of geographies present in the view point of the satellite image. This type of model is very useful in real time imagery, where we can identify the geography of the satellite image simultaneously, and identify the live location of the specified type of the terrain.

9 Individual Analysis - Manshaa Kapoor

• In this paper, we reconstructed two existing CNN models for semantic image segmentation and classification of satellite images from various geographical terrains. Comparing several CNN models for image classification, such as DeepSAT, DEEPSAT V2, StrNet, and UNET.
• We discovered that satellite pictures may be categorized into distinct Geographic Classes based on their attributes. However, because it is not a pixel upsampled semantic segmentation for the various classes, its use is limited to precisely recognizing the picture type rather than distinguishing between pixels of different classes within the image.
• For this, we use another basic U-NET model to do semantic segmentation of distinct types of geographic terrains pixel by pixel. This model assists us in identifying the existence of each sort of geography in the satellite picture. This model is highly important in real-time photography, as it allows us to recognize the geography of the satellite picture while also determining the live position of the required kind of terrain.
• Whereas the latter can be beneficial in passively identifying the geographical type.
• Whereas the future possibilities of a new model might be designed where we have the opportunity to integrate both of them at one place and utilize image classification first, and then based on the type of the picture, we reconstruct the pixel-by-pixel placement of the predicted type in the image. This provides an exact location and identification of the terrain and may be computationally possible.

10 References

• 'Land Cover Classification with U-Net: Satellite Image Multi-Class Semantic Segmentation Task with PyTorch Implementation of U-Net' by Srimannarayana Baratam.
• Feature Augmented Convolutional Neural Nets for Satellite Image Classification' by Qun Liu, Saikat Basu, Sangram Ganguly, Supratik Mukhopadhyay, Robert DiBiano, Manohar Karki, Ramakrishna Nemani.