SATELLITE IMAGE CLASSIFICATION: AN APPLICATION OF CONVOLUTIONAL NEURAL NETWORKS, TRANSFER LEARNING, AND EXPLAINABLE AI

Yene Irvine
yene.irvine@ucalgary.ca

Department of Electrical and Computer Engineering, University of Calgary, Calgary, Canada

ABSTRACT

This study introduces a robust deep learning model leveraging Convolutional Neural Networks (CNNs) and transfer learning to classify a comprehensive dataset of satellite images into a detailed 35-class structure as well as a consolidated 7-class structure. The model achieved exceptional test accuracies, illustrating the effectiveness of transfer learning and the richness of the dataset. A pivotal component of this research was the application of Integrated Gradients, an Explainable AI (XAI) method, which provided transparency into the model's reasoning, particularly highlighting the importance of edge features in geographical classifications. The fusion of high classification accuracy with model interpretability offers valuable insights for enhancing automated satellite image analysis systems, with broader implications for managing and understanding the Earth's multifaceted landscapes.

Index Terms— deep learning, image classification, convolutional neural networks, transfer learning, integrated gradients, Explainable AI

1. INTRODUCTION

Satellite imagery serves as a pivotal tool in how the Earth's surface is observed and interpreted, offering valuable insights for environmental monitoring, urban planning, disaster response, and agriculture among other purposes. The effectiveness of such applications hinges on the accurate classification and interpretation of the images. The challenge lies in the vast amount of data and the complexity of the images involved.

Deep learning, a branch of machine learning, has revolutionized image classification tasks with its ability to discern patterns and features often imperceptible to the human eye. Convolutional neural networks (CNNs), in particular, stand at the forefront of image analysis [1]. They are therefore well-suited for high-resolution satellite imagery classification, such as with the *RSI-CB256* dataset used in this project that classifies various satellite images of the Earth into categories like ‘sea’, ‘mountain’, and ‘forest’ among many others [2].

This project aims to not only develop a deep learning model that proficiently classifies the granular 35-class structure of the original dataset but also to compare its effectiveness against a more generalized classification into the 7 broader labels under which each of the 35 classes are categorized based on the dataset’s folder structure. A key aspect of this study is to unveil the model's decision-making process by implementing Integrated Gradients, an Explainable AI (XAI) technique that helps in understanding the model's reasoning behind its classifications.

This investigation pursues a dual objective: achieving high accuracy in satellite image classification and delving into the interpretability of model predictions. This exploration is vital, as it contributes to the domain of remote sensing by facilitating improved automated image classification systems, which in turn, enhances management and understanding of the Earth's diverse landscapes.

The following sections will outline related work, describe the methodology for developing the CNN model, present the results with classification metrics, and discuss the implications of these findings for future applications.

2. RELATED WORK

The implementation of deep learning frameworks to classify satellite imagery has been completed before in prior works. Abburu and Golla’s 2015 article ‘Satellite Image Classification Methods and Techniques: A Review’, for instance, explores various automated satellite image classification methods [3]. Additional related research highlights an approach that combines the application of CNNs and transfer learning with satellite metadata integration for enhanced accuracy [4].

The purpose of satellite image classification can be highlighted by its crucial role in disaster response scenarios. For instance, Duarte et al. detail the use of CNNs with satellite data for the rapid assessment of building damages from multi-platform remote sensing data, demonstrating the critical application of these technologies in aiding timely and effective disaster relief efforts [5]. This addition underscores the practical implications of the research, demonstrating the value of satellite image classification in a real-world context.

While these related studies achieved success in highlighting the applicability and purpose of CNNs in classifying satellite image data, this project’s approach differs by contrasting the classification effectiveness between granular and broader category levels and employing

Integrated Gradients to interpret model decisions. Integrated gradients serve as a vital tool in explaining deep neural networks, particularly by generating heatmaps that reflect the influential regions in an input image for model decisions. Qi et al.'s work emphasizes the effectiveness of integrated gradients in creating more accurate and interpretable heatmaps, enhancing the understanding of model behavior in image classification tasks [6]. The implementation of XAI techniques like integrated gradients is beneficial in a variety of image classification problems, including the critical field of medical image analysis [7].

Another key tool applied in this study is transfer learning, which mitigates the data collection process and circumvents building an image classification model from scratch. Among the effective pre-trained image classification models is the VGG-16 CNN model. Tammina's 2019 study highlights how transfer learning with the VGG-16 can effectively adapt to image classification tasks, even with limited labeled data in the target domain [8]. Another powerful pre-trained image classification model is EfficientNet, which leverages ImageNet, a database of millions of labeled images, to achieve high accuracy scores with notably few parameters relative to other models [9]. ResNet50 is another image classification model that uses the ImageNet library for pre-training. Studies using this model often highlight its prowess in image detection, such as the 2020 report by Theckedath that highlights its adaptability to tasks beyond object classification such as emotional state detection [10].

3. MATERIALS AND METHODS

3.1. Materials

In the execution of this project, a number of computational tools and resources were used to develop, train, and analyze the performance of deep learning models. The primary development environment was Visual Studio Code enhanced with the RemoteSSH extension. This setup allowed for seamless integration with TALC, the University of Calgary's High-Performance Computing (HPC) cluster, enabling direct coding within a powerful computational environment. Detailed information about TALC, which houses advanced GPU resources, can be found on its official webpage [11].

The deep learning models were constructed and trained using PyTorch, an open-source machine learning library widely acclaimed for its flexibility and efficiency in creating and training neural networks. Alongside PyTorch, Weights & Biases (WandB) played a pivotal role in the project, offering a comprehensive platform for real-time monitoring and retroactive analysis of the models. WandB's dashboard was instrumental in organizing and visualizing various aspects of the model training and evaluation process. It was used to monitor training loss to provide insights into model convergence in real-time, track system utilization data—particularly GPU usage—which is critical for optimizing computational resources, and provides a centralized consolidation of different runs to compare and iterate between.

In addition to PyTorch and WandB, some additional Python libraries were incorporated to support various functions throughout the model:

- 'PIL.Image' from Python Imaging Library (PIL) for image processing tasks,
- 'datetime' for timestamping and scheduling within the script,
- 'sklearn.metrics' which provided tools such as confusion matrix, ROC curve, AUC, precision score, and recall score, essential for model performance evaluation,
- 'numpy' and 'matplotlib.pyplot' for numerical operations and data visualization, respectively,
- 'seaborn' for enhanced data visualization aesthetics and utility, and
- 'IntegratedGradients' from 'captum.attr' for advanced model interpretability.

The application of these libraries, particularly those related to data visualization and model interpretability, is explained in further detail in the following Methods and Results sections, showcasing their contribution to the project's findings.

3.2 Methods

Transfer learning was applied in this project, leveraging pre-trained models to provide a foundation for to be refined and adapted to this specific image classification problem. Three reputable architectures were employed: VGG16, EfficientNet, and ResNet50, each acclaimed for their performance in image classification tasks. The choice to forgo the typical approach of freezing layers during feature extraction was deliberate; preliminary trials yielded exceptionally high accuracy, credited to the substantial dataset comprising 24,747 images. This dataset's richness permitted a comprehensive feature learning, alleviating the need for the layer freezing technique.

The model that initially demonstrated superior performance was subjected to further analysis. This involved a thorough exploration of hyperparameters such as the number of epochs, batch size, learning rate, and the implementation of a learning rate scheduler. Iterative experimentation led to an optimal configuration, although exhaustive trials of various parameters are not itemized here. For instance, batch sizes were varied, ranging from 32 to 90 among various tests, number of epochs was configured between ranges of 1 and 3, and a few learning rates were tested between 0.00001 and 0.0001.

Upon tuning, some key metrics were computed: overall test accuracy, class-wise accuracy, precision, and recall. Additionally, a confusion matrix was generated to visually represent the classification performance across different classes. To delve deeper into the model's decision-making process, Integrated Gradients, a feature attribution method within the domain of XAI, was implemented. This tool provided insight into the model's predictions, highlighting decisions for both correct and incorrect classifications. The algorithm was programmed to select random images for detailed analysis, while also incorporating at least one erroneous prediction for analysis.

The experiment was subsequently replicated for a 'higher-order' classification challenge. The initial 35 classes were grouped into 7 overarching categories, aligning with the dataset folder structure (i.e., the entire original dataset is split into 7 folders, and about 5 subfolders exist within each of those folders). Using the transfer learning model that had previously shown the best results, the same analytical methodology was applied to these aggregated classes, examining whether performance metrics would sustain or improve when faced with broader classifications.

4. RESULTS AND DISCUSSION

In this section, the findings of the model's performance in accurately classifying satellite imagery into distinct classes are presented. These results are first evaluated for the 35-class classification problem, and then subsequently for the 7-class classification problem. For more detailed results, refer to the corresponding Jupyter Notebooks in the linked GitHub [12].

4.1. Comparative Model Performance

An initial 35-class image classifier was employed using three different pre-trained models to serve as a basis for choosing only one for further evaluation. The three models, (1) ResNet50, (2) VGG16, and (3) EfficientNet, yielded test accuracies of 99.5%, 98.8%, and 99.1% respectively. Therefore, the highest accuracy model, ResNet50, was selected for further evaluation due to its superior performance. It is noted that ResNet50's performance in this context may not necessarily translate to superior performance in the 7-class classifier, thereby indicating a potential imitation with this selection procedure, but it was deemed adequate based on the exceptional test accuracy score of 99.5%.

4.2. 35-Class Classifier

A subsequent run using the ResNet50 pre-trained model was performed, this time incorporating a suite of metrics for evaluation of the model's performance. This run yielded an overall test accuracy of 98.98%%, precision score of 99.00%, and recall score of 98.98%. A 35 x 35 confusion matrix illustrating class-level accuracy scores, which was too granular for feasible inclusion in this report due to formatting structure, can be found in the Jupyter Notebook. The lowest individual class accuracy was 'bridge', which scored an accuracy of 88.7%, roughly 7% lower than the second lowest class-specific accuracy. These exceptionally high scores are a testament to the effectiveness of transfer learning as a powerful deep learning tool, as well as the benefit of having such a large dataset.

Because only two epochs of batch size 90 were used for this run, a plot of the training loss or validation loss for each epoch did not seem feasible. Therefore, a graph indicating batch training loss for every 10 batches is shown in **Figure 1**, generated from WandB.

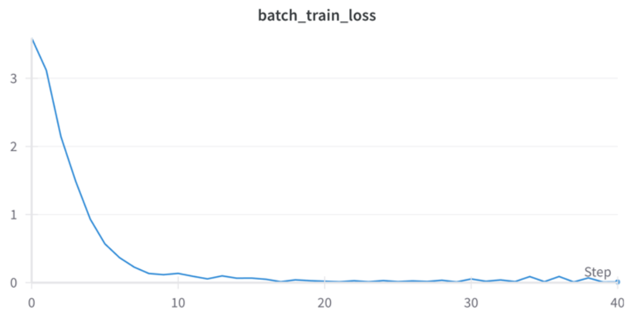

Fig. 1. Batch Training Loss for 35-Class Classifier

The rapid decline in loss at the beginning indicates that the model learned quickly from the data. After the initial sharp drop, the curve flattens out, suggesting that with each batch, the rate of learning or improvement in loss decreases. The stability maintains throughout, indicating that the model does not experience high variance between batches and is likely not overfitting to the training batch data, which is reflected in the exceptional test scores achieved. More detailed results on the model's performance can be found on the run's WandB dashboard [13].

Integrated gradients represent a powerful technique for interpreting machine learning models by attributing the prediction of a model to its input features (i.e., pixel values in image classification). The technique was implemented in this model, applied to five random samples from the test set to highlight which aspects of the image were the most influential in the model's prediction. In **Figure 2**, a random original image from the test set belonging to the class 'lakeshore' is displayed alongside its corresponding attributions map, which is depicted in red tones. The highlighted pixels clearly indicate a strong focus along the coastline, suggesting that this area holds critical information for predicting this type of geographic feature. In this case, the model successfully predicted the label as 'lakeshore.'

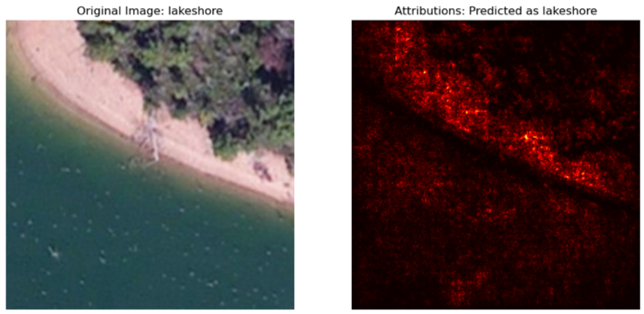
Original Image: lakeshore
Attributions: Predicted as lakeshore

Fig. 2. Integrated Gradients Attribution Map Correct Prediction for 35-Class Classifier

4.3. 7-Class Classifier

The broader 7-class classifier performed similarly exceptionally, yielding a test accuracy, precision, and recall of 99.41% across the board. A 7 x 7 confusion matrix illustrating class-wise results is displayed in **Figure 3**.

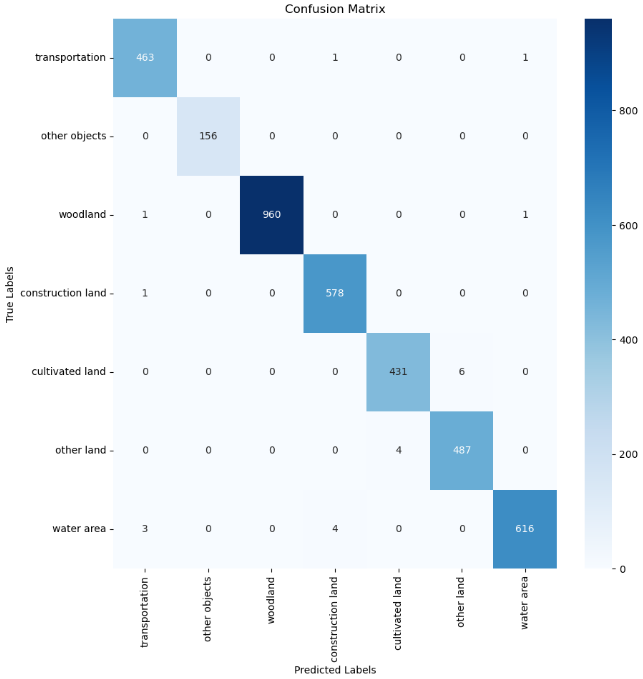

Fig. 3. 7-Class Classifier Confusion Matrix

The training loss curve, which logs training loss at every tenth batch, is shown in **Figure 4** and exhibits similar characteristics to those observed in the 35-class classifier's training loss curve in **Figure 1**. A thorough representation of the results can be found on the WandB dashboard for the 7-class classifier [14].

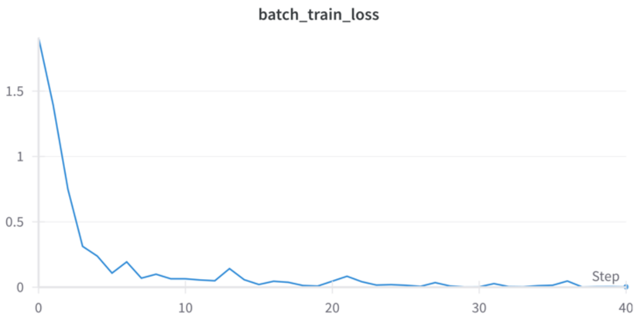

Fig. 4. Batch Training Loss for 7-Class Classifier

In an effort to gather further insight into the model's decision-making process, more integrated gradient attribution maps were generated for the 7-class classifier model. The initially code was manipulated slightly, however, to include at least one incorrect prediction, which is shown in **Figure 5**.

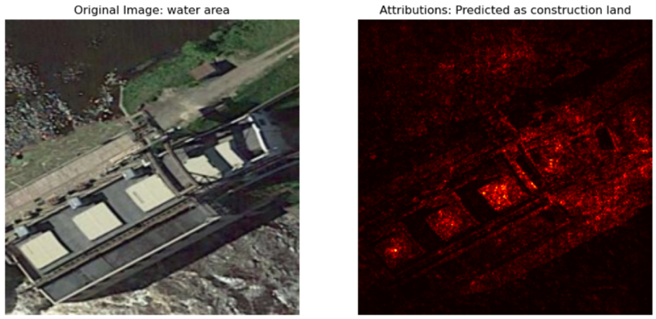
Original Image: water area
Attributions: Predicted as construction land

Fig. 5. Integrated Gradients Attribution Map Incorrect Prediction for 7-Class Classifier

This attribution map clearly indicates the model's focus on certain distinct edges. In this case, it misattributes the 'water area' image as 'construction land' and seems to have been misled by the square shapes that are often associated with aerial views of storage containers or perhaps city buildings, both of which fall under the broader class of 'construction land.' This highlights a limitation in the model's ability to generalize from training data to new unseen examples perfectly every time, particularly in complex scenarios where class features are not distinctly separable. Many more attribution maps of the model's prediction relative to the original image can be found in the corresponding Jupyter Notebook linked on GitHub.

In addition to some of the limitations addressed throughout this report, one of the key challenges associated with this applied methodology (i.e., running a comprehensive 35-class classifier model and then replicating on a 7-class classifier model) came from the excessively long training time for each job, typically ranging anywhere between 5 and 8 hours per job. This added a level of complexity to the workflow, particularly when errors in the code persisted, sometimes at the very final stages of the job after training, like during the integrated gradients attribution mapping process. If a similar project is replicated in the future, a proposed solution to this challenge would be to complete training and validation in one job, on a separate Jupyter Notebook, and then export the best model '.pth' file to a distinct directory that can be loaded up for a new job that consistutes testing and subsequent stages, thus creating a sort of checkpoint.

5. CONCLUSIONS

This investigation into satellite image classification reinforces the significance of comprehensive datasets and advanced learning techniques in improving satellite imagery analysis. It culminated in the development of a deep learning model that not only achieves high accuracy across a complex 35-class problem but also retains effectiveness when generalized to broader classifications. Leveraging the capabilities of CNNs and the strategic application of transfer learning, the model demonstrated an impressive capacity to discern intricate features within the dataset, achieving a test accuracy of 99.5% with the ResNet50 architecture.

The employment of Integrated Gradients has shed light on the model's decision-making, enhancing transparency and offering a valuable tool for interpreting AI-driven classifications. Notably, the model's focus on edge detection in geographical features underscores the potential of XAI in remote sensing applications.

Future endeavors could extend these methodologies to broader datasets and apply them across various types of remote sensing data, potentially enhancing a wide array of applications that benefit from satellite observations of the Earth.

6. REFERENCES

[1] O'shea, Keiron, and Ryan Nash. "An introduction to convolutional neural networks." *arXiv preprint arXiv: 1511.08458* (2015).

[2] Digra, Monia (2023). RSI-CB256. Figshare. Dataset. https://doi.org/10.6084/m9.figshare.22139921.v1. Accessed 2024.

[3] Abburu, Sunitha, and Suresh Babu Golla. "Satellite image classification methods and techniques: A review." *International journal of computer applications* 119.8 (2015).

[4] Pritt, Mark, and Gary Chern. "Satellite image classification with deep learning." *2017 IEEE applied imagery pattern recognition workshop (AIPR)*. IEEE, 2017.

[5] Duarte, D., Nex, F., Kerle, N., and Vosselman, G.: Satellite Image Classification of Building Damages Using Airborne and Satellite Image Samples in a Deep Learning Approach, ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci., IV-2, 89–96, https://doi.org/10.5194/isprs-annals-IV-2-89-2018, 2018.

[6] Qi, Zhongang, Saeed Khorram, and Fuxin Li. "Visualizing Deep Networks by Optimizing with Integrated Gradients." *CVPR workshops*. Vol. 2. 2019.

[7] Van der Velden, Bas HM, et al. "Explainable artificial intelligence (XAI) in deep learning-based medical image analysis." *Medical Image Analysis* 79 (2022): 102470.

[8] Tammina, Srikanth. "Transfer learning using vgg-16 with deep convolutional neural network for classifying images." *International Journal of Scientific and Research Publications (IJSRP)* 9.10 (2019): 143-150.

[9] Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." *International conference on machine learning*. PMLR, 2019.

[10] Theckedath, Dhananjay, and R. R. Sedamkar. "Detecting affect states using VGG16, ResNet50 and SE-ResNet50 networks." *SN Computer Science* 1.2 (2020): 79.

[11] "TALC Cluster." Research Computing Services, University of Calgary, Sept. 2023, rcs.ucalgary.ca/index.php/TALC_Cluster. Accessed 2024.

[12] Irvine, Yene. "Satellite Image Classification." GitHub, yirvine, 2024, https://github.com/yirvine/Satellite-Image-Classification.

[13] "35-Class Classifier Results." Weights & Biases, Weights & Biases, Inc., 2024, https://wandb.ai/yeneirvine/645-Final-Project/runs/zmdpwgy1. Accessed 2024.

[14] "7-Class Classifier Results." Weights & Biases, Weights & Biases, Inc., 2024, https://wandb.ai/yeneirvine/645-Final-Project/runs/fvigl5l0. Accessed 2024.