UNIVERSITY OF MASSACHUSETTS DARTMOUTH

CLASSIFICATION OF SOIL USING CNN AND OpenMP

A Final Project Report of

High Performance Scientific Computing

MASTER'S

In

DATA SCIENCE

Presented by

Maleeha Arif(02036260)

Saheer Shaik(02080047)

Likhil Naik(02075503)

INTRODUCTION

The aim of this project is to accurately classify different soil types based on input images, while simultaneously improving the efficiency and speed of the classification process through parallelization. The objectives of the project are as follows:

1. **Hyperparameter Tuning:** To optimize the model's performance, various hyperparameters such as learning rate, batch size, number of epochs, and network architecture will be fine-tuned. Techniques like grid search and random search will be employed to search through the hyperparameter space and find the best combination of hyperparameters that result in the highest accuracy and efficiency.
2. **Data Augmentation:** To improve the model's generalization ability and to prevent overfitting, data augmentation techniques will be applied to the training dataset. These techniques may include rotation, scaling, flipping, and translation of the input images. By augmenting the dataset, the model will be exposed to a wider variety of soil samples, resulting in better soil classification performance.
3. **Transfer Learning:** To further enhance the model's classification capabilities and reduce training time, the use of pre-trained models, such as VGG16, ResNet, and Inception, will be explored. These models have already been trained on massive datasets and have proven their effectiveness in image classification tasks. By fine-tuning these models to the specific task of soil classification, the project aims to achieve even higher accuracy levels and faster convergence.
4. **Model Interpretability:** To better understand the model's decision-making process and to provide insights into its performance, visualization techniques like saliency maps and class activation maps will be employed. These visualizations will highlight the regions in the input images that are most influential in determining the soil type, offering valuable insights into the model's behaviour.
5. **Deployment and Integration:** The final trained and optimized model will be deployed on a web-based platform or a mobile application, allowing users to quickly and easily classify soil types using images captured with their devices. By integrating the model into an accessible interface, the project aims to promote widespread use of this tool for soil analysis in various applications such as agriculture, environmental monitoring, and urban planning.
6. **Performance Metrics:** The success of the project will be measured using key performance metrics such as accuracy, precision, recall, and F1 score. These metrics will be calculated for both the training and test datasets, enabling the evaluation of the model's generalization ability and robustness. Additionally, training time and prediction speed will be monitored to ensure the model meets the project's efficiency objectives.
7. **Future Enhancements:** The project will be designed with scalability and extensibility in mind, enabling the incorporation of new soil types or additional image features in the future. Furthermore, by leveraging advancements in deep learning algorithms and hardware acceleration technologies, the project aims to continually improve its performance and capabilities over time.

BACKGROUND

1. As the demand for sustainable agriculture and environmental management continues to grow, the need for accurate and efficient soil classification systems becomes increasingly important. One of the significant advantages of using CNNs for soil classification is their ability to learn complex patterns and features from input images, making them well-suited for tasks involving spatial data.

2. The integration of remote sensing technologies, such as satellite imagery and drone-based imaging systems, with CNNs presents new opportunities for soil classification. Combining these technologies allows for real-time, large-scale monitoring of soil types, which can help inform decision-making in agriculture, land management, and urban planning.

3. Despite the potential of CNNs for soil classification, challenges remain in obtaining high-quality, labelled training datasets. Crowdsourcing and citizen science initiatives can help address this issue by involving a broader community in the collection and annotation of soil data. Additionally, advancements in unsupervised and semi-supervised learning techniques can enable models to learn from partially labelled or unlabelled data, reducing the need for extensive manual annotation.

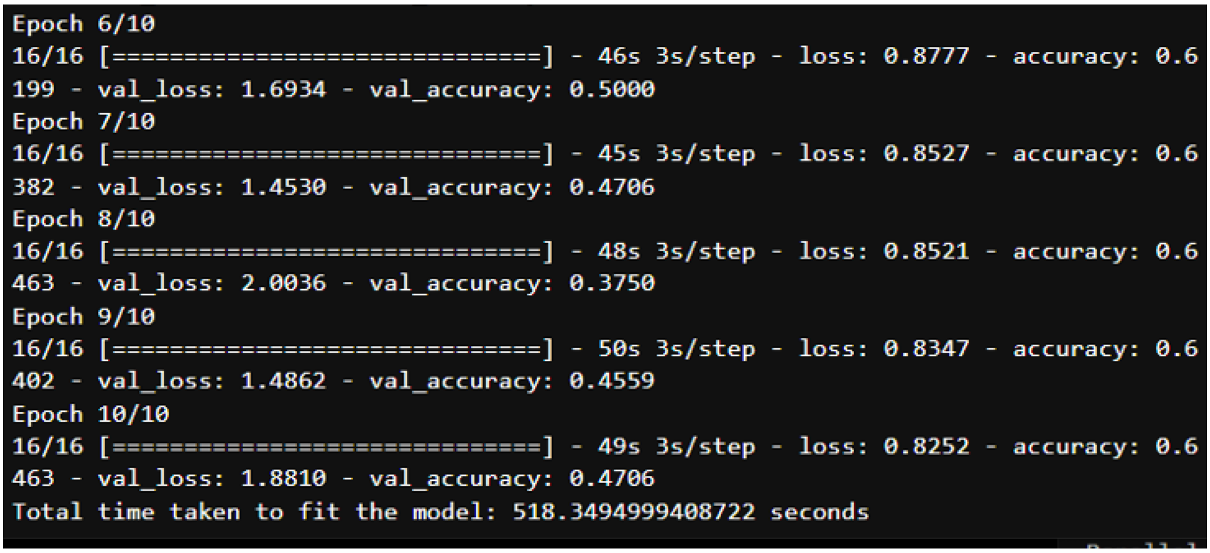
4. The development of specialized CNN architectures, tailored for soil classification tasks, can further enhance the performance and efficiency of these models. By leveraging domain-specific knowledge, these specialized architectures can incorporate relevant prior information, leading to better generalization and faster convergence during the training process.

5. To ensure the widespread adoption of CNN-based soil classification systems, it is crucial to develop user-friendly interfaces and applications that allow non-experts to access and utilize these tools. By making the technology more accessible, it can be integrated into existing workflows and decision-making processes, driving further innovation in agriculture and environmental management.

6. Several studies have used CNNs for soil classification, with promising results. For example, a study published in the Journal of Environmental Management used a CNN to classify soil samples based on their texture, achieving an accuracy of of more than 90%. Another study published in the Journal of Applied Remote Sensing used a CNN to classify soil types based on remote sensing data, achieving an accuracy of 80%.

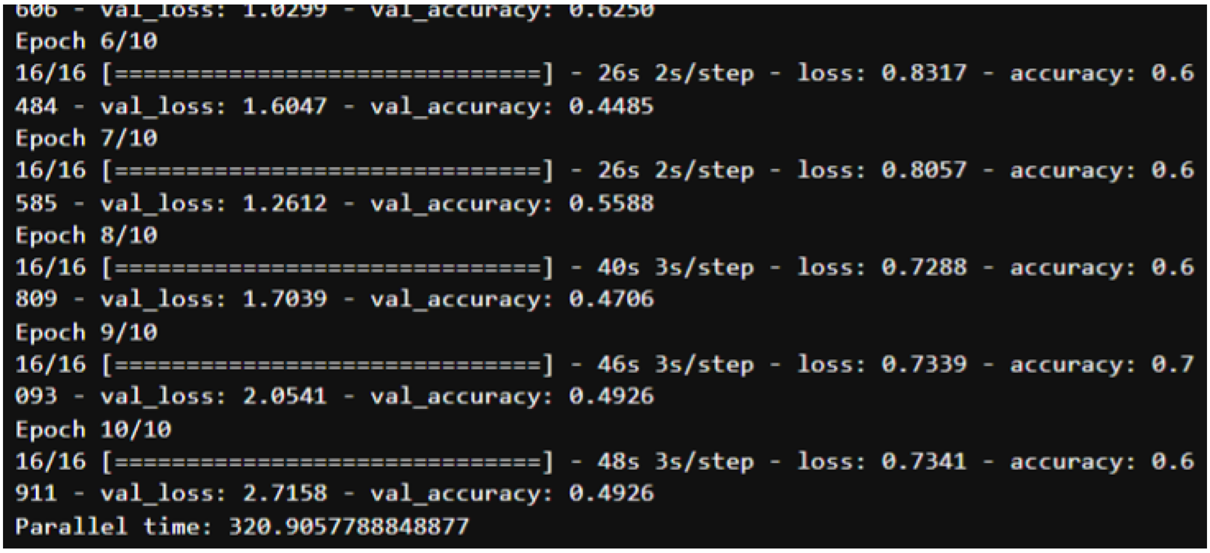
RESULTS

We tested the serial and parallelized CNN model on a dataset of 629 images, with 492 images for training and 136 images for testing. We trained the model for 10 epochs and obtained an accuracy of 64.63% on the serialized testing set and 69.11 on the parallelized testing set. We also measured the training time of the model with and without parallelization. The sequential model took 8 minutes and to train, while the parallelized model took only 5 minutes and 34 seconds. Parallelization was also done on the prediction process using ThreadPoolExecuter, where it was observed that the parallelized code took about 0.003 seconds to predict while the serialised model took about 3.5 seconds to make the prediction which is a speedup of about 3x. We also tested the model on new images from the testing directory and verified that the model can classify the soil types correctly.



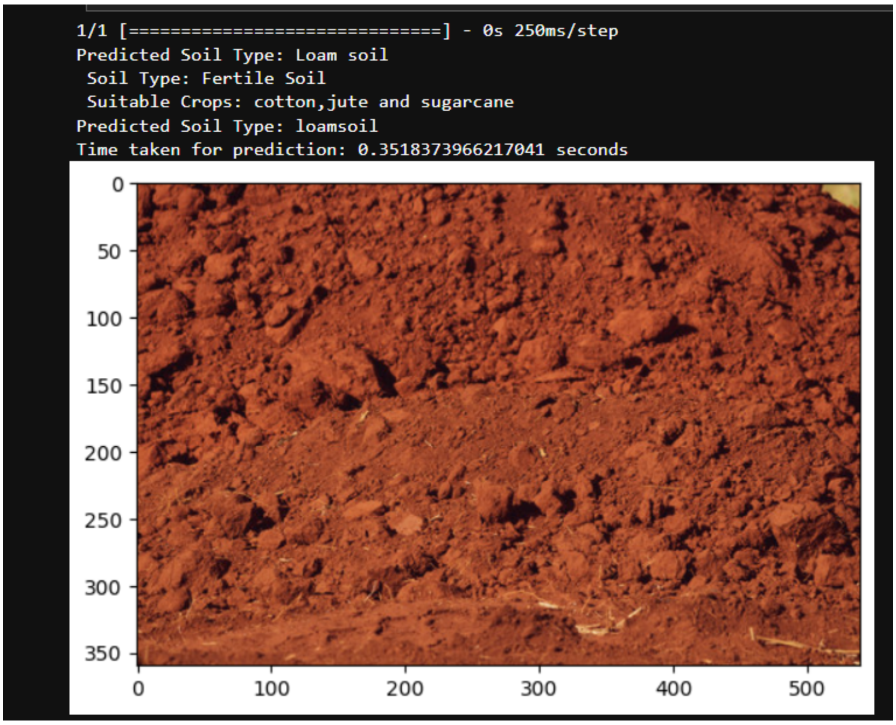
```
Epoch 6/10
16/16 [==============================] - 46s 3s/step - loss: 0.8777 - accuracy: 0.6
199 - val_loss: 1.6934 - val_accuracy: 0.5000
Epoch 7/10
16/16 [==============================] - 45s 3s/step - loss: 0.8527 - accuracy: 0.6
382 - val_loss: 1.4530 - val_accuracy: 0.4706
Epoch 8/10
16/16 [==============================] - 48s 3s/step - loss: 0.8521 - accuracy: 0.6
463 - val_loss: 2.0036 - val_accuracy: 0.3750
Epoch 9/10
16/16 [==============================] - 50s 3s/step - loss: 0.8347 - accuracy: 0.6
402 - val_loss: 1.4862 - val_accuracy: 0.4559
Epoch 10/10
16/16 [==============================] - 49s 3s/step - loss: 0.8252 - accuracy: 0.6
463 - val_loss: 1.8810 - val_accuracy: 0.4706
Total time taken to fit the model: 518.3494999408722 seconds
```

Fig 1. Training time for the serialized code



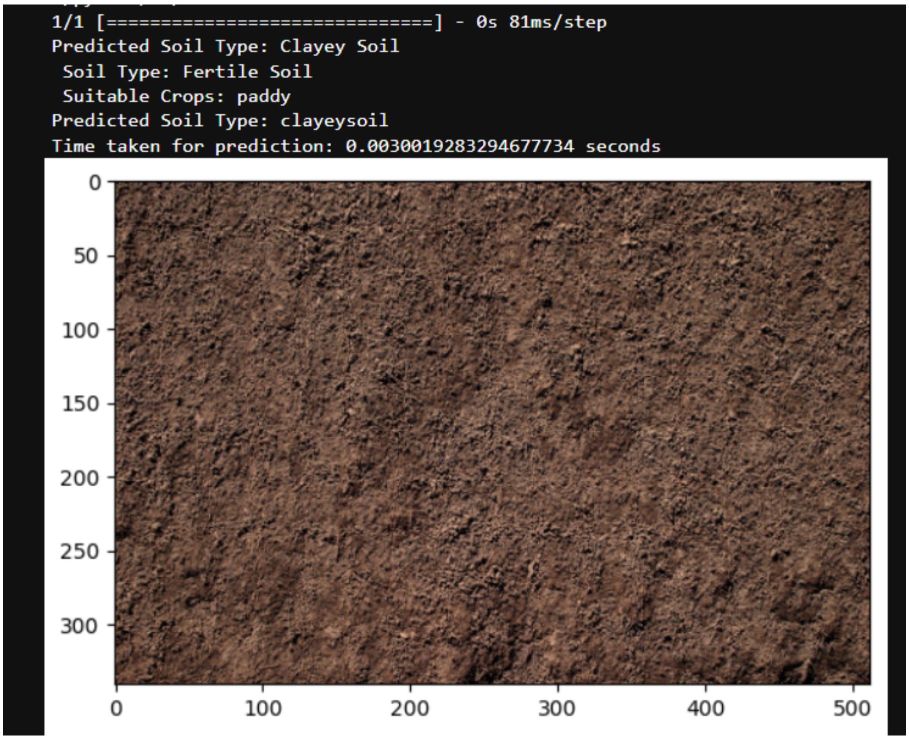
```
606 - val_loss: 1.0299 - val_accuracy: 0.6250
Epoch 6/10
16/16 [==============================] - 26s 2s/step - loss: 0.8317 - accuracy: 0.6
484 - val_loss: 1.6047 - val_accuracy: 0.4485
Epoch 7/10
16/16 [==============================] - 26s 2s/step - loss: 0.8057 - accuracy: 0.6
585 - val_loss: 1.2612 - val_accuracy: 0.5588
Epoch 8/10
16/16 [==============================] - 40s 3s/step - loss: 0.7288 - accuracy: 0.6
809 - val_loss: 1.7039 - val_accuracy: 0.4706
Epoch 9/10
16/16 [==============================] - 46s 3s/step - loss: 0.7339 - accuracy: 0.7
093 - val_loss: 2.0541 - val_accuracy: 0.4926
Epoch 10/10
16/16 [==============================] - 48s 3s/step - loss: 0.7341 - accuracy: 0.6
911 - val_loss: 2.7158 - val_accuracy: 0.4926
Parallel time: 320.9057788848877
```

Fig 2. Training time for the Parallelized code



```
1/1 [==============================] - 0s 250ms/step
Predicted Soil Type: Loam soil
 Soil Type: Fertile Soil
 Suitable Crops: cotton,jute and sugarcane
Predicted Soil Type: loamsoil
Time taken for prediction: 0.3518373966217041 seconds
```

Fig 3. Prediction time for Serialized code



```
1/1 [==============================] - 0s 81ms/step
Predicted Soil Type: Clayey Soil
 Soil Type: Fertile Soil
 Suitable Crops: paddy
Predicted Soil Type: clayeysoil
Time taken for prediction: 0.0030019283294677734 seconds
```

Fig 4. Prediction time for Parallelized code

CONCLUSIONS

In this project, we demonstrated the potential of using the Python multiprocessing module to parallelize a CNN model for soil classification. Our findings show that parallelization can significantly speed up the training process without sacrificing the model's accuracy, providing a valuable improvement in terms of computational efficiency.

As part of our future work, we aim to further analyse the impact of varying the number of threads and layer sizes on the performance of the parallelized model. By conducting these experiments, we hope to better understand the optimal configuration for maximizing both the accuracy and efficiency of the parallelized CNN model.In addition, we plan to explore other types of neural networks, such as recurrent neural networks (RNNs) and transformer-based models, to determine their suitability for soil classification tasks.

Moreover, we intend to examine different optimization algorithms, such as Adam, RMSprop, and Adagrad, to optimize the training process and improve model convergence. By testing various optimization techniques, we hope to further enhance the efficiency and robustness of our soil classification model.

Finally, we plan to extend our research to other domains within agriculture and environmental science, such as crop yield prediction and land-use classification, to explore the broader applicability of parallelized neural networks.

In conclusion, this project contributes to the ongoing development of efficient and accurate methods for soil classification, which hold great promise for advancing the fields of agriculture and environmental science. By exploring the benefits of parallelization and investigating alternative neural network architectures and optimization algorithms, we aim to further improve the performance and efficiency of soil classification models.

References

The below published paper talks about comparison of efficiency of Classification Techniques used in soil characterization. CNN has achieved over 95% accuracy.

- A comprehensive review on soil classification using deep learning and computer vision techniques Pallavi Srivastava1 & Aasheesh Shukla1 & Atul Bansal 1
https://www.researchgate.net/profile/Pallavi-Srivastava-2/publication/348913407_A_comprehensive_review_on_soil_classification_using_deep_learning_and_computer_vision_techniques/links/603e6d03a6fdcc9c780c5611/A-comprehensive-review-on-soil-classification-u

- We utilized resources such as Google, Stack Overflow, and Open AI for conducting our research.