Feature Learning-based Knowledge Distillation to train Teacher and Student Networks

Project report submitted in partial fulfillment
of the requirements for the degree of

Bachelor of Technology
in
Communication and Computer Engineering

by

Akshat Maithani - 21UCC014
Akshay Anand - 21UCC015
Patel Het Manojkumar - 21UCC125

Under Guidance of
Dr. Lal Upendra Pratap Singh

LNMIIT
The LNM Institute of
Information Technology

Department of Communication and Computer Engineering
The LNM Institute of Information Technology, Jaipur

November 2024

The LNM Institute of Information Technology
Jaipur, India

CERTIFICATE

This is to certify that the project entitled "Feature Learning-based Knowledge Distillation to train Teacher and Student Networks" , submitted by Akshat Maithani (21UCC014), Akshay Anand (21UCC015) and Patel Het Manojkumar (21UCC125) in partial fulfillment of the requirement of degree in Bachelor of Technology (B. Tech), is a bonafide record of work carried out by them at the Department of Communication And Computer Engineering, The LNM Institute of Information Technology, Jaipur, (Rajasthan) India, during the academic session 2024-2025 under my supervision and guidance and the same has not been submitted elsewhere for award of any other degree. In my/our opinion, this report is of standard required for the award of the degree of Bachelor of Technology (B. Tech).

Date	Adviser: Dr. Lal Upendra Pratap Singh

Acknowledgments

We would like to express our profound gratitude to our Bachelor's Thesis Project supervisor, **Dr. Lal Upendra Pratap Singh**, for his invaluable guidance and unwavering support throughout our research journey. His extensive expertise in machine learning, combined with his patient mentorship, has been instrumental in shaping this project. Dr. Lal Upendra Pratap Singh's insightful feedback during our regular review meetings and his encouragement during challenging phases of implementation were crucial to the success of our work.

We are particularly thankful for his detailed technical guidance in implementing the teacher-student model architecture and his suggestions regarding the nuclear norm regularisation approach. His commitment to academic excellence and attention to detail have significantly enhanced the quality of this research.

As a team, we extend our sincere appreciation to each other for the collaborative spirit, dedication, and diverse perspectives that enriched our project. The countless hours spent together debugging code, analyzing results, and preparing documentation have fostered both professional growth and lasting friendships.

Abstract

This study explores teacher-student model distillation techniques across four prominent remote sensing datasets: UC Merced Land Use, EuroSat, AID, and NWPU-RESISC45. We address the challenge of developing lightweight deep learning models for remote sensing image classification while maintaining accuracy. Our methodology implements a knowledge distillation framework where a four-layer convolutional neural network teacher model guides a compact three-layer student model, using distillation loss with variable alpha values and nuclear norm regularisation. Experimental results show that the student model achieves comparable or better classification performance compared to the teacher model while significantly reducing computational demands. The successful validation across diverse datasets—from aerial imagery to multi-spectral satellite data—demonstrates the approach's effectiveness for practical remote sensing applications in resource-constrained environments. This work reinforces the proliferation of efforts beyond models towards the minimisation of the hardware requirements. Such models can be deployed on low-resource hardware platforms such as drones or edge devices because of the decrease in the computational overhead. This nuclear norm regularisation improves the generalisation too, so the model can perform well over different distributions of data. Thus, the proposed framework turns out to be especially useful in important areas, where rapid decision making is the key factor, like environmental monitoring, precision agriculture, and disaster relief. This effort increases the global penetration of remote sensing technologies through the combination of efficiency and effectiveness.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The Area of Work

Teacher-student model distillation has become a potent method in machine learning for increasing model efficiency while preserving high-performance levels. This method efficiently transfers knowledge while minimizing computational demands by teaching a student model to learn from a teacher model.[1].

Our BTP aims to apply and assess teacher-student model distillation methodologies across four reputable remote sensing image datasets—UC Merced Land Use, EuroSat, AID, and NWPU-RESISC45.

Each dataset provides unique characteristics and challenges, making them suitable for a comprehensive study.[2]

1.1.1 UC Merced Land Use Dataset

This dataset[3] is used frequently in academic research because of its varied land use classifications and high-resolution aerial images. It provides an excellent platform for creating and testing image classification algorithms, with 21 classes and 100 images (256x256 pixels) per class.

1.1.2 EuroSat Dataset

Exploring model distillation in scenarios involving both RGB and multi-spectral imagery is made possible by the EuroSat dataset[4]. This dataset comes from Sentinel-2 satellite imagery and is ideal for experiments utilizing tiny models because it comprises 64x64 pixel pictures.

1.1.3 AID Dataset

10,000 aerial photos divided into 30 different scene groups comprise the extensive Aerial Image Database (AID) dataset[5]. The dataset is perfect for evaluating the generalizability of teacher-student models in aerial scene classification tasks because of its variety and high-quality annotations.

1.1.4 NWPU-RESISC45 Dataset

NWPU-RESISC45[6] has 31,500 photos of 45 different scene types with different spatial resolutions, making it one of the most extensive datasets for remote sensing image scene categorization. For teacher-student model distillation, the quantity and diversity of the dataset offer both opportunities and difficulties.

1.1.5 Objectives and Research Goals

- **Knowledge Transfer:** Examine how well teacher-student model distillation transfers information from intricate and uncomplicated models.
- **Model Performance:** Compare the distilled models' accuracy, computational efficiency, and memory footprint to their instructor counterparts.
- **Dataset Analysis:** Analyze how distilled models behave on datasets with different scene variety, picture resolutions, and spectral characteristics.
- **Use in Remote Sensing:** Evaluate how well these methods work in practical applications like disaster relief, urban planning, and environmental monitoring.

1.2 Problem Addressed

Remote sensing image classification has several challenges, including large model sizes, powerful processing, and effective resource management to handle high-resolution data and diverse scene types. Even though state-of-the-art deep learning models are incredibly accurate, they often require much computational power, which limits their application in resource-constrained scenarios like embedded systems or edge devices. This becomes very problematic when dealing with large datasets, such as those used in remote sensing, where data volatility and high-resolution photos add to the complexity.

The specific problem this Bachelor Thesis Project (BTP) aims to address is developing and evaluating teacher-student model distillation procedures to generate lightweight and valuable machine learning models without noticeable reductions in classification performance.

1.2.1 Key Challenges Addressed

- **High Computational Cost:** Reducing the computational overhead while maintaining the accuracy of models.
- **Scalability and Deployment:** Ensuring that the distilled models can be deployed efficiently in resource-constrained environments.
- **Generalisation of the Model:** Creating models that generalize well across diverse datasets with varying characteristics.
- **Data Diversity:** Handling the wide variety of scene types, resolutions, and spectral properties in remote sensing data.

1.3 Existing Systems

Deep learning approaches have made significant progress in classifying remote-sensing photos. These systems, primarily based on convolutional neural networks (CNNs) and their variants, have outperformed high-resolution aerial and satellite imaging expectations.

1.3.1 Convolutional Neural Networks (CNNs)

Traditional CNN designs such as VGGNet, ResNet, and Inception have been widely used for remote sensing image categorization problems. These models have successfully identified complicated patterns and extracted spatial properties from high-resolution photos. Nevertheless, they are less feasible for deployment on devices with constraints because of their high processing requirements.[7]

1.3.2 Deep Learning-Based Scene Classification

The UC Merced Land Use and AID datasets are two examples of remote sensing datasets for which deep learning-based scene classification tasks are modified using systems like AlexNet and DenseNet. Even while these models have better accuracy and feature extraction capabilities, they are frequently huge and demand a lot of processing power and training time. These models' performance on high-resolution data has established a standard for accuracy at the price of efficiency.[8]

1.3.3 Transfer Learning with Pretrained Models

Using pre-trained models, such as those from the ImageNet dataset, in conjunction with transfer learning, has become a common strategy to reduce training time and computing load. Researchers can improve classification performance by applying pre-existing knowledge to refine models such as ResNet and MobileNet on remote sensing data. The size and flexibility of these pre-trained models to different spectral features in remote sensing data still need to be improved, notwithstanding their effectiveness.[9]

1.3.4 Multi-Spectral and Multi-Modal Approaches

Models are trained on RGB and other spectral bands, including near-infrared, in multi-spectral and multi-modal approaches, which existing systems have also investigated. For example, algorithms trained on the EuroSat dataset used additional spectral information to improve classification accuracy. However, more powerful, resource-intensive models are frequently required as data complexity increases.

1.3.5 Lightweight and Efficient Architectures

More compact architectures, such as MobileNetV2, SqueezeNet, and EfficientNet, have been investigated recently to strike a compromise between efficiency and performance. Depthwise separable convolutions are used in these systems to minimize computational effort while maintaining respectable performance. Even if these models are a step in the right direction, when used on detailed datasets like NWPU-RESISC45, their performance could still catch up to larger, more complicated networks.

1.3.6 The Gap in Existing Systems

Traditional deep learning models have made great strides, but there is still a need for models that compromise high performance and resource efficiency. A potential solution to this problem is to use teacher-student model distillation, which provides a realistic method of producing more manageable, deployable models while maintaining good classification performance.

This study will help to fill this gap by investigating the efficacy and limitations of knowledge distillation across four sizeable remote sensing datasets.

Chapter 2

Literature Review

Information distillation has emerged as an essential technique in deep learning, allowing information transfer from a larger, pre-trained teacher model to a smaller student model, resulting in comparable performance but dramatically decreasing computational complexity. Hinton et al. (2015) initially presented the concept of distillation loss, which uses a combination of soft-target predictions from the teacher model and complex labels from the dataset to train the student model. This strategy has been widely used in various fields, including computer vision and natural language processing, to improve the efficiency of deep learning models.

2.1 Overview of Previous Research

Early investigations in knowledge distillation focused on traditional teacher-student designs, with the teacher's softened outputs serving as informative gradients for training the student model. These approaches performed well in picture classification, object detection, and sequence modelling tests. Recent research has expanded the framework to encompass changes to architecture, loss function, and data augmentation procedures. For example, attention-based distillation methods improve student learning by transferring intermediate feature maps or attention processes from the teacher to the student. Other approaches, such as cross-modal distillation, have investigated knowledge transfer between modalities such as audio and text.[10]

In remote sensing, the datasets included in our study—UC Merced, NWPU-RESISC45, AID, and EuroSAT—are essential standards for land-use and land-cover categorisation tasks. Yang and Newsam (2010) introduced the UC Merced collection, which includes 21 classes of high-resolution aerial photos. Similarly, Cheng et al. (2017) offered NWPU-RESISC45, and Xia et al. (2017) produced AID, which provides different datasets with considerable intra-class variability, making them perfect for evaluating model generalisation. EuroSAT, which is based on Sentinel-2 satellite imagery (Helber et al., 2019), adds a new dimension by focusing on multispectral data, offering novel challenges for knowledge transfer.

2.2 Comparison and Contrast of Current Trends

Current advancements in knowledge distillation indicate a shift towards using advanced strategies to bridge the performance gap between instructor and student models. Attention-based approaches and feature-matching procedures are gaining popularity due to their capacity to convey more information. However, these methodologies frequently necessitate additional processing costs, which goes against the fundamental purpose of knowledge distillation—efficiency. In contrast, techniques such as nuclear norm regularisation strike a balance by improving student model accuracy while not significantly increasing complexity. While these strategies have been investigated in broader areas, their use with remote sensing datasets is under-represented in the literature.

2.3 Conclusion

Despite significant advances, knowledge distillation in remote sensing still needs to be explored. The problems presented by datasets such as UC Merced, NWPU-RESISC45, AID, and EuroSAT necessitate bespoke approaches that optimise model architecture and training strategies. Addressing gaps in hyperparameter tuning, architectural design, and regularisation will help realise the full potential of knowledge distillation in these fields.

This study intends to contribute to the expanding corpus of research on efficient, high-performance deep learning models for remote sensing by combining insights from previous studies and incorporating unique techniques such as nuclear norm regularisation.

TABLE 2.1: Summary of Remote Sensing Datasets Used in Knowledge Distillation.

Dataset	Key Features	Challenges
UC Merced	21 classes, aerial images	Intra-class variability
NWPU-RESISC45	45 scene types, high diversity	Large size, computational cost
AID	30 scene types, high-quality annotations	Model generalisation
EuroSAT	Multi-spectral data, Sentinel-2 images	Spectral diversity

Chapter 3

Proposed Work

3.1 Proposed Methodology

In our study, we used knowledge distillation to train a lightweight student model under the guidance of a more complicated instructor model. The instructor model, which consists of four convolutional layers, was initially trained on four benchmark datasets: UC Merced, NWPU-RESISC45, AID, and EuroSAT. The student model, which consisted of three convolutional layers, was then trained using distillation loss, which blends the teacher model's predictions with the actual dataset labels. We investigated the impact of soft-label contributions on the performance of the student model by varying the alpha values in the distillation loss function.

Our findings showed that the student model attained accuracy comparable to, and in some cases exceeding, the teacher model, demonstrating the efficacy of knowledge distillation in lowering model size without significantly compromising performance. To increase accuracy even more, we employed nuclear norm regularisation, which encourages compact and discriminative feature representations by penalising the singular values of intermediate feature maps. This change significantly improved the performance of our student model on the previously specified datasets.

Our research contributes to developing efficient and accurate deep-learning models for remote sensing applications by thoroughly assessing the current literature and iterative testing. By proving the efficacy of knowledge distillation and nuclear norm regularisation, we show that these strategies can deploy models in resource-constrained situations while maintaining high classification accuracy.

Chapter 4

Simulation and Results

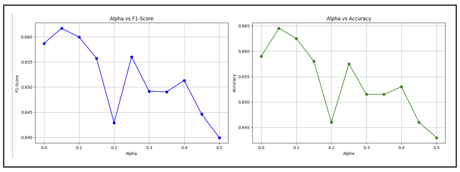
The data illustrates how the hyperparameter α influences student model performance. The plot demonstrates that α and $(1 - \alpha)$ significantly increase the model's effectiveness. Specific α values lead to higher F1 scores and accuracy, whereas others are lower. This demonstrates the balanced role of α and $(1 - \alpha)$ in determining the student model's success.

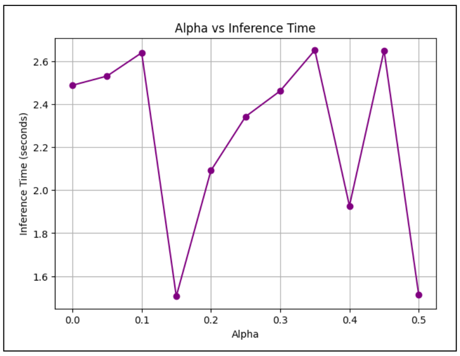
Additional analysis includes data gathered using the nuclear norm. The nuclear norm of a matrix, defined as the sum of its singular values (or absolute eigenvalues for symmetric matrices), provides various advantages in model evaluation and regularisation:

- Encourages Simplicity: Focuses on the most relevant data aspects, resulting in more straightforward representations.
- Noise Reduction: Removes unnecessary characteristics or noise to highlight meaningful patterns.
- Improves Generalisation: Regularising the model's weights reduces overfitting and increases its ability to handle previously unknown data.

The accuracy range of the models can be improved by including the nuclear norm. The teacher model's accuracy ranged from 59% to 83%. In comparison, the student model achieved an accuracy ranging from 60% to 86%. Notably, in many cases, the accuracy of the student model matched or exceeded that of the teacher model. Including nuclear norm results improves the validity of these ranges, as seen in the accompanying charts.

These findings show that the student model can achieve performance levels that are on par with or better than the instructor model while using the nuclear norm's simplicity, noise reduction, and generalisation advantages.

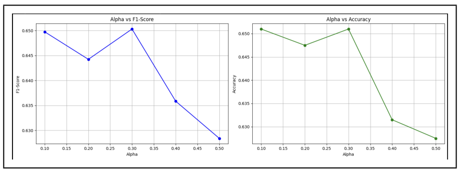


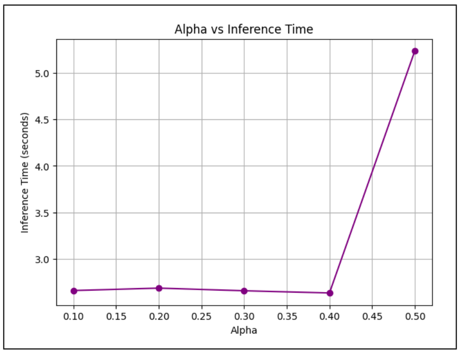


α Value	Accuracy	F1 Score
0.00	0.66	0.66
0.05	0.66	0.66
0.10	0.66	0.66
0.15	0.66	0.66
0.20	0.65	0.64
0.25	0.66	0.66
0.30	0.65	0.65
0.35	0.65	0.65
0.40	0.65	0.65
0.45	0.65	0.64
0.50	0.64	0.64

Model	Precision	Recall	F1 Score	Inference Time (s)
Teacher	0.67	0.65	0.64	10.48
Student	0.66	0.66	0.66	2.63

(a) Comparative analysis of performance metrics across varying α values (n=11)



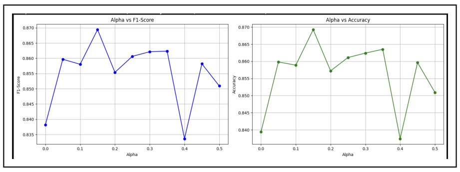


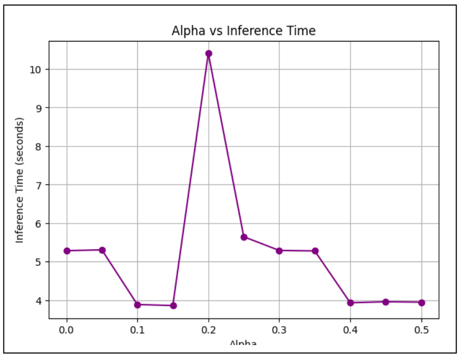
α Value	Accuracy	F1 Score
0.10	0.65	0.65
0.20	0.65	0.64
0.30	0.65	0.65
0.40	0.63	0.64
0.50	0.63	0.63

Model	Precision	Recall	F1 Score	Inference Time (s)
Teacher	0.63	0.62	0.62	10.37
Student	0.65	0.65	0.64	3.46

(b) Comparative analysis of performance metrics across varying α values (n=5)

FIGURE 4.1: AID Dataset Results

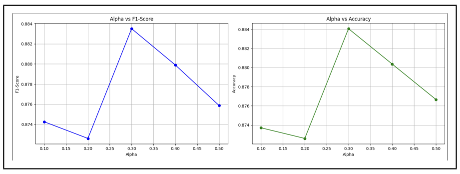


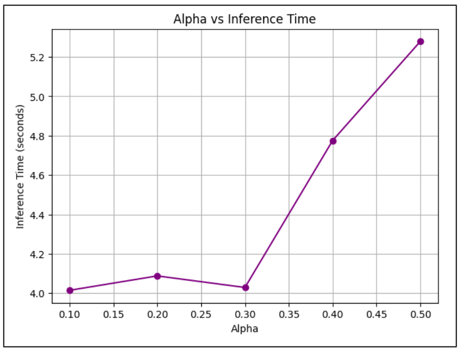


α Value	Accuracy	F1 Score
0.00	0.84	0.84
0.05	0.86	0.86
0.10	0.86	0.86
0.15	0.87	0.87
0.20	0.86	0.86
0.25	0.86	0.86
0.30	0.86	0.86
0.35	0.86	0.86
0.40	0.84	0.83
0.45	0.86	0.86
0.50	0.85	0.85

Model	Precision	Recall	F1 Score	Inference Time (s)
Teacher	0.82	0.80	0.79	41.25
Student	0.86	0.86	0.86	3.89

(a) Comparative analysis of performance metrics across varying α values (n=11)



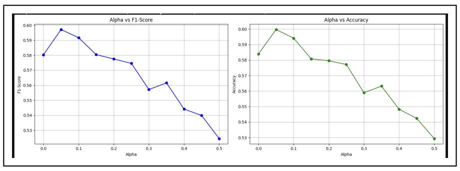


α Value	Accuracy	F1 Score
0.10	0.87	0.87
0.20	0.87	0.87
0.30	0.88	0.88
0.40	0.88	0.88
0.50	0.88	0.88

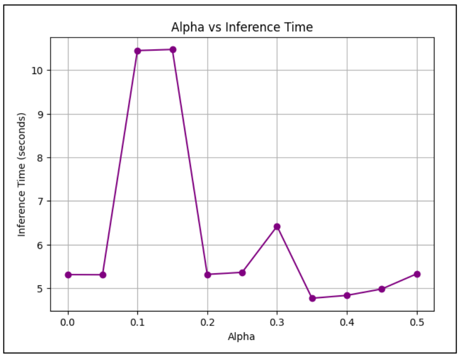
Model	Precision	Recall	F1 Score	Inference Time (s)
Teacher	0.83	0.83	0.83	41.13
Student	0.86	0.86	0.86	5.32

(b) Comparative analysis of performance metrics across varying α values (n=5)

FIGURE 4.2: EUROSAT Dataset Results

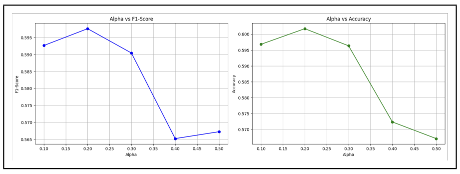


α Value	Accuracy	F1 Score
0.00	0.58	0.58
0.05	0.60	0.60
0.10	0.59	0.59
0.15	0.58	0.58
0.20	0.58	0.58
0.25	0.58	0.57
0.30	0.56	0.56
0.35	0.56	0.56
0.40	0.55	0.54
0.45	0.54	0.54
0.50	0.53	0.52

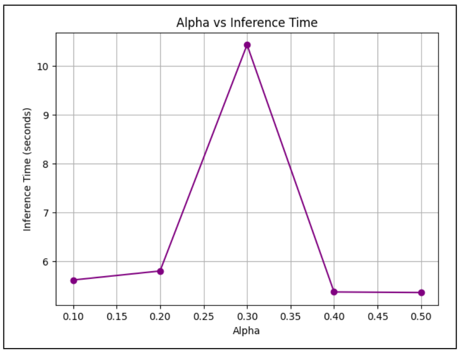


Model	Precision	Recall	F1 Score	Inference Time (s)
Teacher	0.62	0.58	0.58	17.32
Student	0.60	0.59	0.59	10.44

(a) Comparative analysis of performance metrics across varying α values (n=11)



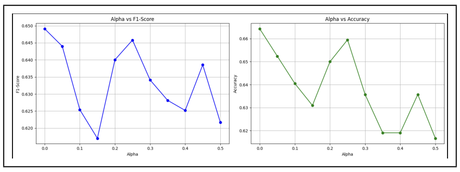
α Value	Accuracy	F1 Score
0.10	0.60	0.59
0.20	0.60	0.60
0.30	0.60	0.59
0.40	0.57	0.57
0.50	0.57	0.57

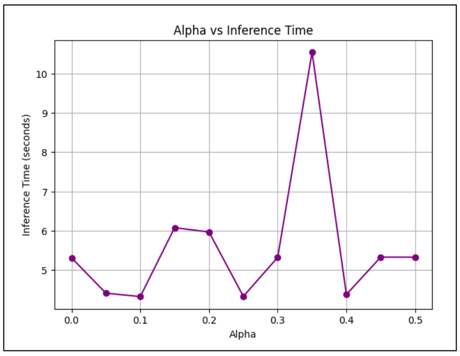


Model	Precision	Recall	F1 Score	Inference Time (s)
Teacher	0.63	0.60	0.60	16.57
Student	0.56	0.55	0.55	5.11

(b) Comparative analysis of performance metrics across varying α values (n=5)

FIGURE 4.3: NWPU Dataset Results

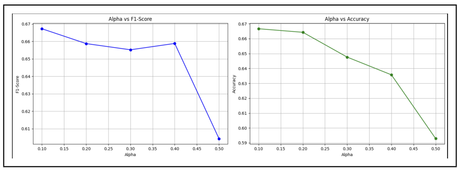


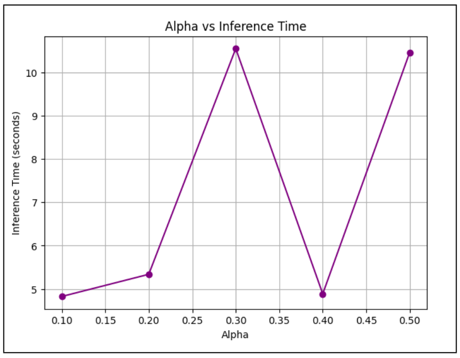


α Value	Accuracy	F1 Score
0.00	0.66	0.65
0.05	0.65	0.65
0.10	0.64	0.63
0.15	0.63	0.62
0.20	0.65	0.64
0.25	0.66	0.65
0.30	0.64	0.63
0.35	0.62	0.63
0.40	0.62	0.63
0.45	0.64	0.65
0.50	0.62	0.62

Model	Precision	Recall	F1 Score	Inference Time (s)
Teacher	0.65	0.64	0.63	41.13
Student	0.65	0.65	0.64	4.26

(a) Comparative analysis of performance metrics across varying α values (n=11)





α Value	Accuracy	F1 Score
0.10	0.67	0.67
0.20	0.66	0.66
0.30	0.65	0.66
0.40	0.64	0.66
0.50	0.59	0.60

Model	Precision	Recall	F1 Score	Inference Time (s)
Teacher	0.68	0.65	0.65	20.66
Student	0.71	0.69	0.69	5.81

(b) Comparative analysis of performance metrics across varying α values (n=5)

FIGURE 4.4: UC-MERCED Dataset Results

Chapter 5

Conclusions and Future Work

5.1 Conclusions

Our study on the influence of hyperparameters, specifically , on teacher-student dynamics has generated encouraging results. The effective use of nuclear norm regularisation to improve model simplicity and generalisation offers up numerous promising options for further investigation.

5.2 Future Work

Several promising directions for future research have been identified:

5.2.1 Expanding Dataset Scope

To enhance our findings, we need to test these methodologies on a more significant number of datasets. This extension will allow us to assess better how well the student model performs under various scenarios and if nuclear norm regularisation remains successful across diverse data patterns.

5.2.2 Investigating Alternative Regularisation Methods

While our work with the nuclear norm has shown promise, we should look into alternative techniques, such as the Frobenius norm or strategic dropout, to see whether they give any further benefits for model performance and adaptability.

5.2.3 Architecture Refinement

We see possibilities in fine-tuning the student model's structure to achieve the best computing efficiency and model intricacy mix.

5.2.4 Industrial Applications

The ultimate measure of our approach is its practicality. To address specific industrial difficulties, we want to use these technologies in various domains, including natural language processing, computer vision, and recommendation engines.

Bibliography

[1] F. A. A. Alkhulaifi and I. Ahmad, "Knowledge distillation in deep learning and its applications," *PeerJ Comput. Sci.*, vol. 7, p. e474, 2021.

[2] X. W. R. W. H. C. L. Wang, Y. Chen and Y. Zhu, "Research on remote sensing image classification based on transfer learning and data augmentation," in *Int. Conf. Knowl. Sci., Eng. Manage.*, pp. 99–111, Springer, 2023.

[3] H. S. J. Liang, J. Xu and L. Fang, "Land-use classification via constrained extreme learning classifier based on cascaded deep convolutional neural networks," *Eur. J. Remote Sens.*, vol. 53, no. 1, pp. 219–232, 2020.

[4] K. T. H. Yassine and M. Jaber, "Improving lulc classification from satellite imagery using deep learning–eurosat dataset," *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, vol. 43, pp. 369–376, 2021.

[5] İnik, "Classification of scenes in aerial images with deep learning models," *Turk. Dog. Fen Derg.*, vol. 12, no. 1, pp. 37–43, 2023.

[6] W. N. X. P. H. Z. X. Yang, W. Yan and M. Zhang, "Object-guided remote sensing image scene classification based on joint use of deep-learning classifier and detector," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2673–2684, 2020.

[7] F. H. X. Liu, K. H. Ghazali and I. I. Mohamed, "Review of cnn in aerial image processing," *Imaging Sci. J.*, vol. 71, no. 1, pp. 1–13, 2023.

[8] T. K. R. Naushad and E. Ghaderpour, "Deep transfer learning for land use and land cover classification: A comparative study," *Sensors*, vol. 21, no. 23, p. 8083, 2021.

[9] A. Alem and S. Kumar, "Transfer learning models for land cover and land use classification in remote sensing image," *Appl. Artif. Intell.*, vol. 36, no. 1, p. 2014192, 2022.

[10] S. J. M. J. Gou, B. Yu and D. Tao, "Knowledge distillation: A survey," *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1789–1819, 2021.