EFFICIENT LAND-COVER IMAGE CLASSIFICATION VIA MIXED BIT-PRECISION QUANTIZATION

Tushar Shinde & Ahmed Silima Vuai
Indian Institute of Technology Madras Zanzibar shinde@iitmz.ac.in

ABSTRACT

Land cover (LC) image classification is essential for monitoring environmental changes, urban planning, and disaster management. Deep neural networks (DNNs) have achieved remarkable success in LC classification; however, their deployment on edge devices is constrained by high computational and storage requirements. Quantizing neural networks to reduce model size has proven effective in achieving low bit-width representations of parameters while maintaining the original network's performance. To facilitate deployment on edge devices, we propose a novel adaptive quantization technique that optimally reduces model size while preserving accuracy. This method evaluates layer importance through statistical measures, enabling the adaptive selection of bit-width precision for each layer. Experimental results show that the proposed quantization strategy effectively balances compression and accuracy for different DNN architectures like VGG19, ResNet18, and ResNet50, providing a practical solution for LC classification on EuroSAT dataset in resource-constrained environments.

1 INTRODUCTION

LULC classification is crucial for applications like environmental monitoring and urban planning Koenig & Gueguen (2016). Deep Neural Networks (DNNs), particularly CNNs Mäyrä et al. (2021); Karra et al. (2021), have significantly improved remote sensing analysis. Recent advancements, including Autoencoders Zhang et al. (2017), multitask deep learning Benhammou et al. (2022), Generative Models Hong et al. (2024), and Transformers Khan et al. (2024), have further enhanced performance of LULC classification tasks.

Despite the remarkable success of DNNs in LULC classification LeCun et al. (2015), their inherent complexity poses significant challenges for deployment on resource-limited edge devices. These models often consist of numerous parameters, making them memory-intensive and computationally demanding Li et al. (2023). Various techniques have been proposed to address the deployment challenges of DNNs Mishra et al. (2020), such as model pruning Li et al. (2016), low-rank factorization Yin et al. (2021), knowledge distillation Hinton et al. (2015), and quantization Kim et al. (2020); Deng et al. (2020). Among these, quantization has proven to be an effective method, as it reduces the bit precision of model weights (and, optionally, activations) from the conventional 32-bit floating-point representation to lower bit-width formats, thus enhancing storage and memory efficiency and making quantized models better suited for scenarios requiring rapid inference Rokh et al. (2023).

However, most existing quantization techniques adopt a uniform bit-width precision across all network layers. While this approach can reduce model size and inference time, it often leads to a notable drop in accuracy, particularly with complex models Menghani (2023). This issue arises because different layers contribute differently to the overall performance of the model Yang et al. (2023). Although mixed-precision quantization techniques Koryakovskiy et al. (2023) have been explored, identifying the optimal bit precision for each layer remains a challenge, often requiring computationally intensive optimization processes Chen et al. (2021); Tang et al. (2022).

To address these challenges, we introduce an adaptive layer-wise quantization framework that assigns bit-widths based on layer importance. Experiments on VGG19, ResNet18, and ResNet50 using EuroSAT Helber et al. (2019) demonstrate its efficiency for edge deployment. The main contributions of this paper include an efficient layer importance computation technique based on

statistical parameters and a novel adaptive layer-wise quantization approach that dynamically assigns bit-widths to layers according to their importance. Additionally, an iterative search algorithm is proposed to refine bit precision for each layer, ensuring an optimal balance between model size and performance.

The remainder of this paper is organized as follows: Section 2 elaborates on the proposed method. Section 3 presents experimental results and analysis. Lastly, Section 4 concludes the paper and discusses potential future work.

2 METHOD

This section elaborates on our adaptive layer-wise quantization approach designed for deep neural networks. The proposed framework encompasses several critical stages aimed at optimizing performance while minimizing model size. As illustrated in Fig. 1, our novel framework enhances land-cover image classification through adaptive quantization. The process begins with the training of deep neural network architectures. Subsequently, we implement adaptive quantization guided by layer importance, which optimizes the precision of weights to achieve model compression without compromising accuracy. This process involves (i) computing layer importance, (ii) performing iterative search optimization for bit precision, and (iii) applying layer-wise quantization. The outcome is a compressed quantized model that significantly reduces both size and inference time while sustaining high classification performance. The final stage involves categorizing the test images into predefined classes. This layer importance-guided adaptive quantization framework presents an effective solution for land-cover image classification, particularly in resource-constrained environments and real-world applications.

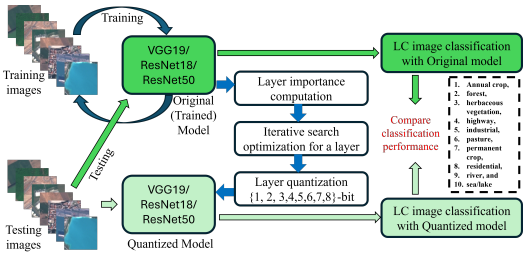

Figure 1: The framework of the proposed approach.

The motivation for our method arises from the observation that different layers exert varying levels of influence on the final accuracy of the model Elkerdawy et al. (2020). To exploit the unique characteristics of each layer, we employ different bit-precisions for quantization. However, determining the order in which layers should be quantized is a challenging task. To address this, we introduce a ranking scheme based on layer-wise importance.

2.1 LAYER IMPORTANCE COMPUTATION SCHEME

We introduce a method to compute the importance of each layer in the network. To guide the compression process, we extend the work of Shinde (2024) by presenting the *Layer Importance Computation* scheme based on following parameters.

Normalized Parameters Proportion: The number of parameters in each layer plays a critical role in the overall model size. For layers with a high parameter count, our algorithm should aim for maximum quantization to reduce model size while maintaining accuracy. Therefore, we define the normalized parameter proportion for a layer l as:

$$N_P(l) = \frac{\text{Number of parameters in layer } l}{\text{Total number of parameters in the model}} \tag{1}$$

Normalized Variance: The distribution of parameters also plays a role in determining bit precision, as layers exhibit different distributions and variances. The variance of parameters provides insight into the distribution's compactness, and leveraging advanced quantization techniques can improve compression performance. The normalized variance is computed as:

$$N_V(l) = \log\left(e - 1 + \frac{\text{Variance of parameters in layer } l}{\max_k\left(\text{Variance of parameters in layer } k\right)}\right) \tag{2}$$

Layer Importance Calculation: We compute the overall layer importance by combining these three components with corresponding weights:

$$Importance(l) = w_P \cdot N_P(l) + w_V \cdot (1 - N_V(l)) \tag{3}$$

where $Importance(l)$ denotes the importance of layer l. Here, w_P, and w_V are weights for the normalized parameter proportion and variance, respectively.

2.2 Layer-wise bit-width Selection Algorithm

The optimal bit-precision for each layer is determined through a search process, ensuring the final model achieves high accuracy with minimal average bit-width. The layer-wise importance helps in identifying which layers are crucial and ranks them based on their $Importance(l)$. The layer with the highest importance is chosen to be quantized first to preserve its significant features. The bit-width search for the chosen layer is carried out sequentially. To this end, we introduce a threshold margin $T_{margin}(l)$, which ensures the chosen bit-width $b(l)$ maintains the quantized model's performance within $T_{margin}(l)$ of the 8-bit quantized model's accuracy. Furthermore, an adaptive $T_{margin}(l)$ strategy is proposed, updating $T_{margin}(l)$ based on the layer's importance $Importance(l)$:

$$T_{margin}(l) = T_{margin}^{init} \cdot Importance(l) \tag{4}$$

where T_{margin}^{init} is empirically determined for the DNN. This adaptive approach ensures optimal bit-width selection, effectively compressing the model while minimizing performance loss.

3 Experimental Results

Dataset: The EuroSAT dataset Helber et al. (2019) is widely used for land use and land cover (LULC) classification in remote sensing. It consists of 27,000 labeled Sentinel-2 images with a 10m spatial resolution, covering ten LULC classes: annual crop, forest, herbaceous vegetation, highway, industrial, pasture, permanent crop, residential, river, and sea/lake. Each 64×64 pixel image contains 13 spectral bands spanning visible, near-infrared, and shortwave infrared wavelengths. With 2000-3000 images per class, EuroSAT serves as a benchmark for deep learning-based multi-spectral classification. Data augmentation, including random flips, affine transformations, and resizing, was applied to enhance training diversity.

Configuration: The experiments were conducted on the Kaggle platform, leveraging NVIDIA Tesla P100 and G4 GPUs, which accelerated the training of our neural networks. Python 3 was used with essential libraries such as PyTorch, NumPy, pandas, matplotlib, and scikit-learn.

Model Training and Evaluation Protocol: We assessed our approach using popular DNN architectures, including VGG19 Simonyan & Zisserman (2014), ResNet18 He et al. (2016), and ResNet50 He et al. (2016), training these models from scratch without pre-trained weights. Training was performed over 120 epochs with a batch size of 16. The Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9 and a weight decay of 1×10^{-6} was used in combination with the cross-entropy loss function, and an initial learning rate of 0.001 was applied. A learning rate scheduler, *ReduceLROnPlateau*, adaptively reduced the learning rate by a factor of 0.1 upon plateauing validation loss. An 80-20 stratified train-test split was applied to validate model performance. Our focus was post-training quantization rather than quantization-aware training for streamlined deployment.

Hyper-parameter Settings: The weights w_P and w_V used for computing layer importance were set equally, totaling 1. The threshold margin T_{margin} was initially set to 0.5%. Recognizing the significant impact of the first and last layers on overall performance Liu et al. (2021), T_{margin} was adjusted to half of the predefined threshold. We concentrated on weight quantization, using 8-bit quantization as the baseline bit-precision.

Evaluation Metrics: We employed several metrics such as accuracy and average bit-width to evaluate the effectiveness of model compression and quantization techniques. The average bit-width indicates the overall bit-precision used across all layers of the model. It reflects the extent of model compression by showing how bit-width varies from one layer to another. For a model with L layers, the average bit-width $\bar{b}$ can be expressed as:

$$\bar{b} = \sum_{l=1}^{L} N_P(l) \cdot b(l) \tag{5}$$

where $N_P(l)$ represents the normalized parameter proportion for layer l, and $b(l)$ denotes the bit-width of the parameters in layer l.

Dataset	VGG19	ResNet18	ResNet50
Original (32-bit)	96.06%	94.78%	94.61%
Fixed (8-bit) Q	96.04%	94.70%	94.70%
Fixed (7-bit) Q	96.07%	94.67%	93.24%
Fixed (6-bit) Q	96.13%	94.63%	93.81%
Fixed (5-bit) Q	95.96%	94.43%	93.83%
Fixed (4-bit) Q	95.52%	90.09%	87.50%
Fixed (3-bit) Q	90.41%	63.93%	29.28%
Fixed (2-bit) Q	9.30%	9.41%	11.11%
Fixed (1-bit) Q	11.11%	9.26%	11.11%
Proposed Adaptive Quantization			
Fixed $T_{margin} = 0.5$	95.78%	93.83%	93.74%
(Average bit-width)	(**1.38**)	(3.73)	(3.74)
Adaptive $T_{margin} = 0.5$	**96.06**%	**95.07**%	**94.44**%
(Average bit-width)	(1.44)	(**3.43**)	(**3.70**)

Table 1: Model accuracy and Average bit-width comparison for different DNN architectures at fixed and adaptive quantization.

Model	Accuracy	Parameters Size
ResNet50 Helber et al. (2019)	96.43%	24M × 32
ResNet50 Khan et al. (2024)	96.57%	24M × 32
ResNet101 Khan et al. (2024)	97.22%	43M × 32
AlexNet Rangel et al. (2024)	83.7%	61M × 32
DenseNet Rangel et al. (2024)	92.5%	29M × 32
MobileNetV3 Rangel et al. (2024)	87.7%	6M × 32
EfficientNetV2 Rangel et al. (2024)	90.8%	24M × 32
ViT32 Rangel et al. (2024)	91.7%	86M × 32
SwinB Rangel et al. (2024)	92.6%	88M × 32
Proposed AQ VGG19	**96.06%**	144M × **1.44**
Proposed AQ ResNet50	**94.44%**	24M × **3.70**
Proposed AQ ResNet18	**95.07%**	12M × **3.43**

Table 2: Comparison of the proposed approach with the existing studies (without pre-trained weights). The parameters size is shown (in bits) as multiplication of number of parameters and the average bit-width of each parameter.

Results and Analysis: Table 1 presents a comparison of model accuracy and average bit-widths across different quantization methods for each architecture. The original full-precision (32-bit) models achieved accuracies of 96.06%, 94.78%, and 94.61% for VGG19, ResNet18, and ResNet50, respectively. As expected, accuracy decreased with reduced bit-width in fixed-precision quantization, with significant drops below 4 bits. The adaptive quantization method exhibited a clear advantage over fixed quantization, successfully balancing model accuracy and bit-width. For instance, with a fixed $T_{margin} = 0.5\%$, VGG19 achieved 95.78% accuracy with an average bit-width of 1.38, ResNet18 reached 93.83% with a bit-width of 3.73, and ResNet50 obtained 93.74% with a bit-width of 3.74. Employing the adaptive T_{margin} led to improved accuracy (96.06%, 95.07%, and 94.44%) while maintaining comparable bit-widths (1.44, 3.43, and 3.70, respectively). Hence, the adaptive quantization strategy successfully reduced model size while preserving high classification performance across various DNN architectures. The outcomes highlight the effectiveness of the adaptive quantization approach, demonstrating that it outperforms fixed-order strategies in achieving a balance between accuracy and bit-width.

Table 2 presents a comparison of the proposed adaptive quantization (AQ) approach against established models (without pre-trained weights) evaluated on the EuroSAT dataset. The results indicate that the proposed AQ method for VGG19, ResNet50, and ResNet18 achieved accuracies of 96.06%, 94.44%, and 95.07%, respectively. Notably, these models exhibit significantly lower average bit-widths compared to traditional models. In contrast, established models such as ResNet50, AlexNet, DenseNet, MobileNetV3, EfficientNetV2, ViT32, and SwinB generally require larger average bit-widths and often demonstrate lower accuracy rates. For instance, while models like MobileNetV3 achieved an accuracy of 87.7%, it has a memory requirement of 6M multiplied by 32 bits. The proposed AQ approach thus illustrates competitive performance, achieving high accuracy while minimizing memory requirements, making it a promising option for applications demanding efficient resource utilization. This comparative analysis underlines the effectiveness of the adaptive quantization strategy in balancing accuracy and memory efficiency, which is critical for deployment in resource-constrained environments.

4 Conclusion

This paper presented an adaptive quantization method based on layer importance for deep neural networks, evaluated on VGG19, ResNet18, and ResNet50 using the EuroSAT dataset. Our technique achieved nearly original accuracy while significantly reducing the average bit-width, outperforming conventional fixed-precision quantization methods. The adaptive T_{margin} strategy further improved the model's performance, highlighting its potential for applications in environments with limited computational resources. In future work, we aim to explore practical applications across various architectures and datasets. We plan to assess the generalizability and scalability of our framework by investigating dynamic quantization during inference and integrating it with other compression techniques like pruning for enhanced efficiency. Additionally, we will utilize a wider range of spectral bands in satellite imagery to improve the reliability and interpretability.

REFERENCES

Yassir Benhammou, Domingo Alcaraz-Segura, Emilio Guirado, Rohaifa Khaldi, Boujemâa Achchab, Francisco Herrera, and Siham Tabik. Sentinel2globallulc: A sentinel-2 rgb image tile dataset for global land use/cover mapping with deep learning. *Scientific Data*, 9(1):681, 2022.

Weihan Chen, Peisong Wang, and Jian Cheng. Towards mixed-precision quantization of neural networks via constrained optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5350–5359, 2021.

Lei Deng, Guoqi Li, Song Han, Luping Shi, and Yuan Xie. Model compression and hardware acceleration for neural networks: A comprehensive survey. *Proceedings of the IEEE*, 108(4): 485–532, 2020.

Sara Elkerdawy, Mostafa Elhoushi, Abhineet Singh, Hong Zhang, and Nilanjan Ray. To filter prune, or to layer prune, that is the question. In *proceedings of the Asian conference on computer vision*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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*, 12(7):2217–2226, 2019.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, et al. Spectralgpt: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Krishna Karra, Caitlin Kontgis, Zoe Statman-Weil, Joseph C Mazzariello, Mark Mathis, and Steven P Brumby. Global land use/land cover with sentinel 2 and deep learning. In *2021 IEEE international geoscience and remote sensing symposium IGARSS*, pp. 4704–4707. IEEE, 2021.

Mehak Khan, Abdul Hanan, Meruyert Kenzhebay, Michele Gazzea, and Reza Arghandeh. Transformer-based land use and land cover classification with explainability using satellite imagery. *Scientific Reports*, 14(1):16744, 2024.

Nahsung Kim, Dongyeob Shin, Wonseok Choi, Geonho Kim, and Jongsun Park. Exploiting retraining-based mixed-precision quantization for low-cost dnn accelerator design. *IEEE Transactions on Neural Networks and Learning Systems*, 32(7):2925–2938, 2020.

Jan Koenig and Lionel Gueguen. A comparison of land use land cover classification using superspectral worldview-3 vs hyperspectral imagery. In *2016 8th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, pp. 1–5. IEEE, 2016.

Ivan Koryakovskiy, Alexandra Yakovleva, Valentin Buchnev, Temur Isaev, and Gleb Odinokikh. One-shot model for mixed-precision quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7939–7949, 2023.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning filters for efficient convnets. *arXiv preprint arXiv:1608.08710*, 2016.

Zhuo Li, Hengyi Li, and Lin Meng. Model compression for deep neural networks: A survey. *Computers*, 12(3):60, 2023.

Hongyang Liu, Sara Elkerdawy, Nilanjan Ray, and Mostafa Elhoushi. Layer importance estimation with imprinting for neural network quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2408–2417, 2021.

Janne Mäyrä, Sarita Keski-Saari, Sonja Kivinen, Topi Tanhuanpää, Pekka Hurskainen, Peter Kullberg, Laura Poikolainen, Arto Viinikka, Sakari Tuominen, Timo Kumpula, et al. Tree species classification from airborne hyperspectral and lidar data using 3d convolutional neural networks. *Remote Sensing of Environment*, 256:112322, 2021.

Gaurav Menghani. Efficient deep learning: A survey on making deep learning models smaller, faster, and better. *ACM Computing Surveys*, 55(12):1–37, 2023.

Rahul Mishra, Hari Prabhat Gupta, and Tanima Dutta. A survey on deep neural network compression: Challenges, overview, and solutions. *arXiv preprint arXiv:2010.03954*, 2020.

Antonio Rangel, Juan Terven, Diana M Cordova-Esparza, and Edgar A Chavez-Urbiola. Land cover image classification. *arXiv preprint arXiv:2401.09607*, 2024.

Babak Rokh, Ali Azarpeyvand, and Alireza Khanteymoori. A comprehensive survey on model quantization for deep neural networks in image classification. *ACM Transactions on Intelligent Systems and Technology*, 14(6):1–50, 2023.

Tushar Shinde. Adaptive quantization and pruning of deep neural networks via layer importance estimation. In *Workshop on Machine Learning and Compression, NeurIPS 2024*, 2024.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Chen Tang, Kai Ouyang, Zhi Wang, Yifei Zhu, Wen Ji, Yaowei Wang, and Wenwu Zhu. Mixed-precision neural network quantization via learned layer-wise importance. In *European Conference on Computer Vision*, pp. 259–275. Springer, 2022.

Guoliang Yang, Shuaiying Yu, Hao Yang, Ziling Nie, and Jixiang Wang. Hmc: Hybrid model compression method based on layer sensitivity grouping. *Plos one*, 18(10):e0292517, 2023.

Miao Yin, Yang Sui, Siyu Liao, and Bo Yuan. Towards efficient tensor decomposition-based dnn model compression with optimization framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10674–10683, 2021.

Xiaodong Zhang, Guanzhou Chen, Wenbo Wang, Qing Wang, and Fan Dai. Object-based land-cover supervised classification for very-high-resolution uav images using stacked denoising autoencoders. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(7):3373–3385, 2017.