



UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Escola Tècnica Superior d'Enginyeria
de Telecomunicació de Barcelona



AI Distillation for Lightweight Models

Bachelor's Degree Thesis

Submitted to the Faculty at the

Escola Tècnica Superior

d'Enginyeria de Telecomunicació de Barcelona

of the Universitat Politècnica de Catalunya

by

{Name of the Author}

In partial fulfillment

of the requirements for the

DEGREE IN TELECOMMUNICATION TECHNOLOGIES AND SERVICES ENGINEERING

Advisor: {Advisor's Name}

Reviewer: {Reviewer's Name (if applicable)}

Barcelona, January 2026

Resum

Aquest projecte té com a objectiu millorar l'eficiència i la usabilitat de grans models d'IA, específicament els models d'aprenentatge profund utilitzats en Visió per Computador quan es despleguen en dispositius perimetraus amb recursos limitats. El nostre objectiu és reduir les demandes computacionals d'aquests models sense sacrificar el rendiment desenvolupant una cadena modular que integra les tècniques de destil·lació de coneixement, l'ajustament de paràmetres amb LoRA i quantificació. El projecte quantifica els compromisos entre el rendiment i l'eficiència del model en cada etapa de la cadena. Això s'aconsegueix integrant tres metodologies bàsiques per reduir un gran model d'IA (ResNet-18) a una versió compacta i eficient (MobileNetV2).

Resumen

Este proyecto busca mejorar la eficiencia y la usabilidad de grandes modelos de IA, en concreto los modelos de aprendizaje profundo utilizados en Visión Artificial al implementarse en dispositivos perimetrales con recursos limitados. Buscamos reducir la demanda computacional de estos modelos sin sacrificar el rendimiento mediante el desarrollo de una canalización modular que integra técnicas de destilación de conocimiento, ajuste de parámetros con LoRA y cuantificación. El proyecto cuantifica el intercambio entre el rendimiento y la eficiencia del modelo en cada etapa de la canalización. Esto se logra integrando tres metodologías principales para reducir un gran modelo de IA (ResNet-18) a una versión compacta y eficiente (MobileNetV2).

Summary

This project aims to enhance the efficiency and usability of large AI models, specifically deep learning models used in Computer Vision when deployed on edge devices with limited resources. We aim to reduce the computational demands of these models without sacrificing performance by developing a modular pipeline that integrates knowledge distillation, LoRA fine-tuning, and quantization techniques. The project quantifies the trade-offs between model performance and efficiency at each stage of the pipeline. This is achieved by integrating three core methodologies to shrink a large AI model (ResNet-18) into a compact, efficient version (MobileNetV2).

Acknowledgements

Revision history and approval record

Revision	Date	Author(s)	Description
1.0	22/10/2025	NPC	Document creation
1.1	dd/mm/yyyy	AME, JPV	Error correction
2.0	dd/mm/yyyy	AME, MLO	Revised after review
4.0	dd/mm/yyyy	AME	Final version

DOCUMENT DISTRIBUTION LIST

Role	Surname(s) and Name
Student	Noa Paguera Contelles
Project Supervisor	Sergi Bermejo
Project Supervisor	Khac-Hoang Ngo

Written by:		Reviewed and approved by:	
Date	22/10/2025	Date	dd/mm/yyyy
Name		Name	
Position	Project Author	Position	Project Supervisor

Contents

Summary	2
Acknowledgements	4
Revision history and approval record	5
Contents	6
List of figures	8
List of tables	9
Abbreviations	10
1 Introduction	11
1.1 Motivation	11
1.2 Work goals	12
1.3 Requirements and specifications	13
1.4 Methods and procedures	14
1.5 Work plan	14
2 State of the art	16
2.1 Deep Learning	16
2.2 Model compression	17
2.3 Knowledge Distillation	17
2.4 Low-Rank Adaptation	18
2.5 Quantization	19
2.6 Federated Edge Learning	19
3 Methodology / Project Development	21
3.1 Data Preparation and Preprocessing	21
3.2 Model Training and Fine-Tuning	21
3.3 Quantization	23
3.4 Benchmark	23
3.5 Evaluation Framework and Testing	23
4 Results	25
4.1 Experiments and Tests	25
4.2 Data Visualization	26
4.3 Discussion	26

5 Sustainability Analysis and Ethical Implications	27
6 Conclusions and Future Work	29
6.1 Conclusions	29
6.2 Future Directions	29
Bibliography	30
A An appendix	32

List of Figures

List of Tables

1.1 Requirements table	13
1.2 Specifications table	13

Abbreviations

AI Artificial Intelligence

CNNs Convolutional Neural Networks

ETSETB Escola Tècnica Superior d'Enginyeria de Telecomunicació de Barcelona

FL Federated Learning

GRETST Grau en Enginyeria de Tecnologies i Serveis de Telecomunicació

LLMs Large Language Models

LoRA Low-Rank Adaption

PEFT Parameter-Efficient Fine-Tuning

Introduction

1.1 Motivation

This project aims to enhance the efficiency and usability of large AI models, specifically deep learning models used in Computer Vision (CV), when deployed on edge devices with limited resources. The aim is to reduce the computational demands of these models without sacrificing performance by developing a modular pipeline that integrates knowledge distillation, LoRA fine-tuning, and quantization techniques. The project systematically quantifies the trade-offs between model performance (accuracy) and efficiency (size, latency, computational cost) at each stage of the pipeline, applying information-theoretic principles for efficient knowledge transfer and exploring its applicability within a federated edge learning context.

This will be achieved by integrating three core methodologies to shrink a large AI model (ResNet-18) into a compact, efficient version (MobileNetV2):

1. Knowledge Distillation: Train a small AI model to mimic a larger, more complex one by transferring the teacher's knowledge into a student.
2. LoRA Fine-Tuning: Adjust the small model for specific tasks without increasing unnecessary size.
3. Quantization: Simplify how the model handles numbers, cutting memory usage and battery drain.

The project empirically validates this pipeline, providing a clear methodology for developing lightweight, task-specialized AI suitable for edge computing applications. The methodology is grounded in information theory to maximize knowledge transfer and is designed with federated edge learning in mind, enabling collaborative training without centralizing user data.

1.2 Work goals

The purpose of this project is to address the deployment of large, powerful AI models on resource-constrained edge devices (e.g., smartphones and IoT sensors) by designing, implementing, and evaluating a modular, three-stage pipeline. This pipeline systematically combines knowledge distillation, Low-Rank Adaptation (LoRA), and quantization to create a lightweight yet functional version of a standard computer vision model.

Furthermore, the project will be grounded in information theory to guide the distillation process, ensuring that the student model learns to compress the input while preserving the maximum information about the teacher's predictions. Knowledge distillation is a mechanism that enables communication-efficient and privacy-preserving collaborative training across multiple devices.

To achieve the main goal of developing and evaluating a compressed student model, the project has the following specific objectives:

1. Hybrid Distillation-LoRA Framework: Develop a two-phase approach where a compact student model is first distilled from a large pre-trained AI model and then fine-tuned using low-rank adaptation to achieve task-specific performance improvements.
 - 1.1 Develop a mini-model for hardware with limited resources: Create a compact model based on a method to distill knowledge from a large AI into a smaller, lighter model that can run efficiently on T4 Colab.
 - 1.2 Optimize this Lightweight Version to minimize memory usage for CV by Fine-Tuning with Low-Rank Adaptation: Use techniques like LoRA to further optimize the smaller model for specific tasks, improving its performance without increasing its size significantly.
2. Reduce Computational Costs through Quantization and Edge Optimization: Apply quantization strategies to shrink the model's size and lower its energy consumption, making it suitable for real-world applications. In other words, implement quantization-aware distillation methods to reduce model size and computational complexity, ensuring viability for edge deployment.
3. Evaluate the resulting AI model against the original using standard metrics for speed, memory usage, and accuracy:
 - 3.1 Evaluate Performance: Test the optimized model on practical tasks (e.g., object classification), analyzing how well it balances size, speed, and accuracy.
 - 3.2 Study Fine-Tuning Dynamics: Investigate how the initial distillation phase impacts the efficiency and effectiveness of subsequent LoRA-based fine-tuning, identifying the most critical elements of the original large AI model knowledge for downstream performance.

The evaluation will be done in parallel with knowledge distillation and fine-tuning.

4. Frame the distillation process within an information-theoretic context: Analyze knowledge transfer within an information-theoretic context, and explore extending the above pipeline to a Federated Learning setting for decentralized, privacy-preserving training.

1.3 Requirements and specifications

A set of requirements has been defined to ensure that the developed system meets the objectives, covering key points such as the model's capabilities and its evaluation.

Requirements

Requirement	Description
Benchmark	The distilled model must retain high accuracy compared to their teacher while being deployable on resource-constrained edge devices.
Efficiency	The lightweight model must demonstrate reduced computational cost and inference time.
Pipeline	The pipeline must integrate knowledge distillation, LoRA fine-tuning and quantization. It must be reproducible and modular for future extensions.
Evaluation	The resulting model must be evaluated on standard CV benchmarks and framed within an information-theoretic perspective.

Table 1.1: Requirements table

Specifications

Specification	Description
Accuracy	The student model must achieve 90% accuracy of the teacher model.
Model size	The model must reduce >75% of the original size.
Time optimization	The model must reduce inference time.
Development platform	The developed system must be compatible with major operating systems such as Linux, macOS, Windows.

Table 1.2: Specifications table

Model Architectures:

The selection of teacher and student models is designed to create a realistic and challenging distillation scenario.

- **Teacher Model: ResNet-18:** This architecture is selected for its known high performance in CV, since it can learn complex, non-linear channel features, establishing a strong upper-bound for accuracy.
- **Student Model: MobileNetV2:** This model is chosen as a prototypical lightweight architecture for edge devices. Its efficiency comes from the use of depthwise separable convolutions, which drastically reduce the parameter count and computational load compared to standard convolutions. The significant architectural differences between ResNet-18 and MobileNetV2 make them excellent test cases for evaluating the effectiveness of different distillation techniques.

1.4 Methods and procedures

The project builds upon existing state-of-the-art technologies, architectures and methods developed by other authors:

- Teacher model: ResNet-18 (He et al., 2015)
- Student model: MobileNetV2 (Sandler et al., 2018)
- Distillation methods: based on Hinton et al. (2015)
- LoRA fine-tuning: Low-rank adaptation methods (Hu et al., 2021)

The project is conducted independently and does not form part of any department or company research or development project. The UPC supervisor provided the foundational concepts and initial ideas, which were then complemented by the supervisor at Linköping University.

1.5 Work plan

The project is structured into three overlapping phases over 17 weeks (September 1, 2025 – December 20, 2025), ensuring a clear progression from foundational development to advanced analysis.

Distillation-LoRA Framework & Initial Evaluation. This phase focuses on implementing the core compression pipeline. The key objective is a functional, fine-tuned student model benchmarked against the original teacher and baseline student.

Quantization & Continued Evaluation. This phase builds on Phase 1 by applying the final compression stage, quantization. The objective is the final, fully compressed model, and a complete set of experimental results.

Theoretical Analysis, Final Report, and Submission. This final phase is dedicated to the theoretical framing of the project, in-depth analysis of the results, and writing the thesis document.

Work Packages

- WP1: Distillation-LoRA Framework & Evaluation
 - Tasks:
 - * T1.1: Comprehensive literature review on KD, LoRA, and quantization.
 - * T1.2: Setup of development environment (PyTorch, Colab).
 - * T1.3: Implementation of data loading and preprocessing for CIFAR-100.
 - * T1.4: Train and benchmark the ResNet-18 teacher and baseline MobileNetV2 student.
 - * T1.5: Implement the knowledge distillation framework and train the distilled student model.
 - * T1.6: Implement LoRA fine-tuning on the distilled student model.
 - * T1.7: Benchmark all intermediate models (teacher, baseline, distilled, LoRA-tuned).
- WP2: Quantization & Theoretical Framing
 - Tasks:
 - * T2.1: Apply post-training INT8 quantization to the LoRA-tuned model.
 - * T2.2: Design and execute a final, comprehensive benchmarking script to evaluate all model versions on accuracy, size, latency, and FLOPs.
 - * T2.3: Generate all tables and visualizations for the results chapter.
- WP3: Project Management & Thesis Writing
 - Tasks:
 - * T3.1: Draft Introduction and State of the Art chapters.
 - * T3.2: Draft Methodology chapter as WP1 and WP2 tasks are completed.
 - * T3.3: Draft Results, Sustainability Analysis, and Conclusion chapters.
 - * T3.4: Final review, formatting, and thesis submission.
 - Primary Deliverables:
 - * Proposal Presentation and TFG Work Plan. deadline: 05/10/2025
 - * Critical Review. deadline: 30/11/2025
 - * Final Bachelor's Thesis document. deadline: 18/01/2026

State of the art of the technology applied in this thesis

In order to achieve better performance, current deep learning models generally are deeper and wider. However, these heavy models are hard to deploy on resource-constrained devices in practice due to computational and memory resource limitations. We aim to develop an AI model that is smaller, faster, and more energy-efficient while maintaining its intelligence. In the following sections, we follow the development of deep learning, making a survey and integrating three core methodologies to shrink a large AI model (ResNet-18) [7] into a compact, efficient version (MobileNetV2) [17]: Knowledge Distillation [8], LoRA Fine-Tuning [9] and Quantization.

This project empirically validates this pipeline, providing a clear methodology for developing lightweight, task-specialized AI suitable for edge computing applications. The methodology is grounded in information theory to maximize knowledge transfer and is designed with Federated Edge Learning (FEL) [24] in mind, enabling collaborative training without centralizing user data.

2.1 Deep Learning

Deep learning is a subset of machine learning that utilizes multilayered neural networks to simulate decision-making of the human brain to perform tasks such as classification, regression and feature learning. Deep neural networks consist of multiple layers of interconnected nodes, a combination of data inputs, weights and bias, which can be expressed as $W^T \cdot X + b$.

Over the past decade, deep learning has revolutionized computer vision. Architectures such as ResNet have set new records in benchmark performance. However, these deep learning models are computationally intensive and require large amounts of memory and processing power, which limits their deployment to high-performance servers with dedicated GPUs or TPUs.

On the other hand, edge devices such as smartphones, IoT sensors, drones, and medical wearables, have resource constraints, including limited processing power, memory, and energy. Despite these constraints, edge computing is becoming increasingly important because it reduces latency, lowers dependence on centralized cloud services, and improves privacy by keeping sensitive data local. A key challenge in this field is finding a way to leverage the power of deep learning models while addressing the limitations of edge devices.

2.2 Model compression

The deployment of advanced AI models on edge devices with limited resources has motivated substantial research into model compression techniques. The central goal of this field is to balance the power of large models with the requirements of efficiency, latency and privacy. This field has evolved from applying individual optimization techniques in isolation to designing integrated, multi-stage pipelines that synergistically combine methods for maximum efficiency.

There are several ways to make deep learning models more efficient without sacrificing accuracy with the following core components of this project: Knowledge Distillation (KD) shrinks the knowledge base [8], Parameter-Efficient Fine-Tuning (PEFT) [25] with a specific focus on Low-Rank Adaptation (LoRA) fine-tunes for specific tasks [9], and Quantization further compresses the model for deployment.

In this integration, we combine multiple AI model compression techniques in sequence for developing methodologies; instead of applying these methods in isolation, we view them as interconnected steps in a single, integrated process, in which the order and interaction of these techniques are considered crucial for maximizing efficiency and achieving the best compression results, as exemplified by approaches that combine distillation, LoRA, and pruning in a multi-stage process.

Moslem's recent work on speech translation combines knowledge distillation with QLoRA and iterative pruning in a multi-stage process [15]. This highlights the importance of technique order, as distillation can create a more robust student model, aiding subsequent quantization or fine-tuning.

2.3 Knowledge Distillation

A critical analysis of seminal and recent works reveals a clear trajectory in the field. Initially, techniques were developed to solve singular problems, e.g. KD for knowledge transfer and LoRA for tuning efficiency.

Hinton, Vinyals, & Dean focused exclusively on the concept of transferring "dark knowledge" from a teacher to a student model, establishing the theoretical groundwork for response-based distillation [8]. Although their work showed that a student model could

learn a richer structure by training on the teacher’s outputs, they didn’t address the efficiency of adapting the distilled model to new tasks.

Knowledge Distillation is a machine learning technique that aims to transfer the learnings of a large pre-trained model to a smaller and more compact one so that the “student” mimics the behaviour of the “teacher” [8] by matching its predictions. It is usually applied to deep neural networks with many layers and learnable parameters. The main idea is to use soft probabilities of the larger network to supervise the smaller one, which reveal more information than the class labels alone. Soft targets can be estimated by a softmax function as:

$$q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

The loss for the student is then a linear combination of cross entropy loss L_{CE} and knowledge distillation loss L_{KD} .

$$L = (1 - \alpha)L_{CE}(y, p_{\text{student}}) + \alpha L_{KD}(p_{\text{teacher}}, p_{\text{student}})$$

Knowledge Distillation techniques have been employed across multiple fields, including speech recognition, natural language processing, image recognition [7] and object detection. KD is an effective means of transferring the capabilities from a large model or set of models to a single smaller model.

AI model’s accuracy and capacity are not enough to make the model useful - it has a limit of time, memory and computational resources. While top performing models are often too large for most applications, small models are faster yet lack the accuracy and knowledge capacity of the first.

2.4 Low-Rank Adaptation

Low-Rank Adaptation (LoRA) was originally developed for efficient fine-tuning in large language models (LLMs). LoRA offers an opportunity to fine-tune compact models without enlarging their size. Hu et al. developed LoRA to address the computational infeasibility of fine-tuning large models, reducing trainable parameters, memory usage and training time by freezing pre-trained weights and training only small, low-rank adapter matrices inserted into existing weight structures, without increasing inference latency [9]. The pre-trained model’s original weight matrix is frozen and only the smaller matrices are updated during training. Yet, the LoRA paper did not focus on reducing the size of the base model itself, which remained a significant barrier for edge deployment.

Instead of retraining the whole model, LoRA freezes the original weights and parameters of the model. On top of this original model, it adds a lightweight addition called a low-rank matrix, which is then applied to new inputs to get results specific to the context. The low-rank matrix adjusts for the weights of the original model so that outputs match the desired use case.

The diagram shows how the additional matrices A and B are updated by using smaller matrices of rank r. Once LoRA training is complete, smaller weights are merged into a new weight matrix, without modifying the original weights W of the pre-trained model and not needing to store a complete copy of the model. In the training process after calculating the loss, the loss is only backpropagated to the LoRA matrices. [insert diagram of LoRA]

Full fine-tuning requires storing optimizer states for all parameters. Since LoRA freezes most of the model’s weights, GPU memory usage is lower than full fine-tuning storage efficiency training speed.

Forward pass:

$$h = W_0x + \Delta Wx = W_0x + BAx$$

2.5 Quantization

The limitations of these isolated approaches naturally led to hybrid solutions such as QLoRA [2], which combines 4-bit quantization and LoRA adapters, enabling the fine-tuning of large models on a single consumer GPU. This unified approach addresses memory issues in both base models and fine-tuning, shifting towards combined compression solutions.

Quantization reduces the precision of model weights and activations (e.g., from 32-bit floating-point to 8-bit integers), which reduces memory and speeds up inference time.

Post-training quantization (PTQ) applies quantization to a fully trained model without additional training. It typically relies on a small calibration dataset to estimate activation ranges and is widely used due to its simplicity and low computational cost. While PTQ can introduce accuracy degradation it is often sufficient for inference-only deployments where retraining is impractical.

In contrast, quantization-aware training (QAT) applies quantization during training, allowing the model to adapt to reduced precision and mitigating accuracy degradation. This approach generally achieves higher accuracy than PTQ but requires access to training data and increased training complexity.

2.6 Federated Edge Learning

In many real-world applications (e.g., healthcare, smart cities), it is impractical to centralize user data for privacy concerns. Federated learning (FL) enables decentralized training across distributed devices while keeping raw data local. However, FL aggravates the need for efficient models because each device must independently train and update parameters under limited resources. By combining KD, LoRA, and quantization, the proposed pipeline aligns with federated edge learning, ensuring that lightweight models can still learn collaboratively without overwhelming device capabilities.

The evolution of compression pipelines is further enriched by integrating principles from information theory and paradigms like Federated Edge Learning (FEL) [24]. Knowledge distillation, seen through an information-theoretic perspective, aims for a student model to compress input while maximizing information from the teacher’s prediction, which implies optimal teachers might be intermediate models, preserving “dark knowledge” – mutual information crucial for generalization often lost in fully converged models [22].

Concurrently, the rise of FEL has positioned knowledge distillation as a critical enabler for privacy-preserving, decentralized machine learning. In Federated Distillation (FD), instead of exchanging high-dimensional model parameters, clients exchange compact model outputs (logits) on a shared, public dataset, allowing heterogeneous models to collaboratively train without sharing private data [24].

This project meets at the intersection of these trends, using information-theoretic concepts to guide the distillation process and structuring the pipeline in a way that is directly applicable to a future federated deployment.

While these multi-stage pipelines are emerging, particularly in natural language processing, a systematic, empirical study of a sequential KD → LoRA → Quantization pipeline for computer vision remains a clear research gap.

Methodology / Project Development

This chapter describes the methodology employed to design, train and evaluate the knowledge distillation framework proposed in this study. It outlines the dataset, data preparation, teacher and student model architectures, training procedures, evaluation metrics and experimental setup.

This research follows an experimental design where a teacher model is used to supervise the training of a smaller student model. A series of experiments evaluate the impact of different distillation strategies.

3.1 Data Preparation and Preprocessing

For the initial evaluations of the model, we conducted experiments to evaluate RestNet18 and MobileNetV2 performance using the public available CIFAR-10 and CIFAR-100 datasets. The CIFAR-10 and CIFAR-100 datasets are labeled subsets of the 80 million tiny images dataset. This dataset was chosen because it is an established computer vision dataset used for object recognition and it allows to quickly try different algorithms, ideal for benchmarking and preliminary evaluations.

Data preparation is a crucial step that transforms the input data into the format required by the model, ensuring it can be correctly processed and learned from.

The images are normalized to ensure consistency in the data format. Data normalization prevents large values from dominating the learning process by ensuring numerical features are on a similar scale for optimal model performance. It avoids numerical instability, speeds up convergence in gradient-based algorithms and ensures features contribute equally.

3.2 Model Training and Fine-Tuning

Training a model involves teaching it to identify patterns within data so that it can make accurate predictions or decisions when presented with new, unseen inputs. The training

process is generally divided into two key stages: pre-training and fine-tuning.

- Pre-training refers to training a model on a large, general-purpose dataset, often using unsupervised or self-supervised learning methods. During this stage, the model develops a foundational understanding of the data by learning to extract features and recognize broad patterns, without focusing on any specific task. The resulting pre-trained models serve as a strong starting point for various downstream applications, significantly reducing both time and computational cost compared to training a model entirely from scratch.
- Fine-tuning, on the other hand, builds upon a pre-trained model by adapting it to a specific task or dataset. This is done by further training the model on a smaller, labeled, and task-specific dataset. Through fine-tuning, the model refines its learned representations and becomes specialized in the target application while still benefiting from the general knowledge acquired during pre-training.

In this work, the ResNet-18 model serves as the teacher network due to its strong performance on image classification tasks. We focus on fine-tuning the ResNet-18 model rather than training it from scratch. This approach is supported by the availability of high-quality pre-trained checkpoints in PyTorch, which have been trained on large-scale datasets. These pre-trained models already capture fundamental visual features and general data representations. However, their learned knowledge remains shallow and tied to the original training data. Fine-tuning enables the model to adapt and develop more task-specific capabilities. In our case, the teacher was pre-trained on ImageNet and fine-tuned on the target CIFAR dataset.

By leveraging the visual representations obtained during pre-training, this method avoids the significant computational expense and environmental impact associated with large-scale model training. Consequently, it aligns with sustainable AI practices. In our implementation, we fine-tune the ResNet-18 model using labeled data to tailor it to the desired task.

In the distillation process, the student model is a MobileNetV2, selected for its compact size and suitability for edge deployment. The student model is trained for 75 epochs using Adam optimizer with learning rate of 0.001. All hyperparameters are tuned based on the validation set. After several runs, the distillation temperature and alpha values chosen are $T=4$ and $\alpha = 0.3$. It is trained under the supervision of the teacher, using both ground truth labels and soft targets learned from the larger model.

Model performance is evaluated using top-1 accuracy, F1-score, and computational metrics such as inference latency.

To select the best checkpoint, we monitor validation accuracy and choose the checkpoint that achieves the highest score. This helps ensure that the model generalizes well to unseen data and avoids overfitting to the training set. Although fine-tuning continues for additional steps, this process ensures that training stops at an optimal point.

3.3 Quantization

Model quantization converts high-precision floating-point values into lower-precision formats such as integers, resulting in faster inference, lower energy consumption, and reduced storage requirements. In particular, model parameters (weights and activations) are approximated using INT8 instead of the more common 32-bit floating-point format (FP32). This is especially beneficial in tasks such as image processing and neural network inference, where computational efficiency and memory optimization are crucial.

3.4 Benchmark

Evaluation is a fundamental component of any research project, as it provides a systematic and objective means of measuring the effectiveness and robustness of a proposed system. A well-designed evaluation framework allows researchers to compare models under consistent conditions, quantify performance using standardized metrics, and identify the strengths and limitations of different approaches.

The quality of the evaluation process depends heavily on the datasets used. These datasets must be both trustworthy and sufficiently diverse to reflect real-world scenarios. Reliable benchmark data ensure that the results are reproducible and not influenced by noise or bias, while variation within the dataset allows for a more comprehensive assessment of the model's ability to generalize beyond the specific examples seen during training.

3.5 Evaluation Framework and Testing

The evaluation framework used in this project is structured to assess the performance and reliability of the MobileNetV2 student model against the teacher. The framework enables consistent comparison of model behavior under different experimental parameters. This structure not only helps verify the model's robustness but also ensures that its performance can be generalized beyond the specific conditions encountered during training.

To conduct evaluation, the model is tested on a dataset that is isolated from the training process, therefore ensuring an unbiased assessment. The testing procedure examines the model's ability to correctly classify inputs and handle variations in data distribution.

Following the evaluation, accuracy is computed as the primary performance metric. Accuracy provides a clear measure of the proportion of correct predictions relative to the total number of samples, making it an appropriate and widely accepted indicator for classification tasks.

However, accuracy is often insufficient for evaluating models trained on imbalanced datasets, where one class substantially outweighs the others. In such cases, a model may achieve deceptively high accuracy simply by consistently predicting the majority class, while failing to detect the minority class altogether. This makes accuracy an unreliable

indicator of true model performance. To address these limitations, additional metrics such as precision, recall, and the F1 score are employed. These metrics provide more informative insights, particularly regarding the model's ability to recognize minority classes and to manage trade-offs between false positives and false negatives.

Precision measures the proportion of correctly predicted positive instances out of all instances predicted as positive. High precision indicates that the model's positive predictions are reliable.

Recall, also known as sensitivity or the true positive rate, quantifies the proportion of actual positive instances that the model successfully identifies. Recall is crucial in scenarios where failing to detect positive cases, producing false negatives, is particularly costly.

Precision and recall frequently exhibit a trade-off. Improving one may reduce the other. Therefore, selecting the appropriate balance depends on the specific requirements and risks associated with the task.

The F1 score provides a single metric that balances both precision and recall, making it particularly useful when false positives and false negatives are equally important. It is defined as the harmonic mean of precision and recall:

$$F1Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

The F1 score ranges from 0 to 1, with higher values indicating stronger overall performance. Achieving an F1 score of 1 would require perfect precision and recall simultaneously, which is a condition rarely met in practice due to the trade-offs between detecting all positive instances and minimizing incorrect positive predictions.

Results

4.1 Experiments and Tests

This chapter provides an overview of the experimental setup used for training and evaluating the model. It specifically addresses the configurations implemented to achieve the results discussed. For further methodological details or prerequisites, please refer to Chapter 3. In this section, the parameters employed for fine-tuning are defined.

Fine-tuning Configuration

Fine-tuning of the MobileNet model is performed using a custom configuration that defines the parameters used in training, optimization, and overall model behavior. This setup ensures the model can adapt to the specific dataset used in our experiments. The configuration remains consistent across all trained models to maintain uniformity.

The model uses a 32-bit precision format (FP32), a storage format for floating-point values suitable for applications where higher precision is not critical. This is particularly advantageous in scenarios like image processing and neural network training, where computational efficiency and memory optimization are crucial.

Fixed random seed values of 42, 518, 1993 are used to ensure reproducibility, enabling consistent results across multiple runs.

To maintain efficiency during training, only the highest-performing checkpoint is retained. Selection of the best checkpoint is based on the validation loss metric, ensuring that the stored model reflects peak performance for future evaluation.

The model is pre-trained on the ImageNet dataset for image classification. Data pre-processing includes normalization to standardize inputs before they are fed into the model.

For optimization, we use the Adam optimizer, a widely adopted choice due to its adaptability to changing gradients during training. The model begins training with an initial learning rate of 0.001, chosen to provide stable updates without causing divergence.

Knowledge Distillation Results

The implementation of the knowledge distillation phase aimed to transfer the predictions of the ResNet18 to the more compact MobileNetV2. Initial results indicate that the student model achieved an accuracy of 62%, successfully recovering 76% of the teacher's performance. This suggests that the soft targets provided by the teacher were effective in regularizing the student's training, despite the significant reduction in total parameter count.

LoRA & Quantization Trade-offs

To further optimize the model for deployment, LoRA fine-tuning was applied with 8-bit quantization. While quantization typically introduces a quantization error that degrades performance, the LoRA adapters effectively compensated for this loss. The final quantized model reached a compressed size of 79% compared to the baseline, with an accuracy degradation of only 12%.

The proposed pipeline demonstrates that the integration of distillation, low-rank adaptation, and quantization provides a robust framework for model compression. The findings confirm that it is possible to achieve a reduction in memory usage while maintaining competitive accuracy levels. This makes the resulting model suitable for resource-constrained environments where the original baseline would be computationally infeasible.

4.2 Data Visualization

Create visual representations of the results (e.g., scatter plots, bar charts). Interpret the visualizations and relate them to the research questions.

4.3 Discussion

Acknowledge any limitations in the data or analysis. Explain how these limitations may have influenced the results.

Sustainability Analysis and Ethical Implications

The sustainability of technological innovations, particularly those involving artificial intelligence and large-scale data processing, requires careful assessment across environmental, social, economic, and ethical dimensions. This section evaluates the broader impacts of the project and discusses potential societal implications associated with advancing this technology.

The development, training, and deployment of machine learning models require substantial computational workloads, which translate into energy consumption and associated environmental costs. In this project, the processing of visual data during the preparation phase and the subsequent fine-tuning and evaluation of the model contribute to the overall computational demand.

Training machine learning models requires considerable electricity due to the intensive operations involved. The NVIDIA T4 available on Google Colab, used for this project, consumes approximately 70W, emphasizing the importance of efficient resource usage during training.

The T4 GPU was selected because GPUs are optimized for parallel computation, enabling them to conduct multiple operations simultaneously. This significantly reduces training time compared to CPU-only systems and makes GPUs particularly effective for large-scale datasets and complex neural network architectures. While GPUs consume more power than CPUs, the computational efficiency they provide justifies their use in scenarios where model training time and throughput are critical.

Storage requirements also contribute to environmental impact. The project requires approximately [Xbytes] of storage, primarily due to the datasets used in model training. Storage infrastructure has its own energy overhead: maintaining 1 terabyte of data in a data center typically consumes around 100 kWh annually. Consequently, managing [Xbytes] of data would result in an estimated annual energy consumption of [X kWh], excluding the additional energy needed for data processing and cooling. Implementing

effective data management and pruning strategies is therefore essential to minimize environmental impact when handling large-scale datasets.

Training AI models contributes to carbon emissions, as the electricity used by computational hardware often originates from carbon-intensive energy sources. Although an exact carbon footprint cannot be calculated without detailed measurements of power usage, hardware efficiency, and local energy mixes, we can estimate the approximate environmental impact associated with the model training process. These estimates help highlight the importance of optimizing training pipelines, adopting energy-efficient hardware, and exploring low-carbon or renewable energy options where available.

Conclusions and Future Work

6.1 Conclusions

- Summarize the main results of your work.
- Discuss the degree of achievement in relation to the objectives set at the beginning of the work.
- Highlight the contributions of your work to the field of study.

6.2 Future Directions

- Identify areas for future research or development based on your work.
- Discuss possible ways to expand or improve the project.
- Consider questions that remained unanswered and opportunities for future exploration.

Bibliography

- [1] Jang Hyun Cho and Bharath Hariharan. *On the Efficacy of Knowledge Distillation*. 2019. arXiv: [1910.01348 \[cs.LG\]](https://arxiv.org/abs/1910.01348). URL: <https://arxiv.org/abs/1910.01348>.
- [2] Tim Dettmers et al. *QLoRA: Efficient Finetuning of Quantized LLMs*. 2023. arXiv: [2305.14314 \[cs.LG\]](https://arxiv.org/abs/2305.14314). URL: <https://arxiv.org/abs/2305.14314>.
- [3] Albert Einstein. “Zur Elektrodynamik bewegter Körper”. In: *Annalen der Physik* 322.10 (1905), pp. 891–921.
- [4] Simon Fear. *booktabs – Publication quality tables in L^AT_EX*. 2020. URL: <http://mirrors.ctan.org/macros/latex/contrib/booktabs/booktabs.pdf> (visited on 11/01/2023).
- [5] Christian Feuersänger. *Manual for Package PGFPLOTS*. URL: <http://mirrors.ctan.org/graphics/pgf/contrib/pgfplots/doc/pgfplots.pdf> (visited on 11/03/2023).
- [6] Richard P. Feynman, Robert B. Leighton, and Matthew Sands. *The Feynman Lectures on Physics*. Vol. I-III. Reading, Massachusetts: Addison-Wesley, 1963. ISBN: 978-0201021165.
- [7] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: [1512.03385 \[cs.CV\]](https://arxiv.org/abs/1512.03385). URL: <https://arxiv.org/abs/1512.03385>.
- [8] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. “Distilling the Knowledge in a Neural Network”. In: (2015). arXiv: [1503.02531 \[stat.ML\]](https://arxiv.org/abs/1503.02531). URL: <https://arxiv.org/abs/1503.02531>.
- [9] Edward J. Hu et al. *LoRA: Low-Rank Adaptation of Large Language Models*. 2021. arXiv: [2106.09685 \[cs.CL\]](https://arxiv.org/abs/2106.09685). URL: <https://arxiv.org/abs/2106.09685>.
- [10] Philip Kime, Moritz Wemheuer, and Philipp Lehman. *The biblatex Package*. URL: <https://ctan.org/pkg/biblatex> (visited on 11/06/2023).
- [11] Donald Knuth. *Knuth: Computers and Typesetting*. URL: <https://www-cs-faculty.stanford.edu/~knuth/abcde.html>.
- [12] Donald E. Knuth. *The T_EXbook*. Reading, Massachusetts: Addison-Wesley, 1986. ISBN: 0-201-13448-9.

- [13] Leslie Lamport. *L^AT_EX: A Document Preparation System, 2nd Edition*. Reading, Massachusetts: Addison-Wesley, 1994. ISBN: 0-201-52983-1.
- [14] Frank Mittelbach and Ulrike Fischer. *The L^AT_EX Companion*. 3rd ed. 2 vols. Reading, Massachusetts: Addison-Wesley, 2023.
- [15] Yasmin Moslem. “Efficient Speech Translation through Model Compression and Knowledge Distillation”. In: *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*. Association for Computational Linguistics, 2025, pp. 379–388. doi: [10.18653/v1/2025.iwslt-1.40](https://doi.org/10.18653/v1/2025.iwslt-1.40). URL: <http://dx.doi.org/10.18653/v1/2025.iwslt-1.40>.
- [16] Tobias Oetiker et al. *The Not So Short Introduction to L^AT_EX*. URL: <https://tobi.oetiker.ch/lshort/lshort.pdf> (visited on 10/31/2023).
- [17] Mark Sandler et al. *MobileNetV2: Inverted Residuals and Linear Bottlenecks*. 2019. arXiv: [1801.04381 \[cs.CV\]](https://arxiv.org/abs/1801.04381). URL: <https://arxiv.org/abs/1801.04381>.
- [18] Claude E. Shannon. “A Mathematical Theory of Communication”. In: *The Bell System Technical Journal* 27.3 (1948), pp. 379–423. doi: [10.1002/j.1538-7305.1948.tb01338.x](https://doi.org/10.1002/j.1538-7305.1948.tb01338.x).
- [19] Wolfgang Skala. *Drawing Gantt Charts in L^AT_EX with TikZ*. URL: <http://osl.ugr.es/CTAN/graphics/pgf/contrib/pgfgantt/pgfgantt.pdf>.
- [20] Thomas F. Sturm. *The tcolorbox package*. URL: <http://mirrors.ctan.org/macros/latex/contrib/tcolorbox/tcolorbox.pdf> (visited on 11/05/2023).
- [21] Till Tantau. *The TikZ and PGF Packages*. URL: <http://mirrors.ctan.org/graphics/pgf/base/doc/pgfmanual.pdf> (visited on 11/02/2023).
- [22] Chaofei Wang et al. *Efficient Knowledge Distillation from Model Checkpoints*. 2022. arXiv: [2210.06458 \[cs.LG\]](https://arxiv.org/abs/2210.06458). URL: <https://arxiv.org/abs/2210.06458>.
- [23] Joseph Wright. *siunitx – A comprehensive (SI) units package*. URL: <https://ctan.org/pkg/siunitx> (visited on 11/08/2023).
- [24] Zhiyuan Wu et al. *Knowledge Distillation in Federated Edge Learning: A Survey*. 2024. arXiv: [2301.05849 \[cs.LG\]](https://arxiv.org/abs/2301.05849). URL: <https://arxiv.org/abs/2301.05849>.
- [25] Dan Zhang et al. *Parameter-Efficient Fine-Tuning for Foundation Models*. 2025. arXiv: [2501.13787 \[cs.CL\]](https://arxiv.org/abs/2501.13787). URL: <https://arxiv.org/abs/2501.13787>.

APPENDIX **A**

An appendix

Appendices may be included in your thesis but it is not a requirement.

