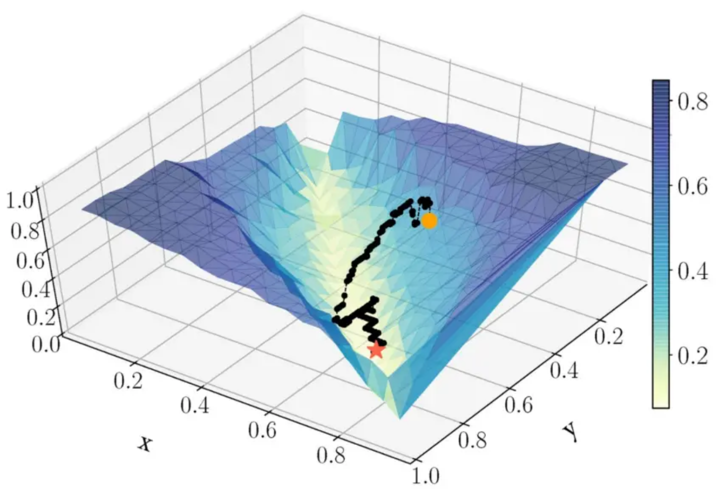
Paris-Saclay University
Institut d'optique Graduate School
ATSI Master 2

Machine Learning

Incremental Learning Learning for Traffic Sign Recognition

Realized by:
Abdelbaki Guir
Omar Mameche
Taqwa Mezaache

Professor:
Stephane Herbin



March 22, 2025

I. Introduction

Incremental learning is a critical area of research in machine learning, particularly for applications where models must adapt to new data over time without forgetting previously learned knowledge. This project focuses on addressing the challenge of catastrophic forgetting in the context of traffic sign recognition using the German Traffic Sign Recognition Benchmark (GTSRB) dataset.

To tackle this issue, the proposed solution combines two complementary techniques: memory rehearsal and knowledge distillation. Memory rehearsal retains a subset of data from previous tasks to "rehearse" old knowledge during training, while knowledge distillation uses a "teacher" model to guide the training of the updated model, ensuring consistency in predictions for old tasks. This document provides a detailed theoretical analysis of the methods employed, their implementation, and the insights gained, highlighting the effectiveness of the proposed approach in mitigating catastrophic forgetting.

The project is publicly available on GitHub at the following link: https://github.com/abdou1579/Incremental-Learning-for-GTSRB. This repository contains the complete implementation, including the code and dataset preprocessing pipeline, enabling reproducibility and further exploration of the proposed approach.

II. Theoretical Background

1. Catastrophic Forgetting

Catastrophic forgetting occurs when a neural network trained on a sequence of tasks loses performance on earlier tasks as it learns new ones. This phenomenon arises because the model's parameters are optimized for the new task, overwriting the knowledge required for previous tasks. Mathematically, this can be described as a loss of information in the weight space of the model:

$$\mathcal{L}(\theta_{\text{new}}) \approx \mathcal{L}(\theta_{\text{old}}) + \Delta\mathcal{L},$$

where θ_{new} and θ_{old} represent the model parameters before and after learning the new task, and $\Delta\mathcal{L}$ quantifies the degradation in performance on the old task.

2. Incremental Learning

Incremental learning aims to mitigate catastrophic forgetting by enabling models to learn new tasks while retaining knowledge of previous ones. This is particularly important in real-world applications like traffic sign recognition, where new sign categories may be introduced over time. The two primary techniques used in this project are:

- **Memory Rehearsal**: Retains a subset of data from previous tasks to "rehearse" old knowledge during training.
- **Knowledge Distillation**: Uses a "teacher" model to guide the training of the updated model, ensuring consistency in predictions for old tasks.

III. Dataset: German Traffic Sign Recognition Benchmark (GTSRB)

The GTSRB dataset consists of 43 classes of traffic signs, each representing a unique category. The dataset is divided into 5 sequential tasks to simulate a real-world scenario where a model must incrementally learn new sign categories. The diversity in lighting conditions, perspectives, and occlusions makes this dataset a challenging benchmark for incremental learning.

To prepare the GTSRB dataset for training, the following preprocessing steps were applied:

- **Resizing**: All images were resized to a fixed resolution of 32×32 pixels to ensure uniformity and reduce computational complexity.
- **Normalization**: The pixel values were normalized using dataset-specific mean and standard deviation values.

IV. Solution Architecture

1. Convolutional Neural Network (CNN) Model

The model architecture is designed to extract hierarchical features from traffic sign images. It consists of:

- Three convolutional layers with 32, 64, and 128 filters, respectively.
- Batch normalization after each convolutional layer to stabilize training.
- Max pooling for dimensionality reduction.
- Dropout (rate = 0.5) to prevent overfitting.
- Two fully connected layers with 512 units and 43 output classes.

2. Memory Rehearsal

Memory rehearsal addresses catastrophic forgetting by maintaining a fixed-size memory buffer of samples from previous tasks. The key theoretical aspects of this approach are:

- **Buffer Size**: The buffer stores up to 150 samples, ensuring a balance between memory efficiency and performance.
- **Balanced Sampling**: Up to 20 samples per class are stored to maintain class diversity.
- **Random Selection**: When the buffer exceeds its capacity, random samples are discarded to ensure uniform representation.

3. Knowledge Distillation

Knowledge distillation is a technique used to transfer knowledge from a pre-trained "teacher" model to a "student" model. In the context of incremental learning, it plays a crucial role in mitigating catastrophic forgetting by ensuring that the updated model retains the behavior of the teacher model on previously learned tasks. The theoretical foundation of knowledge distillation lies in minimizing the difference between the output distributions of the teacher and student models, typically measured using the Kullback-Leibler (KL) divergence.

3..1 Theoretical Foundation

The core idea of knowledge distillation is to use the teacher model's output probabilities as soft targets for training the student model. Unlike hard labels (one-hot encoded vectors), soft targets provide richer information about the relationships between classes, which helps the student model generalize better. The loss function for knowledge distillation is defined as:

$$\mathcal{L}_{\text{distill}} = \text{KL}(p_{\text{teacher}} \| p_{\text{student}}),$$

where:

- p_{teacher} is the softmax output of the teacher model, computed as:

  $$p_{\text{teacher}} = \text{softmax}(z_{\text{teacher}}/T),$$

  where z_{teacher} are the logits (pre-softmax outputs) of the teacher model, and T is the temperature parameter.

- p_{student} is the softmax output of the student model, computed similarly:

  $$p_{\text{student}} = \text{softmax}(z_{\text{student}}/T).$$

- T is a temperature parameter that controls the smoothness of the output distributions. Higher values of T produce softer probability distributions, emphasizing the relationships between classes rather than the dominant class.

The KL divergence measures how one probability distribution diverges from another. In this case, it quantifies the difference between the teacher's and student's output distributions. Minimizing this loss encourages the student model to mimic the teacher's behavior, thereby preserving knowledge of previous tasks.

3..2 Role in Incremental Learning

In incremental learning, the teacher model is typically a copy of the student model before it is updated with new task data. During training, the student model is optimized to:

- Perform well on the new task (using the standard cross-entropy loss with ground truth labels).

- Maintain consistency with the teacher model's predictions on previous tasks (using the distillation loss).

The combined loss function is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \alpha \cdot \mathcal{L}_{\text{distill}},$$

where:

- $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss for the new task.
- α is a hyperparameter that controls the contribution of the distillation loss. In this project, $\alpha = 0.5$.

V. Performance Evaluation

The system's performance is evaluated using key metrics:

1. Task-by-Task Accuracy Matrix

The task accuracy matrix provides a comprehensive view of the model's performance on each task after learning subsequent tasks. The results are summarized below:

Task	After Task 1	After Task 2	After Task 3	After Task 4	After Task 5
Task 1	88.91%	84.13%	80.96%	79.25%	75.84%
Task 2	-	75.44%	70.49%	67.82%	61.18%
Task 3	-	-	59.86%	61.63%	52.91%
Task 4	-	-	-	64.22%	60.68%
Task 5	-	-	-	-	80.96%

Table 1: Task Accuracy Matrix

The matrix shows that the model maintains relatively stable performance on earlier tasks, with gradual degradation as new tasks are introduced.

2. Forgetting Measurements

Forgetting is quantified as the percentage decrease in accuracy for each task after learning subsequent tasks. The results are as follows:

Task	Forgetting (%)
Task 1	13.07%
Task 2	14.26%
Task 3	6.95%
Task 4	3.54%
Task 5	0.00%

Table 2: Forgetting by Task

The average forgetting across all tasks is 7.56%, which demonstrates the effectiveness of the combined memory rehearsal and knowledge distillation approach in mitigating catastrophic forgetting.

3. Task Performance Graphs

The task performance graphs illustrate the evolution of accuracy across tasks during incremental learning. Figure 1 shows the accuracy trends for each task after learning subsequent tasks.

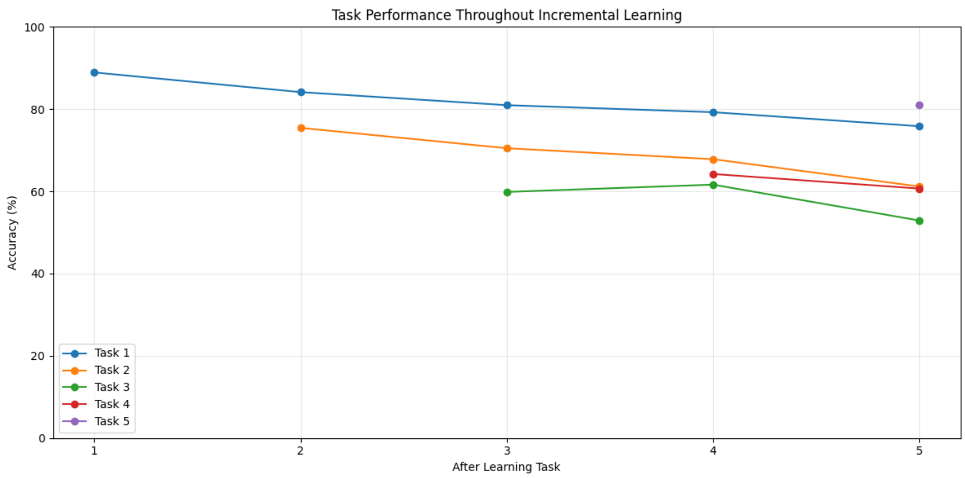


Figure 1: Task Performance Throughout Incremental Learning

The graph highlights the stability of the model's performance on earlier tasks, with only gradual degradation as new tasks are introduced. This is a key indicator of the success of the incremental learning approach.

4. Catastrophic Forgetting by Task

Figure 2 visualizes the forgetting rates for each task. The results show that Task 2 experiences the highest forgetting (14.26%), while Task 5 experiences no forgetting, as it is the most recently learned task.

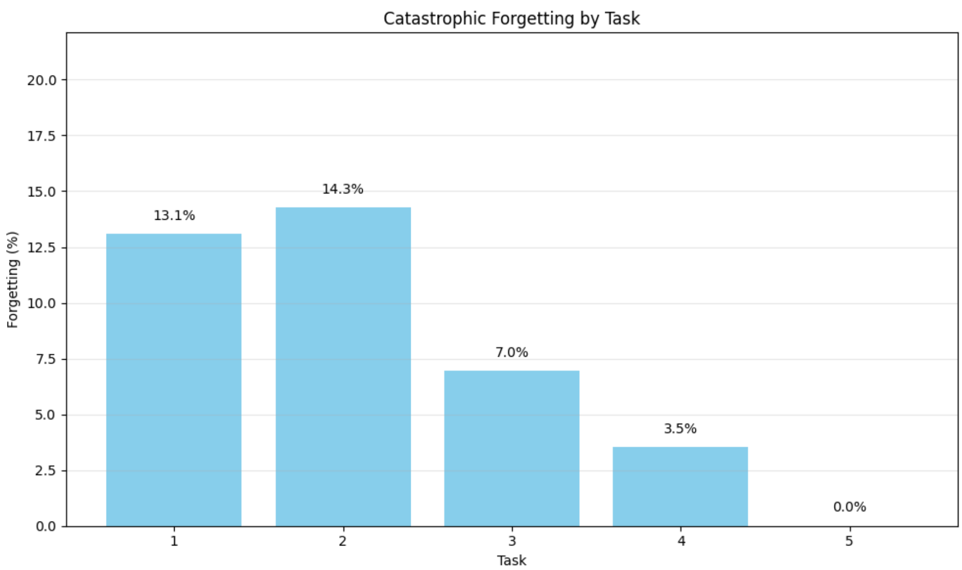


Figure 2: Catastrophic Forgetting by Task

VI. Results and Insights

The combination of memory rehearsal and knowledge distillation effectively mitigates catastrophic forgetting, as demonstrated by the following key insights:

- **Stable Performance Across Tasks**: The task accuracy matrix (Table 1) shows that the model maintains relatively stable performance on earlier tasks, with gradual degradation as new tasks are introduced. For example, the accuracy on Task 1 drops from 88.91% to 75.84% after learning Task 5, indicating a forgetting rate of 13.07%.
- **Low Forgetting Rates**: The forgetting measurements (Table 2) reveal that the average forgetting across all tasks is 7.56%, with Task 2 experiencing the highest forgetting (14.26%) and Task 5 experiencing no forgetting.
- **Visual Confirmation**: The task performance graph (Figure 1) and catastrophic forgetting visualization (Figure 2) provide clear evidence of the effectiveness of the combined approach in maintaining performance on earlier tasks while accommodating new knowledge.

Summary Metrics

The overall performance of the system can be summarized as follows:

- **Average Final Accuracy**: 66.31%
- **Average Forgetting**: 7.56%
- **Forgetting by Task**: 13.07%, 14.26%, 6.95%, 3.54%, 0.00%

These results highlight the success of the proposed approach in addressing the challenges of incremental learning for traffic sign recognition.
Future work could explore:

- **Parameter-Efficient Fine-Tuning**: Reducing memory requirements while maintaining performance.
- **Advanced Memory Management**: Developing more sophisticated strategies for sample selection and retention.
- **Self-Supervised Learning**: Enhancing feature representation for better generalization.

VII. Conclusion

The proposed incremental learning system for traffic sign recognition demonstrates the effectiveness of combining memory rehearsal and knowledge distillation in mitigating catastrophic forgetting. By maintaining a memory buffer of samples from previous tasks and leveraging knowledge distillation to preserve the model's behavior on old tasks, the system achieves stable performance across sequential tasks.

The overall performance underscores the potential of this approach for real-world applications like autonomous vehicles and smart city infrastructure, where models must continuously adapt to new data without losing previously acquired knowledge.