ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Institut für Integrierte Systeme
Integrated Systems Laboratory

Department of Information Technology and Electrical Engineering

Machine Learning on Microcontrollers

227-0155-00G

Exercise 7

Knowledge Distillation & Image Classification with STM CubeAI

Michele Magno, PhD
Marco Giordano
Pietro Bonazzi

1 Introduction

In the previous exercise, we created a project using *STM Cube IDE*, set up our environment, performed real-time inference on the MCU, and explored in-depth quantization and pruning methods—two crucial techniques for Machine Learning on Microcontrollers—and compared their performance results. In this exercise session, we will focus on **Knowledge Distillation (KD)**, another key model compression technique. Besides learning how to train models using KD, we will conduct live inference on our `B-L475E-IOT01A2` board using TensorFlow Lite, similar to our previous session. Additionally, we will explore how to generate optimized code from our *.tflite* or *.h5* models for our board using STM Cube AI, and we will compare outputs across different deployment scenarios.

2 Notation

Student Task: Parts of the exercise that require you to complete a task will be explained in a shaded box like this.

Note: You find notes and remarks in boxes like this one.

3 Knowledge Distillation

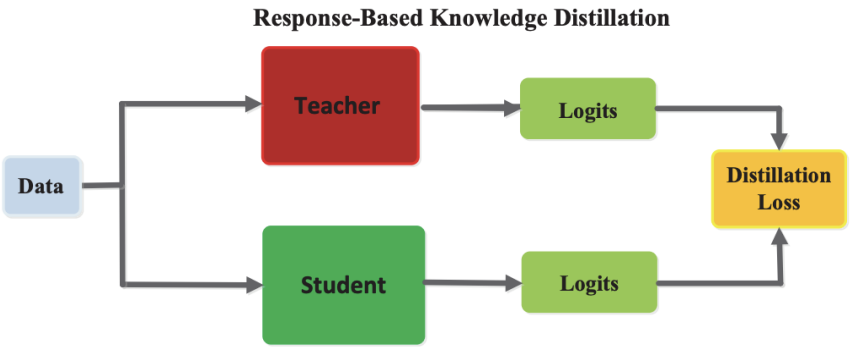

Figure 1: Response Based Knowledge Distillation

Knowledge Distillation (KD) is a powerful model compression technique used to transfer knowledge from a large, high-performing model (often referred to as the ”teacher” model) to a smaller, resource-efficient model (the ”student” model). This approach enables the smaller model to achieve higher performance than it would typically attain through direct training on the data alone.

The main concept behind KD involves leveraging the outputs of the teacher model, typically the soft probabilities (logits), as an additional training signal for the student model. By doing this, the student not only learns directly from the true labels but also captures the nuanced decision-making patterns of the teacher model, including subtle distinctions between classes that may not be evident from hard labels alone.

Practically, Knowledge Distillation is implemented by introducing a softening factor, called the "temperature," (T) which smooths the probability distribution output by the teacher.

$$p(z_i, T) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \tag{1}$$

A higher temperature leads to softer probability distributions, making it easier for the student to learn nuanced relationships between classes. The student model is then trained using a combination of two loss functions: the traditional classification loss based on ground truth labels and a distillation loss based on the softened outputs of the teacher. The global loss is defined with (2).

$$\mathcal{L}_{\text{global}} = (1-\lambda)\,\mathcal{L}_{\text{CE}}\big(\psi(Z_s), y\big) \;+\; \lambda\,\tau^2\,\text{KL}\left(\psi\left(\frac{Z_s}{\tau}\right), \psi\left(\frac{Z_t}{\tau}\right)\right) \tag{2}$$

KD offers several advantages:

1. **Improved Model Efficiency**: Allows deployment of smaller models suitable for microcontrollers or edge devices with limited resources.
2. **Enhanced Generalization**: Helps the student model generalize better by providing rich information from the teacher's learned representations.
3. **Reduced Inference Time**: Smaller models result in faster inference, essential for real-time applications on constrained hardware platforms.

In this exercise, we will apply *Response Based Knowledge Distillation* to train optimized student models, deploy them on our STM32-based hardware platform, and evaluate their performance in real-time inference scenarios. Additionally, we will leverage STM Cube AI and TensorFlow Lite to generate optimized code and compare performance across different deployment strategies.

Note: In knowledge distillation, softmax outputs can be used instead of logits. In this exercise, the teacher and student models include a softmax activation in their final layers, enabling a direct comparison of their outputs. If you would like to use knowledge distillation in your projects, you could also experiment by training teacher models without a softmax in their final layer and compare the results accordingly.

4 Preparation

For this exercise, we will first train a model using Jupyter Notebook. The required TensorFlow version is 2.13.0, please check if you have a compatible version. If not, **please use 'pip install' when you install the correct version, not 'conda install'**. Next, we will briefly train another model and compare it to a previously trained teacher model. Then, keeping the initial model's architecture unchanged, we will retrain our model using knowledge distillation with the teacher model. Since knowledge distillation leverages the teacher model's generalization capabilities to improve the student's generalization, it typically demonstrates greater effectiveness with longer training periods. Therefore, we provide you with the final version of the knowledge-distilled model as pre-trained. Finally, we will compare all models and perform live inference on our STM board using the best-performing model.

5 Training a Convolutional Neural Network

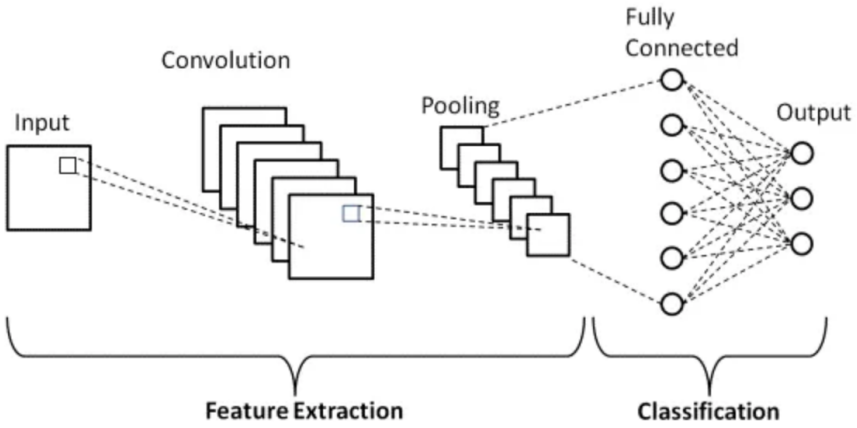

Figure 2: Basic CNN Architecture

A Convolutional Neural Network (CNN) is a type of deep learning model particularly suited for analyzing visual data such as images and videos. CNNs leverage spatial hierarchies through multiple layers of convolution operations, pooling, and fully connected layers to automatically detect meaningful patterns and features in data. The convolution layers apply filters to the input images, enabling the model to recognize features such as edges, textures, or shapes, while pooling layers help reduce spatial dimensions and computational load. CNNs have been widely used in various computer vision tasks, including image classification, object detection, and image segmentation, due to their ability to capture complex spatial structures efficiently.

For the first task, we will train a CNN model using the CIFAR-10 dataset. CIFAR-10 is a standard image classification dataset consisting of 60,000 colored images divided into 10 distinct classes (e.g., airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck). Each image has a resolution of 32x32 pixels. This dataset is widely used as a benchmark for training and evaluating computer vision algorithms and CNN models.

Student Task 1 (Training a CNN):

1. Open the provided *Jupyter Notebook* file and check if you have the correct version of TensorFlow.
2. Go to **Task 1** in the notebook and do all the steps provided.
3. How many parameters does the model have? _____
4. After the model training, what is the
   - Validation accuracy: _____
   - Test accuracy: _____

6 Quantizing the CNN

Last week, we conducted an in-depth analysis of quantization techniques. In this exercise, we will use one of the techniques examined last week: the Post-Training Integer Quantization method.

As a reminder, **Post-Training Quantization (PTQ)** is a technique used to optimize deep learning models by reducing their size and computational requirements while maintaining acceptable accuracy. PTQ is applied after training, converting model weights and activations from floating-point precision (e.g., FP32) to lower-bit representations such as INT8. Post-Training Integer Quantization specifically ensures that both weights and activations are fully quantized to integer values, making it highly efficient for deployment on hardware with limited computational resources, such as microcontrollers and edge devices. This technique significantly improves inference speed and reduces memory usage while enabling compatibility with specialized integer-based accelerators.

Student Task 2 (Quantization of the Model):

1. Go to **Task 2** in the notebook and do all the steps provided.
2. How did the accuracy of the model change after applying the post training quantization?
3. Check the number of parameters in the model. Make a guess if the non-quantized model can fit on your microcontroller. _____
   Hint: Check your microcontroller's datasheet for its memory capacity.
4. Does the quantized network fit into your microcontroller? Calculate it. _____

7 Interpretation of the Teacher Model

In knowledge distillation, a *teacher model* is a larger and more well-trained model used to guide the *student model* (the smaller one). The teacher model can have a different architecture, be trained using different methods, and is generally a well-generalized model for the given dataset. Since the teacher model's generalization ability is transferred to the student, the student learns not only from the original ground truth labels but also from the teacher's predictions, allowing it to benefit from the teacher's learned representations.

However, the **teacher model is not a perfect model**—it has learned from the dataset and retains a certain level of error, meaning it does not always produce values that exactly match the ground truth data. This also means that the teacher model transfers some of its errors to the student. Interestingly, when the teacher is a well-generalized model, **this imperfection can actually improve the student model's generalization**. By learning from the teacher's softened predictions rather than hard labels alone, the student can develop a more robust understanding of the dataset, often leading to better generalization and improved performance in real-world scenarios.

Because training the teacher model may last too long, in this exercise, we provide you with a *pre-trained teacher model* (teacher.h5), which has been trained on the CIFAR-10 dataset and achieves approximately 90% accuracy (The higher the teacher provides accuracy, the better the student model generalizes the data). This teacher model also follows a CNN architecture but is significantly larger and more complex than the student model. Its higher capacity allows it to generalize the dataset better, making it a strong reference for guiding the student model during knowledge distillation.

Student Task 3 (Interpretation of the Teacher Model):

1. Go to **Task 3** in the notebook and do all the steps provided.
2. How many parameters does the teacher model have? _____
3. Teacher model
   - Test accuracy: _____
   - Test loss: _____
4. How did the accuracy of the teacher model changed after quantization? _____
5. Do you think the non-quantized or the quantized teacher model fits on your microcontroller?
6. Visualize and compare the teacher model and small model. Follow the *.summary()* outputs as well.

8 Knowledge Distillation with the Teacher Model

So far, we have trained a small CNN network and compared its results with a previously trained larger model. Now, we will retrain our small network, but this time, we will redefine the loss function according to the distillation loss, as given in (2).

In *knowledge distillation*, two key parameters influence the process: **alpha** and **temperature**.

- Alpha (α) determines the balance between the student model's standard loss and the distillation loss. It controls how much the student model learns from the ground truth labels versus how much it learns from the teacher's predictions.
- Temperature (τ) affects how **softened** the outputs of the teacher and student models are. Higher temperature values produce softer probability distributions, allowing the student to capture more nuanced relationships between classes.

All these definitions and operations have been implemented in **Task 4** within the **Distiller class**.

Student Task 4 (Interpretation of the Teacher Model):

1. Go to **Task 4** in the notebook and read the explanation about **Distiller class**.
2. Try to match the mathematical definition of softmax with temperature (1) and distillation loss definition (2) in the Distiller class. And then, turn back to this sheet.

Now, we will train the model using knowledge distillation while also analyzing the impact of the selected parameters on the training process. By adjusting alpha (α) and temperature (T), we can observe how they influence the learning dynamics of the student model. Specifically, we will evaluate how different values of α affect the balance between direct supervision and distillation, and how varying T impacts the softness of the teacher's predictions. This analysis will help us understand the trade-offs involved in knowledge distillation and optimize the training process for better generalization.

Student Task 5 (Parameter Selection for Knowledge Distillation):

1. Set (*kd_alpha*, *kd_temperature*) as they are given below. First, please make a prediction about the parameters' effects, and then run the training to observe their effects. Compare especially losses of the training sessions (X is don't care).
   - (*kd_alpha*: X, *kd_temperature*: 0) _____
   - (*kd_alpha*: 1, *kd_temperature*: 1) _____
   - (*kd_alpha*: 0, *kd_temperature*: 1) _____

In this exercise, we have conducted an analysis using **Task 4** with different hyperparameter choices to keep the training time short while still observing the effects of knowledge distillation. As seen, even though the **student model's loss is not directly calculated from the ground truth labels**, it can still be trained using the **teacher's predictions**. However, finding the right student-teacher balance requires careful hyperparameter selection and longer training periods. Additionally, although knowledge distillation can be performed directly using the teacher model's logits, the teacher model used in this exercise includes a softmax layer at the end. This means that instead of using raw logits, the student learns from the **softened probability outputs** of the teacher.

Common values for alpha (α) range from 0.3 to 0.7, where lower values emphasize the distillation loss. A typical choice may be $\alpha = 0.5$, which balances both.

For temperature (T), values typically range from 2 to 6, with higher values producing softer probability distributions, making it easier for the student to learn subtle relationships between classes.

To observe the impact of knowledge distillation on the test set, we provide you with a pre-trained model, *'knowledge_distilled_model_final.h5'*, which follows the **same architecture** as the model used in this exercise but has been trained for a much longer period to maximize distillation effectiveness, using $\alpha = 0.4$ and $T = 3$.

Student Task 6 (Final Knowledge Distilled Model):

1. Go to **Task 5** in the notebook and do all the steps provided.
2. How many parameters does the provided student model have? Is it the same with the previous small model? _____
3. Final model
   - Test accuracy: _____
   - Test loss: _____
4. How did the accuracy of the final model changed after quantization? _____
5. With the inference time comparison, _____ model was _____ times faster than the _____ model.

9 Real-time Inference on Microcontroller with the Student Model

At the final stage, we conducted a teacher vs. student speed comparison, where results may vary depending on the computer used. However, this same comparison could not be performed on an MCU because the teacher model is too large to fit into the microcontroller's memory. Instead, the student model was specifically trained using knowledge distillation to be small enough to fit on the MCU while still achieving performance close to the teacher model.

At this stage, we will verify our trained model on the MCU and perform live inference using the provided code that communicates with the board via UART, *live_inference.py*. The codes takes your camera input and resizes it to 32x32 image before sending it to the MCU. After MCU runs inference on the camera frame, it reads the detected class and inference time that comes from MCU and prints it to the terminal. Now, plug in your MCU to your computer.

As we did in previous exercise sessions, follow these steps to set up the project for the B-L475E-IOT01A2 development board:

- Create a new project selecting your board. Make sure that target language is C.
- Reset the pinout configuration (Shortcut: CTRL + P).
- In **Middleware and Software Packages**, ensure that X-CUBE-AI is installed. If it is not installed on your system, download and install the latest version.
- Open the settings interface under **Software Packs** $\longrightarrow$ **X-CUBE-AI** and verify the Core option is enabled and the Application option is set to Validation.
- Go to Middlewares and Software Packages, click on X-CUBE-AI, and confirm that this module is activated.
- In the Add Network section, navigate to the directory containing the trained models and proceed with the following task.

Student Task 7 (Verification of the Model on the Microcontroller):

1. After selecting the proper model type (Keras or TFLite) try to **Analyze** all the models below. Try different **compression and optimization** selections while you are trying to analyze them for the microcontroller. Note your observations about the models.
   - Teacher Model (FP32) _____
   - Teacher Model (INT8) _____
   - Student Model (FP32) _____
   - Student Model (INT8) _____
2. Select the one that fits into your MCU and run **Validate on Target**. From the generated report, calculate the MACC/cycle and observe the inference time while the MCU runs @ 80 MHZ. _____

After analyzing and validating the model on the MCU, you are ready to run real-time inference on the microcontroller. To generate the code using the IDE, go to **Project Manager**, write your project name (if it is empty). Then, name your network as **cifar_model** (to be compatible with the provided main.c) in the part shown in 3. Hit *CTRL + S* or *Generate Code* button to generate your optimized code for the model.

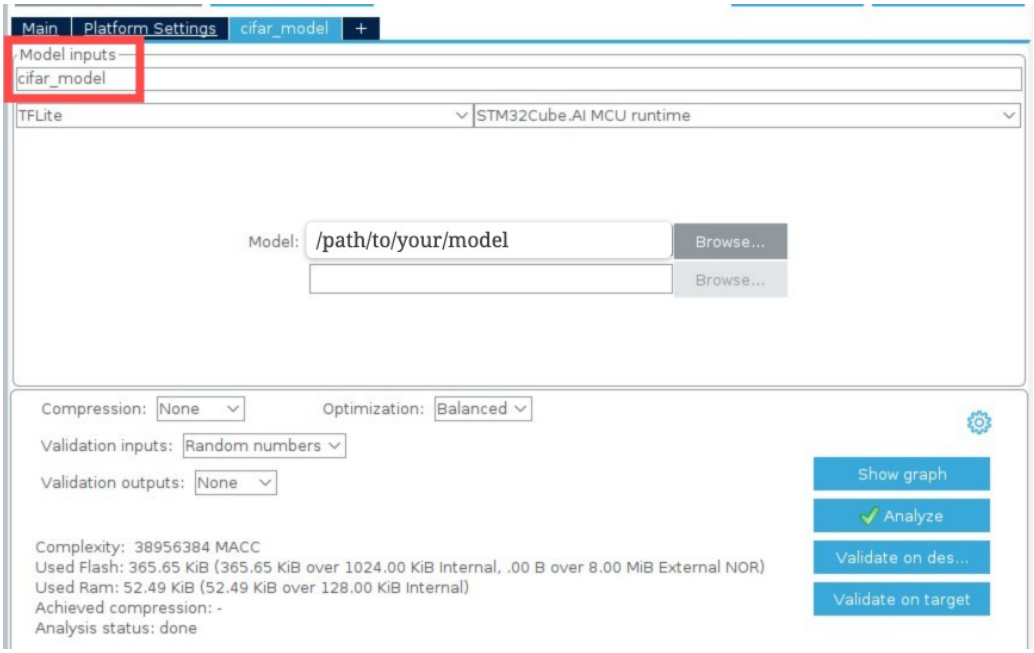

Figure 3: Generating the Code from the Project File

Before starting the task, check your inclusions are the same with 4, include directories may change depending on your OS and project folder, but you should see *Drivers, Middlewares, and X-CUBE-AI* in your included directories.

Figure 4: Project Include Directories

Note: After the code generation, you may face some build or compilation problems if your X-CUBE-AI version is not compatible with the given main.c code. Here, there are some possible issues that you may face and solutions for them:

- If you do not have an automatically generated **Middlewares** folder in your project, copy it from **provided_packages** folder that you have downloaded with this exercise.
- To disable X-CUBE-AI logs, go to *X-CUBE-AI/App/aiTestUtility.c* and find *void lc_print(const char* fmt, ...)* function and directly return it, or change the generated X-CUBE-AI with the provided one in **provided_packages** folder.
- If your IDE is generating seperate .c/.h files for your activated peripherals, you will see errors related to multiple definitions. Go to uart.c (if you have) and comment the lines related to **huart1 and USART1 initialization** since they are already defined in main.c.

If you face any other problem, you can ask for help.

Now, we have to fill the *main.c* file for real-time inference. Copy the provided main.c into the auto-generated one.

Student Task 8 (Real-time Inference on Microcontroller):

1. Go to the directory of your generated code and open the code. Open the generated *main.c* and replace it with the file that is given with this exercise. Familiarize yourself with the code. Fill the *TODO* part in the given main.c.
2. Check and learn the port that your computer communicates with your MCU. And then, run the *live_inference.py* code giving the communication port from the terminal, e.g python3 live_inference.py /dev/ttyUSB0. Now, your camera input is read with this Python code and sent to the MCU. Reset your MCU using its reset button, then you will see a message in your terminal "CIFAR10 Inference Demo is Started" after successful initialization. You may open images corresponding to CIFAR-10 classes on your phone and hold your phone close to your camera to test the model's inference performance.
3. What is the inference time on your MCU? Is it consistent with the STM generated report after the validation on target? _____
4. You have the C++ template from previous weeks for running live inference on MCU. Open up the project again and use the *quantized_final_model.h* as the model. Change app.cpp and make your code available to be used with live_inference.py and run it again (You may also need to increase your tensor size in this code). Compare the inference times of TFLite and X-CUBE-AI implementations. Which one is faster? What do you think causes the difference? _____

$\mathcal{E}$ **Congratulations! You have reached the end of the exercise.**
If you are unsure of your results, discuss them with an assistant. $\mathcal{E}$