Embedded Machine Learning - SS 2024 Example Exam

Bernhard Klein and Holger Fröning, ZITI, Heidelberg University, 01.08.2024

Rules and remarks:

- The duration of the exam is 90 minutes and there are 90 points total.
- As a rough rule, the amount of points indicates the amount of time in minutes for a question. For instance, 3 points for one question indicate a processing time of 3 minutes. You might want to use this as a hint.
- Use the paper provided for the answers. Green paper is for notes, only the white paper will be considered for grading. More paper sheets can be provided upon request.
- If a question is unclear, ask by raising your hand.
- When providing code examples: we are not a compiler, i.e. we don't care about parameter order, minor mistakes etc. However, the code in principle (as pseudo-code) should be correct.
- You can provide the answers in English or in German.
- Turn off your cell phones during the exam.
- No additional material is allowed, all necessary information is provided.
- No electronic devices except for a non-programmable standalone calculator are allowed.
- No exchange of notes or information between participants is allowed.
- Any attempt to disregard these rules will result in a "not passed" mark.

Important: check the exam for completeness (? pages total)

Candidate data:

Name: ______________________________

Student ID (Matrikelnummer): ______________________________

Course of Studies: ______________________________
(Studiengang - MScTI, MScDACS, BScI, BScPhysics, MScPhysics, ...)

Versicherung/ Affirmation 1

Ich versichere die vorliegende Prüfungsleistung selbstständig verfasst und keine anderen als die erlaubten Hilfsmittel benutzt zu haben. Ich habe während der Bearbeitung der Aufgaben in keinerlei Weise mit anderen Personen oder elektronischen Medien über die Aufgaben, mögliche Lösungen und ähnliches kommuniziert. Bei Abgabe einer unwahren Versicherung wird die Prüfung mit "nicht bestanden" bewertet.

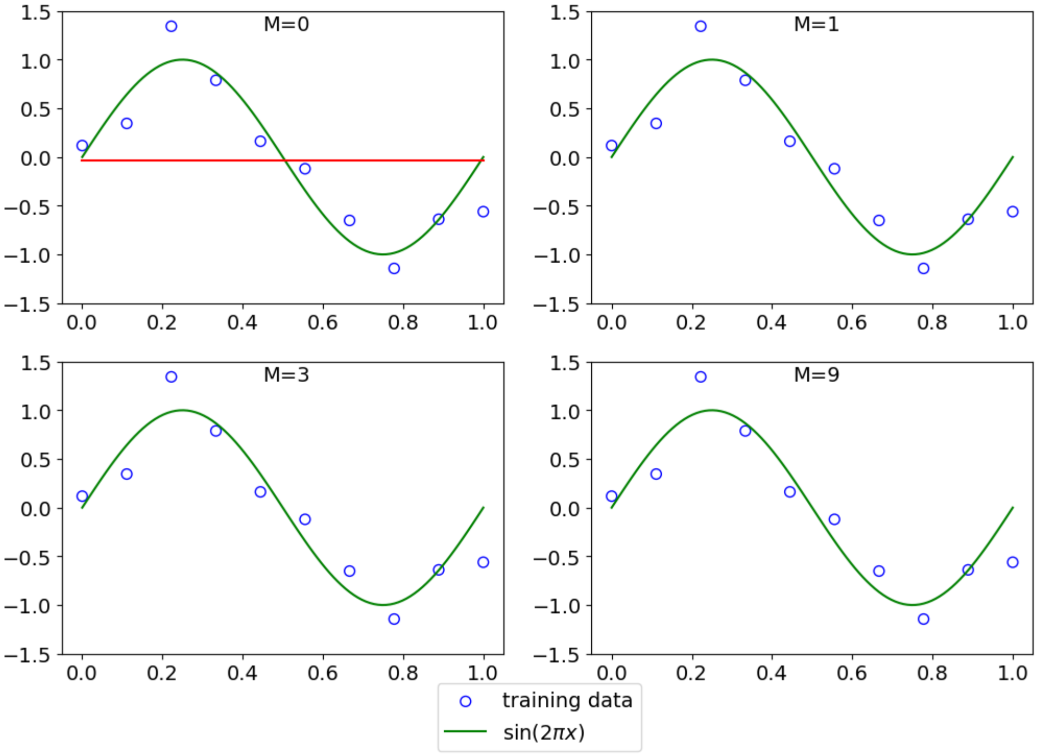
I confirm that I have written this examination independently and have not used any aids other than those permitted. I have not communicated in any way with other persons or electronic media about the tasks, possible solutions and the like while working on the tasks. If an untrue statement is submitted, the examination will be graded as "failed".

____
Date, Signature

I. Basics of Machine Learning (30 points)

1. Fitting and model capacity (10P)

The following figure shows (noisy) measurements in blue, sampled from a ground truth function shown in green. Consider the use of different polynomial-curve-based models of order M.



1. Complete the sketch (you can draw directly into it) by adding estimations of how the fitted function would look like for the other model orders. An example fit is shown for M=0 as a red curve.
2. Explain under/overfitting on this example.
3. What methods could help to overcome overfitting. Discuss one method in detail.

2. Compute Intense Layers (10P)

Name and describe the two major compute intense operations in neural networks used for image classifications, use sketches where appropriate. Compare them with respect to their computational complexity. Discuss requirements, advantages and disadvantages of both operators. Discuss use cases for both operators; when should you prefer which one?

3. Regularization (10P)

1. What is the primary goal of regularization in machine learning?
2. How does L1 regularization differ from L2 regularization? What are the trade-offs between these two approaches?
3. Discuss other regularization approaches which are not based on a penalty term in the loss function.

II. Model Compression (30 points)

4. Trained Tenary Quantization (TTQ) (10P)

1. Describe the different steps in trained ternary quantization. How does it affect the backpropagation of the network?
2. In which scope does the quantization occur and what is the reason?
3. How may the hyperparameter t affect the performance of a neural network regarding accuracy and execution speed?

5. Explain Inception (10P)

Inception blocks have become a fundamental component in many deep neural network architectures. Answer the following questions:

1. Explain the concept of an Inception block. Provide an overview of what an Inception block is, its structure, its key components, and the motivation behind its design in neural networks.
2. What are the advantages of using Inception blocks in deep neural networks? Discuss the benefits of incorporating Inception blocks, such as improved model performance and computational efficiency, in comparison to traditional architectures.
3. What is the role of 1x1 convolutions in an Inception block? Elaborate on the significance of 1x1 convolutions and why this is essential in Inception blocks.

6. Grouped Convolutions (10P)

Explain what grouped convolutions are. Use sketches where appropriate to explain the concept. What are the advantages and disadvantages of grouped convolutions? How do they compare to non-grouped convolutions in classical cost metrics (MACs, Parameters, Activations)?

III. Advanced topics (30+5 points)

7. End-to-End Use of Convolutional Neural Networks (CNNs) for Image Classification (15P for 1-3, 20P for 1-4)

You have been tasked with developing an end-to-end system for classifying images of different species of flowers using a Convolutional Neural Network (CNN). Input images are 50x50 pixels and have 3 RGB channels.

1. **Data Preparation:** Describe the steps you would take to prepare a dataset of flower images for training a CNN. Include details on data collection, preprocessing, and augmentation techniques.
2. **CNN Architecture Design**: Design a CNN architecture suitable for this image classification task. Explain the choice of layers, kernel sizes, skip connections, activation functions, and any regularization techniques you would employ.
3. **Training Process**: Outline the process of training the CNN on the prepared dataset. Discuss the choice of loss function, optimizer, learning rate, and any strategies you would use to prevent overfitting.
4. **Evaluation and Testing**: After training, describe how you would evaluate the performance of your CNN model. What metrics would you use, and how would you interpret them? Include a discussion on how you would handle a situation where the model performs well on the training set but poorly on the validation set.

8. Low-precision operations (15P)

1. Introduce the BOPS metric
2. Conceptually, how would you calculate BOPS for floating-point operations?
3. Given an exemplary processor, draw the roofline for different data types (INT8, FP4, FP16, FP32). While the absolute numbers are of no importance here, please ensure that relative numbers are in line with your experience with the performance of such low-precision operations.