EINDHOVEN UNIVERSITY OF TECHNOLOGY

3EBX0

Machine learning in science

Assignment 3

Hydraulic cross sections

Teacher: dr. A. Corbetta (a.corbetta@tue.nl)

Deadline: 23-06-2024 23:59

Reports/codes uploaded on canvas must state the following in the beginning: students names, id, group name and group number. The report is expected in pdf format and should include the code. Runnable codes must also be uploaded. In case these are missing, the final grade will be reduced by 0.5/10.

1 Introduction

We consider the laminar flow of a viscous fluid through a porous medium, like a "sponge". The porous medium extends along the interior of a long channel (length L). The fluid moves under the action of a pressure drop (ΔP). The system is sketched in Figure 1, which also highlights a sample cross-section (colors representing the magnitude of the flow rate).

The cross-sections of the porous medium come in a wide variety of shapes (see examples in Figure 2). We represent cross-sections with 40×40 binary images in which light pixels (pixel value = 1, "open pixels") indicate locations through which the liquid can flow (in the direction perpendicular to the figure). On the opposite, dark pixels (pixel value = 0, "closed pixels") are impermeable to flow.

Given a cross-section (i.e. a binary image), a crucial challenge in applied porous media physics is to quantify the associated laminar[1] flow rate:

- Q (in SI units measured in $m^3 s^{-1}$).

The flow rate is determined by an interplay of the following variables (listed with SI units):

[1] i.e. non-turbulent

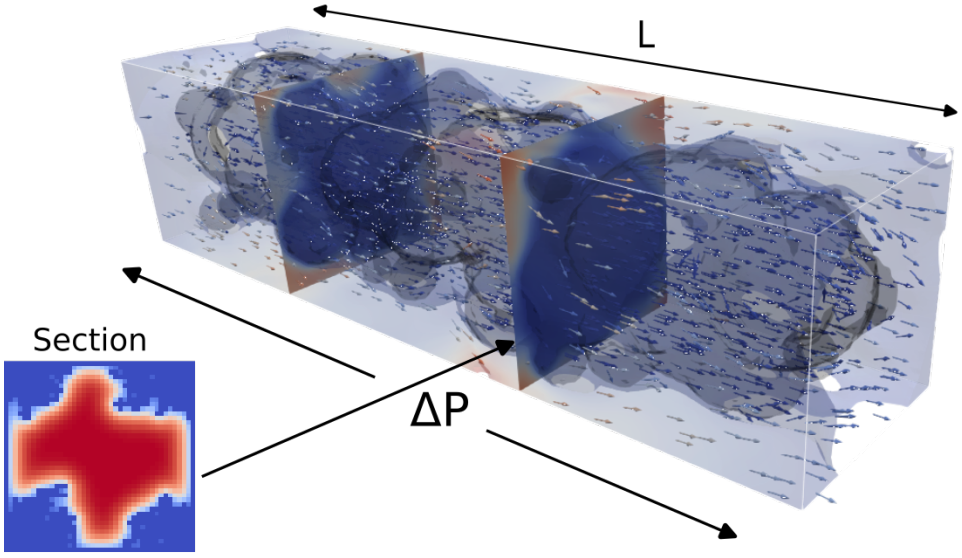

Figure 1: Pressure driven flow through a porous media; the panel of the bottom left shown an example of a section of the domain. The color map represents the magnitude of the flow rate.

- the pressure drop ΔP over the channel (in Pa)
- the length L of the channel (in m)
- the fluid viscosity μ (in $Pa \cdot s$)
- the area A of individual pixels in the binary image (measured in m^2)

It turns out that scaling the flow rate, Q, by the ratio of pressure gradient $\Delta P/L$ and viscosity μ, i.e. considering the **hydraulic throughput**,

$$S = \frac{\mu L Q}{\Delta P}, \tag{1}$$

yields a geometric parameter (measured in m^4) that is **independent on the fluid properties, the channel length, and the pressure drop**. On the opposite S grows with the pixel area (i.e. the overall area of the section).

In this assignment, you are tasked to design and train a (convolutional) feedforward network that takes as input a 40×40 image, and predicts the corresponding hydraulic throughput S. All the symmetries of the problem (invariances/equivariances) can be hardwired in the network.

For training, you are provided with a total of 1660 binary images, each accompanied by their corresponding hydraulic throughput S.

Note: In the images, measurement units are chosen such that the area, A, of individual pixels equals unity.

For verification purposes, you are also provided with 500 binary images for which the individual S values are kept hidden (the test data). This test data also utilizes a unit pixel size. This data set is further split: 40% of the images constitute the public test data (public leaderboard data), and the rest constitutes the private test data.

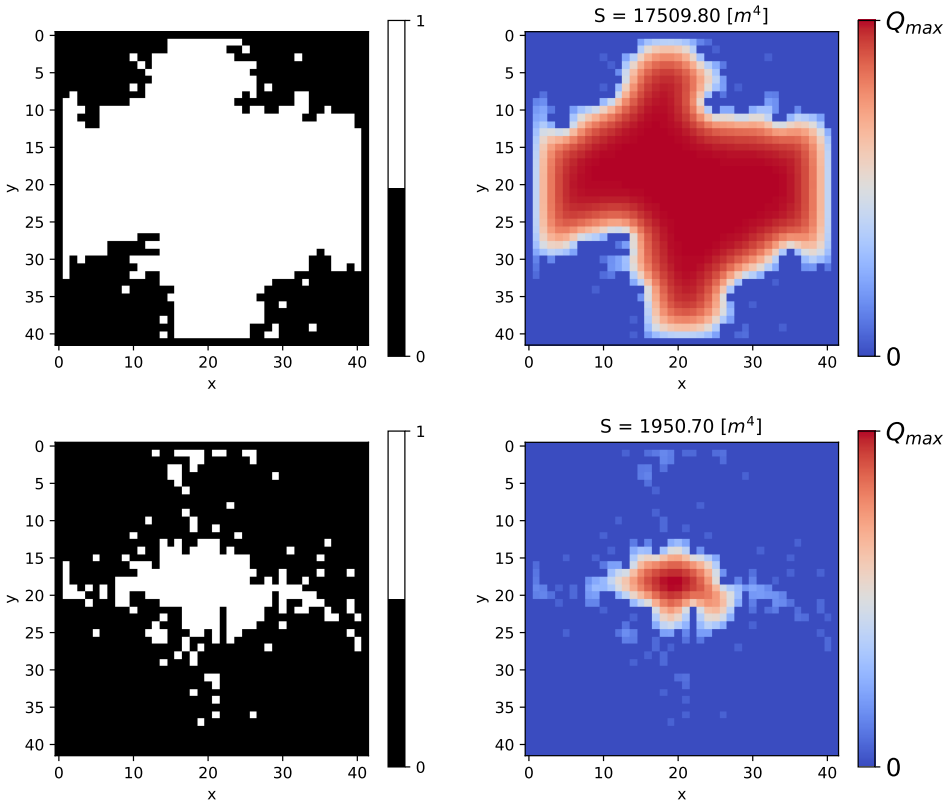

Figure 2: Examples of a cross-sections. In the left column you are given examples of the binary matrices, with "open" (= 1) and "closed" (= 0) pixels. The right column shows the corresponding flow profile. The hydraulic throughput for the first example (top) is $S = 17509$ m^4, while in the second example (bottom) the hydraulic throughput is $S = 1950$ m^4.

By submitting to Kaggle you get feedback on the root-mean-square relative error (RMSPE in Kaggle) over the public test data:

$$RMSPE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{S^{(i)}-T^{(i)}}{T^{(i)}}\right)^2} \tag{2}$$

Part of your grade will be based on the RMSPE error obtained over the private test data (via Kaggle).

1.1 Important Remarks

1. Your training data have inputs in m^2 and output in m^4.
2. You are asked to create neural networks that operate on input data with physical dimension m^2. I.e. any dimension-wise data operation shall only change the output data. The Kaggle **test submission** expects values in m^2 as well.
3. The system has a number of symmetries: all can be hardwired. Please extensively reflect on the symmetry present and how to embed them in the neural network.
4. The training time for this problem is going to be substantially longer than in previous

assignments. Please take this into account and start early. You might want to start with non-augmented data and train with augmented data only your best performing models.

5. We advise to train your networks with the loss in Eq. 2. This loss is not natively available in PyTorch and needs to be implemented manually. The following code implements the loss, use this as your loss function instead of one of the standard functions from `torch.nn`.

```
import torch

def rmspe(y_pred, y_true):
    return (((y_true - y_pred) / y_true) ** 2).mean().sqrt()
```

   As always, make sure that the `y_true`, and `y_pred` vectors have the same shape.

6. No computations of the hydraulic throughput with numerical methods different from a feedforward neural network designed on purpose are admitted.

7. Please start your report with names, student numbers, group name and group number.

A challenge for more advanced python/pytorch users

It is possible, and avised to first-time pytorch users, to solve this assignment **hardwiring** the various symmetries mixing different strategies.
Yet, all the symmetries of this problem can be hardwired in a feedforward neural network exploiting the architecture of CNNs layers beyond their traditional spatially-equivariant purpose. How?
You are welcome to solve the assignment using (also) this strategy and compare results.

AI/ChatGPT usage

While the usage of AI aids is not forbidden, it is however discouraged. Your report must include a statement of which of its parts have been done using the help of AI/GPT tools and how. The responsibility of the final content is in any case on the student.

As importantly, the objective of this assignment is understanding the underlying theory and implementation thereof. Hence, the teachers reserve themselves the right of discussing with the students any components of the report and possibly revise the final mark based on such a discussion.

Kaggle competition links

- Competition link: `https://www.kaggle.com/competitions/hydraulic-cross-sections-24`
- Private key: `https://www.kaggle.com/t/425864ca4d9641219f3877bd3499a030`

2 Grading rubric

You shall aim at minimizing the root-mean-square relative error (RMSPE) in the estimates for the hydraulic throughput S. The maximum grade that can be achieved is 20. Your grade is determined by your performance on the following elements:
A - Performance (over private test set)

- 1 pt for Kaggle submission yielding RMSPE ≤ 0.10, or
- 2 pts for Kaggle submission yielding RMSPE ≤ 0.065, or
- 3 pts for Kaggle submission yielding RMSPE ≤ 0.03, or
- 4 pts for Kaggle submission yielding RMSPE ≤ 0.02, or
- 5 pts for Kaggle submission yielding RMSPE ≤ 0.01, or
- 6 pts for Kaggle submission yielding RMSPE ≤ 0.005.

B - Low complexity

- (1 pt) Kaggle submissions with RMSPE ≤ 0.01 can have a further single point for using a total number of tunable parameters smaller than 20.000.

  Include the output of the function `model.summary()` to be eligible for this point.

Implementation and report

- C - (2 pts) Analysis of physical dimensions, data ranges, and symmetries of the problem. Perform a preliminary analysis of the problem and data you received and elaborate on these aspects.
- D - (4 pts) Reflect on the symmetry group of the problem. Discuss strategies for data augmentation and disambiguation. Discuss how all symmetries of the problem can be hardwired in a network.
- E - (3 pts) Design and train a neural network to estimate the hydraulic throughput (Eq. 1). Elaborate on your architectural choices. Discuss computational efficiency and physical properties.
- F - (1 pt) Show that your neural network fits (and does not overfit) the function we are modeling, and comment on the rationale of your answer.
- G - (1 pt) Code cleanliness and clarity. Present a clear and commented code. Avoid code repetitions and strive for efficiency.
- H - (2 pts) Report quality in terms of clarity, consistency, layout and language.

The report should be clear, self-contained, and comprehensive. The report alone should give a colleague student the opportunity to reproduce all results obtained. Ensure you elaborate on all choices made: what considerations and what experimentation led you to this particular choice of feedforward net architecture, how did you ensure dimensional homogeneity, what experimentation led you to fix the parameters in the network training process, what measures did you put in place to avoid overtraining. Don't hold back in listing considerations, this report is your opportunity to illustrate what you have learned during this course! Details matter: make sure your figures are annotated (no graph axis without label, no figure without caption, etc.). Whether to present data in the form of a table or a graph should be a concious decision.

Please make sure you checked the *Important Remarks* sections carefully.