Wildcard

Nijat Rustamov

Introduction

An accurate understanding of transport mechanisms in nano-confined systems is critical for many applications relevant to energy and sustainability technologies. However, the intricate physics governing the gas flow in confined media challenges the scientific efforts to bridge the gap between continuum and free molecular flow scales. In an effort to mitigate this problem a highly efficient numerical simulation model powered by the lattice Boltzmann method was developed and optimized capable of undertaking extremely large and complex porous media with a wide range of Knudsen number. However, as for any other numerical methods, simulation of large number of samples is restricted by computational time. In this course, I decided to start exploring and implementing what I have learned so far and apply it to the domain of my interest. In this exercise and several subsequent ones, I generate numerous artificial porous media, run the numerical simulations, and try to formulate the problem as a machine learning task to predict the flow behavior.

The numerical simulations are based on the continuous Boltzmann equation given as

$$\frac{\partial f}{\partial t} + \vec{\xi} \cdot \vec{\nabla} f + \frac{\vec{F}}{m} \cdot \vec{\nabla} f = \Omega(f) \tag{1}$$

Where f represent particle distribution function in space and time. The left-hand side describes the movement of particles in space and time and the right-hand side describes the collision dynamics. The discretized multi-relaxation time Boltzmann equation is given by.

$$f_\alpha(x + ce_\alpha \delta t, t + \delta t) = f_\alpha^{eq} + \widetilde{f_\alpha} - \sum_\beta (\boldsymbol{M}^{-1}\boldsymbol{S}\boldsymbol{M})_{\alpha,\beta} \widetilde{f_\beta} + \delta t F_\alpha(x, t) \tag{2}$$

The details of what each term stands for can be found in [2]. With proper boundary conditions the method is capable of simulating flows in confined (nanoscale) media. In this work, Knudsen numbers are used to represent the scale. Knudsen number is the ratio of mean free path of the fluid to the representative pore diameter. In this work, fluid is methane gas flowing through the pores as shown in Figure 1 in Appendix. At high Knudsen numbers the flow goes into slip, transitional and free molecular flow regimes where the slip velocity cannot be neglected.

Dataset description

300 artificial porous media were generated by randomly placing obstacles in the 500x500 domains. Then each of those domains were skeletonized to calculate local pore sizes and eventually Knudsen number distributions which are input to the numerical simulation. Numerical simulation as briefly described in the introduction section is written in C++ and is outside the scope of this work to discuss in detail. However, references [1, 2] are our publications for interested readers. 300 simulations were generated, cleaned, processed, and run over the

period of 2 weeks to generate the data. The output of the simulations are x and y direction velocity distributions as well as density distributions. For simplicity, I am only interested in x direction velocity, since only x component of driving force in the simulation is kept on. Furthermore, to gain time advantage, intermolecular forces are turned off, so the density distributions are also not considered here. In Figure 2, examples for the input and the output of numerical simulation can be found. In the end, velocity from each distribution is averaged to represent this problem as a multi-input regression problem. The distribution of these mean values across all 300 samples are shown in Figure 3.

Experimental Setup

All 300 input images are down sampled to 200x200 to reduce the computational cost. 80% (240 samples) are used for training and the remaining 60 samples for the test. Later these images are normalized between 0 and 1. Bayesian Optimization is used for hyper parameter tuning. The search space is described in Table 1. Keras tuner library is used for the optimization with 50 trial points and 50 epochs each. At the end of optimization, 3 best models were selected and trained with 5 cross validation folds for 500 epochs.

Results and Discussion

Figure 4 illustrates training curves of 3 best models selected by Bayesian optimization. The validation and test scores of Bayesian search are listed in Table 2. **For hyperparameter setting, please refer to the outputs of wildcard.ipynb. Due to the large number of outputs, I did not want to crowd the report with the list of all selected hyperparameters.** Looking at the training curves and later the predictions in Figure 3 from all 3 models it looks like the training went smooth for all 3 models. However, the test predictions cannot be considered acceptably good.

Elements of Exploratory Data Analysis

Throughout the last 3 exercises - hyper parameter optimization, machine learning pipeline optimization and wildcard – I have been trying to solve one problem but formulated differently every time. And there are several observations I have made. First of all, there are not nearly enough samples to draw reasonable conclusions and make good predictions. Secondly, the input images are too large for some models. For instance, for pipeline optimization problem a quarter million pixels are way too large, so I had to down sample the image so much that most of the features are lost and that resulted in a poor performance. Third, implementation of PCA did not prove to be useful as I discussed in hyper-parameter optimization exercise. So, the elements of analysis such as down sampling, dimensionality reduction, normalization, etc. have been considered a lot in these exercises. Moving on, I will keep working on these datasets and expand to produce new samples for my PhD research. The main idea I have is to extract quantitative values from the image that represent connectivity and use them as input features instead of feeding images themselves. I would consider this the most valuable lesson I have learnt from these exercises.

Appendix: Figures and Tables

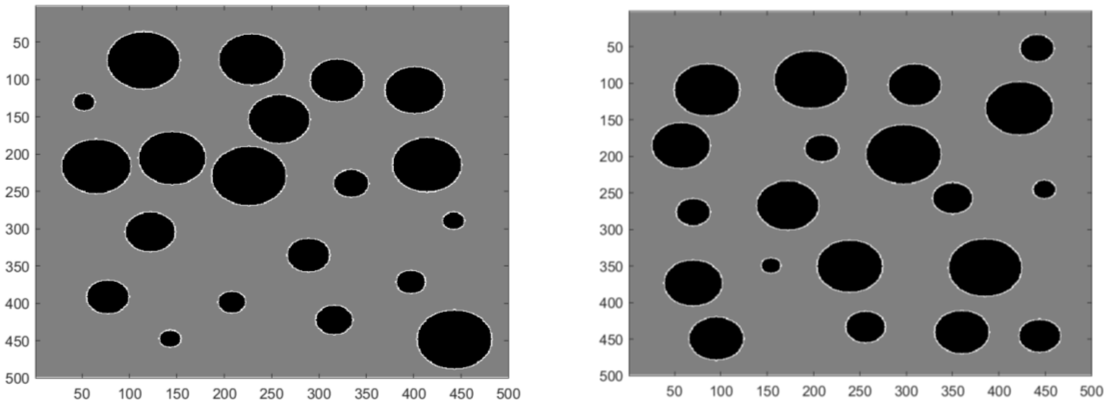

Figure 1. Randomly generated simulation domains. Gray areas indicate pores and black areas (circles) represent grains (boundaries) that fluid cannot penetrate.

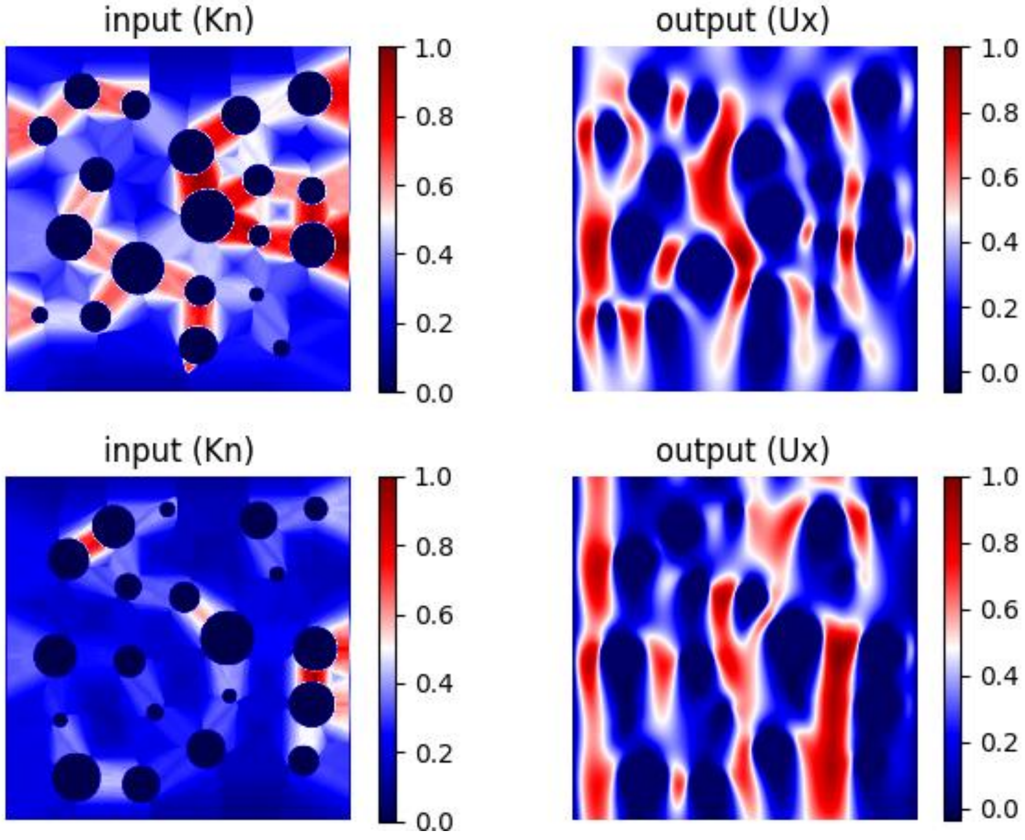

Figure 2. Sample input (left column) Knudsen number distributions and output (velocity) distribution. Data has been normalized between 0 and 1.

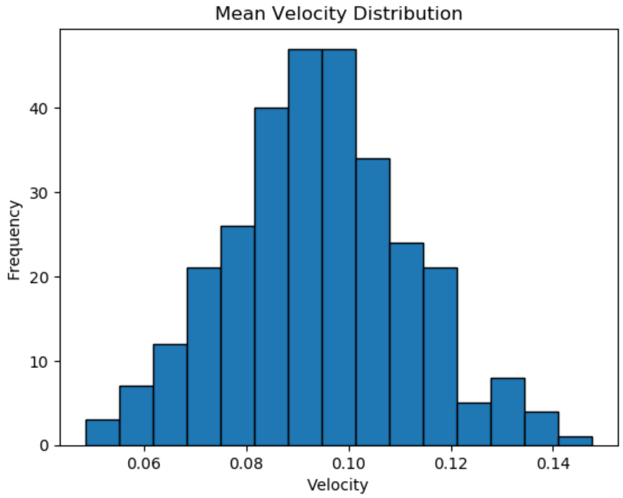

Figure 3. Mean Velocity Distribution

Table 1. Search space for hyper-parameter optimization

Parameter	**Values**
Number of kernels	32, 64, 96, 128
Kernel size	3x3, 4x4, 5x5
Number of Conv2D layers	1, 2, 3, 4, 5
Number of Maxpooling layers	1, 2, 3, 4, 5
Neurons in Dense layer	min = 32, max = 256, step = 32
Learning rate	Min = 1e-4, max = 1e-2, sampling = linear
Optimization setting	50 trials, 50 epochs each

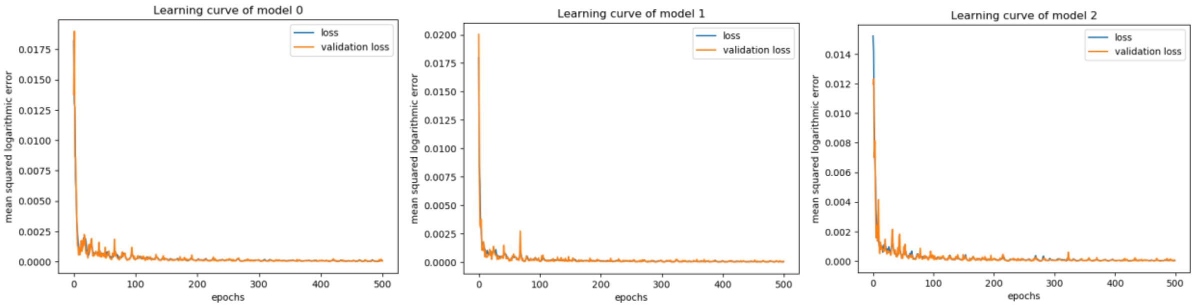

Figure 4. Training curves 3 best models

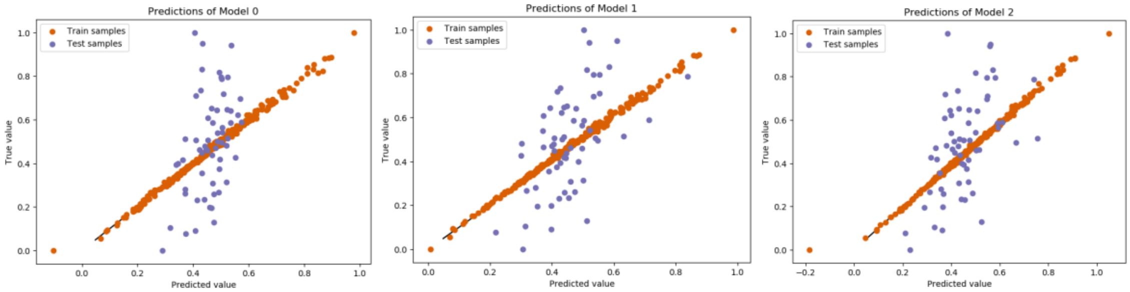

Figure 4. Predictions of best 3 models

Table 2. Results of Bayesian Search

Model	**Validation Score**	**Test Score (MSE)**
0	0.00597	0.0434
1	0.00835	0.0368
2	0.008444	0.0382

References

1. Rustamov, N., Douglas, C. C., & Aryana, S. A. (2022). Scalable simulation of pressure gradient-driven transport of rarefied gases in complex permeable media using lattice Boltzmann method. *Fluids*, *8*(1), 1. https://doi.org/10.3390/fluids8010001
2. Rustamov, N., Liu, L., & Aryana, S. A. (2023). Scalable simulation of coupled adsorption and transport of methane in confined complex porous media with density preconditioning. *Gas Science and Engineering*, *119*, 205131. https://doi.org/10.1016/j.jgsce.2023.205131
3. https://automl.github.io/auto-sklearn/master/