Design Space Exploration of Deep Neural Networks

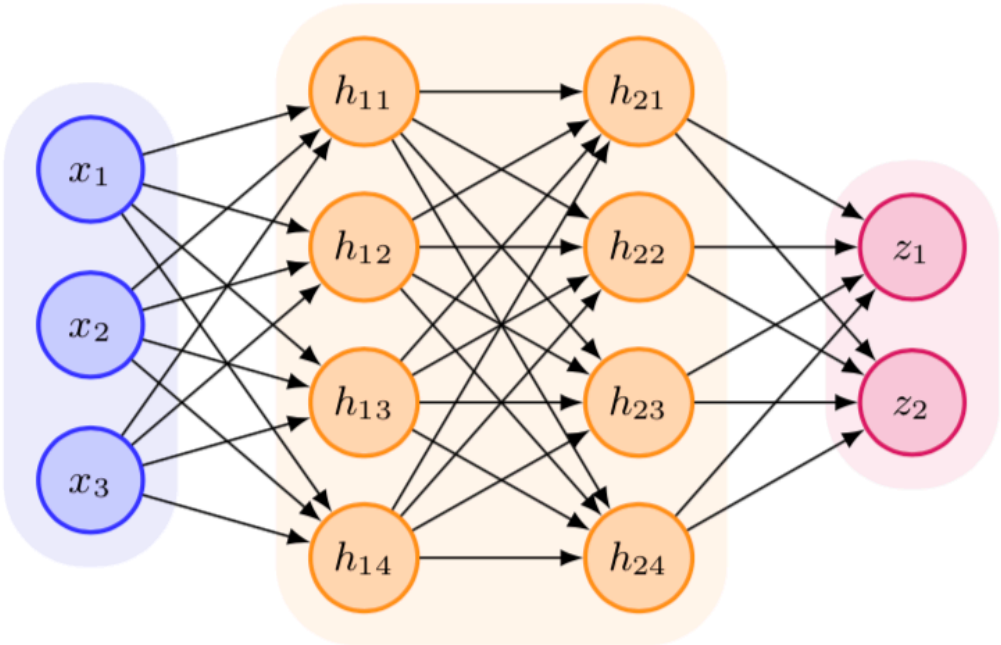
Course Project
CSCE 790
(Machine Learning Systems)

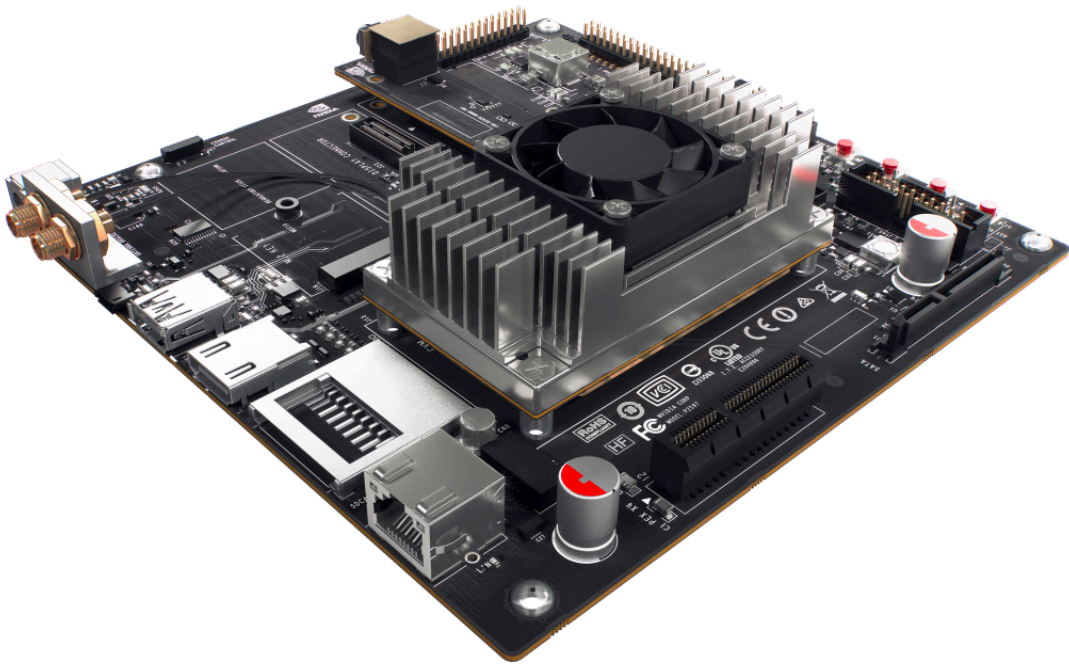
How projects will be evaluated?

- Team up with other 2 students, so each team 3 persons
- Select one project
- No communications between the two teams
- Every teammate should be able to demonstrate her/his contribution
- The outcome will be evaluated based on the quality of the results, report, and final presentation.
- The final report is an iPython notebook that has documentation, results, comparisons, discussions, and related work.
- 60% of your final mark will be evaluated based on the course project.

Project 1: Design Space Exploration of Deep Neural Networks

How the choice of DVFS (e.g., CPU frequency, enable/ disable CPU cores) affect inference time and energy consummations of Deep Neural Networks (DNNs)?





Project description

- The aim of the project is to perform design space exploration of DNNs in resource constrained devices (e.g., Nvidia TK1, TX1).

- The goal is to understand how the choice of hardware configurations in the deployment environment can influence runtime characteristics of DNNs.

Selecting hardware platform

- You first need to select the hardware platform for your experiment
  - I will give you access to an instance of Nvidia TX1, once you created your GitHub repository, send me an email and I will give you access instructions.
  - You are also free to choose the hardware platform of your choice, let me know this choice as well.

Deciding the configuration space

- You need to then choose the configuration space you would like to explore.
- For this, you need to select specific configuration options you can vary on the hardware platform you choose for your experiments. E.g.:
  - CPU frequency
  - Number of enabled cores
  - Modes of CPU cores (General, low-power)
- You should also select few configuration options at the model compiler level, e.g.:
  - TensorFlow options (KMP BLOCKTIME, KMP AFFINITY, OMP NUM THREADS),
  - LLVM options (O1, O2, O3, vectorize slp aggressive),
  - CUDA options (maxrregcount, ptxas options, ftz, prec div, prec sqrt, use fast math, fmad)

Selecting specific DNN architectures

- Select few pre-trained DNN architectures that fit onto your hardware platform, e.g.:
  - Any pre-trained CNN architecture
  - Use available implementations, e.g.,: https://github.com/tensorflow/benchmarks

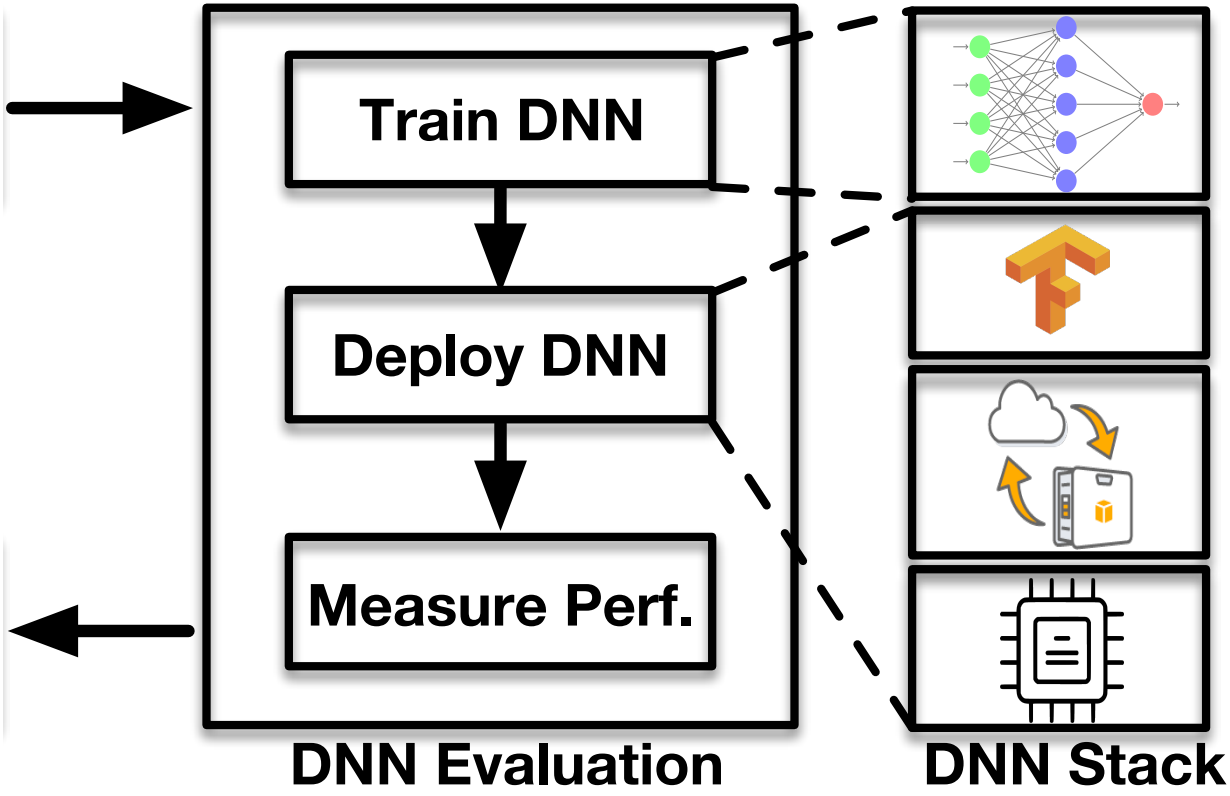
Start measurements

- Once you decided about the configuration space, you need to determine the configurations that you want to measure.
- At this stage you need to discretize the continuous variables.
- And think about using a sampling strategy, e.g., random sampling, or possibly Full factorial design
  - https://en.wikipedia.org/wiki/Design_of_experiments
- Do not forget that you need to measure both Inference time and energy consumption for each configuration

Deciding about workload

- Choose 2 different workloads from existing datasets, e.g. UCI repository, or other available datasets
  - Image
  - Time-series
  - Text
  - etc.

Performance measurements



Analyzing data

- Once you measured configurations, you need to dig into data and find interesting trends.
  - You could look into Pareto-optimal configurations
  - You could find whether the optimal configurations in one DNN architecture is also optimal in other architectures, if not dig into and find out why.
  - You could look into correlation measures across different workloads
  - You may want to have a look at this to get some idea what kinds of analyses you may want to perform: https://arxiv.org/pdf/1709.02280.pdf

Final point

- Use your creativity when it comes to analyzing the results, try to surprise me!
- If you find a very interesting observations and dig into it by providing some insight, you will then get a good score!
- If you also produce very good results, you may also want to think about a potential paper, it’s optional, but I strongly recommend it.