III International Scientific Symposium "Intelligent Solutions" (IntSol-2023).
The Satellite Symposium, XI International School-Seminar "Decision Making Theory" and the VII International Conference "Computational Intelligence"

# On Improving the Use of Dedicated Hardware Resources for Deep Learning Methods on the Small Data

**Vladyslav Kuznetsov** [1] , Yurii Krak [1,2] ,
Serhii Kondratiuk [1,2] , Andriy Stavrovskyi [2]

*V.M. Glushkov Institute of Cybernetics of National Academy of Sciences of Ukraine* [1] ,
*Taras Shevchenko National University of Kyiv* [2],

III International Scientific Symposium "Intelligent Solutions" (IntSol-2023).
The Satellite Symposium, XI International School-Seminar "Decision Making Theory" and the VII International Conference "Computational Intelligence"

# Related work

- A Comparative Study on Algorithms of Computer Vision and Deep Learning for Facial Expressions Analysis [V. Kuznetsov [1] , Y. Krak [1,2] , O. Barmak [3] , A. Kulias [1] , V. Petrovich [1] ] , 2022
- Using the Temporal Data and Three-dimensional Convolutions for Sign Language Alphabet Recognition [S. Kondratiuk [1,2] , Y. Krak [1,2] , V.Kuznetsov [1] , A. Kulias [1] ], 2022
- On classification hidden concepts language in specialized texts based on methods of the intellectual data processing [Y. Krak [1,2] , V.Petrovich[1] , V. Kuznetsov [1] , O.Barmak [3] , A. Kulias [1] ] , 2021
- Detection of early pneumonia on individual CT scans with dilated convolutions [Y. Krak [1,2], O. Barmak [3], P. Radiuk [3]], 2021

*Occupation/work : V.M. Glushkov Institute of Cybernetics of National Academy of Sciences of Ukraine [1] , Taras Shevchenko National University of Kyiv [2] , Khmelnytsky National University [3]*

# What has to be done

- to study the processors' and video adapters' performance for training with such networks:
  - back propagation network and autoencoder;
  - convolutional and recursive conv. network;
  - generative adversarial network.
- suggest a strategy for task, according to behaviour of hardware in different tasks;
- to check out the speed and compare to known results,e.g. Techpowerup specs and GeekBench

III International Scientific Symposium "Intelligent Solutions" (IntSol-2023).
The Satellite Symposium, XI International School-Seminar "Decision Making Theory" and the VII International Conference "Computational Intelligence"

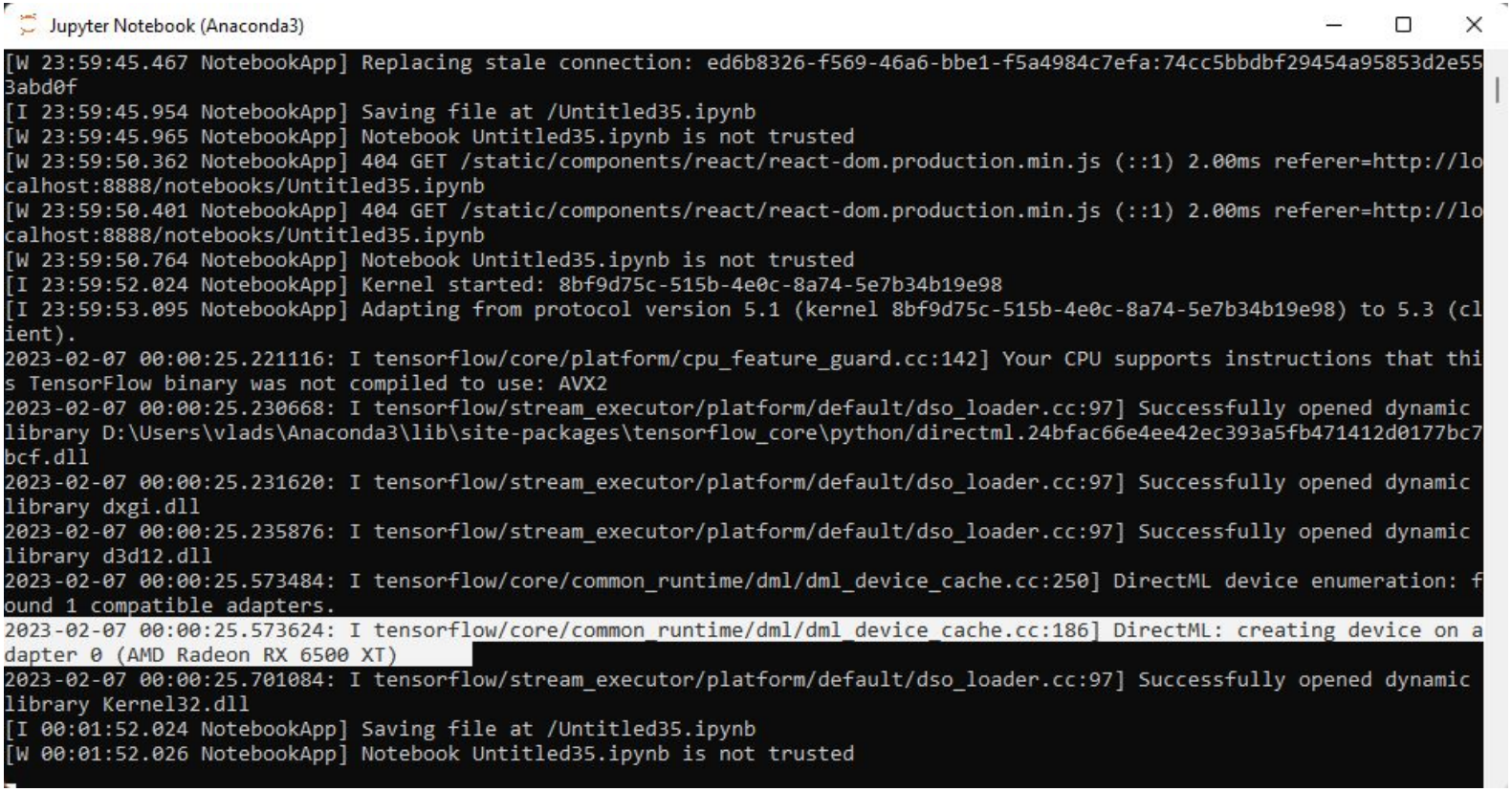
# Tools used in the research

- Computer specs and used setup :
  - Processor: Ryzen 5 3600;
  - Video card: AMD Radeon RX 6500 XT
  - Other hardware:
    - GPUs: Nvidia RTX 2060 mobile
    - CPUs: AMD Athlon X2 series, Intel Core2 E series
- Libraries and software :
  - Python 3.8 (Anaconda);
  - PyTorch-DirectML [1]

DirectML is a cross-platform machine learning library with tools for scaling machine learning models for AMD Radeon, Intel and Nvidia GeForce accelerators , more details here: https://github.com/microsoft/DirectML

III International Scientific Symposium "Intelligent Solutions" (IntSol-2023).
The Satellite Symposium, XI International School-Seminar "Decision Making Theory" and the VII International Conference "Computational Intelligence"

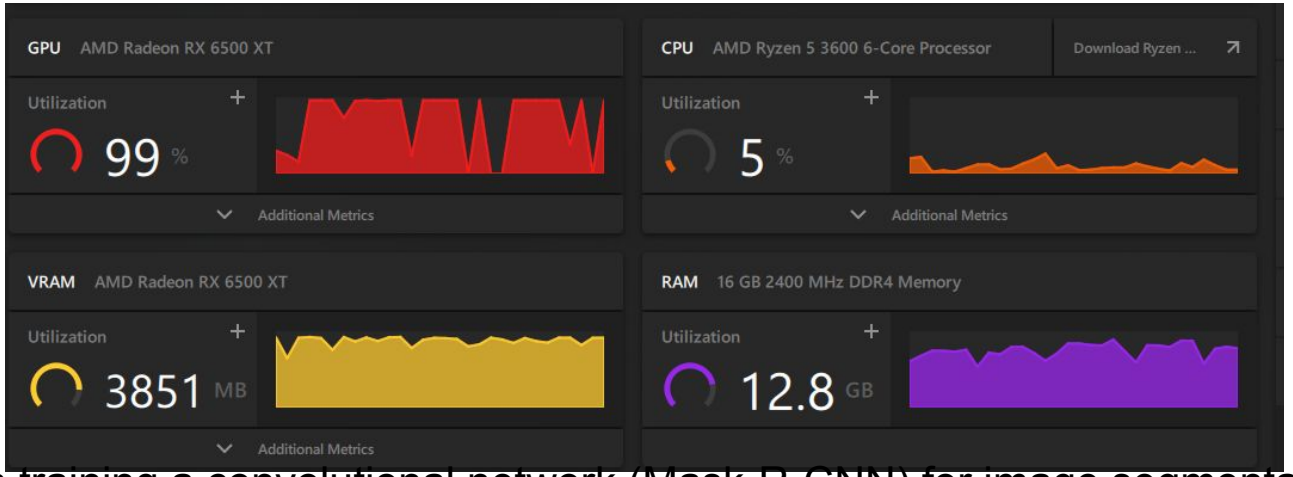
# DirectML. Putting it all together




DirectML enables PyTorch network models and Tensorflow for AMD Radeon, Intel, and Nvidia GeForce accelerators. Each graphics accelerator is treated as a virtual DML device (thanks to the DirectX 12 library), which emulates the CUDA virtual device in standard PyTorch and Tensorflow versions.

III International Scientific Symposium "Intelligent Solutions" (IntSol-2023).
The Satellite Symposium, XI International School-Seminar "Decision Making Theory" and the VII International Conference "Computational Intelligence"

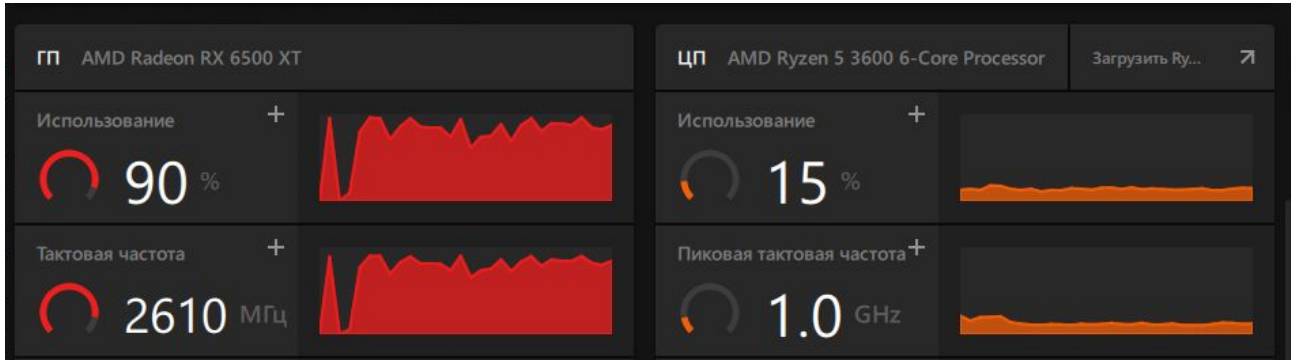
# PyTorch experiments




When training a convolutional network (Mask-R-CNN) for image segmentation, we noticed a significant increase in system load. When estimating the resources of the video accelerator, it seem they are not abundant. The resources of the processor are used to calculate the loss and the gradient is calculated using the video accelerator (also suitable for DCGAN type networks).The GPU memory utilization reaches 3.8GB (4 in total), the shared memory reaches 8GB (16 in total), and the GPU utilization reaches 99%. High resource consumption indicates the need to optimize the modelFor the DCGAN network, there is no difference in load levels; while the execution time ratio of GPU+CPU to CPU is as high as 2.74:1

III International Scientific Symposium "Intelligent Solutions" (IntSol-2023).
The Satellite Symposium, XI International School-Seminar "Decision Making Theory" and the VII International Conference "Computational Intelligence"

# Tensorflow experiments




The number and type of peak compute loads are similar, but there are differences. For example, the speed results obtained here differ from previously calculated values in the GeekBench AI benchmark and GeekBench OpenCL tests, being around 18.9:1, favoring the processor used in the experiments. This leads us to think that performance ratios should be normalized relative to the CPU utilization percentile of the specific library used. The results obtained by normalization differ less from those obtained by previous experiments using the same hardware

GeekBench OpenCL is a standard test for checking the performance of computing devices using the OpenCL library. This benchmark can be used to compare video accelerators: https://browser.geekbench.com/opencl-benchmarks

III International Scientific Symposium "Intelligent Solutions" (IntSol-2023).
The Satellite Symposium, XI International School-Seminar "Decision Making Theory" and the VII International Conference "Computational Intelligence"

# Conclusions and thoughts

- When computing large models, a combination of processor and video accelerator is required.
- Therefore, there is a time penalty in sending memory segments from the graphics card to the processor and back
- However, even then there is a time gain when using a graphics-compatible processor (2.74:1 for the DCGAN training example alone).
- Overall, as a result of this study, a strategy for sharing CPU and video core resources (memory and time) is proposed
- It allows to evaluate the possibility of training large models without using expensive hardware
- DirectML is still under development by some operators, so it may not be available on every network recently developed

DirectML allow a partial subset of operators both for TensorFlow and PyTorch, available for the original non-DML versions. The operator roadmap can be seen on the GitHub page: https://github.com/microsoft/tensorflow-directml/wiki/Roadmap-(Operators)

# Thank you!