Justification for choosing a single-board hardware computing platform for a neural network performing image processing

A E Tolmacheva, D A Ogurtsov and M G Dorrer
Reshetnev Siberian State University of Science and Technology, prospect Krasnoyarsky rabochy, 31, Krasnoyarsk, Russia

E-mail: mdorrer@mail.ru, dim_ogu@mail.ru

Abstract. This article describes testing the operation of the YOLO2 neural network on single-board platforms and processors. The task was to find a computing platform on which the neural network could perform the processing of the images presented to it as quickly as possible. In this case, interference in the operation of the remaining modules of the system should have been minimized. In addition, the task was to minimize the solution price. Based on the results of the comparison, the Neural Compute Stick 2 hardware platform was chosen as a computing platform for the tasks with neural networks. In the future, the selected module can be used instead of the NVIDIA 1070 ti video card for the task of clustering and segmenting objects using the YOLO2 real-time network. This choice will simplify the design of the hardware-software complex and minimize its cost.

1. Introduction
In modern conditions, the tasks of intelligent image processing are relevant, and therefore the question arose for business about choosing a hardware platform for industrial projects. In this work, the task of comparing several platforms that differ in size and computing capabilities, as well as relating to different price segments, was solved.

For testing in this article, two different neural network frameworks were chosen: darknet YOLO [1] and tensorflow [2]. Since the values of the YOLO2 synaptic weights cannot be used directly in the tensorflow framework, the darkflow software add-in was used to ensure compatibility.

The main objective of the study was to find the fastest, most powerful and cheapest device that can be installed in a smart refrigerator as part of the automation of the store’s trading floor.

2. Testing hardware platforms
The 96Boards HiKey 970 Single Board Computer [3] is targeted at developers creating robotic systems and devices for the Internet. The specialized NPU AI Hikey970 platform is built on the basis of the Huawei HiAI architecture, which includes a CPU, GPU and a specialized processor for neural networks (NPU). The platform supports neural networks popular on the market, using NPUs to speed up their work.
Table 1. HiKey970 Module Specifications

| Parameter | Value |
|-----------|-------|
| SoC       | HiSilicon Kirin 970 - HiAI Architecture, Dedicated NPU |
| CPU       | ARM Cortex-A73 MPCore4 @ up to 2.36 GHz, ARM Cortex-A53 MPCore4 @ up to 1.8 GHz |
| GPU       | ARM Mali-G72 MP12 GPU |
| RAM       | 6GB LPDDR4X 1866MHz |
| Cost      | 350-400 $ |

The NVIDIA Jetson Nano platform [4] is designed for mass development of compact, energy-efficient and affordable devices. The solution expands the capabilities of developers in terms of creating embedded IoT applications, including for entry-level DVRs, home robots and smart gateways with analytic capabilities.

Table 2. NVIDIA Jetson Nano Developer Kit Specifications

| Parameter | Value |
|-----------|-------|
| CPU       | Quad-core ARM A57 @ 1.43 GHz |
| GPU       | 128-core Maxwell |
| RAM       | 4 GB 64-bit LPDDR4 25.6 GB/s |
| Cost      | 100 $ |
The video cards of the GeForce GTX 10 series [5] are based on the Pascal architecture and provide up to 3 times higher performance compared to the previous generation video cards. Such video cards are used not only in the gaming industry, but also when working with block chain, as well as in the calculation of artificial neural networks.

Table 3. NVIDIA GEFORCE GTX 1070 Ti Specifications

| Parameter                        | Value          |
|----------------------------------|----------------|
| GPU Architecture                 | Pascal         |
| Cores                            | 2432           |
| Performance                      | 3x             |
| Framebuffer                      | 8 GB GDDR5     |
| Clock frequency (with acceleration) | 1683 MHz       |

In 2018, a solution called Rock Pi 4 was presented [6], made in the format of the popular Raspberry Pi product. The Rock Pi 4 single-board computer is equipped with the Rockchip RK3399 six-core processor in combination with 1 to 4 GB RAM and corresponds to the form factor of the Raspberry Pi 3 and ASUS Tinker boards.
Table 4. Rock Pi 4 model B 4gb Specifications.

| Parameter | Value |
|-----------|-------|
| SoC       | 6-core Rockchip RK3399 |
| CPU       | ARM Cortex-A72 2-core @ up to 1.8GHz, ARM Cortex-A53 MPCore @ up to 1.4GHz |
| GPU       | Mali-T864 with support OpenGL ES1.1 / 2.0 / 3.0 / 3.1, OpenVG1.1, OpenCL, DX11 and AFBC |
| RAM       | 4 Гб LPDDR4 @ 3200 MB / s |
| Cost      | 79.95$ |

Intel® Neural Compute Stick 2 [7] – a development kit with support for the Plug and Play standard for solving problems of artificial intelligence. The module can work without connecting to a cloud environment and allows you to create prototypes using inexpensive end devices such as Raspberry Pi 3 and the like.

Figure 5. Appearance of Intel® Neural Compute Stick 2.

Table 5. Intel® Neural Compute Stick 2 Specifications.

| Parameter     | Value                  |
|---------------|------------------------|
| VPU           | Intel Movidius Myriad X |
| Power Source  | USB 3.0 Type-A         |
| Mechanical    | 72.5 x 27 x 14 mm      |
| Cost          | 100$                   |

Intel® Core™ i3 CPUs [8] and Intel® Core™ i7 [9] CPUs were also among the devices on which performance comparisons were performed when processing artificial neural networks.

Table 6. CPU Parameters.

| Parameter                  | Value for Intel® Core™ i3-9100 | Value for Intel® Core™ i7-7700 |
|----------------------------|---------------------------------|---------------------------------|
| Cores                      | 4                               | 4                               |
| Threads                    | 4                               | 8                               |
| Maximum clock speed with Turbo Boost Technology | 4.20 GHz                         | 4.20 GHz                         |
| Cache memory               | 6 MB Intel® Smart               | 8 MB Intel® Smart               |
Before the testing began, the preference of the researchers was on the side of Jetson Nano, because it was the only ARM of all tested devices containing in its architecture the CUDA scalar cores used by the neural network when analyzing images in the GPU mode.

In preparation for testing the comparable hardware platforms, the following software was installed:

**Table 7. Software installed on comparable platforms.**

| Platform type                              | Software                                           |
|--------------------------------------------|----------------------------------------------------|
| HiKey970                                   | Debian Pre-built Tensorflow For HiKey970, OpenCV 3.3.1 |
| NVIDIA Jetson Nano Developer Kit           | Ubuntu, OpenCV 3.3.1                              |
| Rock Pi 4 model B 4gb                      | Ubuntu, OpenCV 3.3.1                              |
| Intel® Core™ i3-9100                       | Windows 10, OpenCV 3.3.1                          |
| Intel® Core™ i7-7700                       | Windows 10, OpenCV 3.3.1                          |
| NVIDIA 1070ti                              | Ubuntu 1070ti OpenCV 3.3.1                        |

RockPi 4 testing ended after installing the graphical shell of the Ubuntu operating system v. 18.06. The GUI took up all the RAM, and the device overheated, constantly going into a state of failure. This microcomputer did not show good results also during the testing of additional modules.

The table below shows the comparative image processing time on different platforms and with different image size indicators (2592x1944, 1280x1024 and 500x400 pixels), which allows us to conclude their performance for images of different dimensions. The test was conducted on the darknet YOLO framework. Data is presented in decreasing order of performance.

**Table 8. GPU-mode Image Processing Time (milliseconds).**

| Platform                                          | 2592x1944 pixels | 1280x1024 pixels | 500x400 pixels |
|---------------------------------------------------|------------------|-----------------|---------------|
| NVIDIA GTX 1070 TI (GPU)                          | 0.0001           | 0.0001          | 0.0001        |
| NVIDIA Jetson Nano Developer Kit (GPU)            | 0.0504           | 0.0257          | 0.0209        |

The results of platform testing in CPU mode (can be seen in table 9) showed that HiKey970 is the best in image processing speed. If we consider the use of GPU mode for devices that include CUDA scalar cores, there is no doubt that the 1070ti has the best performance. However, the authors were not faced with the pure task of choosing the most productive solution. A minimum price was required with performance metrics that met future challenges. Given this limitation, Jetson Nano was chosen to work with neural networks in GPU mode.

**Table 9. CPU-mode Image Processing Time (milliseconds).**

| Platform          | 2592x1944 pixels | 1280x1024 pixels | 500x400 pixels |
|-------------------|------------------|-----------------|---------------|
| Intel Core i7 (CPU) | 551.838          | 533.932         | 547.628       |
| Intel Core i3 (CPU) | 1067.570         | 1024.772        | 1028.154      |
| HiKey970 (CPU)     | 26.8410          | 26.5670         | 26.6620       |

The data in table 9 show that the CPU solves the problem of image processing using a neural network significantly (sometimes tens of thousands of times) slower than a GPU, which confirms the hypothesis that it is beneficial to use GPU-based solutions containing CUDA scalar cores.
Hikey970 has a powerful GPU, but it was not possible for the darknet framework to check the image processing speed, since its GPU was specialized in the tensorflow framework. In order to compare the computing capabilities of the Jetson Nano and Hikey platforms in GPU operation mode, it was decided on both platforms to launch a ready-made model on the tensorflow framework [10], [11].

Table 10 below shows a comparison of the performance of Jetson Nano and HiKey 970 on the tensorflow framework on the MNIST dataset (60,000 training images, 10,000 test images). It should be noted that during the tests on the devices different tensorflow versions were installed, which was caused by limitations in HiKey 970 compatibility.

| Platform                        | Time for an epoch of teaching (sec.) | Test Time (sec.) |
|---------------------------------|--------------------------------------|-----------------|
| NVIDIA Jetson Nano Developer Kit| 24                                   | 2.17            |
| HiKey 970                       | 43                                   | 4.6             |

You can see that the Jetson Nano is 2 times more efficient (faster), the neural network is being processed, compared to the HiKey 970.

3. Image Processing Results on Various Platforms
The following is (upper left edge of figure 6) the recognition window when using the Neural Compute Stick. In this figure, you can see the data processing speed.

The Neural Compute Stick test results showed that this system processes an image of 1200x900px size in approximately 60 ms and is able to work with models based on yolo_v3. Testing the same network with Jetson Nano showed that Jetson Nano can process an image of 1200x900px size in approximately 600 ms Thus, the Neural Compute Stick is 10 times faster than the Jetson Nano.

It is worth noting that the Neural Compute Stick does not allow you to train models, unlike all the other solutions presented in this article.

4. Conclusions
As the hardware for the operation of a smart refrigerator with a neural network module, it was decided to choose the Neural Compute Stick, using the YOLO v3 model to work.
At the moment, it is worth paying special attention to the solutions associated with the use of VPU (due to their compactness) for neural network image processing in conjunction with a training server based on the GPU.

Hardware solutions on single-board neural network platforms are becoming more widespread in the modern world [12-15]. This confirms the relevance of this study.

References
[1] Darknet Retrieved from: https://pjreddie.com/darknet/
[2] Tensorflow Retrieved from: https://www.tensorflow.org/
[3] HiKey970 Parameters Retrieved from: https://www.96boards.org/product/hikey970/
[4] NVIDIA Jetson Nano Developer Kit Parameters Retrieved from: https://www.nvidia.com/ru/autonomous-machines/embedded-systems/jetson-nano/
[5] NVIDIA GEFORCE GTX 1070 Ti Parameters Retrieved from: https://www.nvidia.com/ru/geforce/products/10series/geforce-gtx-1070/
[6] Rock pi 4 model B 4gb Parameters Retrieved from: http://rockpi.org
[7] Intel® Neural Compute Stick 2 Parameters Retrieved from: https://software.intel.com/ru/neural-compute-stick
[8] Intel® Core™ i3-9100 Parameters Retrieved from: https://ark.intel.com/content/www/ru/ark/products/134870/intel-core-i3-9100-processor-6m-cache-up-to-4-20-ghz.html
[9] Intel® Core™ i7-7700 Parameters Retrieved from: https://ark.intel.com/content/www/ru/ark/products/97128/intel-core-i7-7700-processor-8m-cache-up-to-4-20-ghz.html
[10] Matija B and Miran P 2019 Adapting YOLO Network for Ball and Player Detection pp 845-51
[11] Abu L and Delowar H 2018 Cloud-based Smart Surveillance System using Raspberry Pi and YOLO
[12] Lewis F L et al 1996 Multilayer neural-net robot controller with guaranteed tracking performance IEEE Transactions on Neural Networks 7(2) 388-99
[13] Şimşir M, Bayır R and Uyaraoğlu Y Real-Time Monitoring and Fault Diagnosis of a Low Power Hub Motor Using Feedforward Neural Network Computational Intelligence and Neuroscience 2016 7129376
[14] Wei B, Hamad R A et al. 2019 A Deep-Learning-Driven Light-Weight Phishing Detection Sensor Sensors 19 4258
[15] Fujisawa Y et al. 2019 Deep- learning- based, computer- aided classifier developed with a small dataset of clinical images surpasses board- certified dermatologists in skin tumour diagnosis British Journal of Dermatology 180 373–81