Optimization and Benchmarking of Convolutional Networks with Quantization and OpenVINO in Baggage Image Recognition

N.A. Andriyanov, G. Papakostas*

*naandriyanov@fa.ru

ITNT - 2022

23 – 27 May 2022

Introduction

An important task is the recognition and detection of objects in images. There are a number of optimization approaches that are successfully used in the problem of accelerating the operation of deep neural networks, including the processing of optical images. These include pruning, which removes weights and links from the model, weight quantization, and distillation. Another solution is software/hardware-based acceleration or platform-specific acceleration. Such solutions include an accelerator for working on Intel processors such as OpenVINO Toolkit. It was proposed to study the inference acceleration for X-ray images of baggage and hand luggage.

Training and Inference

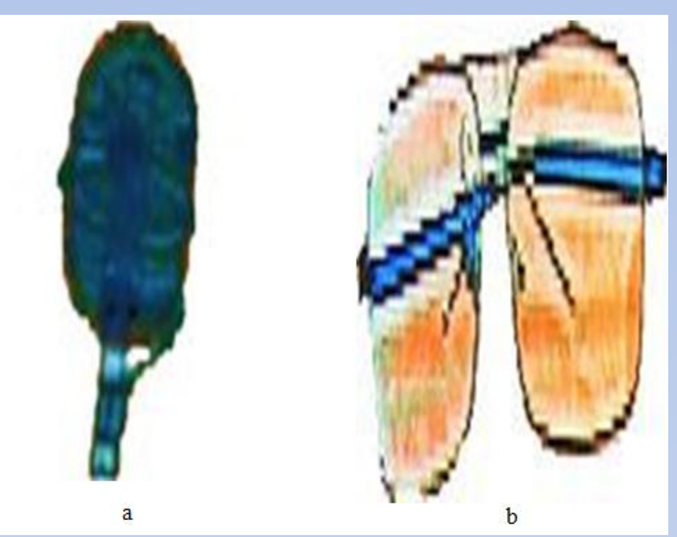

Fig. 1. Prohibited (a) and permitted (b) items

In general, the volume of the image database was 4000 images of permitted objects and 2000 images of prohibited ones. For recognition, two convolutional neural networks were trained from scratch (using the Keras and TensorFlow frameworks).

The training was performed on GPU NVIDIA GeForce GTX 1060. The size of the test sample was 1500 and 800 images for " permitted" and " prohibited " objects, respectively. Image sizes are 200 by 200 pixels.

ResNet and VGG also was used.

Open VINO was used for acceleration of inference (Intel core i7-9700k).

Results

Table 1. Performance Evaluation

Architecture (model)	*FPS*	*Recall*
CNN-3	2.23 ± 0.09	0.79
CNN-5	1.98 ± 0.12	0.84
VGG-16	0.93 ± 0.09	0.91
ResNet-50	0.67 ± 0.10	0.95
ViT	0.23 ± 0.07	0.94
CNN-3q	9.54 ± 0.11	0.68
CNN-5q	7.65 ± 0.10	0.72
VGG-16q	4.82 ± 0.16	0.86
ResNet-50q	3.22 ± 0.17	0.89
CNN-3ov	29.03 ± 5.96	0.79
CNN-5ov	23.69 ± 6.74	0.84
VGG-16ov	11.72 ± 3.22	0.91
ResNet-50ov	7.93 ± 2.69	0.95

Conclusions

Thus, the paper proposes to use accelerators for the inference of neural networks. It is shown that with the help of quantization, an acceleration of approximately 3.5–4 times is performed. At the same time, the recall of prohibited objects recognition drops by about 10%. In the case of using the Intel OpenVINO Toolkit, it is possible to maintain the declared characteristics of the network, while the performance gain is on average about 11-13 times. However, it should be taken into account that the use of OpenVINO provides the inference time with a large spread. A significant improvement in the quality of models can be achieved thanks to transfer learning. Thus, the VGG, ResNet and ViT networks have shown themselves to be much preferable compared to learning from scratch. The average gain in recall was 8-10%.

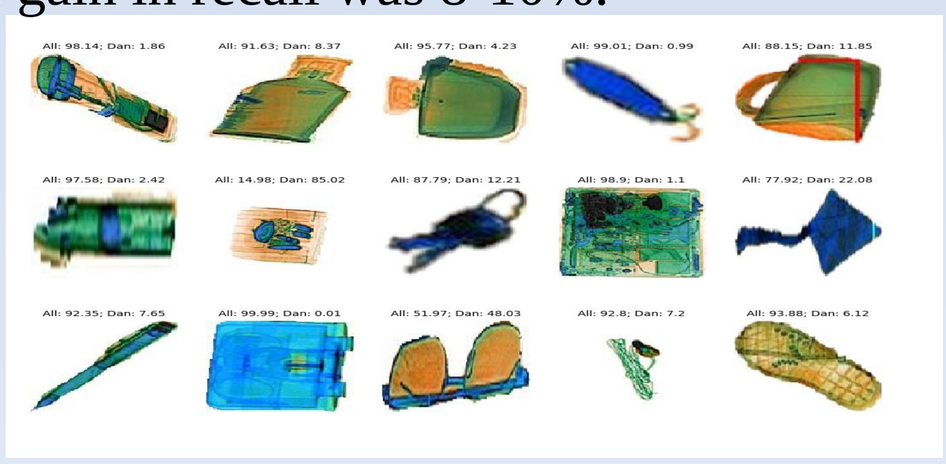

Fig. 2. Processing Examples

Financial University under the Government of the Russian Federation, Moscow, Russia

International Hellenic University, Thessaloniki, Greece

The study was supported by the Council for Grants of the President of the Russian Federation within the framework of the Project on the Scholarship of the President of the Russian Federation for Young Scientists and Postgraduates No. SP-3738.2022.5 and partially with the support of the RFBR grant No. 19-29-09048.