ПЛЕНАРНІ ДОПОВІДІ

UDC 004.93

WHY DO WE NEED A POST-TRAIN ADAPTIVE NEURAL NETWORK?

Kostiantyn Khabarlak, Ph.D., Assistant Professor, Khabarlak.K.S@nmu.one,
Dnipro University of Technology

Neural networks have shown to be effective in many areas. Convolutional neural networks solve computer vision problems, such as classification, detection and segmentation depending on task at human level or better. Recurrent and transformer-based neural networks are actively used for natural language understanding. More recently neural networks have shown high quality in generative tasks in both vision and language domains. All of it results in an increased usage of neural networks. Some of the applications require offline data processing due to privacy requirements, lack of Internet access or high server load which is better to be distributed among user devices to reduce server maintenance cost. However, user devices have significantly different processing power among each other. In many cases it might be preferable to exchange some accuracy for inference speed, especially on low-end devices. While traditional neural networks offer a static architecture, which cannot be changed at runtime, dynamic neural networks have the capability to adjust the number of operations based on system load or input data.

In this work a Post-Train Adaptive (PTA) neural network is proposed as a simple yet effective network, that can change architecture based on user demand without extra training steps. The key element of the network is a PTA block, that serves as a drop-in replacement for 2 Inverted Residual blocks of a MobileNetV2 [1] neural network. It has light and heavy branches (Fig. 1), that can be used either jointly to improve the final quality or separately to reduce system load. By utilizing PTA sampling strategy, branches of a block can be selected dynamically at runtime without extra training steps. Thus, a single trained neural network has several configurations available with different processing speed and quality.

Currently the PTA neural network has been implemented for the classification and image segmentation tasks. In both cases the PTA network has superseded the baseline MobileNetV2 neural network in terms of inference efficiency (computed as recognition quality to inference time ratio). The neural network inference time has been measured across 5 devices with different speed, including edge devices, smartphones, laptops and a GPU.

Experimental results show [2], that for the image classification task inference time can be dynamically adjusted from 80% to 107% relative to the MobileNetV2 network baseline. In the face anti-spoofing task, the lightest configuration of the proposed PTA neural network is not only 20% faster, but also slightly more accuracy.

In the image segmentation task, the PTA blocks have been built-in the U-Net neural network and have allowed to change inference time [3] from 94% to 105% relative to the U-Net baseline, while also offering slightly higher Dice score.

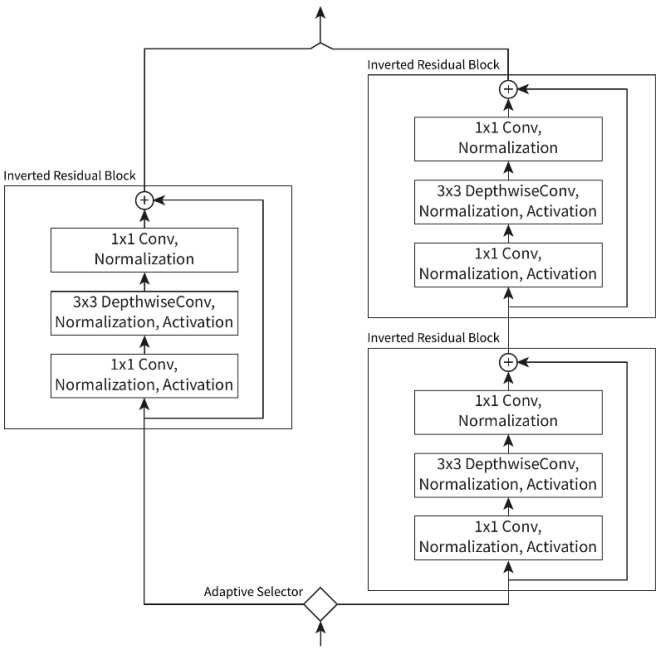

Figure 1 – The proposed PTA block architecture

Conclusions. The PTA network can be trained once and then reconfigured without additional training steps. Based on the conducted experiments, it has shown improved efficiency for the classification and image segmentation tasks. Depending on the dataset, reconfiguration is done with insignificant accuracy reduction or even with accuracy improvement. The network has been applied to edge devices, smartphones, portable PCs and GPUs. Its performance can be adjusted to be from 20% faster to 7% slower relative to the baseline given the required network accuracy and inference speed. In future the network can be applied to other computer vision tasks.

References

1. Sandler M, Howard AG, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: Inverted residuals and linear bottlenecks. In: 2018 IEEE conference on computer vision and pattern recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. Computer Vision Foundation / IEEE Computer Society; 2018. p. 4510–20. Available from: https://arxiv.org/pdf/1801.04381.pdf

2. Khabarlak K. Post-Train Adaptive MobileNet for Fast Anti-Spoofing. In: Proceedings of the 3rd international workshop on intelligent information technologies & systems of information security, Khmelnytskyi, Ukraine, March 23–25 [Internet]. CEUR-WS.org; 2022. p. 44–53. (CEUR workshop proceedings; vol. 3156). Available from: http://ceur-ws.org/Vol-3156/keynote5.pdf

3. Khabarlak K. Post-Train Adaptive U-Net for Image Segmentation. Information Technology: Computer Science, Software Engineering and Cyber Security. 2022;(2):73–8. Available from: https://journals.politehnica.dp.ua/index.php/it/article/view/93

UDC 519.85

DEVELOPING A HYBRID CONTINUOUS-DISCRETE APPROACH FOR OPTIMIZING MEDICAL LOGISTICS THROUGH TWO-STAGE LOCATION PROBLEM SOLVING

Oleksii Serhieiev, postgraduate student, serhieiev.o.s@nmu.one, Dnipro University of Technology

Optimizing logistics is essential for supply chain management, especially in healthcare. It efficiently distributes medical products efficiently and ensures public health and quick response during crises. Technologies and algorithms, like genetic algorithms for two-stage location problems, improve medical logistics by optimizing facility distribution. This enhances decision-making, speeding up operations and making them more cost-effective.

Metaheuristic algorithms are widely used in research on problems akin to those encountered in medical logistics. In [1], the researchers applied a genetic algorithm to analyze a two-stage transportation issue, focusing on a fixed route cost and the movement of goods. Paper [2] aims to improve spatial planning for public health services through location-allocation and accessibility models. The study seeks to determine the best sites for hospitals and healthcare facilities by considering population needs, accessibility, and closeness to other medical centers. Using Lisbon, Portugal's healthcare system as a case study, the research showed that applying these techniques greatly enhanced the quality and cost-effectiveness of healthcare. Study [3] explores optimizing resource allocation for natural disasters featuring several secondary hazards. It introduces a two-stage stochastic optimization model to replicate scenarios with unpredictable timing