UDC 004.8:623.746-519

QUANTIZATION EFFECTS ON DRONES AND UNMANNED AERIAL VEHICALS DETECTION

Bodenchuk-Pastukhov Y.V.
e-mail: yehor.bodenchuk-pastukhov@nure.ua
Supervisor – Candidate of Technical Sciences, Assistant Kobylin I.O.
Kharkiv National University of Radio Electronics, dep. INF
Kharkiv, Ukraine

This work investigates the impact of neural network quantization on the detection of tiny unmanned aerial vehicles (UAVs), including drones. A YOLO-based object detection model is employed to compare the performance of FP32 and INT8 weight representations using the «Drone vs Bird dataset». The evaluation focuses on the detection of small objects, which represent a challenging scenario for modern computer vision systems. The Intersection over Union (IoU) and mean Average Precision (mAP) metrics are used to assess detection accuracy. Experimental evaluation is conducted on an RTX 5070 Ti GPU to analyze the trade-off between model precision and computational efficiency. The results of this study aim to provide insights into the applicability of quantized neural networks for UAV detection tasks, particularly in scenarios involving resource-constrained or edge computing environments.

From 2022, the use of UAVs has significantly increased, particularly in the military domain. The primary purpose of UAV systems is to provide operators with a wide situational overview of a given area or operational sector. For example, long-endurance surveillance platforms such as the Global Hawk and Ultra LEAP UAVs have accumulated more than 18,000 combat flight hours, performing essential reconnaissance and monitoring missions [1].

In addition, the large-scale deployment of drones on the Ukrainian battlefield since 2023 has demonstrated the growing role of UAV technologies in modern warfare. Besides traditional reconnaissance tasks, drones are increasingly used for direct combat operations, including the engagement of personnel and military equipment using explosive payloads [2].

To keep pace with modern technological developments, the integration of neural networks (NNs) into UAV and drone systems has become an important direction for further improvement. However, despite the high accuracy achieved by neural networks, modern deep learning models require significant computational resources, which limits their deployment on resource-constrained systems. One of the widely used approaches for reducing model complexity and accelerating inference is neural network quantization.

Quantization reduces the numerical precision of network parameters and activations, typically converting floating-point representations (FP32) into lower precisions integer formats such as INT8. A common linear quantization scheme maps a floating-point value x to an integer value q according to (1):

$$q = round\left(\frac{x}{s}\right) + z$$

where x is floating-point, q is quantization value, s is scale factor, and z is zero point.

The quantization pipeline can be pictured as follows:

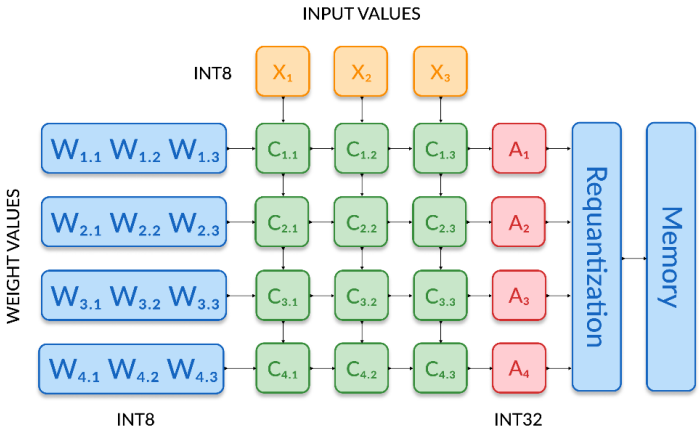

Figure 1 – Quantization pipeline scheme

To evaluate the performance of the YOLO-based model, the Intersection over Union (IoU) and mean Average Precision (mAP) metrics were used. The IoU metric measures the overlap between the predicted bounding box and the ground-truth bounding box and can be computed as follows (2):

$$IoU = \frac{Area(B_{pred} \cap B_{gt})}{Area(B_{pred} \cup B_{gt})}$$

where B_{pred} – predicted bounding box, B_{gt} – ground truth bounding box.

The overall detection performance is further evaluated using the mean Average Precision (mAP), which summarizes the precision-recall performance across detection thresholds and can be calculated as follows (3):

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i$$

where N – number of classes, and AP_i – average precision for class i.

After training the model on the «Drone vs Bird» dataset [3], the FP32 weights were obtained. To evaluate the impact of quantization, the model was converted to INT8 and FP16, and a comparison was performed using detection metrics and visual inspection of the results.

Detection accuracy comparison between FP16 and INT8 TensorRT models

Model	Precision	Recall	mAP50	mAP50-95
TensorRT FP16	0.969	0.957	0.980	0.780
TensorRT INT8	0.971	0.953	0.981	0.770

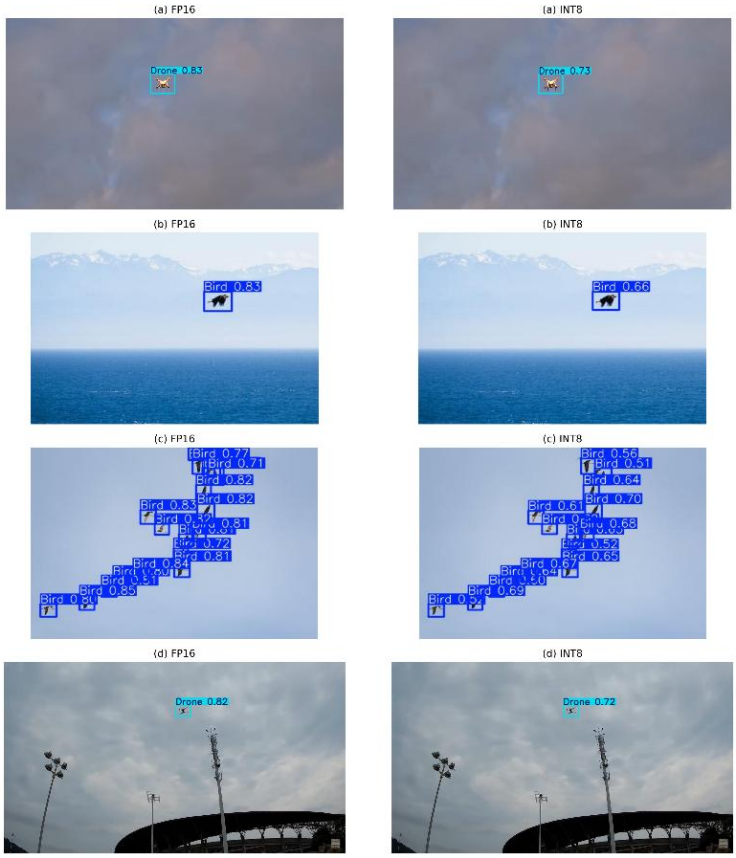

Figure 2 – Comparison of FP16 and INT8

Overall, the results show that quantization has only a minor impact on mAP accuracy, while the INT8 model remains suitable for deployment in research involving neural network integration on edge devices.

References:

1. Adnan W. H., Khamis M. F. Drone use in military and civilian application: risk to national security // Journal of Media and Information Warfare (JMIW). – 2022. – T. 15, № 1. – C. 60–70.

2. Molloy O. Drones in modern warfare: lessons learnt from the war in Ukraine // Australian Army Research Centre. – 2024. – T. 29.

3. drone-vs-bird Dataset [Electronic resource] / dam // Roboflow Universe. – 2023. – Available at: https://universe.roboflow.com/dam-tpuul/drone-vs-bird-lanzg (accessed: 05.03.2026).