A Deep Learning-Based Face Mask Detector for Autonomous Nano-Drones (Student Abstract)

Eiman AlNuaimi[1 2], **Elia Cereda**[3], **Rafail Psiakis**[2],
Suresh Sugumar[2], **Alessandro Giusti**[3], **Daniele Palossi**[3 4]
[1]Department of Electrical Engineering and Computer Science, Khalifa University, United Arab Emirates
[2]Secure Systems Research Centre (SSRC), TII, United Arab Emirates
[3]Dalle Molle Institute for Artificial Intelligence (IDSIA), USI-SUPSI, Switzerland
[4]Integrated Systems Laboratory (IIS), ETHZ, Switzerland,
100044615@ku.ac.ae, Eiman.Alnuaimi@tii.ae

Abstract

We present a deep neural network (DNN) for visually classifying whether a person is wearing a protective face mask or not. Our DNN is compatible with a resource-limited, sub-10-cm nano-drone: this robotic platform is an ideal candidate to fly in human proximity and safely perform ubiquitous visual perception tasks. This paper describes our pipeline, including: dataset collection; selection and training of a full-precision (i.e., float32) DNN; network quantization to int8 precision, enabling the DNN's deployment on a parallel ultra-low power (PULP) system-on-chip aboard a nano-drone. Results demonstrate the efficacy of our pipeline with a mean area under the ROC curve score of 0.81, which drops by only $\sim$2% when quantized to 8-bit for deployment.

Introduction

The COVID19 outbreak has shown the importance of quick social reactions, including monitoring the use of protective face masks among the population. With their sub-10-cm size and a few tens of grams in weight, nano-sized unmanned aerial vehicles (UAVs) are the ideal candidates to safely fly in human proximity and perform ubiquitous visual perception tasks (Palossi et al. 2021), such as the mask detection use case we address in this work (see Figure 1). However, enabling high-level sensing capabilities on nano-UAVs, i.e., without relying on off-board computational resources, is hindered by their small form factor, which limits the on-board computational/memory/sensory resources to the class of ultra-low-power devices. Convolutional neural networks (CNNs) are essential for visual pattern recognition tasks: with reduced memory and computational requirements, these models are a perfect fit for full deployment on resource-constrained embedded platforms, such as those found on small nano-UAVs (Palossi et al. 2021).

In this work, we present a novel CNN for face mask detection, based on MobileNetV2 (Sandler et al. 2018), and conceived to run on a parallel ultra-low power (PULP) GAP8 System-on-Chip (SoC) aboard a commercial Crazyflie nano-UAV. We propose two models: one trained *from-scratch* on our custom dataset collected for the specific mask detection task, and one consisting of a pre-trained model on Imagenet, then *fine-tuned* on our dataset. Then, we apply a quantization stage, to both models, to make their execution compatible with the target SoC. Finally, we contribute *i*) a quantitative comparison between the two models and *ii*) an assessment of the classification performance between the full-precision CNN (float 32-bit) and its quantized (int 8-bit) version, using the PyTorch framework. Our results show that: *i*) the *fine-tuned* model outperforms its *from-scratch* counterpart (30% higher accuracy), and *ii*) the quantized model incurs in a marginal accuracy loss (less than 3%) with respect to the full-precision baseline. Ultimately, our work paves the way towards the full deployment of our CNNs into the PULP SoC available aboard our miniaturized robotic platform.

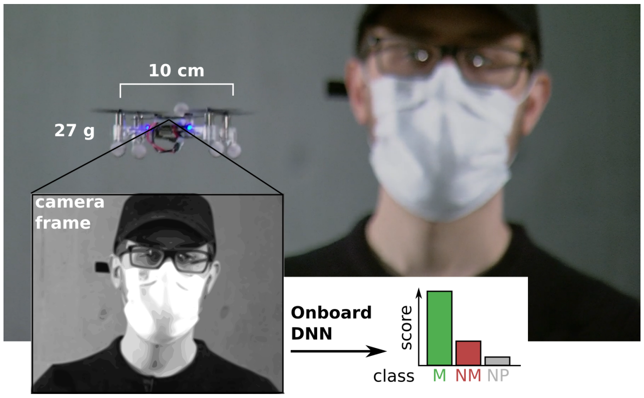

Figure 1: Use case: face mask detection on nano-drones.

Implementation

Datasets are acquired by shooting $\sim$1-minute videos with a camera handheld at eye level, moved in such a way to capture good variability in terms of backgrounds and lighting directions, with small attitude and height variations to mimic those of typical drone flight. The training set is acquired in 3 indoor environments and built from 15 videos: 6 featuring a different masked subject each (class M); 6 featuring each of these subjects not-masked (class NM); 3 (one per environment) with no person (class NP). The test set is acquired

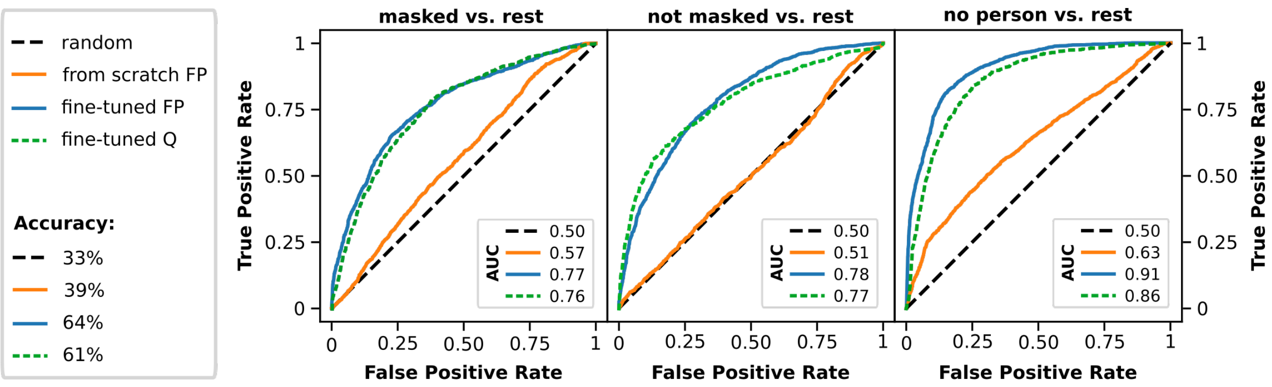

Figure 2: *One vs. rest* AUCs and accuracy of *from-scratch* vs. *fine-tuned* models, full-precision (FP) and quantized (Q).

in 2 different indoor environments featuring 2 new subjects for a total of 6 videos: 2 for each of the three classes. From each video, we extract $\sim$800 frames at $15\,\mathrm{frame/s}$, which yields $\sim$19 k images, 13.5 k for training, and 5.1 k for testing. The three classes (M, NM, NP) have approximately the same prevalence in the testing set.

We use the PyTorch framework to train the two proposed MobileNetV2 models, i.e., *from-scratch* and *fine-tuned*, in full-precision (float 32-bit). Each CNN has $224 \times 224 \times 3$ input and three scalar outputs and uses geometric and photometric data augmentation. Since our envisioned robotic platform features a GreenWaves Technologies GAP8 SoC, which does not provide floating-point hardware support, we need to quantize the full-precision models to 8-bit fixed-point ones. To do so, we use the open-source NEMO library[1] to perform layer-wise quantization, according to the PACT-based quantization strategy (Choi et al. 2018).

Results

In our evaluation, we consider the *masked vs. rest* case, the *not masked vs. rest* counterpart, and the *no person vs. rest* case. The former case measures the ability of the network to discriminate whether there is a correctly masked person in the image. The latter measures whether there is a person at all (masked or not): a simpler task, which represents an upper bound to the achievable performance of *masked vs. rest*. Our comparison is based on two metrics: the prediction accuracy and the area under the receiver operating characteristic curve (AUC).

Figure 2 shows that the *fine-tuned* model vastly outperforms the *from-scratch* one in all cases. On the test set, the former achieves 64% accuracy, and 82% mean AUC, while the latter reaches 39% accuracy and a mean AUC of 57%, which are not far from a random classifier's performance. This significant difference in performance depends on the limited size of our training set (i.e., 19 k images): too small to train an accurate model from scratch, but sufficient to specialize a pre-trained one. Therefore, we proceed with the quantization stage focusing only on the fine-tuned version.

Our second experiment assesses the impact of the quantization process on classification performances. Therefore, we compare the full-precision model (our baseline) against its quantized version. In Figure 2, the quantized model reaches 61% accuracy, and 80% mean AUC, representing a minimal 2% performance drop, in line with expectations of the PACT quantization strategy, making the resulting model compliant with the hardware constraints imposed by our target SoC. In Figure 3, we showcase a few examples from our testing set, coupled with ground-truth labels and predictions of the 8-bit quantized model. Ultimately, we expect the proposed model to achieve an inference throughput of $\sim 3-4$ frame/s when deployed aboard the GAP8 SoC, leaving space for further improvements by *i*) streamlining the MobileNetV2 architecture and *ii*) stronger quantization.

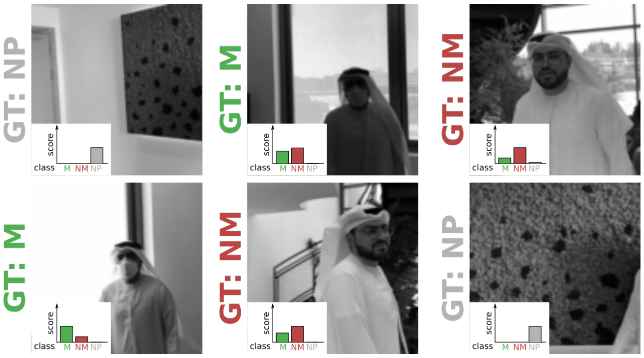

Figure 3: Sample images with ground-truth (GT) and quantized model's predictions (M: Masked; NM: Not Masked; NP: No Person).

References

Choi, J.; et al. 2018. Pact: Parameterized clipping activation for quantized neural networks. *arXiv preprint arXiv:1805.06085*.

Palossi, D.; et al. 2021. Fully Onboard AI-powered Human-Drone Pose Estimation on Ultra-low Power Autonomous Flying Nano-UAVs. *IEEE Internet of Things Journal*, 1–1.

Sandler, M.; et al. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4510–4520.

[1] https://github.com/pulp-platform/nemo