**tinyEOD: Small Deep Neural Networks and Beyond for Embedded Vision Applications**

Christos Kyrkou and Theocharis Theocharides, KIOS Research and Innovation Center of Excellence and Department of Electrical and Computer Engineering, University of Cyprus

Visual edge intelligence is a growing necessity for emerging applications where real-time decision is vital. Object detection in particular, the first step in such applications, achieved tremendous improvements in terms of accuracy due to the emergence of Convolutional Neural Networks (CNNs) and Deep Learning. However, such complex paradigms require extensive resources, which prevents their deployment on resource-constraint mobile and embedded devices that simultaneously need to process high resolution images.

In many applications, the number of targeted object classes is reduced. This provides us with an opportunity to investigate smaller neural network architectures without losing considerable accuracy. We first explore the parameter space (type, number and size of filters, input image size) for convolutional networks to identify the best trade-offs between accuracy, performance, and memory for classification and detection applications. A major challenge when reducing the number and size of filters as well as the size of the input image to extract more performance, is the potential considerable loss of information that negatively affects the accuracy. ***Hence, we introduce a new way to boost the detection accuracy of computationally efficient but resolution-limited CNNs for operating on larger images than their input allows, without changing the underlying network structure.*** The main idea behind this approach is to separate the larger input image into smaller images with size equal to that of the CNN receptive field, called tiles, and selectively process only a subset of them using (a) an attention mechanism and (b) track activity in other tiles using a memory mechanism. In addition, we incorporate domain knowledge and pre-processing filters to further improve the performance and processing time of deep learning models. ***We experimented using a dataset of aerial images, where different models were trained and evaluated on. Our results indicate considerable improvements across multiple applications, ranging from 2-5x speedup across a variety of different platforms while achieving up to 95% detection accuracy. Most importantly, the networks remain relatively small in the range of ~300KBs, while also exhibiting significant power savings.***