Automated Optimization of DNN Models for People Counting with Infrared Sensors Using NAS

Student: Seyed Morteza Mollaei

Supervisors: Prof. Daniele JAHIER PAGLIARI, Dr. Chen XIE, Dr. Matteo RISSO, Dr. Francesco DAGHERO, Dr. Alessio BURRELLO

Introduction

Neural Architecture Search (NAS) has become increasingly popular in recent years due to its outstanding performances in finding automatically accurate models. However, the application of NAS to tasks different from classical Image Recognition has been less explored. This work focuses on the application of a NAS tool, named Pruning In Time (PIT) on the less explored task of people counting based on Ultra-low-resolution Infrared (IR) array sensors. The latter have become increasingly popular due to their low-cost and privacy-preserving nature, as the low resolution does not allow to recognize specific features. However, these sensors are generally too computational and memory bound to run Deep Learning (DL) models locally. Leveraging PIT, we explore 4 DL architectures, finding several models that represent different trade-offs between model size and accuracy. In this way, we show that, when compared to traditional techniques and previous DL approaches, our models improve not only the accuracy up to 2.85% but also 44% decrease of the memory footprint, and at iso-accuracy this reduction goes further to 61.6%. Finally, these enhanced DL models may operate in real time on low-power Internet of Things (IoT) nodes, while also being lightweight in terms of energy, thus requiring less frequent recharges in battery operated devices.

Proposed Method

In this work, we benchmark the effectiveness of the PIT-NAS on 8 state-of-the-art models, handcrafted specifically for a state-of-the-art IR dataset featuring 8x8 arrays, named LINAIGE. By changing the regularization strength, we spawn several models, each representing a different trade-off between accuracy and number of parameters.

The first architecture we benchmark is a single-frame Convolutional Neural Network (CNN), taking as input a frame and predicting the number of people present in it. Noteworthy, while this model is extremely lightweight, it does not exploit the temporal correlation of the different frames (shown in Figure 1), leading to poor accuracy. As a solution, we stack a window W of frames, feeding them as a multi-channel image to a CNN. This multi-frame approach grants increased accuracy, however it requires more computations and additional parameters, growing rapidly with W. An alternative architecture is the Temporal Convolutional Network, belonging to the CNNs family, but being tailored for time-series. This architecture can handle large W thanks to dilation, however, as shown in the results, its accuracy rapidly

deteriorates when its size is pruned with PIT. Finally, we apply a Majority-Voting approach to the single-frame CNN. We deploy a network using a single frame as input, but we use the previous W-1 predictions, together with the current one, for a voting mechanism, selecting the most frequently predicted class.

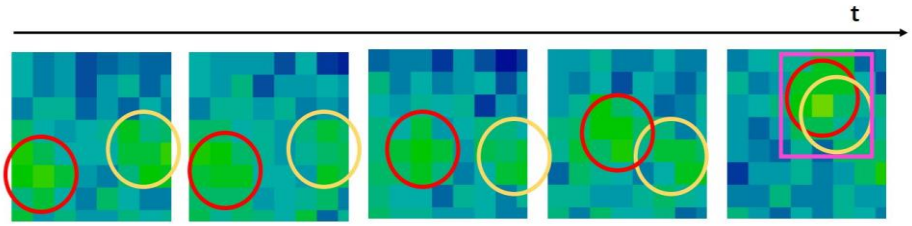

Fig. 1. An example of an IR frame sequence related to two persons moving in close proximity to each other.

As an alternative to handcrafted seed models, in this work we also explore the application of a cascade of two NAS techniques. The first, named SuperNet, allows a coarse exploration of the search space, selecting among few layer types (e.g, depthwise versus standard convolutions) to balance accuracy and model size. Afterward, we apply PIT on the output model of the first stage, that in turn, works on a finer search space.

The advantage of this approach lies in the design speed, as given a new IR dataset, it is not necessary to perform a time-consuming handcrafted search of the hyperparameters, as for the other seed models previously described.

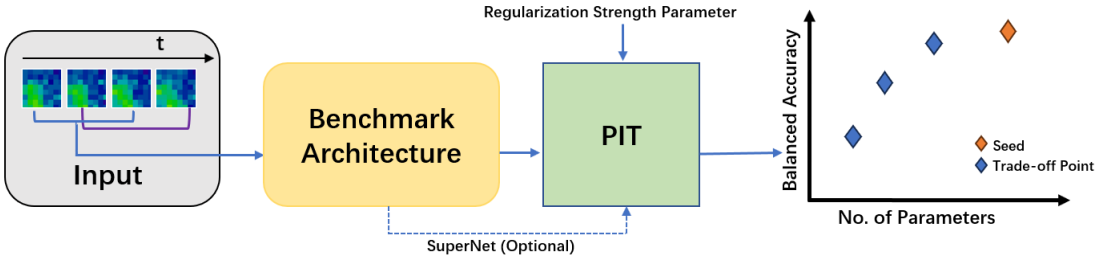

Fig. 2. Discovery of Pareto-Optimal points.

Results and Conclusion

Figure 3 and 4 shows the Pareto curves in terms of total parameters versus balanced accuracy. Each point of these curves is a different model found by changing the regularization strength of PIT. Most of the models deliver satisfactory accuracy with fewer number of parameters w.r.t. their initial seed results. Notably, the Majority-Voting CNN outperforms the other architectures where it obtains iso-accuracy with the seed model while introducing a reduction of nearly 62% in size. Another interesting achievement by this model is to exceed the seed model balanced accuracy by 3.65% with 14.4% decrease in size (highest point). The TCN-based CNN underperforms the others, with an increase of 26% parameters w.r.t to the Majority Voting CNN while introducing 9.1% less accuracy. Regarding SuperNet, our best

performance is achieved with SuperNet-medium, where we attain 10.1% less balanced accuracy than the seed but requiring 90.5% less parameters. In addition, the top point of SuperNet-small-α1 reaches almost 5% less accuracy w.r.t the highest point of SuperNet-medium requiring 26% more parameters.

This work shows that DL improves the accuracy with smaller model size when compared to previous algorithms in this setting. When comparing our results with other state-of-the-art DL approaches for people counting, we can achieve up to 2.85% improved accuracy, while reducing the total number of parameters by up to 44%.

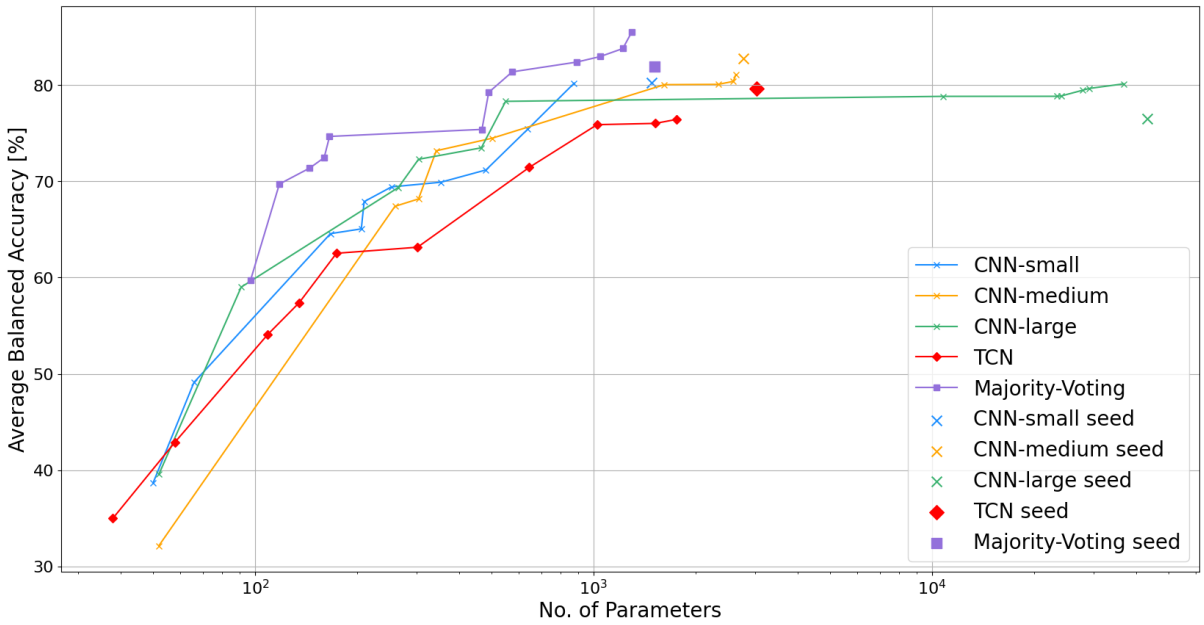

Fig. 3. Pareto-Optimal curves of different CNN seed models obtained by applying PIT.

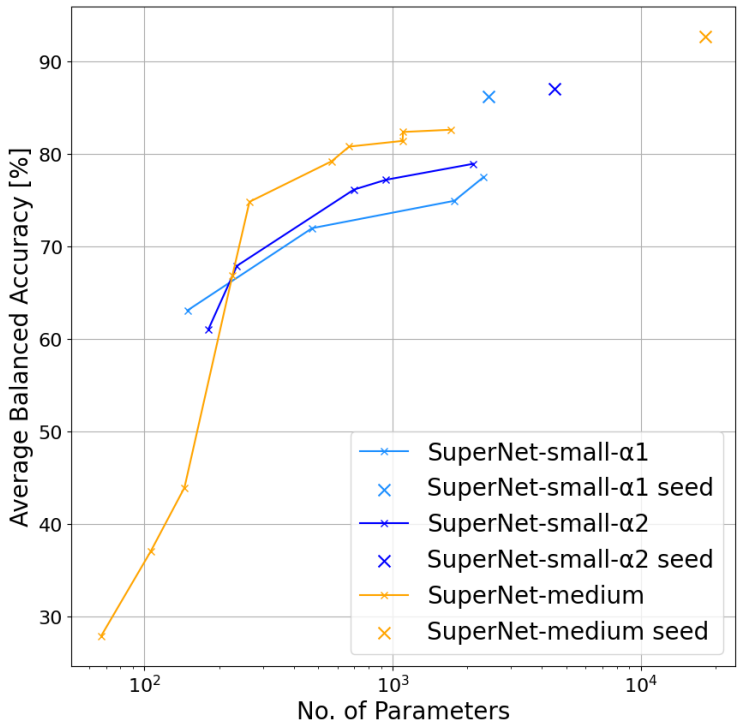

Fig. 4. Pareto-Optimal curves of different SuperNet seed models obtained by applying PIT.