Efficient Deep Network Architecture for Vision-Based Vehicle Detection

Keyvan Kasiri, Mohammad Javad Shafiee, Francis Li — University of Waterloo, ON, Canada
Alexander Wong — University of Waterloo, ON, Canada
Justin Eichel — Miovision Technologies Inc., ON, Canada,

Abstract

With the progress in intelligent transportation systems in smart cities, vision-based vehicle detection is becoming an important issue in the vision-based surveillance systems. With the advent of the big data era, deep learning methods have been increasingly employed in the detection, classification, and recognition applications due to their performance accuracy, however, there are still major concerns regarding deployment of such methods in embedded applications. This paper offers an efficient process leveraging the idea of evolutionary deep intelligence on a state-of-the-art deep neural network. Using this approach, the deep neural network is evolved towards a highly sparse set of synaptic weights and clusters. Experimental results for the task of vehicle detection demonstrate that the evolved deep neural network can achieve a substantial improvement in architecture efficiency adapting for GPU-accelerated applications without significant sacrifices in detection accuracy. The architectural efficiency of $\backsim$4X-fold and $\backsim$2X-fold decrease is obtained in synaptic weights and clusters, respectively, while the accuracy of 92.8% (drop of less than 4% compared to the original network model) is achieved. Detection results and network efficiency for the vehicular application are promising, and opens the door to a wider range of applications in deep learning.

1 Introduction

With the growth of smart cities, the intelligent transportation system (ITS) and technologies deployed in ITS have been developed and updated constantly [1]. As vision-based surveillance devices have been increasingly equipped to transportation systems, visual vehicle detection has become a crucial part in ITS systems as well as a major research issue in topics from traffic analysis to building and deployment real-time surveillance systems [2]. Although new technologies in the computer vision and image processing have drastically been developed recently, deployable real-time methods are still strongly needed for ITS applications, particularly for advanced transportation system and traffic management.

Having a wide range of traditional and new emerging visual traffic sensors has led to collecting enormous amount of transportation data. With the availability of large scale video datasets, learning-based approaches have been significantly improved towards the application of vehicle and pedestrian detection, classification, recognition, and so on. Recently, deep learning methods have drawn a lot of interest from both academic and industrial sides, due to its substantial improvement over the state-of-the-art computer vision methods in many detection and classification applications. Numerous deep neural network models have been proposed recently to accurately detect vehicles and pedestrians [3–5].

Although new deep learning solutions have elevated the performance of detection and classification methods, deployment cost of such approaches is still a major concern in ITS applications, meaning that computational ability, real-time performance requirement, and memory capacity have restricted deep neural networks to be applied in widely in vehicle detection applications. Traditional deep networks suffer from a big structure with structural redundancy, leading to increasing the required memory as well as training and decision time, which is in contrast with the requirements for deploying vehicle detection algorithms in embedded platforms.

Recently, there has been a wide attempt towards attaining efficient deep neural networks in terms of storage, memory bandwidth, computational resources, and power consumption. As there is significant redundancy in parameters of the deep networks, various approaches have been proposed to reduce the redundancy and therefore decrease the amount of computations and memory required. Compressed deep neural network using vector quantization with 1% accuracy loss was proposed by Gong et al. [6]. Network pruning has been also used to lower the network complexity as well as over-fitting. Examples of such approaches to pruning are proposed in [7, 8]. A combination of quantization and pruning was used in [9] to get further improvement. Recently, Shafiee et al. [10, 11] tackled this problem by offering a new framework inspired by the evolutionary approach for synthesizing highly efficient deep neural networks. The proposed evolutionary deep approach follows biological evolution mechanisms to mimic random mutation, natural selection, and heredity in synthesizing successive generations of network models, and as the result more efficient network architectures was achieved.

In this paper, by taking advantage of the evolutionary deep intelligence framework [11], an efficient architecture of deep neural network is presented to improve the efficiency of the deep model based on the architecture presented by Luo et al. [12]. Although the Luo's deep model reached a significant speed gain compared to other state-of-the-art deep learning methods, its size consumes considerable memory and computational resources. Therefore, it is highly desired to deploy such deep learning method in such a way that those resource demands will not become restrictive for embedded applications. The network is trained towards yielding a highly sparse set of synaptic weights and clusters across successive generations of evolution. The generated model is trained and validated over large datasets for the application of vehicle detection demonstrates.

2 Methodology

In this paper, an efficient architecture for a deep neural network model is presented for the vehicle detection application. The network architecture, on the basis of a non-local deep feature (NLDF) model [12], is employed as the detector and the evolutionary synthesis of deep neural networks (Evo-net) [10] is used to achieve an optimized architecture.

The architecture proposed in [12], as a high performance saliency detection method, is based on a 5×4 grid of multiscale convolution and deconvolution blocks to capture local and global context as well as features at different scales of resolution. As illustrated in Fig. 1, the first row consists of a set of five convolutional blocks derived from VGG-16 to learn the global feature map. In the second row of the grid, five more convolutional blocks are used to compute features at each specific resolution. The third row is formed by a set of five contrast layers to emphasize on features with strong local contrast at the specific resolution. In the last row, a set of four deconvolutional blocks is used to up-scale the features to the desired output size, as well as a block to construct the final local feature map. At the end two convolutional layers are employed for the score block by fusing local and global feature maps. Inspiring by the Mumfordshah functional, the idea of penalizing errors on the boundary of objects is used to form a loss function in the convolutional neural network.

The method, employed to achieve an efficient model out of the NLDF deep network [12], is based on the Evo-Net [10] inspiring from biological evolution. In this approach the network model evolves in successive generations into highly efficient deep neural networks. The architectural evolution is formulated using a synaptic probability model, in which new descendant version of network is synthesized based on these synaptic probability models from the ancestor network, as well as computational environmental factors in a random manner. This approach tries to mimic heredity, natural selection, and random mutation from biological evolution.

The genetic encoding approach for the network architecture $\mathcal{H}$ with a set of possible synapses S and a set of the synaptic strength $\mathcal{W}$ is formulated as a conditional probability of the network architecture in generation g given the architecture of its ancestor in generation $g-1$,

$$P(\mathcal{H}_g) = \mathcal{F}(\mathcal{E}) \cdot P(S_g | \mathcal{W}_{g-1}), \tag{1}$$

where $\mathcal{F}(\mathcal{E})$ models the environmental factor, which computationally restricts resources available to descendant networks. The term $\mathcal{F}(\mathcal{E})$ constrains the number of synapses that can be synthesized in the descendant network and is set to $\mathcal{F}(\mathcal{E}) = K$, where the quantity K enforces the highest percentage of synapses desired in the descendant network.

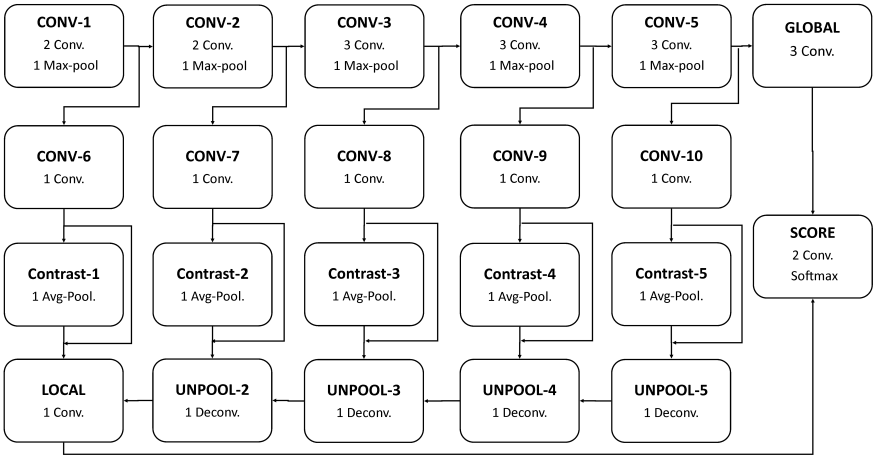

Fig. 1: Architecture of the deep convolutional neural network reproduced from the NLDF [12].

As a more efficient genetic encoding scheme, synaptic clustering proposed in [11] is incorporated to improve the memory and storage requirements, as well as adaptability for parallel computations such as embedded GPUs. Therefore, the synthesis procedure in Eq. 2 is reformulated as

$$P(\mathcal{H}_g) = \prod_{c \in C} [\mathcal{F}_c(\mathcal{E}) P(\bar{s}_{g,c} | \mathcal{W}_{g-1}) \cdot \prod_{i \in c} \mathcal{F}_s(\mathcal{E}) P(s_{g,i} | w_{g-1,i})], \quad (2)$$

where $\mathcal{F}_c(\cdot)$ and $\mathcal{F}_s(\cdot)$ denote the environmental factors enforced at the cluster and synapse levels, respectively. In this equation $s_{g,c} \in S_g$ and $\bar{s}_{g,c} \subset S_g$ represent a particular synapse and a particular cluster of synapses for a given generation g and cluster c, and $w_{g-1,i} \in W_{g-1}$. The particular synaptic cluster in a deep convolutional architecture can be any subset of synapses such as a kernel or a set of kernels.

The synthesized descendant networks are trained and the evolutionary synthesis process is successively repeated to attain successive generations of descendant networks.

3 Experimental Results and Discussion

In order to investigate the efficacy of the presented architecture, the evolutionary synthesis across several generations is performed and the performance of the network for the application of vehicle detection is evaluated.

The dataset of vehicles of all kinds as well as pedestrians and bicycles provided in [13] is used for training the models across different generations. Examples of captured images from this database are shown in Fig. 2. The foreground objects have been identified to provide the corresponding ground truth maps, where the salient objects are labelled with pixel-wise annotation. The detection results are validated using a separate dataset of $25K$ samples of vehicles, bicycles, and pedestrians.

The deep network architecture in [12] is to used to learn discriminant saliency features. The input image to the model is a 352×352 image and the output is a 176×176 saliency map, which is resized to the size of input image using a bilinear interpolation. The network model was implemented in TensorFlow, and was initialized with the pretrained weights of the model in [12]. The Adam optimizer was employed to train the model with the default parameters, learning rate of 10^{-6}, $\beta_1 = 0.9$, and $\beta 2 = 0.999$. The other network settings are set to the default values suggested in the original model in [12]. The inputs were resized to 352×352 for training.

To assess the performance of saliency detection for the evolved deep neural networks at different generations, the detection accuracy metric is computed for each generation. The efficiency of the network architecture over successive generations is presented as synaptic and cluster efficiency for the convolutional and deconvolutional layers. The synaptic efficacy of the architecture is described as the total number of synapses in the original network divided by the one in the network of the current generation. Similarly, cluster efficiency is defined as ration of the number of clusters in the original network over that of the current synthesize network.

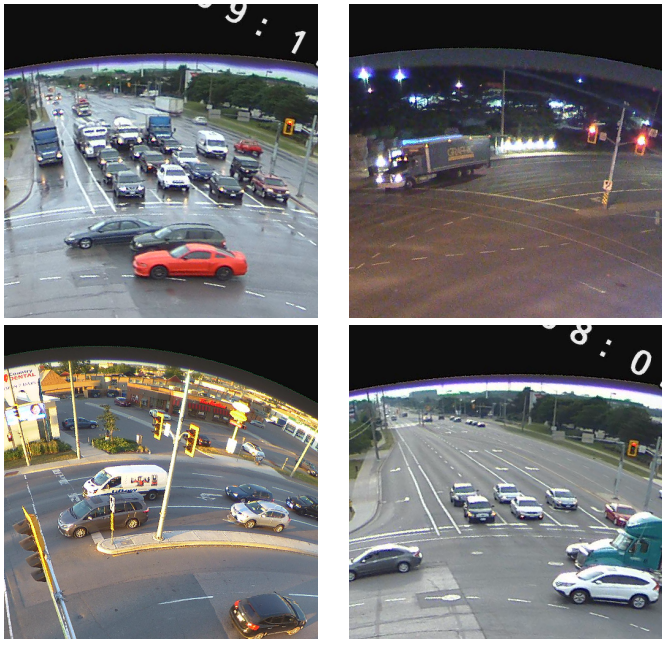

Fig. 2: Sample images from the dataset [13] used for training.

The deep neural network with the above settings is trained over a number of generations. At every generation, the environment factor is forced to the model to yield progressively more efficient network architectures while maintaining the modelling accuracy. In this set of preliminary experiments, each kernel is considered as a synaptic cluster in the synapse probability model. Fig. 3 shows the architectural efficiency of the network over a number of generations. It can be observed that over successive generations of evolution, the architectural efficiency is improved. Particularly, after six generations, the number of preserved synapses and synaptic clusters goes down to less than half and around a quarter of the ones for the original network.

The detection accuracy of the deep neural network in the corresponding generations is shown in Fig. 4. As is observed, the accuracy of the model is decreasing over generations, and after six generations, the cost would be accuracy drop of less than %4 compared to the original network at Generation 0. Sample results of the saliency map detection is shown in Fig. 5, for which, the synthesized network at the sixth generation was employed. In this figure, the purple regions show where the neural network specifies as the detected regions. As is observed, the model is able to clearly detect cars and the motorbike in these samples.

To investigate the effect of network reduction on different layers of the deep neural network, architectural efficiency is considered for each convolutional and deconvolutional layers separately. Table 1 shows the synaptic and cluster efficiency of the synthesized deep neural network obtained after sixth generations. It can be seen that

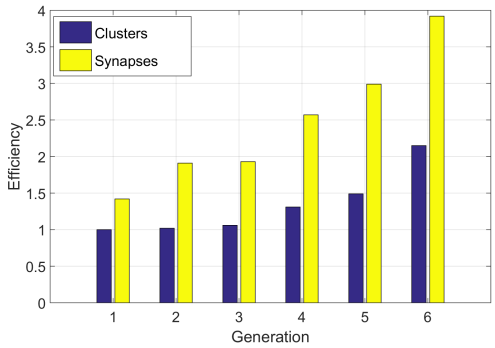

Fig. 3: Efficiency of the network architecture over different generations for synthesized deep networks.

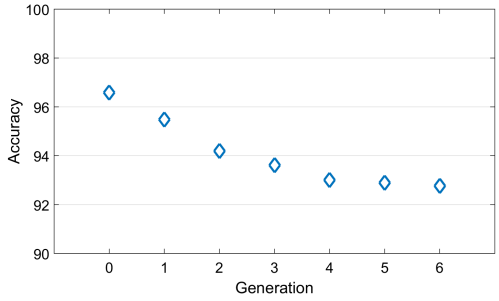

Fig. 4: Accuracy of the deep neural network for saliency map detection over different generations.

all layers ended up with almost the same amount of synapse reduction. However, for the cluster efficiency, reduction is more substantial for the first set of five convolutional blocks. The convolutional layers Conv-6 to Conv-10, which take care of computing features, preserve higher percentage of synaptic clusters compared to the ones on the first row of the network grid.

Quantitative and qualitative assessment of the experiments for the task of vehicle detection demonstrate the architecture efficiency of the network while maintaining detection accuracy of the model. Based on the results in this paper, applying Evo-Net to the NLDF network significantly improves the architecture efficiency of the model in formation of highly sparse synaptic weights and clusters, and therefore facilitates the adaptation for highly parallel computations such as GPUs. Yielding a highly efficient, yet powerful deep models for vehicular applications can lead to a promising direction for future exploration in embedded deep learning.

Acknowledgments

This research is funded by the SOSCIP TalentEdge Post-doctoral Fellowship Program in partnership with Ontario Centres of Excellence (OCE), and supported by Miovision Technologies Inc.

References

[1] Zhang, N., Wang, F. Y., Zhu, F., Zhao, D., Tang, S. DynaCAS: Computational experiments and decision support for ITS. *IEEE Intelligent Systems* (2008).

[2] Zhang, J., Wang, F. Y., Wang, K., Lin, W. H., Xu, X., Chen, C. Data-driven intelligent transportation systems: A survey. *IEEE Trans. Intelligent Transportation Systems* (2011).

[3] Hai, W., Cai, Y., and Chen, L. A vehicle detection algorithm based on deep belief network. *The scientific world journal* (2014).

[4] Yu, S., et al. A model for fine-grained vehicle classification based on deep learning. *Neurocomputing* (2017).

[5] Lv, Y., et al. Traffic flow prediction with big data: a deep learning approach. *IEEE Trans. on Intelligent Transportation Systems* (2015).

[6] Gong, Y., Liu, L., Yang, M., and Bourdev, L. Compressing deep convolutional networks using vector quantization. arXiv preprint arXiv:1412.6115 (2014).

Fig. 5: Sample results of the saliency map detection for the synthesized network at the sixth generation. The purple regions show where the deep neural network specifies as the detected regions.

Table 1: Synaptic and cluster efficiency of the convolutional and deconvolutional layers at the sixth generation of the synthesized deep neural network.

Layers	synaptic efficiency	cluster efficiency
Conv-1-1	3.31X	2.91X
Conv-1-2	3.80X	2.46X
Conv-2-1	3.86X	2.91X
Conv-2-2	3.88X	4.27X
Conv-3-1	3.87X	3.28X
Conv-3-2	3.87X	2.75X
Conv-3-3	3.87X	2.10X
Conv-4-1	3.88X	2.69X
Conv-4-2	3.88X	2.30X
Conv-4-3	3.88X	2.75X
Conv-5-1	3.87X	2.14X
Conv-5-2	3.88X	2.47X
Conv-5-3	3.88X	2.38X
Conv-6	3.50X	1.66X
Conv-7	3.50X	1.78X
Conv-8	3.62X	1.80X
Conv-9	3.68X	2.00X
Conv-10	3.63X	2.03X
Unpool-5	3.77X	1.79X
Unpool-4	3.79X	1.96X
Unpool-3	3.79X	1.93X
Unpool-2	3.78X	1.88X
Local score	3.44X	1.00X
Global-1	3.38X	1.41X
Global-2	3.39X	1.78X
Global-3	3.48X	1.83X
Global score	3.32X	1.00X

[7] Hanson, S. J. and Pratt, L. Y. Comparing biases for minimal network construction with back-propagation. *Advances in neural information processing systems* (1989).

[8] Le Cun, Y., Denker, J. S., and Solla, S. A. Optimal brain damage. *Advances in Neural Information Processing Systems* (1990).

[9] Han, S., Mao, H., and Dally, W. J. Deep compression: Compressing deep neural network with pruning, trained quantization and huffman coding. arXiv preprint arXiv:1510.00149 (2015).

[10] Shafiee, M. J., Mishra, A., Wong, A. Deep learning with Darwin: Evolutionary synthesis of deep neural networks. arXiv preprint arXiv:1606.04393 (2016).

[11] Shafiee, M. J. and Wong, A. Evolutionary synthesis of deep neural networks via synaptic cluster-driven genetic encoding. NIPS Workshop (2016).

[12] Luo, Z and Mishra, A. and Achkar, A. and Eichel, J. and Li, S. Jodoin, P. M. Non-Local Deep Features for Salient Object Detection. *IEEE CVPR* (2017).

[13] MIOvision Traffic Camera Dataset (MIO-TCD), CVPR Traffic Surveillance Workshop and Challenge, http://podoce.dinf.usherbrooke.ca/challenge/dataset/