




# Optimizing Convolutional Neural Networks for Embedded Systems by Means of Neuroevolution

Filip Badan and Lukas Sekanina<sup>(✉)</sup> 

Faculty of Information Technology IT4Innovations Centre of Excellence,  
Brno University of Technology, Božetěchova 2, 612 66 Brno, Czech Republic  
badan.filip@gmail.com, sekanina@fit.vutbr.cz

**Abstract.** Automated design methods for convolutional neural networks (CNNs) have recently been developed in order to increase the design productivity. We propose a neuroevolution method capable of evolving and optimizing CNNs with respect to the classification error and CNN complexity (expressed as the number of tunable CNN parameters), in which the inference phase can partly be executed using fixed point operations to further reduce power consumption. Experimental results are obtained with TinyDNN framework and presented using two common image classification benchmark problems – MNIST and CIFAR-10.

**Keywords:** Evolutionary Algorithm · Convolutional neural network · Neuroevolution · Embedded Systems · Energy Efficiency

## 1 Introduction

*Deep neural networks* (DNNs) currently show an outstanding performance in challenging problems of image, speech and natural language processing as well as in many other applications of machine learning. The design of high-quality DNNs is a hard task even for experienced designers because the state of the art DNNs have large and complex structures with millions of tunable parameters [4, 11]. Automated DNN design approaches, often referred to as the *Neural Architecture Search* (NAS), that have recently been developed, provide networks comparable with DNNs created by human designers.

This paper deals with automated design and optimization of *convolutional neural networks* (CNN), a subclass of DNNs primarily utilized for image classification. Our objective is to design and optimize not only with respect to the classification error, but also with respect to hardware resources needed when the final (trained) CNN is implemented in an embedded system with limited resources. As energy-efficient machine learning is a highly desired technology, various *approximate implementations* of CNNs have been introduced [2, 7]. Contrasted to the existing neuroevolutionary approaches trying to minimize the classification error as much as possible and assuming that CNN is executed using

floating point (FP) operations on a Graphical Processing Unit (GPU) [1,3], our target is a highly optimized CNN whose major parts are executed with reduced precision in fixed point (FX) arithmetic operations.

We propose EA4CNN (Evolutionary Algorithms for Convolutional Neural Networks) – a neuroevolution platform capable of evolving and optimizing CNNs with respect to the classification error and model complexity (expressed as the number of tunable CNN parameters), in which the inference phase can partly be executed using FX operations. One of our goals is to demonstrate that the proposed method is capable of reducing the number of parameters of an already trained CNN and, at the same time, providing good tradeoffs between the classification error and CNN complexity. Experimental results are obtained with TinyDNN framework [6] and presented using two common benchmark problems – the classification of MNIST and CIFAR-10 data sets.

## 2 Related Work

Image classification conducted by CNNs is the state of the approach in the image processing domain. CNNs usually contain from four to tens layers of different types [11]. *Convolutional layers* are capable of extracting useful features from the input data. In these layers, each neuron is connected to a subset of inputs with the same spatial dimensions as the tunable kernels. The convolution is computed as  $y = b + \sum_i \sum_j \sum_k (\mathbf{x}_{i,j,k} \cdot \mathbf{w}_{i,j,k})$ , where  $\mathbf{x}$  is the input subset,  $\mathbf{w}$  is the convolution kernel and  $b$  is a scalar bias. *Pooling layers* combine, e.g. by means of averaging, a set of input values into a small number of output values to reduce the network complexity. *Fully connected* (FC) layers are composed of artificial neurons; each of them sums weighted input signals (coming from a previous layer) and produces a single output. Convolutional layers and fully connected layers are typically followed by non-linear *activation functions* such as  $\tanh(\cdot)$  or rectified linear units (ReLU). The structure of the network is defined by hyperparameters (e.g., the number of layers, filters etc.) and this structure also determines the number of tunable parameters (weights and neuron biases). Modern CNNs also utilize normalization layers, residual connections, dropout layers etc. (see [4,11]).

In the *training phase*, the objective is to optimize the CNN parameters in order to minimize a given *error metric*. The training is a time-consuming iterative procedure which is typically implemented with the standard FP number representation. A trained CNN is then used, for example, for classification, in which an input image (a set of pixels) is classified to one of several classes. This (feed-forward) procedure is called *inference* and only this procedure is typically implemented in low power hardware CNN accelerators [11].

In order to automatically design the architecture (hyperparameters) and the parameters of CNNs, machine learning as well as evolutionary approaches have been proposed. Evolutionary design of neural networks (the so-called *neuroevolution*) that was introduced three decades ago [9], is now being extended for CNN design [5]. As both CNN training and evolutionary optimization are very computationally expensive methods, the key problem of the current neuroevolution

research is to reduce computational requirements and provide competitive CNNs with respect to the human-created CNNs. Most papers are focused on single-objective automated design methods, where the main goal is to minimize the classification error of CNN running on a GPU [5, 8, 10]. Recent works have been focused on multi-objective approaches in which the error is optimized together with the computation requirements [1, 3], but again, for GPU-based platforms. Evolved CNNs are now competitive with human-created CNNs for some challenging data sets; for example, some evolved CNNs achieve a 95% accuracy on CIFAR-10 data set. Note that a CNN with more than 1 million parameters is required in order to reach this accuracy and its training can take days on a GPU cluster [8].

Another research direction is focused on energy efficient (hardware) implementations of CNNs – with the aim of deploying advanced machine learning methods to low power systems such as mobile devices and IoT nodes. The most popular approach is to introduce approximate computing techniques to CNNs and benefit from the fact that the applications utilizing CNNs are highly error resilient (i.e., a huge reduction in energy consumption can be obtained for an acceptable loss in accuracy) [7]. Approximate implementations of CNNs are based on various techniques such as innovative hardware architectures of CNN accelerators, simplified data representation, pruning of less significant neurons, approximate arithmetic operations, approximate memory access, weight compression and “in memory” computing [2, 7, 11]. For example, employing the FX operations has many advantages such as reduced (i) power consumption per arithmetic operation, (ii) memory capacity needed to store the weights and (iii) processor-memory data transfer time.

To best of our knowledge, there has been no research on fully automated design of approximate CNNs by means of neuroevolution. As this is a very computationally expensive approach, we will focus this initial study on automated approximation of middle-size CNNs. Our method is based on simplifying the CNN architecture and reducing the precision of arithmetic operations.

### 3 CNN Design and Optimization with Neuroevolution

The proposed EA4CNN framework exploits an evolutionary algorithm (EA) and TinyDNN library for the design and optimization of CNN-based image classifiers. TinyDNN was chosen because it can easily be modified with respect to the requirements of EA. TinyDNN can, however, be replaced by another suitable CNN library because EA4CNN provides a general interface between EA and CNN implementations. EA4CNN is able to optimize and approximate an existing CNN, but it can also evolve a new CNN from scratch. CNN parameters as well as hyperparameters are optimized together.

#### 3.1 Evolutionary Algorithm

Algorithm 1 presents the EA developed for the design and optimization of CNNs. The EA is initialized with existing or randomly generated CNNs (line 1 in

Algorithm 1) and runs for  $G_{max}$  generations (line 3). It employs a two-member tournament selection (line 6) to determine the parents that later undergo crossover (line 7; with probability  $p_c$ ; see Sect. 3.3 for details) and mutation (line 8; with probability  $p_m$ , see Sect. 3.3). All offspring are continuously stored to the  $Q$  set (line 9) and undergo a training process implemented in TinyDNN (line 11).

Every new population is composed of the individuals selected from the sets of parents ( $P$ ) and offspring ( $Q$ ). The replacement algorithm (line 14) uses a simple speciation mechanism based on the CNN age (Sect. 3.3). To prevent the overfitting, the data set is divided into three parts – training set  $D_{train}$ , test set  $D_{test}$  and validation set  $D_{val}$ . During the evolution, candidate individuals are trained using  $D_{train}$  (line 11), but their fitness score is determined using  $D_{eval}$  (lines 2 and 13). At the end of the evolution process, the best solution is evaluated on the validation set  $D_{val}$  and this result is reported.

---

**Algorithm 1.** Neuroevolution
 

---

```

1:  $P$  = Create Initial Population; // randomly or using existing CNN
2: Evaluate( $P, D_{test}$ ) using TinyDNN;  $i = 0$ ;
3: while ( $i < G_{max}$ ) do
4:    $Q = \emptyset$ ; // a set of offspring
5:   while ( $|P| \neq |Q|$ ) do
6:     ( $a, b$ ) = Tournament Selection ( $P$ );
7:     ( $a', b'$ ) = Crossover( $a, b, p_{cross}$ );
8:      $a'' =$  Mutation( $a', p_{mut}$ );  $b'' =$  Mutation( $b', p_{mut}$ );
9:      $Q = Q \cup \{a''\} \cup \{b''\}$ ;
10:  end while
11:  Run TinyDNN's Training Algorithm for all NNs in  $Q$  with  $D_{train}$ ;
12:  Update the Age counter for all NNs.
13:  Evaluate( $Q, D_{test}$ ) using TinyDNN;
14:   $P$  = Replacement With Speciation ( $P, Q$ );
15:   $i = i + 1$ ;
16: end while

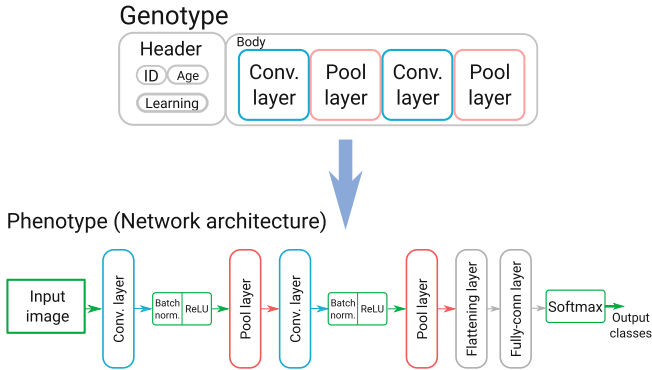
```

---

### 3.2 CNN Encoding

A candidate CNN is represented in the chromosome as a variable-length list of layers with a header containing the chromosome identifier, the age and the learning rate. Two types of layers can occur in the chromosome: (1) *Convolutional layer* with hyperparameters: kernel size, number of filters, stride size and padding. (2) *Pooling layer* with hyperparameters: stride size, subsampling type and subsampling size.

Each convolutional layer is (obligatorily) followed by a batch normalization and ReLU activation. The last (obligatory) layers of each CNN are a convolutional flattening layer and a fully connected layer, followed by a softmax activation to obtain a classifier. These layers are not represented in the chromosome as shown in the example of genotype-phenotype mapping in Fig. 1.



**Fig. 1.** Example of the genotype-phenotype mapping, where some parts of CNN (such as the flattening, fully connected and softmax layers) are not directly represented in the genotype.

### 3.3 Genetic Operators

The mutation operator is applied with the probability  $p_m$  per individual. One of the following mutation options (MO) is chosen with a predefined probability:

1. MO1: Weight reset – all weights of a given layer are randomly generated.
2. MO2: Add a new layer – a randomly generated layer (with randomly generated hyperparameters) is inserted on a randomly chosen position in CNN.
3. MO3: Remove layer – one layer is removed from a randomly chosen position.
4. MO4: Modify layer – some parameters of a randomly selected layer are randomly modified.
5. MO5: Modify hyperparameters of the fully connected layer – the number of connections in the last fully connected layer is increased or decreased.
6. MO6: Modify the learning rate (randomly).

We use a simple one-point crossover operator on each pair of parents obtained with the tournament selection. If a CNN layer is modified by a genetic operator, EA4CNN automatically ensures its correct connection to the previous/next layer. For example, superfluous weights are cut off or missing weights are added and randomly initialized.

### 3.4 Training and Evaluation of Candidate CNNs

As some candidate CNNs exist for many generations while others exist only for a short time, these long-lived CNNs have more opportunities for a good training (line 11 in Algorithm 1). It turns out that candidate CNNs do not have, in principle, the same chance during the selection and replacement process. Hence, inspired in [9], we introduced a speciation mechanism based on the *network age*. A species is defined by all individuals having the same age. The age is increased with every new training process a given candidate CNN undergoes. We define

$age_{max}$  as the maximum age a candidate network can obtain even if it undergoes more than  $age_{max}$  training exercises. The reason for introducing this limit is to increase the selection pressure for networks that were trained many times. A typical setup of  $age_{max}$  is  $\sim |P|/2$ . On the other hand, the network age is reset to the initial value if a given CNN is changed and its fitness is decreased as a consequence of crossover or mutation, e.g., after inserting or removing some layer(s) or changing parameters of the layer. The replacement is independently performed for all selected age levels; for example, if there are 5 age levels and the population size is 15 then 3 best-performing candidate CNNs are selected for each age level and copied to the new population. This algorithm is implemented by ‘Replacement With Speciation’ on line 14 in Algorithm 1.

For the new candidate individuals that are created by mutation or crossover, the principles of *weight inheritance* are applied [8]. All the weights that can be reused in the offspring are copied from the parent(s) to the offspring. If needed, superfluous weights are cut off or missing weights are added and randomly initialized. Before each training phase is executed,  $D_{train}$  is randomly shuffled [4].

### 3.5 Fitness Function

The fitness function is based on the CNN accuracy ( $a$  is the number correctly classified inputs divided by all inputs from the test set  $D_{test}$ ) and the CNN relative size ( $s_{rel}$  is the number of parameters divided by the number of parameters of the best CNN of the initial population):

$$f = \begin{cases} a * (k * \frac{1}{\log(s_{rel}+1)} + 1) & \text{if } a \geq a_{min} \\ 0 & \text{otherwise,} \end{cases} \quad (1)$$

where  $k$  is a coefficient reflecting the impact of CNN size on the final fitness score and  $a_{min}$  is the minimal acceptable accuracy. It is important to introduce  $a_{min}$  as less complex CNNs providing unacceptable (low) classification accuracy would dominate the entire population.

### 3.6 Data Type and CNN Size Optimization

Almost all major CNN design frameworks operate over (32 bit) FP numbers and their computation is optimized for arithmetic FP operations and accelerated using GPUs. In order to enable FX operations (in particular, FX multiplications conducted during the inference phase in convolutional and fully connected layers), we modified relevant parts of TinyDNN source code. When a multiplication has to be executed in these layers, the FP operands are converted to a given FX number format, the multiplication is performed in FX and the product is converted back to FP. While this process emulates the error introduced by FX representation in a low cost hardware, all the remaining CNN steps can be implemented with the (highly optimized) FP operations. Unfortunately, this implementation slows down the CNN simulations approx. 8 times in our case.

When a CNN which should (partly) operate in the FX representation is evolved, we apply the aforementioned procedure in the fitness function (line 13 in Algorithm 1); however, the training is completely conducted in FP.

In our study, a (signed) FX number is implemented using 16 bits, in which 8 bits are fractional. If a 32 bit FP multiplication is replaced with a 16 bit FX multiplication, energy consumption of this operation is reduced approx.  $c_1 = 2.4$  times (for 65 nm technology [2]). Let  $E_{mult}$  denote the energy consumed by all multiplications performed during one inference phase carried out in a CNN embedded accelerator ( $E_{mult}$  is approx. 20% – 40% of the total energy required by the accelerator [11]). If the number of parameters of CNN is reduced from  $par_{orig}$  to  $par_{red}$  by EA4CNN and 16 bit FX instead of 32 bit FP multipliers are employed,  $E_{mult}$  is reduced approx.  $c_1 \times par_{orig}/par_{red}$  times because each parameter is associated with at least one multiplication in CNN.

## 4 Experimental Setup

EA4CNN is implemented in C++. We utilized the parallel training of CNNs supported in TinyDNN (by means of OpenMP and SSE instructions). Experiments were executed on a computer node containing two Intel Xeon E5-2680v3 processors @ 2.5 GHz, 128 GB RAM and 24 threads. As the entire neuroevolution process is very time consuming (an average run in which 750 candidate CNNs are evaluated takes almost 72 h for CIFAR-10), we typically generated only 50 populations of 15 individuals and performed only five independent runs for a particular setup. Hence, most EA parameters and CNN (hyper)parameters were set up on the basis of preliminary results from several test runs.

EA4CNN was evaluated using MNIST (10 digit classes) and CIFAR-10 (10 image classes) classification problems. MNIST consists of  $28 \times 28$  pixel grayscale images of handwritten digits and includes 60 000 training images and 10 000 test images. In CIFAR-10, the numbers of training and test images are 50 000 and 10 000, respectively, and the size of images is  $32 \times 32$  pixels. For our purposes, these data sets were divided into three parts in such a way that there are 75% vectors in  $D_{train}$ , 10% vectors in  $D_{test}$  and 15% in  $D_{val}$ .

The basic setup of EA parameters is as follows:  $G_{max} = 20 - 50$ ,  $|P| = 8 - 15$ ,  $p_{cross} = 0.35$ ,  $p_{mut} = 0.7$ ,  $age_{max} = |P|/2$ ,  $k = 0.5$ ,  $a_{min} = 0.80$  for MNIST and 0.60 for CIFAR-10. Mutation operators MO1 – MO6 are used with the probabilities 0.41, 0.07, 0.03, 0.29, 0.10, and 0.10, respectively.

Table 1 summarizes the initial CNN hyperparameters for both data sets. Randomly generated networks of the initial populations contain from 1 to 8 layers in which all weights are randomly initialized to the close to zero values. TinyDNN utilizes the stochastic gradient descent learning method.

## 5 Results

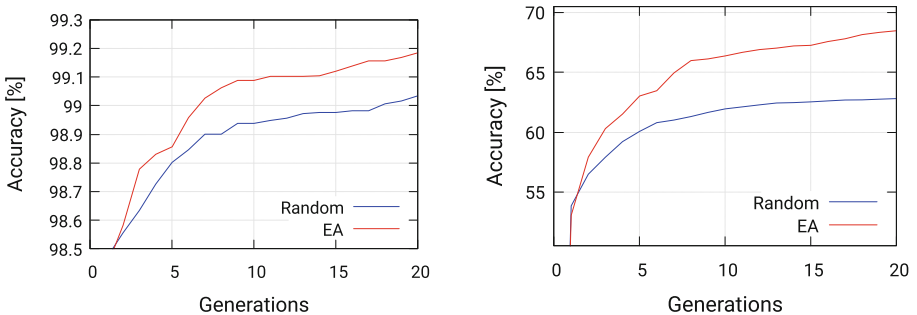
### 5.1 Basic Evaluation Of EA4CNN

In the first experiment, we compared the randomly-initialized EA with a random search of CNNs (RS-CNN). EA used the setup presented in Sect. 4, but

**Table 1.** The initial setting of CNN hyperparameters in EA4CNN. The hyperparameters given in the first part of the table can be modified during the evolution.

Parameter/Data set	MNIST	CIFAR-10
Learning rate	0.1	0.1
Initial number of neurons in FC layers	50	70
Max. filters in a newly added layer	12	20
Max. pooling layer size	4	4
Batch size	32	32
Epochs for training	1	1

$G_{max} = 20$ ,  $|P| = 8$  for MNIST and  $|P| = 12$  for CIFAR-10. RS-CNN starts with  $|P|$  randomly generated CNNs and performs their training for  $G_{max}$  epochs to ensure the same number of training exercises as in EA in which only one epoch of training is conducted for each candidate CNN in each generation. The average accuracy out of 5 independent runs of both algorithms is given in Fig. 2. Because MNIST classification is currently considered as a simple problem for NNs (the best reported accuracy is 99.79% [11]), even randomly generated CNN architectures provide (after their training) almost perfect classification accuracy. The average number of parameters of resulting CNNs is 200k for RS-CNN, but only 58k for EA which indicates that EA can optimize not only accuracy but also the CNN complexity (resulting CNNs have only 1–2 convolutional layers). While EA is only slightly better than RS-CNN for MNIST, the difference in the average accuracy on CIFAR-10 is relatively high (5.7% for 5 runs) which indicates that EA can also effectively increase the CNN size to improve the accuracy.

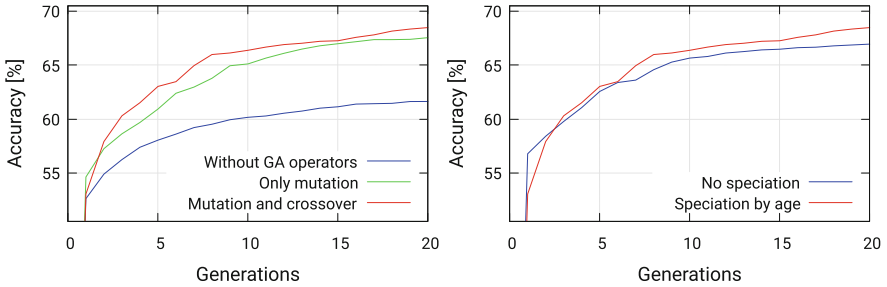
**Fig. 2.** The average accuracy obtained from five EA and five RS-CNN runs for MNIST (left) and CIFAR-10 (right) data sets.

In the second experiment, we investigated the impact of genetic operators on the progress of evolution of CNNs. Let EA1 denote Algorithm 1 in which neither crossover nor mutation are used. EA1, in fact, does not introduce any



new CNN structures, but optimizes how CNNs (randomly generated in the initial population) are selected for training by means of TinyDNN. Note that  $D_{train}$  is randomly shuffled before each training. Higher-scored CNNs can thus undergo more training exercises and improve their fitness score. Let EA2 and EA3 denote EA1 with mutation ( $p_{mut} = 0.80$ ; no crossover) and EA1 with mutation ( $p_{mut} = 0.50$ ) and crossover ( $p_{cross} = 0.35$ ). The other parameters remained as given in Sect. 4. The average classification accuracy out of 5 independent runs of EA1, EA2 and EA3 is given in Fig. 3 (left). Because of limited space only results on CIFAR-10 are reported. One can observe that performance of EA1 is roughly similar with RS-CNN. Incorporating the mutation operator (EA2) and crossover (EA3) leads to a higher classification accuracy of resulting CNNs.

Finally, Fig. 3 (right) illustrates the impact of employing the speciation mechanism on the accuracy during the CNN evolution. If the speciation “is not used” vs. “is used”, the classification accuracy of resulting five CNNs is between 59.93% – 72.96% vs. 62.02% – 73.05%; the average accuracy is 66.93% vs. 68.46%; the average depth of the network is 3.6 vs. 4.6 layers and the average number of parameters is 114k vs. 173k. We can conclude that the EA benefits from the proposed speciation mechanism.

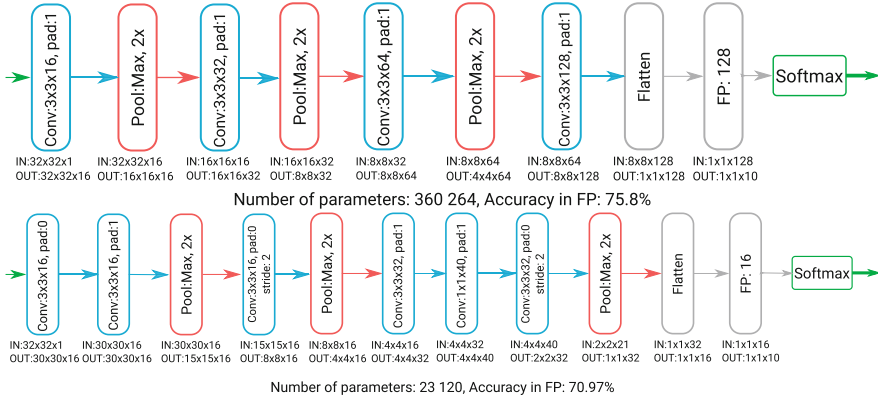


**Fig. 3.** Left: The average classification accuracy if EA uses selection only (EA1); selection and mutation (EA2); selection, mutation and crossover (EA3). Right: The average accuracy for EA3 with and without speciation (on CIFAR-10).

## 5.2 Evolution of Approximate CNNs

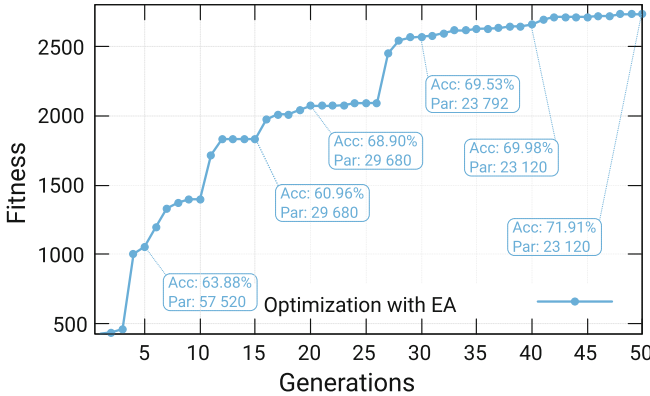
The experiments reported in this section have started with a *baseline CNN* shown in Fig. 4 (top) which contains 360,264 parameters, operates in FP and provides 75.8% accuracy on CIFAR-10 (trained with TinyDNN). We decided to approximate this middle-size CNN as less complex CNNs are our target and our computational resources are limited.

Figure 5 shows the fitness score, classification accuracy and complexity of CNNs obtained from a single run of the proposed EA which was seeded with the baseline CNN. The EA parameters are set according to Sect. 4, but  $k = 1$  to



**Fig. 4.** Hyperparameters and architecture of the baseline CNN (top) and one of the CNNs optimized with EA4CNN (bottom) for CIFAR-10 data set.

find good tradeoffs between the accuracy and the number of CNN parameters;  $G_{max} = 50$ , and  $|P| = 15$ . Note that the accuracy shown in the plot is the test accuracy on  $D_{test}$ . The resulting CNN is presented in Fig. 4 (bottom).



**Fig. 5.** An example run of the evolutionary CNN approximation process on CIFAR-10 data set.

Table 2 summarizes the best tradeoffs obtained from multiple EA runs. CNN\*-FP and CNN\*-FX denote CNNs performing the multiplication operations in FP and FX representation, respectively. Because of limited computing resources, we could only execute 20 generations to evolve CNNs utilizing the FX representation, which negatively influenced the quality of CNN\*-FX networks. For example, a similar classification accuracy ( $\sim 67.5\%$ ) was obtained by CNN2-FP and CNN1-FX, but CNN1-FX needs  $2.9\times$  more parameters and hence

it is less energy efficient despite the usage of FX multiplications. Reduction in the energy ( $E_{mult}$ ) needed for multiplication (calculated according to Sect. 3.6,  $c_1 = 2.4$ ) is clearly traded off for the loss in accuracy. EA4CNN allowed us to obtain this reduction by simplifying the CNN structure (fewer parameters) or/and employing FX operations. A more significant contribution of the FX representation is expected if EA4CNN could prolong the optimization and thus further reduce the number of parameters in CNN\*-FX networks.

Table 2 also presents some CNNs that are available in the NAS literature. CNNs achieving a 90% and higher classification accuracy on CIFAR-10 contain more than one million parameters [8] and their design takes days on a GPU, which is unreachable with our setup. The impact of employing the FX representation was reported for a human-created CNN (based on AlexNET [4]) which exhibits 81.22%, 79.77% and 77.99% accuracy in FP, 16 bit FX and 8 bit FX, respectively. The 16 bit and 8 bit FX implementations reduce the energy requirements approx. 2.5 and 6.8 times, respectively. Contrasted to our approach, these FX designs only implemented the original FP implementation with reduced precision in FX, i.e. without optimizing the CNN architecture.

**Table 2.** Examples of CNNs and their parameters obtained from the evolutionary approximation conducted with EA4CNN and from literature (for CIFAR-10).

CNN	Parameters	Accuracy	Layers	$E_{mult}$ reduction
Evolved with EA4CNN				
Baseline CNN (FP)	360,264	75.80 %	7	1.0
CNN1-FP	8 480	64.33 %	9	42.9×
CNN2-FP	12 784	67.50 %	7	28.1×
CNN3-FP	15 728	68.92 %	8	22.9×
CNN4-FP	23 120	70.97 %	9	15.6×
CNN5-FP	0.17 M	72.96 %	6	2.1×
CNN1-FX (16 bit)	36 720	67.66 %	11	23.6×
CNN2-FX (16 bit)	30 672	66.52 %	8	28.2×
CNN3-FX (16 bit)	19 632	65.63 %	7	44.0×
From literature				
[10] (FP) default scenario	1.68 M	94.02 %	–	–
[10] (FP) small data set (5k)	0.83 M	76.53 %	–	–
[8] (FP)	5.40 M	94.60 %	–	–
ALEX [2] (FP)	~10 M	81.22 %	–	1.0
ALEX [2] (FX, 16 bit)	~10 M	79.77 %	–	~2.5×
ALEX [2] (FX, 8 bit)	~10 M	77.99 %	–	~6.8×

## 6 Conclusions

The proposed EA4CNN platform can automatically evolve a CNN (with 29k parameters) showing almost the state-of-the-art accuracy (99.36%) for the MNIST task. Evolved CNNs for CIFAR-10 are far from the state-of-the-art, but it was expected because we used only the basic CNN techniques (no data augmentation, residual connections, dropout layers etc.) and very limited computing resources. However, we demonstrated that EA4CNN, if seeded with a trained CNN, can find interesting tradeoffs between the accuracy and implementation cost.

Our future work will focus on improving the search quality by incorporating advanced CNN techniques and employing more computing resources. We will also explore more options for optimizing the CNN cost in order to develop a fully automated holistic CNN approximation method.

**Acknowledgments.** This work was supported by the Ministry of Education, Youth and Sports, under the INTER-COST project LTC 18053, NPU II project IT4Innovations excellence in science LQ1602 and by Large Infrastructures for Research, Experimental Development and Innovations project “IT4Innovations National Supercomputing Center – LM2015070”.

## References

1. Dong, J.-D., Cheng, A.-C., Juan, D.-C., Wei, W., Sun, M.: DPP-Net: device-aware progressive search for pareto-optimal neural architectures. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11215, pp. 540–555. Springer, Cham (2018). [https://doi.org/10.1007/978-3-030-01252-6\\_32](https://doi.org/10.1007/978-3-030-01252-6_32)
2. Hashemi, S., Anthony, N., Tann, H., Bahar, R.I., Reda, S.: Understanding the impact of precision quantization on the accuracy and energy of neural networks. In: DATE, pp. 1478–1483. EDAA (2017)
3. Hsu, C., et al.: MONAS: multi-objective neural architecture search using reinforcement learning. CoRR abs/1806.10332 (2018). <http://arxiv.org/abs/1806.10332>
4. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 25, pp. 1097–1105. Curran Associates, Inc. (2012)
5. Miikkulainen, R., et al.: Evolving deep neural networks. CoRR abs/1703.00548 (2017). <http://arxiv.org/abs/1703.00548>
6. Nomi, T.: TinyDNN. <https://github.com/tiny-dnn/tiny-dnn> (2016)
7. Panda, P., et al.: Invited - cross-layer approximations for neuromorphic computing: from devices to circuits and systems. In: 53rd Design Automation Conference, pp. 1–6. IEEE (2016). <https://doi.org/10.1145/2897937.2905009>
8. Real, E., et al.: Large-scale evolution of image classifiers. arXiv e-prints [arXiv:1703.01041](https://arxiv.org/abs/1703.01041) (2017)
9. Stanley, K.O., Miikkulainen, R.: Evolving neural networks through augmenting topologies. *Evol. Comput.* **10**(2), 99–127 (2002)

10. Suganuma, M., Shirakawa, S., Nagao, T.: A genetic programming approach to designing convolutional neural network architectures. In: Proceedings of the Genetic and Evolutionary Computation Conference, GECCO 2017, pp. 497–504. ACM (2017)
11. Sze, V., Chen, Y., Yang, T., Emer, J.S.: Efficient processing of deep neural networks: a tutorial and survey. *Proc. IEEE* **105**(12), 2295–2329 (2017)