

Evolutionary NAS with Gene Expression Programming of Cellular Encoding

Cliford Broni-Bediako

Graduate School of Science and Engineering
Soka University, Hachioji, Tokyo, Japan
e18d5252@soka-u.jp

Luiz H. B. Mormille

Graduate School of Science and Engineering
Soka University, Hachioji, Tokyo, Japan
e18d5251@soka-u.jp

Yuki Murata

Graduate School of Science and Engineering
Soka University, Hachioji, Tokyo, Japan
e19d5202@soka-u.jp

Masayasu Atsumi

Graduate School of Science and Engineering
Soka University, Hachioji, Tokyo, Japan
matsumi@soka.ac.jp

Abstract—The renaissance of neural architecture search (NAS) has seen classical methods such as genetic algorithms (GA) and genetic programming (GP) being exploited for convolutional neural network (CNN) architectures. While recent work have achieved promising performance on visual perception tasks, the direct encoding scheme of both GA and GP has functional complexity deficiency and does not scale well on large architectures like CNN. To address this, we present a new generative encoding scheme—*symbolic linear generative encoding* (SLGE)—simple, yet a powerful scheme which embeds local graph transformations in chromosomes of linear fixed-length string to develop CNN architectures of variant shapes and sizes via an evolutionary process of gene expression programming. In experiments, the effectiveness of SLGE is shown in discovering architectures that improve the performance of the state-of-the-art handcrafted CNN architectures on CIFAR-10 and CIFAR-100 image classification tasks; and achieves a competitive classification error rate with the existing NAS methods using fewer GPU resources.

Index Terms—neural architecture search; convolutional neural network; gene expression programming; cellular encoding

I. INTRODUCTION

Historically, evolutionary neural architecture search (NAS) research has been of interest in the AI community for three decades [1]. Recent growing interest in automation of deep learning has seen a phenomenal development of new NAS methods based on reinforcement learning (RL) and others in the last three years [2]. However, classical NAS methods such as Genetic Algorithms (GA) [3], [4] and Genetic Programming (GP) [5] has also been exploited to develop convolutional neural networks (CNN) for visual perception tasks. Besides achieving promising results, classical NAS methods also consume fewer computational resources [4] than their RL-based competitors [6]. But, the direct encoding scheme of both GA and GP has two inherent limitations: (1) chromosomes of

fixed-length and (2) genotype and phenotype spaces not distinctive separated, which limit their functional complexity [7]. For example, chromosomes of fixed-length scheme employed in Genetic CNN [3] does not scale well on large architectures.

To address these issues, EvoCNN [8] introduced variable-length scheme in GA, and was adopted in CA-CNN [4]. The variable-length scheme is used to encode CNN architectures of variant shapes and sizes similar to nonlinear structures in GP. Though, the variable-length scheme exhibits some certain amount of functional complexity, it is not without difficulty to reproduce with crossover operations. The reason for the crossover limitation in both EvoCNN and CA-CNN, and their alike CGP-CNN [5] is because their genotype and phenotype spaces are not explicitly separated, thus, genetic modification is directly subject to phenotype structural constraint. DENSER [9] combines GA with Grammatical Evolution (GE) to separate genotype and phenotype spaces in analogous to nature. However, GE lacks modularity and it is not flexible to modify the grammar with modules [10]. To this end, we argue that the development of evolutionary NAS methods that separate genotype and phenotype spaces to generate CNN architectures for visual perception tasks is still in its infancy.

In this work, we introduce a new generative encoding scheme, *symbolic linear generative encoding*, which embeds local graph transformations of Cellular Encoding [11] in simple linear fixed-length chromosomes of Gene Expression Programming [7] to develop CNN architectures of variant shapes and sizes. Moreover, to enable the evolutionary process to discover new motifs for the architectures, *regular convolution*¹ operations are employed as the basic search units as opposed to sophisticated building-blocks such as ResNet and DenseNet blocks in CA-CNN [4].

In experiments, the preliminary results show the effectiveness of the proposed method by discovering CNN architecture that obtains 3.74% error on CIFAR-10 image dataset bench-

This work was supported in part by the Soka Makiguchi Foundation for Education, Japan.

¹We define a regular convolution operation as a standard convolution operation with batch normalization and ReLU.

mark and 22.95% error when transfer to CIFAR-100. The results is competitive with the current auto-generated CNN architectures and an improvement to the performance of the state-of-the-art handcrafted ones. The remainder of the paper is organised as follows. Section 2 discusses the related work and Section 3 presents the details of the proposed method. The experimental results are reported in Section 4 and Section 5 concludes this study with future direction.

II. RELATED WORK

A. Evolutionary NAS

Most earlier work in evolutionary NAS evolve both the network architecture and its connection weights at a small scale [1]. Recent neural networks such as CNN have been scaled up with millions of connection weights to improve performance on a given task. And learning the connection weights of these large-scale networks via back-propagation method outperform evolutionary approach. Thus, recent evolutionary NAS work [2] have focused on evolving only the network architectures and using back-propagation method to optimize the connection weights.

Typically, chromosomes of fixed-length is used to represent the network architectures [3], [12], [13]. But, since the best architecture shape and size for a given data is not known, chromosomes of variable-length using direct encoding scheme have been employed for the architectures to adapt their shapes and sizes on a given task [4], [14], [15]. CGP-CNN [5] used Cartesian GP to represent CNN architectures of variable-length structures. To depart from the functional complexity deficiency of direct encoding, DENSER [9] combines GA with GE to adopt a genotype and phenotype spaces distinction and explicitly use grammars to generate phenotypes of CNN architectures.

Generally, the architecture search space is classified into two categories: the global search space which defines the entire architecture (macroarchitecture) [4], [14], and cell-based search space for discovering a microarchitecture (cell) which can be stacked repeatedly to build the entire architecture [13], [15]. The cell is a directed acyclic graphs (DAGs) which is used as building blocks to form the architecture. CNN architectures discovered by cell-based approach are flexible and transferable to other tasks [16], and they perform better than the global ones [17]. The popular cell-based approach is NASNet [16] which involves two types of cells: the normal cell and reduction cell (used to reduce feature resolutions). Zhong *et al.* [18] and Liu *et al.* [15] proposed similar cell-based approach but used max-pooling and separable convolution layers respectively to reduce the feature resolutions. The basic search units in both search spaces are mostly sophisticated convolutions such as depthwise separable and asymmetric convolutions in AmoebaNet [13] or sophisticated blocks such as ResNet and DenseNet in CA-CNN [4]. These search units reduce the search space complexity, however, they may detriment the flexibility and restrict the discovering of new architecture building blocks that can improve the

current handcrafted ones. Thus, in this work, we use regular convolutions (Section III-A).

B. Gene Expression Programming

Gene Expression Programming (GEP) is a full-fledged genotype-phenotype evolutionary method based on both GA and GP. Chromosomes consist of linear fixed-length genes similar to the ones used in GA, and are developed in phenotype space as expression-trees of different shapes and sizes similar to the parse-trees in GP. The genes are structurally organized in a head and a tail format called Karva notation [7]. The separation of genotype and phenotype spaces with distinct functions allows GEP to perform with high effectiveness and efficiency that surpasses GA and GP [19].

Ferreira [7] has proposed that the chromosomes in GEP can completely encode ANN to discover an architecture via evolutionary process. Thus far, GEP is yet to be adopted for complex architectures like CNN. The Achilles' heel of GEP is its Karva notation which does not allow hierarchical composition of candidate solutions, which means an evolved good motifs are mostly destroyed by genetic modifications in subsequent generations [20]. Thus, to adopt GEP for CNN architecture search, we propose a new generative scheme that naturally embeds motifs in Karva expression of GEP as individual chromosomes and consequently a new genotype-phenotype mapping is encapsulated in GEP convention of evolutionary process.

C. Cellular Encoding

Cellular Encoding (CE) is a generative encoding based on simple local graph transformations that control the division of nodes which evolve into ANN of different shapes and sizes. The graph transformations are represented by program symbols with unique names; and these symbols depict a grammar-tree (program) which encapsulates the developing procedure of an ANN via evolutionary process from a single initial unit² that has both input and output nodes. CE has shown its efficiency on a wide range of problems such as evolving ANN for controlling two poles on a cart and locomotion of a 6-legged robot, a review can be found in Gruau [21].

In this work, we adopt four transformation functions of CE and embed them in the fixed-length individual chromosomes of GEP, since their transformations—particularly *CPI* and *CPO* [22]—generate motifs similar to ResNet block which is prominent in deep neural networks. We briefly explain the four functions we adopt and refer Gruau [11] and Gruau and Quatramaran [22] for in-depth details. Fig. 1 depicts the graphical representation of the SEQ, CPO and CPI transformation functions.

- SEQuential division (*SEQ*): it splits current node into two and connects them in serial; the child node inherits the outputs of the parent node.
- CoPy Input division (*CPI*): it performs *SEQ*, then shares the same inputs with parent and child nodes.

²A unit is a single neuron in the original idea of CE, but in this work, it represents a convolutional layer of neurons

- CoPy Output division (*CPO*): it performs *SEQ*, then shares the same outputs with parent and child nodes.
- END program (*END*): it stops the developing process.

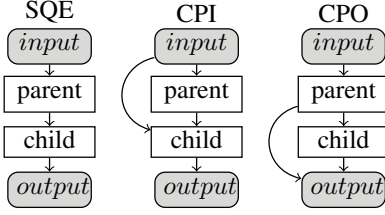


Fig. 1: Illustration of SEQ, CPI and CPO transformation functions.

III. PROPOSED METHOD

We formulate the CNN architecture search problem as follows. Given the problem space $\Psi = \{\mathcal{A}, \mathcal{S}, \mathcal{P}, t\mathcal{D}, v\mathcal{D}\}$, where \mathcal{A} is the architecture search space, \mathcal{S} represents the search strategy, \mathcal{P} denotes the performance measure, and $t\mathcal{D}$ and $v\mathcal{D}$ are the training and validation datasets respectively, the objective is to find a smaller CNN architecture $a^* \in \mathcal{A}$ via the search strategy \mathcal{S} , then after training it on the dataset $t\mathcal{D}$, it maximizes some performance \mathcal{P} (in this case classification accuracy \mathcal{P}_{acc}) on the validation dataset $v\mathcal{D}$. The smaller architecture here means a model with fewer number of parameters θ . Mathematically, the objective function \mathcal{F} can be formulated as:

$$\begin{aligned} \mathcal{F}(\Psi) = \max_{\theta, a} \mathcal{P}_{acc}(\mathcal{L}(a(\theta), t\mathcal{D} \mid \mathcal{S}, a \in \mathcal{A}), v\mathcal{D}) \\ \text{s.t. } \text{ModelSize}(a(\theta)) \leq \mathcal{T}_{params} \end{aligned} \quad (1)$$

where \mathcal{L} represents the training of the model parameters θ with the loss function and \mathcal{T}_{params} denotes the target number of parameters. In this section, we describe the architecture search space and search strategy that we proposed.

A. Search Space

The basic search units in the search space consists of regular convolutions with batch-norm and ReLU and CE program symbols. The regular convolutions units may enable the evolutionary process to find new motifs to form CNN architectures rather than the predefined ones (depthwise separable, asymmetric convolution, ResNet and DenseNet). Since predefined units reduce the complexity of the architecture search space, they may detriment the flexibility of discovering new motifs that can improve the current handcrafted ones. Thus, we adopt cell-based search approach [16] in which each node in the discovered cells is associated with a regular convolution in the search space whereas edges are representation of latent information flow direction. Following Zhong *et al.* [18], we use max-pooling to reduce feature maps resolution and 1×1 convolution is applied when necessary to downsample the input depth.

The regular convolutions include in the search space are: 1×1 , 1×3 , 3×1 , and 3×3 . Each operation has a stride of one and appropriate pad is applied to preserve the spatial

resolution of the feature maps. And the CE program symbols are: *SEQ*, *CPI*, *CPO* and *END*. The complexity of the search space can be expressed as: $(\#program_symbols)^h \times (\#convolution_operations)^{h+1} \times n$ possible architectures, where h is the number of CE program symbols that forms the head of a gene in a chromosome and n is the number of genes in a chromosome. For example, a chromosomes with $h=2$ and $n=3$, the search space contains 3072 possible architectures.

B. Symbolic Linear Generative Encoding

Encoding network architectures into genotypes using a particular encoding scheme is the first stage of evolutionary NAS task. We proposed a new generative encoding scheme, *symbolic linear generative encoding* (SLGE), which embeds local graph transformations of CE in simple linear fixed-length chromosomes of GEP to develop CNN architectures of different shapes and sizes. Chromosomes in SLGE can evolve different motifs via the evolutionary process of GEP to build CNN architectures. SLGE explicitly separates genotype and phenotype spaces in analogue to nature, to benefit from all the advantages of evolutionary process without functional complexity deficiency. It is worth emphasizing how the implementation of SLGE is simple, because of its linear fixed-length structure. We implement SLGE on the top of *geppy*³, a GEP library based on *DEAP*⁴ framework.

1) *Representation*: Genetic representation defines the encoding of a phenotype into a genotype. A simple, but effective and efficient representation significantly affects the overall performance of evolutionary process. In SLGE, the chromosomes are structured similar to the ones in GEP. Chromosomes are made up of genes of equal fixed-length string. Each gene composes of a head which consists of CE program symbols, and a tail of regular convolutions. Given the length of a gene head h , the length of the gene tail t is a function of h expressed as: $t = h + 1$, thus, the length of a gene is $2h + 1$. Fig. 2 is an example of a typical SLGE chromosome of two genes and its phenotype (cell) is presented in Fig. 3.

2) *Mapping and Fitness Function*: The mapping algorithm translates genotypes (chromosomes) into phenotypes (cells)—each one is stacked repeatedly to build candidate CNN architecture—then the fitness function is used to train each architecture for a few epochs and evaluates to determine its fitness quality in the phenotype space, which is then maps back into genotype space where genetic variations occur to produce offspring for the next generation. The representation can be mathematically expressed as follows:

$$\mathcal{F}_g(\varphi_g) : \Phi_g \rightarrow \Phi_p \quad (2)$$

$$\mathcal{F}_p(\varphi_p) : \Phi_p \rightarrow \mathbb{R} \quad (3)$$

where \mathcal{F}_g is the mapping algorithm which maps a set of genotypes $\varphi_g \subset \Phi_g$ into phenotype space Φ_p to form individual architectures and \mathcal{F}_p represents the fitness function that trains each individual architecture (phenotype) $\varphi_p \in \Phi_p$

³<https://geppy.readthedocs.io/en/latest/index.html>

⁴<https://deap.readthedocs.io/en/master/>

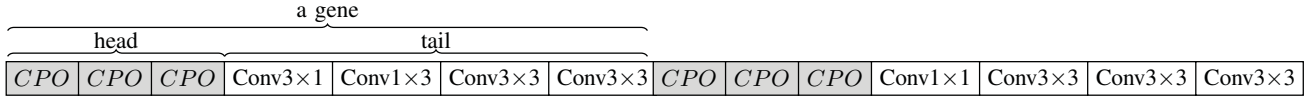


Fig. 2: The structural representation of a typical SLGE chromosome. The chromosome has two genes of equal fixed-length, and each gene has a head of three CE program symbols and a tail of four regular convolution operations. This chromosome is the genotype of the cell (Fig. 3) used to build the best evolutionary-discovered network in Table II.

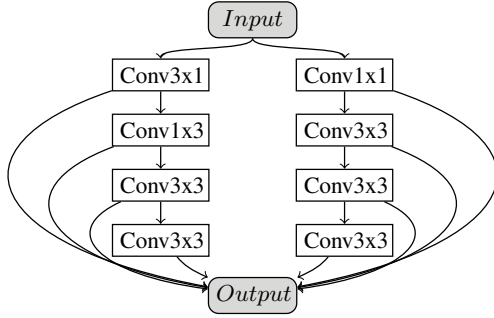


Fig. 3: The representative cell of the chromosome in Fig. 2, and it is the cell used to build the top network in Table II. (In general, the convolution nodes without successor in the cell are depthwise concatenated to provide an output, and if a node has more than one predecessor, the predecessors parameters are add together. This is a common approach in cell-based search [16].)

for a few epochs to determine its fitness value in the fitness space \mathbb{R} . Algorithm 1 and 2 are the mapping algorithm \mathcal{F}_g that is used to translate chromosomes into cells in SLGE. The algorithm takes a chromosome of linear fixed-length string as an input and produces a DAG which represents a candidate cell. For example, given the chromosome in Fig. 2 as an input, Algorithm 1 and 2 will produce the cell in Fig. 3 as an output.

The fitness function \mathcal{F}_p is a surrogate to the objective function \mathcal{F} in equation (1) which aims to find an individual architecture that maximize classification accuracy on validation dataset, subject to model size constraint. However, since we search for cells in the architecture search space during the evolutionary process, the model size constraint of individuals is not considered in their fitness evaluation. Rather, we only apply the model size constraint for the entire architecture. Thus, the function is simply the objective function \mathcal{F} in equation (1) without the size constraint. The loss function is used to train each individual via back-propagation on the training and validation sets, and the fitness function determines its fitness value which represents the probability of the individual survival. The higher the fitness value, the more likely the individual will have progeny and survive into the next generation.

3) *Evolutionary Process*: We adopt the evolutionary process in GEP and refer Ferreira [7] for details of the steps presented here.

Step 1: *Initialization*—randomly generate a population of SLGE chromosomes in a uniform distribution manner.

Step 2: *Mapping*—apply the mapping function \mathcal{F}_g to translate individual chromosomes into cells.

Algorithm 1 The genotype-phenotype mapping function \mathcal{F}_g

Input: Chromosome of linear fixed-length string $\varphi_g \in \Phi_g$

Output: A directed acyclic graph DAG (cell)

```

1:  $n \leftarrow \text{len}(\varphi_g)$  //Number of genes in chromosome  $\varphi_g$ 
2:  $DAG.\text{init}(\text{null})$  //Initialize cell with null node which
   //has input and output nodes
3: for  $i \leftarrow 1$  to  $n$  do
4:    $\varphi_g^{(i)} \leftarrow \varphi_g[i]$  //Get gene  $i$  in chromosome  $\varphi_g$ 
5:   Create queue  $Q$  of all the convolutions in gene  $\varphi_g^{(i)}$ 
6:    $pnode \leftarrow Q.\text{dequeue}(0)$  //Parent node
7:    $cnode \leftarrow Q.\text{next}()$  //Child node
8:    $G \leftarrow \text{subgraph}(pnode)$  //Initialize subgraph of  $DAG$ 
   //for  $\varphi_g^{(i)}$  with  $pnode$  which
   //has input and output nodes
9:    $pos \leftarrow 0$ 
10:  while  $|Q| > 0$  do
11:     $ps \leftarrow \varphi_g^{(i)}[pos]$  //Get CE program symbol
12:    if  $ps = \text{"END"}$  then
13:       $DAG.\text{merge}(G)$  //Merge subgraph  $G$  to  $DAG$ 
      //at input and output nodes
14:    return
15:  end if
16:   $\text{TRANSFORM}(G, ps, pnode, cnode)$  //Algorithm 2
17:   $pnode \leftarrow Q.\text{dequeue}(0)$ 
18:   $cnode \leftarrow Q.\text{next}()$ 
19:   $pos \leftarrow pos + 1$ 
20: end while
21:   $DAG.\text{merge}(G)$ 
22: end for
23: return  $DAG$ 

```

Step 3: *Fitness*—build candidate CNN architectures with each cell, and train each via back-propagation and evaluate its fitness using the fitness function \mathcal{F}_p .

Step 4: *Selection*—select individuals to form next generation population via roulette-wheel strategy with elitism.

Step 5: *Mutation*—randomly mutate all elements in a chromosome. The structural rule must be preserved (e.g. convolution element cannot be assigned to a gene head).

Step 6: *Inversion*—randomly invert some sequence of elements in a gene head of individual chromosomes.

Step 7: *Transposition*—randomly replace some sequence of elements with consecutive elements in the same chromosome. The structural rule must be preserved.

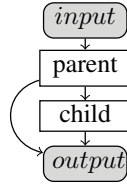
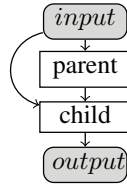
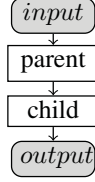
Step 8: *Recombination*—crossover gene elements of two chromosomes using two-point. Since SLGE chromosomes are

Algorithm 2 The CE local graph transformation procedure. It transforms gene subgraph G with parent node $pnode$ to generate child node $cnode$ via CE program ps .

```

1: procedure TRANSFORM( $G, ps, pnode, cnode$ )
2:   if  $ps = \text{"SEQ"}$  then
3:      $succ \leftarrow list(G.successors(pnode))$ 
4:     for  $node$  in  $succ$  do
5:        $G.addEdge(cnode, node)$ 
6:        $G.removeEdge(pnode, node)$ 
7:     end for
8:      $G.addEdge(pnode, cnode)$ 
9:   end if
10:  if  $ps = \text{"CPI"}$  then
11:     $pred = list(G.predecessors(pnode))$ 
12:     $succ = list(G.successors(pnode))$ 
13:    for  $node$  in  $pred$  do
14:       $G.addEdge(node, cnode)$ 
15:    end for
16:    for  $node$  in  $succ$  do
17:       $G.addEdge(cnode, node)$ 
18:       $G.removeEdge(pnode, node)$ 
19:    end for
20:     $G.addEdge(pnode, cnode)$ 
21:  end if
22:  if  $ps = \text{"CPO"}$  then
23:     $succ = list(G.successors(pnode))$ 
24:    for  $node$  in  $succ$  do
25:       $G.addEdge(cnode, node)$ 
26:    end for
27:     $G.addEdge(pnode, cnode)$ 
28:  end if
29: end procedure

```



fixed-length strings similar to the ones in GEP, performing two-point crossover always yields a valid chromosome.

Step 9: Go to Step 2 if max generation not met; else return individual with the highest fitness as best discovered cell.

IV. EXPERIMENTS

The preliminary experiments aimed at verifying the effectiveness of the proposed method to discover cells for CNN architectures that perform well on image classification tasks. We ran eight experiments, two each of four different configurations of chromosome (Table II) and a random search as baseline. The search was conducted on CIFAR-10 [23] dataset, and the best discovered architecture was transferred to CIFAR-100 [23]. The experiments were performed on one 11GB GPU GeForce GTX 1080 Ti machine for 20 days. We implemented all the architectures with PyTorch⁵ and trained using fastai⁶ library. The codes are available at https://github.com/cliffbb/geppy_nn.

⁵<https://pytorch.org/>

⁶<https://docs.fast.ai/>

Parameter	Value
Number of elites	1
Mutation rate	0.044
Inversion and Transposition rate	0.1
Inversion and Transposition elements length	2
Two-point/Gene recombination rate	0.6/0.1

TABLE I: Evolutionary parameters for the search process.

Chromosome	Network	Error	Params
Genes=2, Head=2	$B=[3, 3, 2]$	3.93	3.2M
	$B=[3, 3, 2]$	3.84	2.7M
Genes=2, Head=3	$B=[3, 3, 1]$	3.74	2.8M
	$B=[2, 3, 1]$	3.85	2.7M
Genes=3, Head=2	$B=[3, 3, 1]$	4.24	3.5M
	$B=[3, 3, 1]$	4.03	3.4M
Genes=3, Head=3	$B=[3, 3, 2]$	4.19	3.4M
	$B=[3, 3, 2]$	5.07	3.0M

TABLE II: Experimental results on four different configurations of chromosome. We ran evolutionary search on each configuration twice, after search, trained from scratch with $C=40$ and report classification error rate (%) on CIFAR-10. The boldface is the best result.

A. Evolutionary Settings

We used a small population of 20 individuals since GEP is capable of solving relatively complex problems with small population [24], and generation of 20 to reduce the computation cost of the search process. Other evolutionary parameters settings used for each run are summarized in Table I, and these are defaults settings in GEP.

B. Training Details

Each network begins with 3×3 convolution stem with output channel size C , followed by three blocks denoted as $B=[b_1, b_2, b_3]$ —each block i consist of repeated cell of b_i times with max-pooling layer inserted between them to downsample—then a classifier. The max-pooling reduces the feature maps of each block by a half. The channels are doubled after each max-pooling operation.

During search, the CIFAR-10 50k training set was split into 40k training and 10k validation subsets with normalization. We used relatively small network with $C=16$ and $B=[1, 1, 1]$. Each candidate network was trained on the training subset and evaluated on the validation subset to determine its fitness value using 1-cycle policy in Smith [32] with Adam optimizer. The learning rate was set to go from 0.004 to 0.1 linearly while the momentum goes from 0.95 to 0.85 linearly in phase one, then in phase two, the learning rates follows cosine annealing from 0.1 to 0, whereas the momentum goes from 0.85 to 0.95 with the same annealing. The weight decay was set to 0.0004, batch size to 128 and training epochs to 25.

After the search, we set $C=40$ and each evolutionary-discovered cell was repeatedly stacked to build large CNN architectures subject to model size constraint $\mathcal{T}_{params} \leq 3.5M$. We trained each CNN architecture from scratch on the training and validation subsets for 800 epochs, and reported the classification error on the CIFAR-10 10k test dataset. The learning rate remained as during search for the first 350 epochs, and

Model	CIFAR-10	CIFAR-100	Params	GPU-days	Method	
ResNet-110 [25]	6.61	—	1.7M	—	Handcrafted	
ResNet-1001 (pre-activation) [26]	4.62	22.71	10.2M	—		
FractalNet [27]	5.22	23.30	38.6M	—		
DenseNet ($k=12$) [28]	4.10	20.20	7.0M	—		
DenseNet-BC ($k=40$) [28]	3.46	17.18	25.6M	—		
MetaQNN (top model) [29]	6.92	27.14	11.2M	100	Reinforcement	
NASv2 [6]	6.01	—	2.5M	22,400		
NASv3 [6]	4.47	—	7.1M	22,400		
Block-QNN-S (more filters) [18]	3.54	18.06	39.8M	96		
NASNet-A (6 @ 768) [16]	3.41	—	3.3M	2,000		
ENAS + cutout [17]	2.89	—	4.6M	0.5		
DARTS (first order) + cutout [30]	3.00 ± 0.14	—	3.3M	1.5	Gradient-Based	
DARTS (second order) + cutout [30]	2.76 ± 0.09	—	3.3M	4		
P-DARTS + cutout [31]	2.50	—	3.4M	0.3		
P-DARTS + cutout [31]	—	17.20	3.4M	0.3		
Genetic CNN [3]	7.10	29.05	—	17	Evolutionary	
Large-scale Evolution [14]	5.40	—	5.4M	2,750		
Large-scale Evolution [14]	—	23.0	40.4M	2,750		
Hierarchical Evolution [15]	3.63	—	—	300		
CA-CNN [4]	4.30	—	2.0M	27		
CA-CNN [4]	—	20.85	5.4M	36		
AmoebaNet-A [13]	3.34	—	3.2M	3,150		
SLGE Networks (our approach)	$C=40, B=[3, 3, 1]$	3.74	22.95	2.8M	20	Evolutionary
	$C=40, B=[3, 3, 2]$	3.84	24.47	2.7M	20	Evolutionary
	$C=40, B=[4, 4, 3]$	4.47	—	3.2M	4	Random search

TABLE III: Comparisons between the proposed method and other architecture search methods in terms of classification error (%), number of parameters and computing days of GPU ($\#$ of employed GPUs \times $\#$ of days the algorithms was run) based on CIFAR-10 and CIFAR-100 datasets.

was reset to go from 0.00012 to 0.003 in phase one and 0.003 to 0 in phase two of 1-cycle policy. We augmented the training subset as in He *et al.* [25], and the other hyperparameters remained the same as during search.

C. Search on CIFAR-10

To investigate the effectiveness of SLGE, we conducted evolutionary search on four different configurations of chromosome—2 genes with head of 2 and 3, and 3 genes with head of 2 and 3. Each was run twice on CIFAR-10 dataset. Chromosome of 2 genes with head of 3 discovered the best architecture which obtains 3.74% classification error (Table II). The result is compared with other search methods in Table III.

a) Compared with handcrafted networks: SLGE networks improve the performance of most popular handcrafted ones. SLGE achieves classification error of 0.8% and 1.4% lower than ResNet-1001 and FractalNet respectively with few parameters. However, the DenseNet-BC ($k=40$) performs better than SLGE networks using more parameters.

b) Compared with reinforcement NAS networks: It shows that SLGE achieves the error rate of approximately 3.2% and 2.3% improvement to MetaQNN and NASv2 respectively; and competes with Block-QNN-S (used more parameters) and NASNet-A (consumed more GPU resources). With cutout augmentation and 1.8M parameters more, ENAS achieved 0.9% error rate lower than SLGE using 0.5 GPU days.

c) Compared with gradient-based NAS networks: Using cutout augmentation method, DARTS and P-DARTS used less GPU days to achieve classification error rate of approximately

0.9% and 1.2% lower than SLGE respectively. However, the SLGE networks have fewer parameters than DARTS and P-DARTS.

d) Compared with evolutionary NAS networks: SLGE performs significantly well than Genetic CNN and Large-scale Evolution, but falls slightly behind the state-of-the-art Hierarchical Evolution and AmoebaNet-A networks. SLGE consumes 0.06 and 0.006 GPU computing days of that consumed by Hierarchical Evolution and AmoebaNet-A respectively.

D. Random Search

To evaluate the effectiveness of the SLGE representation scheme, a simple random search was performed as the baseline. We randomly generated a population of ten individual networks from the top chromosome (Table II); trained all from scratch with no evolutionary modifications on CIFAR-10 using the same training settings with 500 epochs. The network with the lowest classification error is reported in Table III. We achieve the mean classification error of 5.85% with standard deviation 1.77% over the ten individual networks. This is a considerable competitive result which demonstrates the effectiveness of our proposed SLGE scheme with regular convolution search units.

E. Evaluation on CIFAR-100

The top architecture learned on CIFAR-10 task was experimented on CIFAR-100 to evaluate the transferability of the evolutionary-discovered cell. The CIFAR-100 has similar features as CIFAR-10, but with 100 classes which makes it rather challenging classification task. We simply transferred

the best architecture from CIFAR-10 (Table II) but trained from scratch with a classifier head of 100 classes, all training settings remained the same. And we achieve the classification error of 4.1% improvement to MetaQNN and 6.1% to Genetic CNN, and compete with other networks (Table III).

F. Discussion

We make a few observations from the preliminary experimental results. The networks developed with chromosomes of 2 genes performed better than the ones with 3 genes, something we least expected. And the evolutionary-discovered cell (Fig. 3) of the top network has topological structure similar to DenseNet with a well-formed asymmetric convolution. This underscore the effectiveness of the proposed method and further experiments will be conducted in future to ascertain its robustness. In general the preliminary results on CIFAR-10 and CIFAR-100 classification tasks are very promising, although no evolutionary parameter was tuned. Thus, the natural future direction is to tune the evolutionary parameters, and extend the task to the most challenging ImageNet classification task to verify the generality of the proposed method.

V. CONCLUSION

We have presented an effective evolutionary method which discovers high-performing cells as building blocks for CNN architectures based on a representation scheme which embeds elements of CE in chromosomes of GEP. We show that our SLGE representation scheme coupled with regular convolution units can achieve significant results even with a simple random search strategy. Our top network obtains a highly competitive performance with the state-of-the-art networks on both CIFAR-10 and CIFAR-100 datasets. In addition to the observations made in Section IV-F, we will continue improving the proposed method and extend it to other visual tasks such as semantic segmentation and object detection.

REFERENCES

- [1] X. Yao, "Evolving artificial neural networks," *Proceedings of the IEEE*, vol. 87, no. 9, pp. 1423–1447, 1999.
- [2] T. Elsken, J. H. Metzen, and F. Hutter, "Neural architecture search: A survey," *JMLR*, vol. 20, no. 55, pp. 1–21, 2019.
- [3] L. Xie and A. Yuille, "Genetic CNN," in *IEEE ICCV*. IEEE, 2017, pp. 1388–1397.
- [4] Y. Sun, B. Xue, M. Zhang, and G. G. Yen, "Completely automated CNN architecture design based on blocks," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2019.
- [5] M. Suganuma, S. Shirakawa, and T. Nagao, "A genetic programming approach to designing convolutional neural network architectures," in *Proceedings of the Twenty-Seventh IJCAI*, 2018, pp. 5369–5373.
- [6] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," in *Proceedings of the 5th International Conference on Learning Representations*, 2017.
- [7] C. Ferreira, *Gene Expression Programming: Mathematical Modeling by an Artificial Intelligence*. Springer, 2006.
- [8] J. M. Swafford, M. O'Neill, M. Nicolau, and A. Brabazon, "Exploring grammatical modification with modules in grammatical evolution," in *Genetic Programming*, 2011, vol. 6621, pp. 310–321.
- [9] Y. Sun, B. Xue, M. Zhang, and G. G. Yen, "Evolving deep convolutional neural networks for image classification," *IEEE Transactions on Evolutionary Computation*, vol. 24, no. 2, pp. 394–407, 2020.
- [10] F. Assunção, N. Lourenço, P. Machado, and B. Ribeiro, "DENSER: deep evolutionary network structured representation," *Genetic Prog. and Evolvable Machines*, vol. 20, no. 1, pp. 5–35, 2019.
- [11] F. Gruau, "Neural network synthesis using cellular encoding and the genetic algorithm." Thesis Doctoral, L'universite Claude Bernard-lyon I, 1994.
- [12] T. Desell, "Large scale evolution of convolutional neural networks using volunteer computing," in *Proc. of the GECCO*, 2017, pp. 127–128.
- [13] E. Real, A. Aggarwal, Y. Huang, and Q. V. Le, "Regularized evolution for image classifier architecture search," in *Proc. of AAAI*, 2019.
- [14] E. Real, S. Moore, A. Selle, S. Saxena, Y. L. Suematsu, J. Tan, Q. V. Le, and A. Kurakin, "Large-scale evolution of image classifiers," in *Proceedings of the 34th JMLR*, vol. 70, 2017, pp. 2902–2911.
- [15] H. Liu, K. Simonyan, O. Vinyals, C. Fernando, and K. Kavukcuoglu, "Hierarchical representations for efficient architecture search," in *Proceedings of Sixth ICLR*, 2018.
- [16] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proceedings of the IEEE Conference on CVPR*, 2018, pp. 8697–8710.
- [17] H. Pham, M. Guan, B. Zoph, Q. Le, and J. Dean, "Efficient neural architecture search via parameters sharing," in *Proceedings of the 35th ICML*, vol. 80, 2018, p. 4095–4104.
- [18] Z. Zhong, J. Yan, W. Wu, J. Shao, and C.-L. Liu, "Practical block-wise neural network architecture generation," in *Proceedings of the IEEE Conference on CVPR*, 2018, pp. 2423–2432.
- [19] J. Zhong, L. Feng, and Y.-S. Ong, "Gene expression programming: A survey," *IEEE Comp Intelligence Magazine*, vol. 12, no. 3, pp. 54–72, 2017.
- [20] X. Li, C. Zhou, W. Xiao, and P. C. Nelson, "Prefix gene expression programming," in *Proceedings of the GECCO*, 2005, p. 25–31.
- [21] F. Gruau, "Artificial cellular development in optimization and compilation," in *Towards Evolvable Hardware*, 1996, vol. 1062, pp. 48–75.
- [22] F. Gruau and K. Quatramaran, "Cellular encoding for interactive evolutionary robotics," CWI, Amsterdam, Tech. Rep., 1996.
- [23] A. Krizhevsky, V. Nair, and G. Hinton, "Learning multiple layers of features from tiny images," 2009 (accessed February 3, 2020). [Online]. Available: <https://www.cs.toronto.edu/~kriz/cifar.html>
- [24] C. Ferreira, "Gene expression programming in problem solving," in *Soft Comp. and Industry: Recent Applications*, 2002, pp. 635–653.
- [25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conf. on CVPR*, 2016, pp. 770–778.
- [26] K. He, X. Zhang, S. Ren, and J. Sun, "Identity Mappings in Deep Residual Networks," in *Proceedings of the ECCV*, 2016, pp. 630–645.
- [27] G. Larsson, M. Maire, and G. Shakhnarovich, "Fractalnet: Ultra-deep neural networks without residuals," in *Proceedings of the ICLR*, 2017.
- [28] G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on CVPR*, 2017, pp. 2261–2269.
- [29] B. Baker, O. Gupta, N. Naik, and R. Raskar, "Designing neural network architectures using reinforcement learning," in *Proceedings of the 5th International Conference on Learning Representations*, 2017.
- [30] H. Liu, K. Simonyan, and Y. Yang, "DARTS: Differentiable architecture search," in *Proceedings of the International Conference on Learning Representations*, 2019.
- [31] X. Chen, L. Xie, J. Wu, and Q. Tian, "Progressive differentiable architecture search: Bridging the depth gap between search and evaluation," in *Proceedings of the International Conference on Computer Vision*, 2019, pp. 1294–1303.
- [32] L. N. Smith, "A disciplined approach to neural network hyperparameters: Part1—learning rate, batch size, momentum, and weight decay," 2018, arXiv:1803.09820v2 [cs.LG].