A Particle Swarm Optimization-based Flexible Convolutional Auto-Encoder for Image Classification

Yanan Sun, Member, IEEE, Bing Xue, Member, IEEE, Mengjie Zhang, Senior Member, IEEE, and Gary G. Yen, Fellow, IEEE

Abstract—Convolutional auto-encoders have shown their remarkable performance in stacking to deep convolutional neural networks for classifying image data during the past several years. However, they are unable to construct the state-of-the-art mainstream searching for the optimal architectures of the proposed approach on four extensively used image classification resource and without any manual intervention. We test the proposed approach on four extensively used image classification datasets. Experimental results show that our proposed approach in this paper significantly outperforms the peer competitors including the state-of-the-art algorithms.

Index Terms—Convolutional auto-encoder, particle swarm optimization, image classification, deep learning, neural networks.

I. INTRODUCTION

UTO-ENCODERS (AEs) [1]–[4] are building blocks of Stacked AE (SAE) [5], [6] that is one of the tri mainstream deep learning algorithms [7] (i.e., others are Deep Belief Networks (DBN) [8] and Convolutional Neural Networks (CNNs) [9], [10]). An AE is a three-layer neural network comprising one input layer, one hidden layer, and one output layer, where the number of units in the input layer is identical to that in the output layer. Typically, the transformation from the input layer to the hidden layer is called the encoder, and that from the hidden layer to the output layer refers to the decoder. The encoder extracts the features/representations from the input data, while the decoder reconstructs the input data from the features/representations. By minimizing the divergences between the input data and the reconstruction, one AE is trained. An SAE is stacked by multiple trained AEs for learning hierarchical representations that have gained more remarkable performance than ever before in the field of image classification [10]–[12].

When image data are fed to the SAE, they must be transformed into the vector-form beforehand, which will change their inherent structures and reduce the consecutive performance in turn. For instance, one image is with the form \( I \in \mathbb{R}^{n \times n} \), where the pixel \( I_{j,k} \) (\( 1 < j < n, 0 < k < n \)) has the close distance to the pixel \( I_{j-1,k} \). When \( j \) is vectorized to \( V \in \mathbb{R}^{n^2} \), the relationship between \( I_{j,k} \) and \( I_{j-1,k} \) will be changed and may not be neighbors anymore in \( V \). Extensive literatures have shown that adjacent information is a key factor in addressing images related problems [9], [10], [13]–[15]. To address this issue, Masci et al. [16] proposed the Convolutional AEs (CAEs), where the image data is directly fed in 2-D form. In CAEs, the encoder is composed of one convolutional layer followed by one pooling layer, and the decoder comprises only one deconvolutional layer. Multiple trained CAEs are stacked to a CNN for learning the hierarchical representations that enhance the final classification performance. Inspired by the advantages of CAEs in addressing data with the original 2-D form, variants of CAEs have been proposed subsequently. For example, Norouzi et al. [17] proposed the Convolutional Restricted Boltzmann Machines (RBM) [9], [18] (CRBM). Lee et al. [19] proposed the convolutional DBN by stacking a group of trained CRBMs. In addition, Zeiler et al. [20], [21] proposed the inverse convolutional ones based on the sparse coding schema [22], which inspired Kavukcuoglu et al. [23] to design the convolutional stacked sparse coding for solving object recognition tasks. Recently, Du et al. [24] proposed the Convolutional Denoising AE (CDAE) by using Denoising AE (DAE) [25] to learn the convolutional filters.

Although experimental results from the CAE and its variants have shown benefits in diverse applications, one major limitation exists in that the architectures of their stacked CNNs are inconsistent with those of state-of-the-art CNNs, such as ResNet [26] and VGGNet [27]. To be specific, because one CAE has one convolutional layer and one pooling layer in the encoder part, the stacked CNN has the same numbers of convolutional layers and pooling layers. However, state-of-the-art CNNs are with non-identical numbers of convolutional layers and pooling layers. Because the architecture of CNN is one key ingredient contributing to the final performance, the restriction on the numbers of convolutional layers and pooling layers of CAEs should be removed. However, choosing the appropriate numbers of convolutional layers and pooling layers is intractable due to the non-differentiable and non-convex characteristics in practice, which is related to the architecture optimization for neural networks.

This work was supported in part by the Marsden Fund of New Zealand Government under Contracts VUW1209, VUW1509 and VUW1615, Huawei Industry Fund E2880/3663, and the University Research Fund at Victoria University of Wellington 209862/3580, and 213150/3662, in part by the National Natural Science Fund of China for Distinguished Young Scholar of China under Grant 61803277.

Yanan Sun, Bing Xue, and Mengjie Zhang are with the School of Engineering and Computer Science, Victoria University of Wellington, PO Box 600, Wellington 6140, New Zealand (e-mails: yanan.sun@ecs.vuw.ac.nz; bing.xue@ecs.vuw.ac.nz; and mengjie.zhang@ecs.vuw.ac.nz).

Gary G. Yen is with the School of Electrical and Computer Engineering, Oklahoma State University, Stillwater, OK 74078 USA (e-mail: gyen@okstate.edu).
Algorithms for automatically searching for the optimal architectures of neural networks can be classified into three different categories. The first refers to the algorithms based on the stochastic system, including Random Search (RS) [28], Bayesian-based Gaussian Process (BGP) [29], [30]. Tree-structured Parzen Estimators (TPE) [31]. Sequential Model-Based Global Optimization (SMBO) [32], Evolving Unsupervised Deep Neural Network (EUDNN) [33], structure learning [34] and sparse feature learning [35]. The second covers the algorithms that are designed specifically for CNNs where multiple different building blocks exist. The Meta-modeling algorithm (MetaQNN) [36] and the Large Evolution for Image Classification (LEIC) algorithm [37] belong to this category. The third refers to the NeuroEvolution of Augmenting Topologies (NEAT) [38] algorithm and its diverse variants, such as [39]–[41]. Above all, there is one method that does not belong to these categories, i.e., the Grid Search method (GS), which tests every combination of the related parameters.

Particle Swarm Optimization (PSO) is a population-based stochastic evolutionary computation algorithm, motivated by the social behavior of fish schooling or bird flocking [42], [43], commonly used for solving optimization problems without requiring domain knowledge. Compared with other heuristic algorithms, PSO is enriched with the features of the simple concept, easy implementation, and computational efficiency. In PSO, the individuals are called particles, each particle maintains the best solution (denoted by pBest_i for the i-th particle) from the memory of itself, and the population records the best solution (denoted by gBest) from the history of all particles. During the process, particles expectedly cooperate and interact with the pBest_i and gBest, enhancing the search ability and pursuing the optimal solutions. Due to the characteristics of no requirements (e.g., convex or differentiable) imposed on the problems to be optimized, PSO has been widely applied to various real-world applications [44]–[46], naturally including the architecture design of neural networks, such as [47]–[52]. In the optimization of neural network architectures, these algorithms employ an implicit method to encode each connection of the neural networks and take PSO or its variants to search for the optimum. However, they cannot be utilized for CAEs and CNNs, even SAEs and DBNs, which are deep learning algorithms, where tremendous numbers of connection weights exist, causing the unaffordable cost for implementation and effective optimization in these existing PSO-based architecture optimization algorithms [53]. As have discussed, CAEs without the constraints on the numbers of the convolutional layer and pooling layers would be greatly preferred for stacking the state-of-the-art CNNs. However, the absolute numbers of these layers are unknown before the architecture is identified. When PSO is employed for the architecture optimization, particles will need to have different lengths. The reasons are that: 1) the length of the particle refers to the number of decision variables of the problem to be solved by PSO; and 2) in the architecture optimization problems, a set of different architectures are involved, and different architectures have different numbers of decision variables. However, the canonical PSO did not provide any way to update the velocity of particles with non-identical lengths. In addition, evaluating particles each of which represents a deep learning algorithm is time-consuming, and will become even more intractable for the population-based updating process. A common way to solve this problem is to employ intensive computational resources and utilize parallel-computation techniques.

The objective of this paper is to design and develop an effective and efficient PSO method to automatically discover the architecture of the flexible convolutional auto-encoder without manual intervention. To achieve this goal, we have specified the four aims as follows:

1) Propose a Flexible CAE (FCAE) where multiple convolutional layers and pooling layers can exist. The FCAE has no requirement on the particular numbers of the convolutional layers and the pooling layers, and have the potential for stacking to different types of CNNs.

2) Design a PSO-based Architecture Optimization (PSOAO) algorithm for the proposed FCAE. In PSOAO, we will propose an efficient encoding strategy to represent the FCAE architectures, which involve hundreds of thousands of parameters, into each particle, and we will also develop an effective velocity updating mechanism for particles with variable lengths.

3) Investigate the performance of the proposed FCAE when its architecture is optimized by the designed PSOAO on image classification benchmark datasets (i.e, the CIFAR-10 dataset [54], the MNIST dataset [9], the STL-10 dataset [55], and the Caltech-101 dataset [56]), compare the classification accuracy to peer competitors and examine the evolution effectiveness of PSOAO.

4) Investigate the effectiveness of the designed velocity updating method through quantitative experiments on the comparisons to its opponents.

The remainder of this paper is organized as follows. Background of the CAE and PSO is reviewed in Section II. This is followed by the details of the proposed PSOAO algorithm in Section III. Then, the experiment design and the result analysis are documented in Sections IV and V respectively. Finally, the conclusions and future work are drawn in Section VI.

II. LITERATURE REVIEW

This work will build FCAE and PSOAO based on CAE and PSO, respectively. Therefore, we would like to provide the skeletons of CAE and PSO as well as their limitations for FCAE in the following subsections, which could help the readers to conveniently understand our work in this paper. In addition, related works are also reviewed and commented, which helps the readers to easily appreciate the importance of our work in this paper.

A. CONVOLUTIONAL AUTO-ENCODER

For the convenience of the development, assuming CAEs are utilized for image classification tasks, and each image \(X \in \mathbb{R}^{w \times h \times c}\), where \(w, h, c\) refer to the image width, height, and number of channels, respectively. Fig. I illustrates the architecture of one CAE [16].

**Convolution:** Given the input data, the convolution operation employs one filter to slide with one defined stride, and
outputs the element that is the sum of the products of the filter and the input data with which this filter overlaps. All elements generated by one filter construct one feature map, and multiple feature maps are allowed in the convolution operations. The convolution operation has the SAME type and the VALID type. The parameters related to the convolutional operation are the filter size (width and height), the stride size (width and height), the convolutional type, and the number of feature maps.

**Pooling:** Pooling operation resembles the convolution operation, in addition to the filter and the way to generate the elements of the corresponding feature map. Specifically, the filter in a pooling operation is called a “kernel”, and no value exists in the kernel. There are two types of statistical indicators in the pooling operation: mean and maximal. Typically, maximal pooling is preferred in CAEs. The pooling operation requires the parameters: the kernel size (width and height), the stride size (width and height), and the pooling type.

**Deconvolution:** The deconvolutional operation is equivalent to the corresponding convolutional operation with inverse parameter settings. Specifically, the deconvolutional operation performs the convolutional operation with the filter and stride, which has been used in the corresponding convolutional operation, on the feature map that is resulted from the corresponding convolutional operation. In order to assure the output of the deconvolutional operation to have the same size as the input of the corresponding convolutional operation, extra zeros may be padded to the input of the deconvolutional operation.

**Learning of CAE:** The mathematical form of the CAE is represented by Equation (1), where the \(\text{conv}(\cdot)\), \(\text{pool}(\cdot)\), and \(\text{de}_\text{conv}(\cdot)\) denote the convolution, pooling, and deconvolution operations, respectively, \(F(\cdot)\) and \(G(\cdot)\) refer to the element-wise nonlinear activation functions, \(b_1\) and \(b_2\) are the corresponding bias terms, \(r\) and \(\hat{X}\) are the learned features and reconstruction of \(X\), \(l(\cdot)\) measures the differences between \(X\) and \(\hat{X}\), and \(\Omega\) is the regularization term to improve the feature quality. By minimizing \(L\), the CAE is trained, and then parameters in convolution operation, bias terms, and deconvolutional operation are identified. Encoders with these parameters from multiple trained CAEs are composed to be a CNN for learning hierarchical features that benefit the final classification performance \[7\].

\[
\begin{align*}
\hat{X} &= G(\text{de}_\text{conv}(r) + b_2) \\
L &= l(X, \hat{X}) + \Omega
\end{align*}
\]

\[\text{pool}(F(\text{conv}(X) + b_1))\]

**B. Motivation of FCAE**

As shown in Fig. 1, a CAE is composed of a convolutional layer followed by a pooling layer and then a deconvolutional layer. The transformation through the convolutional layer and then the pooling layer is called the encoder. The transformation through the deconvolutional layer is called the decoder. During the construction of a CNN by using CAEs, encoders from multiple CAEs are stacked together based on their training orders, and the input data of the current encoder is the output data of the previous one. However, this architecture is not able to form the state-of-the-art CNNs nor the common deep CNNs. Next, we will describe them in detail.

The architectures of a CNN stacked by CAEs and a state-of-the-art CNN named VGGNet \[27\] are shown in Fig. 2a and Fig. 2b respectively. From these two examples, it is evident that CAEs are incapable of stacking into VGGNet. The reason is that the CNN stacked by CAEs are with the same building blocks, i.e., the two-layer network including a convolutional layer followed by a pooling layer. Consequently, the stacked CNN is composed of a series of such building blocks and with the same number of the convolutional layers to that of the pooling layers. In addition, a convolutional layer must be followed by a pooling layer in the CNNs stacked by CAEs. As can be observed from Fig. 2b, there are multiple convolutional layers and pooling layers with the identical numbers, and also a convolutional layer is not necessarily followed by a pooling layer.

![Fig. 2. Architectures of the CNN stacked by CAEs (in Fig. 2a) and VGGNet (in Fig. 2b).](image-url)
Based on the description above, the root of the limitations from CAE exists in its architectures that the encoder of a CAE is composed one convolutional layer and then one pooling layer. Therefore, the concern is naturally raised that the architecture of CAE should be revised, which motivates the design of FCAE, i.e., in the encoder part of an FCAE, non-identical numbers of convolutional layers and pooling layers are allowed, and a pooling layer can follow a series of convolutional layers.

C. Particle Swarm Optimization

A typical PSO has the procedure as follows:

Step 1): Initialize the particles, redefine a maximal generation number maxt, and initialize a counter t = 0;

Step 2): Evaluate the fitness of particles;

Step 3): For each particle, choose the best one, pBestt, from its memory;

Step 4): Choose the best particle gBest from the history of all particles;

Step 5): Calculate the velocity \{v1, ..., vi, ..., vn\} of each particle \{x1, ..., xi, ..., xn\} by Equation (2);

Step 6): Update the position \{p1, ..., pi, ..., pn\} of each particle \{x1, ..., xi, ..., xn\} by Equation (3);

Step 7): Increase t by 1, if t < maxt, go to Step 2; otherwise, go to Step 8;

Step 8): The position of gBest is reported.

\[ v_i = \omega \cdot v_i + c_1 \cdot r_1 \cdot (p_g - p_i) + c_2 \cdot r_2 \cdot (p_p - p_i) \]  \hspace{1cm} (2)

\[ p_i \leftarrow p_i + v_i \]  \hspace{1cm} (3)

In Equation (2), \( \omega \) denotes the inertia weight, \( c_1 \) and \( c_2 \) are acceleration constants, \( r_1 \) and \( r_2 \) are random numbers between 0 and 1, and \( p_g \) as well as \( p_p \) denotes the positions of gBest as well as pBestt, respectively. \( v_i \) and \( p_i \) denote the velocity and position of the \( i \)-th particle \( x_i \), respectively. By integrating the “inertia”, “global search”, and “local search” terms into the velocity updating, the best position is expected to be found by particles.

Noting in Equation (2) that there are two subtraction operations existing in the “global search” and “local search”. In order to better understand how such a subtraction operation works, we will describe the details by taking the term \( p_g - p_i \) in “global search” as an example. Supposing the minimization problem to be solved is formulated as \( f(z_1, z_2, \ldots, z_n) \) where there are \( n \) decision variables \( \{z_1, z_2, \ldots, z_n\} \in \Psi \). When PSO is used to solve this problem, the \( i \)-th particle \( x_i \) will be randomly sampled from \( \Psi \), and its position is determined by a particular value \( p_i = \{z_1, z_2, \ldots, z_n\} \in \Psi \). Through the interaction formulated by Equations (2) and (3), the position of \( x_i \) is updated towards the position of the optimal solution. After a number of iterations, the optimization is solved and the final solution is the position of the global best particle, gBest.

In this example, the length of a particle is \( n \), i.e., the number of decision variables and also the dimension of the position. Obviously, \( p_g - p_i = \{z_1^g - z_1^i, z_2^g - z_2^i, \ldots, z_n^g - z_n^i\} \).

**Limitation of Using PSO for Architecture Design in FCAE**: The velocity updating requires the particle \( x_i \), gBest, and \( pBest_t \) to have the same/fixed length. When PSO is used for the architecture optimization of FCAE, particles represent the potential optimal architectures of FCAE. Because the optimal architecture of FCAE for solving the task at hand is unknown, particles with different variable lengths will emerge. Therefore, a novel velocity updating method needs to be designed in this regard.

D. Related Works

As we have discussed in Section I, algorithms for optimizing the architectures of deep neural networks fall into four different categories.

For the algorithms in the first category, RS [28] evaluates the randomly selected architectures with a predefined maximal trial number, and uses the best one in performance. RS has been reported that it only works under the condition when optimum is in a subspace of the whole search space [58]. However, it is not clear whether this fact equally applies to CNNs or not. Compared to RS, BGP [29], [30] utilizes more knowledge on choosing the potential optimal architecture based on Bayesian inference [30]. However, BGP has extra parameters, such as the kernels, which are hard to tune. TPE [31] works with the assumption that the parameters related to the architecture are independent, while most parameters in CNNs are indeed dependent such as the convolutional layer size and the strides. In addition, EUDNN is designed for unsupervised deep neural networks that are with different architectures from CNNs. Furthermore, methods in [34], [35] are only used for optimizing the particular network connections with the given architectures, such as the sparsity. To this end, algorithms in this category cannot be used to optimize the architectures of CNNs. Consequently, they are not suitable for FCAEs of which the architectures are based on CNNs.

For the algorithms in the second category, i.e., MetaQNN [36] and LEIC [37], both of them are designed specifically for optimizing the architectures of CNNs. Specifically, MetaQNN uses the reinforcement learning technique [59] to heuristically exploit the potential optimal architectures of CNNs, thoroughly evaluate them, and then choose the one that has the best fitness. LEIC employs nearly the same strategy to MetaQNN except for that LEIC used genetic algorithm as the heuristic method. Due to the complete training on each candidate in both algorithms, their deficiencies are also obvious, i.e., their training relies on the extensive computational resources. For instance, MetaQNN employed 10 Graphics Processing Unit (GPU) cards for 8-10 days, while LEIC employed 250 high-performance computers for 20 days on the CIFAR-10 test problem [24]. Unfortunately, sufficient computational resource is not necessarily available to all interested researchers.

Because the algorithms in the third category are all based on NEAT, their working flows are nearly the same. Specifically, the input layer and the output layer are assigned first, and then neurons between these two layers and connections from arbitrary two neurons are heuristically generated. With the
fitness evaluation upon each situation, the better ones are selected and the best one is expected to be found with genetic algorithm. In addition to the limitations occurred in the algorithms from the second category, other limitations from NEAT-based algorithms are that hybrid connections (i.e., the weight connections between the layers which are not adjacent) would be produced, and the configurations of the input layer and the output layer must be specified in advance, which are not allowed or applicable in CAEs.

Theoretically, GS can find the optimal architecture because of its exhaustive nature to try each candidate. However, it is impossible for GS to try each candidate in practice. Recently, experimental investigation [60] shows that GS is only suitable for the problems with no more than four parameters in practice. As have been shown in Subsection II.A a CAE will have more than 10 parameters even it contains only one convolutional layer and one pooling layer. In addition, GS cannot well handle parameters with continuous values because of the “interval” problems [60].

III. THE PROPOSED PSOAO ALGORITHM FOR FCAE

In this section, the details of the proposed PSOAO algorithm for FCAE will be provided. We will describe the encoding strategy which involves the FCAE representation, the particle initialization, the fitness evaluation, and the velocity and position updating mechanism.

A. Algorithm Overview

Algorithm 1: Framework of the PSOAO Algorithm

```plaintext
1  x ← Initialize the particles based on the proposed encoding strategy;
2  t ← 0;
3  while t < the maximal generation number do
4    Evaluate the fitness of each particle in x;
5    Update the pBest, and gBest;
6    Calculate the velocity of each particle;
7    Update the position of each particle;
8    t ← t + 1;
9  end
10  Return gBest for deep training.
```

Algorithm 1 outlines the framework of the designed PSOAO algorithm. Firstly, particles are randomly initialized based on the proposed encoding strategy (line 1). Then, particles start to evolve until the generation number exceeds the predefined one (lines 2-9). Finally, the gBest particle is picked up for obtaining the final performance through the deep training to solve tasks at hand (line 10).

During the evolution, the fitness of each particle is evaluated (line 4) first, and then the pBest, and gBest are updated based on the fitness (line 5). Next, the velocity of each particle is calculated (line 6) and their positions are updated (line 7) for the next generation of evolution. In the following subsections, keys aspects of PSOAO are detailed.

B.Encoding Strategy

For the convenience of the development, the definition of FCAE is given in Definition 1 by generalizing the building blocks in all CNNs. Obviously, the CAE is a special form of FCAE when the numbers of convolutional layers and pooling layers are both set to be 1.

**Definition 1.** A flexible convolutional auto-encoder (FCAE) encompasses one encoder and one decoder. The encoder is composed of the convolutional layers and pooling layers, where these two types of layers are not mixed and their numbers are flexible. The decoder part is the inverse form of the encoder.

We design an encoding strategy through variable-length particles to encode the potential architecture of one FCAE into one particle. Because the decoder part in an FCAE is the inverse form of the encoder, the particle in the proposed variable-length encoding strategy only encodes the encoder part for the reason of reducing the computational complexity. Each particle contains different numbers of convolutional layers and pooling layers. Based on the introduction of the convolution operation and pooling operation in Subsection II-A, all the encoded information of PSOAO for FCAE are summarized in Table I where the l2 denotes the weight decay regularization term for preventing from the overfitting problem [61]. Because only the convolutional layers involve weight parameters, this regularization term is applied only to the convolutional layers. Furthermore, because the output size will not change from the input size with the SAME convolutional layers, which is easy to control in automatic architecture discovering, the designed encoding strategy will not encode the type of the convolutional layers but default to the SAME type. As have mentioned in Subsection II.A that CAE prefers to the max pooling layer, we also don’t need to encode this parameter. In addition, three examples of the generally encoded particles in PSOAO are illustrated in Fig. 3. In the following, we will detail the rationale of this encoding strategy.

In the proposed PSOAO algorithm, a variable-length encoding strategy is designed for the particles representing FCAEs with different architectures. The major reason is that the optimal architecture is unknown prior to the optimization, and the fixed-length encoding strategy often imposing con-
With the designed variable-length encoding strategy, all the information of potential optimal architecture for FCAE can be flexibly represented for exploitation and exploration during the search process without manual intervention. A too small number denoting the maximal length would be inefficient for the optimized architecture of FCAE to solve complex problems. To avoid a large number which would consume much more unnecessary computation, and also results in worse performance within the same predefined evolution generation number. Furthermore, two types of layers exist in each particle, which increases the difficulty of employing the fixed-length encoding strategy. With the designed variable-length encoding strategy, all the information encoded in these two types of layers can be flexibly represented for exploitation and exploration during the search process without manual intervention.

C. Particle Initialization

Algorithm 2: Particle Initialization

Input: The population size \( N \), the maximal number of convolutional layers \( N_c \), and the maximal number of pooling layers \( N_p \).

Output: Initialized population \( x_0 \).

1. \( x \leftarrow \emptyset \).
2. While \( |x| \leq N \) do
   3. \( \text{conv} \_\text{list} \leftarrow \emptyset \);
   4. \( n_c \leftarrow \text{Uniformly generate an integer between} \ [1, N_c] \);
   5. While \( \text{conv} \_\text{list} \leq n_c \) do
      6. \( \text{conv} \_\text{unit} \leftarrow \text{Initialize a convolutional layer} \) with random settings;
      7. \( \text{conv} \_\text{list} \leftarrow \text{conv} \_\text{list} \cup \text{conv} \_\text{unit} \);
   8. End
   9. \( \text{pool} \_\text{list} \leftarrow \emptyset \);
   10. \( n_p \leftarrow \text{Uniformly generate an integer between} \ [1, N_p] \);
   11. While \( |\text{pool} \_\text{list}| \leq n_p \) do
      12. \( \text{pool} \_\text{unit} \leftarrow \text{Initialize a pooling layer} \) with random settings;
      13. \( \text{pool} \_\text{list} \leftarrow \text{pool} \_\text{list} \cup \text{pool} \_\text{unit} \);
   14. End
   15. Use \( \text{conv} \_\text{list} \) and \( \text{pool} \_\text{list} \) to generate a particle \( x \);
   16. \( x \leftarrow x \cup x \);
18. Return \( x \).

Algorithm 3: Fitness Evaluation

Input: The population \( x \), the training set \( D_{\text{train}} \), the number \( N_{\text{train}} \) of training epoch.

Output: The population \( x \) with fitness.

1. Calculate the reconstruction error and \( l_2 \) loss of the FCAE encoded by each particle in \( x \), and train the weights with \( N_{\text{train}} \) epochs;
2. Calculate the reconstruction error of each batch data in \( D_{\text{train}} \) and set the mean reconstruction error as the fitness of the corresponding particle;
3. Return \( x \).

Typically, a deep learning algorithm requires a training epoch number in the magnitude of \( 10^2 \) to \( 10^3 \) to train its weight parameters by gradient-based algorithms. This high computational issue is even worsen in population-based algorithms. In the proposed PSOAO algorithm, this number is specified at a very smaller number (e.g., 5 or 10) for speeding up the training. For example, it will take 2 minutes for training one epoch on the CIFAR-10 dataset (with 50,000 training samples) utilizing one GPU card with the model number of GTX1080. If we train it with \( 10^2 \) epochs for each particle with the population size of 50 for 50 generations, it will take about one year, which is not acceptable for the purpose of general academic research. The widely used solution for easing this adversity is to employ intensive computation resources, such as the LEIC.
Inertia and local search parts in Equation (2) are calculated after the global search part is derived by Algorithm 4, the promising performance of PSOA0 is still maintained. If the number of convolutional layers, pooling layer part (line 15). For the convolutional layer part, the optimum. When the evolution is terminated, the gBest is selected and one-time deep training is performed for reaching the complete velocity is calculated and the particle position is updated by Equation (3). The first is on the convolutional layer part of gBest and x (lines 2-14), and the second is on the pooling layer part (line 15). For the convolutional layer part, if the number of convolutional layers, cg, from gBest is smaller than that from x, new convolutional layers initialized with zero values are padded to the tail of cg. Otherwise, extra convolutional layers are truncated from the tail of cg. The “global search” part of Equation (2). Algorithm 4 shows the details of the x-reference method. Specifically, the x-reference method is applied twice in the same manner in the velocity updating for the “global search” part of Equation (2). The first is on the convolutional layer part of gBest and x (lines 2-14), and the second is on the pooling layer part (line 15). For the convolutional layer part, if the number of convolutional layers, cg, from gBest is smaller than that from x, new convolutional layers initialized with zero values are padded to the tail of cg. Otherwise, extra convolutional layers are truncated from the tail of cg. The “global search” part is derived by Algorithm 4, the “inertia” and “local search” parts in Equation (2) are calculated as normal, then the complete velocity is calculated and the particle position is updated by Equation (3). For an intuitive understanding the proposed x-reference velocity updating method, an example is provided in Fig. 4. Specifically, Fig. 4 displays the gBest and x that are used to do the “global search” part in the velocity updating. In Fig. 4, the convolutional layers and pooling layers are collected from gBest and x. Because the lengths of convolutional layers and pooling layers from x are 2 and 4, and those from gBest are 3 and 2, the last convolutional layer from the convolutional layer part from gBest is truncated, and the other two pooling layers are padded to the tail of the pooling layer part of gBest. Specifically, the running time on the investigated benchmark datasets are shown in Table II, the adopted computational resources for the investigated benchmark datasets are given in Subsection IV.C and the experimental results regarding the performance are shown in Subsections V-A and V-C.

E. Velocity Calculation and Position Update

In PSOA0, the particles are with different lengths, and Equation (2) cannot be directly used. To solve this problem, we design a method named “x-reference” to update the velocity. In x-reference, the lengths of gBest and pBesti refer to the length of the current particle x, i.e., gBest and pBesti keep the same length to that of x. Because the pBesti is selected from the memory of each particle, and the gBest is chosen from all particles, the current particle x is always with the same length of pBesti and the x-reference is applied only to the “global search” part of Equation (2). Algorithm 4 shows the details of the x-reference method.

Specifically, the x-reference method is applied twice in the same manner in the velocity updating for the “global search” part of Equation (2). The first is on the convolutional layer part of gBest and x (lines 2-14), and the second is on the pooling layer part (line 15). For the convolutional layer part, if the number of convolutional layers, cg, from gBest is smaller than that from x, new convolutional layers initialized with zero values are padded to the tail of cg. Otherwise, extra convolutional layers are truncated from the tail of cg. After the “global search” part is derived by Algorithm 4, the “inertia” and “local search” parts in Equation (2) are calculated as normal, then the complete velocity is calculated and the particle position is updated by Equation (3). For an intuitive understanding the proposed x-reference velocity updating method, an example is provided in Fig. 4. Specifically, Fig. 4 displays the gBest and x that are used to do the “global search” part in the velocity updating. In Fig. 4, the convolutional layers and pooling layers are collected from gBest and x. Because the lengths of convolutional layers and pooling layers from x are 2 and 4, and those from gBest are 3 and 2, the last convolutional layer from the convolutional layer part from gBest is truncated, and the other two pooling layers are padded to the tail of the pooling layer part of gBest. In particular, the padded pooling layers are created with the values of encoded information equal to 0. Fig. 4 demonstrates the updating between the convolutional layer part and pooling layer part from gBest and x. Fig. 4 shows the results of this updating.

No matter whether padding or truncating is operated in the designed x-reference velocity updating method, the goal is to make the same length of convolutional layers and pooling layers in gBest to that in x, respectively. The mechanism behind this design is discussed as follows. In PSOA0, there are a group of particles with different lengths in the population, with the same goal of searching for the optimal architectures of FCAEs for solving image classification tasks. If we have each particle follow the length of gBest (i.e., the gBest-reference velocity updating method), all the particles will have the same length to that of gBest from the second generation. Because the pBesti is chosen from the memory of each particle, the gBest, pBesti, and current particle x may all have the same length from the third generation. Consequently, all particles participate in the optimization with one particular depth of FCAE and only change the encoded information. Indeed, the variation of length regarding gBest can be seen as an exploration search behavior, while that of the encoded information is viewed as an exploitation search behavior. When all particles are in the same length, the length of gBest will be constant until it terminates. In this regard, the exploration search ability is lost if we employ the gBest-reference velocity updating method. In addition, keeping the length of x equal to gBest can also be viewed as the losses of diversity, which

| Algorithm 4: The x-reference Velocity Calculation Method |
|--------------------------------------------------------|
| **Input:** The particle x, the gBest, the acceleration |
| constant c1.                                           |
| **Output:** The global search part of velocity updating.|
| 1. r1 ← Randomly sample a number from [0, 1];          |
| 2. cg ← Extract the convolutional layers from gBest;   |
| 3. cx ← Extract the convolutional layers from x;       |
| 4. pos_c ← 0;                                         |
| 5. if |cg| < |cx| then                                     |
| 6. e ← Initialize |cx| − |cg| convolutional layers with |
| encoded information of zeros;                          |
| 7. cg ← Pad e to the tail of cg;                       |
| 8. else                                                |
| 9. cg ← Truncate the last |cg| − |cx| convolutional |
| layers from cg;                                        |
| 10. end                                                |
| 11. for i = 1 to |cx| do                                 |
| 12. p_{cg}, p_x ← Extract the position of the i-th    |
| convolution layer from cg and cx;                      |
| 13. pos_c ← pos_c ∪ c_i · r1 · (p_{cg} − p_x);         |
| 14. end                                                |
| 15. pos_p ← Analogy the operations in lines 2-14 on    |
| the pooling layers of gBest and x;                     |
| 16. **Return** pos_c ∪ pos_p.                          |


would easily lead to the premature convergence in population-based algorithms. Both the loss of exploration search and the premature phenomenon will result in a poor performance. An experiment is conducted in Subsection V-D to further quantitatively investigate this velocity updating design.

**F. Deep Training on gBest**

When the evolution of PSOAO is finished, the best particle, gBest, is picked for deep training. As stated in Subsection III-D that each particle is trained with only a few epochs, which is not the optimal performance for solving real-world applications. To solve this concern, the deep training is necessary. Typically, the process of the deep training is the same to the fitness evaluation in Subsection III-D except for a larger epoch number, say 100 or 200.

**IV. EXPERIMENTAL DESIGN**

In this section, the benchmark datasets, peer competitors, and the parameter settings are detailed for the experiments investigating the performance of the proposed FCAE in which its architecture is optimized by the proposed PSOAO algorithm.

**A. Benchmark Datasets**

The experiments are conducted on four image classification benchmark datasets, which are widely used and specifically for investigating the performance of AEs. They are the CIFAR-10 dataset [54], the MNIST dataset [9], the STL-10 dataset [55], and the Caltech-101 dataset [56]. Fig. 5 shows examples from these benchmark datasets. In the following, these chosen datasets are briefly introduced.

1) **CIFAR-10 dataset**: It contains 50,000 training images and 10,000 test images. Each one is a 3-channel RGB image in the size of $32 \times 32$, and belongs to one of 10 categories of natural objectives (i.e., truck, ship, horse, frog, dog, deer, cat, bird, automobile, and airplane). Each category is with roughly the same number of images. In addition, different objects occupy different areas of the images.

2) **MNIST dataset**: It is a handwritten digit recognition dataset to classify the numeral numbers of $0 \cdots 9$, including 60,000 training images and 10,000 test images. Each one is a 1-channel gray image in the size of $28 \times 28$, and the samples in each category are with different variations, such as rotations. Each category is composed of the same number of image samples.

3) **STL-10 dataset**: It is a widely used dataset for unsupervised learning, containing 100,000 unlabeled images, 5,000 training images, and 8,000 test images from 10-category natural image object recognition (i.e., airplane, bird, car, cat, deer, dog, horse, monkey, ship, truck). Each one is a 3-channel RGB image in the size of $96 \times 96$. In addition, the unlabeled images contain images that are beyond the 10 categories. Due to the small number of training samples, this dataset challenges the feature learning ability of CAEs/AEs and PSOAO.

4) **Caltech-101 dataset**: It is a 101-category image classification dataset where the weights and heights of images vary from 80 to 708 pixels. Most images are 3-channel RGB while occasionally gray, and with different images in each category from 31 to 800. In addition, most images only display a small part of the image, and other areas are occupied by noises for increasing the difficulty in classification. Due to the quite small number of images, and the non-identical numbers of images in
The stacked forms of these AEs are also employed as the peer competitors for a comprehensive comparison. They are the Sparse AE (SAE) [62] and Denoising AE (DAE) [25]. In addition, two widely used variants of AEs are also employed as the peer competitors for a comprehensive comparison. They are the Sparse AE (SAE) [62] and Denoising AE (DAE) [25].

Because this paper aims at proposing an FCAE that could be stacked to the state-of-the-art CNNs, peer competitors for comparisons here also include the stacked forms of these AEs, i.e., the Stacked CAE (S-CAN), the Stacked CRBM (S-CRBM), the Stacked CDAE (S-CDAE), the Stacked SAE (S-SAE), and the stacked DAE (S-DAE). The Stacked form of the proposed FCAE is SFCAE.

C. Parameter Settings

The peer competitors SCAE, SCRBM, S-SAE, and S-DAE have been investigated very recently on the chosen benchmark datasets, and their architectures have been manually tuned with domain expertise [24]. Their classification results are directly cited from the original publications, thus none of their parameter settings needs to be specified. In addition, the state-of-the-art S-CDAE provides the classification results with the usage of only one and two building blocks on the chosen benchmark datasets. In order to do a fair comparison, we also perform experiments on SFCAE with at most two building blocks. Note that the SFCAE is tested on the chosen benchmark datasets without the preprocessing of data augmentation for keeping consistency to its peer competitors. In the following, the parameter settings of PSOAO are provided in detail.

In the designed PSOAO algorithm, PSO related parameters are specified based on their conventions [63], i.e., the inertia weight $w$ is set to be 0.72984, the acceleration constants $c_1$ and $c_2$ are both set to be 1.496172, and the initial velocity is set to be 0. The training sets of MNIST and CIFAR-10, the unlabeled data of STL-10 are naturally used for their fitness evaluations. Due to the non-identical and a quite small number of images in each category of the Caltech-101 dataset, 30 images randomly selected from each category are used for the fitness evaluations and also as the training set based on the suggestions in [24]. Because inappropriate settings of convolutional operations and pooling operations would lead to an unaffordable computational cost and make FCAE being incompetent, in the exploration of each particle, the number of feature maps is set to be $[20, 100]$, the kernel with the same size of width and height is set to be $[2, 5]$, the maximal number of pooling layers is set to be 1, and that of convolutional layer is set to be 5. Note that only the square convolutional filters and pooling kernels are investigated in the experiments, which is based on the conventions from state-of-the-art CNNs [26, 57]. In addition, the coefficient of $l_2$ term is set to be $[0.0001, 0.01]$, which is a commonly utilized range for training neural networks in practice.

FCAE with the architecture determined by PSOAO and the weights initialized by the widely used Xavier method [64], by adding one full connection layer with 512 units and 50% Dropout [65] from the conventions of deep learning community, are used for the deep training. We investigate the classification results by feeding the trained model with the corresponding test set employing the widely used rectifier linear unit [66] as the activation function, the Adam [67] optimizer with its default settings as the training algorithm for weight optimization, and the BatchNorm [68] technique for speeding up the training. For keeping consistency in the results to be compared, the experiments with the trained model on each benchmark dataset are also independently performed 5 times. Due to the extreme imbalance training data exist in the Caltech-101 dataset, we investigate this dataset based on its convention [19], i.e., investigating the classification accuracy on each category of the images, and then reporting the mean and standard derivations over the whole dataset.

The proposed PSOAO algorithm is implemented by Tensorflow [69], and each copy of the source code runs on one GPU card with the same model number of GTX1080. In addition, the architecture configurations of FCAEs optimized by PSOAO for the benchmark datasets in these experiments are provided in Subsection V-E. Training time of the proposed PSOAO algorithm on the benchmark datasets are shown in Table [II].

---

\[2\] The test data of Caltech-101 dataset is the entire dataset excluding from the training set.
TABLE II
CONSUMED TIME (HOURS) OF THE PROPOSED PSOAO ALGORITHM FOR DIFFERENT BENCHMARK DATASETS.

| Algorithm | CIFAR-10 | MNIST | STL-10 | Caltech-101 |
|-----------|----------|-------|--------|-------------|
| SFCAE-1   | 81.5     | 118   | 230    | 22.5        |

V. EXPERIMENTAL RESULTS AND ANALYSIS

A. Overview Performance

Table III shows the mean and standard derivations of the classification accuracies of the proposed FCAE method, whose architecture is optimized by the designed PSOA O algorithm, and the peer competitors on the benchmark datasets. Because literatures do not provide the standard derivations of SCAE on CIFAR-10 and MNIST, SCRB M on CIFAR-10, and SDAE on MNIST, only their mean classification accuracies are shown. The references in Table III denote the sources of the corresponding mean classification accuracies, and the best mean classification results are highlighted in bold. The terms “SFCAE-1” and “SFCAE-2” refer to the SFCAE with one and two building blocks, respectively, which is the same meaning as the terms “SCDAE-1” and “SCDAE-2.”

TABLE III
THE CLASSIFICATION ACCURACY OF THE PROPOSED FCAE METHOD AGAINST ITS PEER COMPETITORS ON THE CHOSEN BENCHMARK DATASETS.

| Algorithm        | CIFAR-10 | MNIST       | STL-10 | Caltech-101 |
|------------------|----------|-------------|--------|-------------|
| SSAE             | 74.0 (0.9) | 96.29 (0.12) | 55.5 (1.2) | 66.2 (1.2)   |
| SDAE             | 70.1 (1.0) | 99.06 (0.25) | 53.5 (1.5) | 59.5 (0.3)   |
| SCAE             | 78.2 (0.9) | 99.29 (0.16) | 40.0 (3.1) | 58.0 (2.0)   |
| SCBM             | 78.9 (0.9) | 99.18 (0.19) | 43.5 (2.3) | 65.4 (0.5)   |
| SCDAE-1          | 75.0 (1.2) | 99.17 (0.10) | 56.6 (0.8) | 71.5 (1.6)   |
| SCDAE-2          | 80.4 (1.1) | 99.38 (0.05) | 60.5 (0.9) | 78.6 (1.2)   |
| SFCAE-1          | 78.9 (0.3) | 99.30 (0.03) | 61.2 (1.2) | 79.8 (0.0)   |
| SFCAE-2          | 83.5 (0.5) | 99.51 (0.09) | 56.8 (0.2) | 79.6 (0.0)   |

It is clearly shown in Table III that FCAE outperforms the traditional AEs (i.e., the SSAE and the SDAE) and traditional CAEs (i.e., the SCAE and the SCBM) on all benchmark datasets. In addition, FCAE also outperforms the state-of-the-art CAEs (i.e., the SCDAE-1 and the SCDAE-2) on these benchmark datasets. Furthermore, the best results on CIFAR-10 and MNIST are reached by SFCAE-2, and those on STL-10 and Caltech-101 are by SFCAE-1. Note that SFCAE-2 performs worse on STL-10 and Caltech-101 than SFCAE-1, which is caused by the much smaller numbers of training instances in these two benchmark datasets, and deeper architectures are suffered from the overfitting problem. Because CIFAR-10 and MNIST are with many more training samples (50,000 in CIFAR-10 and 60,000 in MNIST), a deeper architecture naturally results in the promising classification accuracy. In summary, when the architecture of the proposed FCAE method is optimized by the designed PSOA O algorithm, FCAE shows superior performance among its peer competitors on all the four image classification benchmark datasets.

B. Evolution Trajectory of PSOA O

In order to intuitively investigate the efficacy of the designed PSOA O algorithm in optimizing the architectures of the proposed FCAE method, its evolution trajectories on the chosen benchmark datasets during the training phases are plotted in Fig. 6 where the horizontal axis denotes the number of generations in the evolution, and the vertical axis denotes the fitness values of the gBest.

As can be seen from Figs. 6a, 6b, 6c, and 6d, PSOA O has converged within the specified maximal generation number. Specifically, it has converged since about the 15th generation on all benchmark datasets for both SFCAE-1 and SFCAE-2, and about the 5th generation on the CIFAR-10 and Caltech-101 datasets for SFCAE-2. Note that the reconstruction error of SFCAE-2 is smaller than that of SFCAE-1 on STL-10 dataset (shown in Fig. 6c), which is caused by the different input data for them.

C. Performance on Different Numbers of Training Examples

In this subsection, we investigate the classification performance of the proposed FCAE method whose architecture is optimized by the designed PSOA O algorithm on different numbers of training samples. The peer competitors on the MNIST dataset are SCRB M [19], ULIFH [71], SSE [72], and SCAE [16], and those for the CIFAR-10 dataset are SCAE [16], Mean-cov. RBM [70], and K-means (4k feat) [55]. The reason for choosing these benchmark datasets and peer competitors is that the literature has provided their corresponding information that is widely used by the comparisons between various variants of CAEs.

TABLE IV
THE CLASSIFICATION ACCURACY OF FCAE-2 AGAINST PEER COMPETITORS ON THE DIFFERENT NUMBERS OF TRAINING SAMPLES FROM MNIST.

| # samples | 1K   | 2K   | 3K   | 5K   | 10K  | 60K  |
|-----------|------|------|------|------|------|------|
| SCRB M    | 97.38 | 97.87 | 98.09 | 98.41 | —    | 99.18 |
| ULIFH     | 96.79 | 97.47 | —    | 98.49 | —    | 99.36 |
| SSAE      | 97.27 | —    | 98.17 | —    | 98.50 | —    |
| SCAE      | 92.77 | —    | —    | 98.12 | 99.29 | —    |
| SFCAE     | 97.49 | 98.45 | 98.80 | 99.09 | 99.16 | 99.51 |

Tables IV and V show the experimental results of FCAE-2 with the architecture confirmed in Subsection V-A on different numbers of training samples of the MNIST and CIFAR-10 benchmark datasets. The symbol “—” denotes that there is no result reported in the corresponding literature. The best classification accuracy is highlighted in bold.
As can be seen from Tables IV and V, SFCAE-2 surpasses all peer competitors on these two datasets. Especially, with a much smaller number (1K) of training examples, SFCAE-2 performs better than the seminal work of CAE (SCAE) with 4.72\% classification accuracy improvement on MNIST and 6.09\% on CIFAR-10. These results show the promising scalability of SFCAE-2 in dealing with different numbers of training samples.

D. Investigation on $x$-reference Velocity Calculation

![Graph showing classification accuracy comparisons between gBest-reference and x-reference velocity updating strategy.]

Fig. 7. Classification accuracy comparisons between the $x$-reference and $gBest$-reference velocity updating strategy in the designed PSOA method.
To further investigate the superiority of the proposed x-reference velocity updating method, we replace it with the proposed PSOAO algorithm with the gBest-reference velocity updating method, to compare the performance on the chosen benchmark datasets introduced in Subsection [V-A]. To achieve this, we first let the length of pBesti equal that of gBest. If the length of the convolutional layers in pBesti is less than that of gBest, zeros are padded. Otherwise, truncating the corresponding parts from pBesti. We use this method in the pooling layers of pBesti and these two types of layers in x. Then, we use Equations (2) and (3) to update the position of each particle. The experimental results are shown in Fig. 7, where Fig. 7a displays the results of FSCAE-1 while Fig. 7b displays those of FSCAE-2.

As can be seen from Fig. 7a with the x-reference velocity updating method, SFCAE-1 achieves the classification accuracy improvements of 5.7%, 8.8%, 5.9%, and 7.9% on the CIFAR-10, MNIST, STL-10, and Caltech-101 benchmark datasets, respectively. The same promising performance of SFCAE-2 can also be observed with the classification accuracy improvements of 7.7%, 10.6%, 5.6%, and 9.6% on these chosen benchmark datasets, as shown in Fig. 7b.

In summary, the analysis and experimental results convincingly justify the effectiveness of the proposed x-reference velocity updating method in the proposed PSOAO algorithm.

E. Investigation on The Obtained architectures

Tables VI, VII, VIII, IX, and XII show the architecture configurations of SFCAE-1 for the CIFAR-10, MNIST, STL-10 and Caltech-101 datasets, respectively. Additionally, Tables VII, VIII, IX, XI, and XII show the architecture configurations of SFCAE-2 for the CIFAR-10, MNIST, STL-10 and Caltech-101 datasets, respectively. Noting that the SAME convolutional layers, the max pooling layers and the strides with 1 × 1 are used based on the experimental settings, therefore, these configurations are not shown in these tables.

| TABLE VI | THE ARCHITECTURE CONFIGURATION OF SFCAE-1 ON THE CIFAR-10 DATASET. |
|-----------------|---------------------------------------------------------------|
| layer type | configuration |
| conv | Filter: 2 × 2, # Feature map: 24, L2: 0.0001 |
| conv | Filter: 3 × 3, # Feature map: 57, L2: 0.0001 |
| conv | Filter: 5 × 5, # Feature map: 63, L2: 0.0006 |
| conv | Filter: 5 × 5, # Feature map: 35, L2: 0.0071 |
| conv | Filter: 3 × 3, # Feature map: 76, L2: 0.0015 |
| pooling | Kernel: 2 × 2, Stride: 2 × 2 |

| TABLE VII | THE ARCHITECTURE CONFIGURATION OF SFCAE-2 ON THE CIFAR-10 DATASET. |
|-----------------|---------------------------------------------------------------|
| layer type | configuration |
| conv | Filter: 4 × 4, # Feature map: 36, L2: 0.0001 |
| pooling | Kernel: 2 × 2, Stride: 2 × 2 |

The obtained architectures are based on our designed encoding strategy. The encoding strategy in this paper is designed

| TABLE VIII | THE ARCHITECTURE CONFIGURATION OF SFCAE-1 ON THE MNIST DATASET. |
|-----------------|---------------------------------------------------------------|
| layer type | configuration |
| conv | Filter: 2 × 2, # Feature map: 100, L2: 0.0010 |
| conv | Filter: 2 × 2, # Feature map: 82, L2: 0.0018 |
| conv | Filter: 2 × 2, # Feature map: 100, L2: 0.0001 |
| pooling | Kernel: 2 × 2, Stride: 2 × 2 |

| TABLE IX | THE ARCHITECTURE CONFIGURATION OF SFCAE-2 ON THE MNIST DATASET. |
|-----------------|---------------------------------------------------------------|
| layer type | configuration |
| conv | Filter: 3 × 3, # Feature map: 89, L2: 0.0001 |
| conv | Filter: 3 × 3, # Feature map: 93, L2: 0.0005 |
| pooling | Kernel: 2 × 2, Stride: 2 × 2 |

| TABLE X | THE ARCHITECTURE CONFIGURATION OF SFCAE-1 ON THE STL-10 DATASET. |
|-----------------|---------------------------------------------------------------|
| layer type | configuration |
| conv | Filter: 2 × 2, # Feature map: 83, L2: 0.0096 |
| conv | Filter: 3 × 3, # Feature map: 77, L2: 0.0086 |
| conv | Filter: 3 × 3, # Feature map: 83, L2: 0.0025 |
| pooling | Kernel: 2 × 2, Stride: 2 × 2 |

| TABLE XI | THE ARCHITECTURE CONFIGURATION OF SFCAE-2 ON THE STL-10 DATASET. |
|-----------------|---------------------------------------------------------------|
| layer type | configuration |
| conv | Filter: 2 × 2, # Feature map: 83, L2: 0.0094 |
| conv | Filter: 4 × 4, # Feature map: 49, L2: 0.0087 |
| pooling | Kernel: 2 × 2, Stride: 2 × 2 |

| TABLE XII | THE ARCHITECTURE CONFIGURATION OF SFCAE-1 ON THE CALTECH-101 DATASET. |
|-----------------|---------------------------------------------------------------|
| layer type | configuration |
| conv | Filter: 2 × 2, # Feature map: 14, L2: 0.0084 |
| conv | Filter: 5 × 5, # Feature map: 8, L2: 0.0002 |
| pooling | Kernel: 2 × 2, Stride: 2 × 2 |

| TABLE XIII | THE ARCHITECTURE CONFIGURATION OF SFCAE-2 ON THE CALTECH-101 DATASET. |
|-----------------|---------------------------------------------------------------|
| layer type | configuration |
| conv | Filter: 5 × 5, # Feature map: 15, L2: 0.0051 |
| pooling | Kernel: 2 × 2, Stride: 2 × 2 |
based on the architectures of the traditional convolutional neural networks (CNNs) that are composed of several blocks, and each block is composed of several convolutional layers followed by the pooling layer. Unsurprisingly, these obtained architectures follow the architectures of the traditional CNNs and the architectures of known and manually designed network, such as the VGGNet [27].

Recently, there are two famous types of CNNs, i.e., ResNet [26] and DenseNet [57]. Because these two networks are different from the architectures of the traditional CNNs due to the skip and dense connections, the obtained architectures are different to these networks. Due to the promising performance of ResNet and DenseNet on large-scale image classification tasks shown in their experiments, we will in future investigate a new encoding strategy that is capable of encoding the skip connections of ResNet and dense connections of DenseNet.

VI. CONCLUSIONS AND FUTURE WORK

The goal of the paper is to develop a novel PSO algorithm (namely PSOAO) to automatically discover the optimal architecture of the flexible convolutional auto-encoder (named FCAE) for image classification problems without manual intervention. This goal has been successfully achieved by defining the FCAE that has the potential to construct the state-of-the-art deep convolutional neural networks, designing an efficient encoding strategy that is capable of representing particles with non-identical lengths in PSOAO, and developing an effective velocity updating mechanism for these particles. The FCAE with the “optimal” architecture is achieved by PSOAO, and compared with five peer competitors including the most state-of-the-art algorithms on four benchmark datasets, specifically used by auto-encoders for image classification. The experimental results show that FCAE remarkably outperforms all the compared algorithms on all adopted benchmark datasets in term of their classification accuracies. Furthermore, FCAE with only one building block can surpass the state-of-the-art with two building blocks on the STL-10 and Caltech-101 benchmark datasets. In addition, FCAE reaches the best classification accuracies compared with the four peer competitors when only 1K and 10K training images of the MNIST benchmark dataset are used, and significantly outperforms three peer competitors when only 1K and 10K training images of the CIFAR-10 benchmark dataset are used. Moreover, the proposed PSOAO algorithm also shows the excellent characteristic of fast convergence by investigating its evolution trajectories, and the effective velocity updating mechanism through the quantitative comparison to its opponents.

Although deep CNNs have achieved the current state-of-the-art results for image classification, the architectures and hyper-parameters are largely determined by manual tuning based on domain expertise/knowledge. This paper provides a direction, which shows that such manual work can be replaced by automatic learning through evolutionary approaches. In the future, we will investigate simpler evolutionary methods on more complex CNN models with using much less computational resources.

REFERENCES

[1] D. E. Rumelhart, G. E. Hinton, R. J. Williams et al., “Learning representations by back-propagating errors,” Cognitive modeling, vol. 5, no. 3, p. 1, 1988.
[2] H. Bourlard and Y. Kamp, “Auto-association by multilayer perceptrons and singular value decomposition,” Biological cybernetics, vol. 59, no. 4, pp. 291–294, 1988.
[3] G. E. Hinton and R. S. Zemel, “Autoencoders, minimum description length and helmholtz free energy,” in Advances in neural information processing systems, 1994, pp. 3–10.
[4] H. Schwenk and M. Milgram, “Transformation invariant autoassociation with application to handwritten character recognition,” in Advances in neural information processing systems, 1995, pp. 992–998.
[5] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” science, vol. 313, no. 5786, pp. 504–507, 2006.
[6] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, “Greedy layer-wise training of deep networks,” in Advances in neural information processing systems, 2007, pp. 153–160.
[7] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” Nature, vol. 521, no. 7553, pp. 436–444, 2015.
[8] G. E. Hinton, S. Osindero, and Y.-W. Teh, “A fast learning algorithm for deep belief nets,” Neural computation, vol. 18, no. 7, pp. 1527–1554, 2006.
[9] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.
[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in neural information processing systems, 2012, pp. 1097–1105.
[11] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, “Learning hierarchical features for scene labeling,” IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 8, pp. 1915–1929, 2013.
[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[13] J. P. Pluim, J. A. Maintz, and M. A. Viergever, “Mutual-information-based registration of medical images: a survey,” IEEE transactions on medical imaging, vol. 22, no. 8, pp. 986–1004, 2003.
[14] G. Pass, R. Zabih, and J. Miller, “Comparing images using color coherence vectors,” in Proceedings of the fourth ACM international conference on Multimedia. ACM, 1997, pp. 65–73.
[15] X. Peng, B. Zhao, R. Yan, H. Tang, and Z. Yi, “Bag of events: An efficient probability-based feature extraction method for aer image sensors,” IEEE transactions on neural networks and learning systems, vol. 28, no. 4, pp. 791–803, 2017.
[16] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, “Stacked convolutional auto-encoders for hierarchical feature extraction,” Artificial Neural Networks and Machine Learning–ICANN 2011, pp. 52–59, 2011.
[17] M. Norouzi, M. Ranjarb, and G. Mori, “Stacks of convolutional restricted boltzmann machines for shift-invariant feature learning,” in Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009, pp. 2735–2742.
[18] P. Smolensky, “Information processing in dynamical systems: Foundations of harmony theory,” COLORADO UNIV AT BOULDER DEPT OF COMPUTER SCIENCE, Tech. Rep., 1986.
[19] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, “Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations,” in Proceedings of the 26th annual international conference on machine learning. ACM, 2009, pp. 609–616.
[20] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, “Deconvolutional networks,” in Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010, pp. 2528–2535.
[21] M. D. Zeiler, G. W. Taylor, and R. Fergus, “Adaptive deconvolutional networks for mid and high level feature learning,” in Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, 2011, pp. 2018–2025.
[22] B. A. Olshausen and D. J. Field, “Sparse coding with an overcomplete basis set: A strategy employed by v1?” Vision research, vol. 37, no. 23, pp. 3311–3325, 1997.
[23] X. Kavukcuoglu, P. Sermanet, Y.-L. Boureau, K. Gregor, M. Mathieu, and Y. L. Cun, “Learning convolutional feature hierarchies for visual recognition,” in Advances in neural information processing systems, 2010, pp. 1090–1098.
