IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium | 978-1-7281-6374-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/IGARSS39084.2020.9323429

RSSM-NET: REMOTE SENSING IMAGE SCENE CLASSIFICATION BASED ON MULTI-OBJECTIVE NEURAL ARCHITECTURE SEARCH

Yuting Wan[1], *Yanfei Zhong*[1,*], *Ailong Ma*[1], *JunJue Wang*[1], *Ruyi Feng*[2]

[1]State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing,
Wuhan University, Wuhan 430079, China
[2]School of Computer Science, China University of Geosciences, Wuhan, China
*Corresponding author E-mail: zhongyanfei@whu.edu.cn; Phone: 86-27-68779969

ABSTRACT

The deep learning (DL)-based scene classification methods have been obtained the remarkable attention for the high spatial resolution remote sensing (HRS) imagery. However, from one aspect, the existing DL methods in HRS image scene classification are usually the variations of the natural image processing methods and often the inherent network structures; from another aspect, the strenuous and significant efforts have been devoted to the design of relevant network structures by human experts. In this paper, learning from the natural evolution, the deep neural network is expected to be globally evolved by the machine for automatically adapting the structure of the HRS imagery, a multi-objective neural architecture search based HRS image scene classification method is proposed (RSSM-Net). The two objectives of minimizing a classification error and the computational complexity have been simultaneously optimized through the evolutionary multi-objective method, the competitive neural architectures in a Pareto solution set are then obtained. The effectiveness is proved by the experiment of the UC Merced dataset with several networks designed by human experts.

Index Terms— Remote sensing, scene classification, neural architecture search, evolutionary algorithm, multi-objective optimization

1. INTRODUCTION

Due to the complexity of spatial and structural patterns of the HRS images, the scene classification is still a difficult problem and remains the remarkable attraction [1]. In the past decades, the low-level features, such as the features of color histogram and local binary patterns (LBP), which are extracted from the HRS images based on pre-defined algorithms, and the bag-of-visual-words (BoVW) model is used as the classification method [2]. However, the deep-level and learning features are ignored in these traditional methods with handcrafted features, the essential features of specific HRS images are then difficultly extracted.

Fortunately, with the rapid development of the DL technology, especially the convolutional neural networks (CNNs), have been the powerful tools to discover the intricate structures and extract the essential features of HRS imagery [3]. Moreover, based on the image classification in the ImageNet large-scale visual recognition challenge [4], the GoogLeNet and CaffeNet are then successfully applied in the HRS image scene classification [5].

However, in order to design a satisfactory CNN for scene classification, the comprehensive domain knowledges are required. Fortunately, thanks to the rapid development of hardware, the computing power has been greatly improved, Google releases the AutoML platform, and the mechanisms are collectively referred to as neural architecture search (NAS). The main idea of NAS is: 1) define the search space, 2) find out the candidate network structures through the search strategy and evaluate them, 3) carry out the next iteration according to the feedback. Thus, the structures of CNNs can be obtained through the aspect of the HRS images dataset. For the search strategies, including the gradient-based, reinforcement learning, and evolutionary algorithm (EA) [6]. In this paper, the population-based EA based NAS is further discussed for HRS image scene classification. In addition, considering the computational complexity is also an important metric for CNNs, and it is often conflict with the performance accuracy. Thus, in this paper, they are needed to be simultaneously optimized in the framework, the multi-objective optimization method is employed. More importantly, it is possible to automatically perceive and evolve a global neural network to solve local problems. The contributions are summarized as follows:

(1) A framework of AutoML based multi-objective neural evolution for scene classification, the competitive networks can be obtained for different application scenes.

(2) Gene form-based network structure encoding with more flexible search space, which is similar to the natural selection in biological world to produce the better neural architecture for obtaining better scene classification result.

(3) Search strategy for exact solution. The Bayesian Optimization Algorithm (BOA) is employed for sampling the exact solutions in the search history archive.

In the rest of this paper, Section 2 presents the proposed method, the experimental results are provided in Section 3. Finally, the conclusion is drawn in Section 4.

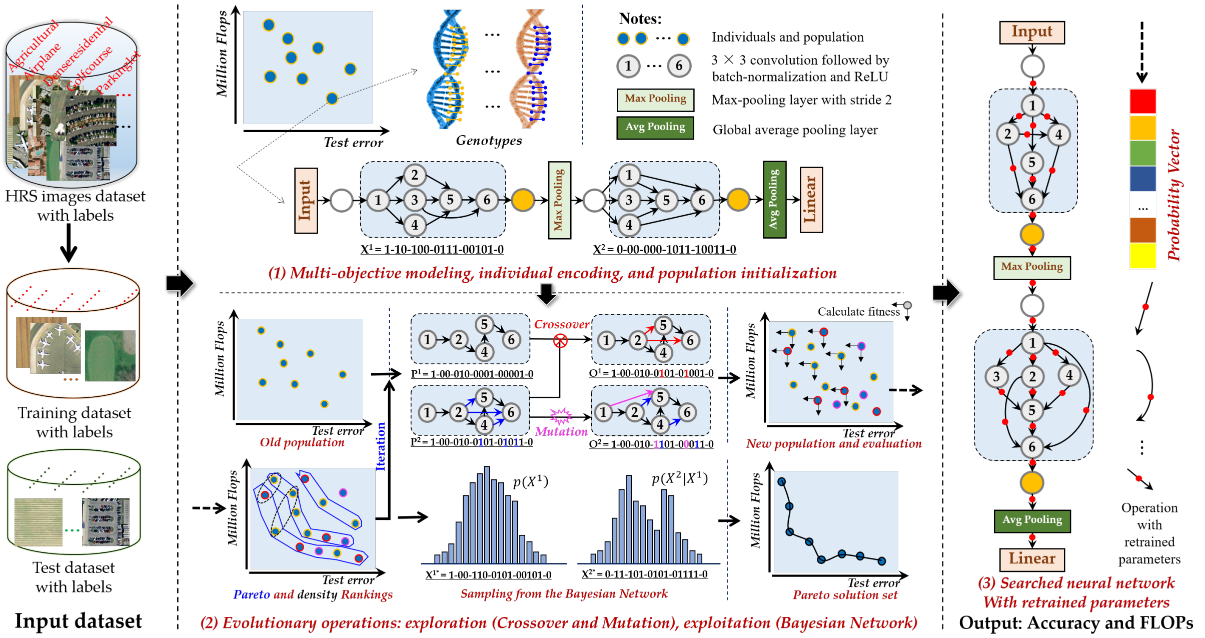

Agricultural
Airplane
Denseresidential
Golfcourse
Parkinglot
HRS images dataset with labels
Training dataset with labels
Test dataset with labels
Input dataset
Million Flops
Test error
Genotypes
Notes:
Individuals and population
3 × 3 convolution followed by batch-normalization and ReLU
Max Pooling
Max-pooling layer with stride 2
Avg Pooling
Global average pooling layer
Input
Linear
X¹ = 1-10-100-0111-00101-0
X² = 0-00-000-1011-10011-0
(1) Multi-objective modeling, individual encoding, and population initialization
Old population
Crossover
Mutation
P¹ = 1-00-010-0001-00001-0
P² = 1-00-010-0101-01011-0
O¹ = 1-00-010-0101-01001-0
O² = 1-00-010-1101-00011-0
Calculate fitness
New population and evaluation
Iteration
Pareto and density Rankings
p(X¹)
p(X²|X¹)
X¹* = 1-00-110-0101-00101-0
X²* = 0-11-101-0101-01111-0
Sampling from the Bayesian Network
Pareto solution set
(2) Evolutionary operations: exploration (Crossover and Mutation), exploitation (Bayesian Network)
Probability Vector
Operation with retrained parameters
(3) Searched neural network With retrained parameters
Output: Accuracy and FLOPs

Fig. 1. The overall framework of the proposed RSSM-Net method.

2. PROPOSED RSSM-NET

In the proposed method, the desirable neural architectures can be obtained by the input dataset and evolutionary search. In addition, the input dataset is divided into the training dataset and test dataset, which are used for training the searched neural net structures and given the test error and values of the computational complexity for multi-objective comparison of the different solutions in a population. The main steps in RSSM-Net are introduced as follows.

2.1. Multi-Objective Modeling and Initialization

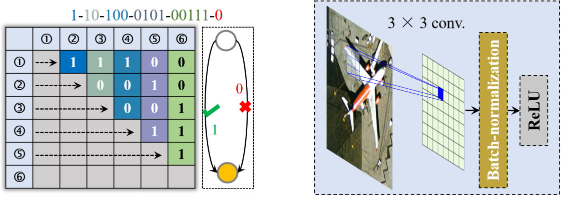
Different to single-objective optimization method, there are two objective functions $\min F(X) = (f_1(X), f_2(X))$ needed to be simultaneously optimized in the proposed RSSM-Net. Moreover, the multi-objective modeling is presented in (1) and (2). For the first objective function, the test error is adopted; for the second objective function, the Float point Operations (FLOPs) is counted and the Giga FLOPs (GFLOPs) = 10^9×FLOPs.

$$f(1) = test\_error = \frac{num\_incorrected\_samples}{num\_all\_test\_samples} \qquad (1)$$

$$\begin{aligned} f(2)_{\text{conv.}} &= FLOPs = 2HW\left(C_{in}K^2+1\right)C_{out} \\ f(2)_{\text{FC.}} &= FLOPs = \left(2I-1\right)O \end{aligned} \qquad (2)$$

where H and W are height and width of the input feature map, the C_{in} is the number of channels of the input feature map, K is the kernel width, and C_{out} is the number of channels of the output feature map when calculating the FLOPs of the convolutional operation; I and O represent the number of input and output channels when calculating the FLOPs of the fully connection layer.

For the initialization step, as shown in Fig. 1-(1), there are eight neural net structures (individuals) in a population; and for each individual, there are several phases with several nodes in the entire structure. Learning from the gene coding on chromosome, the binary encoding is used to encode the connection between nodes, for example, there are six nodes in the first phase, it can be inferred that the X^1 = 1-10-100-0101-00111-0. In addition, each node in a phase carries the same sequence of operations: 3 × 3 convolution followed by batch-normalization and ReLU. As shown in Fig. 2.



	①	②	③	④	⑤	⑥
①	--->	1	1	1	0	0
②	--------	-->	0	0	1	0
③	------------	----	-->	0	0	1
④	----------------	----	----	-->	1	1
⑤	--------------------	----	----	----	-->	1
⑥						

(a) (b)

Fig. 2. Phase coding and node definition. (a) Binary coding for each phase. (b) The sequence of operations for each node.

Based on the initialization, the search space can be obtained:

$$\Omega = n_p \times 2^{n(n-1)/2+1} \qquad (3)$$

where n_p is the number of the phases, n is the number of the nodes in a phase.

2.2. Evolutionary Operations

As shown in the Fig. 1-(2), the exploration with crossover and mutation, the multi-objective evaluation and ranking, and exploitation with Bayesian optimization are included in the evolutionary operations.

2.2.1. Exploration: crossover and mutation

The global search is more reflected in the exploration, imitating the crossover and mutation of the chromosome, individuals and populations can be evolved, and the better neural network structures can be obtained for HRS image scene classification task. For the crossover, two parent individuals are used to produce a new offspring individual, for example, the P^1 = *1-00-010-0001-00001-0* and P^2 = *1-00-010-0101-01011-0*, the crossover offspring O^1 = *1-00-010-0101-01001-0* is then obtained. The mutation means the binary change of a certain gene position, for example, the P^2 is changed to O^2 = *1-00-010-1101-00011-0*.

2.2.2. Multi-objective evaluation and ranking

After the crossover and mutation, a new population with more individuals is obtained. In order to keep the size of the population in a constant, the individuals are needed to be evaluated and ranked. For the multi-objective evaluation, based on the trained network by the HRS image training dataset, the HRS image test dataset is utilized to obtain the classification error and the number of FLOPs that the network carries out during a forward pass, thus the values of objective functions (1) and (2) are counted. In addition, the Pareto ranking (4) and density ranking (5) used in non-dominated sorting genetic algorithm II (NSGA-II) are utilized for individuals ranking. After the ranking, the superior individuals are selected into next iteration.

$$\forall i = 1,2, f_i(I_1) \le f_i(I_2) \wedge \exists i = 1,2, f_i(I_1) < f_i(I_2) \tag{4}$$

$$Density_i = \frac{f_1(I_{i-1}) - f_1(I_{i+1})}{f_1^{\max} - f_1^{\min}} + \frac{f_2(I_{i-1}) - f_2(I_{i+1})}{f_2^{\max} - f_2^{\min}} \tag{5}$$

where I is the individuals in the population, I_1 dominates I_2 in (4), $f_1^{\max}$, $f_1^{\min}$, $f_2^{\max}$, and $f_2^{\min}$ are the maximum and minimum values of the two objective functions.

2.2.3. Exploitation: Bayesian network

After the iteration of the steps 2.2.1-2.2.2, as shown in Fig. 1, the exploitation is conducted after the exploration, the goal of this stage is to exploit and reinforce the patterns commonly shared among the past successful architectures explored in the previous stage. In addition, the BOA has been utilized. For example, there are two phases in a network, which are X^1 and X^2. In the Bayesian network, the distributions $P(X^1)$ and $P(X^2|X^1)$ are estimated through the search history, and the new offspring solutions X^{1*} and X^{2*} are obtained by sampling from this Bayesian network. After this step, the Pareto optimal solution set is obtained.

2.3. Neural Network Retraining and Test

After the evolutionary operations, a competitive HRS image scene classification neural architectures set with non-dominated test error and computational complexity can be obtained for different application scenes. From the Pareto solution set, the network with minimum test error is selected for retraining and testing. As shown in Fig. 1-(3), the searched neural network with retrained parameters is obtained, and the test accuracy and FLOPs are output.

3. EXPERIMENTS AND ANALYSES

In order to verify the effectiveness of the RSSM-Net, it is compared with several major state-of-the-art handcrafted networks, which are AlexNet [4], VGG16 [7], ResNet34 [8], and GoogLeNet [5]. In addition, the well-known UC Merced land-use dataset [9] is utilized, which is composed of 2,100 overhead scene images and divided into 21 land-use scene classes. Each class consists of 100 aerial images measuring 256 × 256 pixels, with a spatial resolution of 0.3 m, some examples are presented in Fig. 3.

For the experimental settings, phase = 3 and node = 8 in a network, population size = 20; the batch size = 32, learning_rate = 0.03 momentum = 0.9, weight_decay = 3e-4, and training epoch = 200 for each searched network in the search stage; and the batch size = 40, a cosine annealing based learning rate that decayed from 0.03 to 0.001, and five-fold cross-validation is utilized while the epoch = 1000.

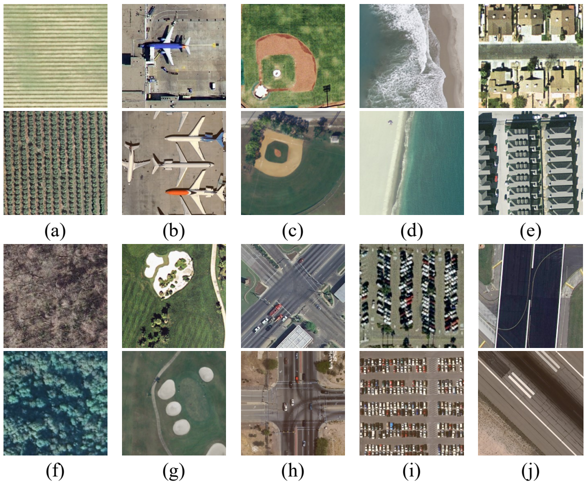


Fig. 3. Examples from the UC Merced land-use dataset: (a) Agricultural. (b) Airplane. (c) Baseballdiamond. (d) Beach. (e) Denseresidential. (f) Forest. (g) Golfcourse. (h) Intersection. (i) Parkinglot. (j) Runway.

The mean accuracy and kappa of five-fold cross-validation for the human expert and evolutionary DL methods is given in Table I, it can be found that the searched network obtains the best performance than other benchmark classification CNNs, which can inferred that the more suitable learning network for the HRS image scene classification can be automatically searched through neural evolution based NAS.

In addition, a NVIDIA RTX 2080 accelerator is used for about 3 days in the search stage. Moreover, the comparison of the efficiency for the human expert and evolutionary searched networks is given in Table II, the corresponding metrics are all tested in a NVIDIA Tesla P100 GPU accelerator. It can be seen that the searched network obtains the least amount of theoretical parameter size and GPU memory occupation.

TABLE I
THE COMPARISON OF THE PERFORMANCE ACCURACY FOR THE HUMAN EXPERT AND EVOLUTIONARY DEEP LEARNING METHODS

Model	Overall Accuracy (%)	Kappa Coefficient	Search Method
AlexNet	90.76 ± 0.72	0.8963 ± 0.0076	*Human experts*
VGG16	96.76 ± 1.50	0.9669 ± 0.0177	*Human experts*
ResNet34	95.48 ± 0.67	0.9524 ± 0.0071	*Human experts*
GoogLeNet	94.19 ± 0.65	0.9419 ± 0.0148	*Human experts*
RSSM-Net	**99.10 ± 0.31**	**0.9905 ± 0.0033**	*Evolution*

TABLE II
THE COMPARISON OF THE EFFICIENCY FOR THE HUMAN EXPERT AND EVOLUTIONARY DEEP LEARNING METHODS

Model	Params (M)	Memory (M)	GFLOPs	Speed (sample/sec)
AlexNet	61.106	1027	0.724	1215
VGG16	138.377	1857	15.45	151
ResNet34	21.259	945	4.15	447
GoogLeNet	6.646	875	1.51	635
RSSM-Net	1.076	526	10.03	298

A Pareto solution set with competitive networks for different custom choices of users between the computation complexity and the accuracy, as shown in Fig. 4. In addition, specific structures of the first two searched networks are given in Fig. 5.

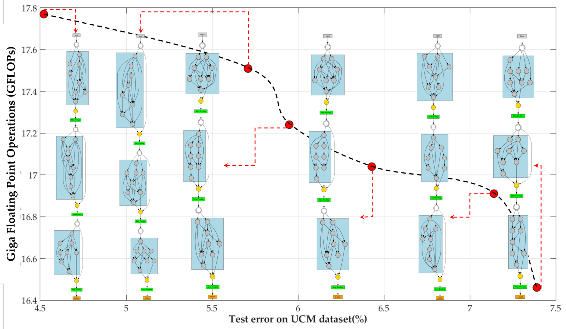
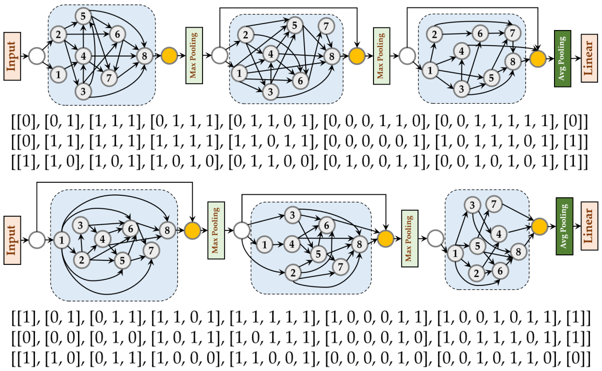


Fig. 4. Several evolutionary searched network on Pareto solution set.



[[0], [0, 1], [1, 1, 1], [0, 1, 1, 1], [0, 1, 1, 0, 1], [0, 0, 0, 1, 1, 0], [0, 0, 1, 1, 1, 1, 1], [0]]
[[0], [1, 1], [1, 1, 1], [1, 1, 1, 1], [1, 1, 0, 1, 1], [0, 0, 0, 0, 0, 1], [1, 0, 1, 1, 1, 0, 1], [1]]
[[1], [1, 0], [1, 0, 1], [1, 0, 1, 0], [0, 1, 1, 0, 0], [0, 1, 0, 0, 1, 1], [0, 0, 1, 0, 1, 0, 1], [1]]

[[1], [0, 1], [0, 1, 1], [1, 1, 0, 1], [1, 1, 1, 1, 1], [1, 0, 0, 0, 1, 1], [1, 0, 0, 1, 0, 1, 1], [1]]
[[0], [0, 0], [0, 1, 0], [1, 0, 1, 1], [1, 0, 1, 1, 1], [1, 0, 0, 0, 1, 0], [1, 0, 1, 1, 1, 0, 1], [1]]
[[1], [1, 0], [0, 1, 1], [1, 0, 0, 0], [1, 1, 0, 0, 1], [0, 0, 0, 0, 1, 0], [0, 0, 1, 0, 1, 1, 0], [0]]

Fig. 5. Specific structure of the two searched neural evolutionary networks.

4. CONCLUSION

A framework of multi-objective neural evolution for HRS image scene classification was proposed, the competitive neural networks can be obtained for different demands. In addition, gene form-based network structure coding with more flexible search space, which is similar to the natural selection in biological world to produce the better neural network for obtaining better HRS image scene classification result. Moreover, the BOA based exploitation has refined the searched networks. From the experimental results, the advantages of the proposed RSSM-Net can be demonstrated. More importantly, based on the remote sensing big data and from the evolutionary perspective, a global and robust deep neural network could be obtained for local solutions.

5. ACKNOWLEDGEMENTS

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFB0504202, in part by the National Natural Science Foundation of China under Grant 41801267, in part by the China Postdoctoral Science Foundation under Grant 2017M622522.

6. REFERENCES

[1] G. Cheng, J. Han, and X. Lu, “Remote sensing image scene classification: Benchmark and state of the art,” *Proceedings of the IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.

[2] Q. Zhu, Y. Zhong, Y. Liu, L. Zhang, and D. Li, “A deep-local-global feature fusion framework for high spatial resolution imagery scene classification,” *Remote Sens.*, vol. 10, no. 4, Article number: 568, Apr. 2018.

[3] Q. Zou, L. Ni, T. Zhang, and Q. Wang, “Deep learning based feature selection for remote sensing scene classification,” *IEEE Geosci. and Remote Sens. Lett.*, vol. 12, no. 11, pp. 2321–2325, Nov. 2015.

[4] 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.

[5] M. Castelluccio, G. Poggi, C. Sansone, and L. Verdoliva, “Land use classification in remote sensing images by convolutional neural networks,” *arXiv preprint arXiv:1508.00092*, 2015.

[6] Z. Lu *et al.*, “NSGA-Net: neural architecture search using multi-objective genetic algorithm. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2019)*, Jul. 2019, pp. 419–427.

[7] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[8] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[9] Y. Yang and S. Newsam, “Bag-of-visual-words and spatial extensions for land-use classification,” in *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM, 2010, pp. 270–279.