

## RESEARCH ARTICLE

# An Ordered Aggregation-Based Ensemble Selection Method of Lightweight Deep Neural Networks With Random Initialization

LIN HE<sup>1,2</sup>, LIJUN PENG<sup>1,3</sup>, AND LILE HE<sup>1</sup><sup>1</sup>School of Mechanical and Electrical Engineering, Xi'an University of Architecture and Technology, Xi'an 710055, China<sup>2</sup>School of Science, Xi'an University of Architecture and Technology, Xi'an 710055, China<sup>3</sup>Engineering Comprehensive Training Center, Xi'an University of Architecture and Technology, Xi'an 710055, China

Corresponding author: Lijun Peng (penglijun683@xauat.edu.cn)

This work was supported in part by the Special Scientific Research Project of Education Department of Shaanxi Provincial Government of China under Grant 21JK0732, in part by the Natural Science Special Project of Xi'an University of Architecture and Technology under Grant ZR19058, and in part by the Key Research and Development Program of Shaanxi Province of China under Grant 2022NY-094.


**ABSTRACT** Due to the popularity of 5G connectivity and The Internet of Things sensors, deep learning algorithms are being extended to edge devices. Compared with AI(Artificial Intelligence) cloud platforms, the deployment of deep neural networks on edge devices must focus on low power consumption, low latency, stability and reliability. In recent years, the development of lightweight deep neural network architecture has provided a basis for the deployment of deep neural networks on edge devices. However, the shortcomings of deep neural networks, such as overconfidence, vulnerability to adversarial attack, and easy over fitting when samples are insufficient, still limit their applications in many fields. One of the ways to compensate for these defects is to use deep ensemble. An ordered aggregation-based ensemble selection algorithm is proposed, which uses soft-margin as the importance assessment metric to take full advantage of the diversity and complementarity of lightweight deep neural networks obtained from different initialization training, so as to improve the overall performance of multiple edge devices. The experimental results show that this algorithm has a significant improvement in generalization performance compared with random ensemble and ordered aggregation algorithms based on accuracy or diversity, and provides a new complementary idea for the deployment of lightweight deep neural networks on edge devices.

**INDEX TERMS** DNN, deep ensemble, selective ensemble, ordered aggregation, soft-margin.

## I. INTRODUCTION

At present, deep learning models with multi-layer processing architectures show better performance than shallow or traditional classification models. With the development of 5G connectivity and the IoT (Internet of Things) sensors, deep learning algorithms are expanding to edge devices, the deployment of DNNs (Deep Neural Networks) on edge devices must focus on low power consumption, low latency, stability and reliability. In recent years, the development of lightweight deep neural network architecture has provided a basis for the deployment of DNNs on edge devices. However, the shortcomings of DNNs, such as overconfidence, vulnera-

bility to adversarial attack, and easy over fitting when samples are insufficient, still limit their applications in many fields. One of the ways to compensate for these defects is to use deep ensemble. Deep ensemble learning combine the advantages of deep learning and ensemble learning to improve generalization performance and robustness by training multiple models and aggregating their predictions. The members of a good ensemble model should be both accurate and error-independent. The loss surfaces of DNNs are non-convex and depend on millions of parameters, and the geometry of these loss surfaces is not well understood. Even for simple networks, the number of local optima and saddle points is large and can grow exponentially in the number of parameters [1], [2]. Moreover, the loss is high along a line segment connecting two optima [3], [4]. These two observations suggest

The associate editor coordinating the review of this manuscript and approving it for publication was Mahdi Zareei .

that the local optima are isolated. Meanwhile, in the process of DNNs training, SGD ( Stochastic Gradient Descent ) [5] and its variant Adam [6] are the most common optimization algorithms. The random noise of mini-batch data sampling in sgd-like algorithms and the random initialization of deep neural networks, coupled with the existence of various local minimum solutions in high-dimensional optimization problems, show that DNNs trained with different random seeds can converge to very different local minima, although they have similar error rates [7], [8], [9]. That is, DNNs trained with different random seeds usually do not produce the same error in the test set, even if the models have converged, they may produce inconsistent predictions given the same input [10]. So theoretically, it is feasible to ensemble DNNs trained with different initialization to improve the prediction performance. Furthermore, single model in an ensemble can be distributed to multiple end devices, which can further speed up inference and potentially simplify the design of specialized hardware.

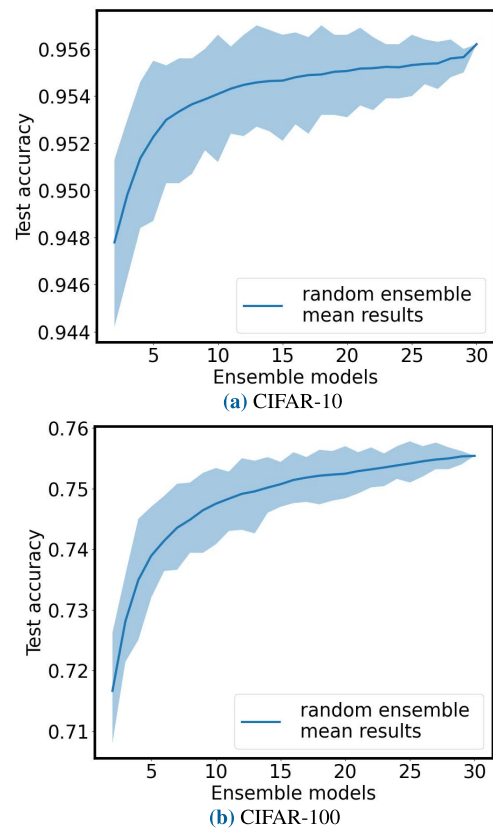
MobileNetV2 is a classic lightweight deep neural network architecture that seeks to perform well on mobile devices. 30 MobileNetV2 models with different initialization were trained on the training set of the CIFAR-10 and CIFAR-100 datasets in this paper. Then 2, 3, 4, . . . 29, 30 models were randomly selected from the 30 models for ensemble. Next, the accuracy of each ensemble model was calculated on the test set. This process was repeated 100 times to get the corresponding test results. As shown in Figure. 1(a) and Figure. 1(b), the solid line represents the mean of 100 random ensemble results at each ensemble scale, and the shaded area represents the variation range of 100 random ensemble results at each ensemble scale. There is still a large room for optimization in the random ensemble results.

In this paper, the selective ensemble method of lightweight DNNs based on ordered aggregation is studied, and an ordered aggregation algorithm based on soft-margin to improve the performance of the ensemble is proposed. The experimental results show that the ensemble model obtained by this algorithm has a significant improvement in generalization performance compared with random ensemble and ordered aggregation algorithms based on accuracy and diversity. It makes the ensemble of multiple lightweight DNNs better at the same computational efficiency. This ordered aggregation-based algorithm is simple to use and does not require architectural tuning, which provides a new idea for the ensemble design and deployment of lightweight DNNs. The main contributions of this study are as follows:

1) The ensemble results of lightweight DNNs obtained by random initialization training under different ensemble scales are analyzed, and it is proposed that the diversity and complementarity brought by random initialization can be fully utilized to optimize the final ensemble performance.

2) According to the margin theory, soft-margin is proposed as one of the importance assessment metric for selecting the base lightweight DNNs.

3) Ordered aggregation of greedy heuristics algorithms based on soft-margin, accuracy and diversity assessment



**FIGURE 1.** Variation range of random ensemble results for MobileNetV2 at different ensemble scales on CIFAR-10/CIFAR-100.

metric are proposed and compared to random ensemble, the algorithm based on soft-margin can get the optimal generalization performance.

The rest of this article is organized as follows. The second section introduces the related work on deep neural network ensemble and selective ensemble methods. In the third section, the ordered aggregation-based ensemble selection method of lightweight DNNs with random initialization is given in detail. In the fourth section, the experiment is carried out and the results of ordered aggregation method based on different metrics are analyzed objectively. The fifth section is the conclusion of this article.

## II. RELATED WORK

### A. DEEP NEURAL NETWORK ENSEMBLE

In implicit ensembles, the parameters of the models are shared, and the single unthinned network at test time approximates the ensemble averaging. However, explicit ensembles do not share model parameters, and the ensemble output is a combination of the predictions of the ensemble models using different approaches, such as majority vote, averaging and so on.

During the training of the network, dropout [11] removes hidden nodes from the network in order to create an ensemble network. During the testing phase, all nodes are active. Using

dropout, the network is regularized to avoid overfitting, and the output vectors become sparse. Dropconnect [12] provides a generalization of dropout. It randomly drops each connection, unlike dropout, which drops output unit. This causes sparsity in the weight parameters of the model. Both dropout and dropconnect require a lot of training time. As a solution, deep networks with stochastic depth [13] were designed to reduce the network depth during training while keeping it unchanged during testing. Stochastic depth is an improvement on ResNet [14], where residual blocks are removed randomly during training and these transformation blocks are bypassed via skip connections. Swapout [15] involves dropping individual units or skipping randomly through blocks, it is a generalization of dropout and stochastic depth.

All the aforementioned methods create an ensemble of networks by sharing the weights. Some researchers have explored explicit ensembles that do not share weights between models. Huang et al. [16] exploits good and bad local minima and let the SGD converge  $M$ -times to local minima along the optimization path, and take the snapshots when the model reaches the minimum, these snapshots are then ensembled by averaging for object recognition. The training time of the ensemble model is the same as that of the single model. The ensemble output is taken as the average of the output of the snapshot models' outputs. Random vector functional link network [17] has also been explored to create explicit ensemble, in which different random initialization of hidden layer weights in the hierarchy makes the ensemble prediction diversified.

Fast Geometric Ensembling (FGE) [18], shows that it is possible to collect models that are spatially close to each other but produce different predictions using cyclic learning rates. They use the collected models to train the ensemble, and there is no computational overhead compared with training a single DNN. An effective method of Bayesian neural network model averaging is also discussed in [19]. SWA [20] is inspired by the development track of FGE scheme. The purpose is to find a single model that can approximate FGE set, but provides stronger interpretability, convenience and scalability during testing.

The above DNN ensemble methods mainly take into account how to reduce training costs, without considering the screening of base models, so they are essentially random ensemble. As can be seen in Fig.1, random ensemble results have a large range of variation, and it is difficult to ensure the optimal results. This paper mainly explores the selective ensemble of different base models under the condition of random initialization, and makes full use of the diversity and complementarity brought by the random initialization of DNNs to obtain better ensemble performance, so that the lightweight DNNs can get better overall performance when deployed on multiple edge devices.

## B. SELECTIVE ENSEMBLE

In order to achieve the ideal generalization performance, the ensemble learning algorithm usually generates a large

number of base models to form an ensemble system. However, it is not that the more base models participating in the ensemble, the better the generalization performance of the ensemble system, including the following reasons: (1) Some generated models may have lower accuracy, and their participation in the ensemble will reduce the generalization performance of the final ensemble system. (2) Some generated models may be similar to each other, that is, they usually give the same results for the same samples, and ensembling some similar models will not improve their generalization performance. (3) Ensembling a large number of models requires a lot of storage and computing overhead and reduces the prediction speed of the ensemble system.

For the above reasons, the base models need to be screened. The base models ensemble selection is to select an approximate optimal subset from the initial base model pool according to some performance evaluation metric, and use the ensemble subset as the final ensemble system. In the past ten years, scholars have carried out a series of research work and proposed many ensemble selection algorithms. In general, these algorithms can be divided into three categories: (1) Ensemble selection based on ordered aggregation [21], [22], [23]; (2) Ensemble selection based on clustering algorithm [24], [25], [26]; (3) Ensemble selection based on optimization [27], [28], [29]. Optimization-based algorithms can usually select an ideal subset of ensembles, but at the cost of significant computational and time overhead, especially for ensembles of DNNs. Therefore, this paper mainly discusses the ensemble selection method based on ordered aggregation.

The ensemble methods in traditional ensemble learning mostly use simple base models like decision trees, and they mostly use majority voting for ensemble. These methods are not fully applicable to DNNs. The complexity of the DNN base model is higher, and the output is in the form of probability, so this work studies ordered aggregation method based on three importance assessment metric and compared to random ensemble, the method based on soft-margin gives the best performance.

## III. ORDERED AGGREGATION-BASED ENSEMBLE SELECTION OF LIGHT DNNs WITH RANDOM INITIALIZATION

### A. ORDERED AGGREGATION AND IMPORTANCE ASSESSMENT

The basic idea of ensemble selection methods based on ordered aggregation is as follows. First, the performance of each base model in the initial ensemble system is evaluated separately according to some importance assessment metrics. Then, the base models are reordered according to the corresponding assessment values obtained to get a new model sequence, in which those base models that are evaluated to have good performance are ranked in the front of the sequence. Finally, some strategy is used to select the top  $k$  ( $0 < k < T$ ) models from the new sequence to form the final ensemble subset.

For an ordered aggregation-based ensemble selection algorithm, the importance assessment metric used by the algorithm determines the performance it can achieve. It is well known that accuracy and diversity are the two most commonly used importance assessment metrics. To measure ensemble diversity, a classical approach is to measure the pairwise similarity/dissimilarity between two learners, and then average all the pairwise measurements for the overall diversity. The representative paired metrics are Disagreement, Q-Statistic, Correlation Coefficient, Kappa-Statistic and Double-Fault etc. Non-pairwise measures try to assess the ensemble diversity directly, rather than by averaging pairwise measurements. The representative unpaired metrics are Kohavi-Wolpert Variance, Interrater agreement, Entropy, Difficulty, Generalized Diversity and Coincident Failure etc. Although there are many diversity metrics, the exact form and measurement of diversity has not been solved, and the optimization of existing diversity metrics cannot guarantee the learner to obtain good generalization performance.

## B. MARGIN THEORY

Schapire et al. [30] introduced the margin-based explanation to AdaBoost, Formally, in the context of binary classification i.e.,  $f(\mathbf{x}) \in (-1, +1)$ , the margin of the classifier  $h$  on the instance  $\mathbf{x}$ , is defined as  $f(\mathbf{x})h(\mathbf{x})$ , the margin of the ensemble  $H(\mathbf{x}) = \sum_{t=1}^T \alpha_t h_t(\mathbf{x})$  is  $f(\mathbf{x})H(\mathbf{x}) = \sum_{t=1}^T \alpha_t f(\mathbf{x})h_t(\mathbf{x})$ , while the normalized margin of the ensemble is

$$f(\mathbf{x})H(\mathbf{x}) = \frac{\sum_{t=1}^T \alpha_t f(\mathbf{x})h_t(\mathbf{x})}{\sum_{t=1}^T \alpha_t} \quad (1)$$

Based on the definition of the margin, Schapire et al [30], Breiman [31], Gao and Zhou [32] successively gave the upper bound of the generalization error of the ensemble model.

*Theorem 1:* Schapire et al. [30] For any  $\delta > 0$  and  $\theta > 0$ , with probability at least  $1 - \theta$  over the random choice of sample  $S$  with size  $m$ , every voting classifier  $f \in C(H)$  satisfies the following bound:

$$\Pr_D [yf(\mathbf{x}) < 0] \leq \Pr_S [yf(\mathbf{x}) \leq \theta] + O \left[ \frac{1}{\sqrt{m}} \left( \frac{\ln m \ln |H|}{\theta^2} + \ln \frac{1}{\delta} \right)^{\frac{1}{2}} \right] \quad (2)$$

*Theorem 2:* Breiman [31] For any  $\delta > 0$ , with probability at least  $1 - \theta$  over the random choice of sample  $S$  with size  $m$ , every voting classifier  $f \in C(H)$  satisfies the following bound:

$$\Pr_D [yf(\mathbf{x}) < 0] \leq R \left( \ln(2m) + \ln \frac{1}{R} + 1 \right) + \frac{1}{m} \ln \frac{|H|}{\delta} \quad (3)$$

where  $\theta = \hat{y}_1 f(\hat{\mathbf{x}}_1) > 4\sqrt{\frac{2}{|H|}}$ ,  $R = \frac{32 \ln 2 |H|}{m \theta^2} \leq 2m$ .

*Theorem 3:* Gao and Zhou [32] For any  $\delta > 0$ , with probability at least  $1 - \delta$  over the random choice of sample  $S$  with size  $m \geq 5$ , every voting classifier  $f \in C(H)$  satisfies

the following bound:

$$\Pr_D [yf(\mathbf{x}) < 0] \leq \frac{2}{m} + \inf_{\theta \in (0, 1]} \left[ \Pr_S [yf(\mathbf{x}) < \theta] + \frac{7\mu + 3\sqrt{3}\mu}{3m} + \sqrt{\frac{3\mu}{m}} \Pr_D [yf(\mathbf{x}) < \theta] \right] \quad (4)$$

where  $\mu = \frac{8}{\theta^2} \ln m \ln(2|H|) + \ln \frac{2|H|}{\delta}$ .

In 2019, A Grønlund, L Kamma et al. published a paper at the NeurIPS conference [33], proving that Gao and Zhou gave almost the tightest upper bound on the generalization error, improving at most one log factor. And this upper bound has been matched with the lower bound, and theoretically impossible to get a better result.

The margin theory is a very effective theoretical tool to analyze the generalization performance of ensemble models. From Eqs.(2)-(4), when other variables are fixed, the larger the margin over the training examples, the better the generalization performance. Therefore, if a base model is more beneficial to increase the margin of the ensemble model on the training samples, then it is more conducive to improve the generalization performance of the ensemble model. Inspired by the theory described above, margin is tried to be used as an important assessment metric for ordered aggregation of individual DNN in this paper, and is compared with accuracy and diversity assessment metrics.

Since the final output of the deep convolutional neural network model is based on the class probability (or confidence). Therefore, the soft voting method is usually used, and the individual classifier  $h_i$  outputs a 1-dimensional vector  $(h_i^1(\mathbf{x}), \dots, h_i^L(\mathbf{x}))^T$  for the example  $\mathbf{x}$ . Among them,  $h_i^j(\mathbf{x}) \in [0, 1]$  can be regarded as the estimated result of the posterior probability  $P(c_j | \mathbf{x})$ . The final output of category  $c_j$  can be written as  $H^j(\mathbf{x}) = \frac{1}{T} \sum_{i=1}^T h_i^j(\mathbf{x})$ .

Assume that  $\mathbf{V} = \{v^{(1)}, v^{(2)}, \dots, v^{(N)}\}$  &  $\mathbf{v} = [v_1^{(i)}, v_2^{(i)}, \dots, v_L^{(i)}]$ ,  $i = 1, 2, \dots, N$  is a set of vectors where  $v_j^{(i)}$  is the predictions probabilities for the  $j$ th label on example  $\mathbf{x}_i$  combined by soft voting, label  $y \in \{1, 2, \dots, L\}$ .

Based on the above theoretical analysis and the ensemble method of DNNs, the soft-margin of the given example  $\mathbf{x}_i$  can be written as  $\frac{1}{M} (v_{y_i}^{(x_i)} - (M - v_{y_i}^{(x_i)})) = \frac{1}{M} (2v_{y_i}^{(i)} - M)$ , soft-margin is in the range  $[-1, 1]$ .

## C. THE ORDERING AGGREGATION OF GREEDY HEURISTICS BASED ON DIFFERENT IMPORTANCE ASSESSMENT METRICS

The selection of an optimal subensemble from a given base model pool is a difficult combinatorial optimization problem. With the limited computational resources, only approximate solutions are accessible for ensembles of realistic size. Reference [22] pointed out that the generalization performance of an ensemble cannot be improved by pruning technology based on individual attributes of the ensemble members, exhaustive search confirms that the greedy ordering heuristics

**TABLE 1.** The ordering aggregation of greedy heuristics algorithm.

**Input:**The trained models  $H = \{h_i | i = 1, 2, \dots, M\}$  and the validation set  $D_{val} = \{(x_i, y_i) | i = 1, 2, \dots, N\}$  for computing the importance assessment

**Output:** The ordered list ES of  $M$  trained models

**Initialize :**A list of vectors

$$V = \{v^{(1)}, v^{(2)}, \dots, v^{(N)} | v^{(i)} = [v_1^{(i)}, v_2^{(i)}, \dots, v_L^{(i)}]$$

$i = 1, 2, \dots, N\}$  where  $v_j^{(i)} = 0$  is initial number of predictions in label  $j$  on the example  $x_i$  in  $D_{val}$ ,  $L$  is the number of class labels, and ES is an empty list.

**Model ordering :**

**For** each  $h_i \in H$  **do**

Compute the the accuracy of  $h_i$  in  $D_{val}$

**End For**

**return** the most accurate model  $h_{best}$

Append  $h_{best}$  to ES and remove  $h_{best}$  from  $H$

Compute all the base models' softmax output on  $D_{val}$ , obtain a 3D tensor  $V$  which is ( $M$ , validation numbers, classes)

**Function** Compute Soft-Margin()

For  $i$  in range (validation numbers)

$y =$  sample  $i$ 's lable

soft-margin = ( $2V[i][y]$  - ensemble model numbers) / ensemble model numbers

**End Function**

**While** ( $H$  is not empty) **do**

**For**  $h_k$  in  $H$

**If** use soft-margin assessment

Ensemble  $h_k$  and the models in ES

Compute the ensemble model's soft margin with Compute-Soft-Margin() on  $D_{val}$

**If** use accuracy assessment

Ensemble  $h_k$  and the models in ES

Compute the ensemble model's accuracy on  $D_{val}$

**If** use diversity assessment

count the samples of  $D_{val}$  that ensemble model (without  $h_k$ ) is wrong but  $h_k$  is right

**end if**

**end For**

**return** the best  $h_k$

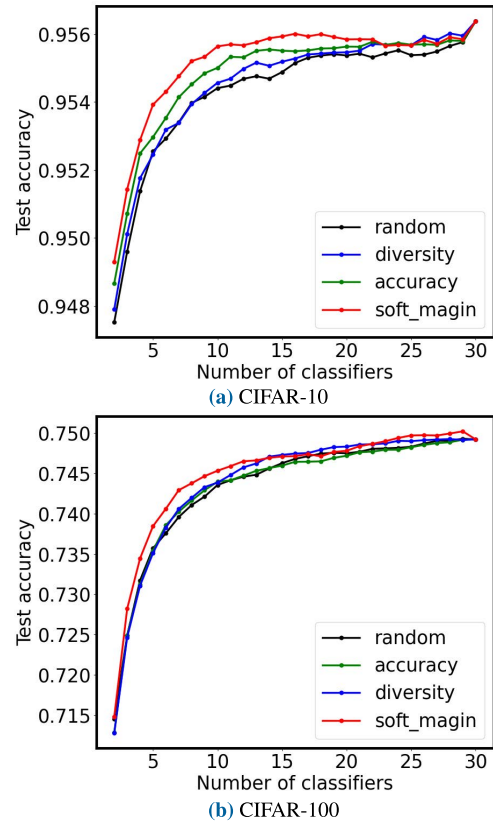
append the best  $h_k$  to ES and remove the best  $h_k$  from  $H$

**end While**

**return** the list ES

devised can efficiently identify near-optimal subensembles of increasing size.

In this paper, the ordered aggregation of lightweight DNNs based on greedy algorithm is studied. First, the most accurate base model in the validation set was selected, and then the base model from the remaining base models was chosen to add to the ensemble in each round, which makes the accuracy, diversity or soft-margin assessment of the ensemble model the best. For details on the ordered aggregation of greedy heuristics algorithm, see Table 1.



**FIGURE 2.** The average accuracy of the random ensemble method, ordered aggregation method based on accuracy, diversity and soft-margin under each ensemble scale on CIFAR-10/CIFAR-100.

Compared with the random ensemble method of DNNs, the ensemble selection method needs to train more base models, but more base models can bring more diversity and complementarity to improve the overall ensemble performance, this work studies three importance assessment metric for lightweight DNNs selection, it is found that methods based on soft-margin proposed by this paper can select better base models than methods based on accuracy and diversity. Compared with random ensemble, it has a significant improvement in generalization performance. It further proves the correctness and validity of the margin theory, and provides a basis for the further research of soft-margin in lightweight DNNs ensemble.

#### IV. EXPERIMENTS AND RESULTS

##### A. EXPERIMENTAL ENVIRONMENT AND CONFIGURATION

In the experiments, the public datasets CIFAR-10 and CIFAR-100 are used to test the proposed method. In order to reduce the size and computation cost of the ensemble model, the classic lightweight DNN MobileNetV2 is used as the base model for training. Four groups of experiments were carried out, as shown in Table 2, each group of experiments was conducted 50 times. In order to reduce the random influence caused by different distribution of validation set and test set, the test set in the original data set is randomly divided into

TABLE 2. Four groups of contrast experiments.

Ensemble method	Operational details	Time consumption (s)
Random ensemble	Randomly select 2, 3, ..., 30 models from 30 base models for ensemble, and calculate the accuracy on the test set of each ensemble scale.	-
Ensemble by accuracy metric	Use the accuracy assessment metric and greedy heuristics algorithm to sort the 30 base models on the validation set, and calculate the accuracy of the ensemble of the first 2, 3, ..., 30 models on the test set.	0.26
Ensemble by diversity metric	Use the diversity assessment metric and greedy heuristics algorithm to sort the 30 base models on the validation set, and calculate the accuracy of the ensemble of the first 2, 3, ..., 30 models on the test set.	36.38
Ensemble by soft-margin metric	Use the soft-margin assessment metric and greedy heuristics algorithm to sort the 30 base models on the validation set, and calculate the accuracy of the ensemble of the first 2, 3, ..., 30 models on the test set.	193.67

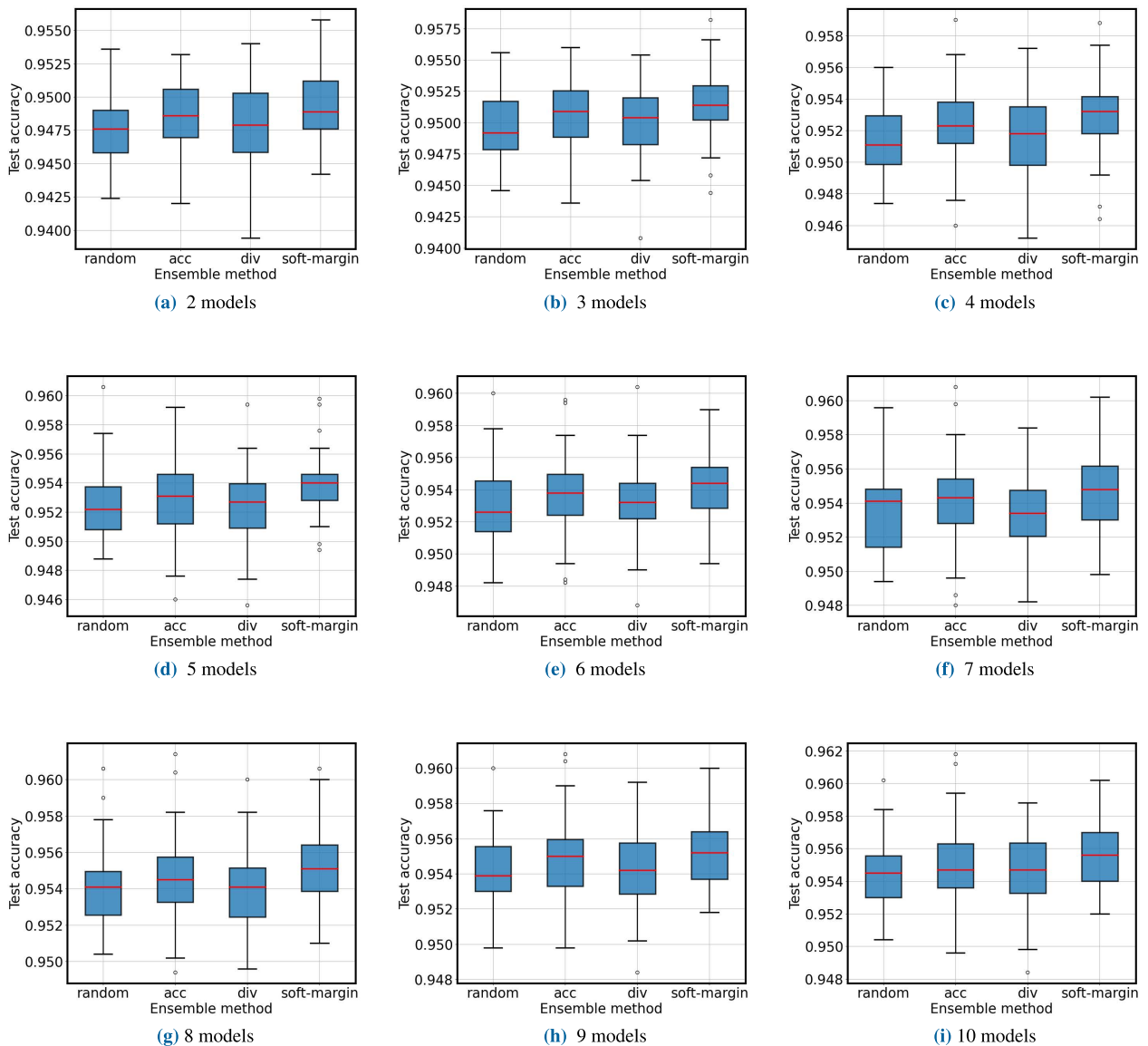
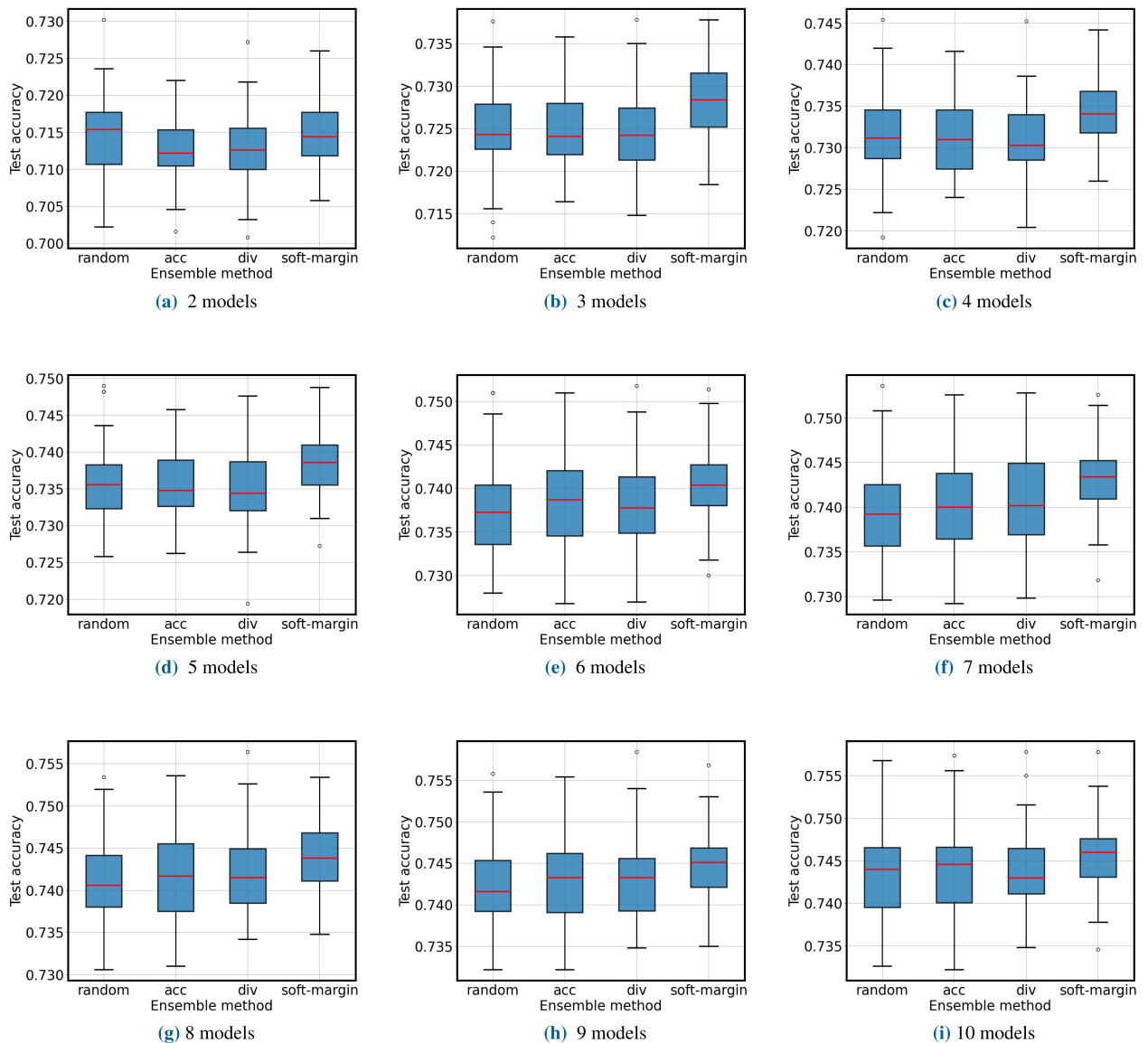


FIGURE 3. The boxplots of the prediction results of the ordered aggregation method based on three different importance assessment metrics and random ensemble method on the CIFAR-10 dataset when the ensemble size is 2-10 models.



**FIGURE 4.** The boxplots of the prediction results of the ordered aggregation method based on three different importance assessment metrics and random ensemble method on the CIFAR-100 dataset when the ensembled size is 2-10 models.

the validation set and the test set according to 1:1, and the number of samples in each category of the validation set and test set is the same, the validation set and test set used were different each time. 50 rounds of calculations are performed and the final results are counted.

### B. ENSEMBLE RESULTS OF ORDERED AGGREGATION METHOD BASED ON DIFFERENT METRICS

The total number of samples in the datasets CIFAR-10 and CIFAR-100 is 60,000. The number of training set samples in the original dataset is 50,000. The training set samples was used to train multiple randomly initialized lightweight DNNs. The purpose of this paper is to make full use of the diversity brought by random initialization, so the samples in the training set did not change in the experiment.

The remaining 10,000 samples were divided into validation set with 5,000 samples and test set with 5,000 samples. Ensemble selection algorithm used validation set to screen base models. Since the validation set has a great impact on the results of base models selection, in order to ensure the reliability of experimental results and the effectiveness of proposed method, in each experiment, the validation set and test set samples were mixed and then redivided randomly according to 1:1 again. 50 experiments were conducted in total, and the validation set and test set in each experiment were different. Very common hyperparameter settings were used to train the models, because compared to the accuracy of single model, the difference in the accuracy of different ensemble models is worthy of more attention. which init lr = 0.1 with warm up and divide by 5 at 60th, 120th, 160th

**TABLE 3. Boxplot statistics of the four algorithms on the CIFAR-10 dataset.**

Ensemble size	Statistics	Random	Accuracy	Diversity	Soft-margin
2	Max	95.36	95.32	95.40	<b>95.58</b>
	Q3	94.90	95.06	95.03	<b>95.12</b>
	Median	94.76	94.86	94.79	<b>94.89</b>
	Mean	94.75	94.87	94.79	<b>94.93</b>
	Q1	94.58	94.69	94.58	<b>94.76</b>
	Min	94.24	94.20	93.94	<b>94.42</b>
3	Max	95.56	95.60	95.54	<b>95.66</b>
	Q3	95.17	95.26	95.20	<b>95.30</b>
	Median	94.92	95.09	95.04	<b>95.14</b>
	Mean	94.96	95.07	95.01	<b>95.14</b>
	Q1	94.78	94.89	94.82	<b>95.02</b>
	Min	94.78	94.35	94.54	<b>94.72</b>
4	Max	95.56	95.68	95.72	<b>95.74</b>
	Q3	95.30	95.38	95.34	<b>95.42</b>
	Median	95.11	95.23	95.18	<b>95.32</b>
	Mean	95.14	95.25	95.18	<b>95.29</b>
	Q1	94.99	95.12	94.98	<b>95.18</b>
	Min	94.74	94.76	94.52	<b>94.92</b>
5	Max	95.74	<b>95.92</b>	95.64	95.64
	Q3	95.38	95.46	95.40	<b>95.46</b>
	Median	95.22	95.31	95.27	<b>95.40</b>
	Mean	95.25	95.30	95.24	<b>95.39</b>
	Q1	94.08	95.12	95.09	<b>95.28</b>
	Min	94.88	94.76	94.74	<b>95.10</b>
6	Max	95.78	95.74	95.74	<b>95.90</b>
	Q3	95.45	95.49	95.44	<b>95.54</b>
	Median	95.26	95.38	95.32	<b>95.44</b>
	Mean	95.29	95.35	95.32	<b>95.43</b>
	Q1	95.14	95.24	95.22	<b>95.28</b>
	Min	94.82	94.94	94.90	<b>94.94</b>
7	Max	95.96	95.80	95.84	<b>96.02</b>
	Q3	95.48	95.54	95.48	<b>95.61</b>
	Median	95.41	95.43	95.34	<b>95.48</b>
	Mean	95.34	95.41	95.34	<b>95.48</b>
	Q1	95.14	95.28	95.21	<b>95.30</b>
	Min	94.94	94.96	94.82	<b>94.98</b>
8	Max	95.78	95.82	95.82	<b>96.00</b>
	Q3	95.49	95.58	95.52	<b>95.64</b>
	Median	95.41	95.45	95.41	<b>95.51</b>
	Mean	95.40	95.45	95.39	<b>95.52</b>
	Q1	95.26	95.32	95.25	<b>95.39</b>
	Min	95.04	95.02	94.96	<b>95.10</b>
9	Max	95.76	95.90	95.92	<b>96.00</b>
	Q3	95.56	95.59	95.58	<b>95.63</b>
	Median	95.39	95.50	95.42	<b>95.52</b>
	Mean	95.42	95.48	95.43	<b>95.53</b>
	Q1	95.30	95.33	95.28	<b>95.37</b>
	Min	94.98	94.98	95.02	<b>95.18</b>
10	Max	95.84	95.94	95.88	<b>96.02</b>
	Q3	95.56	95.63	95.63	<b>95.70</b>
	Median	95.45	95.47	95.47	<b>95.56</b>
	Mean	95.44	95.50	95.46	<b>95.56</b>
	Q1	95.30	95.36	95.32	<b>95.40</b>
	Min	95.04	94.96	94.98	<b>95.20</b>

epochs, trained for 200 epochs with batchsize 256 and weight decay 5e-4, Nesterov momentum of 0.9.

30 MobileNetV2 models with different initialization were trained separately on the training set of CIFAR-10 and CIFAR-100 using the same hyperparameter settings. Then, the algorithm in Table 1 and the random ensemble method were used to get ensemble models of different sizes, and the accuracy of the ensemble model on the test set under each scale was calculated. After that, the validation set and test set were divided randomly again, and the above ensemble selection steps were repeated for 50 times, the results of

**TABLE 4. Boxplot statistics of the four algorithms on the CIFAR-100 dataset.**

Ensemble size	Statistics	Random	Accuracy	Diversity	Soft-margin
2	Max	72.35	72.20	72.18	<b>72.60</b>
	Q3	<b>71.77</b>	71.53	71.56	<b>71.77</b>
	Median	<b>71.54</b>	71.22	71.26	71.44
	Mean	71.46	71.29	71.28	<b>71.48</b>
	Q1	71.06	71.04	71.00	<b>71.18</b>
	Min	70.22	70.46	70.32	<b>70.58</b>
3	Max	73.46	73.58	73.50	<b>73.78</b>
	Q3	72.79	72.79	72.74	<b>73.15</b>
	Median	72.43	72.41	72.42	<b>72.84</b>
	Mean	72.49	72.49	72.47	<b>72.82</b>
	Q1	72.26	72.20	72.13	<b>72.52</b>
	Min	71.55	71.64	71.48	<b>71.84</b>
4	Max	74.19	74.16	73.86	<b>74.41</b>
	Q3	73.45	73.45	73.40	<b>73.68</b>
	Median	73.12	73.10	73.03	<b>73.41</b>
	Mean	73.17	73.12	73.11	<b>73.44</b>
	Q1	72.87	72.74	72.85	<b>73.18</b>
	Min	72.21	72.40	72.04	<b>72.06</b>
5	Max	74.36	754.57	74.76	<b>74.88</b>
	Q3	73.82	73.89	73.87	<b>74.10</b>
	Median	753.56	73.48	73.44	<b>73.86</b>
	Mean	73.57	73.54	73.51	<b>73.84</b>
	Q1	73.23	73.26	73.21	<b>73.55</b>
	Min	72.57	72.62	72.63	<b>73.10</b>
6	Max	74.86	<b>75.10</b>	74.87	74.98
	Q3	74.04	74.20	74.13	<b>74.28</b>
	Median	73.73	73.87	73.78	<b>74.03</b>
	Mean	73.76	73.86	73.83	<b>74.06</b>
	Q1	73.36	73.45	73.49	<b>73.80</b>
	Min	72.80	72.68	72.69	<b>73.18</b>
7	Max	75.10	<b>75.26</b>	75.28	75.14
	Q3	74.26	74.38	74.49	<b>74.53</b>
	Median	73.92	74.00	74.02	<b>74.34</b>
	Mean	73.96	74.02	74.06	<b>74.29</b>
	Q1	73.57	73.65	73.69	<b>74.09</b>
	Min	72.96	72.92	72.98	<b>73.58</b>
8	Max	75.20	<b>75.35</b>	75.26	75.34
	Q3	74.42	74.55	74.49	<b>74.68</b>
	Median	74.06	74.17	74.15	<b>74.38</b>
	Mean	74.11	74.16	74.19	<b>74.37</b>
	Q1	73.80	73.75	73.85	<b>74.11</b>
	Min	73.06	73.10	73.42	<b>73.48</b>
9	Max	75.35	<b>75.54</b>	75.40	75.30
	Q3	74.54	74.62	74.56	<b>74.68</b>
	Median	74.16	74.33	74.33	<b>74.51</b>
	Mean	74.21	74.29	74.33	<b>74.46</b>
	Q1	73.92	73.91	73.93	<b>74.21</b>
	Min	73.22	73.22	73.48	<b>73.50</b>
10	Max	<b>75.68</b>	75.56	75.15	75.38
	Q3	74.65	74.66	74.64	<b>74.76</b>
	Median	74.40	74.46	74.30	<b>74.60</b>
	Mean	74.35	74.39	74.38	<b>74.53</b>
	Q1	74.35	74.01	74.11	<b>74.31</b>
	Min	73.26	73.22	73.48	<b>73.77</b>

50 ensembles under each scale were counted. Figure. 2(a) and Figure. 2(b) show the average accuracy of the random ensemble method, ordered aggregation method based on accuracy, diversity and soft-margin under each ensemble scale. The results show that, for CIFAR-10, since the accuracy of single base model is already relatively high, the diversity among the basic models becomes smaller. Fig. 2(a) shows that the base models screening method using diversity is not as good as the method using accuracy. For CIFAR-100, because the accuracy of single base model is not high, the diversity among the base models is relatively large. From Figure 2(b), it can



be seen that the method of using diversity to screen the base models is slightly better than using the accuracy. However, the accuracy and diversity assessment metric cannot fully guarantee the overall generalization performance of the ensemble model, so the ensemble performance of these two methods is not significantly improved compared to the random ensemble method. In the scale where a large ensemble gain can be achieved, that is, when the number of base models is about ten or less, The method using soft-margin based on margin theory proposed in this paper can achieve the best performance no matter in the CIFAR-10 dataset with high accuracy of single base model or on the CIFAR-100 dataset with low accuracy of single base model.

In order to compare the pros and cons of different methods more comprehensively, the boxplots of the prediction results of the ordered aggregation method based on accuracy, diversity and soft-margin and the random ensemble method when the ensemble scale is 2-10 were drawn, as shown in Figure. 3 and Figure. 4.

Boxplot is a statistical graph that describes the degree of dispersion of a set of data, which can reflect the overall pros and cons and stability of the effects of different algorithms. It can be seen from Figure. 3 and Figure. 4 that the IQR (interquartile range) differences between the ordered aggregation method and the random ensemble method are not obvious, it shows that the four algorithms have little difference in the degree of dispersion when randomly dividing the validation set and test set. In 50 runs, when the number of ensembled base models is 2-10, for CIFAR-10 dataset, the maximum value, upper quartile (Q3), median, lower quartile (Q1) and minimum value of the ordered aggregation algorithm based on soft-margin are greater than the other three algorithms except for the maximum value in the ensemble of 5 models; for CIFAR-100 dataset, the maximum, Q3, median, Q1 and minimum value of the ordered aggregation algorithm based on soft-margin are greater than the other three algorithms except for the median value in the ensemble of 2 models and the maximum value in the ensemble of 6-10 models. It shows that under the ensemble of the same scale, the ordered aggregation algorithm based on soft-margin can select the subensemble model with the best generalization performance. The specific statistics are shown in Table 3 and Table 4 where the best indicators are deepened in bold.

## V. CONCLUSION

This paper proposes a soft-margin based selective ensemble method for lightweight DNNs. The ensemble selection strategy of this work argues that the soft-margin of the ensemble model on the validation set is effective in building an ensemble model with stronger generalization ability. First, select the most accurate base model on the validation set, and then use the ordered aggregation of greedy heuristics algorithm to sequentially add base model which can maximize the soft-margin of the ensemble to form a new larger ensemble model. The method is compared with the ensemble selection method based on accuracy and diversity and the random

ensemble method. The experimental results on the CIFAR-10 and CIFAR-100 datasets show that before the model ensemble gains tends to converge, the soft-margin based lightweight DNNs ensemble selection method can achieve significantly the best generalization results compared to the other three methods, no matter the base model has a high or low accuracy.

Compared with the random ensemble method of deep neural network, the method in this paper needs to train more base models, and these base models will consume more training resources. But more base models can bring more diversity and complementarity to improve the overall ensemble performance. Therefore, the method in this paper is more optimal when it comes to ensemble selection of lightweight models, but for large models with very high training costs, the excessive training cost may not be suitable for the method in this paper.

The next step, the research team intends to further explore how to associate soft-margin with diversity to improve the generalization performance of lightweight deep ensemble models. Based on the analysis and experiments in this paper, the proper combination of soft-margin and diversity can achieve good results, which may provide new research ideas for integration selection.

## REFERENCES

- [1] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio, "Identifying and attacking the saddle point problem in high-dimensional non-convex optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.
- [2] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun, "The loss surfaces of multilayer networks," in *Artificial Intelligence and Statistics*. New York, NY, USA: PMLR, 2015, pp. 192–204.
- [3] I. J. Goodfellow, O. Vinyals, and A. M. Saxe, "Qualitatively characterizing neural network optimization problems," 2014, *arXiv:1412.6544*.
- [4] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," 2016, *arXiv:1609.04836*.
- [5] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. COMPSTAT*. Paris, France: Springer, 2010, pp. 177–186.
- [6] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
- [7] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points—online stochastic gradient for tensor decomposition," in *Proc. Conf. Learn. Theory*, 2015, pp. 797–842.
- [8] K. Kawaguchi, "Deep learning without poor local minima," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.
- [9] Y. Wen, K. Luk, M. Gazeau, G. Zhang, H. Chan, and J. Ba, "An empirical study of large-batch stochastic gradient descent with structured covariance noise," 2019, *arXiv:1902.08234*.
- [10] S. Fort, H. Hu, and B. Lakshminarayanan, "Deep ensembles: A loss landscape perspective," 2019, *arXiv:1912.02757*.
- [11] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014.
- [12] L. Wan, M. Zeiler, S. Zhang, Y. Le Cun, and R. Fergus, "Regularization of neural networks using dropconnect," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1058–1066.
- [13] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, "Deep networks with stochastic depth," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 646–661.
- [14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

- [15] S. Singh, D. Hoiem, and D. Forsyth, "Swapout: Learning an ensemble of deep architectures," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.
- [16] G. Huang, Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, and K. Q. Weinberger, "Snapshot ensembles: Train 1, get M for free," 2017, *arXiv:1704.00109*.
- [17] R. Katuwal, P. N. Suganthan, and M. Tanveer, "Random vector functional link neural network based ensemble deep learning," 2019, *arXiv:1907.00350*.
- [18] T. Garipov, P. Izmailov, D. Podoprikin, D. P. Vetrov, and A. G. Wilson, "Loss surfaces, mode connectivity, and fast ensembling of DNNs," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–10.
- [19] K. Neklyudov, D. Molchanov, A. Ashukha, and D. Vetrov, "Variance networks: When expectation does not meet your expectations," 2018, *arXiv:1803.03764*.
- [20] P. Izmailov, D. Podoprikin, T. Garipov, D. Vetrov, and A. G. Wilson, "Averaging weights leads to wider optima and better generalization," 2018, *arXiv:1803.05407*.
- [21] Q. Dai, T. Zhang, and N. Liu, "A new reverse reduce-error ensemble pruning algorithm," *Appl. Soft Comput.*, vol. 28, pp. 237–249, Mar. 2015.
- [22] G. Martínez-Muñoz, D. Hernández-Lobato, and A. Suárez, "An analysis of ensemble pruning techniques based on ordered aggregation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 245–259, Feb. 2009.
- [23] L. Guo and S. Boukir, "Margin-based ordered aggregation for ensemble pruning," *Pattern Recognit. Lett.*, vol. 34, no. 6, pp. 603–609, Apr. 2013.
- [24] A. Jurek, Y. Bi, S. Wu, and C. D. Nugent, "Clustering-based ensembles as an alternative to stacking," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 9, pp. 2120–2137, Sep. 2014.
- [25] C. Lin, W. Chen, C. Qiu, Y. Wu, S. Krishnan, and Q. Zou, "LibD3C: Ensemble classifiers with a clustering and dynamic selection strategy," *Neurocomputing*, vol. 123, pp. 424–435, Jan. 2014.
- [26] H. Zhang and L. Cao, "A spectral clustering based ensemble pruning approach," *Neurocomputing*, vol. 139, pp. 289–297, Sep. 2014.
- [27] L. Li, R. Stolkin, L. Jiao, F. Liu, and S. Wang, "A compressed sensing approach for efficient ensemble learning," *Pattern Recognit.*, vol. 47, no. 10, pp. 3451–3465, Oct. 2014.
- [28] Z.-H. Zhou, J. Wu, and W. Tang, "Ensembling neural networks: Many could be better than all," *Artif. Intell.*, vol. 137, nos. 1–2, pp. 239–263, May 2002.
- [29] I. Partalas, G. Tsoumakas, and I. Vlahavas, "An ensemble uncertainty aware measure for directed Hill climbing ensemble pruning," *Mach. Learn.*, vol. 81, no. 3, pp. 257–282, 2010.
- [30] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee, "Boosting the margin: A new explanation for the effectiveness of voting methods," *Ann. Statist.*, vol. 26, no. 5, pp. 1651–1686, 1998.
- [31] L. Breiman, "Prediction games and arcing algorithms," *Neural Comput.*, vol. 11, no. 7, pp. 1493–1517, Oct. 1999.
- [32] W. Gao and Z.-H. Zhou, "On the doubt about margin explanation of boosting," *Artif. Intell.*, vol. 203, no. 5, pp. 1–18, 2013.
- [33] A. Grönlund, L. Kamma, K. Green Larsen, A. Mathiasen, and J. Nelson, "Margin-based generalization lower bounds for boosted classifiers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–10.



**LIN HE** received the B.S. degree in electrical engineering from the Xi'an University of Architecture and Technology, Xi'an, China, in 2008, and the M.S. degree in instrument and meter engineering from Northwestern Polytechnical University, Xi'an, in 2013. He is currently pursuing the Ph.D. degree with the Xi'an University of Architecture and Technology.

He is a Lecturer with the Xi'an University of Architecture and Technology. His research interests include deep learning theory and its application.



**LIJUN PENG** received the B.S. and M.S. degrees from the Xi'an University of Architecture and Technology, in 2008 and 2011, respectively. She is currently pursuing the Ph.D. degree with the Xi'an University of Architecture and Technology.

She is currently an Intermediate Engineer with the Xi'an University of Architecture and Technology. Her research interests include machine learning and industrial soft sensing technology.



**LILE HE** received the B.S. and M.S. degrees from the Xi'an University of Architecture and Technology, in 1985 and 1995, respectively, and the Ph.D. degree from the Xi'an University of Technology, in 2006.

He is currently a Full Professor with the Xi'an University of Architecture and Technology. His main research interests include electromechanical system monitoring, model predictive control, and intelligent robot technology.

• • •