Boosting Deep Ensemble Performance with Hierarchical Pruning

Yanzhao Wu, Ling Liu
School of Computer Science
Georgia Institute of Technology
Atlanta, Georgia 30332
Email: yanzhaowu@gatech.edu, lingliu@cc.gatech.edu

***Abstract*—Deep neural network ensembles have become attractive learning techniques with better generalizability over individual models. Some mission critical applications may require a large number of deep neural networks to achieve desirable accuracy and generalizability, making the ensemble execution costly with respect to runtime and space. This paper proposes a novel hierarchical ensemble pruning approach, which can effectively examine a given pool of M base models and identify smaller high quality deep ensembles of size S ($\ll M$) with higher ensemble accuracy than the entire deep ensemble of all M models. Our hierarchical pruning approach, coined as HQ, combines three novel techniques. First, we show that the focal diversity metrics is innovative and can accurately capture the negative correlation among the member models of an ensemble, and the use of focal diversity metrics can boost ensemble accuracy. Second, we introduce a focal-diversity based hierarchical pruning algorithm to progressively identify low-cost ensembles with high ensemble diversity and accuracy. Third, we design a focal diversity consensus method to find smaller deep ensembles with low negative correlation. We demonstrate such ensembles offer high accuracy and high robustness while being more time and space efficient in ensemble decision making. Evaluated using two benchmark datasets, we show that the proposed focal diversity powered hierarchical pruning can find significantly smaller ensembles of deep neural network models while achieving the same or better classification generalizability.**

I. Introduction

Deep neural network ensembles have gained increasing popularity in the deep learning community for improving the generalizability and robustness with the combined wisdom of multiple deep neural networks. Some mission critical applications may require a large number of deep neural network models to achieve a desirable accuracy and robustness, making the ensemble execution both space and time consuming. Several recent works have shown that deep ensembles with high diversity are expected to be more failure independent, which can be critical for improving the overall ensemble accuracy and robustness, including under adverse situations [1], [2]. However, in practice, deep neural network models that are trained with different neural network backbones may not have high ensemble diversity and failure independence [1], [3], [4]. Therefore, given a deep ensemble of a large size such as 10, it is often possible and beneficial to find significantly smaller deep ensembles (e.g., 3 or 4 models) with the same or better classification generalizability as that of the entire deep ensemble [5]–[8]. This motivates us to propose an effective ensemble pruning approach to utilize our enhanced focal diversity metrics to identify small and yet highly diverse deep ensembles that can achieve the same or better ensemble accuracy while improving space and time efficiency of ensemble execution, compared with the large entire ensemble.

Ensemble learning evolves from the three complementary threads of efforts. The first thread of work designs the algorithms for training ensembles using a set of weak models, such as bagging [9], boosting [10] and random forests [11]. It tends to require large ensembles of tens or hundreds of models, represented by the popular random forests in real world deployment. In recent years, a new initiative is to identify high quality deep ensembles from a pool of strong base models with high individual model accuracy and different neural network backbones. These efforts focus on choosing the high quality ensembles that can minimize the correlation of member models on negative samples by using ensemble diversity based methods [1], [3], [7], [8]. This paper contributes to this second thread. The third thread of work centers on voting methods to obtain the ensemble prediction based on member models of an ensemble, such as simple averaging, majority voting, and plurality voting [8], [12]. Prior to 2015, most of the studies [5], [6], [13] have centered on learning and selection of ensembles for traditional machine learning models. Only recently, we have seen several research efforts on deep neural network (DNN) ensembles, most of which focused on training multiple models jointly, such as using diversity via weighted kernels [14]–[16] and employing deep ensembles in real-world applications, such as leveraging deep ensembles to enhance the robustness and resilience of DNNs against adversarial examples [1], [2], [4], [14], [17]. However, very few to the best of our knowledge have put forward solutions to efficient pruning of deep ensembles, aiming to substantially reduce the space and time cost of real time ensemble execution and deploying deep ensembles on edge devices.

In this paper, we present a holistic approach to efficiently pruning deep neural network ensembles. Given a large deep ensemble of M individual DNN models, to find significantly smaller deep ensembles with the same or better ensemble accuracy than the entire deep ensemble of M models, we propose a hierarchical deep ensemble pruning framework, HQ, by combining three novel ensemble selection techniques.

2374-8486/21/$31.00 ©2021 IEEE
DOI 10.1109/ICDM51629.2021.00184

© 2021 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

TABLE I: Example Deep Ensembles for CIFAR-10 and ImageNet

Dataset	CIFAR-10					ImageNet				
Ensemble Team	0123456789	0123	01238	123	1234	0123456789	12345	2345	1234	124
Ensemble Acc (%)	96.33	**97.15**	**96.87**	**96.81**	**96.63**	79.82	**80.77**	**80.70**	**80.29**	**79.84**
Acc Improvement (%)	0	0.82	0.54	0.48	0.30	0	0.95	0.88	0.47	0.02
Team size	10	4	5	3	4	10	5	4	4	3
Cost	100%	**40%**	**50%**	**30%**	**40%**	100%	**50%**	**40%**	**40%**	**30%**

TABLE II: All Individual Member Models for Two Datasets

Dataset	CIFAR-10		ImageNet	
	10,000 testing samples		50,000 testing samples	
Model ID	Name	Acc (%)	Name	Acc (%)
0	**DenseNet190**	**96.68**	AlexNet	56.63
1	DenseNet100	95.46	DenseNet	77.15
2	ResNeXt	96.23	EfficientNet-B0	75.80
3	WRN	96.21	ResNeXt50	77.40
4	VGG19	93.34	Inception3	77.25
5	ResNet20	91.73	**ResNet152**	**78.25**
6	ResNet32	92.63	ResNet18	69.64
7	ResNet44	93.10	SqueezeNet	58.00
8	ResNet56	93.39	VGG16	71.63
9	ResNet110	93.68	VGG19-BN	74.22

First, we use focal diversity metrics [7] to capture the negative correlation among member models of a deep ensemble. The lower focal diversity score indicates the lower level of negative correlation among member models of an ensemble. This is the first innovative method to compare different deep ensembles based on their focal diversity scores and choose the high diversity ensemble that has the low level of negative correlation. Second, we introduce a novel focal-diversity based hierarchical pruning method, which iteratively identifies and prunes out subsets of member models from the entire deep ensemble team, which tend to make similar prediction errors and display high negative correlation. Third, we leverage multiple competing focal diversity metrics to design diversity consensus voting based focal diversity pruning. This enables us to further refine the ensemble recommendations from hierarchical pruning. Comprehensive experiments are conducted on two benchmark datasets: CIFAR-10 [18] and ImageNet [19]. The results show that our hierarchical pruning approach is effective in finding significantly smaller deep ensembles with the same or better accuracy than the entire deep ensemble.

II. Hierarchical Pruning with Focal Diversity

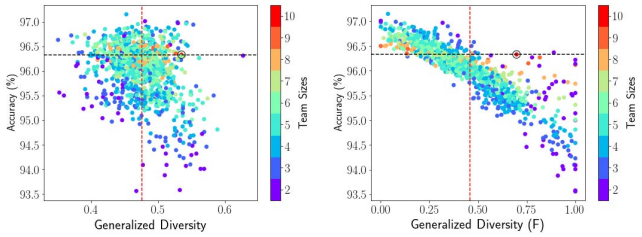
Consider an entire large deep ensemble of M individual member models ($F_i, i \in \{0, 1, ..., M-1\}$) for a learning task on a given dataset, denoted by $F_0F_1...F_{M-1}$. Let $EnsSet(F_0F_1...F_{M-1})$ ($EnsSet$ for short) denote the set of all candidate deep ensembles of size S composed from any proper subset of the given M individual models ($S = 2, \ldots, M-1$). Hence, we have a total of $\sum_{S=2}^{M-1} \binom{M}{S}$ candidate deep ensembles, i.e., $|EnsSet| = \binom{M}{2} + \binom{M}{3} + ... + \binom{M}{M-1} = 2^M - (2+M)$. The cardinality of $EnsSet$ grows exponentially with M. For $M = 3$, we have $|EnsSet| = 3$. For $M = 5, 10, 20$, $|EnsSet| = 25, 1012, 1048554$ respectively. When M is large, an exhaustive search of all candidate deep ensembles in $EnsSet$ may not be feasible. Therefore, efficiently pruning the entire deep ensemble of M individual models can be characterized as the problem of pruning duplicate or low diversity candidate ensembles from the candidate deep ensemble set $EnsSet$ and obtaining the set $GEnsSet$ of good deep ensembles with high ensemble diversity.

Table I lists 5 example deep ensembles for CIFAR-10 and ImageNet respectively, including the entire deep ensemble of 10 models and 4 smaller deep ensembles recommended by our hierarchical ensemble pruning approach. This set of experiments is performed by using the 10 pre-trained models for each of the two datasets given in Table II. For both CIFAR-10 and ImageNet, comparing the 10 model ensemble team, the four deep ensembles recommended by our hierarchical pruning approach are much smaller with 3, 4 or 5 individual member models, while offering better ensemble accuracy than the given large size entire deep ensemble of 10 models. The reduction in the entire ensemble execution cost is about 50%$\sim$70% of the execution cost for the given 10-model ensemble in both CIFAR-10 and ImageNet respectively. Consider the given ensemble team of size $M = 10$, there will be a total of 1012 candidate deep ensembles to be examined for pruning. How should we design our hierarchical pruning approach for efficiently identifying and selecting such time and space efficient small size ensembles with the same or better ensemble accuracy, which can notably reduce both runtime and space cost for ensemble execution?

Problems with Baseline Ensemble Pruning. Recall our discussion in related work, several recent studies have leveraged deep ensembles to improve the robustness of DNN models against adversarial examples [1], [2], [14], [17]. Most of these approaches utilize the Cohen's Kappa (CK) [20] because early studies [21], [22] show that both pairwise metrics such as Cohen's Kappa (CK) [20], Binary Disagreement (BD) [23], and non-pairwise metrics, such as Kohavi-Wolpert Variance (KW) [21], [24], Generalized Diversity (GD) [25], share similar evaluation results with respect to ensemble accuracy. In this section, we argue that all above diversity measures are originally designed for comparing ensembles instead of selecting ensembles. Hence, they tend to fail when used to select high quality small size ensembles defined by ensemble accuracy and ensemble execution efficiency with respect to runtime and space cost of ensemble execution.

We first use the naïve diversity as the baseline ensemble pruning method and analyze its inherent problems. Given a diversity metric, such as BD or GD, the baseline diversity-based pruning method is to calculate the diversity score for each deep ensemble team in the set ($EnsSet$) of all candidate ensembles using a set of negative samples ($NegSampSet$) on which any one or more models make prediction errors ($\bigcup_{i=0}^{M-1} NegSampSet(F_i)$) following [21]. A baseline method for pruning a given entire deep ensemble of M individual models can be performed in three steps: (1) For every

possible candidate deep ensemble in $EnsSet$, we compute the ensemble diversity score using CK, BD, KW or GD, for example; (2) we compute the mean diversity threshold; and (3) we select those ensembles in $EnsSet$ with their diversity scores below the threshold (high ensemble diversity) and place them into the good deep ensemble set $GEnsSet$, and the remaining ensembles with their diversity scores higher than the threshold (low ensemble diversity) will be pruned out.



(a) **GD**, Pruning 427 out of 1012 (b) **F-GD**, Pruning 517 out of 1012

Fig. 1: Pruning all ensembles by mean threshold on CIFAR-10: (a) Naïve Diversity (GD), (b) Focal Diversity (F-GD)

Figure 1a shows the relationship between the naïve GD scores and ensemble accuracy for all 1013 $(1012 + 1)$ deep ensembles on CIFAR-10, including the entire deep ensemble. Each dot represents a deep ensemble team with team sizes color-coded by the color diagram on the right. The vertical red dashed line represents the GD mean thresholds of 0.476. We can visually see that those deep ensembles on the left of the vertical red line will be selected by the baseline GD diversity pruning and placed into $GEnsSet$. The horizontal black dashed line represents the ensemble accuracy 96.33% of the entire deep ensemble. The given entire deep ensemble of $M = 10$ models for CIFAR-10 is marked by the black circle on this horizontal line. For a given large ensemble of M models, we evaluate the performance of each deep ensemble pruning algorithm in terms of four metrics: (1) The ensemble accuracy range for selected deep ensembles in $GEnsSet$; (2) the precision, measured by the ratio of the number of selected ensembles whose ensemble accuracy is equal to or higher than the accuracy of the given entire ensemble (target accuracy, which is 96.33% for the ensemble of 10 models on CIFAR-10) over the total number of all selected ensembles; (3) the recall, measured by the ratio of the number of selected ensembles in $GEnsSet$ over the total number of the candidate ensembles in $EnsSet$, whose ensemble accuracy are equal to or higher than the target accuracy of the entire ensemble; and (4) the cost, measured by the reduction of the ensemble size against the entire ensemble size M. Recall Table I on example ensembles for CIFAR-10, the smaller ensemble of 123 has the cost of 30%, computed by $3/M$, where $M = 10$. In Table III, we show that BD-based baseline pruning in the 2nd column suffers from very low precision of 6.18% and low recall of 11.72%. Although GD-based baseline pruning in the 4th column has the precision of 36.90%, it is still lower than the acceptable ratio of over 50%. Furthermore, both of the baseline diversity pruning algorithms cannot ensure that all selected ensembles have the same or better ensemble accuracy than the entire deep ensemble (96.33% accuracy) given the lower accuracy bound of 93.56%. Similar observations are observed on other diversity metrics, i.e., CK and KW. This motivates us to re-examine the way that ensemble diversity should be measured and to improve the baseline mean threshold based pruning.

TABLE III: Prune ensembles by mean threshold (CIFAR-10)

Methods	BD<0.661 (baseline)	F-BD<0.398 (optimized)	GD<0.476 (baseline)	F-GD<0.457 (baseline)
Acc Range (%) for $GEnsSet$	93.56∼96.72	**95.71∼97.15**	93.56∼97.15	**95.71∼97.15**
Precision (%)	6.18	**63.53**	36.90	**53.90**
Recall (%)	11.72	**95.51**	63.10	**97.59**

Focal Diversity based Hierarchical Pruning. In this study, we leverage three novel techniques to optimize ensemble diversity measurement and improve ensemble pruning results. First, focal model enhanced diversity metrics are used to improve the ensemble diversity measurement by leveraging the concept of focal models to precisely capture the failure independence of member models in a deep ensemble team. Second, a novel focal diversity based hierarchical ensemble pruning algorithm is proposed to progressively identify and prune out subsets of redundant member models with high failure dependency from the entire deep ensemble. Third, the diversity consensus voting is used to combine multiple focal diversity metrics to further improve our focal diversity based hierarchical pruning.

III. Focal Diversity Based Hierarchical Pruning

Focal Diversity Concept. We first introduce the concept of focal model enhanced diversity measure and define focal model enhanced pairwise and non-pairwise diversity metrics, coined as F-CK, F-BD, F-KW and F-GD. The design of our focal diversity metrics aims to more accurately capture the failure independence based diversity of an ensemble of S member models for classification. Concretely, given an ensemble of size S, we use each of the S member models as the focal model to collect negative samples and compute focal-model based diversity score. Unlike conventional approaches to evaluate ensembles, which randomly draw negative samples from any one or more models ($NegSampSet$) for evaluating an ensemble of S member models, such as the naïve diversity metrics, we randomly select negative samples from a specific focal model ($NegSampSet(F_f)$) and then calculate the focal model based diversity score. For an ensemble of S models, we will have S focal model based diversity scores corresponding to this ensemble. We then combine these S focal diversity scores by taking the average as the final focal diversity score of this ensemble of size S. In the context of adversarial robustness with ensemble defense [2], [17], our focal diversity metrics can be viewed as taking the victim model or the attack target model as the focal model for evaluating failure independence of the defense ensemble using negative samples. Our preliminary results have shown some very promising results [7], [8].

Figure 1b shows a visualization comparison of using the focal model enhanced F-GD metric compared to Figure 1a using the naïve GD metric. It is visually clear that F-GD based

pruning even with mean threshold shows the pruning efficiency improvement over the baseline GD based pruning, identifying larger percentage of smaller ensembles with ensemble accuracy above the target accuracy 96.33%.

Table III shows the comparison of focal diversity based pruning algorithms F-BD and F-GD in the 3rd and 5th columns respectively using our pruning efficiency evaluation metrics. We observe that both F-BD and F-GD significantly improve the baseline BD and GD pruning in terms of the accuracy range of the selected deep ensembles (in $GEnsSet$), the precision and the recall. Similar improvements are observed for the F-CK and F-KW based pruning algorithms as well. Although our focal model enhanced diversity pruning has very high recall, over 95% for both F-BD and F-GD, their precision of 63.53% for F-BD and 53.90% for F-GD can still be further improved for achieving very high ensemble pruning efficiency. This motivates us to explore new solutions to replace the mean diversity threshold based pruning.

Hierarchical Pruning Overview. Given a focal diversity measure, say F-GD, it exhibits some anti-monotonicity property. Concretely, if an ensemble of S member models has a large F-GD diversity score (say $[F_5, F_6]$), it often indicates insufficient ensemble diversity, high correlation for prediction errors, thus redundant for the entire ensemble. Therefore, those ensembles that have a larger size than S and contain this small ensemble of low diversity (e.g., $[F_0, F_5, F_6]$, $[F_5, F_6, F_7]$, $[F_0, F_5, F_6, F_7]$, and $[F_0, F_5, F_6, F_8]$) tend to have insufficient ensemble diversity (i.e., larger F-GD diversity scores) compared to those deep ensembles of the same size with other highly diverse member models. This motivates us to design a hierarchical focal diversity based pruning algorithm. It is an iterative process of composing deep ensemble teams for a given desired team size S_d. First, we start with the set of ensembles of two individual models, say $S = 2$, $|EnsSet(S = 2)| = \binom{M}{2} = M(M-1)/2$ candidate ensembles. For $M = 10$ we will have 45 ensembles of size 2. Given a focal diversity metric, we first sort the ensembles of small size S, say $S = 2$, by their focal diversity scores in a decreasing order, and then choose the top β (percentage) of ensembles of size S with large diversity scores as our pruning targets for ensembles of size S. β can be dynamically set to adapt to the specific number of candidate ensembles and diversity measurements. In general, we recommend a conservative approach to setting a small β (e.g., $\beta = 10\%$ by default) to achieve high precision and good recall. We first preemptively prune out the $\beta(\%)$ of the ensembles with large diversity scores (i.e., low ensemble diversity) and then prune all subsequent ensembles that are super-sets of these $\beta(\%)$ of pruned ensembles.

Figure 2 shows a hierarchical structure with all M member models on the top, followed by all ensembles of the smallest size 2, and each tier we add one additional model to the ensemble teams such that all teams of size $S+1$ are placed in the next tier. This way of constructing deep ensembles enables us to efficiently form high quality deep ensembles step by step and strategically prune out low diversity ensembles. The bottom tier will be the ensemble teams of the desired size S_d. For

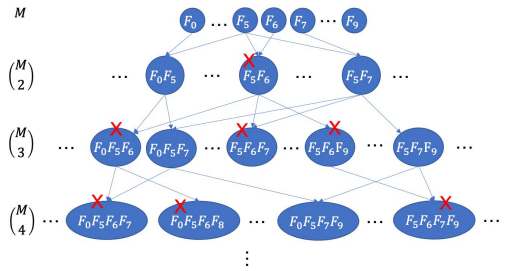
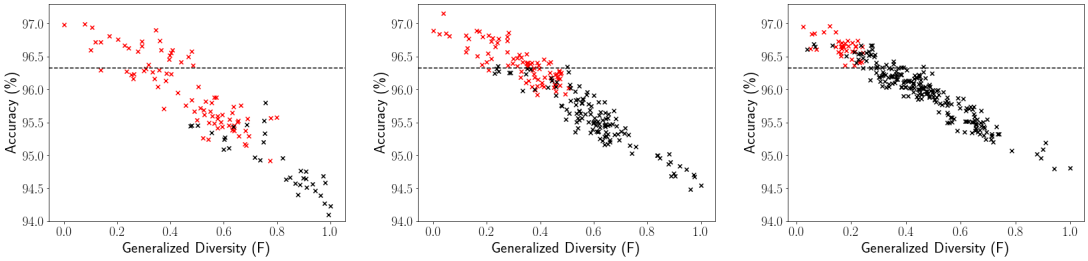


Fig. 2: Hierarchical Pruning

each ensemble being pruned out (e.g., F_5F_6), our hierarchical pruning algorithm will cut off the branches of ensembles that are super-sets of this removed ensemble, e.g., all ensembles containing F_5F_6, marked by red cross in Figure 2. Hence, this algorithm can significantly avoid exploring unpromising branches in searching for high performance deep ensembles.

Figure 3 shows the visualization of applying the F-GD hierarchical pruning algorithms respectively with $\beta = 10\%, S_d = 5$ on CIFAR-10. The black dots denote the ensemble teams pruned out by using the hierarchical pruning and the red dots are the ensembles selected. We highlight three interesting observations. *First,* F-GD diversity metrics display an interesting property: the ensembles with low focal diversity scores tend to give high ensemble accuracy, especially when comparing ensembles of the same sizes, such as $S = 3,\ 4,\ 5$. *Second,* our hierarchical pruning can effectively prune out those ensembles with insufficient diversity (large focal diversity scores). This is because the focal diversity based hierarchical pruning promotes more fair comparison of focal ensemble diversity among ensembles of the same size S. Comparing Figure 3 and Figure 1, the correlation between focal diversity and ensemble accuracy is visually much clearer for the ensembles of a fixed size S in Figure 3. *Third,* the desired ensemble team size S_d can be set to ultimately bound the time and space complexity of ensemble execution cost in our hierarchical pruning algorithm. If the goal of pruning the entire deep ensemble of M individual member models is to obtain significantly smaller ensembles (e.g., one half or one third of the size M), which still provide equal or better ensemble accuracy than the entire ensemble of the M models, we can set the desired ensemble size S_d to be up to $50\% \times M$. *Finally,* we want to note that from Figure 3c, all of the selected ensembles achieve the ensemble accuracy above 96.33%, resulting in the perfect precision of 100% for our focal diversity based hierarchical pruning algorithm with F-GD. Similar observations are found consistently for the other three focal diversity metrics (F-CK, F-BD and F-KW).

Focal Diversity Pruning by Diversity Consensus Voting. Given the set of focal diversity metrics, such as F-CK, F-BD, F-KW, F-GD, the focal diversity pruning using different metrics may recommend different sets of ensembles. Moreover, we observe that those ensembles that are selected by all or a majority of the focal diversity pruning algorithms (e.g., F-CK pruning, F-GD pruning, etc.) tend to have more consistent performance. This motivates us to perform the third step in our



(a) S=3, (Pruning 35 out of 120) (b) S=4, (Pruning 124 out of 210) (c) S=5, (Pruning 215 out of 252)

Fig. 3: Deep Ensembles of size $S = 3, 4, 5$ on CIFAR-10 (**F-GD**, $\beta = 10\%, S_d = 5$)

HQ ensemble pruning framework, which further leverages the diverse strength of these focal diversity metrics for ensemble pruning by consolidating the ensembles recommended by different focal diversity pruning algorithms using a majority voting scheme: an ensemble of size S_d is added to the final selection if it is selected by at least three focal diversity metrics via hierarchical pruning. This third step further improves the efficiency of our focal diversity based hierarchical ensemble pruning in terms of ensemble accuracy and robustness as well as ensemble execution efficiency (space and time).

IV. Experimental Evaluation

Extensive experiments on two benchmark datasets (CIFAR-10 and ImageNet) are conducted with the given entire ensemble of 10 individual member models for each dataset (see Table II). All experiments were conducted on an Intel Xeon E5-1620 server with Nvidia GTX 1080Ti on Ubuntu 16.04.

Efficiency of Pruning with Varying β. Our hierarchical pruning algorithm includes a hyperparameter β to determine the percentage of ensembles to be pruned out for each ensemble size S. Intuitively, a higher β will prune out more ensembles, and hence potentially result in lowering the recall score. The high precision score indicates that the pruning algorithm can correctly and effectively identify the small ensembles with the same or better ensemble accuracy than the target accuracy of the entire deep ensemble. Figure 4 shows the impact of varying β from 5% to 35% on the pruned ensembles of CIFAR-10 and ImageNet respectively, with the desired ensemble size $S_d = 4$ using the F-GD hierarchical pruning. It shows indeed that as the β increases, the pruning precision increases while the recall decreases for the pruned deep ensembles of both datasets. Since the design of our focal diversity based hierarchical pruning approach aims at achieving a very high precision with a good recall, such as 50% recall, for CIFAR-10 with $S_d = 4$, we choose $\beta = 20\%$ to achieve the pruning efficiency measured by 81.25% precision and 52% recall. We follow the same principle to determine the β for other experiments.

We then compare the four focal diversity based hierarchical pruning methods and the diversity consensus voting based focal diversity pruning in terms of pruning efficiency measured by precision and recall on two benchmark datasets.

CIFAR-10. Table IV shows the evaluation of our hierarchical pruning on CIFAR-10 by comparing the four focal model

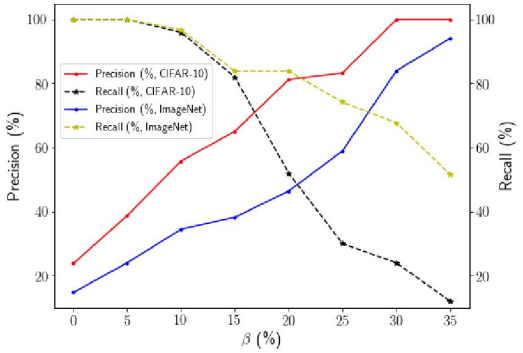


Fig. 4: Impact of β on Precision and Recall ($S_d = 4$, F-GD)

TABLE IV: Hierarchical Pruning with S_d=5 on CIFAR-10

Methods	F-CK	F-BD	F-KW	F-GD	MAJ-F
Precision (%)	85.71	**100**	**100**	**100**	**100**
Recall (%)	17.14	51.43	51.43	52.86	47.14

TABLE V: Hierarchical Pruning with S_d=4 on CIFAR-10

Methods	F-CK	F-BD	F-KW	F-GD	MAJ-F
Precision (%)	26.09	**81.25**	**81.25**	**81.25**	**81.25**
Recall (%)	12.00	52.00	52.00	52.00	52.00

enhanced diversity metrics and the diversity consensus voting based focal diversity pruning over the four different focal diversity pruning methods, with $\beta = 10\%$ and $S_d = 5$ (Cost: 50%), where F-GD corresponds to Figure 3. We highlight two interesting observations. *First*, our hierarchical pruning algorithm performs very well with all four focal diversity pruning methods, achieving over 85% precision in identifying smaller ensembles of the same or higher ensemble accuracy than the entire ensemble. In particular, F-BD, F-KW and F-GD achieved 100% precision for pruning the entire deep ensemble. *Second*, the diversity consensus voting based focal diversity pruning, denoted by MAJ-F, maintains the 100% precision. Its slightly lower recall compared to F-BD, F-GD, F-KW is expected since it further refines the pruning results by removing those ensembles not being selected by at least three focal diversity metrics, thus, slightly reducing the number of selected ensembles. Table V shows similar observations by changing $\beta = 20\%$ and $S_d = 4$ (Cost: 40%). In this setting, the F-CK pruning performs worse than the other three focal diversity pruning methods, while the diversity consensus

voting based focal diversity pruning method MAJ-F can still maintain 81.25% precision, demonstrating its robustness.

TABLE VI: Hierarchical Pruning with S_d=5 on ImageNet

Methods	F-CK	F-BD	F-KW	F-GD	MAJ-F
Precision (%)	0	75.61	75.61	74.42	**78.95**
Recall (%)	0	64.58	64.58	66.67	62.50

ImageNet. Table VI shows the evaluation comparison experiments on ImageNet with $\beta = 10\%$ and $S_d = 5$ (Cost: 50%). We make two highlights. *First,* the hierarchical pruning methods using F-BD, F-KW and F-GD all achieved very high precision of above 74.42%, successfully identifying the smaller ensembles of 50% smaller than $M = 10$ and with ensemble accuracy above the target accuracy 79.82% of the entire deep ensemble. However, F-CK failed to find even one satisfying ensemble for ImageNet, showing the limitation of the CK diversity even optimized by focal model based enhancement. *Second,* the diversity consensus voting based focal diversity pruning method (MAJ-F) can further improve the precision from 74.42%$\sim$75.61% to 78.95%, demonstrating the robustness and effectiveness of combining the focal diversity metrics via majority voting.

V. Conclusion

We have presented a focal diversity based hierarchical ensemble pruning approach, which can find significantly smaller deep ensembles while still retain the same or achieve even better ensemble accuracy than the entire deep ensemble. This paper makes three original contributions. First, we use the focal diversity metrics to accurately capture the negative correlation among member models of an ensemble team. This is the first approach to efficiently comparing ensemble diversity and identifying high quality deep ensembles with high ensemble diversity, and we show this approach can significantly boost the ensemble accuracy for these ensembles selected by using diversity metrics. Second, we present a hierarchical pruning algorithm by leveraging the focal diversity property, which can progressively identify and remove the ensembles with high negative correlation. Third, we present our HQ approach to finding high quality ensembles with a small size (low cost) and high ensemble accuracy for a given pool of multiple base models. It consists of focal diversity metrics, hierarchical pruning algorithms and the diversity consensus voting based ensemble pruning method. Comprehensive experiments conducted on two benchmark datasets of CIFAR-10 and ImageNet show that our focal diversity based hierarchical pruning can effectively prune out ensembles of highly similar individual member models and find substantially smaller deep ensembles with the same or better ensemble accuracy than that of the entire ensemble, effectively reducing the space and time cost for ensemble execution.

Acknowledgment

This research is partially sponsored by National Science Foundation under NSF 1564097, NSF 2038029, a Cisco grant, and an IBM faculty award.

References

[1] L. Liu, W. Wei, K. Chow, M. Loper, E. Gursoy, S. Truex, and Y. Wu, "Deep neural network ensembles against deception: Ensemble diversity, accuracy and robustness," in *2019 IEEE 16th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, 2019, pp. 274–282.

[2] W. Wei and L. Liu, "Robust deep learning ensemble against deception," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 04, pp. 1513–1527, July 2021.

[3] S. Fort, H. Hu, and B. Lakshminarayanan, "Deep Ensembles: A Loss Landscape Perspective," *arXiv e-prints*, p. arXiv:1912.02757, Dec. 2019.

[4] W. Wei, L. Liu, M. Loper, K. Chow, E. Gursoy, S. Truex, and Y. Wu, "Cross-layer strategic ensemble defense against adversarial examples," in *2020 International Conference on Computing, Networking and Communications (ICNC)*, 2020, pp. 456–460.

[5] A. Lazarevic and Z. Obradovic, "Effective pruning of neural network classifier ensembles," in *Proceedings of IJCNN'01. International Joint Conference on Neural Networks.*, vol. 2, 2001, pp. 796–801 vol.2.

[6] G. Martínez-Muñoz, D. Hernández-Lobato, and A. Suárez, "An analysis of ensemble pruning techniques based on ordered aggregation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 245–259, 2009.

[7] Y. Wu, L. Liu, Z. Xie, K.-H. Chow, and W. Wei, "Boosting ensemble accuracy by revisiting ensemble diversity metrics," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2021.

[8] Y. Wu, L. Liu, Z. Xie, J. Bae, K.-H. Chow, and W. Wei, "Promoting high diversity ensemble learning with ensemblebench," in *2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI)*, 2020, pp. 208–217.

[9] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[10] L. Breiman *et al.*, "Arcing classifier (with discussion and a rejoinder by the author)," *The annals of statistics*, vol. 26, no. 3, pp. 801–849, 1998.

[11] L. Breiman, "Random forests," in *Machine Learning*, 2001, pp. 5–32.

[12] C. Ju, A. Bibaut, and M. Laan, "The relative performance of ensemble methods with deep convolutional neural networks for image classification," *Journal of Applied Statistics*, vol. 45, 04 2017.

[13] G. Tsoumakas, I. Partalas, and I. Vlahavas, *An Ensemble Pruning Primer*. Berlin, Heidelberg: Springer, 2009, pp. 1–13.

[14] K.-H. Chow and L. Liu, "Robust object detection fusion against deception," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2021.

[15] Y. Bian, Y. Wang, Y. Yao, and H. Chen, "Ensemble pruning based on objection maximization with a general distributed framework," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 9, pp. 3766–3774, 2020.

[16] X.-C. Yin, C. Yang, and H.-W. Hao, "Learning to diversify via weighted kernels for classifier ensemble," *arXiv preprint arXiv:1406.1167*, 2014.

[17] K. Chow, W. Wei, Y. Wu, and L. Liu, "Denoising and verification cross-layer ensemble against black-box adversarial attacks," in *2019 IEEE International Conference on Big Data (Big Data)*, 2019, pp. 1282–1291.

[18] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," University of Toronto, Tech. Rep., 2009.

[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.

[20] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia medica*, vol. 22, no. 3, p. 276—282, 2012.

[21] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Mach. Learn.*, vol. 51, no. 2, p. 181–207, May 2003.

[22] E. K. Tang, P. N. Suganthan, and X. Yao, "An analysis of diversity measures," *Machine learning*, vol. 65, no. 1, pp. 247–271, 2006.

[23] D. B. Skalak, "The sources of increased accuracy for two proposed boosting algorithms," in *In Proc. AAAI-96, Integrating Multiple Learned Models Workshop*, 1996, pp. 120–125.

[24] R. Kohavi and D. Wolpert, "Bias plus variance decomposition for zero-one loss functions," in *Proceedings of the 13th International Conference on International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1996, p. 275–283.

[25] D. Partridge and W. Krzanowski, "Software diversity: practical statistics for its measurement and exploitation," *Information and Software Technology*, vol. 39, no. 10, pp. 707 – 717, 1997.