

Boosting Ensemble Accuracy by Revisiting Ensemble Diversity Metrics

Yanzhao Wu, Ling Liu, Zhongwei Xie, Ka-Ho Chow, Wenqi Wei
School of Computer Science
Georgia Institute of Technology
Atlanta, Georgia 30332

yanzhaowu@gatech.edu, lingliu@cc.gatech.edu, {zhongweixie, khchow, wenqiwei}@gatech.edu

Abstract

Neural network ensembles are gaining popularity by harnessing the complementary wisdom of multiple base models. Ensemble teams with high diversity promote high failure independence, which is effective for boosting the overall ensemble accuracy. This paper provides an in-depth study on how to design and compute ensemble diversity, which can capture the complementary decision capacity of ensemble member models. We make three original contributions. First, we revisit the ensemble diversity metrics in the literature and analyze the inherent problems of poor correlation between ensemble diversity and ensemble accuracy, which leads to the low quality ensemble selection using such diversity metrics. Second, instead of computing diversity scores for ensemble teams of different sizes using the same criteria, we introduce focal model based ensemble diversity metrics, coined as FQ-diversity metrics. Our new metrics significantly improve the intrinsic correlation between high ensemble diversity and high ensemble accuracy. Third, we introduce a diversity fusion method, coined as the EQ-diversity metric, by integrating the top three most representative FQ-diversity metrics. Comprehensive experiments on two benchmark datasets (CIFAR-10 and ImageNet) show that our FQ and EQ diversity metrics are effective for selecting high diversity ensemble teams to boost overall ensemble accuracy.

1. Introduction

Ensemble learning aims to produce a strong model by harnessing the combined and complementary wisdom of multiple base models. There are two broad categories of approaches to construct high quality ensemble teams: (1) data driven or model driven training of multiple models to form an ensemble and (2) selecting ensemble teams from a given pool of diverse base models (learners). The former is represented by boosting algorithms [3, 17], bagging methods [1], and random forests [2]. The latter is repre-

sented by ensembles of base models, which are trained using diverse neural network structures and diverse settings of hyperparameters [7, 11, 19, 23, 24, 27], including those pre-trained models in public domains. This paper is dedicated to the second category, namely the problem of selecting high quality ensemble teams from a base model pool. Given a pool of M diverse base models, there are M exponential number of possible candidate ensemble teams, a large portion of which may not offer high ensemble performance due to insufficient failure independence among their member models [13, 14, 15, 18, 25]. Ensemble diversity metrics are widely regarded as representative methods for capturing failure independence among member models of ensemble teams and expected to have stable correlation with ensemble accuracy.

1.1. Related Work

Both pairwise and non-pairwise diversity metrics have been proposed in the literature. The pairwise diversity metrics are represented by Cohen’s Kappa (CK) [14], Q Statistics (QS) [26], and Binary Disagreement (BD) [18]. The non-pairwise diversity metrics are represented by Fleiss’ Kappa (FK) [5], Kohavi-Wolpert Variance (KW) [9, 12], and Generalized Diversity (GD) [15]. For presentation brevity, we refer to these existing diversity metrics as Q-metrics. These diversity metrics and their relationship to ensemble accuracy have been studied over traditional machine learning models [12] and over trained neural network models [25]. These Q-diversity metrics share one common property: they compute the diversity of ensemble teams of different sizes using a common criterion.

1.2. Scope and Contributions

In this paper, we revisit and analyze the inherent problems of choosing ensembles using the existing diversity metrics and why these diversity metrics are inefficient to capture failure independence among member models of an ensemble team. We introduce six new diversity metrics, coined as FQ-diversity metrics. The main idea of FQ-

diversity metrics is three folds: (1) For a base model pool of size M , we divide all candidate ensemble teams into $M - 1$ partitions with each representing the set of ensemble teams of equal team size S , ranging from 2 to M , such that the ensemble diversity scores are computed and compared among the ensemble teams of equal size. (2) To accurately capture the failure independence of member models of ensemble teams and the correlation between ensemble diversity and ensemble accuracy, we introduce the focal model based negative sampling for computing and combining FQ diversity scores for each ensemble team. (3) Instead of using a pre-defined mean diversity threshold to partition all candidate ensemble teams into high diversity and low diversity partitions, we leverage a binary clustering method with strategic initialization of the two centroids to automatically divide each of the $M - 1$ partitions of ensembles of size S ($2 \leq S \leq M$) into two clusters: keeping the cluster of ensemble teams with low FQ scores (high ensemble diversity) and high ensemble accuracy, and removing the other cluster of ensemble teams with high FQ scores (low ensemble diversity). In addition, we propose a diversity fusion mechanism by introducing the EQ-diversity measure, which integrates the top three FQ scores for selecting high quality ensembles. Comprehensive experiments are conducted on two benchmark datasets (CIFAR-10 [10] and ImageNet [16]) with ten base models each with the standard soft voting (model averaging) for producing ensemble consensus predictions [8, 20]. The results show that our FQ and EQ diversity metrics are effective in identifying and selecting high quality ensemble teams and boosting overall ensemble accuracy.

2. Q-diversity based Ensemble Selection

Given a pool of M base models for a learning task and its training dataset \mathcal{D} , we have $BMSet(\mathcal{D}) = \{F_0, \dots, F_{M-1}\}$. Let $EnsSet$ denote the set of all possible ensemble teams with team size S ranging from 2 to M , composed from $BMSet(\mathcal{D})$. We have $|EnsSet| = \sum_{S=2}^M \binom{M}{S} = \binom{M}{2} + \binom{M}{3} + \dots + \binom{M}{M} = 2^M - (1 + M)$. Consider $M = 5$, we have $|EnsSet| = 26$. The number of ensemble teams in $EnsSet$ increases exponentially with M . For a larger M , such as $M = 10$ or $M = 20$, $|EnsSet| = 1013$ or 1,048,555 respectively. Table 1 lists ten base models for each of the two benchmark datasets used in this paper.

Evaluation Metrics. Let $GEnsSet$ denote the set of good quality ensemble teams selected from the candidate set $EnsSet$, according to a diversity metric. One indicator for high quality ensemble teams is that they can outperform the maximum model accuracy of their member models (**m_max**). Let $(min_GEnsSet, max_GEnsSet)$ denote the ensemble accuracy range for the selected ensembles in $GEnsSet$. Another indicator of an efficient ensemble selection algorithm is measured by the expected accu-

Dataset	CIFAR-10		ImageNet	
	10,000 testing samples		50,000 testing samples	
Model ID	Models	Accuracy (%)	Models	Accuracy (%)
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
MIN (p_min)	ResNet20	91.73	AlexNet	56.63
AVG (p_avg)		94.25		71.60
MAX (p_max)	DenseNet190	96.68	ResNet152	78.25

Table 1: Base Model Pools

racy range of selected ensemble teams. For example, it is a good indicator if the **lower bound** ($min_GEnsSet$) of the selected ensemble teams in $GEnsSet$ is higher than the average accuracy of the base models in the pool, i.e., $p_avg = avg_BMSet(\mathcal{D})$, $min_GEnsSet \geq p_avg$. In addition, we can also measure the number of selected ensemble teams in $GEnsSet$, which outperform the base model with the maximum accuracy in the pool (**p_max**).

Methods	#EnsSet	#GEnsSet	Ensemble Acc Range (%)
Q-CK	1013	555	61.39~80.50
Q-QS	1013	483	61.39~80.54
Q-BD	1013	554	61.39~80.54
Q-FK	1013	553	61.39~80.50
Q-KW	1013	647	68.72~80.56
Q-GD	1013	530	70.79~80.60

Table 2: Ensemble Selection by Q-diversity on ImageNet

Consider the ten base models of ImageNet in Table 1. We provide the set of ensemble teams selected using the six Q-diversity metrics in Table 2. We observe that among the total of 1013 ensemble teams in the candidate set $EnsSet$, the four Q-diversity metrics: Q-CK, Q-QS, Q-BD, and Q-FK all have poor lower bound of 61.39%, in terms of the accuracy range of the selected ensembles in $GEnsSet$, about 10% lower than the average accuracy of the base model pool of 71.60% (p_avg), showing low quality of ensemble selections. Although in comparison, Q-GD diversity metric has the highest lower bound of 70.79%, which is still lower than the average accuracy of 71.60% over the total of 10 base models. This motivates us to further analyze the inherent problems of using Q-diversity metrics to perform ensemble selection.

Let Y denote a random variable, representing the proportion of models (i.e., i out of S) that fail to recognize a random input sample \mathbf{x} . The probability of $Y = \frac{i}{S}$ is denoted as p_i . In Formula (1), $p(1)$ represents the expected probability of one randomly picked model failing while $p(2)$ de-

notes the expected probability of both two randomly picked models failing. Formula (1) presents the definition of the generalized diversity (GD) metric [15].

$$p(1) = \sum_{i=1}^S \frac{i}{S} p_i, p(2) = \sum_{i=1}^S \frac{i(i-1)}{S(S-1)} p_i, GD = 1 - \frac{p(2)}{p(1)} \quad (1)$$

GD varies from 0 to 1. The maximum diversity score of 1 occurs when the failure of one model is accompanied by the correct recognition by the other model, that is $p(2) = 0$. When both two models fail, we have $p(1) = p(2)$, leading to the minimum diversity score of 0.

Let *NegSampSet* denote a small set of negative samples randomly sampled from the set of negative samples of the base model pool, such as 100 negative samples randomly drawn from the training data, on which one or more base models make errors. This *NegSampSet* will be used to compute the Q-GD diversity score for ensemble teams in the candidate set *EnsSet*. Due to the space constraint, the formal definitions of the other five diversity metrics are given in the supplementary material. To present a consistent view of all six diversity metrics such that the low value corresponds to high ensemble diversity, we apply (1-value) when calculating the diversity scores using *BD*, *KW* and *GD* metrics.

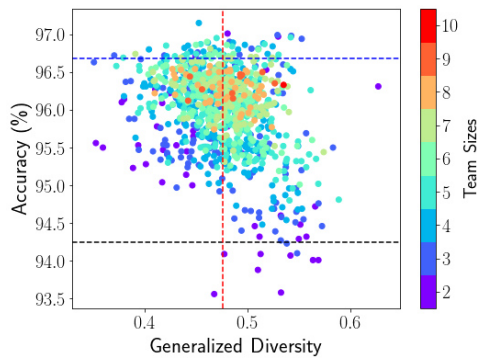


Figure 1: Q-GD, 1013 Teams (CIFAR-10)

Figure 1 shows the GD metric and its relationship with ensemble accuracy for the 1013 candidate ensemble teams in *EnsSet* for the 10 base models of CIFAR-10 (recall Table 1). Each dot represents one ensemble team in *EnsSet* and the color indicates the team size S , ranging from 2 to 10 ($M = 10$), according to the color diagram on the right. The horizontal blue and black dashed lines are the maximum base model accuracy 96.68% (p_{max}) and the average accuracy 94.25% (p_{avg}) of all $M = 10$ base models respectively. We use such two accuracy bounds to analyze the quality of the ensemble teams selected. It is visually clear that using the mean Q-GD diversity value as the cut-off threshold, as indicated by the red vertical dashed line in Figure 1, is not effective for selecting good ensembles

from the 1013 candidate teams for two reasons: (1) there is no clear correlation between ensemble diversity and ensemble accuracy among those selected ensembles in *GEEnsSet*, which have diversity scores below the mean threshold; and (2) among those remaining ensemble teams that are discarded, some ensemble teams with high Q-GD diversity scores also have high ensemble accuracy. Similar observations can be found in other five Q-diversity metrics as well. This motivates us to revisit two design components in computing Q-diversity metrics: (1) the Q-diversity scores are computed for ensembles of different team sizes. Intuitively, it may not be meaningful to compare the ensemble teams of different team sizes in terms of their ensemble diversity; and (2) the negative samples randomly selected from the collection of negative samples from all base models are used uniformly to compute the Q-diversity scores [12, 25] of all candidate ensembles. Such design may not adequately capture the complementary capacity among the member models of all candidate ensemble teams of different sizes.

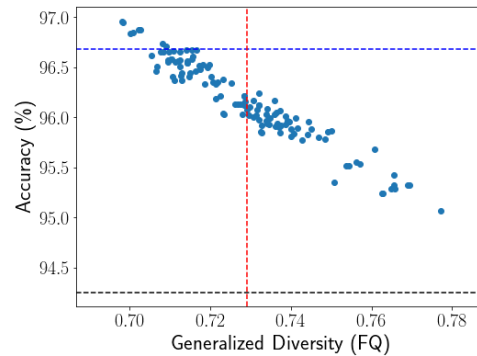


Figure 2: FQ-GD, $S=5, focal=1, 126$ Teams (CIFAR-10)

To address the above problems, we propose the concept of focal model for negative sampling, and compute and compare the diversity scores for ensemble teams of equal size for ensemble selection. We conduct a preliminary experiment for CIFAR-10 by only computing and comparing all candidate ensemble teams of equal size $S = 5$ using our FQ-GD diversity metric. Figure 2 shows the set of candidate ensemble teams of size 5 measured with the negative samples from the focal model F_1 (denoted by $focal=1$), where the red vertical dashed line marks the mean threshold 0.729. From Figure 2, it is visually clear that the focal model based FQ diversity metrics capture the close to linear correlation between ensemble diversity scores and ensemble accuracy.

3. FQ-diversity based Ensemble Selection

Based on the analysis of the inherent problems with Q-diversity metrics and the encouraging preliminary result in Figure 2, we propose to extend the existing diversity met-

rics with six new ensemble diversity metrics, coined as FQ-diversity metrics. Unlike Q-diversity metrics, the FQ-diversity metrics compute and compare the diversity scores among the ensemble teams of a fixed size S with the focal model based negative sampling and the FQ-diversity based binary partitioning.

(1) Equal Size Ensembles: Given the total of M based models in the base model pool, we further divide the candidate ensemble team set $EnsSet$ into $M - 1$ partitions, each consists of ensemble teams of equal size S , denoted by $EnsSet(S)$ ($2 \leq S \leq M$). Given $M = 10$, we will have a total of 1013 candidate ensembles in $EnsSet$ and a total of 252 ensembles in $EnsSet(S)$ for $S = 5$.

(2) Focal Models: The idea of using a focal model is motivated by ensemble defense against adversarial attacks [13, 4, 21, 22], where the ensemble teams are composed to protect a target victim (focal) model. We introduce the concept of focal model, say F_{focal} , and use the negative samples from the focal model to compute the FQ diversity scores for all ensemble teams of a fixed size S , which have the same F_{focal} as a member, denoted by $EnsSet(F_{focal}, S)$. Recall Figure 2, with $focal = 1$ and $S = 5$, the total number of ensemble teams in $EnsSet(F_{focal}, S)$ will be 126. In addition to the focal model based negative sampling, the FQ-diversity based ensemble selection will also perform the binary clustering over $EnsSet(F_{focal}, S)$ to select the good ensemble teams with respect to each focal model and its corresponding ensembles of equal size S , instead of performing binary clustering over all ensembles in $EnsSet$.

(3) Focal Model Based Ensemble Selection: The focal model based ensemble selection performs two tasks. First, given a fixed team size S , a focal model F_{focal} , and a FQ-diversity metric, say FQ-GD, we first compute the FQ-diversity value (q) and the accuracy (acc) for each ensemble team in $EnsSet(F_{focal}, S)$ using the negative samples of the focal model F_{focal} ($NegSampSet(F_{focal})$), and the output set is denoted as $DA(Q) = \{(q_i, acc_i) | T_i \in EnsSet(F_{focal}, S)\}$. We then employ a binary cluster algorithm, such as K-means, to partition $DA(Q)$ into two clusters, with $K=2$ and two initial centroids chosen strategically based on FQ-diversity. Concretely, we choose two specific points in the 2D space of $DA(Q)$ as the two centroids for the K-means clustering algorithm. The first centroid should have the smallest FQ-value and the highest ensemble accuracy measure over $DA(Q)$, denoted as (q_{min}^1, acc_{max}^1) , such that $\forall (q_i, acc_i) \in DA(Q), q_{min}^1 \leq q_i$ and $acc_{max}^1 \geq acc_i$, and $\exists j, k \in \{1, 2, \dots, |EnsSet(F_{focal}, S)|\}, q_{min}^1 = q_j, acc_{max}^1 = acc_k$. Correspondingly, the second centroid should have the largest FQ-value and the lowest accuracy over $DA(Q)$, denoted as (q_{max}^2, acc_{min}^2) . The K-means clustering algorithm will partition $DA(Q)$ into two clusters: $Cluster_1$ with the centroid (q^1, acc^1) and $Cluster_2$ with the centroid (q^2, acc^2) , satisfying the following prop-

erty: $q^1 \leq q^2$ and $acc^1 \geq acc^2$.

The second task is to leverage the binary clustering to define the proper threshold for a given FQ diversity metric, e.g., FQ-GD. Let $min_{div}(Cluster_2)$ be the lowest FQ-value in $Cluster_2$ and $mean_{div}(DA(Q))$ denote the mean value of all FQ-diversity values in $DA(Q)$. We compute the FQ-diversity threshold $\theta_{FQ}(F_{focal}, S, Q) = min(min_{div}(Cluster_2), mean_{div}(DA(Q)))$. By the FQ-diversity based ensemble selection algorithm, which is provided in the supplementary material, all ensemble teams in $Cluster_1$ whose FQ-diversity scores are below the FQ diversity threshold $\theta_{FQ}(F_{focal}, S, Q)$ are selected and placed into the selected set of good ensemble teams, denoted by $GEnsSet(F_{focal}, S, Q)$, i.e., $\forall T_i \in EnsSet(F_{focal}, S)$, if $q_i < \theta_{FQ}(F_{focal}, S, Q)$, then $T_i \in GEnsSet(F_{focal}, S, Q)$.

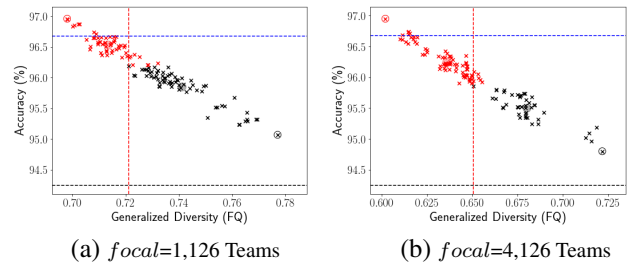


Figure 3: K-Means Thresholds for Different Focal Models (CIFAR-10, $S = 5$, FQ-GD)

Figure 3a and Figure 3b provide two visual illustrations of FQ-GD diversity computation on CIFAR-10 with $S=5$ and F_1 (DenseNet100) and F_4 (VGG19) as the focal models respectively. The 10 base models on CIFAR-10 are given in Table 1. The blue and black horizontal dashed lines indicate $p_{max}=96.68\%$ and $p_{avg}=94.25\%$. The red vertical dashed line marks the diversity threshold computed by using the K-means binary clustering. Figure 3a shows red and black clusters for FQ-GD with $focal = 1$, and the threshold 0.721 is computed using the K-means algorithm with the red and black unfilled circles as the initial centroids and the red and black solid circles as the two calculated centroids. Based on the computed threshold marked by the red vertical line, we select the ensemble teams with FQ-GD scores below the threshold. Similar analysis for FQ-GD with $focal=4$ in Figure 3b. From these two sets of experiments, we make three observations: (1) Both focal model cases exhibit the close to linear correlation between ensemble diversity (FQ-GD) and ensemble accuracy. (2) Focal model based ensemble selection can effectively prune out low quality ensemble teams by leveraging the close to linear correlation between ensemble diversity and ensemble accuracy. (3) Interestingly, not all the good ensemble teams selected by FQ-GD with focal model F_1 in Figure 3a are

also good ensemble teams selected by FQ-GD with a different focal model F_4 . For example, the ensemble team $F_0F_1F_2F_3F_5$ with 96.85% ensemble accuracy is selected by FQ-GD with focal model F_1 , but it is not even a legitimate candidate ensemble in $EnsSet(F_4, S=5)$, in which all ensemble candidates will include the focal model F_4 as a member model. There are a total of 16 ensemble teams in the intersection of $GEnsSet(F_1, S=5, FQ-GD)$ (48 teams) and $GEnsSet(F_4, S=5, FQ-GD)$ (77 teams). To assign one diversity score to each ensemble team, we need to further combine different focal model based FQ-diversity scores to provide efficient ensemble selections.

(4) Combining Different Focal Models: For each of the FQ-diversity metrics, say FQ-GD, we have performed focal model based ensemble selection by the steps (1) to (3). At the end of step (3), we obtain $GEnsSet(F_{focal}, S, FQ-GD)$, which contains the good ensembles selected by FQ-GD from the focal model based candidate ensemble set $EnsSet(F_{focal}, S)$ with respect to the focal model F_{focal} and the ensemble teams of equal size S . Consider a given ensemble team of size $S = 3$, say $F_0F_1F_2$, this ensemble will have three focal model based FQ-GD scores, corresponding to the three focal models. After the first three steps (1)~(3), there are four possible results for this specific ensemble team: (a) it is selected by only one of the three focal model based ensemble selections, say $GEnsSet(F_0, S=3, FQ-GD)$; (b) it is selected by two out of three focal model based ensemble selections, say $GEnsSet(F_0, S=3, FQ-GD)$ and $GEnsSet(F_1, S=3, FQ-GD)$; (c) it is selected by all three focal model based ensemble selections, i.e., $\bigcap_{i=0}^2 GEnsSet(F_i, S=3, FQ-GD)$; (d) it is pruned out by all three focal model based ensemble selections.

In order to produce one unifying FQ-GD score for each ensemble team, we first scale the FQ-GD scores in each $EnsSet(F_{focal}, S)$ to $[0, 1]$ and then use the simple averaging over the S number of scaled FQ-GD scores to produce the unifying FQ-GD score of each ensemble team. Here, we use $EnsSet_{unifyFQ}(S, FQ-GD) = \bigcup_{focal=0}^{M-1} GEnsSet(F_{focal}, S, FQ-GD)$ to denote the set of selected ensemble teams of size S , each with the unifying FQ-GD score. We then use the unifying FQ-GD scores to obtain $DA(FQ-GD) = \{(q_i, acc_i) | T_i \in EnsSet_{unifyFQ}(S, FQ-GD)\}$. Next we perform binary clustering over $DA(FQ-GD)$ to obtain the set of good ensemble teams using their unifying FQ-GD scores, denoted by $GEnsSet(S, FQ-GD)$.

Figure 4a shows a visualization to illustrate the final ensemble selection using the unifying FQ-GD scores for candidate ensembles of equal size $S = 5$. The red vertical dashed line marks the learned diversity threshold (K-means threshold for short), based on the binary clustering on $EnsSet_{unifyFQ}(S, FQ-GD)$ using the K-means

algorithm with $K=2$. The black dots on the right side represent the ensemble teams pruned out by this threshold while the red dots mark the selected ensemble teams in $GEnsSet(S, FQ-GD)$ from the candidate ensemble set $EnsSet_{unifyFQ}(S, FQ-GD)$. The yellow dots mark those ensemble teams that were removed already by focal model specific pruning in step (3). The unifying FQ-GD ensemble selection is built on top of the set of S focal model specific FQ-GD scores by unifying them with simple averaging for each ensemble team. From Figure 4a, it is clear that the ensemble selection using the unifying FQ-GD scores further improves the overall quality of the ensemble teams selected by further pruning out some low quality ensemble teams. For example, the unifying FQ-GD further pruned out 58 out of 172 ensemble teams and increased the ensemble accuracy lower bound from 95.84% to 95.90%. In summary, we have $GEnsSet(FQ-GD) = \bigcup_{S=2}^{M-1} GEnsSet(S, FQ-GD)$ as the set of selected ensembles by using our unifying FQ-GD ensemble selection algorithm. Similar ensemble selection can be obtained by using the other five FQ-diversity metrics, improving the corresponding Q-diversity metrics.

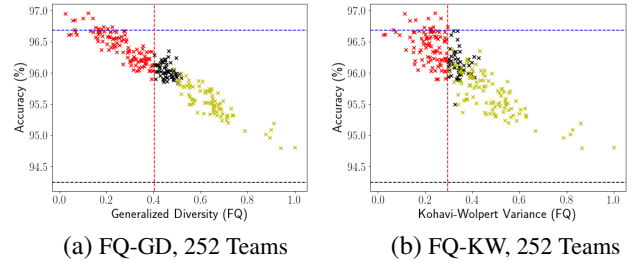


Figure 4: K-Means Thresholds on Equal Size Ensembles for Different FQ Metrics (CIFAR-10, $S = 5$)

(5) FQ Fusion Based Ensemble Selection: With the unifying FQ scores, ensemble teams of the equal size can be directly compared for ensemble selection based on FQ diversity. We have shown the FQ-GD based ensemble selection for ensemble teams of size $S=5$ in Figure 4a. The same process can be applied to the other five FQ-diversity metrics. Figure 4b shows the ensemble team selection with equal size ensembles ($S = 5$) using FQ-KW scores. Similarly, the red vertical dashed line marks the learned diversity threshold based on the K-means binary clustering. The black dots on the right side represent the ensemble teams pruned out by this threshold while the red dots mark the selected ensemble teams in $GEnsSet(S, FQ-KW)$ from the candidate ensemble teams in $EnsSet_{unifyFQ}(S, FQ-KW)$. Similarly, we observe that ensemble selection using the unifying FQ-KW scores further prunes out low quality ensemble teams for FQ-KW as well. For example, the FQ-KW unifying ensemble selection further pruned out

Methods	#EnsSet	#GEnsSet	Ensemble Acc Range (%)	Ensemble Acc Avg (%)	# (Acc \geq m_max)	% (Acc \geq m_max)	# (Acc \geq 96.68% p_max)	% (Acc \geq 96.68% p_max)
FQ-CK	1013	72	95.04~97.01	96.04	13	18.06%	6	8.33%
FQ-QS	1013	42	94.40~96.73	96.07	9	21.43%	2	4.76%
FQ-BD	1013	369	95.23~97.15	96.37	109	29.54%	60	16.26%
FQ-FK	1013	75	95.04~97.01	96.03	13	17.33%	6	8.00%
FQ-KW	1013	370	95.09~97.15	96.36	110	29.73%	61	16.49%
FQ-GD	1013	443	95.23~97.15	96.38	120	27.09%	66	14.90%
EQ (BD+KW+GD)	1013	336	95.23~97.15	96.40	107	31.85%	60	17.86%

Table 3: The Experimental Comparison of FQ and EQ-diversity based Ensemble Selection on CIFAR-10

50 out of 145 ensemble teams and increased the ensemble accuracy lower bound from 95.50% to 95.90%. From Figure 4a and Figure 4b, we also observe that FQ-GD selected more teams (114 teams) than FQ-KW (95 teams), showing that different FQ-diversity metrics may select different ensemble teams. Concretely, the intersection of $GEnsSet(S, FQ-GD)$ (114 teams) and $GEnsSet(S, FQ-KW)$ (95 teams) contains 92 ensemble teams. Hence, we can further prune out a fair number of low quality teams, that is 22 teams from $GEnsSet(S, FQ-GD)$ and 3 teams from $GEnsSet(S, FQ-KW)$, which are all below the $p_max=96.68\%$ accuracy, by using the intersection of FQ-GD and FQ-KW. This motivates us to design the EQ-diversity measure to combine the top performing FQ-diversity metrics and further boost the overall ensemble accuracy of selected ensemble teams.

EQ-diversity Ensemble Selection. For each FQ-diversity metric, we obtain a set of high diversity ensembles, denoted by $GEnsSet(FQ-Q)$, by combining the selections for different team sizes ($2 \leq S \leq M - 1$). Let $GEnsSet_{FQ}$ denote the union of $GEnsSet(FQ-Q)$ for six FQ-diversity metrics. A simple and yet representative approach to combine the six FQ-diversity scores for all ensemble teams in $GEnsSet_{FQ}$ is to simply take the set intersection (\cap) of the teams selected by the top three FQ-diversity metrics, based on the accuracy range defined by min and max ensemble accuracy of all teams selected under a given FQ metric. An example could be $GEnsSet_{FQ}(CK + KW + GD) = GEnsSet(FQ-CK) \cap GEnsSet(FQ-KW) \cap GEnsSet(FQ-GD)$. This approach removes those ensemble teams that are not included in the intersection of the ensemble teams chosen by the top 3 FQ-diversity metrics. Our experimental results show that the ensembles selected by this EQ-diversity measure can outperform ensembles chosen by both FQ-diversity metrics and Q-diversity metrics with consistent performance for booting ensemble accuracy.

4. Experimental Analysis

We conduct extensive experiments on two benchmark datasets (CIFAR-10 [10] and ImageNet [16]), each with ten base models as the base model pool (see Table 1), to compare and evaluate the effectiveness of high quality ensemble selection using the proposed FQ and EQ-diversity metrics.

All experiments are conducted on an Intel Xeon E5-1620 server with the NVIDIA GeForce GTX 1080 Ti (11GB) GPU, installed with Ubuntu 16.04 LTS, CUDA 8.0.

4.1. Performance of FQ Metrics on CIFAR-10

Table 3 shows the experimental comparison of the ensemble teams selected by using ensemble diversity of FQ and EQ for CIFAR-10. We use the evaluation metrics introduced in Section 2 to evaluate the quality of selected ensemble teams, including the ensemble accuracy range, ensemble accuracy average and the number and percentage of selected ensemble teams that outperform their best member model (m_max) and the maximum base model in the pool (p_max). The max single base model accuracy from the 10 base models is 96.68% from DenseNet190. We highlight three observations. *First*, the ensemble teams selected by all FQ/EQ-diversity metrics provide a high ensemble accuracy lower bound of 94.40%, which is a significant improvement over the average accuracy 94.25% of 10 base models and also over the lower bound of 93.56% of using the corresponding Q-diversity metrics. 5 out of 6 FQ-diversity metrics can further increase this ensemble accuracy lower bound to above 95.04%. In particular, FQ-BD, FQ-KW, and FQ-GD can identify over 100 teams with ensemble accuracy above the max member model accuracy (m_max) and over 60 teams with ensemble accuracy above p_max of 96.68%, the max accuracy of the 10 base models for CIFAR-10. *Second*, the upper bound for 5 out of 6 FQ-diversity metrics is above 97.01%. It further demonstrates that our FQ-diversity based ensemble selection can select high quality ensemble teams while pruning out low quality ensemble teams. *Third*, the EQ-diversity method can leverage FQ-diversity fusion to further improve the quality of ensemble selection compared to FQ metrics, further boosting the overall ensemble accuracy of the selected ensemble teams, by pruning out those ensembles that are not in the intersection of the selected teams by the top three FQ metrics. For CIFAR-10, the EQ-diversity with BD+KW+GD improved the average ensemble accuracy to 96.40%.

4.2. Performance of FQ Metrics on ImageNet

We performed the same set of experiments on ImageNet. Table 4 shows the experimental results. For ImageNet, the maximum single base model accuracy is 78.25%

Methods	#EnsSet	#GEnsSet	Ensemble Acc Range (%)	Ensemble Acc Avg (%)	# (Acc >= m_max)	% (Acc >= m_max)	# (Acc >= 78.25% p_max)	% (Acc >= 78.25% p_max)
FQ-CK	1013	30	73.26~79.55	77.92	22	73.33%	18	60.00%
FQ-QS	1013	127	74.51~80.54	79.11	119	93.70%	108	85.04%
FQ-BD	1013	550	74.65~80.77	79.47	541	98.36%	510	92.73%
FQ-FK	1013	30	73.26~79.55	77.92	22	73.33%	18	60.00%
FQ-KW	1013	563	74.65~80.77	79.45	554	98.40%	521	92.54%
FQ-GD	1013	539	75.27~80.77	79.51	531	98.52%	504	93.51%
EQ (BD+KW+GD)	1013	512	75.27~80.77	79.52	504	98.44%	479	93.55%

Table 4: The Experimental Comparison of FQ and EQ-diversity based Ensemble Selection on ImageNet

Ensemble Team	12345	2345	245	1234	12348	1248	124	123467	128	1289
Ensemble Accuracy (%)	80.77	80.70	80.42	80.29	80.15	79.86	79.84	79.45	78.67	78.62
Highest Member Accuracy (%)	78.25	78.25	78.25	77.40	77.40	77.25	77.25	77.4	77.15	77.15
Highest Member Model	F_5	F_5	F_5	F_3	F_3	F_4	F_4	F_3	F_1	F_1
Accuracy Improvement (%)	2.52	2.45	2.17	2.89	2.75	2.61	2.59	2.05	1.52	1.47

Table 5: 10 examples of good ensemble teams identified by our FQ-GD metric on ImageNet

(ResNet152), and the average accuracy of the 10 base models is 71.60% (see details on Table 1). We highlight three observations. *First*, all FQ/EQ-diversity metrics provide a high ensemble accuracy lower bound of 73.26%, compared to the average accuracy of 71.60% of the base model pool (p_avg). 4 out of 6 FQ-diversity metrics further improved this ensemble accuracy lower bound to above 74.51%. Moreover, FQ-BD, FQ-KW, and FQ-GD perform consistently better than the other three FQ-diversity metrics, and over 92% of the ensembles selected by each of the three FQ-diversity metrics achieve higher ensemble accuracy than the maximum base model accuracy (p_max). *Second*, the percentage of ensemble teams selected by all FQ/EQ-diversity metrics, which are above their respective member max accuracy (m_max) among the selected ensemble teams, is very high. FQ-BD, FQ-KW and FQ-GD are in the range of 98.36% to 98.52% with 93.70% for FQ-QS and 73.33% for FQ-CK and FQ-FK. *Third*, the EQ-diversity further improved the quality of selected ensemble teams by leveraging FQ diversity fusion. For ImageNet, the EQ-diversity with BD+KW+GD improved the ensemble accuracy average to 79.52%. In particular, 98.44% of the ensemble teams selected are achieving higher ensemble accuracy than their respective member max accuracy (m_max). Furthermore, 93.55% of ensembles selected by our EQ-diversity method have ensemble accuracy over p_max=78.25%.

Quality of Selected Ensemble Teams: Table 5 gives 10 examples of ImageNet ensemble teams selected by using our FQ-GD diversity metric as a case study. We use the notation of 128 to denote the ensemble team with base models F_1, F_2, F_8 . Among the total of 539 selected ensemble teams, 504 out of 539 have ensemble accuracy on par or higher than p_max=78.25% (see Table 4). We randomly choose 10 ensemble teams from the group of 504 teams in Table 5. We also include the highest accuracy of the member models for each of the 10 ensemble teams. It

is observed that (1) these 10 selected ensemble teams successfully improve their highest member model accuracy by at least 1.47%, and (2) an ensemble team without the model F_5 with the p_max accuracy (78.25%) in the pool, such as 1234, 1248 and 128, can outperform not only its highest member accuracy but also the p_max=78.25% accuracy. It takes about 20 minutes on a single PC for selecting these high quality ensemble teams with our FQ-GD diversity metric. Comparing to the typical time for designing, training or searching for novel neural networks, which takes several days or even several months [6, 11, 27], our proposed method of using high diversity ensemble teams for improving prediction performance is more cost-effective, and it can be potentially applied to many computer vision tasks.

Visualization of FQ-GD Ensemble Selection. Table 6 presents the visualization of 4 examples from ImageNet to illustrate the effectiveness of our FQ-diversity metrics and FQ-diversity based ensemble selection using FQ-GD. We show the prediction results with Top-5 classification confidence from two ensemble teams selected by our FQ-GD ensemble selection algorithm in Table 5. They are $F_2F_4F_5$ and $F_1F_2F_3F_4$ with their best member model F_5 and F_3 respectively. For all 4 images, the best member models fail to give the correct prediction, whereas the FQ-GD selected ensemble teams can generate the correct predictions, repair the wrong predictions by its high performing member model (with maximum member model accuracy), and boost the overall ensemble prediction accuracy.

4.3. Computation Time Comparison

Table 7 shows a comparison of the total time in seconds for performing ensemble selection using the Q-diversity metrics and FQ-diversity metrics respectively on the two benchmark datasets. Although computing FQ-diversity scores is more involved in first computing the focal model specific FQ-diversity scores and then the final ensemble se-





Image				
Ground Truth Label	beacon	measuring cup	sports car	table lamp
F_5 (ResNet152) p_max, 78.25%	breakwater, 0.90 seashore, 0.07 beacon, 0.02 sandbar, 0.00 lakeside, 0.00 ❌	strainer, 0.94 measuring cup, 0.05 wooden spoon, 0.01 Crock Pot, 0.00 mixing bowl, 0.00 ❌	car wheel, 0.83 sports car, 0.11 grille, 0.02 convertible, 0.02 beach wagon, 0.01 ❌	lampshade, 0.56 table lamp, 0.43 pedestal, 0.00 goblet, 0.00 wall clock, 0.00 ❌
$F_2F_4F_5$, 80.42% (EfficientNet-B0, Inception3, ResNet152)	beacon, 0.53 breakwater, 0.42 seashore, 0.03 sandbar, 0.00 lakeside, 0.00 ✅	measuring cup, 0.56 strainer, 0.32 beaker, 0.01 wooden spoon, 0.00 rule, 0.00 ✅	sports car, 0.46 car wheel, 0.38 convertible, 0.06 grille, 0.01 racer, 0.01 ✅	table lamp, 0.54 lampshade, 0.40 altar, 0.00 pedestal, 0.00 four-poster, 0.00 ✅
F_3 (ResNeXt50) m_max, 77.40%	breakwater, 0.61 beacon, 0.10 seashore, 0.05 lakeside, 0.00 pier, 0.00 ❌	strainer, 0.54 measuring cup, 0.06 ladle, 0.00 wooden spoon, 0.00 Crock Pot, 0.00 ❌	car wheel, 0.59 sports car, 0.12 grille, 0.03 convertible, 0.02 beach wagon, 0.02 ❌	lampshade, 0.51 table lamp, 0.20 wall clock, 0.01 desk, 0.00 four-poster, 0.00 ❌
$F_1F_2F_3F_4$, 80.29% (DenseNet, EfficientNet-B0, ResNeXt50, Inception3)	beacon, 0.45 breakwater, 0.25 seashore, 0.20 promontory, 0.01 sandbar, 0.00 ✅	measuring cup, 0.67 strainer, 0.14 beaker, 0.01 rule, 0.00 wooden spoon, 0.00 ✅	sports car, 0.48 car wheel, 0.28 convertible, 0.05 grille, 0.04 beach wagon, 0.03 ✅	table lamp, 0.48 lampshade, 0.41 altar, 0.00 wall clock, 0.00 pedestal, 0.00 ✅

Table 6: Examples on ImageNet and Top-5 Classification Confidence

lection with the unifying FQ scores for ensemble teams of equal size S for $S=2, \dots, M-1$, this table shows that the time cost of FQ diversity computation is about $2 \sim 3 \times$ compared to the Q diversity computation, such as 89.93 seconds for CIFAR-10 by FQ diversity, which is about 1.5 minutes, compared to 27.72 seconds spent on average for the Q diversity based ensemble selection. The dominating cost in FQ-diversity based computation is the time of running complex DNN models (e.g., ImageNet) to compute the focal model based ensemble diversity on each focal model based negative sample set. The actual cost of making the ensemble selection based on the computed FQ diversity scores is relatively small. Furthermore, serving a query of whether an ensemble team with a new base model should be selected is much faster for both datasets.

Computation Time (s)	CIFAR-10	ImageNet
Q-Diversity	27.72	2480.15
FQ-Diversity	89.93	4876.84
Ensemble Query	0.25	14.29

Table 7: Computation time for Q/FQ

5. Conclusion

We have presented new ensemble diversity metrics, coined as FQ-diversity, which extend the respective Q-diversity metrics with three optimizations: (1) separately

computing the FQ-diversity scores for ensembles of equal size, (2) leveraging the concept of focal model for both negative sampling and for computing the FQ-diversity scores to better capture the failure independence among member models of an ensemble team, and (3) utilizing binary clustering with strategic centroid selection to partition candidate ensemble teams of equal size S for $S = 2, \dots, M-1$, for a given base model pool of size M , and select high quality ensemble teams below the K-means FQ-diversity threshold. In addition, we introduce the EQ-diversity as a fusion of the top three performing FQ-diversity metrics to further boost the overall ensemble accuracy of the selected ensemble teams. Extensive experiments on ten base models for each of the two datasets (CIFAR-10 and ImageNet) show that our FQ and EQ diversity metrics are effective for selecting high diversity ensemble teams and boosting overall ensemble accuracy.

Acknowledgment

This research is partially sponsored by National Science Foundation under NSF 1564097, NSF 2038029, a Cisco grant, and an IBM faculty award. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or other funding agencies and companies mentioned above.

References

- [1] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996. 1
- [2] Leo Breiman. Random forests. In *Machine Learning*, pages 5–32, 2001. 1
- [3] Leo Breiman et al. Arcing classifier (with discussion and a rejoinder by the author). *The annals of statistics*, 26(3):801–849, 1998. 1
- [4] 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)*, pages 1282–1291, 2019. 4
- [5] Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. *Statistical methods for rates and proportions*. John Wiley & Sons, 2013. 1
- [6] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. 7
- [7] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E. Hopcroft, and Kilian Q. Weinberger. Snapshot ensembles: Train 1, get M for free. *CoRR*, abs/1704.00109, 2017. 1
- [8] Cheng Ju, Aurélien Bibaut, and Mark Laan. The relative performance of ensemble methods with deep convolutional neural networks for image classification. *Journal of Applied Statistics*, 45, 04 2017. 2
- [9] Ron Kohavi and David Wolpert. Bias plus variance decomposition for zero-one loss functions. In *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning, ICML’96*, page 275–283, San Francisco, CA, USA, 1996. Morgan Kaufmann Publishers Inc. 1
- [10] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009. 2, 6
- [11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. 1, 7
- [12] Ludmila I. Kuncheva and Christopher J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach. Learn.*, 51(2):181–207, May 2003. 1, 3
- [13] L. Liu, W. Wei, K. Chow, M. Loper, E. Gurosoy, 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)*, pages 274–282, 2019. 1, 4
- [14] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276—282, 2012. 1
- [15] D. Partridge and W. Krzanowski. Software diversity: practical statistics for its measurement and exploitation. *Information and Software Technology*, 39(10):707 – 717, 1997. 1, 3
- [16] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 2, 6
- [17] Robert E. Schapire. A brief introduction to boosting. IJ-CAI’99, page 1401–1406, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. 1
- [18] David B. Skalak. The sources of increased accuracy for two proposed boosting algorithms. In *In Proc. American Association for Arti Intelligence, AAAI-96, Integrating Multiple Learned Models Workshop*, pages 120–125, 1996. 1
- [19] L. N. Smith. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 464–472, 2017. 1
- [20] M. van Erp, L. Vuurpijl, and L. Schomaker. An overview and comparison of voting methods for pattern recognition. In *Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition*, pages 195–200, 2002. 2
- [21] W. Wei and L. Liu. Robust deep learning ensemble against deception. *IEEE Transactions on Dependable and Secure Computing*, pages 1–1, 2020. 4
- [22] W. Wei, L. Liu, M. Loper, K. Chow, E. Gurosoy, S. Truex, and Y. Wu. Cross-layer strategic ensemble defense against adversarial examples. In *2020 International Conference on Computing, Networking and Communications (ICNC)*, pages 456–460, 2020. 4
- [23] Yanzhao Wu, , Wenqi Cao, Semih Sahin, and Ling Liu. Experimental Characterizations and Analysis of Deep Learning Frameworks. In *2018 IEEE 38th International Conference on Big Data*, December 2018. 1
- [24] Y. Wu, L. Liu, J. Bae, K. Chow, A. Iyengar, C. Pu, W. Wei, L. Yu, and Q. Zhang. Demystifying learning rate policies for high accuracy training of deep neural networks. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 1971–1980, 2019. 1
- [25] 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)*, pages 208–217, 2020. 1, 3
- [26] G. Udny Yule. On the association of attributes in statistics: With illustrations from the material of the childhood society, c. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 194:257–319, 1900. 1
- [27] Barret Zoph and Quoc V. Le. Neural architecture search with reinforcement learning. *CoRR*, abs/1611.01578, 2016. 1, 7