Reviewer #1

Maximizing feature map difference? Thanks for your question. The Eqs. (3) and (4) do relate to the diversity maximization among individual learners. While here we clarify that our diversity maximization mechanism has nothing to the feature map part. The feature map or say the extracted features from the raw input data is shared by all the individual learners as their input. We will emphasize this fact in our final version.

Individual model must be weak? and the necessity of ensemble for NN? Thanks for your question which gives us good opportunity to further clarify our motivation and the position of this paper. We note that one possible reason for the lack use of neural networks for ensemble in the community may be due to their strong fitting ability. This on one hand reduces the need for ensemble and on the other hand, the ensemble can in turn cause over-fitting. In this paper, we address this issue in two folds: i) we introduce weight sharing to control (reduce) the model capacity of each network leaner; ii) we introduce the diversity loss to mitigate the over-fitting and enhance the generalization ability.

Our approach allows some space to control the complexity of each leaner and the generalization ability on new dataset. Considering the readily available network modules and fast development in this area, it opens more possibility for further research opportunities. The applicability can also be justified considering the use of the GPU power.

Reviewer #2

Errorbar. The following table shows accuracy variation by increasing N. When N gets larger, the accuracy fluctuation becomes smaller. The accuracy in Table 3 in our paper is the mean of 10 trials. We will add errorbars in final version.

dataset	N=1	N=2	N=3	N=4	N=5	N=6
KDD10	-0.46 **77.14** +0.40	-0.31 **78.51** +0.30	-0.31 **78.87** +0.22	-0.22 **79.19** +0.16	-0.15 **79.02** +0.23	-0.15 **78.76** +0.09
Cifar-10	-0.47 **93.56** +0.43	-0.23 **94.01** +0.37	-0.25 **94.32** +0.26	-0.21 **94.89** +0.18	-0.16 **94.81** +0.12	-0.14 **94.66** +0.08

Table 1: Accuracy (%) by different N: (difference of lowest and mean, mean, difference of highest and mean).

Will pre-trained VGG and Resnet limit the diversity? Note that the diversity is fulfilled by the learned individual learners (i.e. their own parameters) **after the shared layers**. VGG or Resnet can be regarded as part of the weight sharing parts hence weather it is pre-trained or random initialized has nothing to do with our diversity maximization mechanism. In another word, the pretrained models serve as a feature extractor which can be application dependent.

Reviewer #3

Difference to the two previous works. The two mentioned papers are both application dependent whose design are much coupled with the specific application: defocus blur detection and network robustifying.

1) Difference to DBD-CENet: **i) Methodology.** The mentioned paper DBD-CENet is more like a boosting ensemble strategy, which also combines the idea of knowledge distillation (teacher-student network) and co-training. While our ensemble layer is more like blending and bagging method. **ii) Techniques.** DBD-CENet uses one branch to estimate the residual of the other branch iteratively. Finally, they finetune the two branches of DBD-CENet in an iterative way with every epoch (mentioned in Eq. (9) and Eq. (10) in that paper). These adhoc practices are all not used in our work.

2) Difference to adaptive diversity promoting (ADP): **i) Methodology.** ADP's diversity is fulfilled by dot product in embedding layer. Hence it is difficult to be applied on tabular data sets and regression task. While our ensemble layer gives a more universal approach to quantify diversity, and it can be trivially applied to these tasks. Besides, since ADP directly use original data as input to ensemble, its time and space overhead can increase exponentially, which also limits its applicability. **ii) Techniques.** One notable difference for handling the diversity and accuracy loss is that ADP keeps the weights of the two loss constant (see in Eq. (5) in that paper) while in our method the weights of the two loss change over time for more effective learning. We will discuss the differences more detailedly in our final version.

Reviewer #4

Number of layers for the ensemble layer? In all our experiments, we only use one layer as the ensemble layer, and we once have tested by increasing the number of its layers while no performance improvement is observed. This may suggest that the diversity can be readily realized by a moderate sized model i.e one layer network, and a more complex model can even hurt the performance. We will add comparison and discussion in final version by the 9th extra page.

Generalization test? The numbers for training and testing loss curve will be added in our final version.

Compared methods tuning? Yes, all the methods are finetuned to the best performance according to a validation set, such as max_depth, num_leaves, learning_rate, etc. This will be clarified in the final version.

Relevant literature. Thanks and the mentioned papers will be added in final version. Their difference will be discussed.