We thank the reviewers for their insightful comments. Below we respond to the comments point-by-point.

1 Response to Reviewer 1:

The improvement achieved by MMA. We appreciate the comments from the reviewer. Firstly, the difference of ours from previous researches is clearly discussed in Section 2. Secondly, we argue that the improvement over other methods is significant. For example, MMA regularization improves the TOP-1 accuracy of VGG19-BN by 1.65%, while the orthogonal regularization improves just by 0.75%. Besides, as a simple plug-in regularizer with negligible computational overhead, it is shown to be architecture-agnostic and produces consistent performance improvement on many tasks. Therefore, the key advantage of MMA regularization highlighted in this paper is not the significance but the robustness.

Combined with other regularization methods. Firstly, weight decay and data augmentation have been applied to all the classification models in our paper, as we exactly follow the same settings in the original papers proposing the models. Secondly, to further demonstrate this advantage, we perform a comparative experiment with the WideResNet-28-10 on CIFAR100: the accuracy of applying the AutoAugment [1] (SOTA data augmentation) is 82.65%, and combining the MMA regularization with AutoAugment gets a higher accuracy of 83.11%. Thirdly, the models of ArcFace and ArcFace+ have used weight decay and dropout. All these demonstrate that MMA can be combined with other regularization methods to further improve the test performance.

[1] Cubuk E D, Zoph B, Mane D, et al. Autoaugment: Learning augmentation strategies from data[C]//CVPR2019.

Source of improvement. Firstly, the claim that the lack of neuron diversity can hurt the generalization performance has been discussed in many previous work ([34, 35, 48, 20, 18, 31, 47, 19] of the paper's References), and we propose a simple and powerful method to solve this issue. Secondly, in all the comparative experiments, the hyperparameters and settings, including the random seed, are the same between models with and without MMA, so the comparisons are absolutely fair. Thirdly, as a plug-in regularizer without changing the architecture, the only effect of MMA is enhancing the neuron diversity. Therefore, we argue that the improvement of accuracy comes from increasing the neuron diversity.

Training procedure. Due to the limited space, we do not list the training details, but we illustrate that the settings follow the original papers proposing the models exactly. The batch size is set to 64 for DenseNet and 128 for other models. The learning rate is initially set to 0.1 and decayed by specific schedules. The optimizer is SGD with a momentum of 0.9. More details of the training procedure will be elaborated by a table in the final version.

2 Response to Reviewer 2:

Regularization coefficient. We appreciate the comments from the reviewer. Follow this suggestion, we report the average results of five runs with different random seeds in Figure 1 and Figure 2, and the influences become stable.

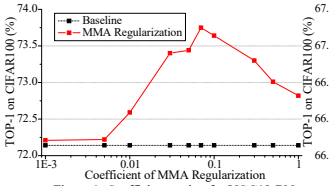

Figure 1: Coefficient tuning for VGG19-BN.

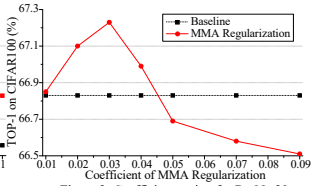

Figure 2: Coefficient tuning for ResNet20.

Training curves. Due to the limited space, we only supplement the training curves of VGG19-BN on CIFAR100 as in Figure 3. The MMA gets persistently higher TOP-1 accuracy and lower loss than the baseline. Besides, the convergence speed and stability are not changed. We will plot more training curves in the final version.

More related work. We thank the reviewer for providing some related work. "Network Decoupling" and "Speeding Up" focus on accelerating the training and the evaluation of CNN respectively. "Rethinking Depthwise" introduces the BSConv based on intra-kernel correlations, while our MMA regularization targets at decreasing the inter-kernel correlations. We will discuss the details in the final version.

3 Response to Reviewer 3:

Hyperparameters for baseline methods. We appreciate the comments from the reviewer. We set the hyperparameters to the same as the papers proposing the methods, which conduct cross-validation experiments. To be more rigorous, we agree with the reviewer and will implement cross-validation experiments by ourself in the final version.

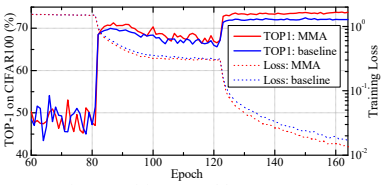

Figure 3: Training curves of the VGG19-BN.

The cases of small lambda. As the reviewer 2 stated, this is because the influences of different coefficients are unstable in a single experiment. To better clarify the influences, we report the average results of five runs with different random seeds in Figure 1 and Figure 2, in which the performance of small lambda approaches vanilla methods from above.

4 Response to Reviewer 4:

More insights. We appreciate the comments from the reviewer. Firstly, based on the Tammes problem, the uniform distribution means the minimal angle is maximized, therefore the angular diversity is boosted to the utmost extent. Secondly, many previous researches ([34, 38, 20] of the paper's References) demonstrate that large angular diversity can decrease filter correlation. Therefore, we argue that uniformly distributed filters are the way to remove filter correlation.

Comparison of MMA and orthogonal regularization. Firstly, MMA focuses on the uniform distribution, while orthogonal regularization pursues orthogonality for all the pairwise vectors. Secondly, the gradient analysis of orthogonal regularization is similar to the one of cosine loss in Eq. (5), whose gradient is not as stable as the MMA. Thirdly, Figure 1 of the supplementary material demonstrates that MMA gets the largest minimal pairwise angle across all layers, and therefore the most diverse weight vectors. Fourthly, as discussed in [20] of the paper's References, orthogonal regularization tends to group neurons closer, especially when the number of neurons is greater than the dimension.

The direct effect on features. For output layer, MMA aims to maximize the distances between classifier neurons. Each classifier neuron plays the role of ground truth centre of features falling into the corresponding category. Therefore, MMA regularization can maximize the inter-class feature separability. ArcFace can enhance the intra-class compactness explicitly and inter-class discrepancy implicitly, while MMA regularization only focuses on the inter-class separability. So, it is not fair to compare MMA alone with ArcFace. However, the MMA regularization can further boost the accuracy of ArcFace and NormFace. We will investigate the effects of MMA on other feature learning methods and tasks.