Data-free Model Fusion with Generator Assistants Supplementary Material

Luyao Shi Prashanth Vijayaraghavan Ehsan Degan
IBM Research, Almaden Research Center, San Jose, CA, USA
luyao.shi@ibm.com prashanthv@ibm.com edehgha@us.ibm.com

1. Model Accuracy over Epochs

Fig. 1 shows the fused model (student) accuracy over training epochs for different methods on the CIFAR datasets. Only one training session is shown for each method for better illustration. Our DFMF not only obtains the highest model accuracy but also converges much faster than the other generative methods. Note that the results of non-generative approaches CMI and CMI (w/ GA) are also shown here as reference, but they should not be compared with the generative approaches directly regarding convergence speed as no warmup training is used here.

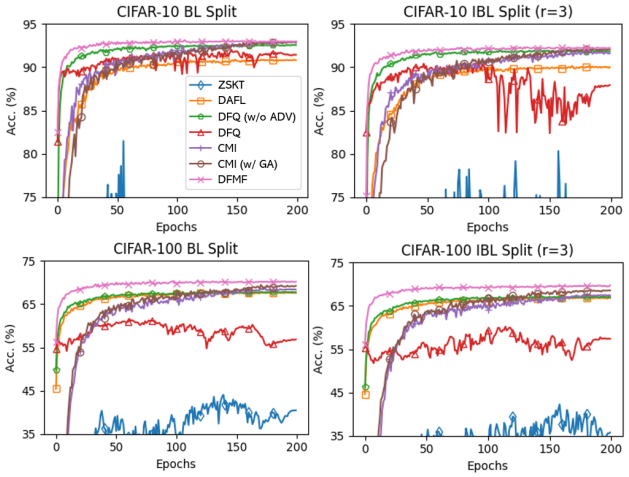

Figure 1. Examples of student model accuracy vs. training epochs for different methods.

2. Model Accuracy on Class-based Subsets

In the Stanford Dogs study, we further split the testing dataset into multiple subsets by class, and compared model accuracy on different testing subsets. The results for the two-party and four-party scenarios are shown in Fig. 2 and Fig. 3, respectively. For DFMF and CMI (w/ GA), each model was trained three times and the mean results are displayed. In the balanced data split (BL) studies, the teachers perform almost equally well on all the class subsets. In the imbalanced data split (IBL) studies, since the teachers were specialized at certain classes, they only perform well on the classes that are the frequent classes of their corresponding training data splits, but perform poorly on the other classes. The performance asymmetry is more significant as the imbalance ratio r increases. The gold standard (one model trained on the whole training set without data split) and teacher ensembles are generally good at all the classes. Our proposed DFMF obtains results very close to the gold standard and teacher emsemble results, and substantially outperforms the compared CMI (w/ GA).

3. Sensitivity Analysis

The sensitivity analysis is not thoroughly explored in this paper. Our method is developed based on the architecture of DFQ, therefore we used most of the hyperparameters the same as those used in the original paper/codes, as they were already proven to be working well. The purpose of this study is mainly focusing on demonstrating the effectiveness of GAs in the data-free model fusion problem, for both generative approaches (e.g. DFQ) and non-generative approaches (CMI). Fine-tuning these hyperparameters is beyond the scope of this study and will be investigated in future work.

The only new hyperparameter introduced is the frequency of GA updates (M_{GA}). In our experiments we update GAs with half the frequency of the student update ($M_{GA} = 5$, $M_S = 10$) to reach a balance between computation speed and performance, as GAs only help with the generators and do not need to have the same high quality as the student. We tried to update GAs with the same frequency as the student but got similar results, as shown in Tab. 1. As a next step, we will investigate whether further reducing M_{GA} can still maintain the same performance.

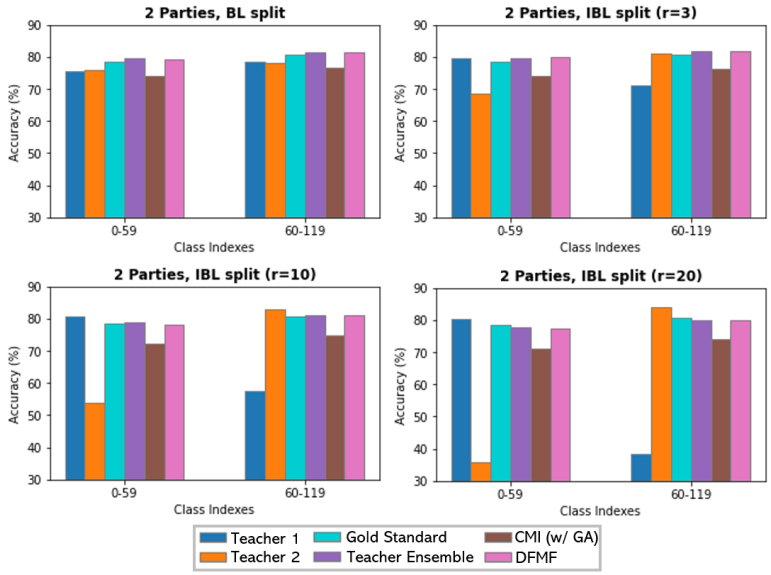

Figure 2. Comparison of model accuracy on class-based subsets of testing data in the two-party scenario.

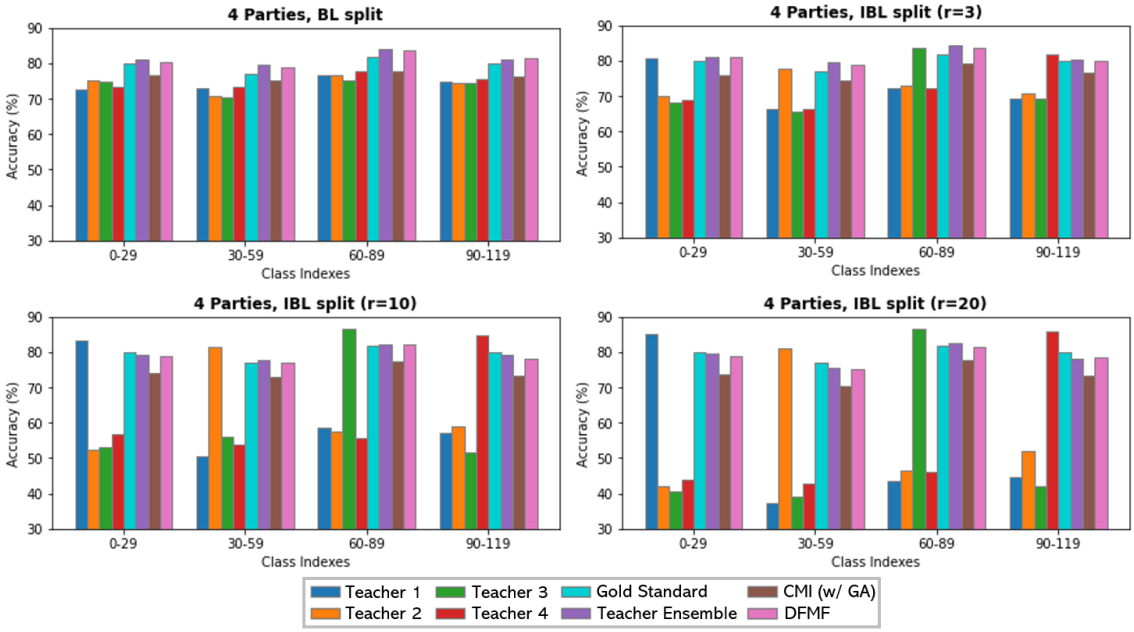

Figure 3. Comparison of model accuracy on class-based subsets of testing data in the four-party scenario.

DFMF/ CIFAR-10	Accuracy (%)	
	BL split	IBL split (r=3)
$M_{GA} = 5$	93.10±0.11	92.27±0.19
$M_{GA} = 10$	93.16±0.08	92.21±0.14

Table 1. DFMF results on the CIFAR-10 dataset with 2 parties. The teachers and the student are all using the ResNet-18 architecture. Results with different M_{GA} are compared.

4. Additional Implementation Details

In addition to the implementation details provided in the main paper, we also provide some additional details regarding image augmentation here. All of our image augmentation are based on *torchvision.transforms* functions.

For training the teachers, *RandomCrop(32, padding=4)* and *RandomHorizontalFlip()* are used on CIFAR-10 and CIFAR-100 images. *RandomResizedCrop(224)* and *Ran-*

domHorizontalFlip() are used on the Stanford Dogs images.

For training the student with knowledge distillation using the generated images, the following augmentations are applied on the generated images: *RandomRotation(degrees=10)*, *RandomResizedCrop(*img_size*, scale = (0.8, 1.0))*, *RandomHorizontalFlip(p=0.5)* and *ColorJitter(brightness=0.1, contrast=0.1, saturation=0.1, hue=0.1)* with a probability of 0.2, where img_size is 32 for CIFAR studies and 224 for Stanford Dogs studies. Image augmentation is not applied when training the GAs using the generated images.