Data-free Model Fusion with Generator Assistants

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

Abstract

In a Model Marketplace, multiple parties may submit pretrained neural networks that accomplish similar tasks. These networks usually have different architectures and are trained on different datasets. It would be significantly beneficial to fuse all these models into a single model with superior performance and lower cost of execution. However, the training parties may be unwilling to share their training data, due to data privacy and confidentiality, and may not agree to participate in a federated learning collaboration. As an alternative, we propose a data-free model fusion framework, based on knowledge distillation, to combine several pretrained models into a superior model without the need for the raw training data. We employ a generative approach to synthesize data for knowledge distillation. The data generator needs to be trained to produce a diverse set of samples that have a similar distribution to that of the training data. Generating samples that cause student-teacher disagreement can expand the coverage of the data distribution, reduce chance of mode collapse and, improve data-free knowledge distillation. However, we found that in a multi-teacher setting, encouraging disagreements between the teachers and the student causes confusion for the generators and deteriorates the results. To tackle this, we introduce Generator Assistants (GA), which keep the generators evolving without causing confusion. Experiments on CIFAR-10, CIFAR-100 and Stanford Dogs datasets show that our method greatly improves the data-free model fusion performance compared to the prior art.

1. Introduction

Recent years have witnessed an unprecedented development and usage of deep Convolutional Neural Networks (CNNs) that achieved state-of-the-art results in various computer vision tasks, including classification and segmentation. In order to achieve generalizable models, a diverse set of data from various sources is required. However, due to privacy and security concerns, data sets from various sources often cannot be moved to a centralized database for training a model. Therefore, distributed learning mechanisms such a federated learning and model fusion have gained a lot of attention. A related scenario is a Model Marketplace, where multiple parties can submit their pretrained AI models to be used by other users for inference. In a Model Marketplace it is common to have multiple pretrained models with different architectures, trained by different parties using different datasets that perform a similar task. If users want to use these models for inference, they should either choose one of the models without knowing which model would perform the best on their data or use multiple models in an ensemble [20, 21] and incur higher cost. Model ensembling is also less interpretable compared to a single model. It would be greatly beneficial if the marketplace owner could fuse all these models into a single model to provide the users with superior performance and lower cost of execution. As mentioned, it might be impossible to obtain the training data of each model due to privacy and confidentiality issues. In addition, the training parties may not agree to participate in a federated learning collaboration [12], where they need to retrain a model, repeatedly. As a practical solution, we propose a data-free model fusion framework, based on knowledge distillation, that can be used to combine pretrained models into one superior model in terms of performance (accuracy, robustness, scalability) and infrastructure usage (time, memory, number of GPUs), without the need for the raw training data.

We propose to treat this problem as a Data-free Knowledge Distillation (DFKD) with multiple teachers, where synthesized samples are used for distillation. We consider a general case in which the teachers are trained on different datasets and may have different architectures. For data synthesis, we employ a generative approach where we train a dedicated generator for each teacher. Each generator needs to synthesize samples that have a distribution similar to that of the original training data of its corresponding teacher (i.e. have similar statistics to that of the training data). The samples should also be diverse to sufficiently cover the training distribution and avoid mode collapse. Several loss functions have been proposed to satisfy these conditions. Among them, the adversarial (ADV)

loss [3, 5, 25] is especially effective in producing statistically diverse samples for single-teacher knowledge distillation. The ADV loss encourages the generator to synthesize images that cause student-teacher disagreement, which iteratively expands the distributional coverage of the samples during training. However, expanding the ADV loss to a multi-teacher setting is not a trivial problem. Our experiments show that rewarding dissimilarity between multiple teachers and one student causes confusion for the generators and deteriorates the results. Without the adversarial loss, the generator does not expand the distributional coverage of the generated images which results in sub-optimal performance. We address this problem by introducing Generator Assistants (GA), and extend the application of adversarial loss to multi-teacher settings. Each Generator Assistant is a dedicated model to a teacher-generator pair and is trained to mimic a teacher using samples synthesized by its corresponding generator. Therefore, it can be used to calculate a meaningful adversarial loss to train a generator to produce diverse samples and mitigate mode-collapse.

We summarize our novelty and contribution here: 1) Although DFKD has been well studied, Data-free Model Fusion (DFMF) with multi-talent teachers and different architectures are not well explored in the literature. Through extensive experiments, we illustrate the superior performance of our DFMF over several compared methods, and the fused model attains greater accuracy when compared to individual teacher networks; 2) Our main novel contribution is the introduction of Generator Assistant (GA) to address the challenge that uniquely exists in data-free model fusion.

2. Related Work

Knowledge Distillation (KD) [9] was originally proposed to learn a compact student model from a large teacher model. Instead of learning directly from labeled data, the student can learn from the soft labels generated by the teacher on unlabeled data. Later, training a student using multiple teachers was also proposed where the student either learns from the ensemble of the teachers' outputs [14, 19, 22, 27, 28] or the output of the most confident one [23]. To improve learning efficiency and effectiveness, most of these methods also proposed various losses to incorporate learning from the intermediate features, in addition to the final soft labels. In [14, 19, 27, 28] the teachers are trained on the same dataset, while in [22, 23] the teachers are trained on different ones (multi-talent teachers). Although these methods do not require labeled datasets, they all use real data for knowledge distillation. In fact, most of these methods used the original training data of the teachers [14, 19, 27, 28]. As discussed before, in many applications, the raw training data are not accessible for model fusion. It should be mentioned that using out-of-domain or even same-domain but out-of-distribution data as replacement for original data during KD usually leads to unsatisfactory results. Therefore, data-free model fusion using synthetic samples is the only practical solution in many cases.

Data-free model fusion can be achieved via Data-free Knowledge Distillation (DFKD). These methods synthesize samples for KD using: (1) the non-generative approaches [5, 6, 25] that produce data samples, batch-by-batch, based on model inversion, or (2) the generative approaches [2, 3, 7, 16, 17, 26] that train a generative model for synthesis. DeepInversion [25] "inverts" a trained network (teacher) to synthesize class-conditional images starting from random noise. A BatchNorm feature distribution regularization loss (BN loss) was proposed to optimize the input while regularizing the distribution of intermediate feature maps using information stored in the batch normalization layers of the teacher. An iterative competition scheme using an adversarial (ADV) loss was also proposed to encourage the synthesized images to cause student-teacher disagreement and improve sample diversity. CMI [5] improves data diversity using a contrastive learning objective that encourages the newly synthesized instances to be distinguishable from the ones synthesized in previous batches. Although impressive image synthesis and KD results were obtained, these methods suffer from long image synthesis time, making them less practical. An alternative line of work focuses on training a generator that can synthesize samples with faster speed. In DAFL [2], a generator was trained to take random noise as input and generate images that can produce strong one-hot predictions with the teacher classifier. KEGNET [26] is similar to DAFL but used a class-conditional generator. They further introduced a decoder to recover the input noise from the generated images, and preventing the generator from converging to a naïve solution, with a collapsed mode. ZSKT [17] trains an adversarial generator to search for images on which the student poorly matches the teacher. Han *et al*. [7] attempted to increase the diversity of the generated images by proposing diversity seeking regularization. Since these works mainly rely on one-hot prediction of the teacher classifier as regularization, the generated samples do not necessarily follow the same distribution as the original training data and could result in degraded KD performance. To alleviate this problem, in [3, 16], the authors introduced a batch normalization (BN) loss to train generators that are able to produce higher-quality and more realistic images. In [16], multiple class-conditional generators were trained for each class to address the mode-collapse problem. However, this method does not scale well when there are many classes. In DFQ [3], class-unconditional generators were trained that scale well with the number of classes. ADV loss was used to iteratively improve the generators. Variational Information Distillation (VID) loss [1] was also used to improve student learning efficiency and effectiveness by matching intermediate

layer outputs the student and the teacher. The latest work in DFKD is FastDFKD [6], which learns a meta-synthesizer that seeks common features as the initialization for the fast data synthesis and significantly accelerate DFKD.

There are studies exploring data-free model fusion using other approaches. In [24], the authors constructed group-stack generative adversarial networks with dual generators. The target network is obtained by regrouping the trained dual part generator. Although multiple teachers were present, only one generator was trained. Additionally, BN regularization was not used for training the dual generators. One limitation of this approach is that the teacher and target network should have similar architecture with the same number of block numbers. In their study, the same architecture is used for all the networks.

3. Methods

Let T_{θ_k} and θ_k represent the k-th teacher and its learnable parameters and K be the number of teachers. The pretrained teacher T_{θ^*}, with parameters θ^*, is trained on its corresponding labeled dataset and will not be updated during model fusion. Let S_ϕ be the student with learnable parameters ϕ. For simplicity, we assume that all the teachers and the student are trained for the same task of C-category classification. Since the teachers are pretrained on different datasets, for each teacher we train a corresponding generator network G_{ψ_k} to synthesize samples with a distribution similar to that of the teacher's training data. To help with the process, for each teacher, we train a generator assistant (GA) network A_{χ_k}, with learnable parameters χ_k. A diagram of our data-free model fusion scheme is given in Fig. 1.

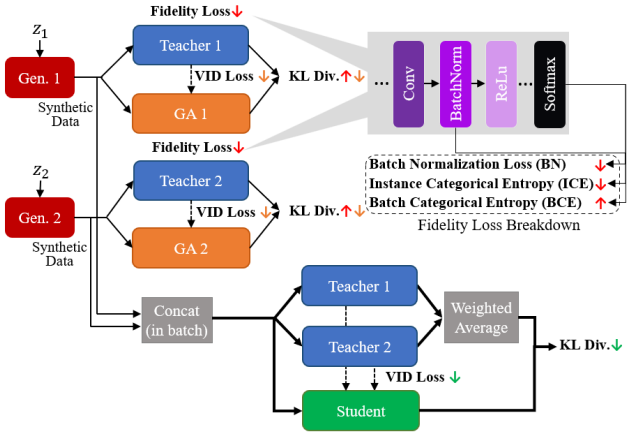

Figure 1. A diagram of our DFMF scheme. Only 2 teachers are included in this example. The colored arrows indicate which networks the loss functions are trying to minimize or maximize.

3.1. Training the Student

The student S_ϕ learns from multiple teachers through KD using the synthetic samples produced by all of the generators. The student is updated by minimizing the loss below:

$$L_\phi^S = \mathbb{E}_x\Big[\mathcal{D}\big(\sum_{k=1}^{K} w_k T_{\theta_k^*}(x), S_\phi(x)\big) + \beta \sum_{k=1}^{K} w_k L_{\phi,\theta_k^*}^{\mathrm{VID}}(x)\Big],$$
$$x \in \bigcup_{k=1}^{K} G_{\psi_k}(z) \tag{1}$$

In this equation, x is an image generated by one of the generators, $T_{\theta_k^*}(x)$ and $S_\phi(x)$ are the soft labels produced by the k-th teacher and the student, respectively where a softmax with temperature τ is applied to the teachers and the students outputs. $\mathbb{E}$ represents expected value and $\mathcal{D}$ stands for Kullback-Leibler (KL) divergence. In addition to the KL divergence loss, we also use variational information distillation (VID) to match intermediate layer outputs between the teachers and the student. $L_{\phi,\theta^*}^{\mathrm{VID}}$ is the VID loss which is one of the state-of-the-art KD variants that formulates knowledge transfer as maximizing the mutual information between the teacher and the student networks, which yields better student accuracy with faster convergence. Details about VID loss can be found at [1].

Additionally, β is a weighting factor for the VID loss, w_k is the weight to control the importance of the k-th teacher in the learning process and $\sum_{k=1}^{K} w_k = 1$. z is the random input to a generator and $p(z)$ is its Gaussian probability distribution. $G_{\psi_k}(z)$ is the image produced by the k-th generator. Here we use the union of all the generators' outputs as the input images x for KD. This means that an image produced by any of the generators goes through all the teachers during the KD process.

3.2. Training the GAs

Each GA A_{χ_k} learns from its matched teacher $T_{\theta_k^*}$ through knowledge distillation using the synthetic samples generated by its corresponding generator G_{ψ_k}. Each GA is updated by minimizing the following loss:

$$L_{\chi_k}^A = \mathbb{E}_{x_k}[\mathcal{D}(T_{\theta_k^*}(x_k), A_{\chi_k}(x_k)) + \beta L_{\chi_k,\theta_k^*}^{\mathrm{VID}}(x_k)], \tag{2}$$

where x_k is an image produced by the k-th generator and $A_{\chi_k}(x_k)$ is the output of the k-th GA, where a softmax with temperature τ is applied to the teacher and GA's outputs.

3.3. Training the Generators

The loss function to update the generators is comprised of two parts: fidelity loss and adversarial loss. The fidelity loss encourages each generator to synthesize samples that are similar to the training data of its corresponding teacher.

The fidelity loss is further comprised of Batch Normalization (BN) loss, Instance Categorical Entropy (ICE) loss and Batch Categorical Entropy (BCE) loss [3] as in:

$$L^F_{\psi_k} = \mathbb{E}_{p(z)}\Big[\sum_l \mathcal{D}_{\mathcal{N}}((\hat{\mu}_{l,\psi_k}(z), \hat{\sigma}^2_{l,\psi_k}(z)), (\mu_l, \sigma^2_l))\Big] + \mathbb{E}_{p(z)}[H(T_{\theta^*_k}(G_{\psi_k}(z)))] - H(\mathbb{E}_{p(z)}[T_{\theta^*_k}(G_{\psi_k}(z))]). \quad (3)$$

The first term in Eq. (3) is the BN loss which encourages the generator to produce samples that result in features with the same mean and variance values as those stored in the BN layers of the teacher. Here, μ_l and σ^2_l are the mean and variance values stored in the l-th normalization layer and, hence, are learned from the original training data. $\hat{\mu}_{l,\psi_k}(z)$ and $\hat{\sigma}^2_{l,\psi_k}(z)$ are the corresponding mean and variance computed over the synthesized samples $G_{\psi_k}(z)$. We assume the distributions are Gaussian and use the Kullback-Leibler (KL) divergence between two Gaussian distributions as:

$$\mathcal{D}_{\mathcal{N}}((\hat{\mu}, \hat{\sigma}^2), (\mu, \sigma^2)) = \frac{(\hat{\mu} - \mu)^2 + \hat{\sigma}^2}{2\sigma^2} - log\frac{\hat{\sigma}}{\sigma} - \frac{1}{2}. \quad (4)$$

The second term in Eq. (3) minimizes the instance categorical entropy, denoted by $H(\cdot)$. Assuming the teacher is well trained for accurate classification, a good generator should produce samples that yield low entropy outputs from the teacher (the probability for one category should be high). The third term maximizes the batch categorical entropy. Because there is no prior knowledge for the categorical probability distribution of the original training data, it is reasonable to assume the classes appearing in the dataset follow uniform distribution. This prevents the generator from producing samples of only one or a few classes, and is achieved by maximizing the entropy of the teacher's averaged outputs over any batch.

The adversarial loss encourages a generator to produce samples that cause disagreement between each teacher and its corresponding GA and hence, avoid producing repeated or similar samples. This expands the distribution of the generated images and helps with mitigating the mode collapse problem. The adversarial loss is given by:

$$L^A_{\psi_k} = -\mathbb{E}_{p(z)}[\mathcal{D}(T_{\theta^*_k}(G_{\psi_k}(z)), A_{\chi_k}(G_{\psi_k}(z)))] \quad (5)$$

The overall loss to train a generator is given by:

$$L^G_{\psi_k} = L^A_{\psi_k} + \alpha L^F_{\psi_k}, \quad (6)$$

where $\alpha \geq 0$ is a weighting factor.

3.4. Implementation

Our proposed data-free model fusion scheme is summarized in Algorithm 1. z^B denotes the random input batch of size B to the generators, and $L(z^B)$ denotes the loss averaged over the batch. As suggested in [3], we perform warm-up training for the generators using only the fidelity loss in Eq. (3). The pre-training procedure reduces generation of unreliable samples in the early phase. The number of epochs for warm-up and main training are $N_{\text{warm-up}}$ and N, respectively. M is the overall number of iterations per epoch. We update the student more frequently than the generators to reduce the chance of falling into local minima, which is controlled by M_S. By adjusting M_{GA}, we also update GAs less frequently than the student to balance between speed and performance, as GAs only help the generators and do not need to be as high-quality as the student.

Algorithm 1 Data-free model fusion

Input: Pretrained teachers $\{T_{\theta^*_k}\}_{k=1}^K$, randomly initialized student S_ϕ
Output: An optimized student S_{ϕ^*}

Randomly initialize generators $\{G_{\psi_k}\}_{k=1}^K$ and GAs $\{A_{\chi_k}\}_{k=1}^K$
$b \leftarrow \lfloor B/K \rfloor$
for k: 1 to K **do** ▷ Warm-up training for generators
  for n: 1 to $N_{\text{warm-up}}$ **do**
    for m: 1 to M **do**
      $z^B \leftarrow [\mathcal{N}(0, I)]^B$
      $\psi_k \leftarrow \psi_k - \eta_G \nabla_{\psi_k} L^F_{\psi_k}(z^B)$
    end for
  end for
end for
for n: 1 to N **do**
  for m: 1 to M **do**
    for k: 1 to K **do** ▷ Update generators
      $z^B \leftarrow [\mathcal{N}(0, I)]^B$
      $\psi_k \leftarrow \psi_k - \eta_G \nabla_{\psi_k} L^F_{\psi_k}(z^B)$
    end for
    for k: 1 to K **do**
      for m: 1 to M_{GA} **do** ▷ Update GAs
        $z^B \leftarrow [\mathcal{N}(0, I)]^B$
        $\chi_k \leftarrow \chi_k - \eta_A \nabla_{\chi_k} L^A_{\chi_k}(z^B)$
      end for
    end for
    for m: 1 to M_S **do** ▷ Update student
      for k: 1 to K **do**
        $z^b \leftarrow [\mathcal{N}(0, I)]^b$
        $x^{bk} \leftarrow \text{concatenate}(x^{b(k-1)}, G_{\psi_k}(z^b))$
      end for
      $\phi \leftarrow \phi - \eta_S \nabla_\phi L^S_\phi(x^{bK})$
    end for
  end for
end for
$\phi^* \leftarrow \phi$

4. Experiments and Results

4.1. Experimental Settings

Datasets. We evaluated our method on CIFAR-10 (10 categories), CIFAR-100 (100 categories) [13] and Stanford Dogs (120 categories) [10] datasets to demonstrate the efficacy of our method in fusing models with different number of output labels and input image resolution. Each CIFAR dataset contains 50,000 training samples and 10,000 testing samples with image resolution of 32 × 32. Stanford

Dogs consists of 12,000 training samples and 8,580 testing samples, with image resolution of 200 $\times$ 200 or larger that are resized to 224 $\times$ 224 in our study [1]. The training samples are only used to train the teacher networks, and the testing data are used to evaluate model accuracy. The CIFAR datasets are used to investigate two-party ($K = 2$) scenario, whereas the Stanford Dogs dataset is used to investigate both two-party and four-party ($K = 2$ and $K = 4$) scenarios. We also investigated balanced (BL) and imbalanced (IBL) data split modes. In the BL data split, the dataset is evenly split between multiple parties hence, class distribution imbalance between parties is unlikely (roughly same number of samples per party for each class). The IBL mode simulates imbalanced class distribution where each pretrained teacher is better at some classes than the other ones. For each party, $1/K$ of all the classes are selected as the frequent classes and the remaining classes are the infrequent classes. For each selected frequent class, p samples are selected for this party and the remaining samples (used as infrequent class samples for the other parties) are evenly split between the other $K - 1$ parties so that each one contains q samples. We investigated different imbalance ratios where the imbalance ratio r is defined as $r = p/q$. This way we make sure that there is no data overlap for each party nor missing data for the whole training set. For example, when $K = 2$ and $r = 3$, the first party is composed of 75% of the data for the first half of the classes and 25% of the data for the last half of the classes ($r = 75\%/25\% = 3$), whereas the second part keeps the rest of the data. When $K = 4$ and $r = 10$, the first party is composed of 76.9% of the data for the first quarter of the classes and 7.69% of the data for the rest of the classes; the second party is composed of 76.9% of the data for the second quarter of the classes and 7.69% of the data for the rest of the classes, etc. In this paper, we use $r = 3$ in the CIFAR studies. For the Stanford Dogs study, we use different values for $r \in \{3, 10, 20\}$.

Compared methods. Since there are limited studies exploring data-free model fusion, we scaled up various DFKD methods from a single-teacher context to a multi-teacher scenario, and compared our method with them. Our baselines include generative DFKD methods such as ZSKT [17], DAFL [2] and DFQ [3], and also the state-of-the-art non-generative method CMI [5]. We notice that when we apply the state-of-the-art generative DFKD method, DFQ, to the multi-teacher setting, the ADV loss causes confusion and leads to worse results. Therefore, we also compare to DFQ without ADV loss (w/o ADV). Since the non-generative method CMI also uses the ADV loss, we took one step further and applied GAs to CMI (w/ GA) as well for comparison. Tab. 1 shows the usage of different loss functions and GAs for each method. All the loss functions can be found in Sec. 3 except for the activation loss (ACT) used in DAFL and the contrastive learning loss (CR) used in CMI, which can be found in [2] and [5], respectively. Note that DFQ is essentially the ablation study of our DFMF without GA. For fair comparison, VID loss is used in all the generative methods, even though it is not originally used by ZSKT and DAFL. For all the methods, we combine all the generators' synthesized samples as a batch for KD, as described in Sec. 3.1.

Method	BN	ICE	BCE	ADV	ACT	CR	GA
ZSKT	-	-	-	✓	-	-	-
DAFL	-	✓	✓	-	✓	-	-
DFQ (w/o ADV)	✓	✓	✓	-	-	-	-
DFQ	✓	✓	✓	✓	-	-	-
CMI	✓	✓	-	✓	-	✓	-
CMI (w/ GA)	✓	✓	-	✓	-	✓	✓
DFMF (ours)	✓	✓	✓	✓	-	-	✓

Table 1. Comparison of different methods based on usage of various loss functions and generator assistants.

Implementation details. We use ResNet-18, ResNet-34 and ResNet-50 [8] for our teachers and students. For our generators, we use the same architecture as in [3] which is one fully connected layer followed by three upsampling (nearest neighbor interpolation) plus convolutional layers. For pre-training the teacher networks, we use Nesterov accelerated gradient [18] with weight decay and the momentum set to 5×10^{-4} and 0.9, respectively, as suggested by [2]. We train the teacher networks for 200 epochs. For the CIFAR datasets, the teachers are trained from scratch with a batch size of 256. The learning rate is initialized at 0.1 and divided by 10 at epochs 80 and 120. Stanford Dogs dataset has only 100 images per class in its training set. Training each teacher from scratch using only a portion of this data set results in poor performance. To improve performance, we use ImageNet [4] pretrained ResNet weights as initialization before training the teacher models on the Stanford Dogs samples. A batch size of 64 is used. The learning rate is initialized at 0.01 and divided by 10 at epochs 50, 100 and 150. For training the generators, we use the Adam optimizer [11] with learning rate of $\eta_G = 10^{-3}$ and momentum of 0.5. Similar to the teachers, the GAs and the student are initialized with ImageNet pretrained weights before training on the Stanford Dogs data set, but are trained from scratch for CIFAR studies. We use Nesterov accelerated gradient with cosine decaying [15] learning rate of $\eta_A = \eta_S = 0.05$ and momentum 0.9. Random rotation, crop, flipping and color jitter (details provided in Supplementary Material) are used on the generated images for data augmentation. The original VID loss [1] was proposed to match every intermediate layer between the teacher and the student. In this study, we only match the last convolutional layer outputs between the

[1]Due to our organization's internal policies we cannot use ImageNet for evaluation. We use the Stanford Dogs dataset as an alternative for a more challenging dataset than the CIFAR datasets.

Dataset	Method	Accuracy (%)	
		BL split	IBL split (r=3)
CIFAR10	Gold Standard	94.92	
	Teachers	92.41/92.26	90.43/89.79
	Teacher Ensemble	93.85	93.12
	ZSKT	56.34±11.61	63.44±16.36
	DAFL	89.04±2.87	89.76±0.24
	DFQ (w/o ADV)	92.66±0.10	91.81±0.09
	DFQ	90.64±1.47	88.12±2.42
	CMI	92.89±0.07	91.74±0.11
	CMI (w/ GA)	92.99±0.06	92.02±0.12
	DFMF (ours)	**93.10±0.11**	**92.27±0.19**
CIFAR100	Gold Standard	76.85	
	Teachers	67.9/68.6	65.22/65.48
	Teacher Ensemble	73.04	73.3
	ZSKT	41.13±0.77	36.04±6.60
	DAFL	67.46±0.15	63.98±4.69
	DFQ (w/o ADV)	67.75±0.21	66.97±0.04
	DFQ	57.73±2.87	57.26±2.06
	CMI	68.46±0.24	67.57±0.18
	CMI (w/ GA)	69.10±0.37	68.32±0.25
	DFMF (ours)	**70.05±0.13**	**69.67±0.09**

Table 2. Data-free model fusion results with homogeneous networks on CIFAR-10 and CIFAR-100 datasets. The two teachers and the student are all using the ResNet-18 architecture.

teacher and the student (or GA) as these models can have different number of layers. Here, we use uniform weights of $w_k = 1/K$ in Eq. (1). The remaining hyperparameters for our DFMF are: $N_{\text{warm-up}} = 50$, $N = 200$, $M = 400$, $M_{GA} = 5$, $M_S = 10$, $B = 256$, $\alpha = 0.1$, $\beta = 1.0$ and $\tau = 3.0$. For the compared generative methods (ZSKT, DAFL, DFQ) listed in Sec. 3, we used the same parameters $N_{\text{warm-up}}$, N, M, M_S and B for warm-up and knowledge distillation. The weights to balance different loss function terms are equal to the ones described in their original papers. Note that for DFQ (w/o ADV), we found that pre-training the generators sufficiently and keeping them frozen during KD performs better as compared to updating the generators iteratively. To have a fair comparison with the same number of generator updates, for DFQ (w/o ADV) we use $N_{\text{warm-up}} = 250$ in the warm-up training stage and keep the generators frozen later on. For CMI, we used the hyperparameters in the authors' GitHub implementation [6]. For CMI with GAs, we used 200 iterations to update the GAs and 400 iterations to update the student in each epoch. For each method, we trained all the models three times and reported the results as mean ± (sample) standard deviation.

4.2. Performance on CIFAR-10 and CIFAR-100

Homogeneous Network Results. In this section, the teachers, GAs and the student all use ResNet-18 architecture. DFMF is compared with other methods and the results are shown in Tab. 2. Gold Standard is the result of a ResNet-18 network trained on all the training data without data split.

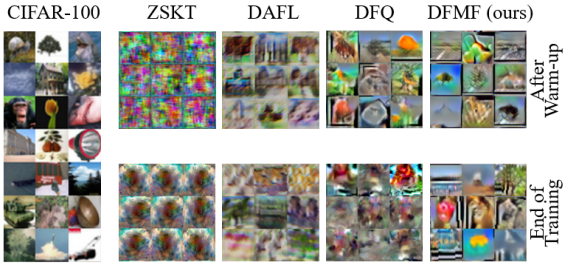

Figure 2. Example images produced by one of the generators from different methods in the CIFAR-100 'IBL split (r=3)' experiment. Sample images from generators after warm-up training and at the end of the training process are compared.

Two teachers are trained on half of the training data each. Teacher Ensemble is the ensemble of the two teachers.

Our proposed DFMF with GAs produces the best results compared to the other data free fusion methods. The fused model accuracy is also substantially higher than either of the teachers. Although an ensemble shows higher accuracy than a fused model, it is more expensive in terms of inference time and computational resources. Among all the compared methods, ZSKT performs the worst. This is expected because ZSKT only uses ADV loss to update the generators. Although using the ADV loss alone works fine in the single-teacher setting, introducing disagreements between multiple teachers and one student causes confusion for the generators and worsens the results. DAFL performs better than ZSKT, but its performance is limited due to the lack of the BN loss. Without using GAs, DFQ also performs poorly with the ADV loss. DFQ (w/o ADV) performs much better than DFQ, however its performance is still limited as the generator will no longer expand the distributional coverage of the generated images. CMI performs better than the other generative baseline methods, and adding the GAs further improves its performance. This suggests that GAs can help the non-generative methods as well. Note that the performance improvement is more substantial on CIFAR-100 compared to CIFAR-10. We also plot the fused model accuracy over training epochs for different methods in the Supplementary Material, where we show that DFMF training is more stable and converges faster than the other generative methods.

Fig. 2 shows a few example images produced by one of the generators from different generative methods in the CIFAR-100 'IBL split (r=3)' experiment. Sample images from generators after warm-up training and at the end of the training process are compared (CMI samples are not shown here as CMI is a non-generative method and does not fit into this paradigm). It can be observed that both ZSKT and DAFL produce unrealistic images due to incomplete fidelity losses. DFQ generator produces good images after warm-up training without the ADV loss, but the im-

Method	Accuracy (%)	
	BL split	IBL split (r=3)
Teachers 1 (ResNet18)	67.90	65.22
Teachers 2 (ResNet34)	67.80	61.50
Teacher Ensemble	73.15	71.70
DFMF (GA=T)	**71.16±0.08**	**69.05±0.22**
DFMF (GA=S)	70.50±0.68	67.54±1.33

Table 3. Data-free model fusion results with heterogeneous networks on the CIFAR-100 dataset. The two teachers are using ResNet-18 and ResNet-34, and the student is using ResNet-50.

age quality deteriorates towards the end of training because of the ADV loss limitation in the multi-teacher setting. Our DFMF generators manage to produce high-quality synthetic images throughout the whole training process.

Heterogeneous Network Results. In real-world applications, it is not uncommon for the pretrained teacher networks to have different architectures. Depending on our needs and available computational resources, we might need a large and powerful or a compressed student network. Therefore, we also investigate the performance of DFMF under the heterogeneous network scenario, where the teachers and the student have different network architectures. In such a setting, we can use GAs that have a network architecture similar to their corresponding teacher (GA=T) or to the student (GA=S). We conducted experiments to investigate the performance of the student in each of these scenarios. To this end, we use ResNet-18 and ResNet-34 for the two teachers and ResNet-50 for the student. We run DFMF (GA=T) and DFMF (GA=S) on CIFAR-100 dataset three times and show the mean ± (sample) standard deviation results. Tab. 3 suggests that GAs with similar architectures to their corresponding teachers result in improved performance. Similar to the homogeneous case, the fused model using DFMF outperforms each of the pretrained teachers.

4.3. Performance on Stanford Dogs

In this section, ResNet-18 is used for all the teachers, GAs and the student. The DFMF performance is shown in Tab. 4. It can be seen that on a more challenging dataset with more classes, higher image resolution, and a larger number of parties (teachers), our DFMF still produces outstanding results significantly outperforming the teachers and CMI (w/ GA). In BL and IBL studies with a mild imbalance ratio ($r = 3$), our DFMF outperforms the gold standard. As we further increase the imbalance ratio ($r = 10$ and $r = 20$), DFMF still produces results close to the gold standard and ensemble results. More detailed analysis regarding the model accuracy on different class-based subsets are provided in the Supplementary Material, where we show that the teacher models are only specialized at certain classes, whereas our DFMF, like the gold standard and ensemble models, are good at all the classes. Fig. 3 shows that DFMF can still produce high-quality synthetic images when the images have higher resolution.

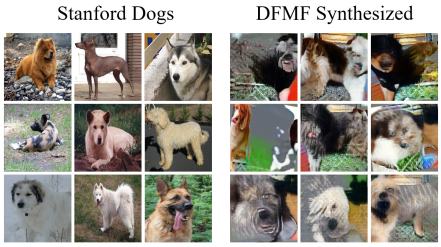

Figure 3. Example images produced by one of the DFMF generators in the Dogs 'IBL split (r=3)' experiment with 2 parties, as compared to the original Stanford Dogs images.

5. Discussion

Our experiments showed that, in a multi-teacher setting without generator assistants, using an adversarial loss to train the generators deteriorates the results. This was demonstrated by lower performance of DFQ with adversarial loss in Tab. 2. In a single-teacher KD scenario, after some epochs, the student learns to follow the teacher very closely on samples similar to the ones that have been used for KD up to that point. Therefore, dissimilarity between the outputs of the teacher and the student for an input sample is an indicator of the sample novelty for the student. As a result, we can use adversarial loss to encourage the generator to produce samples that cause dissimilarity between the student and the teacher, and hence, produce a more diverse set of samples. This will improve the coverage of the original training data distribution and decrease chances of mode collapse. However, in model fusion as a multi-teacher KD scenario, the student cannot closely resemble all of the teachers, as the teachers' outputs may be different from each other for any given sample. Therefore, there is always some dissimilarity between each teacher and the student regardless of the novelty of the input sample. As such, encouraging the generators to produce samples that result in more student-teacher dissimilarity does not lead to better coverage of the original training data distribution or enhanced KD.

This is illustrated in in Fig. 4 where we plot the ADV loss (averaged across multiple parties) vs. training epochs for DFQ and DFMF in a multi-teacher setting. Only one training session is shown for each method for better illustration. The ADV loss between the teachers and the student for DFQ (loss_T_S_DFQ) shows abnormally (negative) large values with significant fluctuations as the student cannot resemble all of the teachers. Rewarding this dramatic dissimilarity between multiple teachers and one student causes confusion for the generators and deterio-

Party Num.	Method	Accuracy (%)			
		BL split	IBL split		
			r=3	r=10	r=20
1	Gold Standard	79.50			
2	Teachers	76.82/76.98	75.69/74.30	70.00/67.25	61.12/57.94
	Ensemble	80.56	80.71	79.90	78.81
	CMI (w/ GA)	75.60±0.09	75.11±0.04	73.34±0.01	72.65±0.22
	DFMF (ours)	**80.44±0.15**	**80.78±0.15**	**79.45±0.13**	**78.60±0.23**
4	Teachers	74.14/74.15/ 73.60/74.84	72.30/72.74/ 70.87/72.30	54.02/55.48/ 49.77/54.95	54.02/55.48/ 49.77/54.95
	Ensemble	81.33	81.08	79.45	78.76
	CMI (w/ GA)	76.22±0.13	76.11±0.20	74.33±0.09	73.64±0.28
	DFMF (ours)	**81.01±0.23**	**80.75±0.15**	**78.94±0.07**	**78.31±0.06**

Table 4. Data-free model fusion results on the Stanford Dogs dataset with 2 and 4 parties. The teachers and the student are all using the ResNet-18 architecture.

rate the results. In comparison, the ADV loss between the teachers and GAs (loss_T_A_DFMF) shows manageably small values with minimal fluctuations, which also leads to much more mild ADV loss plots between the teachers and the student (loss_T_S_DFMF, only calculated for comparison purpose but not contributing to back-propagation here) compared with the DFQ plots, suggesting more stabilized student training. This supports our approach to pair each teacher with a generator assistant to resemble a single-teacher KD scenario for each generator and bring back the benefits of the adversarial loss.

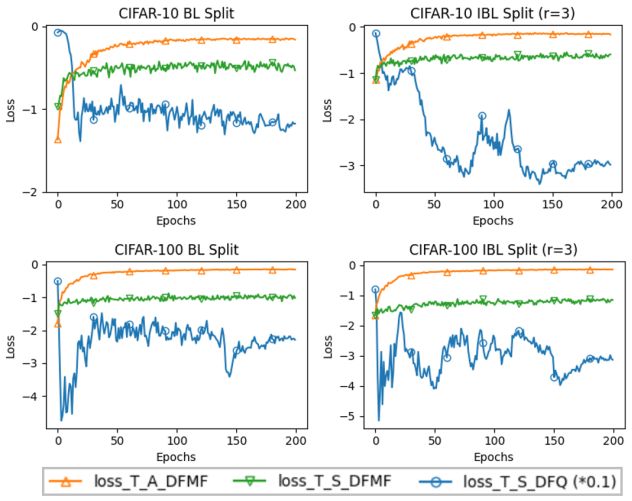

Figure 4. Examples of ADV loss vs. training epochs for DFQ and DFMF (loss_T_S_DFQ is multipled by 0.1 for better illustration).

In this study, we use the VID loss by only matching the last convolutional layer outputs between the teacher and the student (or GA). Although this might sacrifice performance slightly, it has the potential for better scalability when the teachers and student have different architectures and layer-wise matching is difficult. In the future, we will explore heterogeneous network settings beyond ResNets.

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 updating GAs with the same frequency as the student but got similar results (see Supplementary Material for numerical results). As a next step, we will investigate whether further reducing M_{GA} can still maintain the same performance. In this study, we used uniform weights for w_k in Eq. (1), where $w_k = 1/K$. In more complicated scenarios where there are mixed-quality teachers, adaptive sample-wise weight assignment [28] may be beneficial. For example, teacher predictions close to one-hot labels can be assigned with larger weights.

Last but not least, we focused on the scenario where the teachers are all trained to predict the same classes in this study. DFMF with multi-talent teachers that specialize on different classification tasks is our on-going research, where we expect the GAs will also be helpful.

6. Conclusion

In this paper, we present a novel data-free model fusion method (DFMF) that combines multiple pretrained models into one model for superior performance without the need for the raw training data. We propose to use generator assistants to improve the generators for producing adversarial samples, which are helpful for multi-teacher knowledge distillation. Experiments on the CIFAR-10, CIFAR-100 and Stanford Dogs datasets demonstrate that our proposed DFMF achieves substantial improvement compared to the prior art. We also show that the proposed DFMF works well on both homogeneous and heterogeneous network settings.

References

[1] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9163–9171, 2019. 2, 3, 5

[2] Hanting Chen, Yunhe Wang, Chang Xu, Zhaohui Yang, Chuanjian Liu, Boxin Shi, Chunjing Xu, Chao Xu, and Qi Tian. Data-free learning of student networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3514–3522, 2019. 2, 5

[3] Yoojin Choi, Jihwan Choi, Mostafa El-Khamy, and Jungwon Lee. Data-free network quantization with adversarial knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 710–711, 2020. 2, 4, 5

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 5

[5] Gongfan Fang, Jie Song, Xinchao Wang, Chengchao Shen, Xingen Wang, and Mingli Song. Contrastive model inversion for data-free knowledge distillation. *arXiv preprint arXiv:2105.08584*, 2021. 2, 5

[6] Gongfan Fang, Kanya Mo, Xinchao Wang, Jie Song, Shitao Bei, Haofei Zhang, and Mingli Song. Up to 100x faster data-free knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6597–6604, 2022. 2, 3, 6

[7] Pengchao Han, Jihong Park, Shiqiang Wang, and Yejun Liu. Robustness and diversity seeking data-free knowledge distillation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2740–2744. IEEE, 2021. 2

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[9] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 2

[10] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, 2011. 4

[11] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015. 5

[12] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016. 1

[13] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009. 4

[14] Yuang Liu, Wei Zhang, and Jun Wang. Adaptive multi-teacher multi-level knowledge distillation. *Neurocomputing*, 415:106–113, 2020. 2

[15] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[16] Liangchen Luo, Mark Sandler, Zi Lin, Andrey Zhmoginov, and Andrew Howard. Large-scale generative data-free distillation. *arXiv preprint arXiv:2012.05578*, 2020. 2

[17] Paul Micaelli and Amos J Storkey. Zero-shot knowledge transfer via adversarial belief matching. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 5

[18] Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence o (1/kˆ 2). In *Doklady an ussr*, pages 543–547, 1983. 5

[19] SeongUk Park and Nojun Kwak. Feature-level ensemble knowledge distillation for aggregating knowledge from multiple networks. In *ECAI 2020*, pages 1411–1418. IOS Press, 2020. 2

[20] Robi Polikar. Ensemble based systems in decision making. *IEEE Circuits and systems magazine*, 6(3):21–45, 2006. 1

[21] Lior Rokach. Ensemble-based classifiers. *Artificial intelligence review*, 33(1):1–39, 2010. 1

[22] Chengchao Shen, Xinchao Wang, Jie Song, Li Sun, and Mingli Song. Amalgamating knowledge towards comprehensive classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3068–3075, 2019. 2

[23] Chengchao Shen, Mengqi Xue, Xinchao Wang, Jie Song, Li Sun, and Mingli Song. Customizing student networks from heterogeneous teachers via adaptive knowledge amalgamation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3504–3513, 2019. 2

[24] Jingwen Ye, Yixin Ji, Xinchao Wang, Xin Gao, and Mingli Song. Data-free knowledge amalgamation via group-stack dual-gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12516–12525, 2020. 3

[25] Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8715–8724, 2020. 2

[26] Jaemin Yoo, Minyong Cho, Taebum Kim, and U Kang. Knowledge extraction with no observable data. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[27] Shan You, Chang Xu, Chao Xu, and Dacheng Tao. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1285–1294, 2017. 2

[28] Hailin Zhang, Defang Chen, and Can Wang. Confidence-aware multi-teacher knowledge distillation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4498–4502. IEEE, 2022. 2, 8