Data-Free Knowledge Amalgamation via Group-Stack Dual-GAN
– *Supplementary Material* –

Jingwen Ye[1], Yixin Ji[1], Xinchao Wang[2], Xin Gao[3], Mingli Song[1]
[1]College of Computer Science and Technology, Zhejiang University, Hangzhou, China
[2]Department of Computer Science, Stevens Institute of Technology, New Jersey, United States
[3]Alibaba Group, Hangzhou, China
{yejingwen, jiyixin, brooksong}@zju.edu.cn, xinchao.w@gmail.com, zimu.gx@alibaba-inc.com

Here we provide the additional details and results that are left in the main text to this supplementary material. First, we provide more details of the proposed approach and the experimental implementation. We then show additional results on the 'Customized Tasks of Multiple Teachers' as well as the discuss the influence of the group number of the generator.

1. More Details of Method

Due to the page length limit, some details are omitted in the main manuscript. Here, we provide more information on the adversarial loss and the teacher-level filter.

1.1. Adversarial Loss

Note that in Sec. 4.1 of the main paper, the adversarial loss is calculated to update the generator as:

$$\mathcal{L}_{gan} = \mathcal{L}_{oh} + \alpha\mathcal{L}_a + \beta\mathcal{L}_{ie} + \gamma\mathcal{L}_{dis}, \tag{1}$$

where $\mathcal{L}_a$ and $\mathcal{L}_{ie}$ are the activation loss function and the information entropy loss function respectively. These two losses are justified from the work of [1], which aims at solving the multi-class classification problem. In this paper, some modifications are made to fit the multi-label classification problem.

Given a set of random vector $\{z(n)\}_{n=1}^{N}$, images generated from these vectors are denoted as $\{\mathcal{I}_{gan}(n)\}_{n=1}^{N}$. For z^n, and the features extracted before the fully connected layer of the M-target discriminator D are donated as $\{F_1(n), F_2(n), ..., F_M(n)\}$. Then the activation loss is calculated as:

$$\mathcal{L}_a = -\frac{1}{NM}\sum_{n}\sum_{m}||F_m(n)||_1, \tag{2}$$

where $||\cdot||_1$ is the L1 norm. This multi-label activation loss forces the feature maps to receive higher activations with the real image input.

In order to construct a balanced dataset by generator G for the multi-label classification task, the information entropy loss is utilized. Recall that the raw output from the discriminator is $\mathcal{O}^n = \{y^1(n), y^2(n), ..., y^C(n)\}$ with $0 \leq y \leq 1$. The multi-label information entropy loss, which maximizes the information entropy of generated image set, is organized as:

$$\mathcal{L}_{ie} = \frac{1}{C}\sum_{c}\mathcal{H}_{info}(\frac{1}{N}\sum_{n}y^c(n)), \tag{3}$$

where $\mathcal{H}_{info}(p) = p\log p + (1-p)\log(1-p)$.

1.2. Teacher-level Filtering

As can be seen in Eq. 10 of the main manuscript, we utilize the teacher-level filter f_m^j for the j-th group generator G^j to amalgamate the knowledge of teacher $\mathcal{A}_m$. The filter is constructed by a global pooling layer and two fully connected layers, which is modified from the channel-attention module in [2].

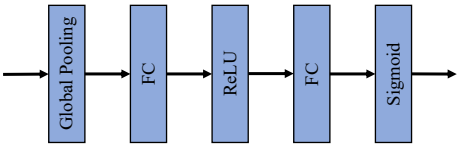

Figure 1. The learnable teacher-lever filter, which consists of two fully connected layers and one global pooling layer.

More concrete architecture of the teacher-level filter is depicted in Fig. 1. Let c denotes the number of channels of the output feature maps in G^j and thus also the number of channels fed to the teacher-level filter $\{f_1^j, f_2^j, ..., f_M^j\}$. Within each filter, the first fully connected layer reduces the channel number to c/r, and then the second fully connected layer reverts the number to c. In the experiments, we set $r = 4$ for all the groups of the generator G and the dual-generator $\mathcal{T}$. Besides, the proposed filter is very light in size and increases the total number of parameters by less than 4%, leading to very low computation cost.

2. More Details in Implement

2.1. Architecture of GAN

In this paper, we construct the GAN in the dual architecture, where the generator G produces the generated images as well as the intermediate features from z and the dual-generator $\mathcal{T}$ does the image-to-vector work. As for the discriminator part, we regroup the off-the-shelf teacher models as the well-trained discriminators.

In Fig. 2, we show the architecture of one single group, which is repeated to form a multi-group generator. In the main manuscript, we set the total group number of G to be B, which is the same as the total block number of the teacher $\mathcal{A}$. This setting makes the proposed method easy to follow. In real-world application, however, the group number can be less than B, and a 2-group stack generator may already yield satisfactory results.

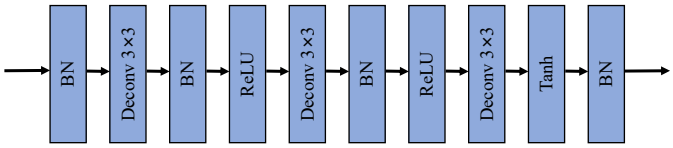

Figure 2. The architecture of one single group of the generator G.

2.2. Parameter Settings

We provide here more details on our parameter settings.

Weights for Loss $\mathcal{L}_{gan}$. The balancing weights setting for the adversarial loss $\mathcal{L}_{gan}$ in Eq. 1 is depicted in Table 1, which is utilized for all the datasets (VOC 2007 and MS-COCO) in this paper.

Table 1. Balancing weights α, β and γ setting in the adversarial loss.

Balancing Weights	$\mathcal{L}_a - (\alpha)$	$\mathcal{L}_{ie} - (\beta)$	$\mathcal{L}_{dis} - (\gamma)$
Value	1	1	5

Weights for Loss $\mathcal{L}_u^b$. $\mathcal{L}_u^b = \lambda_{in}^1 \mathcal{L}_{dual}^{b,m}(F_{in}^1) + \lambda_{in}^2 \mathcal{L}_{dual}^{b,m}(F_{in}^2)$ is utilized to train the b-th block of the dual-generator $\mathcal{T}$, where the balance weights are listed in Table 2.

Table 2. Balancing weights λ_{in}^1 and λ_{in}^2 for dual adversarial loss $\mathcal{L}_u^b$.

Balancing Weights	$\mathcal{L}_{dual}^{b,m}(F_{in}^1) - (\lambda_{in}^1)$	$\mathcal{L}_{dual}^{b,m}(F_{in}^2) - (\lambda_{in}^2)$
Value	1	1

3. Additional Results

In this section, we perform additional experiments to show the influence of the group number and the case of 'Customized Tasks of Multiple Teachers'.

3.1. Group Number of the Generator.

As we discuss in the Sec. 2.1, the group number of the generator G can be an arbitrary number no greater than B. In this experiment, we discuss the influence of the group number of the generator, which is performed on the VOC 2007 dataset with the same setting of 'Customized Tasks of Single Teacher' in the main text.

The classification results for different group numbers of the generator are shown in Table 3, where the block number is $B = 4$ for the teachers (ResNet-101). The accuracy results fluctuate strongly with the different group number settings. The 2-stack generator outperforms with 73.6% other the 1-stack and 4-stack generator. Therefore an important conclusion can be drawn that the final accuracies do not monotonically increase with the group number, and it is of vital importance to choose the exact group number.

Table 3. The classification results (mAP%) on VOC 2007 with respect to different group numbers of the generator. The generators with one group, 2 groups and 4 groups are compared.

Group Number of G	1-stack	2-stack	4-stack
mAP	49.3	73.6	65.2

3.2. Customized Tasks of Multiple Teachers

In the main manuscript, we show the experimental results of the customized tasks of single teacher and multiple teachers on whole label set (Sec. 5.2.1). In this section, we show the performance of the proposed method of the customized tasks of multiple teachers, which is conducted on the VOC 2007 dataset. We randomly separate a total of 20 labels into two sets, each learned by one teacher network:
Teacher-1: $Y_1 \leftarrow$ {'plane', 'bike', 'bird', 'boat', 'bus', 'car', 'horse', 'motor', 'person', 'train'};
Teacher-2: $Y_2 \leftarrow$ {'bottle', 'cat', 'chair', 'cow', 'table', 'dog', 'plant', 'sheep', 'sofa', 'tv'}.

And the accuracies for all the labels learned in the teachers are depicted in Table 4.

Table 4. The classification results (AP in %) on the randomly selected 10-label set learned in the teachers $\mathcal{A}_1$ and $\mathcal{A}_2$. And $\{y^1, y^2..., y^{10}\}$ denote the labels in Y_1 and Y_2.

Teachers	y^1	y^2	y^3	y^4	y^5	y^6	y^7	y^8	y^9	y^{10}	mAP
$\mathcal{A}_1$	94.3	82.0	82.4	81.9	76.3	89.2	89.5	82.4	93.2	90.0	86.1
$\mathcal{A}_2$	38.5	80.0	56.5	62.9	73.3	79.0	51.7	74.0	67.9	71.6	65.5

Then the task level filters are taken to be:

$$Y_{cst} = g_1(Y_1) \cup g_2(Y_2) \leftarrow \{\text{'bus', 'table'}\}, \tag{4}$$

which means that we train the TargetNet to justify the existence of the labels 'bus' and 'table'.

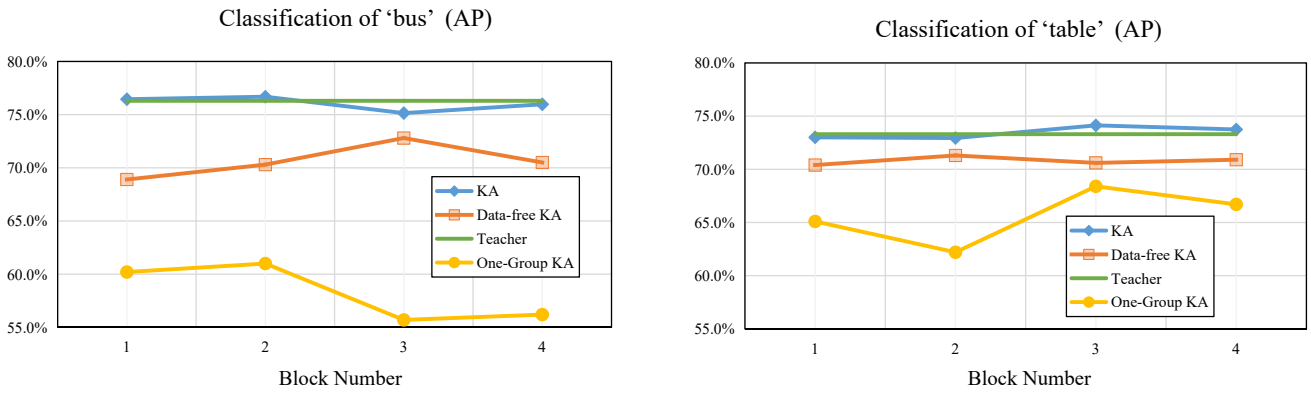

Figure 3. The classification results on label 'bus' and 'table' during the block-wise training process.

Note that the block-wise training strategy is utilized for the dual-generator (TargetNet), the comparative experimental results for each block is shown in Fig. 3. We compare the performance of the orginal KA('KA'), one-group KA and the

proposed group-stack KA ('Data-free KA'), where the original KA trained with unlabeled dataset surpasses the teachers' performance. The proposed 'Data-free KA' obtains the competitive results compared with 'KA', which shows the effectiveness of our method. Recall that at the last step of acquiring the satisfying TargetNet, we branch out the trained network with the consideration of the loss convergence values. In our implementation, the branching out takes place at the 3-th block of the TargetNet, which is proved to produce high accuracies for both 'bus' and 'table' in the chart.

Table 5 displays the final results of the branched TargetNet (branching out at block-3 for both 'bus' and 'table') and compares the parameters of the final models. The trained TargetNet is smaller than the teachers, and the performance are approaching the teachers'. Compared with the method trained with random noise, the proposed method shows great superiority.

Table 5. Comparative results of the teacher networks, the branched TargetNet (ours) and random noise. The parameters of the final networks are also depicted.

Methods	Parameters	table	bus	mAP
Teacher	$\sim$ 88.7 M	73.3	76.3	74.8
Random Noise	-	5.1	4.8	5.0
Ours	$\sim$ 59.8 M	70.6	72.8	71.7

References

[1] Hanting Chen, Yunhe Wang, Chang Xu, Zhaohui Yang, Chuanjian Liu, Boxin Shi, Chunjing Xu, Chao Xu, and Qi Tian. Dafl: Data-free learning of student networks. In *International Conference on Computer Vision*, 2019. 1

[2] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 1