Supplementary Material for Effectiveness of Arbitrary Transfer Sets for Data-free Knowledge Distillation

Gaurav Kumar Nayak*
Indian Institute of Science
Bangalore, India
gauravnayak@iisc.ac.in

Konda Reddy Mopuri*
Indian Institute of Technology
Tirupati, India
kmopuri@iittp.ac.in

Anirban Chakraborty
Indian Institute of Science
Bangalore, India
anirban@iisc.ac.in

1. Importance of Class Balance: Training Data

We considered AlexNet CNN trained on CIFAR-10 as the *Teacher* and AlexNet-Half as the *Student*. We carefully composed multiple transfer sets representing only a subset of classification regions learned by the *Teacher* in order to perform the distillation. In other words, we performed distillation using the samples from different number of classes. We varied the number of classes present in the transfer set, fixing its size (total number of samples).

More specifically, we fixed the size of the transfer set to approximately[1] 10000 and varied the class composition from 2 to 10. Note that as the number of representing classes increases, number of samples per class decrease to meet the fixed size criterion. Table 1 shows the distillation performance of the transfer sets in terms of the *Student* classification accuracy on the test set that consists of samples from all the 10 classes. Clearly, the performance increases monotonically with the class balance in the transfer set. In other words, with better representation of the classification regions the transfer set achieves better distillation between the *Teacher* and *Student*.

	# classes in the transfer set				
	2	4	6	8	10
Student Acc.	19.56	34.50	46.08	58.94	71.24
Feature dist.	52.25	49.26	44.45	34.72	8.97

Table 1. Importance of the class balance in a fixed-size transfer set when training data is used for transfer. Test accuracy of the *Student* trained via distillation on samples from different number of CIFAR-10 classes. The table also presents the Hausdorff distance between transfer set and training set computed in the feature space.

We can also verify the effectiveness of a transfer set via measuring its similarity (or distance) to the target data distribution. Note that the data on which the *Teacher* model is trained and tested is assumed to be sampled from the target data distribution. Therefore, along with the *Student*'s accuracy we also compute the distance between the transfer set and the training set. Since those are sets of images, we compute the Hausdorff distance [5] between the corresponding feature sets. We consider the deepest embedding (before the softmax layer) learned by the *Teacher* model as the feature. Bottom row of Table 1 shows the computed distances. Note that the distance monotonically decreases as the balance in the transfer set improves. In other words, as the similarity of the transfer set to the target set improves, the distillation performance improves.

2. Importance of Class Balance: Arbitrary Data

In this subsection we demonstrate the importance of class balance in the case of arbitrary transfer set. We consider setup similar to that in section 1 with AlexNet trained on CIFAR-10 as *Teacher* and AlexNet-Half as *Student*. However, here we consider an arbitrary transfer set composed with samples from SVHN and TinyImageNet datasets. Similar to the previous subsection, we fixed the size of the transfer set approximately to 10000 and analysed the effect of class balance. We varied the

*denotes equal contribution

[1] Training data per each class is not exactly equal to 5000 but very close to it.

class representation from 2 to 10 in the transfer set and investigated the distillation performance. Table 2 shows the *Student*'s classification accuracy on the 10000 CIFAR-10 test set that has almost equal number of samples from all the 10 classes. Note that the performance monotonically increases with the class-balance in the arbitrary transfer set.

Similar to section 1 we also verified the similarity of the transfer sets to the target set via measuring the feature similarity. Bottom row of Table 2 shows the Hausdorff distance measured between the corresponding feature sets. It is evident that with better class-balance in the transfer set, it is more similar to the target dataset and results in better distillation. Thus, sections 1 and 2 clearly support the proposed hypothesis that class-balance improves the distillation performance.

	# classes in the transfer set				
	2	4	6	8	10
Student Acc.	23.36	26.51	32.08	39.06	44.42
Feature dist.	50.25	49.99	44.73	38.65	38.65

Table 2. Importance of the class balance in the fixed-size transfer set when arbitrary data (SVHN+Tiny Imagenet) is used. Note that the table also presents the Hausdorff distance between the transfer set and the training set computed in the feature space.

3. Base Transfer Set Matters

Transfer set	Balance	Distillation Performance
SVHN	✗	40.46
SVHN+Tiny Imagenet	✓	64.90
Tiny Imagenet	✗	66.94
Tiny Imagenet + SVHN	✓	70.58

Table 3. Distillation performance with different base transfer sets on the CIFAR-10 dataset. Note that in order to have a fair comparison, amount of transfer set does not exceed the total count of original training data. The training is done without using any augmentations.

Along with class balancing, the choice of base transfer set is also an important factor on which the effectiveness of the KD depends. From Table 3, it can be observed that there is a significant improvement in the distillation performance when a natural dataset such as TinyImageNet is considered as base transfer set in comparison to SVHN. Moreover, balanced TinyImageNet gives further improvement in the accuracy which is approximately 6% more than using balanced SVHN. Therefore, the order in which the arbitrary datasets are mixed to create a transfer set also matters.

4. Type of Dataset Used for Balancing also Matters

Unbalanced Dataset	Dataset which is added to balance class count			
	No dataset added	Random Noise	Mid-Air (Synthetic)	Tiny Imagenet (Natural)
Random Noise	14.38	31.03	38.03	67.29
Clevr	36.58	41.53	47.18	60.04
SVHN	40.46	46.77	49.65	64.90

Table 4. Ablation on different types of dataset when mixed with an unbalanced transfer set to increase the target class balance on CIFAR-10 *Teacher*. The values represent the distillation performance.

From Table 4, we can observe that an unbalanced transfer set when used directly always gives lower distillation performance in comparison to addition of dataset for increasing the class balance. Please note that we always ensure that count of samples in a balanced transfer set does not exceed the count of samples in unbalanced dataset. Also, the maximum number of samples taken from unbalanced transfer set is limited by the amount of original training data which was used for training the *teacher* model in oder to have a fair comparison. This shows that class-balanced transfer sets are more effective than unbalanced transfer set. Even when random noise is added to unbalanced transfer set like random data, synthetic data or natural data improves the distillation accuracy significantly. Moreover, we can notice that the distillation performance also depends on type of dataset which is being added for target class balancing. The distillation performance is best when a natural dataset (Tiny Imagenet) is added to several unbalanced transfer sets like Random Noise, Clevr and SVHN data.

5. What Makes a Better Transfer Set ?

It is clear that not all arbitrary transfer sets are equally effective. Even a pair of target class-balanced transfer sets need not be equally effective. Despite the balance, random noise, synthetic, and natural arbitrary transfer sets result in significantly different transfer performances. This naturally rises the question, *In the absence of the training data, what makes an arbitrary transfer set effective?* By now, intuitively one can expect that "more the similarity of the transfer set to the training set, better the transfer". Although it is not a sufficient condition, it is consistently observed that better similarity results in effective transfer performance. For instance, Table 5 shows this similarity (in terms of distance) against the corresponding transfer performance for random noise, synthetic, and natural data as transfer set. As the distance at which the transfer set lies from the manifold of the training dataset, its effectiveness decreases. Note that it is not possible to find such similarity easily in the absence of training data. However, this simple observation can guide and influence the future data-free knowledge transfer objectives that attempt to create proxy transfer set (e.g. [4, 3, 1]).

Transfer Set	Distillation Acc.	Feature Distance
Random Noise	67.40	36.40
Synthetic data	76.92	29.91
Natural data	79.19	28.53

Table 5. Distillation performance using different types of transfer sets for distilling the knowledge from CIFAR-10 *Teacher*.

6. Class Frequencies in Transfer Set

In this section we present the class frequencies in the transfer set before and after the balancing. In other words, we show the number of samples in the transfer set that are classified into each of the classes in the *Teacher*'s training data before and after the proposed class-balancing (Algorithm 1 in the main draft). Note that these counts are related to the performances in Table 1 of the main draft without performing augmentation. Tables 6, 7, 8 show the counts for MNIST, FMNIST, and CIFAR-10 datasets with various arbitrary transfer sets. Note that though the achieved balance is not perfect, it is very significant compared to the unbalanced arbitrary transfer set t which results in the distillation performance (Table 1 of the main draft).

Class Label	**Random Noise**		**Synthetic**		**Natural data**	
	Unbalanced	Balanced	Unbalanced (Clevr)	Balanced (Clevr + Mid-Air)	Unbalanced (SVHN)	Balanced (SVHN+Tiny ImageNet)
0	933	6000	2868	4415	3295	6000
1	233	6000	3376	5450	6758	6000
2	3721	6000	156	6000	1588	6000
3	4835	6000	898	6000	1092	6000
4	5910	6000	31100	6000	31368	6000
5	10435	6000	2026	6000	1424	6000
6	419	6000	2117	4745	2828	6000
7	2886	6000	5884	6000	6014	6000
8	29280	6000	9443	6000	3849	6000
9	1348	6000	2132	6000	1784	5773

Table 6. Class frequencies (number of samples in each class) before and after class-balancing various transfer sets for performing KD on the MNIST.

Class Label	Random Noise		Synthetic		Natural data	
	Unbalanced	Balanced	Unbalanced (Clevr)	Balanced (Clevr + Mid-Air)	Unbalanced (SVHN)	Balanced (SVHN+Tiny ImageNet)
0	2751	6000	6959	6000	8015	6000
1	3501	6000	988	6000	3544	6000
2	917	6000	5698	6000	4625	6000
3	404	6000	656	6000	4892	6000
4	263	6000	40	1000	390	2240
5	99	6000	551	6000	1113	6000
6	22614	6000	1413	6000	13500	6000
7	0	6000	0	92	20	178
8	29451	6000	43681	6000	22578	6000
9	0	6000	14	844	1323	2702

Table 7. Class frequencies (number of samples in each class) before and after class-balancing various transfer sets for performing KD on the FMNIST.

Class Label	Random Noise		Synthetic		Natural data	
	Unbalanced	Balanced	Unbalanced (Clevr)	Balanced (Clevr + Mid-Air)	Unbalanced (SVHN)	Balanced (SVHN+Tiny ImageNet)
0	0	5000	11205	5000	1719	5000
1	0	5000	461	2490	34	2493
2	15	5000	1416	5000	7886	5000
3	1	5000	5937	5000	27073	5000
4	13	5000	129	5000	1803	5000
5	0	5000	22472	5000	7761	5000
6	49971	5000	1680	5000	23	5000
7	0	5000	2186	4727	2509	5000
8	0	5000	138	5000	928	4190
9	0	5000	4376	5000	264	5000

Table 8. Class frequencies (number of samples in each class) before and after class-balancing various transfer sets for performing KD on the CIFAR-10.

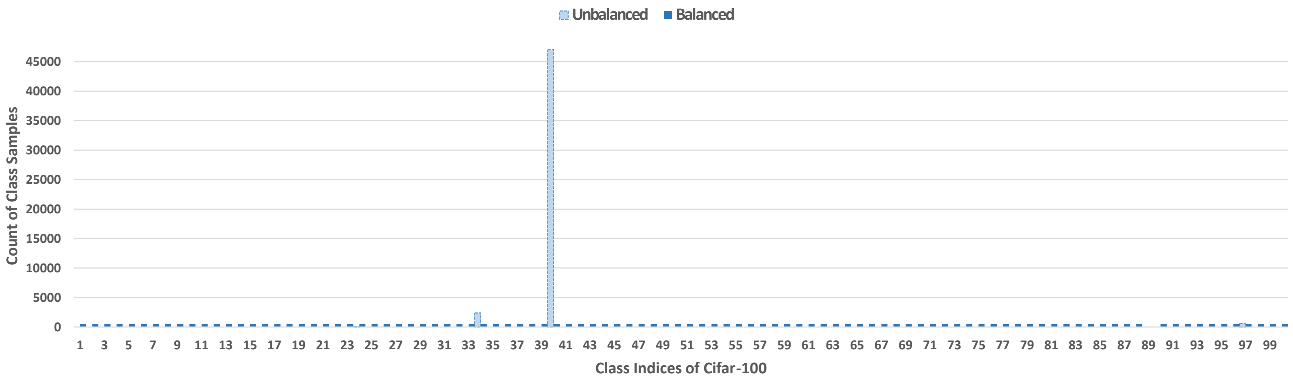

Figure 1. Class frequencies (number of samples in each class) before and after class-balancing the random noise data for performing KD on the CIFAR-100.

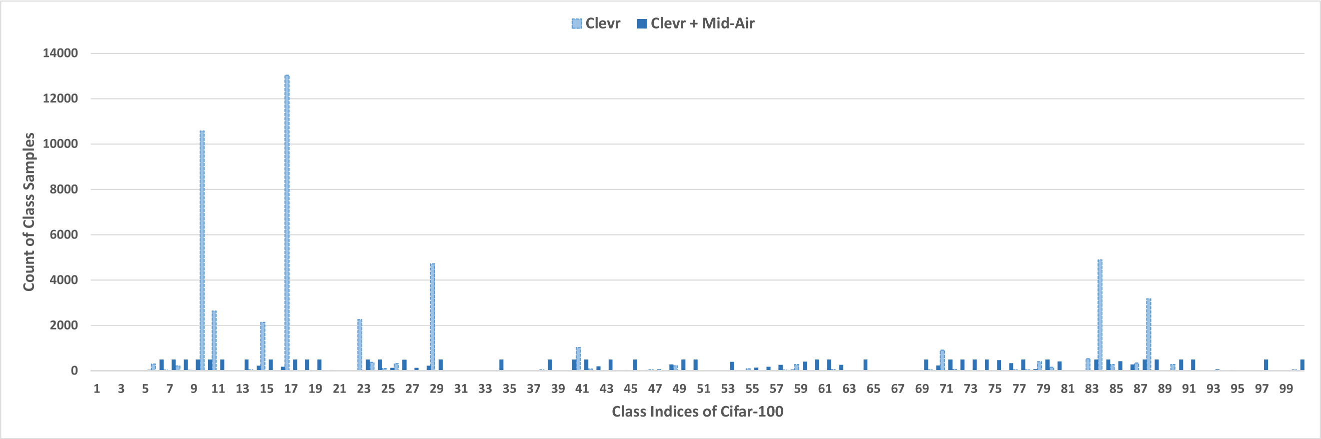

Figure 2. Class frequencies (number of samples in each class) before and after class-balancing the synthetic data for performing KD on the CIFAR-100.

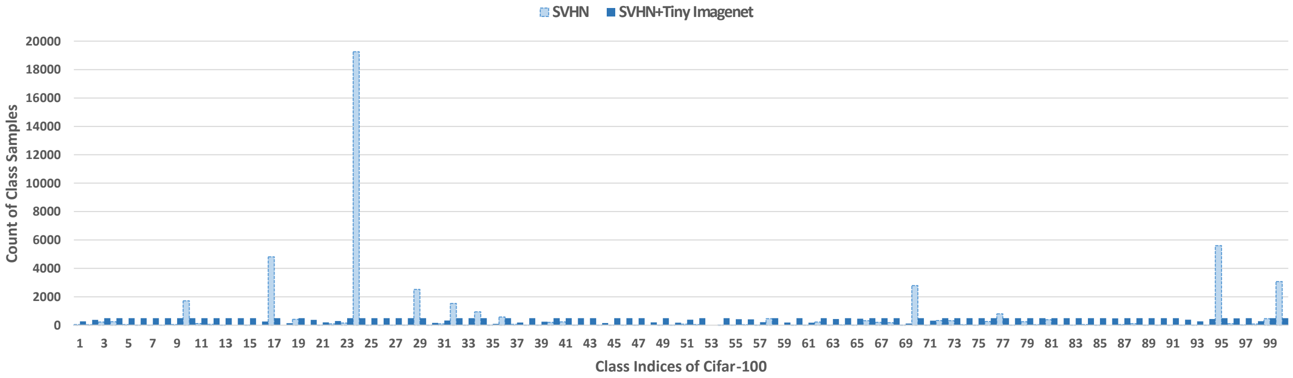

Figure 3. Class frequencies (number of samples in each class) before and after class-balancing the natural data for performing KD on the CIFAR-100.

Figures 1, 2 and 3 show the class frequencies for CIFAR-100 dataset when the arbitrary transfer sets are Random noise, Synthetic, and Natural datasets respectively. However, note that it is challenging to achieve perfect class balance using any arbitrary transfer set due to the large number of classes present in it. Even after mixing 'Mid-Air' dataset on the base transfer set 'Clevr', it does not achieve the class balance significantly which can also be observed in Figure 2. In order to retain the same transfer set across several other datasets such as MNIST, FMNIST and CIFAR-10, we did not choose any other synthetic datasets which could have significantly improved the balance of 'Clevr' and hence achieved better distillation performance.

7. Overall Size of Transfer Sets Used in Experiments

Transfer Sets	Balanced	MNIST	FMNIST	CIFAR-10	CIFAR-100
Random Noise	✗	60000	60000	50000	50000
Random Noise	✓	60000	60000	50000	50000
Clevr	✗	60000	60000	50000	50000
Clevr + Mid-Air	✓	56610	43936	47217	26407
SVHN	✗	60000	60000	50000	50000
SVHN + Tiny Imagenet	✓	59773	47120	46683	43026

Table 9. Total number of samples in transfer sets used for distilling the knowledge from *Teacher* model trained on MNIST, FMNIST, CIFAR-10 and CIFAR-100.

The amount of original training samples used to train the *Teacher* network on MNIST, FMNIST, CIFAR-10 and CIFAR-100 are 60000, 60000, 50000 and 50000 respectively. Due to privacy and safety concerns, we assume the unavailability of these original samples as motivated in several works [4, 3, 2]. Thus, in order to train the lightweight models called *Student* network, we leverage on the availability of arbitrary data. This arbitrary data acts as a transfer set for distilling the knowledge of the pretrained *Teacher* network. It is evident from the Table 9 that size of arbitrary transfer set does not exceed the amount of original training samples to have a fair comparison. Also, the amount of samples per target class in case of balanced transfer sets, does not exceed the amount of samples per class in original training data which is 6000 for MNIST and FMNIST and 5000 for CIFAR-10 and CIFAR-100 respectively. For the experiments, we limit ourselves to mixture of two arbitrary datasets for obtaining balanced transfer sets. Therefore, we sometimes end up in having non-perfectly balanced transfer sets which have lower transfer set size in comparison to original training samples. Even then, we have shown that these transfer sets achieve better distillation accuracy than the unbalanced transfer sets.

8. Summary

Finally, we summarize the major advantages of our proposed approach as follows:

1. The proposed method is intuitive in the sense that higher the number of classes and their population, the teacher network is able to transfer more information on to the student network which will reflect in its generalization capabilities.

2. It achieves competitive distillation performance even with an arbitrary transfer set in the absence of original training data.

3. It does not require any complicated training procedure, or generative models such as the GANs. The arbitrary transfer sets are used in their original forms (albeit with augmentations applied to them).

4. Only the distillation loss is used. There is no other additional/auxiliary loss which otherwise needs to be properly weighted with the distillation loss.

Hence the proposed approach can be used as a simple baseline, particularly for Data free Knowledge Distillation research works and can act as an alternative to computationally expensive approaches.

References

[1] 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 *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[2] Raphael Gontijo Lopes, Stefano Fenu, and Thad Starner. Data-free knowledge distillation for deep neural networks. In *LLD Workshop at Neural Information Processing Systems (NIPS)*, 2017.

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

[4] Gaurav Kumar Nayak, Konda Reddy Mopuri, Vaisakh Shaj, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty. Zero-shot knowledge distillation in deep networks. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[5] A. A. Taha and A. Hanbury. An efficient algorithm for calculating the exact hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 37(11):2153–2163, Nov 2015.