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

Abstract

Knowledge Distillation is an effective method to transfer the learning across deep neural networks. Typically, the dataset originally used for training the Teacher model is chosen as the "Transfer Set" to conduct the knowledge transfer to the Student. However, this original training data may not always be freely available due to privacy or sensitivity concerns. In such scenarios, existing approaches either iteratively compose a synthetic set representative of the original training dataset, one sample at a time or learn a generative model to compose such a transfer set. However, both these approaches involve complex optimization (GAN training or several backpropagation steps to synthesize one sample) and are often computationally expensive. In this paper, as a simple alternative, we investigate the effectiveness of "arbitrary transfer sets" such as random noise, publicly available synthetic, and natural datasets, all of which are completely unrelated to the original training dataset in terms of their visual or semantic contents. Through extensive experiments on multiple benchmark datasets such as MNIST, FMNIST, CIFAR-10 and CIFAR-100, we discover and validate surprising effectiveness of using arbitrary data to conduct knowledge distillation when this dataset is "target-class balanced". We believe that this important observation can potentially lead to designing baselines for the data-free knowledge distillation task.

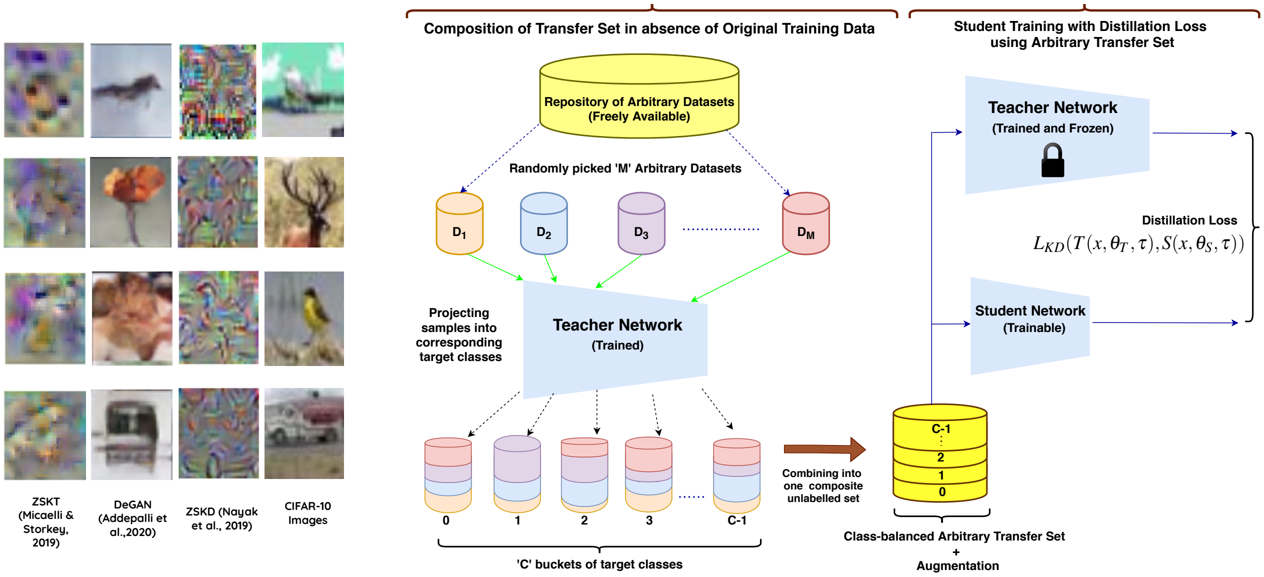
1. Introduction

Knowledge Distillation (KD) [3, 7] is a contemporary technique for transferring learning across neural network models. Typically, knowledge from one or more complex and deep models (called *Teacher*s) is distilled into a relatively lightweight model (called *Student*). The core idea of Knowledge Distillation, as discussed in the seminal paper by Hinton *et al*. [7], is to transfer the (input to output) learned mapping function from *Teacher* to *Student* via sharing the "dark knowledge" extracted by the *Teacher* on the training images. This typically is achieved via matching the soft targets (or soft labels, i.e., output of softmax layer) predicted by the *Student* to that of the *Teacher* for the same inputs. This is the distillation mechanism that enables transfer of the better generalization capability (i.e., the "knowledge") of the *Teacher* to the *Student*. Thus, Knowledge Distillation has established itself as a very useful and practical tool because of its simplicity and potential.

The samples used for performing distillation constitute the "Transfer set", which is typically required to be constructed using the data sampled from the target distribution. Therefore, the most commonly used transfer set is the original training dataset on which the Teacher model was trained. However, this requirement has been identified as a limitation (e.g. [17, 15]) since it is common now-a-days that many popular pre-trained models are released without providing access to the training data (e.g. Facebook's Deepface model trained on 4M confidential face images). This is due to one or more practical constraints such as (i) privacy (e.g. models trained on patients' data from hospitals), (ii) property (proprietary data of companies that invest on collection and annotation), and (iii) transience (observations from the training of a reinforcement learning environment do not exist).

To handle this "data-free" (or zero-shot) distillation scenario, most of the approaches broadly follow either of the two ways: (i) compose a synthetic transfer set by directly utilizing the trained *Teacher* model that acts as a proxy to the target data (e.g. [17, 14]), or (ii) capture the target data distribution using generative models (e.g. [15, 4, 1]). Both these approaches suffer from heavy computational overhead: iteratively crafting synthetic samples via several steps of backpropagation through the *Teacher* or learning a complex GAN like generator framework that involves complicated optimization. Some of these approaches (e.g. [14]) additionally need to store meta-data about the original training dataset (e.g. feature statistics of the *Teacher* model) for generating the synthetic transfer set. Further, in case of image data, the generated samples are often observed to

[*] denotes equal contribution.


ZSKT (Micaelli & Storkey, 2019)
DeGAN (Addepalli et al.,2020)
ZSKD (Nayak et al., 2019)
CIFAR-10 Images
Composition of Transfer Set in absence of Original Training Data
Repository of Arbitrary Datasets (Freely Available)
Randomly picked 'M' Arbitrary Datasets
D_1
D_2
D_3
D_M
Projecting samples into corresponding target classes
Teacher Network (Trained)
0
1
2
3
C-1
'C' buckets of target classes
Combining into one composite unlabelled set
C-1
2
1
0
Class-balanced Arbitrary Transfer Set + Augmentation
Student Training with Distillation Loss using Arbitrary Transfer Set
Teacher Network (Trained and Frozen)
Distillation Loss
$L_{KD}(T(x,\theta_T,\tau),S(x,\theta_S,\tau))$
Student Network (Trainable)

(a)

(b)

Figure 1. (a) Example of pseudo samples generated by existing data-free KD approaches for the CIFAR-10 dataset compared against actual target data samples (rightmost column), (b) Our proposed KD Baseline: An approach depicting simple and effective way of performing KD in absence of the original training data by utilizing arbitrary data samples to construct the transfer set.

be visually quite dissimilar (Fig. 1(a)) to the training data samples. That means, they do not lie close to the training samples in the data manifold. At the same time, it is unclear how or why these samples, despite seemingly being "out-of-distribution" and "far-from-real", enable effective transfer between the models, as evidenced by the reported results.

These observations motivate us to investigate the effectiveness of any *arbitrary transfer set* towards the task of knowledge distillation, despite it being unrelated to the original training data. If proven effective, such datasets can in fact be used to design important and often strong baselines for the KD tasks, while saving us the large overhead of composing synthetic transfer sets, as incurred by the existing data-free distillation approaches. This is especially true for text/image domains, where it is easy to collect large volume of unlabeled arbitrary data from ubiquitous publicly available sources. More importantly, this investigation can uncover important insights into the mechanism of the distillation process.

Therefore, in this work, we consider a wide range of unlabelled stimuli from different content worlds in the context of distillation. Specifically, we consider (i) random noise inputs, (ii) arbitrary synthetic datasets, and (iii) arbitrary natural datasets towards composing the transfer set. However, it is observed (refer to Sec. 3.2) that the deep neural networks often partition the arbitrary input domain into disproportionate classification regions. In other words, arbitrary data samples may not be projected uniformly into the learned classification regions of the *Teacher*. This imbalance in the classification regions results the *Student* to overfit the classification boundaries during the distillation. In other words, it can not preserve the class decision boundaries learnt from the original training data, thereby seriously affecting the *Student*'s generalization performance. These observations lead to the hypothesis that an ideal transfer set should equally represent all the classification regions of the *Teacher* model which can minimize the distortion in decision boundary and hence would help in achieving effective knowledge transfer. In other words, the arbitrary transfer set needs to be "target-class balanced" in order to successfully impart *Teacher*'s learning to the *Student*.

In summary, the contributions of this work are, as follows:

- For the first time in the literature, we show that arbitrary transfer sets, unrelated to the target data set, can be effectively utilized for the task of knowledge distillation in the "data-free" scenario.
- To maximize the efficacy of distillation using such transfer sets, we present a simple yet effective approach of making them "class-balanced".
- We empirically demonstrate the effectiveness of the proposed approach on multiple benchmark datasets such as MNIST, FMNIST, CIFAR-10 and CIFAR-100, as we achieve performance comparable to state-of-the-art data-free distillation approaches.

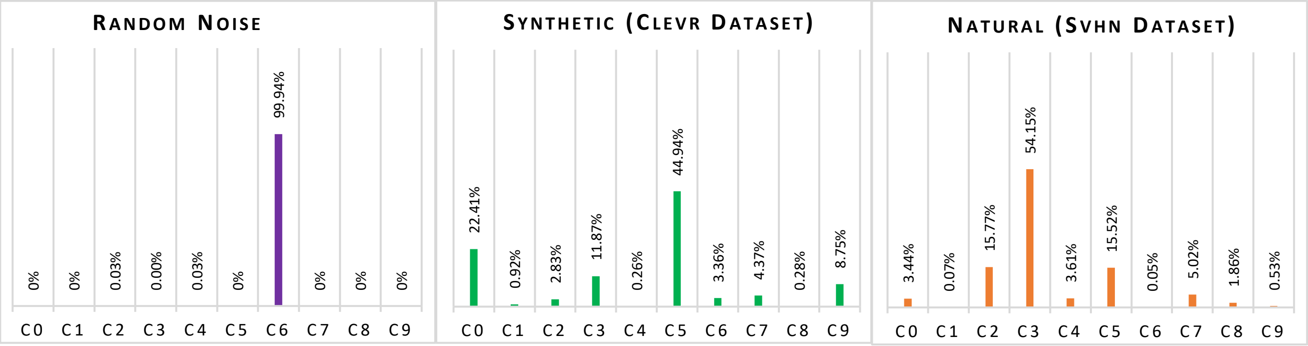


Figure 2. Percentage of the total number of arbitrary samples distributed over the set of the target classes (of CIFAR-10) by the trained *Teacher* model for different arbitrary datasets.

2. Related Works

Our work is broadly related to the data-free Knowledge Distillation. Early works (e.g. [3, 7]) use the entire training data as the transfer set. Buciluă *et al.* [3] suggest to meaningfully augment the training data for effectively transferring the knowledge of an ensemble onto a smaller model. Recently, there have been multiple approaches to perform knowledge transfer in the absence of training data. They can be broadly categorized into two branches: (i) methods that extract samples related to the training data from the *Teacher* model, (ii) methods that attempt to learn the training distribution (e.g., using GAN-like generative models).

General idea of the first category is to iteratively modify a randomly initialised stimulus (input) via back propagation in order to maximize the *Teacher*'s class confidence. Nayak *et al.* [17] compose the synthetic transfer set, which is carefully crafted by modelling the soft-label space. Lopes *et al.* [14] and Bharadwaj *et al.* [2] save information in the form of *Teacher*'s feature statistics in order to acquire transfer set that is closer to the training data. They further add diversity to the set via adding small noise to the saved statistics and generating the samples. These methods are computationally expensive requiring thousands of back propagation iterations per sample. Further, some of them require to store the feature statistics of the *Teacher* model which may not be available.

Another direction of research for handling the absence of training data is to train a generative model that can seed the proxy samples. [15, 4, 1] show that properly optimized generative models can generate samples to be strongly classified by the *Teacher* models. After learning such GAN-like models, generated samples can be used as a transfer set for performing the Knowledge Distillation. Despite generating *far from real* and *out-of-distribution* samples (Figure 1(a)) these methods are observed to successfully do the knowledge transfer. [1] attempts to use a proxy natural dataset for influencing the generations to be similar to proxy data distribution while bringing its features close to original training data distribution. These methods involve training generative models that require careful balancing of multiple terms in the objective. In this work, we take a different direction to study the effectiveness of cheaply available, unlabelled arbitrary data as transfer set. We put forward an intuitive strategy to compose effective transfer sets that can yield potential baselines and empirically demonstrate its efficacy.

3. Class-balanced Arbitrary Transfer Sets

We first briefly review the principles of Knowledge Distillation, and subsequently introduce our framework to compose an effective transfer set from unlabelled arbitrary data sources.

3.1. Knowledge Distillation (KD)

Knowledge Distillation typically uses the original training data as the transfer set on which the *Teacher* model is trained. Let us denote the *Teacher* as T, *Student* as S, their parameters as θ_T and θ_S respectively, and the transfer set as $\mathcal{D}$ that consists of the input-target tuples denoted by (x, y). Note that typically *Student*'s capacity would be smaller compared to that of *Teacher*, i.e., $|\theta_S| \ll |\theta_T|$. The objective of Knowledge Distillation is to train the *Student* in order to match the soft labels produced by the *Teacher* along with learning to predict the correct hard labels on the training set. This objective can be realized via minimizing

$$L = \sum_{(x,y)\in\mathcal{D}} L_{KD}(S(x, \theta_S, \tau), T(x, \theta_T, \tau)) + \lambda \cdot L_{CE}(\hat{y}, y) \tag{1}$$

where, L_{KD} is distillation (e.g. l_2 or cross-entropy) loss computed between the soft labels of T and S, L_{CE} is cross-entropy loss comparing the ground truth (y) with the prediction $(\hat{y})$ by S, τ is the temperature used in distillation, and λ is a hyper-parameter balancing the loss terms.

3.2. KD with Arbitrary Transfer Set

In absence of the original training data (refer to sec. 1 for such scenarios) multiple approaches (e.g. [15, 17, 2, 1]) compose synthetic transfer set and achieve effective distillation. However, it is clear (Figure 1(a)) that these samples are visually very different from the training samples and hence may not actually lie on the data manifold.

Motivated from these observations, we consider investigating the effectiveness of transfer sets composed of arbitrary samples towards conducting KD. That is, if the transfer sets are composed via random picking of samples from a limited supply of publicly available datasets, as opposed to careful crafting or selection. For instance, in case of the object recognition models trained on CIFAR-10 [9] target dataset, Clevr [8] and SVHN [18] are a pair of candidate arbitrary datasets. Note that the later two are unrelated to the target dataset, i.e., they do not share category labels or similar visual/semantic information with the target dataset.

When we compose a transfer set, generally we attempt to ensure that there are samples from all the classification regions of the *Teacher*. However, it is unlikely that an arbitrarily composed transfer set will have samples from all the classification regions representing the data distribution. That means, the distribution of labels predicted by the *Teacher* can be extremely unbalanced. As a consequence, the decision boundaries learnt by the *Student* model using arbitrary data as a transfer set will be distorted with respect to that learnt using the original training samples (i.e. boundaries learnt by the *Teacher*). For instance, Figure 2 shows the distribution of labels predicted by AlexNet *Teacher* trained on CIFAR-10 for three different arbitrary datasets: Random noise, Clevr and SVHN. It can be noticed that these distributions are far from uniform. This might be attributed to the disproportionate classification regions labelled by pretrained deep models on arbitrary transfer sets. Clearly, it is unlikely that arbitrary data would be able to preserve the same class boundaries learnt using the original training data. Hence, the randomly composed transfer sets are not ideal for performing the distillation (section 4).

Algorithm 1: KD with class balanced arbitrary transfer set

```
Input: Teacher T, Student S, arbitrary datasets :
       $\{D_1, D_2, \dots D_M\}$, intended maximum size
       of the transfer set $N$
Output: Trained Student model weights $\theta_S$,
        $\bar{D}$: Target-class balanced Transfer set
Obtain $C$: Number of categories from the
  output-space dimension of T, and Initialize $\bar{D} = \phi$
Initialize sample counts of each target class in $\bar{D}$:
  $c_i = 0, \forall \ i \ \in \ \{0, 1, \dots C-1\}$
for i=1:M do
    while $\exists \ j$ such that $c_j < \lfloor N/C \rfloor$ and $D_i \neq \phi$ do
        sample $x_k \in D_i$
        $D_i \leftarrow D_i \setminus \{x_k\}$
        $l \leftarrow$ class-label predicted by $T$ on $x_k$
        if $c_l < \lfloor N/C \rfloor$ then
            $\bar{D} \leftarrow \bar{D} \cup \{x_k\}$
            $c_l \leftarrow c_l + 1$
        end
    end
    if $c_j = \lfloor N/C \rfloor \ \forall \ j \ \in \{0, 1, \dots C-1\}$ then
      break;
end
Perform the Distillation via optimizing for $\theta_S^* =$
  $\underset{\theta_S}{\text{argmin}} \sum_{x \in \bar{D}} L_{KD}(T(x, \theta_T, \tau), S(x, \theta_S, \tau))$
```

In light of such important observations, we propose a simple but effective strategy to ensure representation of all the classification regions in the transfer set, which helps to mitigate the distortion in the decision boundaries. While composing an arbitrary transfer set, we enforce it to have a label distribution closer to uniform over the set of target labels by design. Note that the predicted labels would still be completely unrelated to the visual patterns, i.e., one may not expect any semantic/visual similarity between a datapoint in the arbitrary transfer set and the original training set even for the same predicted label. Despite being unrelated samples, we aim to have these spanned uniformly across all the classification regions, thereby forming the 'Target class-balanced' arbitrary transfer set. Algorithm 1 (also Figure 1(b)) shows the steps for composing such a transfer set from a repository of freely available unlabelled datasets using our hypothesis. Note that it may not be possible to get exactly uniform predicted label distribution with the finite supply of arbitrary samples. However, the goal here is to carefully avoid the aforementioned extreme imbalance that creeps in with a random composition. Unlike existing data-free KD approaches such as [17, 4, 1], we do not generate any synthetic samples and use the existing unlabelled datasets in their original forms. While composing the arbitrary transfer set, we only require a single forward pass for each sample through the *Teacher* model and there are no backpropagations involved which makes our proposed algorithm much less compute-intensive, especially when compared to existing methods [17, 4, 1]. After composing such a transfer set ($\bar{D}$), we perform distillation via optimizing the *Student* model (θ_S)

$$\theta_S^* = \underset{\theta_S}{\text{argmin}} \sum_{x \in \bar{D}} L_{KD}(T(x, \theta_T, \tau), S(x, \theta_S, \tau)) \quad (2)$$

where L_{KD} is the distillation objective (cross entropy loss is used in our experiments), τ is the temperature used in softmax layers of T and S. Note that, unlike eq. 1, eq. 2 does not contain the classification loss (L_{CE}) because the

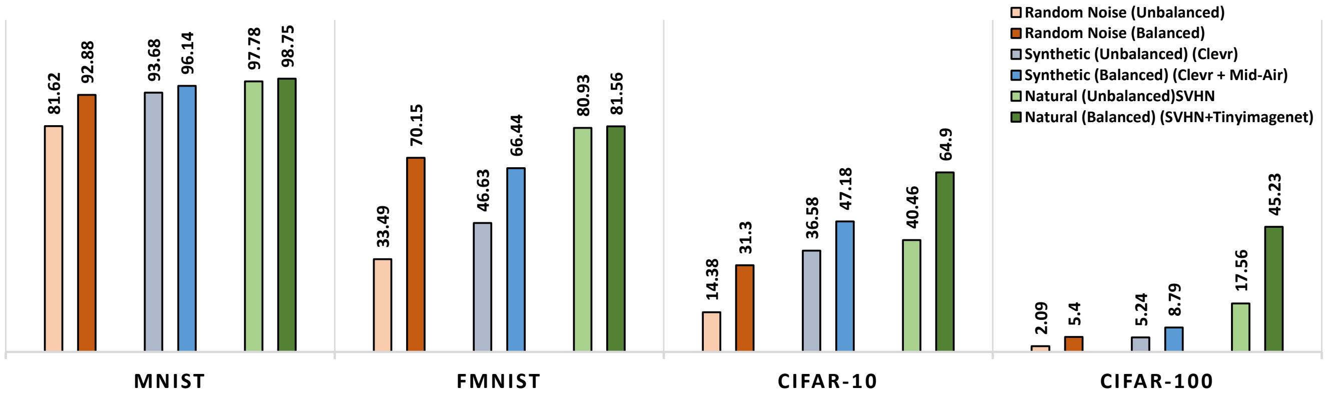


Figure 3. Comparison of the distillation performance using unbalanced and balanced arbitrary transfer sets. Balanced set outperforms its unbalanced counterpart across all the three different varieties of arbitrary datasets: noise, synthetic and unrelated natural data.

transfer set is arbitrary, i.e., unrelated to the target classes and hence forcing hard labels on these samples is counter-intuitive.

4. Experiments

In this section, we empirically demonstrate the importance of target-balanced arbitrary transfer sets as a strong baseline for performing distillation. Before we present the experimental results, we describe the CNN classifiers and the datasets we used in our experiments.

MNIST/FMNIST and LeNet-5: The MNIST [12] dataset contains images of handwritten digits. FMNIST [20] has images of several fashion items. Both the datasets contain 60000 training images, 10000 test gray scale images. Lenet-5 is taken as *Teacher* and Lenet-5-Half as *Student* (identical setting to [14, 17]).

CIFAR-10 and AlexNet/ResNet: CIFAR-10 [9] dataset has colour images of size 32×32. AlexNet [10] is taken as *Teacher* and AlexNet-Half as *Student* to have a fair comparison with [17, 1]. Similar to [4], we also perform experiments with Resnet-34 as *Teacher* model and Resnet-18 as *Student* which are bigger networks in comparison to AlexNet and AlexNet-Half.

CIFAR-100 and Inception-v3: In order to demonstrate the validity of our hypothesis even on large scale datasets, we also experiment on CIFAR-100 [9] which is similar to CIFAR-10 but contains 100 classes instead of 10. Train data has 500 images per class while test data contains 100 images per class. Similar to [1], we take Inception-V3 [19] as *Teacher* and ResNet-18 [6] as *Student*.

We use the following datasets as *Transfer* sets in the absence of original training data:

Random noise: Uniform random noise in $[0, 1]$ for each pixel in an image is used to construct transfer set for MNIST, FMNIST and CIFAR-10. For CIFAR-100, we found that creating balanced set using Gaussian noise $(\mu = 0.5, \sigma = 0.1)$ (clipped to $[0, 1]$) is faster.

Synthetic datasets: We use publicly available synthetic dataset 'Clevr' [8] as unbalanced transfer set. This dataset contains 70000 images. In order to improve the class-balance of Clevr, we add another synthetic dataset, called Mid-Air [5] on top of it. This dataset contains images under different climate conditions. We resize each image to 32×32.

Natural datasets: We utilize SVHN [18] dataset as unbalanced arbitrary natural transfer set. It has 73257 images of street view house numbers. We use the cropped version of the dataset where each image is resized to 32×32. To achieve class balance while keeping the "naturalness" of images in the transfer set, we add samples from TinyImageNet [11] on top of it.

4.1. Arbitrary Transfer Sets for Distillation

In all our experiments, size of the transfer set is kept approximately equal to that of corresponding *Teacher*'s training set. This enables a fair comparison of the distillation performance between arbitrary and training datasets. The exact size of the transfer sets used in the experiments along with exact number of arbitrary samples labelled in each of the target classes is provided in the supplementary material. We present and analyse the experimental results separately for each of the different stimuli.

Random noise stimuli: We consider the following two scenarios: (i) Unbalanced: randomly sampled noise samples, (ii) Balanced: noise set on which *Teacher* predicts all the target labels almost uniformly. In Fig. 3, the first two bar graphs for each dataset show the distillation performance of the random noise transfer set. In the case of MNIST, we observe that even an arbitrary random noise gives a decent performance of 81.62%. By class-balancing the random noise, we can increase the accuracy to 92.88%. In the case of FMNIST and CIFAR-10 we get a significant improvement of close to 36% and 17% respectively on target-class balancing while 3.3% gain in case of CIFAR-100.

Transfer Set	Balanced	MNIST		FMNIST		CIFAR-10		CIFAR-100	
		w/o Aug	w/ Aug	w/o Aug	w/ Aug	w/o Aug	w/ Aug	w/o Aug	w/ Aug
Random Noise	✗	81.62	89.01	33.49	38.37	14.38	47.50	2.09	3.30
Random Noise	✓	92.88	95.76	70.15	74.33	31.30	67.40	5.40	18.20
Clevr	✗	93.68	97.11	46.63	72.68	36.58	72.35	5.24	17.45
Clevr+Mid-Air	✓	96.14	98.53	66.44	83.38	47.18	75.76	8.79	22.28
SVHN	✗	97.78	98.81	80.93	83.85	40.46	72.33	17.56	40.59
SVHN + Tiny	✓	98.75	98.96	81.56	84.75	64.90	79.19	45.23	67.18

Table 1. Effect of Augmentation: Distillation performance using unlabelled arbitrary transfer sets on multiple datasets with and without augmentation (Tiny stands for 'TinyImageNet').

Algorithm	MNIST	FMNIST	CIFAR-10	CIFAR-100
Teacher	99.34	90.84	83.03	79.05
Student-KD [7]	99.25	89.66	81.78	69.65
ZSKD [17]	98.77	79.62	69.56	–
DeGAN [1]	–	83.79	**80.55**	65.25
Ours (SVHN + TinyImageNet)	**98.96**	**84.75**	79.19	**67.18**

Algorithm	CIFAR-10
Teacher (ResNet 34)	95.58
Student-KD (ResNet 18) [7]	94.34
DAFL [4]	92.22
Ours (TinyImageNet + SVHN)	**92.92**

Table 2. Comparison with SOTA : Performance of proposed method in comparison with ZSKD [17] and DeGAN [1] (table on the left), & DAFL [4] (table on the right)

Synthetic stimuli: For the unbalanced case we consider the Clevr dataset to perform the distillation without looking at their predicted labels by the *Teacher*. For the class-balanced case, we add samples from another synthetic dataset, Mid-Air, towards obtaining approximately equal number of samples in each of the target classes, as described in Algorithm 1. The third and fourth bar graphs (for each dataset) in Figure 3 show the distillation performance of the synthetic stimuli. We get a decent improvement of 2.46% and 3.55% in case of MNIST and CIFAR-100 while significant gain of 19.81% and 10.60% for FMNIST and CIFAR-10 respectively by using a class-balanced Synthetic dataset.

Natural image data stimuli: We consider SVHN as the arbitrary transfer set. For achieving class-balance, we use samples from TinyImageNet which are added on top of SVHN as described in Algorithm 1. The last two bar graphs for each dataset in Figure 3 show the results with natural data as transfer set. We get a small gain of 0.97%, 0.63% for MNIST and FMNIST due to the (relatively) low imbalance while a significant gain of 24.44% and 27.67% for CIFAR-10 and CIFAR-100 respectively due to high imbalance in target classes, where class balancing has a profound effect.

4.2. Augmentation Helps the Underrepresented Classes

Note that in all the three scenarios, practically it is very difficult to achieve perfect class-balance (identical number of samples in each target class) with limited supply of arbitrary data. We noticed that more frequently the classification regions learned by the deep models are heavily out of proportion and that makes it difficult to have arbitrary samples representing all the target classes equally. However, the objective of achieving better distillation performance is to reduce the extreme class imbalance and compose a transfer set that represents all the classification regions. Therefore, after achieving some level of balance (as in Algorithm 1), we further improve the representation from the under populated classes via performing augmentations during the distillation process. Our augmentation includes scaling, rotation, flipping, adding random noise (Gaussian, salt and pepper), and multiple combinations of them. Although augmentations add diversity to all the target classes, underrepresented classes get more benefited than the relatively better populated ones resulting in further less distortion of the class decision boundaries and hence, improved distillation performance. Table 1 shows the results with and without augmentation across all the transfer scenarios over multiple datasets. It is evident that augmentation consistently results in better distillation performance via boosting the underrepresented classes in the transfer set.

4.3. Comparison With the State-of-the-art

In this subsection, we compare the proposed baseline against the state-of-the-art data-free knowledge distillation approaches. Table 2 (on left) presents the comparison of distillation performance of the proposed approach against ZSKD [17] and DeGAN [1] on multiple datasets. In order to have a fair comparison, we experimented with the same models used in [17] and [1]. The proposed baseline clearly performs better on MNIST, FMNIST and CIFAR-

100. In case of CIFAR-10, it is close to DeGAN's [1] performance. But, DeGAN achieves better performance on CIFAR-10 when samples from the classes of CIFAR-100 are used. These CIFAR-100 samples are substantially more similar to CIFAR-10 than our arbitrary dataset (SVHN+TinyImageNet), thereby resulting in this improved distillation performance. If an unrelated dataset is used for distillation from the *Teacher* network on CIFAR-10, proposed baseline outperforms the DeGAN significantly (as shown in Figure 4). We have also compared our proposed baseline with DAFL [4]. Again, to have a fair comparison, we have used the ResNet-34 *Teacher* and ResNet-18 *Student* used in [4]. The performance comparison is shown in Table 2 (on right). Our baseline on arbitrary/unrelated transfer set performs slightly better. Further, unlike DAFL, it does not require any complicated GAN training.

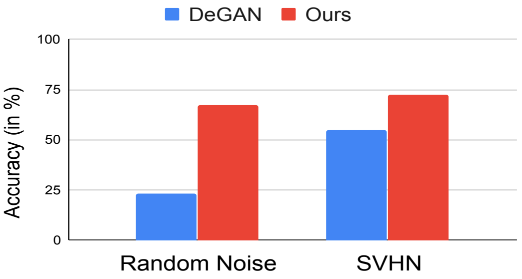


Figure 4. Comparison of our proposed approach with DeGAN [1] when unrelated transfer sets are used to distill the knowledge from *Teacher* model trained with CIFAR-10.

4.4. Strict Arbitrary Transfer Sets : Explicit Removal of Overlapping Classes

In this subsection, we explicitly make sure that the arbitrary transfer sets do not have any overlap with the target categories and is semantically very dissimilar with the target data samples. We investigate the true potential of such strict arbitrary transfer sets when they are *target-class balanced* and used for knowledge distillation in complete absence of original training data.

We chose TinyImageNet as arbitrary data in section 4.1 due to its relatively larger size, which can help to make the transfer set better balanced. Also, this dataset is a widely used publicly available dataset. However, it may contain a few overlapping classes with CIFAR. Therefore, we further perform experiments to show that the distillation performance is close to that using TinyImageNet even if we consider arbitrary datasets that do not share any categories.

SVHN contains images of digits which are dissimilar to CIFAR-10 and are not balanced across target classes of trained *Teacher* model on CIFAR-10. Several different unrelated datasets are added on top of SVHN to improve the target-class balance using our proposed Algorithm 1. Instead of using TinyImageNet which may have samples related to target categories, we add CelebA [13] dataset on top of it. CelebA has images of celebrity faces and does not have any overlap with CIFAR-10 classes. Please note that in all our experiments we make sure that the number of samples in arbitrary transfer set do not exceed the per class sample count of original training data in order to have a fair comparison. The results are reported in Table 3 where completely unrelated arbitrary transfer sets along with augmentations are used for distillation. SVHN mixed with CelebA when used as a transfer set, gives a decent accuracy of 75.87 %. We further observe gain in the distillation performance when the target-class balance is improved by mixing another unrelated dataset i.e. Fruits360 [16] that contains images of fruits and vegetables. Please note that it is still not perfectly target-class balanced and one can further improve the accuracy by adding more non-overlapping datasets and ensuring equal amount of arbitrary samples in each of the target classes labelled by the pretrained *Teacher*.

Transfer Set	Distillation Accuracy
SVHN (Digits) + CelebA (celebrity faces)	75.87 %
SVHN + CelebA + Fruits360 (fruits & vegetables)	76.33 %
SVHN + CIFAR-100 (non-overlapping)	79.13 %
SVHN + TinyImageNet	79.19 %

Table 3. Distillation performance when completely unrelated and non-overlapping arbitrary transfer sets are used to distill knowledge from *Teacher* model trained on CIFAR-10.

DeGAN [1] avoids the overlapping classes of CIFAR-100 with CIFAR-10 and reports results using the non overlapping 90 classes of CIFAR-100. For a strictly fair comparison with DeGAN, we also take the same non-overlapping CIFAR-100 samples to balance the transfer set synthesized using SVHN samples. From Table 3, we observe similar performance (SVHN+TinyImageNet) even by using the non-overlapping CIFAR-100 as the arbitrary dataset. However, since CIFAR-100 samples are added on top of SVHN, we effectively utilize only 18818 samples from CIFAR-100 as opposed to DeGAN which uses all the 45000 samples.

We, thus, empirically observe that it is possible to achieve decent distillation performance even with strictly arbitrary transfer sets (completely unrelated to and non-overlapping with original training data), when these transfer sets are target class-balanced. However, one can further improve the distillation performance by carefully selecting the arbitrary data sources by leveraging the domain knowledge and the knowledge of the task at hand while utilizing our proposed strategy (see Algo 1) to compose the transfer sets.

Size of Transfer Set (SVHN)	**Binary-MNIST** (Teacher Accuracy: 99.44%)		**Binary-FMNIST** (Teacher Accuracy: 93.38%)	
	Unbalanced	Balanced	Unbalanced	Balanced
16000	88.89 ± 2.05	**89.70** ± 1.86	75.04 ± 1.83	**77.68** ± 1.15
8000	72.60 ± 2.41	**82.69** ± 1.73	76.28 ± 1.30	**78.10** ± 0.66
4000	84.66 ± 2.27	**91.38** ± 1.81	77.58 ± 0.67	**79.94** ± 0.89
2000	**83.84** ± 3.06	82.93 ± 3.00	68.35 ± 2.10	**73.37** ± 0.54
1000	81.48 ± 1.32	**82.38** ± 1.35	72.86 ± 0.89	**75.21** ± 0.47
500	**83.67** ± 0.25	83.12 ± 1.23	72.95 ± 0.70	**74.67** ± 0.31

Table 4. Comparison of distillation performance (i.e., Student accuracy in %) with unbalanced and balanced arbitrary transfer set (SVHN), when the Teacher network is trained on unbalanced binary MNIST and FMNIST training samples.

4.5. Unbalanced Target Dataset: Generality of the Proposed Strategy

Until now we have experimented with target datasets that are class-balanced. That is, our *Teacher* models are trained on equal number of samples from each class of the target dataset. In this subsection, we demonstrate the effectiveness of the balanced arbitrary transfer sets towards KD even when *Teacher* is trained on unbalanced target (training) datasets. For this purpose, we have created binary classification tasks out of MNIST and FMNIST datasets separately, referred to as *Binary-MNIST* and *Binary-FMNIST* repectively. We merged samples from three of the ten classes (labels $0, 1$, and 2) into one set, referred to as the 'minority' class and the rest seven into another set, referred to as the 'majority' class. Therefore the resulting binary classification datasets will have a $3 : 7$ class imbalance. We train LeNet *Teacher*s on the corresponding datasets with **balanced** mini-batches. Trained *Teacher*s report test accuracies of 99.44% and 93.38% respectively. Note that the test sets comprise equal number of majority and minority class samples from the corresponding original test data. Since there would be more test datapoints for the majority class, we picked samples equal to that in the minority class test set. Also, in order to ensure maximal diversity within the 'majority' test set, equal number of samples from the constituent MNIST (or FMNIST) labels (3 to 9) are considered.

We then conduct distillation with (i) random (unbalanced), and (ii) balanced arbitrary transfer sets and present the *Student*'s accuracy on the test sets in Table 4. Note that the transfer sets of varying size (from 500 to 16000) are composed from the SVHN dataset and the accuracies are reported across 20 runs. We can clearly observe that the target-balanced arbitrary transfer sets outperform the randomly composed counterparts in vast majority of the cases, thereby validating the generality of the proposed baseline.

5. Conclusion

Distillation enables a low capacity model (*Student*) to learn a sophisticated mapping which is not possible otherwise (via normal cross entropy training). In order to cope with the constrained operational conditions, recent efforts [15, 4, 17, 1] attempt to distill in a data-free scenario via artificially generated transfer set. Despite using out-of-distribution samples that are visually far away from the actual training data, these methods have reported competitive distillation performance. Motivated by these observations, in this work, we explore (i) if a simple baseline can be obtained for data-free KD by leveraging publicly available arbitrary data, and (ii) whether this baseline can be an alternative to the substantially more complex approaches. Further, we presented a simple strategy based on intuitive hypothesis to maximize the transfer performance of such sets. Upon extensively experimenting with multiple datasets and model architectures, we bring out the following observations:

- Arbitrary (unrelated to the target data) transfer sets can be leveraged to deliver competitive KD performance, when compared with the computationally expensive state-of-the-art data-free distillation methods. Thus, such transfer sets can lead to the design of important baselines for the data-free knowledge distillation task.
- For any arbitrary transfer set, being 'target class-balanced' maximizes the transfer performance.
- Though class-balancing improves the transfer performance, it depends on the similarity of the transfer set to the original training data. In other words, as the transfer set lies closer to the target data manifold, knowledge transfer improves. (Please refer to the supplementary materials).

Hinton *et al.* [7] attributed the effectiveness of distillation process to the dark knowledge extracted out of the training data. However, it is very intriguing to understand how even a completely unrelated transfer set with only the distillation objective can help a *Student* to achieve competitive generalization. Also, in the data-free KD setting, it needs to be investigated how the similarity of an arbitrary dataset with the target data distribution can be estimated, especially when multiple such datasets are available for facilitating knowledge distillation. We leave these two aspects of the data-free distillation for future research..

References

[1] Sravanti Addepalli, Gaurav Kumar Nayak, Anirban Chakraborty, and R. Venkatesh Babu. DeGAN : Data-Enriching gan for retrieving representative samples from a trained classifier. In *Proceedings of the AAAI*, 2020.

[2] Kartikeya Bhardwaj, Naveen Suda, and Radu Marculescu. Dream distillation: A data-independent model compression framework. In *ODML-CDNNR Workshop at ICML*, 2019.

[3] Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *International Conference on Knowledge discovery and Data mining*. ACM, 2006.

[4] 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.

[5] Michael Fonder and Marc Van Droogenbroeck. Mid-air: A multi-modal dataset for extremely low altitude drone flights. In *Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, June 2019.

[6] 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*, pages 770–778, 2016.

[7] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *Deep Learning Workshop at NIPS*, 2014.

[8] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017.

[9] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Canadian Institute for Advanced Research, 2009.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[11] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7, 2015.

[12] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[13] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[14] 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.

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

[16] Horea Mureşan and Mihai Oltean. Fruit recognition from images using deep learning. *Acta Universitatis Sapientiae, Informatica*, 10(1):26–42, 2018.

[17] 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.

[18] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning, 2011.

[19] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[20] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.