Rectifying the Data Bias in Knowledge Distillation

Boxiao Liu[1,2], Shenghan Zhang[3], Guanglu Song[3], Haihang You[1,2], and Yu Liu[3,†]

[1]State Key Laboratory of Computer Architecture, Institute of Computing Technology, CAS
[2]University of Chinese Academy of Sciences
[3]SenseTime Research

{liuboxiao,youhaihang}@ict.ac.cn,
{zhangshenghan,songguanglu}@sensetime.com,liuyuisanai@gmail.com

Abstract

Knowledge distillation is a representative technique for model compression and acceleration, which is important for deploying neural networks on resource limited devices. The knowledge transferred from teacher to student is the mapping of teacher model, or represented by all the input-output pairs. However, in practice the student model only learns from data pairs of the dataset that may be biased, and we think this limits the performance of knowledge distillation. In this paper, we first quantitatively define the uniformity of the sampled data for training, providing a unified view for methods that learn from biased data. Then we evaluate the uniformity on real world dataset and show that existing methods actually improve the uniformity of data. We further introduce two uniformity-oriented methods for rectifying the bias of data for knowledge distillation. Extensive experiments conducted on Face Recognition and Person Re-identification have shown the effectiveness of our method. Moreover, we analyze the sampled data on Face Recognition and show that better balance is achieved between races and between easy and hard samples. And this effect can be also confirmed in training the student model from scratch, resulting in a comparable performance with standard knowledge distillation.

1. Introduction

Deep neural networks have achieved remarkable success in many vision tasks, such as image classification [16, 24, 10], object detection [36, 26, 5, 41], and even under imperfect labels [11, 12]. However, the performance of lightweight networks is limited compared with large ones. To improve the performance of light networks, knowledge distillation is a simple yet effective way by transferring knowledge from the big teacher network into the light student. Since being first introduced in [18], knowledge distillation itself has witnessed notable advances [52, 27]. Many existing works focus on making the student approximate different aspects of the teacher's behavior, such as classification

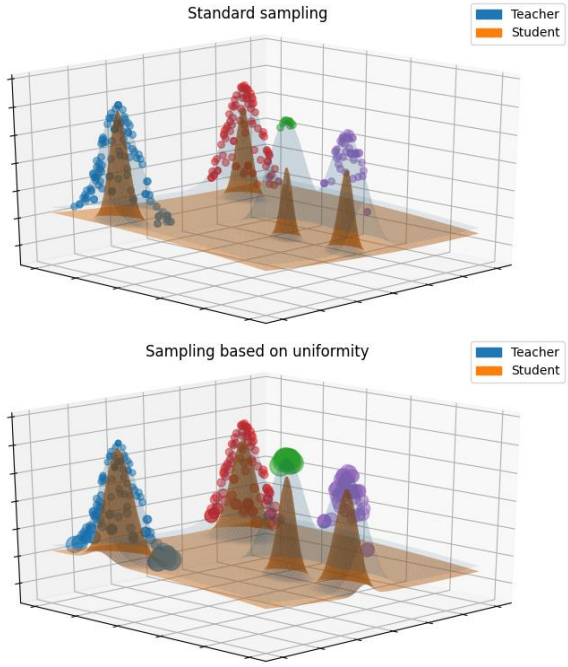

Figure 1. An qualitative illustration for the influence of data bias on knowledge distillation. The data points indicate training samples in dataset, and the points are in the same color if belonging to one class. The radius of the data point indicates the number of sampled times. With the proposed methods based on uniformity, the student can benefit from bias-rectifying and approximate the mapping of teacher better across the input space.

†Corresponding author.

logits, intermediate feature maps [21], or even optimization trajectories [23]. Yet the essence of the transferred knowledge still lies on the mapping of teacher from input space into output space, as like the initial definition in [18].

To make it more concrete, we represent the knowledge by all the pairs of input and its corresponding output of the teacher model. We can not utilize all the pairs to train the student, as it is impossible to directly sample data from the domain of teacher model, i.e. the set of all meaningful images in computer vision tasks. A common practice is to evenly sample a list of images from a collected dataset. Thus the actually transferred knowledge is limited by the dataset. It is expected the dataset is evenly distributed in the domain, so as to help the student approximate the teacher's output across different inputs.

However, this is often not the case. The datasets we have access to are often biased, especially for large-scale automatically generated ones [42, 57]. For example, the popular MS1MV3 [14, 9] dataset for face recognition is collected by searching names of celebrities in the search engine and gathering the retrieved images. As shown in Figure 1, the number of images for each identity varies from 2 to more than 200. Also, clear front face images dominate the dataset while images with large pose are rare. This bias may result in that the student fits the teacher's output well in the subspace containing many input samples, but poorly in the subspace where few samples appear.

Learning from biased data has been studied for a long time in the deep learning area [15, 39, 4]. Some existing works focus on imbalanced classification, and propose to resample the input data [40, 3] or adjust the loss for different samples [26, 48]. Some researchers pay attention to mining hard yet informative samples among easy ones to accelerate the convergence of neural network [26]. For either the classes with few images or the hard samples, we can generally denote them by minorities, then the basic idea behind the methods mentioned above is to over-sample or overweight the minorities. However, emphasizing the minorities too much may bring the model a risk of overfitting [7]. So these methods heavily rely on parameter tuning and could introduce extra computation burden during training process.

In this paper, we focus on knowledge distillation and introduce a unified view towards bias-resistant learning. We find that the balance between classes or between easy and hard samples can be seen as special cases of the uniformity of data, which is important to ensure the knowledge distillation performance. Moreover, to rectify the bias in data equals to improve the uniformity. Motivated by this observation, we propose a quantitative and intrinsic definition of the uniformity in this paper. With this definition, we can identify the bias in data of various types and avoid the overfitting risk, as both of them will bring about a lower uniformity. Towards improving the uniformity of data directly, we further propose two novel methods, called Extrinsic Sampling and Intrinsic Sampling, based on reusing the data in existing biased dataset but changing the sample strategy. Both of our methods are simple and robust, and can be easily integrated in other methods.

We conduct extensive experiments on large-scale face recognition and person re-identification. With the proposed methods, the performance of student model is significantly boosted compared with baselines and other bias-resistant training methods. Notice that our methods are simple and cost-free during training. By visualizing the sampled data, we find that better balance between classes are achieved, as well as between easy and hard samples. This helps the student approximate the teacher better in the sparse subspace.

To sum up, the contribution of this paper are four folds:

1. We provide a unified view towards different types of bias based on uniformity, and introduce a quantitative and intrinsic definition. With this metric, we can evaluate the bias of data and avoid over-emphasizing the minorities.

2. We propose two novel sampling methods, Extrinsic Sampling and Intrinsic Sampling. Our methods are simple yet effective, and can be easily integrated in other methods.

3. Extensive experiments are conducted to show the effectiveness of proposed methods. By simply changing the sample strategy, the performance of the student model is significantly boosted on several tasks.

4. We further analyze the sampled list and provide some insight on how the proposed methods help to improve the student's performance. We also show that the effectiveness can generalize to training the student from-scratch, approximating the performance of standard knowledge distillation.

2. Related Work

2.1. Knowledge Distillation

Knowledge distillation was first introduced in [18] to compress the knowledge of an ensemble of models into a single model, which now denoted as the teacher model and the student. Following researches mainly focus on the three key components in knowledge distillation, distilled knowledge [21, 1, 52, 27, 23], distillation algorithm [35, 17], and teacher-student architecture [25, 51, 13]. Also, extensions inspired by knowledge distillation have been proposed, such as mutual learning [55], assistant teaching [32] and self-learning [53]. The effectiveness of knowledge distillation have been verified in many areas in deep learning, such as image classification [18], object detection [22, 45, 46], face

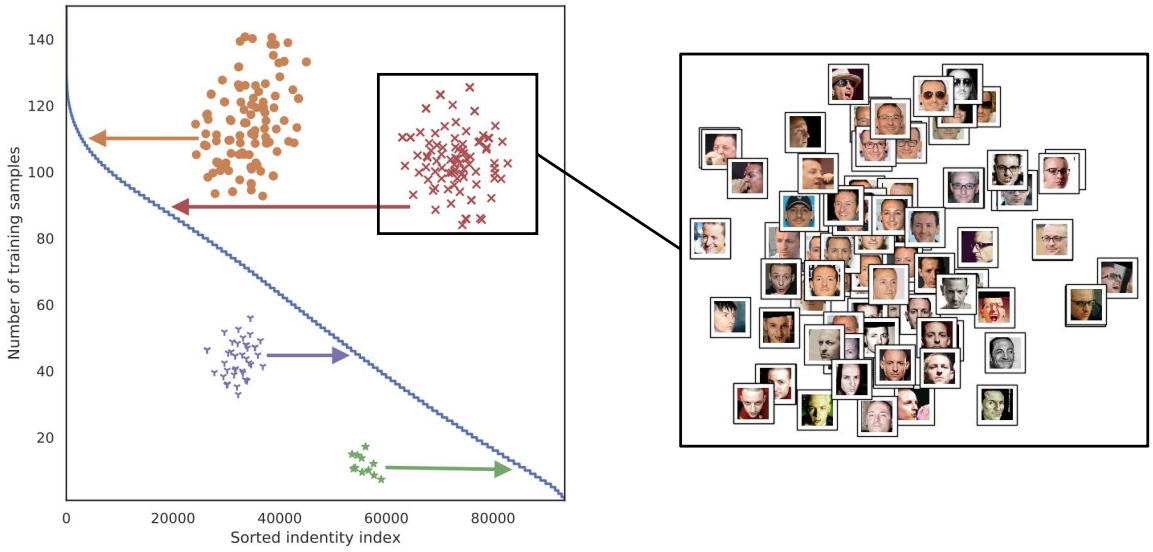

Figure 2. A demonstration of the skewed distribution of data in real-world dataset. The data points are images sampled from MS1MV3, and we projecte them down to the plane by t-SNE, according to the teacher's output.

recognition [54, 47], etc. However, how the data distribution affect the distillation process is under-explored. In this paper, we propose to rectify the bias in training data and demonstrate the benefit of improving uniformity of data.

2.2. Learning from Biased Data

Many real-world datasets exhibit biased distribution of class labels and difficulty, especially large-scale ones [42]. A number of studies have aimed at alleviating the challenge of biased data [15, 39, 4, 44, 43, 33]. Most existing algorithms focus on two ways: re-sampling [2, 3] and re-weighting [19, 48]. In re-sampling, the minor classes or examples are repeated and the frequent ones are under-sampled, which can lead to over-fitting [7]. Thus stronger data augmentation is needed to improve the diversity of minorities [58]. Another way is to re-weight the samples in a cost-sensitive manner. Cui *et al*. [7] propose a novel method to measure data overlap, and a re-weighting scheme is further designed to re-balance the loss, called Class-Balanced Loss. Cao *et al*. [4] regularize the minority classes more strongly than the frequent classes through adjust the margin in loss, and introduce Label-Distribution-Aware Margin loss. Re-weighting methods make the optimization process more difficult and can result in poor performance on frequent data [20]. To alleviate the issue of overfitting, we provide a metric to evaluate the uniformity of data that can ensure a balance between minor and frequent ones. Further, two novel uniformity improving methods are introduced to rectify the data bias, which improve the performance of student in knowledge distillation.

3. The Uniformity of Data

3.1. Preliminaries

In knowledge distillation, we denote the teacher model by a function $f_t : \mathbb{R}^d \to \mathbb{R}^n$ that maps an input x into some output y. The student model is denoted by f_s as like. The knowledge transferred from teacher to student is defined as the mapping f_t itself, and the student is optimized to mimic the teacher's output across the domain of teacher model D_{f_t}. Then the optimization target of knowledge distillation can be written as

$$L_{\text{KD}} = \mathbb{E}_{x \sim D_{f_t}}[\mathcal{M}(f_t(x), f_s(x))], \tag{1}$$

where $\mathcal{M}$ denotes some metric evaluating the distance between the outputs of f_t and f_s.

In practice, we usually replace D_{f_t} by a uniform distribution on the dataset for training, and we denote this distribution by X_{train}. The actually used loss function for knowledge distillation is

$$L_{\text{KD}} = \mathbb{E}_{x \sim X_{train}}[\mathcal{M}(f_t(x), f_s(x))]. \tag{2}$$

We expect the distribution X_{train} to be uniformly distributed in D_{f_t}. However, some datasets are collected without considering this uniformity constraint and may be sig-

nificantly biased, especially for large scale open-set learning. For example, the MS1MV3 [9] dataset used for training face recognition models is collected by searching the names of celebrities in the search engine and gathering the retrieved images. The images collected for a celebrity is affected by the age, popularity and even occupation.

To show the bias in MS1MV3, we sort the identities according to the number of images and present the result in Figure 2. The distribution of numbers is obviously skewed. Also, we project the embedding of images sampled from several identities into 2-dimension plane in Figure 2 using t-SNE [30]. It can be seen that easy and frontal faces gather together and form the majority of the input data.

3.2. The Definition of Uniformity

Before the definition of uniformity, we first define the distance of the input space of neural network. For vision tasks, the input space is usually the RGB space. It is hard to define a meaningful distance between two images based on the RGB value of pixels. Thanks to the knowledge distillation scheme, we can easily get the representation of each image by the powerful teacher model. Then, the distance in the teacher's output space, i.e. the KL divergence between two distributions or the angle between two normalized vectors, can serve as a good metric.

Formally, we denote the distance of images x_i and x_j as $D(f_t(x_i), f_t(x_j))$, in which f_t indicates the teacher model and D denotes specific distance metric in the output space.

Based on the distance between data, we need to estimate the continue probability density for each image from discrete training samples. We introduce a non-parametric estimation method in statistics called Kernel Density Estimation [38, 34]. To be specific, the estimated probability density is given by

$$\hat{f}_h(x) = \frac{1}{E}\sum_{i=1}^{n} K(\frac{D(x, x_i)}{h}). \tag{3}$$

In Eq 3, n is the size of the training set, E is the normalization factor and K denotes a kernel function. In this paper, we choose Gaussian basis function as the kernel function. h, also called bandwidth, is a smoothing parameter to balance between the bias of the estimator and its variance. With a proper h, we can get an approximation of the density distribution of the training set.

Finally, we define the uniformity as the entropy of estimated density distribution of the training set, which is

$$U(X) = -\sum_{x_i \in X} \hat{f}_h(x_i) \log \hat{f}_h(x_i). \tag{4}$$

For clearity, we use a simple 1-dimension example to show the effectiveness of the uniformity above. The first row in Figure 3 shows the histogram of random samples

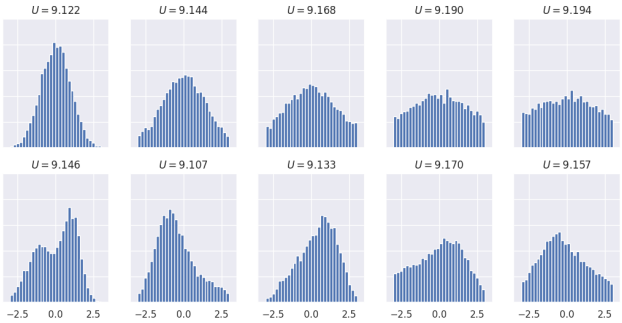

Figure 3. The 1-dimensional example to show the effectiveness of proposed uniformity.

from a truncated Gaussian distribution with increasing variance. As the distribution become more similar to uniform distribution, the uniformity increases monotonously, which shows the validity of the proposed definition. We also demonstrate several Gaussian mixture distribution examples in the second row with their estimated uniformity for comparison.

3.3. Bandwidth Selection

The hyper-parameter h in the definition of uniformity exhibits a remarkable influence on the estimated result. In traditional non-parametric estimation, a rule-of-thumb estimator for h is

$$h = 0.9\sigma n^{-1/5}, \tag{5}$$

if the underlying density is Gaussian. The σ in Eq 5 indicates the standard deviation of the data.

For real world datasets, the data are high-dimensional vectors and the distribution is sophisticated. We observe that the output of the teacher is Gaussian-like for samples in a single class. Thus we propose to estimate the bandwidth for each class and use the average for the final selection. The σ_i for class i is

$$\sigma_i = \sqrt{\frac{1}{m_i}\sum^{m_i} D(x, \bar{x}_i)^2}, \tag{6}$$

where m_i indicates the number of samples and $\bar{x}_i$ indicates the estimated center for class i. The final h is computed as

$$h = \frac{1}{n}\sum_{i}^{n} 0.9\sigma_i m_i^{-1/5}. \tag{7}$$

Notice that the estimated h can only serve as a good start point for parameter tuning instead of the optimal choice.

3.4. Evaluation on Real World Dataset

In this subsection, we apply the uniformity in real world datasets. We use large-scale Face Recognition as example.

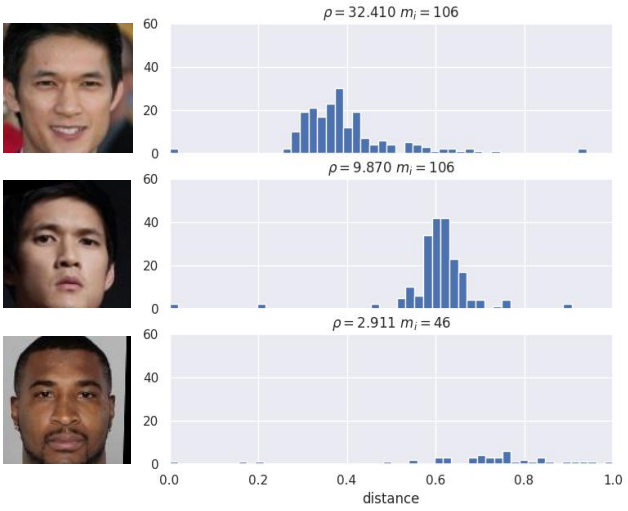

Figure 4. The histogram of distance between an image and others in the same class, and the corresponding images are shown at the left.

Specifically, we choose MS1MV3 [9], a semi-automatic refined version of the MS-Celeb-1M [14] dataset, as the training set and the teacher model is ResNet50 [16] trained on MS1MV3. The output of the teacher model is a 512-*D* embedding feature normalized by its L2 norm. It is easy to see that the output space of the teacher model is a high-dimensional hyper-sphere. The distance in the output space is set to the angle between two embeddings of images, as in

$$D(x_i, x_j) = \arccos\left(\frac{\langle f_t(x_i), f_t(x_j)\rangle}{||f_t(x_i)||_2||f_t(x_j)||_2}\right). \quad (8)$$

Then we put the definition of D into the Equation 3 to estimate the density of each training image and then compute the uniformity according to Equation 4. h is set to $12°$.

Firstly, we show the distribution of distance between several probe images with all the others in Figure 4. The first two probe images are sampled from an identity with about 100 images, while the third image from an identity with only 46 images. It can be observed that the distance distribution for images of different identities and different difficulties varies significantly, and the corresponding estimated density provides a way to discriminate between these images.

Then, we evaluate the uniformity of the dataset and its variant in Table 1. The variant is constrained by uniform sampling on identities. To be specific, we first random sample an identity, then sample an image form this identity. The number of images for each identity in the resampled list is about the same. Note that the computation of the uniformity is a little different with Equation 3. For the resampled list, we count the numbers of sampled times and normalize them to be the weight ω_i for image x_i. Then the distance is multiplied by the weight of data, which result in

method	USI	ES	IS
uniformity	+0.032	+0.089	**+0.093**

Table 1. Comparison of the uniformity for different methods. "USI" indicates uniform sampling on identities, "ES" indicates Extrinsic Sampling and "IS" indicates Intrinsic Sampling.

$$\hat{f}_h(x) = \frac{1}{E}\sum_{i=1}^{N}(\omega_i K(\frac{D(x, x_i)}{h})). \quad (9)$$

In Table 1, we can see that uniformly sampling on identities actually lead to larger uniformity, and we believe this is the essence behind its effectiveness. However, it only considers one aspect of the bias in dataset, and is proposed without considering the uniformity. This leads to an inferior effect for improving uniformity. In this paper, we propose two methods oriented to achieve uniform sampling, which will be detailed in the next section.

4. Uniformity Improvement

For learning knowledge from the teacher better, the training data for the student should be uniformly distributed in the support set. However, the datasets are usually collected without considering the uniformity, especially in large scale metric learning tasks. To improve the uniformity of the dataset, we propose two sample-based methods, which we called Extrinsic Sampling and Intrinsic Sampling.

4.1. Extrinsic Sampling

For special high-dimension space, i.e. hyper-sphere, there exists an easy way to generate a list of vectors that are ensured to be uniformly distributed in the space [50]. Formally, let $a = [a_1, a_2, \ldots, a_n]$, where $a_i \sim \mathcal{N}$ and let $a^* = \frac{a}{||a||_2}$, then a^* is uniformly distributed on the unit hyper-sphere.

So, we propose to find the nearest embedding vector a' to a^*, and use the corresponding image I' as the training data, instead of I^*. The resampled image I' is an approximation of the original I^*, and is more uniformly distributed than the training set. By repeating this sample process, we can generate a list of resampled images and replace evenly choosing from the dataset.

4.2. Intrinsic Sampling

Extrinsic Sampling is a simple and parameter-free method to generate more uniform training data. However, it may be very hard to find a way to generate uniform vectors in some output space. Also, the resampled X' can be easily affected by the distribution of the dataset. To alleviate these issues, we propose a new Intrinsic Sampling method, which only depends on the distance in the output space.

Student Model	Use KD	Use ES	Use IS	IJB-B 1e-5	IJB-B 1e-4	IJB-B 1e-3	IJB-C 1e-5	IJB-C 1e-4	IJB-C 1e-3
MobileFaceNet				88.30	93.02	95.59	92.19	94.75	96.74
MobileFaceNet	✓			89.45	93.48(+0.46)	95.74	92.93	95.19(+0.44)	96.78
MobileFaceNet	✓	✓		89.38	93.60(+0.58)	95.88	93.20	95.39(+0.64)	97.03
MobileFaceNet	✓		✓	89.35	93.76(+**0.74**)	96.04	93.25	95.49(+**0.74**)	97.05
R18				87.92	92.86	95.33	91.96	94.72	96.50
R18	✓			88.62	93.07(+0.21)	95.53	92.54	94.90(+0.18)	96.67
R18	✓	✓		88.08	93.40(+**0.54**)	95.76	92.61	95.08(+**0.36**)	96.81
R18	✓		✓	88.52	93.22(+0.36)	95.70	92.68	95.04(+0.32)	96.74

Table 2. Results on IJB-B and IJB-C datasets. "KD" indicates Knowledge Distillation, "ES" indicates Extrinsic Sampling and "IS" indicates Intrinsic Sampling. The improvement compared with the baseline of training from scratch is shown in brackets.

To be formal, we first estimate the density $\{\hat{f}_h\}$ of training set as in Eq 3. Notice that the bandwidth h here can be different with that in evaluation of the uniformity. As the training data with bigger density should be sampled less, we get the sample probability of training data x_i by

$$p_i = \frac{1}{E} \cdot \frac{1}{\hat{f}_h(x_i)}, \tag{10}$$

where $E = \sum_{x_i \in X} 1/\hat{f}_h(x_i)$ is the normalization factor to ensure p is a valid probability. Then we can generate the resampled list according to the probability of each training data.

5. Experiments

We evaluated our methods on the large-scale metric learning tasks, including face recognition and person re-identification. We first show the effectiveness of the proposed methods, and examine the influence of parameter h for Intrinsic Sampling. Then, comparison with existing methods are provided. Finally, we show some visualization examples of the resampled data and discuss the generalization ability of our methods on training from scratch.

5.1. Implementation Details

Face Recognition The training dataset we use is a semi-automatically refined version of MS-Celeb-1M dataset [14], named MS1MV3 [9]. MS1MV3 contains about 93K identities and 5.2M images. To evaluate the performance, we employ IJB-B [49] and IJB-C[31], and th $Tar@Far = \{1e-3, 1e-4, 1e-5\}$ is reported.

We follow the data pre-possessing method proposed in ArcFace [8] to get the aligned face crops. A ResNet-like [16] network R50 is used as the teacher model, and we choose MobileFaceNet [6] and R18 as the student model. The input size of the network is set to (112×112). The learning rate is 0.1 at the begin of the training and multiplied by 0.1 at 100K, 160K and 220K iterations. The training process ends after 240K iterations. The weight decay is set to 5e-4 and the momentum of SGD optimizer is 0.9. We train the teacher model by using the popular ArcFace method. For knowledge distillation method, we use a simple yet effective strategy, which uses the ArcFace loss along with the L2 loss between the embedding feature of the teacher and the student. It is worthy noting that our methods only change the data sampling strategy and can be easily combined with other distillation methods. The parameter h in the proposed Intrinsic Sampling is set to $12°$.

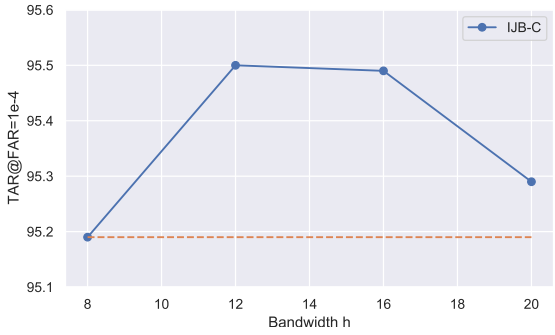

Figure 5. The results of student model on IJB-C with different h for Intrinsic Sampling method. The dashed line indicates the baseline of uniform sampling from dataset.

Person Re-identification We employ Market1501 [56] and DukeMTMC [37] to evaluate the effectiveness of the proposed methods. The top1 accuracy and mean averaged precision is used to evaluate the performance.

We employ the tricks proposed in [28] for our experiments. The teacher model is ResNet 50, and we choose ShuffleNet V2 [29] with scale 0.5x as the student model. Both the teacher and student models are pretrained on ImageNet. The learning rate is 3.5e-4 and multiplied by 0.1 at 70 epochs and 100 epochs. The training process ends after 120 epochs. The weight decay is set to 5e-4 and the momentum of SGD optimizer is 0.9. Following [28], we train the teacher model with a combination of triplet loss, center loss

Student Model	Use KD	Use ES	Use IS	Top1 Acc	MAP
ShuffleNet				86.5	69.2
ShuffleNet	✓			90.1	78.2
ShuffleNet	✓	✓		90.6	78.4
ShuffleNet	✓		✓	**90.9**	**78.5**

Table 3. The results of person re-identification on Market1501 dataset. "ShuffleNet" indicates ShuffleNet V2 with scale 0.5x.

Student Model	Use KD	Use ES	Use IS	Top1 Acc	MAP
ShuffleNet				78.0	61.4
ShuffleNet	✓			82.4	67.9
ShuffleNet	✓	✓		81.7	68.1
ShuffleNet	✓		✓	**82.5**	**68.3**

Table 4. The results of person re-identification on DukeMTMC dataset. "ShuffleNet" indicates ShuffleNet V2 with scale 0.5x.

and softmax loss. For the knowledge distillation method, we keep the softmax loss and introduce an additional L2 loss between the embedding features of the teacher and the student. The parameter h in the proposed Intrinsic Sampling is set to 12°.

5.2. Ablation Study

Effectiveness of Uniformity Improvement We report the results on face recognition in Table 2, with MobileFaceNet and R18 as the student model respectively. By using R50 as teacher, the performance of student MobileFaceNet for IJB-C at 1e-4 can be enhanced from 94.75% to 95.19%, with an improvement of 0.44%. By using the proposed Extrinsic Sampling method, this improvement goes up to 0.64%. The performance of using Intrinsic Sampling is the best compared to naive knowledge distillation and Extrinsic Sampling, improving the baseline by a margin of 0.74%. The results on IJB-B show a similar trend. As for the student R18, Extrinsic Sampling outperforms Intrinsic Sampling on both IJB-B and IJB-C.

The results on person re-identification are shown in Table 3 and Table 4, for Market1501 and DukeMTMC respectively. Notice that the Market1501 and DukeMTMC are relatively small datasets, and the model trained on them is easy to be overfitting. Thus knowledge distillation can help the student model improve generalization by mimic the behavior of the teacher, and improve the performance of the student by a large margin. By using our uniform improving methods, the student can fit the mapping of teacher better across the input space. As shown in Table 3, the Intrinsic Sampling method can boost the top-1 accuracy on Market1501 from 90.1 to 90.9, narrowing the gap between the student and the teacher significantly. Moreover, the results on DukeMTMC also demonstrate the effectiveness of the proposed methods through enhance the knowledge distillation performance by a remarkable margin.

Influence of Parameters As the proposed Extrinsic Sampling method is parameter-free, we examine the influence of the parameter h in the Intrinsic Sampling method. When h goes down, the estimation of density will only consider samples that are closer, and the density finally shrink to an identical value for every sample. Also, the estimated density will become greater as h increases and approach an maximum value. Thus, the resampled list will degenerate into uniform sampling when h is too small or too large.

With Eq 7, we find that $h \approx 12^\circ$ in Face Recognition and Person Re-Identification tasks. To examine the influence on performance, we choose different values for h and conduct ablation experiments on Face Recognition. The results are shown in Figure 5. We can see that the performance under $h = 12^\circ$ achieves the top. Also, the performance of Intrinsic Sampling is insensitive to parameter h, and can always surpass the naive distillation baseline.

method	type	IJB-B	IJB-C
KD		93.48	95.19
USI	sampling	93.30	95.18
Focal Loss	re-weighting	93.48	95.21
CB Loss	re-weighting	74.91	77.86
LDAM	re-weighting	93.42	95.24
ES	sampling	93.60	95.39
IS	sampling	**93.76**	**95.49**

Table 5. Results on Face Recognition for comparison with existing methods. "USI" indicates uniform sampling on identities, "CB Loss" indicates Class-Balanced Loss and "LDAM" indicates Label-Distribution-Aware Margin loss.

Compare with Existing Methods We compare our methods with two existing methods, uniform sampling on identities, Focal Loss [26], Class-Balanced Loss [7] and Label-Distribution-Aware Margin loss [4]. The results on Face Recognition are shown in Table 5, with the MobileFaceNet as the student. It can be observed that uniform sampling on identities do not help to improve the performance of distillation, although it ensures the balance between identities. This invalidity is related to over-emphasis on the minorities and the resulted overfitting. The rest of the existing methods are all based on loss re-weighting. Focal Loss and Label-Distribution-Aware Margin loss can boost the performance by a little, while the Class-Balanced Loss even leads to a significantly inferior result. This phenomenon coincides with the former discussion about the issues of re-weighting methods. On the contrary, the two methods proposed in this paper both improve the performance by a considerable margin. It is noteworthy that the Extrinsic Sampling method is parameter-free and shows good generalization.

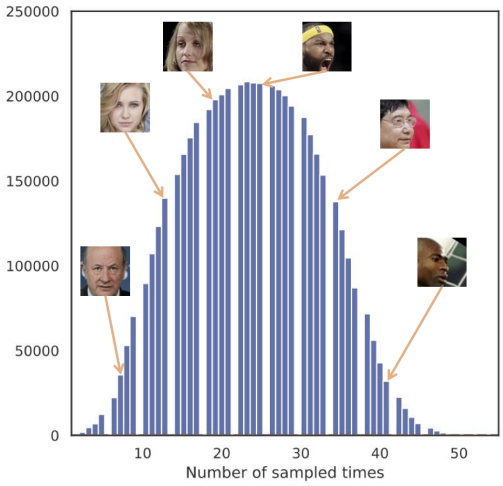

Figure 6. Some images in the resampled data list by the proposed Intrinsic Sampling method.

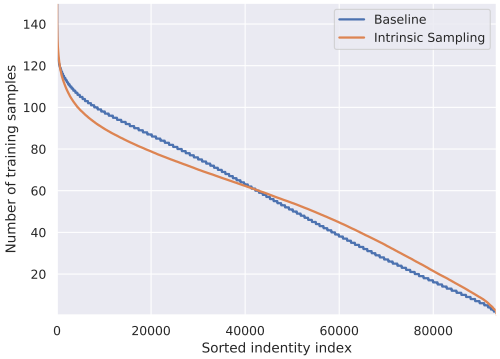

Figure 7. Demonstration for the number of images per identity with different sampling method. "Baseline" indicates uniform sampling from dataset.

5.3. Discussion

In this subsection, we show some of the resampled images by our methods and try to explain how the resampled data help to improve the performance of the student.

Visualization of resampled data Firstly, we show the most and least frequently sampled images in Figure 6. We can see that faces with top occurrence are often in large pose or blurry, yet the face with least frequency are often under front-view and good illumination condition. This demonstrates that our methods can mine hard examples and compress easy examples at the same time.

Also, we recount the number of images per identity in the generated data list and compare the histogram with original dataset in Figure 7. The issue of imbalanced class is alleviated significantly. Moreover, the numbers of identities with many images are not reduced drastically for that the information contained in these images are still rich and should be used to train the network.

model	Use KD	Use ES	Use IS	IJB-B 1e-4	IJB-C 1e-4
MobileFaceNet				93.02	94.75
MobileFaceNet		✓		93.29	94.99
MobileFaceNet			✓	**93.47**	**95.15**
MobileFaceNet	✓			93.48	95.19

Table 6. The demonstration of generalization to training from scratch for the proposed methods.

Generalization to Training from scratch As the explanation above has no clear relation with knowledge distillation, we wonder whether our methods can generalize to other optimization form, like training from scratch. Notice that the analysis about the uniformity is indeed dependent on knowledge distillation.

We show the result of training the student model from scratch with the proposed methods on face recognition as an example. Notice that the resampled list of data is still based on the output of the teacher. From Table 6, we can see that both our methods can improve the performance of the student model, and, surprisingly, the performance with Intrinsic Sampling is comparable with standard knowledge distillation. We assume that the data distribution itself contains latent knowledge that can be transferred to the student even by training from scratch.

6. Conclusion

In this paper, we first examine the problem of bias in the dataset and argue that this can harm the knowledge distillation process. Then, we introduce uniformity as a unified view toward the bias issue and provide a quantitative definition of uniformity. Uniformity check on realworld dataset is also presented to show the rationality of the definition. Further, we propose two uniformity-oriented sampling methods, Extrinsic Sampling and Intrinsic Sampling, to generate more uniform training data. It is noteworthy that both our methods are simple and easy to tune parameter. Extensive experiments are conducted on large-scale face recognition and person re-identification to verify the effectiveness of the proposed methods. Finally, we qualitatively analyze how the resampled data help to improve the performance and check the generalization ability to training from scratch.

Acknowledgments

This research is partially sponsored by the grant from the National Key Research and Development Program (2018YFC1406200) and Natural Science Foundation of China Grant 41930110.

References

[1] S. Ahn, Shell Xu Hu, A. Damianou, N. Lawrence, and Z. Dai. Variational information distillation for knowledge transfer. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9155–9163, 2019.

[2] Mateusz Buda, A. Maki, and M. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural networks : the official journal of the International Neural Network Society*, 106:249–259, 2018.

[3] Jonathon Byrd and Zachary Chase Lipton. What is the effect of importance weighting in deep learning? In *ICML*, 2019.

[4] Kaidi Cao, Colin Wei, Adrien Gaidon, N. Aréchiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*, 2019.

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[6] S. Chen, Y. Liu, Xiang Gao, and Zhen Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. *ArXiv*, abs/1804.07573, 2018.

[7] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge J. Belongie. Class-balanced loss based on effective number of samples. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9260–9269, 2019.

[8] Jiankang Deng, J. Guo, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694, 2019.

[9] Jiankang Deng, J. Guo, Debing Zhang, Yafeng Deng, Xiangju Lu, and S. Shi. Lightweight face recognition challenge. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 2638–2646, 2019.

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[11] Yan Gao, Boxiao Liu, Nan Guo, Xiaochun Ye, Fang Wan, Haihang You, and Dongrui Fan. C-midn: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9834–9843, 2019.

[12] Yan Gao, Boxiao Liu, Nan Guo, Xiaochun Ye, Fang Wan, Haihang You, and Dongrui Fan. Utilizing the instability in weakly supervised object detection. *arXiv preprint arXiv:1906.06023*, 2019.

[13] Jindong Gu and Volker Tresp. Search for better students to learn distilled knowledge. *ArXiv*, abs/2001.11612, 2020.

[14] Yandong Guo, Lei Zhang, Yuxiao Hu, X. He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *ECCV*, 2016.

[15] H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21:1263–1284, 2009.

[16] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[17] Byeongho Heo, Minsik Lee, Sangdoo Yun, and J. Choi. Knowledge distillation with adversarial samples supporting decision boundary. In *AAAI*, 2019.

[18] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[19] C. Huang, Y. Li, Chen Change Loy, and X. Tang. Learning deep representation for imbalanced classification. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5375–5384, 2016.

[20] C. Huang, Y. Li, Chen Change Loy, and X. Tang. Deep imbalanced learning for face recognition and attribute prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:2781–2794, 2020.

[21] Z. Huang and Naiyan Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *ArXiv*, abs/1707.01219, 2017.

[22] Zeyi Huang, Yang Zou, B. V. Kumar, and D. Huang. Comprehensive attention self-distillation for weakly-supervised object detection. *ArXiv*, abs/2010.12023, 2020.

[23] Xiao Jin, Baoyun Peng, Y. Wu, Yu Liu, Jiaheng Liu, Ding Liang, J. Yan, and Xiaolin Hu. Knowledge distillation via route constrained optimization. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1345–1354, 2019.

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

[25] Xu Lan, Xiatian Zhu, and S. Gong. Knowledge distillation by on-the-fly native ensemble. In *NeurIPS*, 2018.

[26] Tsung-Yi Lin, Priyal Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42:318–327, 2020.

[27] Junjie Liu, Dongchao Wen, Hongxing Gao, Wei Tao, Tse-Wei Chen, Kinya Osa, and M. Kato. Knowledge representing: Efficient, sparse representation of prior knowledge for knowledge distillation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 638–646, 2019.

[28] Hao Luo, Youzhi Gu, Xingyu Liao, S. Lai, and W. Jiang. Bag of tricks and a strong baseline for deep person re-identification. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1487–1495, 2019.

[29] Ningning Ma, X. Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *ECCV*, 2018.

[30] L. V. D. Maaten and Geoffrey E. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[31] Brianna Maze, J. Adams, J. A. Duncan, Nathan D. Kalka, T. Miller, Charles Otto, Anil K. Jain, W. T. Niggel, J. Anderson, J. Cheney, and P. Grother. Iarpa janus benchmark - c: Face dataset and protocol. *2018 International Conference on Biometrics (ICB)*, pages 158–165, 2018.

[32] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *AAAI*, 2020.

[33] Aythami Morales, Julian Fierrez, Ruben Vera-Rodriguez, and Ruben Tolosana. Sensitivenets: Learning agnostic representations with application to face images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6):2158–2164, 2020.

[34] E. Parzen. On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962.

[35] N. Passalis and A. Tefas. Learning deep representations with probabilistic knowledge transfer. In *ECCV*, 2018.

[36] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015.

[37] Ergys Ristani, Francesco Solera, Roger S. Zou, R. Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. *ArXiv*, abs/1609.01775, 2016.

[38] M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27:832–837, 1956.

[39] Li Shen, Zhouchen Lin, and Q. Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *ECCV*, 2016.

[40] Abhinav Shrivastava, A. Gupta, and Ross B. Girshick. Training region-based object detectors with online hard example mining. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 761–769, 2016.

[41] Guanglu Song, Yu Liu, and Xiaogang Wang. Revisiting the sibling head in object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11563–11572, 2020.

[42] Fei Wang, L. Chen, Cheng Li, Shiyao Huang, Yanjie Chen, Chen Qian, and Chen Change Loy. The devil of face recognition is in the noise. In *ECCV*, 2018.

[43] Mei Wang and Weihong Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9322–9331, 2020.

[44] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 692–702, 2019.

[45] Tao Wang, L. Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4928–4937, 2019.

[46] W. Wang, Wei Hong, Feng Wang, and Jinke Yu. Gan-knowledge distillation for one-stage object detection. *IEEE Access*, 8:60719–60727, 2020.

[47] Xiaobo Wang, Tianyu Fu, Shengcai Liao, S. Wang, Zhen Lei, and Tao Mei. Exclusivity-consistency regularized knowledge distillation for face recognition. In *ECCV*, 2020.

[48] Yu-Xiong Wang, D. Ramanan, and M. Hebert. Learning to model the tail. In *NIPS*, 2017.

[49] Cameron Whitelam, Emma Taborsky, Austin Blanton, Brianna Maze, J. Adams, T. Miller, Nathan D. Kalka, Anil K. Jain, J. A. Duncan, Kristen Allen, J. Cheney, and P. Grother. Iarpa janus benchmark-b face dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 592–600, 2017.

[50] Wikipedia contributors. Normal distribution — Wikipedia, the free encyclopedia, 2021. [Online; accessed 31-March-2021].

[51] Qizhe Xie, E. Hovy, Minh-Thang Luong, and Quoc V. Le. Self-training with noisy student improves imagenet classification. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2020.

[52] Lu Yu, V. O. Yazici, X. Liu, Joost van de Weijer, Y. Cheng, and Arnau Ramisa. Learning metrics from teachers: Compact networks for image embedding. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2902–2911, 2019.

[53] L. Yuan, F. Tay, G. Li, T. Wang, and Jiashi Feng. Revisit knowledge distillation: a teacher-free framework. *ArXiv*, abs/1909.11723, 2019.

[54] Manyuan Zhang, Guanglu Song, Hang Zhou, and Y. Liu. Discriminability distillation in group representation learning. In *ECCV*, 2020.

[55] Y. Zhang, T. Xiang, Timothy M. Hospedales, and H. Lu. Deep mutual learning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4320–4328, 2018.

[56] L. Zheng, L. Shen, L. Tian, S. Wang, Jingdong Wang, and Q. Tian. Scalable person re-identification: A benchmark. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1116–1124, 2015.

[57] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10492–10502, 2021.

[58] Yang Zou, Zhiding Yu, B. V. Kumar, and J. Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *ECCV*, 2018.