Rethinking Feature-based Knowledge Distillation for Face Recognition – Supplementary Material

Jingzhi Li[†,*,1], Zidong Guo[*,1], Hui Li[1], Seungju Han[2], Ji-won Baek[2], Min Yang[1], Ran Yang[1], Sungjoo Suh[2]
[1]Samsung R&D Institute China Xi'an (SRCX)
[2]Samsung Advanced Institute of Technology (SAIT), South Korea
jingzhi.li, zidong.guo, hui01.li, sj75.han, jw0328.baek, min16.yang, ran01.yang, sungjoo.suh@samsung.com

1. Introduction

In this supplementary material, we first present the details of the estimation of the intrinsic dimension. Then to affirm the universality of tailored teachers, we include additional experiments on *ReFO+*. Finally, we present the effect of using different β in reverse distillation and feature-only distillation to show the insensitivity on changes in distillation weight.

2. Estimating Intrinsic Dimension with TwoNN

To estimate the intrinsic dimension of the network, we follow the TwoNN method [2] as applied in [1].

Theory. TwoNN is a global intrinsic dimension estimator based on the distances of the first two nearest neighbors of each point in the space. Let r_1 and r_2 denote the distances to the nearest and the second nearest neighbors. The volume of the hyperspherical shell enclosed by the two neighbors is related to the intrinsic dimension by,

$$\Delta v = w_d(r_2^d - r_1^d), \tag{1}$$

where d is the intrinsic dimension and w_d is the volume of a unit d-sphere. It is proven in [2] that for uniformly sampled points, the ratio $\mu = \frac{r_2}{r_1}$ follows Pareto distribution with parameter $d+1$ as,

$$f(\mu|d) = d\mu^{-(d+1)}. \tag{2}$$

d can then by simply computed by maximizing the likelihood,

$$P(\boldsymbol{\mu}|d) = d^N \prod_{i=1}^{N} \mu_i^{-(d+1)}, \tag{3}$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_N)$ is the vector of sampled ratios.

Implementation. The actual implementation is based on the code released by [1]. To compute the intrinsic dimension for common face recognition models, we first train the model on MS1MV2 using standard ArcFace loss and obtain the corresponding embeddings. Then we randomly sample 1% of the embeddings from the dataset, containing 58k data points. The embeddings are normalized before the dimension estimation. To avoid disturbances from the unevenness of the dataset distribution, we recommend using fixed random seed at this stage to ensure the same images are used for each model or repeating the whole estimation for a few times for an averaged number. From the 1% of the embeddings, we randomly sample 90% to calculate intrinsic dimension. This is repeated for 20 times on the same subset of selected embeddings. The estimated intrinsic dimension is the mean value of the 20 experiments. The typical value for standard deviation is around $2e^{-2}$.

3. Universality of *ReFO+*

In the main text, we have shown the universality of *ReFO* using two IR100 teachers tailored to IR18 and MFN respectively. In Tab. 1, we present the same set of experiments using *ReFO+*. It is clear that both tailored teachers bring performance improvements on students of different structure. Comparing to *ReFO*, all students from *ReFO+* show higher

Table 1. Teachers tailored to MFN and IR18 by *ReFO+* show universal improvements on students' accuracies (%) with FO distillation, evaluated on MR-all. A←B refers to A tailored by B.

Teacher	Student			
	MFN	MNv2	IR18	IR34
IR100	53.86	58.32	61.70	73.16
IR100 $\xleftarrow{ReFO+}$ MFN	58.66	64.20	67.26	76.40
IR100 $\xleftarrow{ReFO+}$ IR18	59.64	64.53	68.56	77.38

*Equal contribution. † Corresponding author.

accuracies with a mean absolute performance improvement of 1.4% on top of students trained by *ReFO*.

4. Effects of β_1 in Reverse Distillation

In reverse distillation (Sec.3.3, Algorithm 1, Step 3), the teacher is trained by the loss

$$L = L_{cls} + \beta_1 L_{emb}(\boldsymbol{f}_t, \boldsymbol{f}_{s'}). \tag{4}$$

There is a balancing weight β_1 controlling the emphasis on student feature space awareness. We generally keep it small to allow the teacher to focus on the optimization of the main task. The default value of β_1 is 0.5 for normalized embeddings and 0.001 for un-normalized embeddings. In Tab. 2, we show the results of using other values of β_1 on the IR100-IR18 pair with normalized embeddings, evaluated on the three tracks of ICCV21-MFR. On the largest MR-all track, all values of β_1 are rather comparable with $\beta_1 = 1.0$ perform slightly better. The smaller Mask track and Children track show more variations in performance, and we opt for $\beta_1 = 0.5$ for its better overall results.

Table 2. Accuracies of different β_1 during reverse distillation for IR18 on ICCV21-MFR (%). Teacher: IR100. The best results are in **bold**.

Track	β_1				
	0.1	0.2	0.5	1.0	2.0
MR-all	66.09	66.01	66.13	**66.46**	66.15
Mask	46.10	45.23	**47.09**	47.07	46.98
Children	39.42	40.13	**40.46**	39.05	37.66

5. Effects of β_2 in Feature-Only Distillation

In the second stage of feature-only distillation (Sec.3.3, Algorithm 1, Step 5), the final student is trained by the loss

$$L = \beta_2 L_{emb}(\boldsymbol{f}_s, \boldsymbol{f}_t). \tag{5}$$

There is only one parameter β_2. In principle changing its value is equivalent to simultaneously adjusting the learning rate and the weight decay parameter. To keep the optimization parameters consistent across experiments, we allow β_2 to change instead. In Tab. 3, we show the results of *ReFO+* with different choices of β_2 on two teacher-student pairs, IR50-MFN and IR100-IR18. While both teacher-student pairs show the best performance at $\beta_2 = 5$, the IR100-IR18 pair appears more insensitive to different values of β_2.

Table 3. Accuracies of different β_2 during FO distillation on MR-all (%). The best results are in **bold**.

Model	β_2				
	1	2	5	10	20
MFN	58.11	58.24	**59.17**	58.80	58.21
IR18	68.37	68.22	**68.56**	68.46	68.52

References

[1] Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[2] Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific reports*, 7(1):1–8, 2017. 1