Multi-Level Feature Distillation of Joint Teachers Trained on Distinct Image Datasets – Supplementary

Adrian Iordache[1], Bogdan Alexe[1,2], Radu Tudor Ionescu[1]
[1]Department of Computer Science, University of Bucharest, Bucharest, Romania
[2]"Gheorghe Mihoc – Caius Iacob" Institute of Mathematical Statistics and Applied Mathematics of the Romanian Academy, Bucharest, Romania.

1. Overview

In order to offer a better understanding of the generality of our method, we include additional experiments. First of all, we aim to show that our method applies to a distinct collection of datasets. Second of all, we aim to quantify the impact of the number of datasets m on the joint model, as well as on the distilled students. To this end, we follow the setup described in the main paper, but with the following differences: (i) we increase the number of datasets from three to four; (ii) we change the entire collection of datasets; (iii) we train the joint teacher by varying the number of datasets, from two to four.

In another set of experiments, we compare with a single-dataset KD. This experiment aims to assess the benefits of distilling from multiple datasets.

We also aim to determine the applicability of our framework to distinct tasks. To this end, we conduct experiments on three action recognition datasets. In this setup, we start from pre-trained action recognition models, which allows us to demonstrate the effectiveness of our approach in conjunction with pre-training.

We further present ablation results where the joint teacher is either based on different backbones trained on the same dataset, or the same backbone trained on different datasets. Another dataset-related ablation is performed to demonstrate the generalization capacity of the MLFD students.

Next, we analyze the feature space generated by our approach. By visualizing the embeddings via t-SNE, we are able to determine that our approach leads to robust and disentangled representations. Finally, we discuss the time and space limitations of our framework.

2. Additional Image Classification Results

2.1. Datasets

Caltech-101. The Caltech-101 dataset [1] consists of $7,315$ training images and $1,829$ test images. It was originally proposed to test the ability of models to learn from few examples. The images belong to 101 object categories.

Flowers-102. The Flowers-102 dataset [8] contains $7,169$ training images and $1,020$ test images. The dataset contains 102 classes of flowers that typically grow in the United Kingdom.

CUB-200-2011. The Caltech-UCSD Birds 200 2011 (CUB-200) dataset [13] is formed of $9,430$ training images and $2,358$ test images. The images represent 200 distinct species of birds.

Oxford Pets. The Oxford-IIIT Pets dataset [9] consists of $5,906$ training images and $1,477$ test images. The dataset contains images for 37 breeds of cats and dogs.

2.2. Models

Since the datasets are distinct, we evaluate our method on a new set of individual teachers, denoted as $\mathcal{T}_3$, containing four models, namely a ResNet-18 [2] trained on Caltech-101, an EfficientNet-B0 [11] trained on Flowers-102, a SEResNeXt-26D [4] trained on CUB-200, and a ResNet-18 trained on Oxford Pets. $\mathcal{T}_3$ contains similar models to $\mathcal{T}_1$, but trained (from scratch) on distinct datasets.

2.3. Results

Table 1 shows the results obtained when the joint teacher is trained on two, three and four datasets, respectively. The joint teacher is first optimized on Caltech-101 and Flowers-102. The next version adds the CUB-200-2011 dataset into the mix, and the last version is trained on all four datasets. Notably, the results indicate that two datasets are enough to reach substantial performance gains. We report additional gains when using more datasets, although the relative improvements tend to saturate with the number of datasets, as shown in Figure 1. Regardless of the number of datasets, our multi-level distillation framework brings significant performance improvements on all four datasets. Even if the chosen datasets are typically small, the reported gains are still high, suggesting that our framework plays a very important role in increasing the performance and generaliza-

Model	#Datasets	Caltech-101		Flowers-102		CUB-200		Oxford Pets	
		acc@1	acc@5	acc@1	acc@5	acc@1	acc@5	acc@1	acc@5
Dataset-specific models	1	74.79	86.39	78.04	93.14	51.40	77.06	63.91	89.10
Students (our)	2	79.66	92.29	80.78	94.22	-	-	-	-
	3	**80.32**	91.31	79.41	95.10	58.27	82.65	-	-
	4	80.04	**92.45**	**81.76**	**95.29**	**59.29**	**83.29**	**68.79**	**92.42**
Joint teacher (ours)	2	82.56	93.00	80.39	93.73	-	-	-	-
	3	**83.11**	93.06	81.76	**94.61**	57.21	**81.30**	-	-
	4	83.00	**93.71**	**81.86**	94.51	**57.29**	80.66	**66.08**	**89.91**

Table 1. Results for the set $\mathcal{T}_3$, including the dataset-specific baselines, the students obtained by employing multi-level distillation using embeddings extracted at two levels ($\mathbf{L}_2$), and the corresponding joint teachers. The number of datasets used to train the joint teachers is gradually increased from two to four. The results of the best student and the best teacher on each dataset are highlighted in bold.

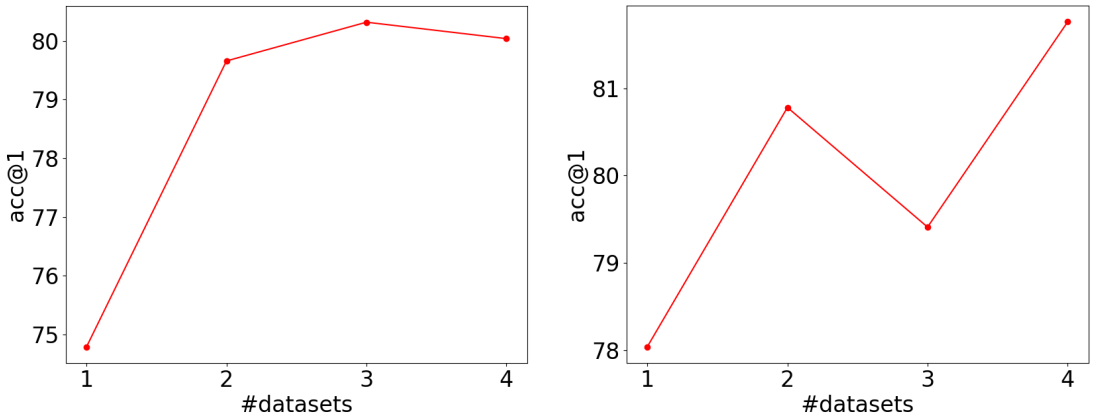

Figure 1. Accuracy rates of the student models on Caltech-101 (left) and Flowers-102 (right) when the number of datasets is increased from one to four. Best viewed in color.

tion capacity of neural models.

3. Comparison with Single-Dataset Distillation

In Table 2, we present additional results with students based on standard (single-dataset) distillation. For the single-dataset distillation, we consider two distinct sets of teacher. On the one hand, we distill from the teachers included in $\mathcal{T}_1$, for a direct comparison with our multi-dataset approach. On the other hand, we use another set of more powerful teachers, namely $\mathcal{T}_4$, which is composed of the following models: ResNet-50 for CIFAR-100, EfficientNet-B1 for Tiny ImageNet, and SEResNeXt-50D for ImageNet-Sketch. The latter set of teachers is considered because it is common to use deeper teachers in teacher-student training setups. Nevertheless, the students trained with our multi-dataset distillation approach reach much better results than both single-dataset student versions. Our empirical results suggest that it is more effective to distill from lighter teachers trained on multiple datasets than distilling from a deeper teacher trained on a single dataset.

4. Additional Action Recognition Results

4.1. Models

The action recognition models are taken from the MMAction2[1] toolbox, which provides various models pre-trained on different datasets. For ActivityNet, the individual teacher is a pre-trained Temporal Segment Network [14] architecture, which is based on a ResNet-50 backbone with 8 segments. For HMDB-51 and UCF-101, the teachers are based on a pre-trained ResNet-50 with Temporal Shift Module [7]. All selected teachers are first pre-trained on the Kinetics-400 dataset [5]. Then, each teacher is fine-tuned on its own target dataset. From this point on, we employ our multi-dataset distillation approach.

4.2. Datasets

ActivityNet. The ActivityNet dataset [3] comprises 19,994 videos labeled with 200 activity classes. Following standard evaluation practices, we report results on the official validation set, since there are no publicly-available labels for the test set.

[1] https://github.com/open-mmlab/mmaction2

Model	CIFAR-100		Tiny ImageNet		ImageNet-Sketch	
	acc@1	acc@5	acc@1	acc@5	acc@1	acc@5
Dataset-specific models (light=$\mathcal{T}_1$)	55.07	81.62	45.09	70.52	49.67	71.39
Dataset-specific models (deep=$\mathcal{T}_4$)	64.38	88.30	43.47	69.48	57.75	76.38
Single-dataset KD (from $\mathcal{T}_1$ to $\mathcal{T}_1$)	56.44	82.63	48.15	74.02	46.87	70.40
Single-dataset KD (from $\mathcal{T}_4$ to $\mathcal{T}_1$)	59.17	83.68	47.16	73.06	54.21	75.45
Multi-dataset KD (from $\mathcal{T}_1$ to $\mathcal{T}_1$)	**62.25**	**84.64**	**51.61**	**76.46**	**61.31**	**78.36**

Table 2. Single-dataset KD (using individual dataset-specific models as teachers) versus our multi-dataset KD. Students are always light (and identical for both single-dataset and multi-dataset distillation). $\mathcal{T}_1$: ResNet-18, EfficientNet-B0, SEResNeXt-26D. $\mathcal{T}_4$: ResNet-50, EfficientNet-B1 and SEResNeXt-50D.

Model	ActivityNet		HMDB-51		UCF-101	
	acc@1	mAP	acc@1	mAP	acc@1	mAP
Dataset-specific models	73.81	42.83	73.60	61.27	94.63	86.32
Students ($\mathbf{L}_1$)	**81.56**	**82.89**	**75.32**	**76.81**	**95.97**	**97.96**
Joint teacher ($\mathbf{L}_1$)	88.50	90.33	78.67	79.13	98.05	99.20

Table 3. Action recognition results on ActivityNet, HMDB-51 and UCF-101 for a new set of models (ResNet-50 with Temporal Segment Network for ActivityNet; ResNet-50 with Temporal Shift Module for HMDB-51 and UCF-101), including the dataset-specific baselines, our students based on $\mathbf{L}_1$ embeddings, and the corresponding joint teacher. The dataset-specific models and individual teachers are pre-trained on Kinetics-400. Our students outperform the dataset-specific models by large mAP gaps on all datasets.

Model	Distillation level	TinyImageNet		ImageNet-Sketch	
		acc@1	acc@5	acc@1	acc@5
Dataset-specific models	-	45.09	70.52	49.67	71.39
Students (same dataset)	$\mathbf{L}_2$	50.07	75.45	54.16	75.55
Students (different datasets)	$\mathbf{L}_2$	**51.61**	**76.46**	**61.31**	**78.36**

Table 4. Comparison between students distilled from joint teachers of identical capacity, but using same or distinct datasets. The same-dataset students are distilled from a joint teacher that combines different backbones which are all trained on the target dataset. The students trained on different datasets are based on our unmodified MLFD framework.

HMDB-51. The HMDB-51 dataset [6] consists of $7,000$ clips distributed in 51 action classes. The official evaluation procedure uses three different data splits. We consider the first split in our experiments.

UCF-101 The UCF-101 dataset [10] contains $13,320$ YouTube videos from 101 action classes. As for HMDB-51, there are three data splits and we select the first one for our evaluation.

4.3. Results

We present action recognition results with $\mathbf{L}_1$ students in Table 3. Although we start from pre-trained individual teachers, the joint teacher leads to significant performance gains. Distilling knowledge from the joint teacher into the student models is also beneficial. In the end, we obtain student models that are identical in terms of architecture to the dataset-specific models, but the action recognition performance of our students is significantly higher, especially in terms of mAP.

5. Additional Ablations

5.1. Distillation from Same-Dataset Joint Teacher

To demonstrate the utility of training the joint teacher on a diversity of datasets, we perform an ablation study where the joint teacher is based on the same variety of architectures, but all teachers are trained on the same dataset. In Table 4, we compare the students based on $\mathbf{L}_2$ distillation for $\mathcal{T}_1$ teachers on TinyImageNet and ImageNet-Sketch. Although both kinds of students surpass the dataset-specific models, our multi-dataset students clearly benefit from the more diverse datasets used to train the joint teacher. The results are consistent on both TinyImageNet and ImageNet-Sketch.

5.2. Distillation from Same-Architecture Joint Teacher

To show that our multi-dataset distillation works even if the joint teacher uses the same architecture across different datasets, we perform an experiment where the joint

Model	Distillation level	Caltech-101		Oxford Pets	
		acc@1	acc@5	acc@1	acc@5
Dataset-specific models	-	74.79	86.39	63.91	89.10
Students (same architecture)	$\mathbf{L}_2$	**80.75**	**92.56**	**68.99**	**93.09**
Joint teacher (same architecture)	$\mathbf{L}_2$	82.61	93.27	64.92	89.10

Table 5. Results with a joint teacher based on the same architecture (ResNet-18) trained on different datasets (Caltech-101 and Oxford Pets). The corresponding students are also based on ResNet-18.

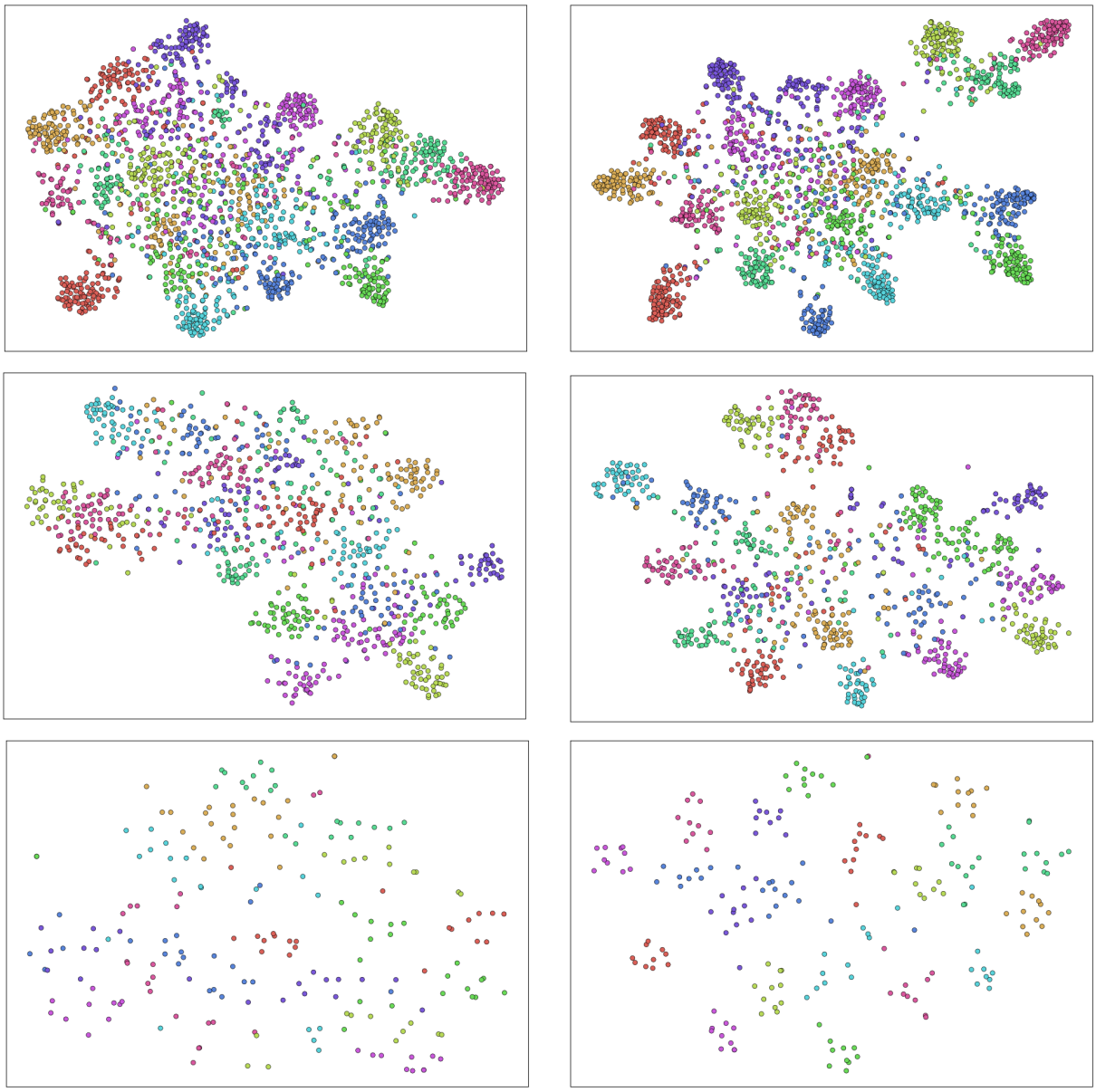

Figure 2. Visualizations based on t-SNE projections of image embeddings learned by the dataset-specific models (left) and those learned by our student models (right) for the three datasets: CIFAR-100 (first row), TinyImageNet (second row), ImageNet-Sketch (third row). Best viewed in color.

teacher comprises a ResNet-18 trained on Caltech-101, and a ResNet-18 trained on Oxford Pets. The corresponding students are also based on ResNet-18. We report the results of the dataset-specific models, the joint teacher and

Model	Oxford Pets	
	acc@1	acc@5
Dataset-specific model	63.91	89.10
Student **w/o** Oxford Pets	**66.14**	**91.94**

Table 6. Results on Oxford Pets with the dataset-specific model versus an $\mathbf{L}_2$ student distilled from a joint teacher which is trained on three datasets: Caltech-101, Flowers-102, and CUB-200.

the $\mathbf{L}_2$ students in Table 5. Both teacher and student models outperform the dataset-specific models, confirming that our multi-dataset distillation performs well, even when the same architecture is employed across all datasets.

5.3. Cross-Dataset Generalization

To showcase the generalization capacity of our framework, we present cross-dataset results by training a joint teacher on Caltech-101, Flowers-102, CUB-200 and distilling the knowledge into an $\mathbf{L}_2$ student for Oxford Pets. In Table 6, we compare this student with the dataset-specific model on Oxford Pets. Our cross-dataset student surpasses the dataset-specific model, thus showing a higher generalization capacity.

6. Feature Analysis

To gain a deeper understanding of our method, we compare the discriminative power of the embeddings learned by the dataset-specific models and those learned by our student models. We achieve this using the t-SNE [12] visualization tool and plot the 2D projections of test data points (images) from different classes, as obtained after applying the corresponding embedding (model). As there are at least 100 classes in each of the three considered datasets in the main paper, we plot the projections of test data points for only a fraction of the total number of classes, to improve clarity. Figure 2 shows the 2D projections obtained using the t-SNE tool, for each of the three datasets. The visualization reveals that the embeddings learned by our student models cluster data points from the same class much better than the dataset-specific models, thus demonstrating a higher discriminative power.

6.1. Limitations

Time complexity. A possible limitation of our method is the wall-clock training time. The number of models that need to be trained is proportional with the number of datasets m. The m individual teachers specific to each dataset can be trained in parallel, whereas the joint teacher and the student models need to be trained sequentially. In general practice, the training time can also be reduced by using pre-trained networks as individual teachers. In our experiments, the joint teachers obtain stable performance after roughly 1/5 of the training time of the individual teachers, since they can harness the information learned by individual teachers. Based on these insights, the total training time required to obtain a student ranges between 1.2 (if all individual teachers are pre-trained) and $m + 1.2$ (if all individual teachers need to be trained), where m is the number of data sets. Notice that, during inference, the wall clock-time remains the same, *i.e*. there is no difference between the dataset-specific models and our students.

Space complexity. Regarding space complexity, the only limitation is the storage required for caching latent representations of the individual teachers, especially when m is larger. In practice, we can limit the number of individual teacher models and datasets to a manageable size, *e.g*. 2-4, to avoid using too much storage space. Our results show that $m = 3$ is enough to bring significant performance gains, up to 12%.

References

[1] Li Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):594–611, 2006. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of CVPR*, pages 770–778, 2016. 1

[3] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In *Proceedings of CVPR*, pages 961–970, 2015. 2

[4] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-Excitation Networks. In *Proceedings of CVPR*, pages 7132–7141, 2018. 1

[5] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The Kinetics Human Action Video Dataset. *arXiv preprint arXiv:1705.06950*, 2017. 2

[6] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proceedings of ICCV*, 2011. 3

[7] Ji Lin, Chuang Gan, and Song Han. TSM: Temporal Shift Module for Efficient Video Understanding. In *Proceedings of ICCV*, pages 7082–7092, 2019. 2

[8] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Proceedings of ICVGIP*, 2008. 1

[9] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *Proceedings of CVPR*, pages 3498–3505, 2012. 1

[10] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 3

[11] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of ICML*, pages 6105–6114, 2019. 1

[12] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. 5

[13] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 1

[14] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal Segment Networks: Towards good practices for deep action recognition. In *Proceedings of ECCV*, pages 20–36, 2016. 2