Divide and Conquer the Embedding Space for Metric Learning
- Supplementary Material -

Artsiom Sanakoyeu* Vadim Tschernezki* Uta Büchler Björn Ommer
Heidelberg Collaboratory for Image Processing, IWR, Heidelberg University
firstname.lastname@iwr.uni-heidelberg.de

1. Implementation details

Re-clustering every T epochs: As pointed out in Sec. 3.3 of the main submission, we update the data partitioning by re-clustering every T epochs using the full embedding space, composed by concatenating the embeddings produced by the individual learners. To maintain consistency, each learner is associated to the cluster, which is most similar to the cluster assigned to this learner in the previous iteration (i.e. in epoch $t-T$). This amounts to solving a linear assignment problem where similarity between clusters is measured in terms of IoU of points belonging to the clusters.

The source code is available at https://bit.ly/dcesml.

2. Additional ablation study

As discussed in the main paper, our approach facilitates the learning of decorrelated representations of individual learners. To show this, we conduct an additional ablation study where we evaluate the performance of individual learners and compute the correlation between their embeddings. In the same way as in the main paper, we use the Stanford Online Products dataset [3] and train our model with Margin loss [7], $K = 8$ and embedding size $d = 128$.

We computed Recall@1 on the entire test set for every individual learner, each of which operates in a 16-dimensional embedding subspace. However, the baseline model was trained with only *one* learner operating in the embedding space with 128 dimensions. Hence, for comparison with the learners of our model, we split the embedding of the baseline model on 8 non-overlapping slices of 16 dimensions each and evaluate them separately. In Tab. S1 we can see that each individual learner trained using our approach is weaker in average than slices of the baseline model embedding. However, when we concatenate the embeddings of all individual learners together they yield significantly higher Recall@1 than the baseline model (3.2% higher in absolute values). In Fig. S2 we also show how the performance changes when we use together only $1, 2, \ldots 7$ or all 8 learners for evaluation: one learner corresponds to 16 out of 128 dimensions, two learners to 32 out of 128 dimensions and so on; 8 learners correspond to all 128 dimensions. We observe a larger gain compared to the baseline when more learners are used together for evaluation. This shows that the learners trained by our approach learn complementary features.

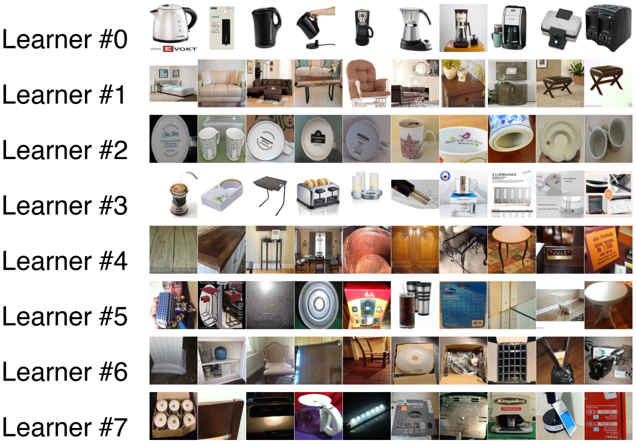

Figure S1: Representative images for the learners and their corresponding subspaces. The model was trained on the Stanford Online Products dataset with $K = 8$. Best viewed zoomed in.

Moreover, in Tab. S1 we directly computed the correlation coefficient between the embedding produced by different learners. The correlation coefficient between the learners in our model is lower than between the slices of the baseline model embedding. This evidence supports our claim that the learners proposed by our approach learn less correlated features and, hence, utilize the embedding space in a more efficient way.

To visualize what is captured in each embedding subspace, in Fig. S1 we show representative images for different learners. Every row shows 10 query images, which are the easiest in terms of recall for one learner (R@1 = 1) but extremely difficult (R@30 = 0) for any other learner. We can see that every subspace has its own abstract ”specialization”.

	Baseline	Ours	Emb. dimensions
Learner 1	37.0	29.6	1..16
Learner 2	37.0	29.7	17..32
Learner 3	36.5	29.5	33..48
Learner 4	36.5	29.4	49..64
Learner 5	36.3	29.1	65..80
Learner 6	37.4	29.7	81..96
Learner 7	36.7	29.4	97..112
Learner 8	37.1	29.9	113..128
Mean	36.8	29.5	-
All together (↑)	72.7	**75.9**	1..128
Corr. coeff. (↓)	0.0602	**0.0498**	-

Table S1: Evaluation of the individual learners. Recall@1 for every individual learner on the entire test set of Stanford Online Products [3]. The last column shows the indices of the corresponding dimensions of the embedding space assigned to the learners. The individual learners of our model yield significantly higher Recall@1 than the baseline model when they are concatenated and evaluated all together, since they learn less correlated representations.

The 1st focuses on the electrical appliances, the 2nd – on furniture, the 3rd – on plates and mugs, etc.

Recall@k	1	5	10	mAP
HAP2S_P [8]	84.5	-	-	69.7
PSE ** [4]	87.7	94.5	96.8	69.0
HA-CNN [2]	91.2	-	-	75.7
DGS [5]	92.7	96.9	98.1	82.5
DNN+CRF [1]	93.5	97.7	-	81.6
MGN ** [6]	95.7	-	-	86.9
Margin baseline* [7]	98.2	99.3	99.3	87.9
Ours (Margin)	**98.9**	**99.5**	**99.7**	**88.8**

Table S2: Recall@k for $k = 1, 5, 10$ and mean average precision (mAP) on Market-1501 [9] with single-query mode. * denotes our own implementation based on ResNet-50 with $d = 128$. ** denotes methods that use ResNet-50 as backbone.

3. Additional quantitative evaluation on person re-identification

In this section, we additionally evaluate our approach and compare to the state-of-the-art methods on Market-1501 [9] dataset for person re-identification.

Market-1501 [9] contains $32,668$ images of $1,501$ identities captured by six cameras in front of a supermarket. The $1,501$ identities are divided into a training set consisting of $12,936$ images of 751 identities and a testing set containing the other $19,732$ images of 750 identities. The query set contains $3,368$ images with each identity having at most 6 queries. For evaluation, we follow the standard protocol of [9] and report the mean average precision (mAP) and Recall@1, Recall@5 and Recall@10. In Tab. S2 we demonstrate the comparison of our approach to other methods, where we can see the superior performance of the proposed approach.

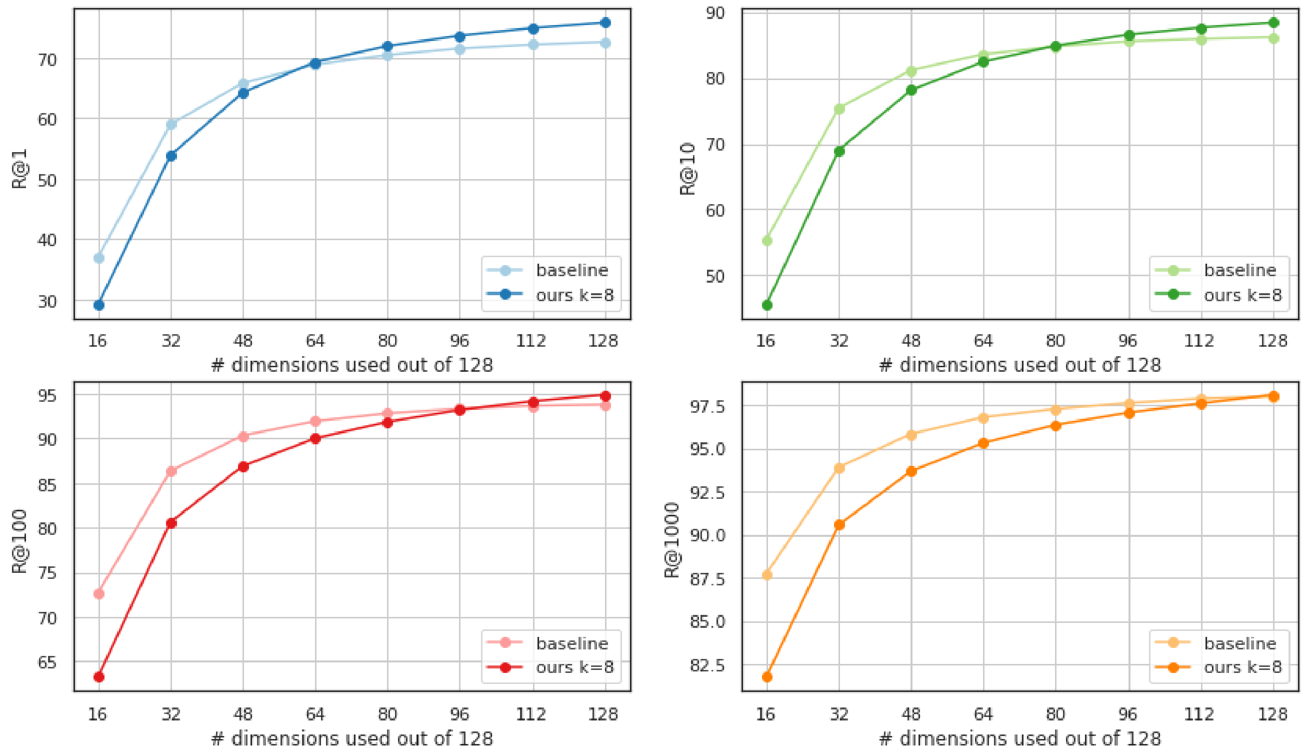

Figure S2: Evaluation of the individual learners. We trained our model with $K = 8$ learners and embedding size $d = 128$ on the Stanford Online Products dataset [3]. The plots show the the Recall@k score when we use only the first m out of 128 dimensions of the embedding layer ($m = \{16, 32, \ldots, 128\}$) for evaluation. Adding another 16 dimensions corresponds to using one more learner $\mathbf{f}_{m/16}$ during the evaluation of our model. In case of the baseline model we do not have any learners, but for a fair comparison we also use only the first m dimensions of the embedding layer. We see a higher performance of our approach compared to the baseline when more dimensions are used together, which shows that the individual learners in our model produce less correlated embeddings.

References

[1] D. Chen, D. Xu, H. Li, N. Sebe, and X. Wang. Group consistent similarity learning via deep crf for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[2] W. Li, X. Zhu, and S. Gong. Harmonious attention network for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[3] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4004–4012, 2016. 1, 2, 3

[4] M. Saquib Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[5] Y. Shen, H. Li, T. Xiao, S. Yi, D. Chen, and X. Wang. Deep group-shuffling random walk for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[6] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou. Learning discriminative features with multiple granularities for person re-identification. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM ’18, pages 274–282, New York, NY, USA, 2018. ACM. 2

[7] C.-Y. Wu, R. Manmatha, A. J. Smola, and P. Krähenbühl. Sampling matters in deep embedding learning. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017. 1, 2

[8] R. Yu, Z. Dou, S. Bai, Z. Zhang, Y. Xu, and X. Bai. Hard-aware point-to-set deep metric for person re-identification. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2

[9] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian. Scalable person re-identification: A benchmark. In *Computer Vision, IEEE International Conference on*, 2015. 2