Towards Universal Image Embeddings: A Large-Scale Dataset and Challenge for Generic Image Representations Supplementary material

Nikolaos-Antonios Ypsilantis[1] Kaifeng Chen[2] Bingyi Cao[2] Mário Lipovský[2]
Pelin Dogan-Schönberger[2] Grzegorz Makosa[2] Boris Bluntschli[2]
Mojtaba Seyedhosseini[2] Ondřej Chum[1] André Araujo[2]

[1]VRG, FEE, Czech Technical University in Prague [2]Google

In the Supplementary material, we include an Appendix with results for mAP metric, an ablation on architecture size, results for a setting where evaluation is performed per domain separately, results for specialists when used as universal embeddings and a comparison of PCA-Whitening (PCAw) with random linear projection. All experiments in the supplementary material are reported for 1 seed (except stated otherwise). Additionally, we provide a collage of image samples coming from the domains the proposed UnED dataset covers.

A. Appendix

A.1. mAP results

As discussed in Section 3.2, we additionally present results using the mean Average Precision (mAP) metric. In particular, we compute mAP@100, where only the top 100 retrieved images contribute to the score. As in [1], this metric is defined as:

$$\mathrm{mAP@100} = \frac{1}{Q}\sum_{q=1}^{Q}\mathrm{AP@100(q)}, \tag{1}$$

where

$$\mathrm{AP@100(q)} = \frac{1}{\min(m_q, 100)}\sum_{k=1}^{\min(n_q,100)}\mathrm{P}_q(k)\mathrm{rel}_q(k) \tag{2}$$

Here, Q is the total number of query images, m_q is the number of index images containing an object in common with the query image q (images from the same class in the index), n_q is the number of predictions made by the system for query q (for our case it is always 100 as we always retrieve 100 images for this metric), $\mathrm{P}_q(k)$ is the precision at rank k for the q-th query; and $\mathrm{rel}_q(k)$ is a binary indicator function denoting the relevance of prediction k for the q-th query.

Results are presented in Table S1. It can be observed that there is high correlation between the mAP and the metrics reported in the main paper. For example, the highest performing method in all cases is obtained with CLIP pretraining and the oracle specialist. The three universal models based on CLIP pre-training perform very similarly: their relative ranking remains the same as the one of the mMP@5 metric. The same holds for the relative ranking of the universal models with IN pretraining. Additionally, for most domains, mMP@5 and mAP agree on the best model. We conclude that all metrics capture similar trends, while specifically mMP@5 and mAP are very correlated. To improve metric interpretability and simplicity, as discussed in Section 3.2, we thus decide to establish the two main metrics in our benchmark as mMP@5 and R@1.

A.2. Architecture ablation

We study the effect of the ViT architecture size, by comparing the performance of ViT-Small, ViT-Base (used in the main paper) and ViT-Large on our evaluation benchmark. Each of them has larger number of parameters than the previous one, being more memory and computationally expensive. We compare them by training with the UJCRR method (explained in the main paper), starting from IN pretraining.

Results shown in Table S2 justify our choice of ViT-Base as our main backbone; it is a good tradeoff for size and performance, performing as well as the larger ViT-Large, but a lot better than the smaller ViT-Small.

A.3. Separate index evaluation

We include results for an evaluation where each domain's queries are tested against the index of the same domain, instead of the merged index set, which is the main evaluation of our proposed benchmark. It corresponds to the setting where an Oracle is available, that restricts the index to images from the same domain as the one of the query image. For

this evaluation, only the CLIP pretraining is used.

Results are shown in Table S3, and each entry in the table can only be equal or greater than the corresponding one in the main paper. This is because all cross-domain mistakes are avoided in the current setting. We observe that the universal models and the oracle specialist performs slightly better on average in this setting, with the highest increase being in the Met domain. This could be caused by the fact that the Met domain contains artworks that can also be considered roughly parts of the other domains as well, *e.g*. clothing pieces, depictions of animals or landmarks in paintings, therefore making it easier to have cross-domain mistakes for Met queries. Additionally, CLIP+PCAw performance is also a lot higher, showing that naive unsupervised projection with PCA-Whitening produces a lot of cross-domain mistakes.

A.4. Specialists as universal embedding models

We present evaluation results for specialist models used as universal embedding models in Table S4, the highest values for each column are highlighted in bold, and the lowest in red. For this evaluation, only the CLIP pretraining is used.

As expected, for each domain, the best performing specialist model is the one trained on the corresponding training set, and the best pretraining for that domain corresponds to the one reported in Table 4 of the main paper. We also note that the best performing models are the specialist models finetuned on Met and Rp2k domains, though the performance of these models is still a lot worse than the best universal model reported in the main paper. Interestingly, finetuning on the GLDv2 domain performs the worse on average, for both types of pretraining.

A.5. PCA-Whitening vs Random Projection

We present a comparison of PCA-whitening as a means to reduce the dimensionality of the off-the-shelf embeddings shown in the Table 4 of the main paper versus Random Linear projection to 64-D, in Table S5. Results on the original dimensionality results are also shown for reference. The random linear projection results are averaged over 3 seeds. PCA-Whitening has been trained on the union of subsets of $\sim$9k images of each domain. We observe that for ImageNet pretraining, the random projection performs better on average than PCA-Whitening, while for CLIP pretraining it underperforms the former.

A.6. Visual presentation of all domains

In Figure S1 we show a collective presentation of example images from the different domains the UnED dataset covers.

References

[1] T. Weyand, A. Araujo, B. Cao, and J. Sim. Google Landmarks Dataset v2 - A Large-Scale Benchmark for Instance-Level Recognition and Retrieval. In *Proc. CVPR*, 2020. 1

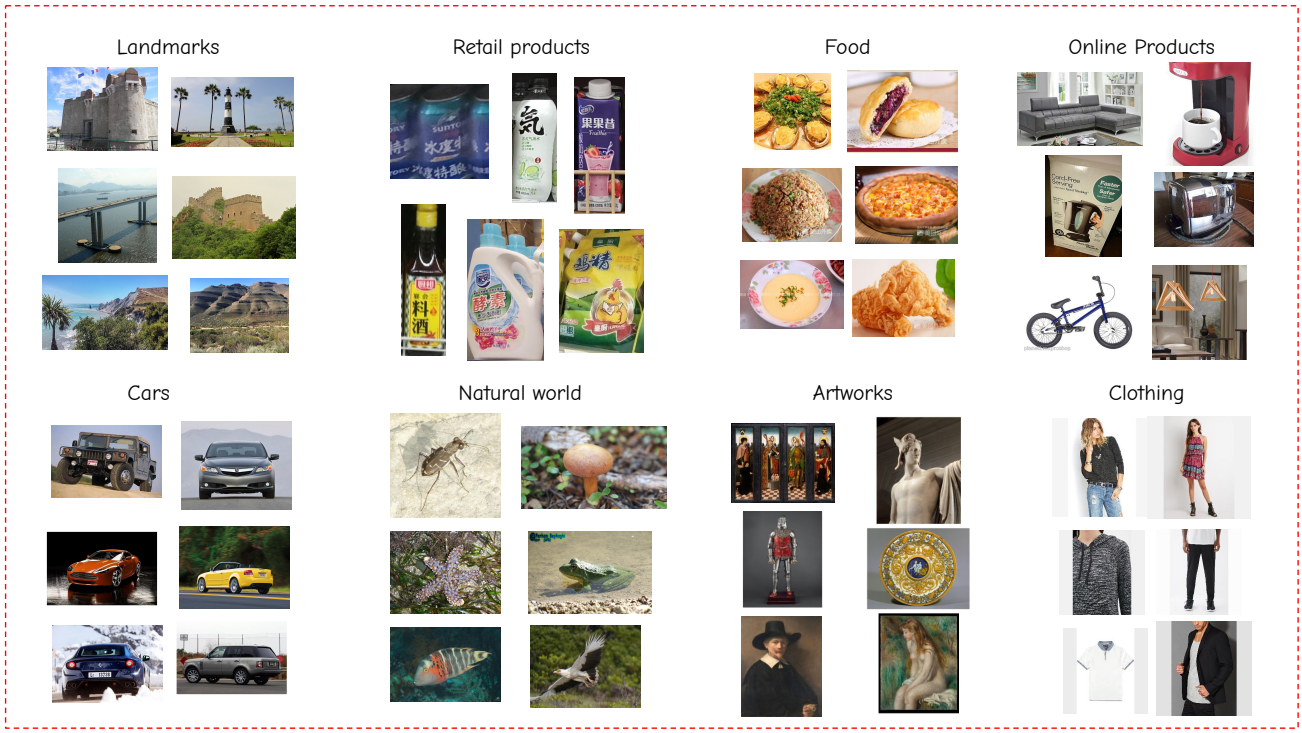

Figure S1: Example images from the different domains that the proposed UnED dataset covers.

	Food2k	CARS196	SOP	InShop	iNat	Met	GLDv2	Rp2k	Mean
Model	mAP@100								
Off-the-shelf									
IN (768-D)	23.6	9.6	38.5	34.5	43.1	22.6	7.8	41.9	27.7
CLIP (768-D)	21.4	29.0	39.3	36.0	24.3	23.7	11.6	28.6	26.7
IN + PCAw	13.6	5.9	26.3	18.2	27.7	9.0	3.6	28.4	16.6
CLIP + PCAw	17.1	20.0	32.2	25.9	17.9	14.0	6.4	23.4	19.6
Specialists									
IN+Oracle	42.7	19.9	56.6	64.8	**49.4**	24.1	19.4	64.8	42.7
CLIP+Oracle	43.7	**40.5**	62.6	66.2	43.9	**27.4**	**23.1**	59.5	**45.9**
Universal models									
IN+UJCDS	44.3	15.4	51.7	58.8	48.0	4.7	17.2	65.7	38.2
CLIP+UJCDS	**45.2**	33.4	55.6	63.1	41.3	2.6	21.2	62.2	40.6
IN+UJCRR	42.7	23.2	61.5	73.6	48.1	5.9	12.0	**66.1**	41.6
CLIP+UJCRR	43.9	39.5	**65.8**	**76.7**	40.4	5.9	15.2	61.5	43.6
IN+USCRR	42.2	16.7	58.2	69.5	48.6	8.0	13.2	65.0	40.2
CLIP+USCRR	41.7	36.2	61.6	71.7	40.5	9.7	15.4	62.8	42.4
IN+USCSS	40.8	13.3	57.2	65.6	47.0	11.5	16.6	64.2	39.5
CLIP+USCSS	42.1	33.5	62.9	70.2	42.5	8.5	20.3	61.9	42.7

Table S1: Corresponding mAP@100 for baselines presented in Table 4 on the main paper. Color coding follows Table 4.

	Mean	
Model	mMP@5	R@1
ViT-S (IN)	48.3	58.9
ViT-B (IN)	52.4	62.6
ViT-L (IN)	52.4	62.6

Table S2: Ablation for the model architecture. All models are finetuned with the UJCRR method described in the main paper.

	Food2k		CARS196		SOP		InShop		iNat		Met		GLDv2		Rp2k		Mean	
Model	mMP@5	R@1	mMP@5	R@1	mMP@5	R@1	mMP@5	R@1	mMP@5	R@1	mMP@5	R@1	mMP@5	R@1	mMP@5	R@1	mMP@5	R@1
Off-the-shelf																		
CLIP (768-D)	29.4	42.9	74.8	82.2	44.4	65.5	37.2	56.0	53.4	62.8	27.7	37.5	20.4	31.0	38.6	59.9	40.7	54.7
CLIP + PCAw	29.9	41.5	67.9	76.3	40.7	61.3	39.8	57.8	52.0	60.3	19.5	25.7	16.5	23.2	40.5	59.6	38.4	50.7
Specialists																		
CLIP+Oracle	52.9	64.4	83.3	88.6	67.5	82.1	69.2	86.9	69.3	74.7	33.1	39.8	36.0	47.7	71.1	85.1	60.3	71.2
Universal models																		
CLIP+UJCDS	51.3	62.9	76.1	82.4	58.9	75.7	62.7	80.3	65.0	70.7	5.7	7.0	33.3	45.6	70.2	84.2	52.9	63.6
CLIP+UJCRR	50.0	62.0	80.3	86.0	68.6	82.7	77.2	90.9	64.6	70.3	9.8	12.3	25.5	36.0	69.8	83.9	55.7	65.5
CLIP+USCRR	50.0	61.8	80.4	85.8	66.7	81.7	73.6	89.7	65.7	71.6	12.3	15.9	25.6	36.3	71.9	85.5	55.8	66.0
CLIP+USCSS	50.1	61.9	78.6	85.0	68.1	82.5	72.5	89.3	67.3	73.2	10.6	14.2	32.6	43.9	71.3	85.1	56.4	66.9

Table S3: Corresponding separate index evaluation for baselines presented in Table 4 on the main paper.

	Food2k		CARS196		SOP		InShop		iNat		Met		GLDv2		Rp2k		Mean	
Model	mMP@5	R@1	mMP@5	R@1	mMP@5	R@1	mMP@5	R@1	mMP@5	R@1	mMP@5	R@1	mMP@5	R@1	mMP@5	R@1	mMP@5	R@1
Specialists																		
IN+Food2k	50.3	63.0	30.9	40.5	29.4	49.1	21.8	34.7	53.5	60.2	7.8	7.0	8.8	13.6	40.5	61.9	30.1	41.2
CLIP+Food2k	**51.3**	**63.5**	70.1	78.8	29.6	49.0	25.5	40.9	42.9	50.3	4.6	6.0	16.7	24.4	34.9	56.0	34.4	46.1
IN+CARS196	19.6	29.4	62.4	72.0	26.1	45.1	20.3	33.1	54.8	61.3	11.6	16.6	8.1	12.7	38.6	60.1	30.2	41.3
CLIP+CARS196	19.0	28.7	**82.6**	**88.4**	29.1	48.4	24.5	39.9	42.5	50.1	10.0	13.8	14.7	22.2	27.6	46.5	31.2	42.2
IN+SOP	13.1	20.9	22.2	32.0	61.2	78.3	29.3	45.9	44.9	52.6	2.5	3.1	5.6	9.1	44.0	66.1	27.8	38.5
CLIP+SOP	10.7	17.9	44.3	56.5	**66.2**	**81.4**	32.1	50.0	30.2	38.3	3.0	4.3	8.6	13.5	37.0	59.2	29.0	40.1
IN+InShop	13.4	21.6	23.6	32.8	33.5	54.7	66.6	86.1	45.8	53.2	5.5	7.2	6.8	11.8	40.0	62.2	29.4	41.2
CLIP+InShop	13.1	21.0	61.0	70.6	31.7	51.8	**67.8**	**86.2**	35.0	42.5	6.3	8.3	12.1	19.8	31.0	52.1	32.2	44.0
IN+iNat	24.1	34.8	34.4	44.0	29.6	49.3	24.6	39.0	**70.0**	**75.1**	13.8	20.7	10.2	16.2	41.8	62.6	31.1	42.7
CLIP+iNat	17.6	27.5	61.4	71.0	30.6	50.1	27.4	43.4	67.1	72.7	10.1	13.6	11.3	16.9	34.0	54.7	32.4	43.7
IN+Met	14.7	23.7	28.5	39.4	38.0	59.5	33.8	52.8	43.2	51.0	21.7	25.9	9.6	16.1	48.6	70.3	29.8	42.3
CLIP+Met	16.1	25.4	59.6	70.2	43.8	64.5	40.5	61.5	36.9	45.1	**25.7**	**30.8**	16.1	24.6	44.6	66.9	35.4	**48.6**
IN+GLDv2	12.7	19.9	13.7	22.6	36.6	57.8	25.7	40.1	43.5	50.9	3.2	4.3	31.6	43.8	41.2	63.3	26.0	37.8
CLIP+GLDv2	9.7	16.4	23.4	33.3	33.3	53.6	22.3	36.5	26.1	33.7	3.3	4.4	**35.6**	**46.7**	26.8	46.3	22.6	33.9
IN+Rp2k	21.4	31.8	34.8	45.7	34.7	56.3	27.1	42.8	54.4	61.6	14.3	19.8	10.4	17.4	**73.6**	**87.2**	33.8	45.3
CLIP+Rp2k	19.0	29.1	62.8	71.6	34.6	55.6	29.3	46.0	38.9	47.1	15.5	20.5	15.4	25.2	69.6	84.6	**35.6**	47.5

Table S4: Results for specialist models when used as universal embeddings on our benchmark. Model column has the format : {Pretraining}+{Finetuning dataset}.

	Food2k		CARS196		SOP		InShop		iNat		Met		GLDv2		Rp2k		Mean	
Model	mMP@5	R@1	mMP@5	R@1	mMP@5	R@1	mMP@5	R@1	mMP@5	R@1	mMP@5	R@1	mMP@5	R@1	mMP@5	R@1	mMP@5	R@1
Off-the-shelf																		
IN (768-D)	31.1	44.1	41.4	54.1	43.7	65.6	35.5	53.9	67.1	74.2	21.1	30.8	14.8	25.2	52.9	74.3	38.4	52.8
CLIP (768-D)	29.4	42.9	74.7	82.2	44.2	65.4	37.2	56.0	52.4	61.9	21.4	28.5	20.4	31.0	38.6	59.9	39.8	53.5
IN+PCAw	19.1	29.1	29.0	37.8	30.5	51.2	19.6	31.6	50.9	57.9	8.0	11.0	8.3	13.2	37.6	57.8	25.4	36.2
CLIP+PCAw	23.4	34.6	62.8	72.7	36.5	57.0	27.0	41.8	42.7	51.1	12.1	15.8	11.9	17.6	32.0	51.8	31.0	42.8
IN+Rand.Proj.	19.4$\pm$0.5	29.5$\pm$0.8	31.0$\pm$0.3	41.7$\pm$0.5	33.1$\pm$0.1	54.5$\pm$0.3	25.7$\pm$0.6	40.4$\pm$0.3	54.4$\pm$0.2	61.5$\pm$0.2	8.8$\pm$0.5	12.2$\pm$0.7	8.7$\pm$0.2	14.8$\pm$0.7	38.7$\pm$0.1	60.0$\pm$0.2	27.5$\pm$0.1	39.3$\pm$0.1
CLIP+Rand.Proj.	18.1$\pm$0.7	28.5$\pm$0.8	61.7$\pm$1.3	71.7$\pm$0.8	34.5$\pm$0.3	55.0$\pm$0.5	26.8$\pm$0.5	42.0$\pm$0.8	41.3$\pm$0.2	49.8$\pm$0.2	9.7$\pm$0.7	13.2$\pm$1.0	12.5$\pm$0.5	18.5$\pm$1.2	29.3$\pm$0.6	47.7$\pm$0.5	29.2$\pm$0.5	40.8$\pm$0.7

Table S5: Comparison of PCA-Whitening vs Random linear projection. For the latter, the average of 3 seeds is shown.