Investigating CLIP Performance for Meta-data Generation in AD Datasets Supplementary Material

Sujan Sai Gannamaneni[1], Arwin Sadaghiani[1], Rohil Prakash Rao[2], Michael Mock[1], Maram Akila[1]
[1]Fraunhofer IAIS, [2]University of Bonn
{sujan.sai.gannamaneni,arwin.sadaghiani,michael.mock,maram.akila}@iais.fraunhofer.de,
s6roraoo@uni-bonn.de

1. Comparison of the ensembling approaches

In this supplementary material, we provide an evaluation of the two ensembling approaches discussed in our work. In the CLIP [2] based ensembling approach, described in section 3.1.4 of their paper, multiple prompts for each class are used. They take the form of "A photo of a big {label}" and "A photo of a small {label}", where label is the class we wish to identify. Here, the adjectives "big" and "small" will only provide more context about the class and do not modify it. To perform the ensembling, the mean of the text embeddings per-class is calculated, and then the similarity scores with image embeddings are obtained. As a final step, a softmax function is applied over the similarity scores. The resulting mean representation effectively acts as a linear classifier as the average is taken over the text embedding space. Due to this, many aspects of the distribution of the original embeddings, for instance, their spread, is lost.

In our approach, we first determine the cosine similarity of all text embeddings with a given image embedding, apply the softmax function across all resulting similarity scores, and then obtain the mean per-class. Here, the softmax effectively creates a non-linearity within the classifier. While, without softmax, taking an average over text representations or over similarity scores would be equivalent (up to a potential normalization of representations), applying a softmax directly on the level of per text-prompt similarities strengthens the contributions of those prompts that are closer to the image representation. This non-linearity could help in better classifying specific samples, for instance, if their semantic content is near one of the augmentations of the prompts, *e.g.*, "small" in the above example. To perform qualitative evaluations, similar to Tab. 2. in our main paper, we evaluate the performance of CLIP on the CelebA [1] dataset for both ensembling approaches, and the results are provided in Tab. 1. Here, by focusing on the F1 scores, we can see that in almost all dimensions, our ensembling approach either does not harm the performance or leads to improvement, which is specifically notable for the "goatee" attribute. Only in the dimension *bald* do we see a decreased performance.

References

[1] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 1

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1

Semantics	Attribute	Counts	Linear Ensemble (CLIP)				Non-linear Ensemble (Ours)			
			Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 score
Age	Young	156734	0.85	0.90	0.90	0.90	0.86	0.91	0.91	0.91
	Not-young	45865		0.66	0.68	0.67		0.70	0.70	0.70
Gender	Male	84434	0.99	0.99	0.98	0.99	0.99	0.99	0.98	0.99
	Not-male	118165		0.99	0.99	0.99		0.99	0.99	0.99
Skin-color	Pale	8701	0.50	0.07	0.89	0.13	0.56	0.07	0.81	0.13
	Not-Pale	193898		0.99	0.48	0.65		0.98	0.54	0.70
Hair-color	Black	47323	0.77	0.93	0.65	0.77	0.78	0.94	0.65	0.77
	Blond	28252		0.81	0.93	0.87		0.83	0.93	0.87
	Gray	7928		0.74	0.61	0.67		0.81	0.65	0.72
	Brown	39167		0.65	0.84	0.73		0.64	0.86	0.73
Misc.	Eyeglasses	13193	0.97	0.74	0.80	0.77	0.97	0.74	0.86	0.80
	No eyeglasses	189406		0.99	0.98	0.98		0.99	0.98	0.98
	Hat	9818	0.96	0.57	0.64	0.60	0.96	0.56	0.74	0.64
	No Hat	192781		0.98	0.98	0.98		0.99	0.97	0.98
	Bald	4547	0.95	0.22	0.49	0.30	0.93	0.19	0.60	0.29
	Not Bald	198052		0.99	0.96	0.97		0.99	0.94	0.96
	Goatee	12716	0.84	0.12	0.24	0.16	0.90	0.26	0.30	0.28
	No Goatee	189883		0.95	0.88	0.91		0.95	0.94	0.95
	Beard	33441	0.83	0.41	0.09	0.14	0.84	0.69	0.10	0.18
	No Beard	169158		0.84	0.98	0.90		0.85	0.99	0.91
	Smiling	97669	0.84	0.94	0.70	0.80	0.87	0.88	0.86	0.87
	Not-smiling	104930		0.78	0.96	0.86		0.87	0.89	0.88

Table 1. Comparison of the ensembling approach by CLIP and our proposed ensembling approach.