Positive-First Most Ambiguous: A Simple Active Learning Criterion for Interactive Retrieval of Rare Categories

Supplementary Material

7. Performance of Representativeness-based AL Strategies Across Iterations

Tab. 5 shows the performance of representative-based AL methods for different iterations. The strong performance of the uncertainty-based sampling strategy is consistent from the early retrieval stage.

Table 5. Performance comparison of representativeness-based AL methods vs. uncertainty-based selection on CIFAR100-LT at iterations 5, 15 and 25.

metric	*Rand*	*DAL*	*CoreSet*	*MA*
cov_5	0.05	0.039	0.039	**0.504**
$f1_5$	0.182	0.102	0.07	**0.828**
pos_5	0.022	0.019	0.02	**0.3**
cov_{15}	0.081	0.066	0.058	**0.839**
$f1_{15}$	0.294	0.215	0.099	**0.857**
pos_{15}	0.029	0.024	0.026	**0.537**
cov_{25}	0.111	0.096	0.087	**0.899**
$f1_{25}$	0.381	0.321	0.164	**0.857**
pos_{25}	0.036	0.032	0.035	**0.61**

8. Influence of Coverage Granularity over Iterations

Tab. 6 extends the results of the effect of coverage granularity on each method's performance to earlier iterations. The results are obtained on ImageNet-LT, using DINOv2 features. Our methods consistently shows strong and stable results since the early retrieval stages, rendering it robust to the choice of K. In contrast, other methods are sensitive to this value.

Table 6. Class coverage scores at iterations 5, 15 and 25 for different K values. Second best is underlined.

metric	method	K = 16	32	64
cov_5	*MA*	<u>0.55</u>	<u>0.423</u>	0.327
	MP	0.413	0.396	<u>0.399</u>
	ALAMP	0.49	0.387	0.315
	MA-S	0.51	0.377	0.28
	MA-D	0.513	0.355	0.246
	MP-S	0.427	0.381	0.343
	MP-D	0.447	0.378	0.297
	PF-MA	**0.586**	**0.493**	**0.4**
cov_{15}	*MA*	<u>0.807</u>	<u>0.736</u>	0.644
	MP	0.69	0.688	<u>0.679</u>
	ALAMP	0.757	0.7	0.622
	MA-S	0.799	0.722	0.62
	MA-D	0.774	0.653	0.527
	MP-S	0.688	0.681	0.666
	MP-D	0.68	0.647	0.582
	PF-MA	**0.812**	**0.79**	**0.778**
cov_{25}	*MA*	<u>0.848</u>	<u>0.804</u>	0.738
	MP	0.776	0.776	<u>0.764</u>
	ALAMP	0.808	0.771	0.712
	MA-S	<u>0.848</u>	0.802	0.731
	MA-D	0.833	0.751	0.648
	MP-S	0.779	0.777	<u>0.764</u>
	MP-D	0.773	0.76	0.723
	PF-MA	**0.849**	**0.844**	**0.843**

9. Performance Across Multiple Datasets and Feature Descriptors

9.1. Class coverage scores

We report the class coverage scores of the different methods, datasets and feature extractors in Tab. 7, across different iterations. The results highlight consistently strong performance of *PF-MA* since the early stages of retrieval. Other methods do not keep consistent rankings. This further guarantees that our method promotes very early user satisfaction.

9.2. Classifier performance

Tab. 8 reports the f1-score of the classifier across different methods, datasets and feature extractors. *PF-MA* is always among the top-3 methods in terms of classifier generalization, with minimal degradation, although our main objective is retrieval. Other competing methods have much lower coverage results as discussed before. Thus, *PF-MA* is able to maintain a fairly generalizable classifier, while achieving better coverage results.

10. Performance Across the Search Iterations and Per range of Class Size

10.1. Class coverage scores

Fig. 7 and Fig. 8 show the strong performance of *PF-MA* across iterations and per range of class size, for different datasets. The results are less pronounced on PlantNet300K because of its specific nature. *MP* and variants perform well on smaller classes, but they struggle for larger ones, where *MA* perform better thanks to the less imbalanced distribution of the negative vs. positive class problem. Our method yields consistently higher coverage across iterations, and

Table 7. Class coverage scores at iterations 5, 15 and 25 for different AL methods, datasets and descriptors. Second best is underlined.

metric	method	Cifar100-LT		ImageNet-LT		PlantNet300K	
		CLIP	DINOv2	CLIP	DINOv2	CLIP	DINOv2
cov_5	*MA*	0.402	<u>0.502</u>	0.351	<u>0.423</u>	0.203	0.282
	MP	<u>0.41</u>	0.454	**0.368**	0.396	**0.212**	0.29
	ALAMP	0.349	0.471	0.289	0.387	0.148	0.235
	MA-S	0.363	0.442	0.321	0.377	0.188	0.257
	MA-D	0.388	0.407	0.327	0.355	0.193	0.284
	MP-S	0.381	0.435	0.34	0.381	0.193	0.27
	MP-D	0.409	0.429	0.347	0.378	0.198	<u>0.294</u>
	PF-MA	**0.411**	**0.56**	<u>0.36</u>	**0.493**	<u>0.204</u>	**0.298**
cov_{15}	*MA*	0.801	<u>0.838</u>	<u>0.76</u>	<u>0.736</u>	<u>0.487</u>	<u>0.59</u>
	MP	0.773	0.785	0.713	0.688	0.456	0.537
	ALAMP	0.704	0.83	0.638	0.7	0.318	0.514
	MA-S	0.775	0.821	0.735	0.722	0.465	0.57
	MA-D	<u>0.815</u>	0.74	0.734	0.653	0.48	0.588
	MP-S	0.756	0.785	0.694	0.681	0.435	0.523
	MP-D	0.774	0.736	0.697	0.647	0.447	0.542
	PF-MA	**0.824**	**0.915**	**0.773**	**0.79**	**0.488**	**0.604**
cov_{25}	*MA*	0.89	<u>0.899</u>	<u>0.852</u>	<u>0.804</u>	**0.596**	0.678
	MP	0.875	0.877	0.821	0.776	0.565	0.628
	ALAMP	0.811	0.892	0.744	0.771	0.406	0.605
	MA-S	0.88	0.895	0.844	0.802	0.583	0.671
	MA-D	<u>0.903</u>	0.84	0.842	0.751	<u>0.595</u>	<u>0.679</u>
	MP-S	0.869	0.882	0.815	0.777	0.552	0.625
	MP-D	0.881	0.86	0.816	0.76	0.563	0.639
	PF-MA	**0.908**	**0.954**	**0.861**	**0.844**	**0.596**	**0.684**

presents the best compromise between smaller and larger classes. This highlights its practicality at test time, where early user satisfaction is desired, and the actual size of the user-defined concept in unknown.

10.2. Classifier performance

Both Fig. 9 and Fig. 10 show the ability of *PF-MA* to achieve competitive classifier generalization performance. This performance is steady across iterations, and across class sizes. *MP* on the other hand, has very poor classifier performance overall, always with a significant drop on larger class sizes.

Table 8. F1 scores at iterations 5, 15 and 25 for different AL methods, datasets and descriptors. Top-3 results are underlined. A star (*) is used only when our method achieves the top-1 result.

metric	method	Cifar100-LT		ImageNet-LT		PlantNet300K	
		CLIP	DINOv2	CLIP	DINOv2	CLIP	DINOv2
$f1_5$	*MA*	0.604	0.83	0.504	0.675	0.191	0.413
	MP	0.258	0.597	0.187	0.573	0.077	0.221
	ALAMP	0.56	0.806	0.45	0.646	0.147	0.373
	MA-S	0.57	0.816	0.476	0.667	0.176	0.398
	MA-D	0.517	0.803	0.437	0.661	0.159	0.398
	MP-S	0.325	0.666	0.241	0.59	0.091	0.279
	MP-D	0.358	0.637	0.245	0.579	0.093	0.296
	PF-MA	0.601	0.814	0.505*	0.67	0.191*	0.41
$f1_{15}$	*MA*	0.659	0.86	0.59	0.727	0.296	0.498
	MP	0.431	0.71	0.378	0.63	0.174	0.388
	ALAMP	0.644	0.858	0.56	0.709	0.237	0.476
	MA-S	0.656	0.858	0.584	0.725	0.283	0.493
	MA-D	0.637	0.85	0.571	0.723	0.268	0.492
	MP-S	0.445	0.727	0.387	0.631	0.174	0.39
	MP-D	0.438	0.725	0.383	0.624	0.17	0.396
	PF-MA	0.665*	0.861*	0.587	0.723	0.296*	0.495
$f1_{25}$	*MA*	0.648	0.859	0.593	0.735	0.298	0.508
	MP	0.456	0.73	0.416	0.63	0.196	0.421
	ALAMP	0.647	0.864	0.58	0.722	0.266	0.493
	MA-S	0.649	0.862	0.593	0.735	0.295	0.506
	MA-D	0.648	0.856	0.589	0.734	0.29	0.508
	MP-S	0.472	0.746	0.427	0.632	0.202	0.429
	MP-D	0.49	0.742	0.434	0.632	0.207	0.434
	PF-MA	0.649*	0.859	0.583	0.73	0.299*	0.507

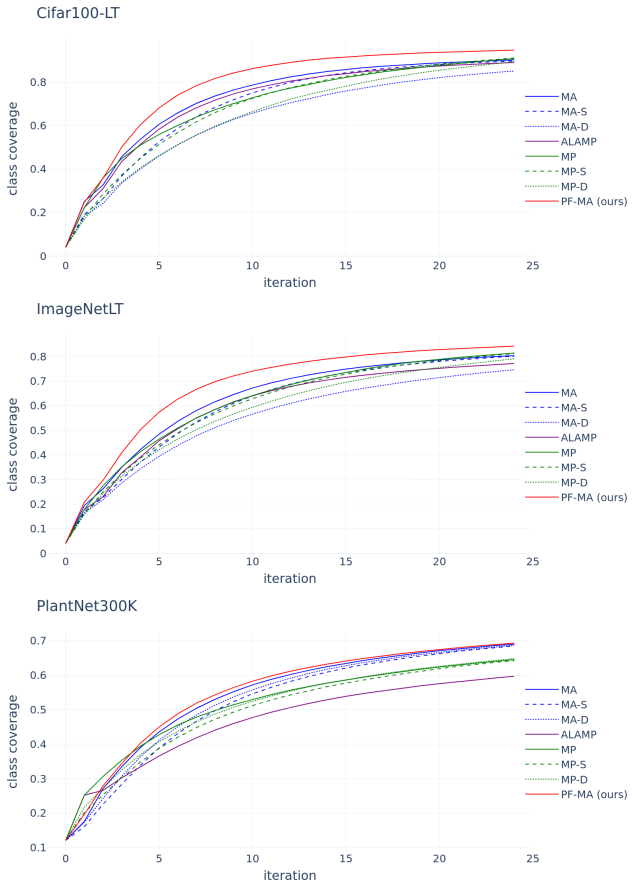

Figure 7. Class coverage results per iteration.

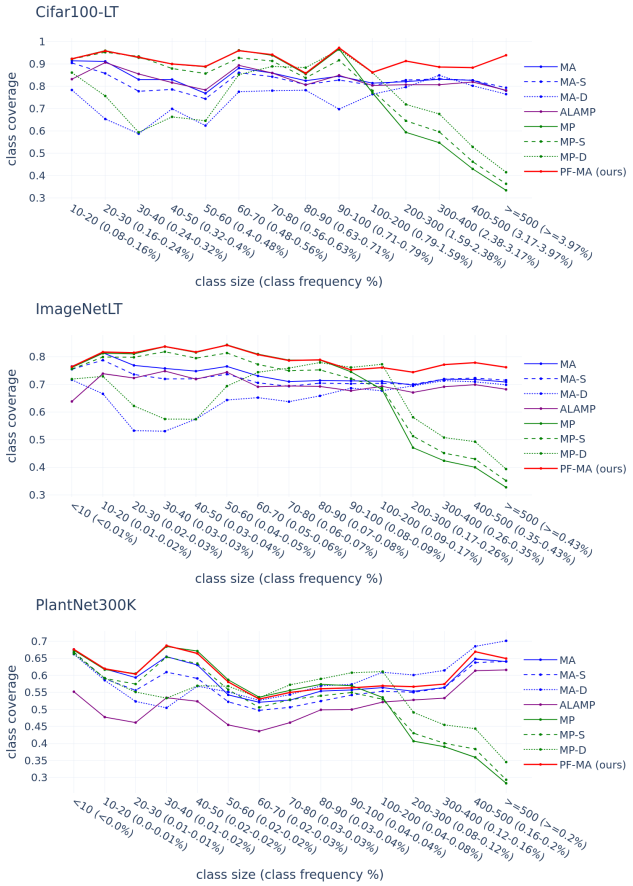

Figure 8. Class coverage results at iteration 15 per range of class size.

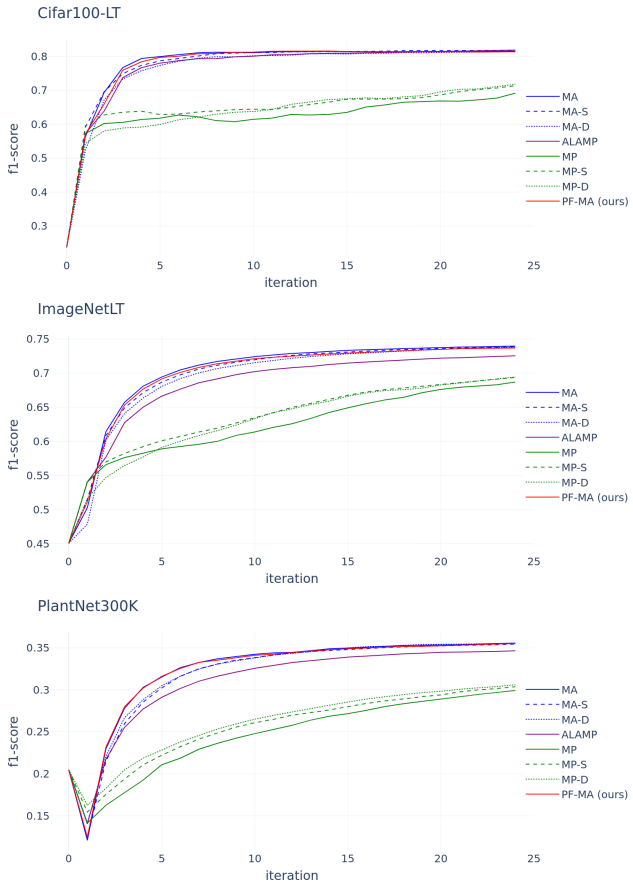

Figure 9. F1 score results per iteration.

Figure 10. F1 score results at iteration 15 per range of class size.