Supplementary Material: Diverse Image Annotation

Baoyuan Wu[†,‡] Fan Jia[†] Wei Liu[‡] Bernard Ghanem[†]
[†]King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia
[‡]Tencent AI Lab, Shenzhen, China
wubaoyuan1987@gmail.com fan.jia@kaust.edu.sa wliu@ee.columbia.edu bernard.ghanem@kaust.edu.sa

1. Evaluation Results by Conventional Metrics

Here we present the results of different methods, evaluated by conventional metrics, including precision (P), recall (R) and F_1 score, on ESP Game and IAPRTC-12, as shown in Table 2. ML-MG shows the best performance in all cases, which has also verified in [1]. In contrast, DPP-S-sampling gives the worst performance in all cases. The obvious reason is MLMG picks the most representative tags in top-k tags, such that more positive tags could be retrieved. In contrast, the diversity encourages DPP-S-sampling to cover tags from different semantic paths, such that some negative labels maybe included. However, as shown in the main manuscript, the conventional metrics are much less consistent with human evaluation than the semantic metrics.

data	metric→ / method↓	3 tags			5 tags		
		P	R	F_1	P	R	F_1
ESP Game	ML-MG [1]	**71.59**	**31.68**	**41.94**	**61.62**	**44.13**	**48.98**
	LEML [2]	62.77	27.48	36.41	53.24	37.80	42.06
	DPP-I-topk	68.63	30.8	40.44	59.28	43.06	47.37
	DPP-S-topk	68.59	30.7	40.38	59.18	43.05	47.32
	DPP-S-sampling	48.41	20.85	27.74	40.57	27.31	30.93
IAPRTC-12	ML-MG [1]	**79.59**	**28.52**	**39.44**	**70.32**	**40.44**	**48.29**
	LEML [2]	74.81	26.35	36.71	65.9	37.23	44.85
	DPP-I-topk	76.65	27.32	37.88	67.6	38.62	46.26
	DPP-S-topk	76.46	27.18	37.76	67.5	38.6	46.21
	DPP-S-sampling	56.46	18.53	26.5	48.57	24.72	30.8

Table 1. Results (%) evaluated by conventional metrics, including precision (P), recall (R) and F_1 score, on ESP Game (top row) and IAPRTC-12 (bottom row). The highest value in each column is highlighted in bold.

2. Combining Our Sampling with Traditional Methods

Here we add an experiment combining our sampling algorithm (see Algorithm 1 in the main manuscript) with the compared traditional image annotation methods, i.e., ML-MG and LEML [2]. The results are presented in Table 2. Our sampling is based on a DPP distribution, so the quality score in DPP should be replaced by the square root of the posterior score produced by ML-MG or LEML. The results of ML-MG are in the range $[0, 1]$, thus they can be directly used as the quality score. The results of LEML range from large negative to large positive values, so we normalize them to [0,1]. As shown in Table 2, in most cases ML-MG + sampling improves over ML-MG without sampling. Also, the value changes of different metrics are reasonable according to the characteristics of the original ML-MG scores. Specifically, the improvements of R_{sp} are 10.58% and 7.74% in the case of 3 tags and 5 tags respectively. The reason is that ML-MG puts the most representative but redundant tags in the top-k tag list, thus its R_{sp} value is lower than the one of other methods. With our sampling, the redundant tags will be removed, giving the chance to add other tags from different semantic paths, leading to the increase of R_{sp}. The improvements of P_{sp} is relatively small, and even negative in the case of 5 tags. This is not strange, because the removed redundant tags is likely to be relevant, while the new added tags may be irrelevant. Then the precision P_{sp} could decrease. This comparison could demonstrate that our sampling algorithm could help other traditional image annotation methods to increase the diversity. Besides, the performance of ML-MG + sampling is still worse than our DPP-S-sampling. This verifies that both our model learning and sampling contribute to the diversity performance. LEML performs worse in most cases. We find that lots of normalized scores of LEML are round 0.5, which should be the main reason for poor sampling.

data	metric→ / method↓	3 tags			5 tags		
		P_{sp}	R_{sp}	F_{1-sp}	P_{sp}	R_{sp}	F_{1-sp}
ESP Game	ML-MG	30.51	16.55	19.73	36.61	29.63	30.59
	ML-MG + sampling	37.4	**27.13**	**29.7**	34.73	**37.37**	34.09
	LEML	**45.16**	23.61	28.31	**41.82**	33.87	**34.58**
	LEML + sampling	34.53	24.17	26.89	29.67	34.12	30.08
IAPRTC-12	ML-MG	35.74	17.99	21.89	41.95	29.56	31.98
	ML-MG + sampling	41.94	**25.24**	**29.47**	38.33	**34.46**	**34.09**
	LEML	**43.03**	19.54	24.86	**47.27**	29.76	33.67
	LEML + sampling	40.82	22.89	27.74	34.87	31.92	31.55

Table 2. Results of combining sampling with ML-MG and LEML.

3. Quality Results

Some quality results are shown in Figure 1 and 2. For each image, we provide the complete tags, and the predicted

Figure 1. Some quality results on ESP Game data. For each image, we present the ground-truth tag subset, the tag subsets with 3 and 5 tags produced by three methods, and the F_{1-sp} scores.

Figure 2. Some quality results on IAPRTC-12 data. For each image, we present the ground-truth tag subset, the tag subsets with 3 and 5 tags produced by three methods, and the F_{1-sp} scores.

tags of ML-MG, LEML and DPP-S-sampling in both cases of 3 and 5 tags, as well as their F_{1-sp} values. We can see that in most cases DPP-S-sampling produces more representative and diverse tags than ML-MG and LEML, with the larger F_{1-sp} values.

References

[1] B. Wu, S. Lyu, and B. Ghanem. Ml-mg: Multi-label learning with missing labels using a mixed graph. In *ICCV*, pages 4157–4165, 2015. 1

[2] H.-F. Yu, P. Jain, P. Kar, and I. Dhillon. Large-scale multi-label learning with missing labels. In *ICML*, pages 593–601, 2014. 1