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.

			2.4			5.	
data	$metric \rightarrow$		3 tags			5 tags	
	method↓	Р	R	F_1	Р	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

data	$metric \rightarrow$		3 tags			5 tags	
	method↓	\mathbf{P}_{sp}	R_{sp}	$F_{1-\mathit{sp}}$	\mathbf{P}_{sp}	R_{sp}	$F_{1-\mathit{sp}}$
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
	LEML + sampling	40.82	22.89	24.80 27.74	47.27 34.87	29.76 31.92	31.55

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

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.

3. Quality Results

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

	The complete tags: 'brown building colors home house man people person room table wall white wood' 3 tags: ML-MG: 'people person colors' (0.6233) LEML: 'people person man' (0.2857) DPP-5-sampling: 'wood man room' (0.6667) 5 tags: ML-MG: 'people person colors man white' (0.2646) LEML: 'people person man building room' (0.5389) DPP-S-sampling: 'room table man chair wood' (0.7273)	The complete tags: 'building car house road shop side street' 3 tags: ML-MG: 'building sky side' (0.2694) LEML: 'building road car' (0.3569) DPP-S-sampling: 'street house car' (0.75) 5 tags: ML-MG: 'building sky side plant road' (0.2856) LEML: 'building road car street house' (0.75) DPP-S-sampling: 'people car street house side' (0.8)
	The complete tags: 'bug car colors grass green plant road sky white yellow' 3 tags: ML-MG: 'people person colors' (0.0262) LEML: 'person people plant' (0.0712) DPP-S-sampling: 'green man tree' (0.675) 5 tags: ML-MG: 'people person colors man woman' (0.3024) LEML: 'person people plant tree woman' (0.2533) DPP-S-sampling: 'man tree green woman' (0.6)	The complete tags: 'desert dune ridge formation group people person salt sand team tourist' 3 tags: ML-MG: 'tourist desert' (0.0251) LEML: 'people formation group' (0.0071) DPP-S-sampling: 'formation sand group' (0.8) 5 tags: ML-MG: 'people formation group water side' (0.0188) LEML: 'formation sand group ridge dune' (0.28) DPP-S-sampling: 'mountain tourist desert' (0.6667)
RETRATOS DE UNA OBSESION	The complete tags: 'colors man movie people person poster red show' 3 tags: ML-MG: 'colors people person' (0.4024) LEML: 'person people red' (0.2617) DPP-5-sampling: 'red man black' (0.75) 5 tags: ML-MG: 'colors people person man show' (0.03) LEML: 'person people red dark black' (0.299) DPP-S-sampling: 'red movie black man' (0.5714)	The complete tags: 'adult cloth clothes man people person shirt table tee-shirt woman' 3 tags: ML-MG: 'person people adult'(0.233) LEML: 'adult person table' (0.4857) DPP-S-sampling: 'woman table man' (0.8571) 5 tags: ML-MG: 'person people adult cloth man' (0.6714) LEML: 'adult person table people woman' (0.6714) DPP-S-sampling: 'man clothes cloth table woman' (0.898)
	The complete tags: 'art dance group hair man party people person woman' 3 tags: ML-MG: 'people person colors' (0.0133) LEML: 'people person woman' (0.0133) DPP-S-sampling: 'man hair woman' (0.6667) 5 tags: ML-MG: 'people person colors man woman' (0.3333) LEML: 'people person woman man group' (0.38) DPP-S-sampling: 'woman group hair man girl' (0.6114)	The complete tags: 'grass group horse meadow people plant sky tree' 3 tags: ML-MG: 'plant people grass' (0.4111) LEML: 'plant group grass' (0.0778) DPP-S-sampling: 'people meadow horse' (0.8571) 5 tags: ML-MG: 'plant people grass horse group'(0.5583) LEML: 'plant group grass meadow people'(0.6667) DPP-S-sampling:'horse vegetation meadow people tree' (0.75)
	The complete tags: 'car colors man people person truck white ' 3 tags: ML-MG: 'colors people person' (0.03) LEML: 'man car people' (0.6667) DPP-S-sampling: 'car white man' (0.8571) 5 tags: ML-MG: 'colors people person man car' (0.5881) LEML: 'man car people person wheel' (0.5714) DPP-S-sampling: 'white blue road man car' (0.75)	The complete tags: 'cloth clothes grass group meadow people plant polo shirt tree' 3 tags: ML-MG: 'plant people group' (0.2456) LEML: 'plant group tree' (0.2568) DPP-S-sampling: 'meadow tree people' (0.75) 5 tags: ML-MG: 'plant people group grass sky' (0.2622) LEML: 'plant group tree grass meadow' (0.5714) DPP-S-sampling: 'tree people clothes meadow sky'(0.6233)
	The complete tags: 'animal beak bird brown colors eye face mouth' 3 tags: ML-MG: 'colors animal face' (0.0767) LEML: 'animal face eye' (0.3774) DPP-5-sampling: 'brown bird eye' (0.8571) 5 tags: ML-MG: 'colors animal face eye brown' (0.6603) DPP-5-sampling: 'nose ear bird brown eye' (0.8571)	The complete tags: 'bell building front people person side spectator surfer wall' 3 tags: ML-MG: 'building wall people' (0.3513) LEML: 'person front adult' (0.3513) DPP-S-sampling: 'spectator side wall' (0.6214) 5 tags: ML-MG: 'building wall people person spectator'(0.6667) LEML: 'person front adult side people'(0.3513) DPP-S-sampling: 'spectator front surfer adult wall'(0.8571)

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.

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

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

Figure 2. Some quality results on IAPRTC-12 data. For each im-

age, 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.

2014. 1