Tagging like Humans: Diverse and Distinct Image Annotation (Supplementary Material)

Baoyuan Wu[1], Weidong Chen[1], Peng Sun[1], Wei Liu[1], Bernard Ghanem[2], and Siwei Lyu[3]

[1]Tencent AI Lab [2]KAUST [3]University at Albany, SUNY

wubaoyuan1987@gmail.com, powerchen@tencent.com, pengsun000@gmail.com, wliu@ee.columbia.edu, bernard.ghanem@kaust.edu.sa, slyu@albany.edu

1. Qualitative Results

Here we present some qualitative results of to compare the performance of DIA [3] and our proposed method D^{2}IA-GAN. The results on ESP Game [2] data are shown in Fig. 3, and those on IAPRTC-12 [1] data are shown in Fig. 4 (**see the last page**). In each sub-figure, the left is the original image, while the right shows the ground-truth complete tag list and the annotation results of DIA and D^{2}IA-GAN. For each method, we present two cases, including 3 tags (*i.e.*, each single tag subset includes at most 3 tags) and 5 tags (*i.e.*, each single tag subset includes at most 5 tags). At each case, we present at most 3 tag subsets. As described in Section 5.5 of the main manuscript, the DPP sampling process in each method (DIA or D^{2}IA-GAN) is run 10 times to obtain 10 tag subsets. Then 3 subsets with the largest tag weight summations are picked as the presented results. If there are same subsets in these 3 subsets, the redundant subsets are removed. Furthermore, we also present the F_{1-sp} score of the ensemble subset of these picked subsets. From these qualitative results, we can see that (1) in each single tag subset of both DIA and D^{2}IA-GAN, the included tags are semantically distinct to each other; (2) the tag subsets of D^{2}IA-GAN are more diverse than those of DIA, so the corresponding ensemble tag subset of D^{2}IA-GAN covers more semantic meanings than that of DIA.

2. Analysis of Human Annotations

As demonstrated in Introduction of the main manuscript, we have conducted a human annotation experiment by asking 3 human annotators to independently annotate the first 1000 test images in the IAPRTC-12 dataset, with the requirement of 'describing the main contents of one image using as few tags as possible. Here we present some analysis about these human annotation results.

The first analysis focuses on the single tag subset from each human annotator. As shown in Fig. 1, the statistics of the sizes of single tag subsets of 3 human annotators are plotted in 3 sub-figures separately. The average sizes of single tag subsets from 3 human annotations are $3.99, 5.15, 4.43$, respectively. This demonstrates that the single human annotator tends to describe the image content using a few **relevant** and **distinct** tags.

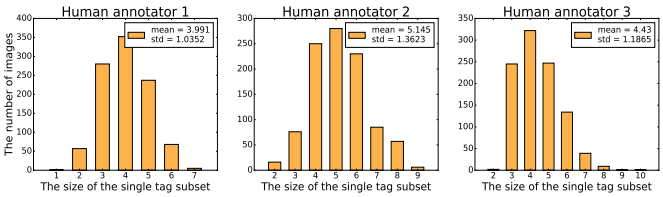


Figure 1. Statistics of the sizes of single tag subsets of 3 human annotations, on the first 1000 test images of IAPRTC-12 [1].

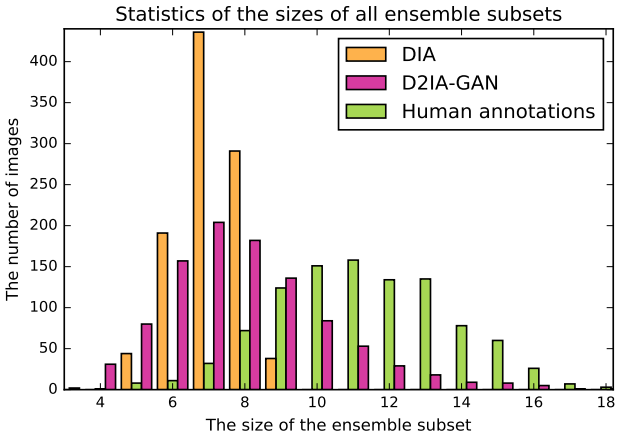


Figure 2. Statistics of the sizes of ensemble tag subsets of human annotations, DIA [3] and D^{2}IA-GAN, on the first 1000 test images of IAPRTC-12 [1].

The second analysis is about the ensemble tag subset, derived by merging 3 human annotated tag subsets for the

same image and removing the repeated tags. The statistics of the sizes of ensemble tag subsets of human annotations are presented as the green bar in Fig. 2. The average size of 1000 human annotated ensemble tag subsets is 11.22. The gap between the average size of single tag subsets and that of ensemble tag subsets reveals that different human annotators tend to give different tag subsets, *i.e.*, **diverse**. Moreover, we also present the statistics of the sizes of ensemble tag subsets produced by DIA [3] and our proposed method D^{2}IA-GAN in Fig. 2, distinguished by the yellow and purple colors, respectively. The ensemble tag subset is derived by merging 5 tag subsets produced by DIA or D^{2}IA-GAN and removing the repeated tags. These 5 subsets are exactly the results evaluated in our experiments (see Section 5.2 in the main manuscript), and each subset includes at most 5 tags. Specifically, the sizes of ensemble tag subsets of DIA range from 5 to 9, and the average size is 7.09. In contrast, the size range of D^{2}IA-GAN is from 3 to 18, and the average size is 7.95. This comparison indicates that the diversity between the tag subsets produced by D^{2}IA-GAN is larger than that by DIA, and the ensemble tag subset of D^{2}IA-GAN can cover more semantic meanings than that of DIA. However, the gap between the average size of ensemble tag subsets produced by human annotations and that by D^{2}IA-GAN reminds us that the diversity of our automatic tagging results is still smaller than the diversity of human annotations. The main reason is that both the training and inference of D^{2}IA-GAN are constrained in a fixed set of candidate tags provided by the dataset (291 candidate tags in IAPRTC-12, and 268 candidate tags in ESP Game [2]). In contrast, the tags used by human annotators are unconstrained, and many are out of the range of the fixed set we used. It is expected that D^{2}IA-GAN could produce more human-like tags if we train it on a larger set of candidate tags derived from the collection of real human annotated tags. This is our future research work.

References

[1] M. Grubinger, P. Clough, H. Müller, and T. Deselaers. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *International Workshop OntoImage*, pages 13–23, 2006. 2

[2] L. Von Ahn and L. Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326. ACM, 2004. 2, 3

[3] B. Wu, F. Jia, W. Liu, and B. Ghanem. Diverse image annotation. In *CVPR*, 2017. 2

Figure 3. Some qualitative results on ESP Game. The value in brackets at the end of each row indicates the F_{1-sp} score of the ensemble subset of the subsets in the same row.

The ground-truth complete tag list: {wood area cloth building curtain glass table bed door entrance room wall}
DIA:
3 tags: {cloth room window} {bed cloth window} {room window bed} (0.2982)
5 tags: {window room curtain table bed} {area curtain bed window table} {room cloth window table bed} (0.4706)
D2IA-GAN:
3 tags: {room bed curtain} {curtain bed wall} {room curtain wall} (0.6154)
5 tags: {room window bed lamp curtain} {curtain wood room window bed} (0.5)

The ground-truth complete tag list: {building frame table restaurant people wall window}
DIA:
3 tags: {tourist table area} {wall people table} {person table wall} (0.54)
5 tags: {window area tourist table} {person area window table} (0.5455)
D2IA-GAN:
3 tags: {tourist table wall} {table window person} {window people table} (0.8)
5 tags: {window tourist table area } {tourist desk window area } {window restaurant tourist table area} (0.7273)

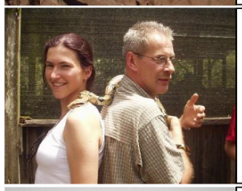
The ground-truth complete tag list: {mountain range front side tourist formation people person group}
DIA:
3 tags: {cloud people group} {mountain people cloud} {building mountain cloud} (0.4108)
5 tags: {cloud mountain group tourist plant} {plant tourist mountain cloud building} {people cloud plant mountain group} (0.5)
D2IA-GAN:
3 tags: {plant cloud mountain} {cloud tourist group} {mountain cloud group} (0.6)
5 tags: {terrace range group tourist cloud} {group mountain range tourist cloud} {tourist group mountain terrace cloud} (0.7273)

The ground-truth complete tag list: {airplane airport desert formation landscape ridge dune group plane field sand}
DIA:
3 tags: {airplane mountain group} {airplane sky group} {sky mountain group} (0.214)
5 tags: {airport airplane ridge group mountain} {airport ridge sky mountain airplane} {airport sky ridge group mountain} (0.3722)
D2IA-GAN:
3 tags: {airport airplane group} {airplane dune group} (0.6309)
5 tags: {group sky airplane dune field} {dune airport sky group airplane} {dune sky airplane group range} (0.4957)

The ground-truth complete tag list: {adult forest railing people person vegetation group plant man}

DIA:
3 tags: {plant man} {adult tree} {tree people} (0.2286)
5 tags: {man forest tree} {forest person tree} {adult vegetation tree} (0.6)
D2IA-GAN:
3 tags: {man tree} {man tree vegetation} {man forest tree} (0.75)
5 tags: {man tree forest} {tree forest man} {man forest tree } (0.6)

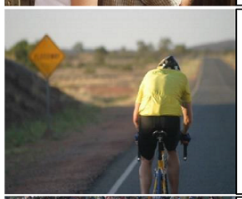
The ground-truth complete tag list: {formation desert mountain landscape ridge dune group sand}
DIA:
3 tags: {cloud landscape mountain} {sky mountain landscape} {mountain ridge landscape} (0.4327)
5 tags: {desert cloud mountain landscape} {landscape desert sky mountain} {sand ridge cloud landscape mountain} (0.7692)
D2IA-GAN:
3 tags: {sky mountain landscape} {mountain landscape cloud} {desert cloud landscape} (0.8333)
5 tags: {dune landscape cloud mountain} {desert cloud mountain landscape} (0.9091)



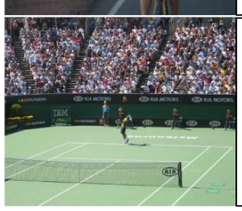
The ground-truth complete tag list: {adult woman shirt people person clothes man}

DIA:
3 tags: {adult clothes} (0.3267)
5 tags: {clothes man group woman} {woman group shirt man} {adult shirt group} (0.75)

D2IA-GAN:
3 tags: {adult clothes } {man clothes woman} (0.7056)
5 tags: {woman man group shirt } (0.8571)



The ground-truth complete tag list: {bicycle country shirt plate short road racing','jersey cyclist side people person clothes hand bike body helmet cycling}
DIA:
3 tags: {cyclist jersey cycling} {cyclist cycling shirt} {cloth jersey cyclist} (0.375)
5 tags: {jersey cloth cycling cyclist helmet} {cycling helmet cyclist shirt bike} {people jersey helmet bike cycling} (0.5263)
D2IA-GAN:
3 tags: {cycling cyclist helmet} {bike jersey cyclist} {bike jersey cycling} (0.625)
5 tags: {helmet cycling jersey cyclist short} {short bike helmet jersey cycling} {bike jersey helmet cyclist cycling} (0.7059)



The ground-truth complete tag list: {court area net tennis people person stadium room player}

DIA:
3 tags: {seat player room} {court seat player} {court space player} (0.4444)
5 tags: {court player seat tennis} {tennis seat player room} (0.6)
D2IA-GAN:
3 tags: {seat court player} {player seat stadium} {tennis seat court} (0.8)
5 tags: {court seat tennis stadium player} {grandstand player court tennis} (0.6667)

Figure 4. Some qualitative results on IAPRTC-12. The value in brackets at the end of each row indicates the F_{1-sp} score of the ensemble subset of the subsets in the same row.