Supplementary: Labelless Scene Classification with Semantic Matching

BMVC 2017 Submission # 457

1 Additional Experimental Results: Validation of the Visual and Semantic Consistency Assumption

The proposed semantic matching methodology is based on one fundamental assumption: *The visual relationships of high-level visual concepts, i.e., objects and scenes, are typically consistent with their semantic relationships expressed in text descriptions.* Hence if some objects such as *'microwave', 'stove'* and *'dishwasher'* typically appear in the images with scene category *'kitchen'*, one would expect that the semantic embedding vectors, φ('*microwave*'), φ('*stove*') and φ('*dishwasher*'), should be more similar to φ('*kitchen*') than the embedding vectors of other scene categories such as φ('*office*') and φ('*bathroom*'). To verify this assumption, we computed two types of statistics from the images and the word embedding vectors respectively on the MIT-Indoor dataset. We first computed the appearance frequency rate of the detected objects in images from each scene category. For each category, we kept the top five most frequent objects, their appearance frequency rates and names. Then for each object, we used its name embeddings to compute its semantic similarities to the embedding vectors of each scene category. For each object, we ranked the scene categories according to the similarity values (in decreasing order), and kept the ranking indices. We expect that each scene category to have small ranking index values in the ranking lists for its top five most frequent objects.

The collected statistics for four scene categories, *bathroom, kitchen, office* and *buffet* are reported in Figure 1. We can see that the similarity rank indices of all the scene categories regarding each of their top five most frequent objects are typically small, and their similarity rank indices in their top-1 most frequent objects are all 1s. Moreover, the average similarity rank index values of the four scene categories across their top five most frequent objects are 1.8, 1.2, 4.6 and 4.2 respectively. Compared to the largest rank index, i.e., the total number of scene categories, 30, all these numbers are reasonably small. All these results support our consistency assumption over the visual and semantic relationships of objects and scenes, and shows the proposed methodology has a reasonable foundation.

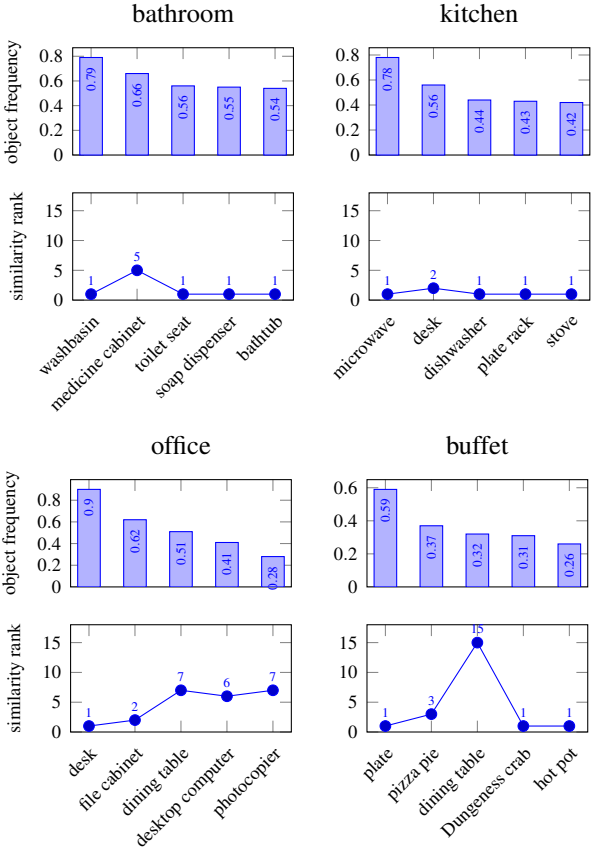

Figure 1: Statistics of scenes and objects for scene categories in MIT-Indoor dataset. Top row: the appearance frequency rate of the top 5 most frequently detected objects in each scene category. Bottom row: similarity ranking indices of the scenes with respect to each of their most frequently detected objects.