Supplementary Material: Fine-grained Image Classification and Retrieval by Combining Visual and Locally Pooled Textual Features

1. Introduction

In this document, as supplementary material, we present the following sections: Section 2 shows the results of mAP per class of our model compared with previous state of the art methods. The next Section 3 presents qualitative results obtained from the retrieval experiments performed. The Sub-section 3.1 uses the same image queries but defines two scenarios. The first one, in which the query image has all the text instances contained in the image blurred. The second one, contains all non-text containing areas blurred. A respective analysis is presented on the experiments regarding each section.

2. Results on Classification per Class

Tables 1 and 2 show the classification results in terms of average precision (AP%) per class on the Con-Text and the Drink Bottle datasets respectively. Significant improvement over previous state of the art is achieved due to the robustness of the employed descriptor. The improvement is considerably higher among most of the classes in the Drink Bottles dataset, whereas is smaller in the Con-Text dataset. Nonetheless, we do not employ an ensemble of classifiers which can specialize on specific classes. We can see in Table 1 that our model does not perform significantly better or worse in any obvious case except in the class "Bistro", in which the proposed pipeline achieves a considerable improvement. Images that belong to the class "Bistro" contain visually similar features that are shared with other classes aside from difficult to recognize text instances. A model that can leverage morphology rather than semantics is specially useful in such cases. A design, such as the presented in this work, that leverages syntax in a word helps to construct a powerful and discriminative model over fine-grained classes.

3. Qualitative Retrieval Results

The application task of Query by example (QbE) employs the output probability vector of the model as the global features that describes a given image. The distance score employed to measure the similarity between two images was the cosine similarity. The images are sorted in a ranked list according to the similarity score obtained when compared to the image that serves as a query.

We can observe in Figure 1 and Figure 2, that the model is able to retrieve images that contain text that is present in the queried image, even if the visual characteristics of the text are diverse. In some queries, the retrieved samples may or may not contain text. This fact enforces the conjecture that the model is able to learn a separable space in which the morphology of text lies close to the corresponding visual features.

3.1. Importance of Text at Retrieval

We provide two additional figures per dataset to further investigate the significance of text in the proposed retrieval application. We run a QbE scenario by blurring the query images with two different approaches, first of which is manually blurring the text instances in the image. Notice that this task is extremely hard to perform even for humans due to the importance of text. The second approach is to blur everything except the text instances. In order to compare each of the proposed scenarios, the queries were performed using the same images found in Figures 1 and 2.

It can be seen in Figure 3 and Figure 5, when the text from the query is blurred, the model retrieves mostly wrong samples. This effect is more profound in the Con-Text dataset since in order to differentiate buildings and storefronts textual features are significantly more important than visual ones. Some of the Drink Bottle dataset classes contain regular visual features which can be generalized even if the text is not present. This notion is reinforced by the results presented in the Sub-section 4.4 found in the main document.

On the other hand, Figure 4 and 6 shows retrieved samples when all the regions of a query are blurred but the ones that contain text. We can observe that the retrieval performs reasonably well in the Con-Text dataset, whereas in the Drink Bottle dataset, in the depicted samples, textual features alone can yield a close to ideal retrieval. These results proof our conjecture that reading text is extremely

Table 1: Classification performance for previous state of the art and our method on the Con-Text dataset. The depicted values are presented in terms of the Average Precision % (AP). Method depicted by * employs an ensemble model.

Class \ Method	Kar.[3]	Kar.[2]	Bai[1]	Bai[1]*	**Ours**
Massage Parlor	34.9	-	-	81.8	**82.2**
Pet shop	45.2	-	-	**89.5**	84.7
Restaurant	51.1	-	-	78.6	**83.0**
Computer Store	33.8	-	-	80.6	**86.2**
Theatre	48.5	-	-	**92.4**	90.1
Repair Shop	17.8	-	-	80.1	**83.2**
Bookstore	60.0	-	-	**94.2**	93.5
Bakery	37.8	-	-	**89.6**	87.4
Medical Center	48.5	-	-	**83.6**	82.9
Barbershop	55.2	-	-	**95.8**	92.3
Pizzeria	55.2	-	-	**90.4**	87.3
Diner	43.4	-	-	**86.7**	85.6
Hotspot	65.7	-	-	80.3	**85.2**
Bistro	9.1	-	-	32.4	**49.8**
Teahouse	12.5	-	-	68.9	**72.7**
School	44.3	-	-	81.3	**81.6**
Pharmacy	60.8	-	-	**88.4**	86.6
Funeral	43.0	-	-	**88.4**	85.1
Country Store	35.2	-	-	78.5	**80.2**
Tavern	10.6	-	-	52.2	**52.8**
Motel	53.0	-	-	**93.3**	88.8
Packinghouse	38.2	-	-	**85.0**	82.7
Cafe	16.2	-	-	57.0	**62.1**
Tobacco Shop	29.0	-	-	72.3	**76.3**
Dry Cleaner	50.9	-	-	**93.3**	88.1
Discount House	18.7	-	-	51.7	**54.0**
Steakhouse	28.1	-	-	**74.3**	72.4
Pawnshop	44.5	-	-	**87.0**	79.9
mAP	39.0	77.3	78.9	79.6	**80.2**

useful when dealing with visually similar classes. It can be seen on the last query on Figure 4, that an error in the PHOC prediction of the text in the query yields non relevant images, thus enhancing the notion of text relevance and precision when embedding text as an incremental step for future work.

Table 2: Classification performance for previous state of the art and our method on the Drink Bottle dataset. The depicted values are presented in terms of the Average Precision % (AP). Method depicted by * employs an ensemble model.

Class \ Method	Bai[1]*	**Ours**
Rootbeer	76.8	**83.7**
Gingerale	70.3	**76.6**
Coke	**95.8**	94.9
Pepsi	**96.9**	94.1
Cream soda	52.5	**61.7**
Egg cream	73.2	**79.9**
Birch beer	42.9	**61.2**
Quinine water	65.6	**71.8**
Sarsaparilla	57.8	**60.6**
Orange soda	**87.4**	87.2
Pulque	61.7	**67.9**
Kvass	39.4	**48.64**
Bitter	77.3	**82.0**
Guiness	**96.7**	93.4
Ouzo	63.8	**70.4**
Slivovitz	54.1	**61.9**
Drambuie	79.9	**83.9**
Vodka	85.7	**85.8**
Chablis	90.2	**91.6**
Sauterne	88.7	**89.3**
mAP	72.8	**77.4**

Figure 1: Qualitative results in Con-Text Dataset. The first image corresponds to the queried image class. The images are ranked from left to right. Red border represents a mistaken retrieved image that do not correspond to the queried class. (Best viewed in pdf).

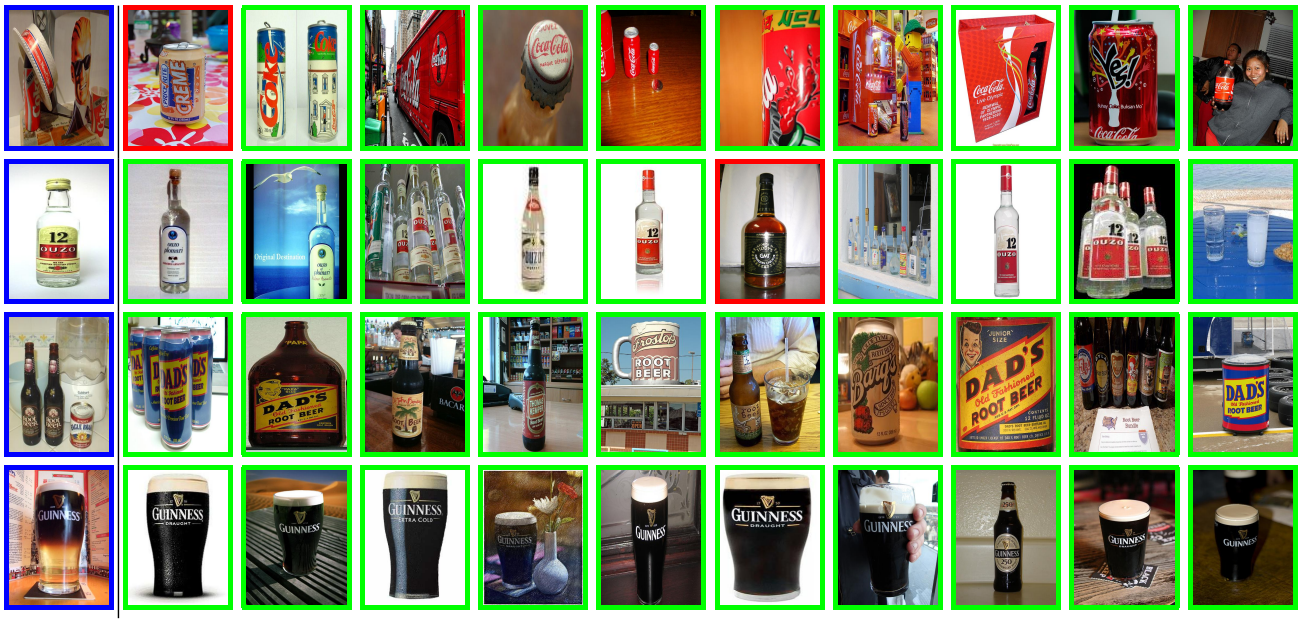

Figure 2: Qualitative results in the Drink Bottle Dataset. The first image corresponds to the queried image class. The images are ranked from left to right. Red border represents a mistaken retrieved image that do not correspond to the queried class. (Best viewed in pdf).

Figure 3: Qualitative results in Con-Text Dataset when the text in the queried image is blurred. (Best viewed in pdf).

Figure 4: Qualitative results in Con-Text Dataset. Results obtained when everything but the text is blurred in a queried image. (Best viewed in pdf).

Figure 5: Qualitative results in the Drink Bottle dataset when the text in the queried image is blurred. (Best viewed in pdf).

Figure 6: Qualitative results in Con-Text Dataset. Results obtained when everything but the text is blurred in a queried image. (Best viewed in pdf).

References

[1] X. Bai, M. Yang, P. Lyu, Y. Xu, and J. Luo. Integrating scene text and visual appearance for fine-grained image classification. *IEEE Access*, 6:66322–66335, 2018. 2

[2] S. Karaoglu, R. Tao, T. Gevers, and A. W. Smeulders. Words matter: Scene text for image classification and retrieval. *IEEE Transactions on Multimedia*, 19(5):1063–1076, 2017. 2

[3] S. Karaoglu, J. C. van Gemert, and T. Gevers. Con-text: text detection using background connectivity for fine-grained object classification. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 757–760. ACM, 2013. 2