Supplementary Material: Multi-Modal Reasoning Graph for Scene-Text Based Fine-Grained Image Classification and Retrieval - Paper ID: 709

1. Introduction

In this document, 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. Section 4 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. Finally, on Section 5 we present some samples of the attention maps obtained and the visualization of the relations between visual and textual regions.

2. Results on Classification per Class

We present in Tables 1 and 2 the predicted classification results in terms of average precision (AP%) per class on the Con-Text and the Drink Bottle datasets respectively. Due do the several layers of the GCN, a set of enriched semantic visual and textual features is obtained. The results obtained by using these features is considerably higher than previous methods. In the Con-Text dataset, the proposed model performs consistently better than previous state-of-the-art in 26 out of 28 classes which is translated in a big improvement in the fine-grained classification task. The classification of "Hotspot" and "Funeral" is a complex task due to some hard to recognize text instances and ambiguity between classes. The low score among all the methods in "Cafe", "Tavern" and "Bistro" can be explained by the lack of visual or textual cues in some test set images that lead to uncertainty when classifying such images even to humans. A similar effect is obtained among the classes in the Drink-Bottle dataset. However, due to the lower-quality of the images and narrower context of images (images focused solely in bottles), the ROIs obtained do not provide signals as strong as the ones in the Con-Text dataset. This previous notion, added up to the specific text instances that often encapsulate brand names, make the task of finding a semantic space harder than in Con-Text. Despite this limitations, we still perform superior than previous state-of-the-art approaches in 13 out of 20 classes.

3. Qualitative Retrieval Results

We present qualitative results based on the Query by Example (QbE) task on Figures 1 and 2 for the Con-Text and Drink-Bottle datasets respectively. In a QbE scenario, a system receives as an input a specific image, which belongs to a class seen in training time. The goal of the system is to retrieve a ranked list of the closest images that belong to the queried class. In order to measure the distance between the queried image and the retrieved samples, we employ the cosine similarity as it is described in the Fine-Grained Image Retrieval section. In all the Figures showcased, the first image (blue border) represents the image employed as query. Images with a green border represent a correct retrieval, whereas images with a red border represent wrongly retrieved samples. We can observe successful retrieval results in both Figures except in the third row of Figure 1, in which the model fails to retrieve the first sample. This effect is due to an incorrect OCR recognition of the text from the queried image. Added this to the fact that the visual features closely resemble the wrongly retrieved sample, outcomes in a wrong prediction. The rest of the samples show that a appropriate space that clusters similar labels is learned by employing textual along with visual features.

4. Relevance of Textual Features at Retrieval

To further provide insights of the relevance of scene text as discriminative features, we perform qualitative experiments in two scenarios. In the first scenario, we blurry the text found in an area given by a text detector and preserve the remaining visual features. In the second scenario, we blurry all the non-textual regions in a queried image and preserve only the scene text. In both scenarios, we used the same queried images as in Figures 1 and 2. Depending on the images used as query in the first scenario, blurring the text makes the retrieval task a very complicated problem to be solved even by humans. Figures 3 and 5 show the retrieved images for the Con-Text and Drink-Bottle datasets

in the first scenario. It is worth noting the significant drop in the retrieved images in both Figures. Specifically, in the first row in Figure 3, we observe that the model learns to recognize pastry in the storefront, which results in some correct retrieved samples. On the third row, the effect is similar based on visual cues alone. The remaining rows contain all wrongly retrieved images. A similar effect is found in Figure 5, which produces correct retrieved images in the first and fourth row due to the shared visual features between samples but incorrect retrievals at the second and third rows. Figures 4 and 6 depicts the outcomes of the second mentioned scenario. By using textual regions only, we can obtain better results than in the first scenario, strongly suggesting that there are several cases in images that contain scene text, in which textual features can be more discriminative than visual ones. Nonetheles in the third row in Figure 4, wrong OCR recognition produces erroneous retrieval of samples. The effect is similar in Figure 6 on the fourth row in the case of the Drink-Bottle Dataset.

5. Visualization

To offer understanding of the effect that the learned attention maps and the MMR module have on the predictions of the model, we show in Figure 7 and Figure 8 the original images, attention maps and affinity visualizations of the Con-Text and Drink-Bottle dataset respectively. The attention map is simply a self-learned mask by the CNN over a 7×7 grid. The affinity visualizations are defined by selecting the highest regions that present semantic correlation in the affinity matrix R in the last layer of the Graph-based MMR module. It is interesting to note that in Figure 7, third row, a strong semantic correlation is learned by attending at text regions and visual regions. For example, in the first two images, the text "Barber" is highly correlated to people cutting hair and to the red, white and blue barber pole. This effect is also evident in the text region that contains the text "Liquor" and the barrel located at the storefront. In the fifth sample, the text "pet shop" contains a strong semantic relation with a cat shown in the image. In the Drink-Bottle dataset, due to its noisy and low quality nature, the local visual regions extracted do not contain as rich features as the ones in the Con-Text dataset. The model learns to attend to textual regions as well as the more salient visual regions that generalize to a specific class of image.

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)

Class \ Method	Kar.[2]	Bai[1]	Mafla[3]	**Ours**
Massage Parlor	34.9	81.8	82.2	**87.9**
Pet shop	45.2	89.5	84.7	**96.2**
Restaurant	51.1	78.6	83.0	**82.4**
Computer Store	33.8	80.6	86.2	**92.2**
Theatre	48.5	92.4	90.1	**92.6**
Repair Shop	17.8	80.1	83.2	**89.1**
Bookstore	60.0	94.2	93.5	**97.4**
Bakery	37.8	89.6	87.4	**92.3**
Medical Center	48.5	83.6	82.9	**84.9**
Barbershop	55.2	95.8	92.3	**96.2**
Pizzeria	55.2	90.4	87.3	**91.8**
Diner	43.4	86.7	85.6	**89.3**
Hotspot	65.7	80.3	**85.2**	75.1
Bistro	9.1	32.4	49.8	**63.3**
Teahouse	12.5	68.9	72.7	**73.0**
School	44.3	81.3	81.6	**85.6**
Pharmacy	60.8	88.4	86.6	**90.0**
Funeral	43.0	**88.4**	85.1	87.5
Country Store	35.2	78.5	80.2	**85.9**
Tavern	10.6	52.2	52.8	**71.5**
Motel	53.0	93.3	88.8	**94.2**
Packinghouse	38.2	85.0	82.7	**88.1**
Cafe	16.2	57.0	62.1	**73.9**
Tobacco Shop	29.0	72.3	76.3	**82.4**
Dry Cleaner	50.9	93.3	88.1	**94.1**
Discount House	18.7	51.7	54.0	**73.2**
Steakhouse	28.1	74.3	72.4	**82.3**
Pawnshop	44.5	87.0	79.9	**92.3**
mAP	39.0	79.6	80.2	**85.81**

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).

Class \ Method	Bai[1]	Mafla[3]	**Ours**
Rootbeer	76.8	**83.7**	82.9
Gingerale	70.3	76.6	**76.8**
Coke	95.8	94.9	**96.9**
Pepsi	**96.9**	94.1	95.9
Cream soda	52.5	61.7	**66.7**
Egg cream	73.2	79.9	**83.8**
Birch beer	42.9	**61.2**	58.9
Quinine water	65.6	71.8	**72.9**
Sarsaparilla	57.8	60.6	**70.8**
Orange soda	87.4	87.2	**92.1**
Pulque	61.7	67.9	**72.7**
Kvass	39.4	48.64	**58.5**
Bitter	77.3	82.0	**86.9**
Guiness	**96.7**	93.4	95.22
Ouzo	63.8	**70.4**	69.5
Slivovitz	54.1	61.9	**66.35**
Drambuie	79.9	**83.9**	82.1
Vodka	85.7	85.8	**87.8**
Chablis	90.2	91.6	**94.6**
Sauterne	88.7	**89.3**	85.7
mAP	72.8	77.4	**79.87**

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 color).

Figure 2: Qualitative results in 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 color).

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

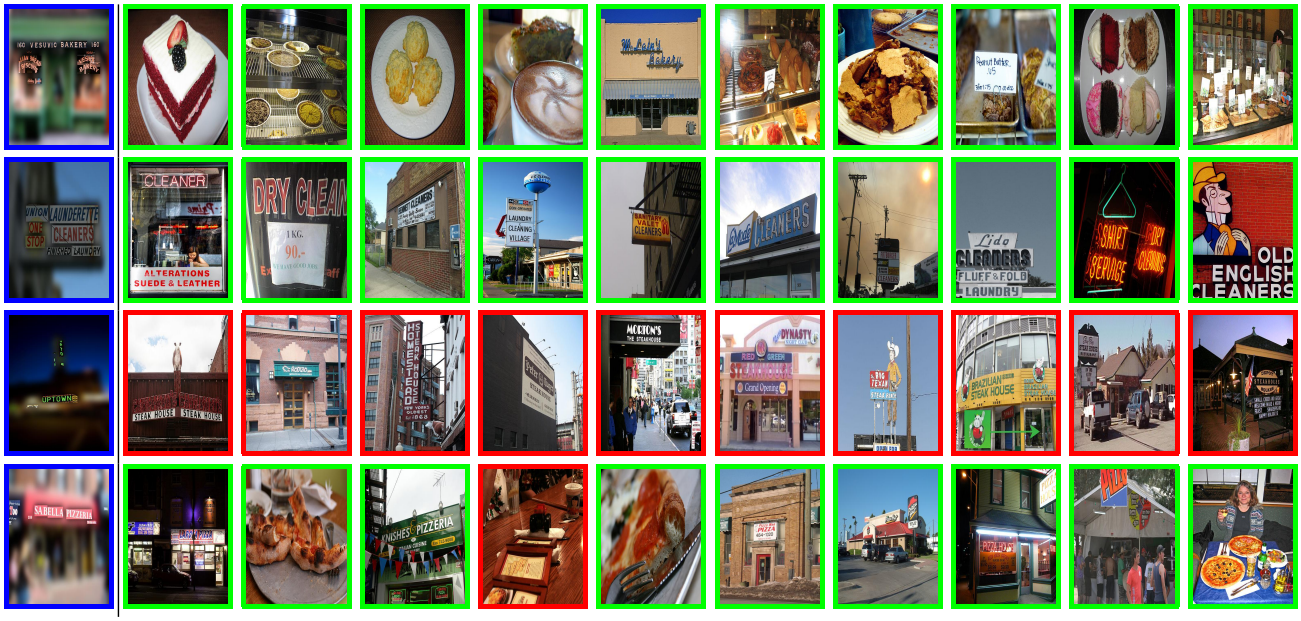

Figure 4: Qualitative results in Con-Text Dataset. Results obtained when everything but the text is blurred in a queried image. (Best viewed in color).

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

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

Figure 7: Visualization of the learned attention and enriched nodes of the model in the Con-Text dataset. First row: Original input images, second row: attention masks learned, third row: highest semantically correlated regions. (Best viewed in color).

Figure 8: Visualization of the learned attention and enriched nodes of the model in the Drink-Bottle dataset. First row: Original input images, second row: attention masks learned, third row: highest semantically correlated regions. (Best viewed in color).

References

[1] Xiang Bai, Mingkun Yang, Pengyuan Lyu, Yongchao Xu, and Jiebo Luo. Integrating scene text and visual appearance for fine-grained image classification. *IEEE Access*, 6:66322–66335, 2018. 3

[2] Sezer Karaoglu, Jan C van Gemert, and Theo Gevers. Context: 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. 3

[3] Andres Mafla, Sounak Dey, Ali Furkan Biten, Lluis Gomez, and Dimosthenis Karatzas. Fine-grained image classification and retrieval by combining visual and locally pooled textual features. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2950–2959, 2020. 3