Murat Ambarkutuk {murata@vt.edu}, 04/19/2016

Labeling Images with a Computer Game

In their paper, the authors propose a game in which Internet images are inteded to be labeled by paired users. In the game, users are paired up randomly and are shown a random subset of-random-images one by one. The users are then asked to find what the other user may have typed for the shown image. If the agreement found on the user input, the users are rewarded with some score. Even though the users are not explicitly asked to describe images, the authors claim that the users tend to describe the images to obtain the score corresponding to the shown image. The authors mention that using a pair of users rather than one user makes sure that the labels generated by the users are cross-validated and coherent.

The authors also point out that the system has a built-in function where once the agreement on the keywords is established, then the keyword becomes a taboo word. By doing so, the authors try to make sure that there will be enough number of keywords captured by the game.

Results and Conclusions

The authors evaluate the overall success of the game by conducting two different experiments.

The first experiment evaluates the labeling quality by comparing the generated labels with some test users' responses. Among 1023 images with 5 or more labels, twenty were randomly chosen to evaluate the label generation abilities of the game. 15 users who have never played the game before were asked to describe the images with some keywords. The authors reported that for every 6 labels generated by the game, 5 labels were covered by the users. Moreover, for all randomly chosen images at least 3 keywords match was observed.

The second experiment was a manual assessment for the label generation ability of the game. The users were shown 2 questions with the same 20 random images and corresponding generated keywords. The first question evaluated how accurate the labels were, while the second question addresses the incorrect labeling. The authors found that more than 85% of the labels were shown to be correct, while the second question showed only less than 1% of the labels marked had nothing to do with the images.

The rationale behind the game is to find the most descriptive words for an Internet image by restricting the use of the words on the taboo list. However, another strategy to game the game that users could select words as simple as possible to collect as many points in shorter time frames. This can cause labels to be too generic, which may not be a favorable situation all the time. For instance, a picture of "The Red Vineyard" by the famous artist Vincent van Gogh should not be marked with generic words like rural, agriculture and farmers. I believe the significance of the painting is that the painting was the only instance of art the artist sold in his lifetime. Thus, the game I believe is not able to capture these details as the gist of the game define the picture as simple as possible to acquire the most points.

Another possible shortcoming of the game may be the fact that the game does not seem to capture emotions. For instance, whenever a graphical-designer searches for stock photos, one usually tries to be explicit about the picture to make sure that the picture represents the situation accurately. One example for the case may be "happy family enjoying barbecue in the field". During the game the picture a user searches for inherently may be labelled with "barbecue" and "field", there is no way to recover the emotions from this kind of labeling. Given that one proposed application of the game is to create coherent image retrieval results for search engines, I believe this shortcoming would limit the success of the system.

Finally, the system heavily relies on the simplicity of the images and/or the labels. As the image contents gets more complex, it is doubtful that the user-pairs can agree on one word to describe the image, if the image is not passed. Figure 3 shows the scenario in the last image: Even though the image shows a toy truck, the users inputted the word "car" to the system rather than a toy truck since it would be simple enough to obtain the score.