Murat Ambarkutuk {murata@vt.edu}, 04/14/2016

Labeling Images with a Computer Game

In their paper, the authors propose a game in which Internet images are inteded to label by paired users. In the game, users are paired up randomly and are shown random images. The user are then asked to find what 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 user rather than one user makes sure that the labels generated by the users are 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, random 20 were chosen to evaluate the label generation abilities of the game. 15 users who never play 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 asked 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 believed to be correct, while the second question showed only less than 1% of the labels marked had nothing to do with the images.

Internet people are mean, the language is too delicate to compansate for it. Microsoft and Twitter bot is an example.

The rationale behind the game is to find the most descriptive words for an Internet image by enforcing taboo list. However, my intuition points me that users should select words as simple as possible to collect as many points in shorter time frames. This can lead labels to be too generic, which is not a favorable situation all the time. For instance, a picture of "The Red Vineyard" by the famous artist Vincent van Gogh should not be mark as generic words like rural, agriculture and so on. 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 feelings. For instance, whenever I search for stock photos I try to be explicit about the picture to make sure that the picture represent the situation accurately. One example for the case may be "happy family enjoying barbecue in the field". Since during the game the picture one 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.

The last point should be mentioned that the system heavily relies on the simplicity of the images and/or the descriptors. As the image contents gets more complex, it is doubtful that the user-pairs can agree on one word to describe the image, if not passed. Figure 3 shows the mentioned case with the last image: Even though the image shows a toy truck, user prompted the system with keyword car because it would be simple enough to score.