R03/Sarthak Mittal/200050129

January 14, 2023

The paper points out the requirement of image descriptions by several applications. The current techniques work under the assumption that the image would be related to text appearing along with it. They say that the only method for obtaining "precise image descriptions" is **manual labeling**, which is tedious and costly. So they developed an interactive game where the people who played it labeled images for them. The labeling of images through their game could be useful for visually impaired individuals, for providing datasets for computer vision, for content filtering, and so on. They make use of people's **abilities** and **desire** in order to achieve this, instead of using computer vision.

The game that they developed was supposed to be played in pairs, but without communicating with their partners. For the image displayed to a pair, they need to guess what the other is typing. Once both players have typed the same string at some point, they **"agree on the image"** and move to the next. The game offers incentives in the form of points for every image, with a large bonus on 15 images. So in a way they encourage players to continue, and allow them to opt out on difficult images. However, the game only asks the players to "think like each other", and not "describe the image". Hence it turn out that the string on which both players agree is a **good label** for the image.

The game also has **"taboo words"**, which are words that the players previously agreed on for that image. They did this to make the players think beyond the general guesses, and also to get different labels for an image. The game ignores the labels that players give other than the ones they agree on. They also use a **threshold** for declaring a label as taboo for an image, where at least X pairs should have agreed on the label. They considered an image completely labeled when it was consistently passed on by the players. They also re-inserted fully labeled images after several months to account for changes in the language.

They also added a **pre-recorded** play option, where a single player could play against a partner that uses guesses recorded from an earlier session. This session of player and bot could further help in labeling. To avoid players using cheap tricks (for example, typing "a" on every image), they took several steps - random pairing can ensure players have no idea who their partner is, using large number of pre-recorded actions, keeping track of average agreement time to detect massive global agreement, taboo words and good label threshold.

They had initially added **350000** images which they collected from random

websites. The game also has a 73000-word English dictionary to handle spelling mistakes. They also acknowledged the fact that wide-range audience could result in more general labels. That is why they suggested **"theme rooms"**, where the labels obtained would be more specific.

To evaluate their method, they analysed the results of searching for random keywords. They found that the using game-generated labels gave highly appropriate results. They also analysed the labeling rate (number of labels collected per minute). They used **three evaluation metrics** - measuring precision when using labels as search, comparing labels generated by game to labels generated by experimental participants, and asking experimental participants whether labels generated by game were appropriate or not.

The **precision of searching** for images using their labels generated by the game was extremely high. The objective of the ESP game was actually to maximize score, and not giving relevant search terms. The **comparison** of labels given by random participants were also largely in agreement with labels produced by game. Their **manual assessment** also indicated that most labels were actually associated with the image.

They say that their game can fulfil the large database requirement for computer vision training. It can also improve image retrieval and search results. They claim that having proper labels for freely available images can also help in improving content filters. They concluded with the thought that such games could be made for other problems as well where the most effective method is to rely on **human perception**.