Rethinking the ESP Game

Stephen Robertson
Microsoft Research Cambridge
7 J J Thomson Avenue
CB3 0FB Cambridge, UK
ser@microsoft.com

Milan Vojnovic
Microsoft Research Cambridge
7 J J Thomson Avenue
CB3 0FB Cambridge, UK
milanv@microsoft.com

Ingmar Weber
Ecole Polytechnique Fédérale de
Lausanne
LTAA, Station 14
CH-1015 Lausanne, Switzerland
ingmar.weber@epfl.ch

CHI 2009, April 4 – 9, 2009, Boston, MA, USA
ACM 978-1-60558-247-4/09/04.

Abstract

The ESP Game [7] was designed to harvest human intelligence to assign labels to images - a task which is still difficult for even the most advanced systems in image processing [3]. However, the ESP Game as it is currently implemented encourages players to assign “obvious” labels, which can be easily predicted given previously assigned labels.

We present a language model which can assign probabilities to the next label to be added. This model is then used in a program, which plays the ESP game without looking at the image. Even *without any use* of the actual image, the program manages to agree with the randomly assigned human partner on a label for 69% of all images, and for 81% of images which have at least one “off-limits” term assigned to them.

We discuss how the scoring system and the design of the ESP game can be improved to encourage users to add less predictable labels, thereby improving the quality of the collected information.

Keywords

ESP Game, Image Labeler, Tagging.

ACM Classification Keywords

H.1.2 Information Systems: User/Machine Systems.

Introduction

The ESP game is one of the best examples of how to harvest the intelligence of thousands of contributors for a task which is still difficult for machines: labeling images [3]. Labeling images is useful as otherwise there is little chance to retrieve an image relevant for a given query. In [7] evidence is presented that images are generally relevant for the queries corresponding to their labels. However, it is questionable, how much a large image repository benefits if the label "car" is correctly assigned to an unlabeled image. Microsoft's Live Image Search currently returns 150 million results for this query. If the main purpose of the ESP Game is indeed to label images for search purposes, then adding informative tags such as "red bmw" or even "talbot 1923" seems more valuable.

The ESP Game in its most popular implementation (http://images.google.com/imagelabeler) fails to collect such informative labels. We show that the sets of tags already present can be generated from a low entropy distribution and new tags added by players are highly predictable given only the "off-limits" terms, which are the tags already assigned to the image.

Google apparently noted these shortcomings and introduced different scores between 50 and 150 points for different labels according to their "specificity". However, (i) this is not a strong enough differentiation, (ii) it punishes terms too much which are globally unspecific, but which add relevant information for the particular context, and (iii) the current scores are not directly linked to the degree of predictability of a label. We show how to redesign the ESP game to improve the quality of the collected labels.

Related Work

Apart from the ESP Game, where two players are randomly paired up and have to agree on appropriate labels for images, games "with a purpose" have also been applied to various other settings [8]. All of these games share the properties that (i) players share the common goal of "agreeing" on certain things, (ii) players are matched randomly, and (iii) no communication is allowed, as this would make the agreement trivial and prone to spamming.

Closely related to our study of inferring the next label to be added, is the issue of tag suggestion [2, 6]. In fact, the probabilistic model employed by our robot to play the ESP Game is taken from [2].

The issue of what actually constitutes a "good" label for an object was investigated in [4], where the focus was on which kinds of labels might be (globally) found useful by a user for at least some object, rather than a particular one. Related to the issue of the quality of labels is the question of why users tag [1] and how their usage patterns are influenced by other users and suggestions made by the tagging system [5].

Shortcomings of the ESP Game

If one looks at how people label images via the ESP Game, then one notices the following.

There is a lot of redundancy in the tag sets because of the use of synonymous labels. Of all 496 (out of 14.5K) images labeled with "guy" 81% were also labeled with "man".

Some labels tend to co-occur with other labels, even in cases when they are not synonyms, such as “water” and “blue”, or “sky” and “clouds”. For instance, we observed that 68% of all the 85 images labeled with “clouds” had also been labeled with “sky”.

There is a tendency to match on colors and over 10% of all off-limits labels are colors, with 3.3% of all tag occurrences being “black”.

People tend to add more generic labels such as “building” as opposed to “terraced house”.

The common reason for these points is that it is far more likely for two people to agree on a general term, than to agree on more specific terms. Even if a player knows that a particular image depicts an oak, it will be pointless for her to enter this term, as the chances of agreement with her partner are considerably smaller than for “tree”, “leaves” or “green”. Note that none of the assigned tags are wrong, but the question arises, whether one has to rely on humans to obtain them. For these general terms there are already millions of publicly available and searchable images online.

A Probabilistic Tagging Model

In this section we describe our Naïve Bayes model for predicting the next tags, given a set *S* of tags already assigned. This model was presented in [2]. As the goal of the model is to test what can be predicted *without* using the image, we will *only* use the set *S* and no other information related to the image.

The probability we are interested in can be written using Bayes’ formula as follows.

P(‘*t* is next label’ | ‘set *T* already present’)

= *P*(‘set *T* already present’ | ‘*t* is next label’) * *P*(‘*t* is next label’)/*P*(‘set *T* already present’)

For our applications, we assume that the probability that *some* label will be added next is 1. This allows us to drop the denominator from consideration as we know that the expression, summed over all possible terms *t*, has to yield 1.

The probability of *P*(‘*t* is next label’) can be estimated by the number of occurrences of the tag *t* among the observed tag sets, divided by the total number of observed tags. So the only probability left to estimate is *P*(‘set *T* already present’ | ‘*t* is next label’). If we had enough training data for every possible set *T*, we could directly estimate this probability. But given the unavoidable problem of data sparsity due to an exponential number of possible sets, we make the following conditional independence assumption.

P(‘set *T* already present’ | ‘*t* is next label’)

= $\Pi_{s\in T}$ *P*(‘*s* is already present’ | ‘*t* is next label’)

The individual probabilities *P*(‘*s* is already present’ | ‘*t* is next label’) are estimated by dividing the number of tag sets in which the label *s* occurs before *t* by the total number of tag sets containing *t*. To avoid zero probability estimates the probability *P* is replaced by a smoothed variant *Ᵽ*.

Ᵽ(‘*s* is already present’ | ‘*t* is next label’)

= (1 - λ) *P*(‘*s* is already present’ | ‘*t* is next label’) + λ *P*(‘*s* is already present’)

The *P*(‘*s* is already present’) is estimated as the number of observed tag sets containing *s* divided by the total number of tag sets. In all of our experiments, we used a value of λ = 0.85, chosen using a validation set not part of the results (Table 1). With this smoothing, we then obtain the following probabilistic model.

P(‘*t* is next label’ | ‘set *T* already present’)

= $\Pi_{s \in T}$ $\tilde{P}$(‘*s* is already present’ | ‘*t* is next label’) * *P*(‘*t* is next label’)/*C* (1)

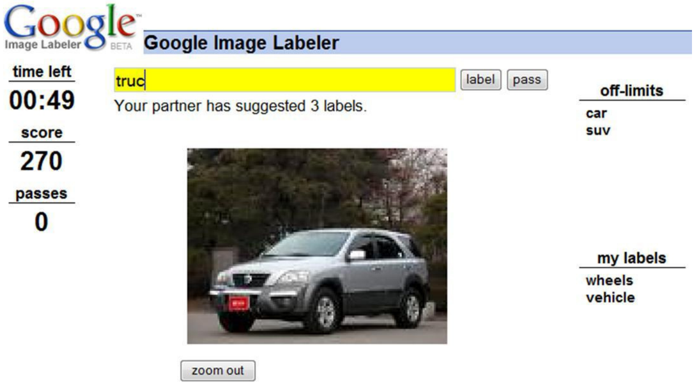
where *C* is a normalizing constant such that the sum over all *t* is 1. In settings where *T* is empty, we use the probability *P*(‘*t* is next label’).

Ultimately, the actual model used is not crucial as our main objective is simply to show that the labels on the ESP game are predictable using only the off-limits terms. An improvement in the model would only strengthen this claim.

A Robot Playing the ESP Game

We used the model presented in the previous section to implement a robot which plays the game without extracting any information from the image itself, demonstrating that labels are highly predictable.

The input rate was throttled to play more human-like. Averaged over all the 2,600 games played, our robot suggested around 4.3 labels per image before (i) finding an agreement, (ii) passing or (iii) running out of time. This corresponds to an input rate of 4.4 seconds per label entered, compared to 5.1 seconds for human players. Figure 1 shows a screen-shot of the robot playing the game.



time left
00:49
score
270
passes
0

truc

Your partner has suggested 3 labels.

off-limits
car
suv

my labels
wheels
vehicle

Figure 1. A screen-shot of our robot playing the ESP game. Using the off-limits terms “car” and “suv”, it has produced the list “wheels, vehicle” and is entering “truck”. This will then lead to a match for 120 points.

If both the probability and the points for a potential match are known, the robot could either try to maximize the number of matches, always choosing the most likely matching label, or it could play to maximize its score, weighting the probability of a match by its number of points. We experimented with both strategies. Surprisingly, the scores used by Google were only weakly (negatively) correlated with the global frequencies of the tags and the two strategies agreed in 95% of the cases on the first tag to enter. Table 1 gives a summary of the robot’s performance when it tries to maximize the number of matches. It finds a match for 81% of images with at least one off-limits term (OLT). For images, where a match was found, the match was usually between 2nd and 3rd in the list of suggestions made by the human. This indicates that the robot manages to “read the human’s mind” and does not match on strange terms.

Number of	
- games played	205
- images encountered	1,335
- images with OLTs	1,105
Percentage with match	
- all images	69%
- only images with OLTs	81%
- all entered tags	17%
Average score	
- per game	467
- per image	72
- per image with OLTs	85
- per match	104
Average number of labels entered	
- per image	4.1
- per game	26.7
Agreement index	
- mean	2.6
- median	2.0

Table 1. Statistics for the test phase of the robot consisting of 205 game instances. By "agreement index" we mean the index, starting at 1 in the partner's list of suggestions, for which we found an agreement. A low agreement index indicates that we did not rely on the partner to enter several tags. Only images with a match are taken into account for this.

Amount of Information Assigned Labels

In this section, we will quantify, how much information was added at each step, as the list of off-limits terms grew from empty to (up to) 5 terms. For each label position, we measured the average information defined as -$\log_2 p(t)$, where $p(t)$ is the probability of tag t being added next as predicted by Equation 1. To avoid assigning zero probability if the next label was previously unseen, we allowed an unseen label to be generated next with a probability equal to the probability of the rarest tag being next.

Table 2 shows that there is an effect of "diminishing returns", where later terms add less and less information to the set already present.

Average information per position of label in tag set				
1	2	3	4	5
9.2	8.5	8.0	7.7	7.5

Table 2. As the previous labels of a set are used to predict the next label, the labels become more and more predictable. An equidistribution over all seen 4,958 words would correspond to 12.3 bits. If terms were independent of the previous labels the information at any position would be 9.3 bits.

To see if humans work their way towards less and less predictable terms, as they think of more labels to add, we looked at the information gain, with respect to the off-limits terms, of the labels suggested by humans as a function of their position among the tag sequence. Table 3 shows that the information does indeed go up for later labels.

Av. information per position of human suggestions				
1	2	3	4	5+
8.7	9.4	10.0	10.6	11.7

Table 3. The average information, when all the off-limits terms are used to predict the next label, goes up with every additionally suggested label. This indicates

that (i) a player has to think of less and less obvious tags to suggest and that (ii) the notion of "obvious" is indeed correlated with the notion of "high probability".

Re-designing the ESP Game

In this section we discuss several approaches to encourage players to enter more informative terms.

The probability estimates of our model could be used to award points proportional to the amount of information ($-\log_2(p)$). A term such as "red" would bring fewer points in the context of "sunset, sky", where it is predictable, and more points in the context of "face, nose", where it adds more information to the system.

The more advanced scoring system above has a nice selling point. Players could be (correctly) told that the goal of the game is to outwit the machine. E.g., if two players find an obvious match, they would (i) get fewer points and (ii) a smug message could be displayed "Haha! I saw that coming!". On the other hand, if they agree on an informative term, the system could say "Oh, you caught me by surprise!", and the players would be awarded more points.

As an alternative way of enforcing more informative tags, terms could come with a time limit before they are activated. That is, an informative term such as "frigate bird" could still lead to an immediate match, but it takes, say, 10 seconds before the term "black" becomes active and can lead to a match.

Off-limits terms could be hidden and only their number revealed to the players. If they then agree on such an unknown taboo term they get zero points. If they agree on a *non-taboo* term, they would be awarded a unit score, independent of the match. This way, both players would probably start with less obvious labels, as they are less likely to be already present, and then work their way up to more predictable ones.

Conclusions and Future Work

Our robot shows that, in its current implementation, no understanding of the image at all is required to agree on labels in the ESP game. Thus, modifications should be explored to get the most out of the human effort. For future work, it is of interest to evaluate different reward schemes through user studies.

Bibliography

[1] Ames, M. and Naaman, M. Why we tag: motivations for annotation in mobile and online media. In *Proc. CHI'07* (2007), pages 971–980.

[2] Garg, N. and Weber, I. Personalized, interactive tag recommendation for flickr. In *Proc. RecSys'08* (2008), pages 67–74.

[3] Jeon, J., Lavrenko, V. and Manmatha, R. Automatic image annotation and retrieval using cross-media relevance models. In *Proc. SIGIR'03* (2003), 119–126.

[4] Sen, S., Harper, F. M., LaPitz, A. and Riedl, J. The quest for quality tags. In *Proc. GROUP'07* (2007), 361–370.

[5] Sen, S., Lam, S. K., Rashid, A. M., Cosley, D., Frankowski, D., Osterhouse, J., Harper, F. M. and Riedl, J. Tagging, communities, vocabulary, evolution. In *Proc. CSCW'06* (2006), 181–190.

[6] Sigurbjörnsson, B. and van Zwol, R. Flickr tag recommendation based on collective knowledge. In *Proc. WWW'08* (2008), 327–336.

[7] von Ahn, L. and Dabbish, L. Labeling images with a computer game. In *Proc. CHI'04* (2004), 319–326.

[8] von Ahn, L. Games with a purpose. In *Computer* (2006), Vol. 39, No. 6, 92-94.