LIA @ MediaEval 2013 Crowdsourcing Task: Metadata or not Metadata? That is a Fashion Question

Mohamed Morchid, Richard Dufour, Mohamed Bouallegue,
Georges Linarès and Driss Matrouf
LIA - University of Avignon (France)
{firstname.lastname}@univ-avignon.fr

ABSTRACT

In this paper, we describe the LIA system proposed for the MediaEval 2013 Crowdsourcing for Social Multimedia task. The aim is to associate an accurate label to an image among multiple noisy labels collected from a crowdsourcing platform. In particular, the task participants have to predict two types of binary labels for each considered image. The first one mentions that an image is truly fashion-related or not, while the second label indicates that the fashion tag assigned to the image is related to the content of the image or not. The proposed system combines noisy crowdsourcing labels, image metadata and external resources.

1. INTRODUCTION

Since the advent of *Web 2.0* [6], Internet users actively participate in information construction and propagation. Thus, many companies rely on users to give their opinion on movies or on musics, to annotate specific data... In June 2006, Jeff Howe [1] defined for the first time this new behavior with the term of *crowdsourcing*.

Thus, crowdsourcing makes it possible to rapidly and cheaply collect supervised labeled data. Nonetheless, annotation quality is uneven since it depends of the annotator, its implication, or its expertise level. As part of the collected data can be very noisy and inaccurate, solutions should be proposed to evaluate the relevance of the labeled data with a minimum time cost.

For these reasons, the Crowdsourcing task becomes a more and more popular and helpful task [3, 2]. In this paper, we describe the LIA system presented at the MediaEval 2013 Crowdsourcing task. The proposed system uses different parts of the image metadata: the annotator contribution (confidence[1], label), the image context, its geographic coordinates, and text descriptors (title, tags). All this described image metadata will be used in our system to decide if an image is truly fashion-related or not (*Label1*) while the relevance of the fashion category (*Label2*) will be chosen using the annotator content only.

2. PROPOSED APPROACH

The proposed approach focuses on the textual part of the images such as the image metadata, its geo-localization and the metadata of the images occurring in the same context. Five systems (*i.e.*, runs) are submitted. Each run is divided into two subtasks. The first one is to evaluate a score $\tilde{x}_i$ (with $\{i = 1, \ldots, 5\}$) for each label j ($j = 1$ for the label *Yes* and $j = 0$ for *No*) and then to select the label l_i with the highest score $\tilde{x}_i$ as:

$$l_i = \begin{cases} Yes & \text{if } \tilde{x}_i^0 \leq \tilde{x}_i^1 \\ No & \text{otherwise} \end{cases} \quad (1)$$

- RUN 1: **Crowdsourcing Annotation**
  The crowdsourcing annotation contributed by the general crowd is used to decide if an image is fashion-related or not, and if it has been correctly tagged. The best label $\tilde{x_1}$ knowing a set of workers $\mathbf{W}$ for both questions is estimated in a similar way with a Naive Bayes method. This *Yes/No* classification is expressed by:

$$\begin{aligned} \tilde{x}_1^j &= \underset{x}{\text{argmax}}\, P(X^j = x | Y_w^j) \\ &\propto P(Y_w^j | X^j) P(X^j) \\ &= \prod_{w \in \mathbf{W}} P(Y_w^j | X^j) P(X^j) \end{aligned} \quad (2)$$

  where we assume that labels of each worker are conditionally independent. Y_w^j corresponds to one label of the worker w; X^j also corresponds to one label and x is the true label. Thus, each image is labeled with the label l_1 as explained in the previous section.

- RUN 2: **Context**
  Each image is potentially related to other images: they have been grouped into sets or pools c according to annotators whether they were annotated by the same person. The annotator label of these images is used to estimate a score of the image context $\tilde{x}_2^j$ for fashion-relatedness:

$$\tilde{x}_2^j = \tilde{x}_1^j \times \sum_{i=1}^{c} \sum_{k=0}^{d} P(X^j | d_k^i) \quad (3)$$

  where d_k^i is the k^{th} image of the context (set or pool) i, X^j is a label (*Yes* or *No*), and $\tilde{x}_1^j$ the crowdsourcing annotation score.

- RUN 3: **Geo-Localization**
  An image could be localized close to one or several other ones. Here, a geo-localization score $\tilde{x}_3^j$ is defined. This score is calculated for the images that have a distance of zero to the current image. Furthermore, only images that have a label probability X^j greater than

[1]Self-reported confidence of the image category.

MediaEval 2013 Workshop, October 18-19, 2013, Barcelona, Spain

$\frac{2}{3}$ (estimated on the development set) are considered close and relevant for the fashion-relatedness question:

$$\tilde{x}_3^j = \tilde{x}_1^j \times \sum_{i=1}^{s} P(X^j|d_i) \quad (4)$$

where d_i is an image from the set s of images with a distance equals to 0, and $\tilde{x}_1^j$ the crowdsourcing annotation score.

- RUN 4: **External Resources**
  Each image comes with metadata (title and tags[2]) and a search rank on the hosting service. The first step is to collect a set of relevant pictures that are fashion-related. The Flickr API is used to compose a new set of 4,328 images that responds to the query *fashion*. The metadata are extracted from this set of images. The second step computes the probability score of each word based on the frequency of the word in the metadata. Thus, a metadata score $\tilde{x}_4^j$ is calculated for an image d by using the probabilistic model and the rank r of the image for the question 1 (fashion or not):

$$\tilde{x}_4^j = \tilde{x}_1^j + \frac{(-1)^{j+1}}{r} \sum_{i=1}^{|d|} P(w_i|d) \quad (5)$$

  where $P(w_i|d)$ is the probability of a word w_i from the image metadata of d knowing the model m of term frequencies. An image is labeled with the label l_4 knowing the annotator score $\tilde{x}_1^j$ and the term frequency model.

- RUN 5: **Combination**
  A final run using a combination of the 4 scores $\tilde{x}_i^j$ described above is submitted. A label is assigned to a picture if this label responds to the different aspects of the image metadata:

$$\tilde{x}_5^j = \tilde{x}_1^j \times x_2^j \times x_3^j \times x_4^j \quad (6)$$

  This score allows to associate an image with the general label l_5. We can notice that the score x_i^j is:

$$x_i^j = \frac{\tilde{x}_i^j}{\tilde{x}_1^j} \quad with \;\; i \neq 1$$

3. EXPERIMENTS

The proposed system is evaluated in the MediaEval 2013 benchmark [4]. In this task, we use the fashion social dataset [5]. Figure 3 presents 3 images extracted from the train dataset with different combinations of label mode-classification: (a) is not fashion-related (Culottes); (b) is not fashion-related but well categorized (Androgyny) and (c) is fashion-related and well categorized (Cowboy hat).

Table 1 presents results obtained in fashion-relatedness classification of images (*Label1*) in terms of F-measure. We can see that the best results are obtained using the crowdsourcing annotation only. Although slight gains were observed during the development phase of our systems, we have to note that the use of metadata information sources in addition to the crowdsourcing annotation do not improve the classification performance on the test set in comparison to the use of the crowdsourcing annotation only. This demonstrates that the crowdsourcing annotation is at least as reliable as the metadata provided with the images, or that these other information sources should be better handled (other approaches or data selection) to expect classification gains. On the fashion tag classification (*Label2*), a F-measure of 0.7175 has been obtained (no image metadata used).

[2]Note that experiments on the development set showed that description or personal notes do not improve the results.

(a) (b) (c)

Figure 1: Sample of pictures from train dataset.

Table 1: Classification results for fashion-relatedness (*Label_1*) classification.

Run Id	Submission	*F1_Label1*
1	Crowdsourcing Annotation	**0.7239**
2	Context	0.7171
3	Geo-Localization	**0.7236**
4	External Resources	0.7176
5	Combination	0.7183

4. CONCLUSIONS

In this paper, an automatic image classification system has been proposed. This system combines different aspects of the metadata content. The main observation is that best results are obtained with the crowdsourcing annotation only (*Label1*). No gains have been observed using metadata on the test set, but efforts on new approaches and metadata selection should be continued to improve classification performance. Finally, the use of the image metadata will be explored for the fashion tag task (*Label2*).

5. ACKNOWLEDGMENTS

This work was funded by the SUMACC project supported by the French National Research Agency (ANR) under contract ANR-10-CORD-007.

6. REFERENCES

[1] J. Howe. The rise of crowdsourcing. *Wired magazine*, 14(6):1–4, 2006.

[2] M. Lease and O. Alonso. Crowdsourcing for search evaluation and social-algorithmic search. In *ACM SIGIR conference on Research and development in information retrieval*, pages 1180–1180. ACM, 2012.

[3] M. Lease, V. Carvalho, and E. Yilmaz, editors. *Workshop on Crowdsourcing for Search and Data Mining (CSDM)*. February 2011.

[4] B. Loni, M. Larson, A. Bozzon, and L. Gottlieb. Crowdsourcing for Social Multimedia at MediaEval 2013: Challenges, data set, and evaluation. In *MediaEval 2013 Workshop*, Barcelona, Spain, 2013.

[5] B. Loni, M. Menendez, M. Georgescu, L. Galli, C. Massari, I. S. Altingovde, D. Martinenghi, M. Melenhorst, R. Vliegendhart, and M. Larson. Fashion-focused creative commons social dataset. In *ACM Multimedia Systems Conference*, pages 72–77, 2013.

[6] T. O'Reilly. What is Web 2.0, 2005.