Crowdsourcing for Social Multimedia at MediaEval 2013: Challenges, Data set, and Evaluation

Babak Loni[1], Martha Larson[1], Alessandro Bozzon[1], Luke Gottlieb[2]

[1]Delft University of Technology, Netherlands
[2]International Computer Science Institute, Berkeley, CA, USA
{b.loni, m.a.larson, a.bozzon}@tudelft.nl, luke@icsi.berkeley.edu

ABSTRACT

This paper provides an overview of the Crowdsourcing for Multimedia Task at MediaEval 2013 multimedia benchmarking initiative. The main goal of this task is to assess the potential of hybrid human/conventional computation techniques to generate accurate labels for social multimedia content. The task data are *fashion-related images*, collected from the Web-based photo sharing platform Flickr. Each image is accompanied by *a*) its metadata (e.g., title, description, and tags), and *b*) a set of 'basic human labels' collected from human annotators using a microtask with a basic quality control mechanism that is run on the Amazon Mechanical Turk crowdsourcing platform. The labels reflect whether or not the image depicts fashion, and whether or not the image matches its 'category' (i.e., the fashion-related query that returned the image from Flickr). The 'basic human labels' were collected such that their noise levels would be characteristic of data gathered from crowdsourcing workers without using highly sophisticated quality control. The task asks participants to predict high-quality labels, either by aggregating the 'basic human labels' or by combining them with the context (i.e., the metadata) and/or the content (i.e., visual features) of the image.

1. INTRODUCTION

Creating accurate labels for multimedia content is conventionally a tedious, time consuming and potentially high-cost process. Recently, however, commercial crowdsourcing platforms such as Amazon Mechanical Turk (AMT) have opened up new possibilities for collecting labels that describe multimedia from human annotators. The challenge of effectively exploiting such platforms lies in deriving one reliable label from multiple noisy annotations contributed the crowdsourcing workers. The annotations may be noisy because workers are unserious, because the task is difficult, or because of natural variation in the judgments of the worker population. The creation of a single accurate label from noisy annotations is far from being a trivial task.

Simple aggregation algorithms like majority voting can, to some extent, filter noisy annotations [3]. These require several annotations per object to create acceptable quality, incurring relatively high costs. Ipeirotis et al. [1] developed a quality management method which assigns a scalar value to the workers that reflects the quality of the workers' answers. This score can be used as a weight for a single label, allowing more accurate estimation of the final aggregated label.

Hybrid human/conventional computing approaches combine human contributed annotations with automatically generated annotations in order to achieve a better overall result. Although the Crowsourcing Task does allow for investigation of techniques that rely only on information from human labels, its main goal is to investigate the potential of intelligently combining human effort with conventional computation.

In the following sections we present the overview of the task, and describe the dataset, ground truth and evaluation method it uses.

2. TASK OVERVIEW

The task requires participants to predict labels for a set of *fashion*-related images, retrieved from the Web photo-sharing platform Flickr[1]. Each image belongs to a given fashion category (e.g., dress, trousers, tuxedo). The name of the fashion category of the image is the fashion-related query that was used to retrieve the image from Flickr at the time that the data set was collected. The process is described in further detail below. For each image listed in the test set, participants predict two binary labels. *Label1* indicates whether or not the image is fashion-related, and *Label2* indicates whether or not the fashion category of the image correctly characterizes its depicted content. Three sources of information can be exploited to infer the correct label of an image: *a*) a set of 'basic human labels', which are annotations collected from crowdworkers using an AMT microtask with a basic quality control mechanism; *b*) the metadata of the images (such as title, description, comments, geo-tags, notes and context); *c*) the visual content of the image. Participants in the task were encouraged to use visual content analysis methods to infer useful information from the image. They were also allowed to collect labels by designing their own microtask (including the quality control mechanism) and running it on a crowdsourcing platform.

3. TASK DATASET

The dataset for the MediaEval 2013 Crowdsourcing Task consists of two collections of images. Both collections contain images collected from the Flickr photo-sharing platform. We collected only images with a Creative Commons Attribution license so that the dataset could be used for research and commercial purposes. The images are crawled us-

[1]www.flickr.com

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

ing the Flickr search API. A list of fashion items was drawn from the Wikipedia index page devoted to the topic of fashion[2]. These items are then used to query the Flickr search API, and then the resulting images along with their metadata are downloaded. As mentioned above, the query that returned the image is assigned to the image as its fashion category. The resulting collections contain a large number of images that are related to fashion, but also images that are returned in response to the fashion item queries, but are not related to fashion.

The first collection was published in the MMSys 2013 data set track [2] and is referred to as 'MMSys 2013'. It consists of 4815 images, their metadata (e.g., title, description, geotags, notes), and two sets of human generated labels. The first set of labels (referred to as 'basic human labels' or 'low fidelity ground truth') has been generated by AMT workers under the application of basic quality control. The second set of labels (referred to as the 'ground truth ground truth' or 'high-fidelity ground truth') contains more reliable labels that were created created by trusted annotators.

The second collection, 'Fashion 10000", contains 31,077 images, their metadata (which is parallel to that of the MMSys 2013 collection) and 'basic human labels' (i.e., low-fidelity ground truth) generated by AMT workers using basic quality control. The images in the collection are associated with 262 fashion categories. 'Fashion 10000' received its name because the original aim was to create a data set containing at least 10,000 fashion-related images. In the final data set, nearly two-thirds of the images are related to fashion. The 'Fashion 10000' collection is divided into three parts: *Dev60* containing 60% of the images, *Dev20* containing another 20%, and an independent test set containing the remaining 20% of images. The 'Fashion 10000' collection was not issued with high-fidelity ground truth. It was recommended that participants apply semi-supervised machine learning approaches to be able to make use of the combination of the high-fidelity ground truth in the MMSys 2013 collection combined with the low-fidelity ground truth to optimize their approaches.

4. 'BASIC HUMAN LABELS'

Each image in the Crowdsourcing Task data is associated with 'basic human labels' collected from three crowdworkers who provided judgments on both *Label1* and *Label2*. The crowdworkers could choose from three options: 'yes', 'no' or 'not sure' options. Each microtask contained four images belonging to same fashion category. Workers were provided with some positive and negative examples of fashion-related images in order to help them to make consistent decisions in cases that could be interpreted as ambiguous. In addition, they were asked about their familiarity with the fashion category of the image. Since workers might not be familiar with some fashion categories, each microtask also provided a short definition as well as a sample image, taken from Wikipedia, to describe the fashion category that images are taken from. Work that did not pass a basic quality control mechanism was not included in the data set. Crowdworkers must have answered all questions in the microtask and the answers were required to be consistent. This simple quality control mechanism was designed so that the 'basic human labels' produced using this microtask would have noise levels characteristic of human annotations generated without a sophisticated mechanism for quality control. The use of only a basic mechanism made it possible for participants of the task to explore more advanced quality control mechanisms in their approaches to the task. More information about the generation of 'basic human labels' for the MMSys 2013 data can be found in [2]. The annotation of the 'Fashion 10000' collection was carried out in a comparable manner.

[2]http://en.wikipedia.org/wiki/List_of_fashion_topics

5. GROUND TRUTH AND EVALUATION

The task was evaluated on the 'Fashion 10000' test set (20% 'Fashion 10000' collection, as mentioned above). Images that were labeled 'not sure' by the majority of workers for either *Label1* or *Label2* are not included in the test set. The evaluation was carried out with respect to a high-fidelity ground truth generated using an additional crowdsourcing task. This second task used a more advanced quality control mechanism: each microtask included one question for which a gold standard ground truth label was already available. This question served as a validation question. If it was not answered correctly, the labels collected in that microtask were discarded. The final high-fidelity ground truth labels were derived by applying a majority vote on three worker labels collected using this additional crowdsourcing task. The official evaluation metric of this task is the F1 score, the harmonic mean of precision and recall.

6. OUTLOOK

The Crowdsourcing for Multimedia Task ran in MediaEval 2013 in its first year as a so-called 'Brave New Task'. The results of the task will inform the development of possible future tasks. In particular, we are interested in understanding how to collect 'basic human labels' in the way most useful for experimentation. We are also interested in understanding how best to create high-fidelity ground truth against which predicted labels can be evaluated. We hope that the experiences this year will help us to develop better methods for studying hybrid human/conventional computation.

7. ACKNOWLEDGMENTS

This task is partly supported by funding from the European Commission's 7th Framework Programme under grant agreement N° 287704 (CUbRIK) and also by the Dutch National COMMIT program.

8. REFERENCES

[1] P. G. Ipeirotis, F. Provost, and J. Wang. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, pages 64–67, 2010.

[2] 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 *Proceedings of the 4th ACM Multimedia Systems Conference*, MMSys '13, pages 72–77, 2013.

[3] S. Nowak and S. Rüger. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, MIR '10, pages 557–566, 2010.