SocialSensor at MediaEval Placing Task 2014

Giorgos Kordopatis-Zilos, Giorgos Orfanidis, Symeon Papadopoulos, Yiannis Kompatsiaris
Information Technologies Institute (CERTH-ITI), Thessaloniki, Greece
{georgekordopatis, g.orfanidis, papadop, ikom}@iti.gr

ABSTRACT

We describe the participation of the SocialSensor team in the Placing Task of MediaEval 2014. We submitted three runs based on tag information for the full test set, using extensions over an existing language modelling approach, and two runs (one based on the full test set and the other on the 25,500 subset) based on visual content, using geospatial clustering and supervised-learning. Our best performance (median error 230km, 23% at 1km) was achieved with the use of tag features, using only internal training data.

1. INTRODUCTION

The goal of the task is to produce location estimates for a set of 510K images using a set of over 5M geotagged images and their metadata for training [1]. For the tag-based runs, we built upon the scheme of [4], extending it with the use of the Similarity Search method, introduced in [6]. We also devised an internal grid technique and a Gaussian distribution model based on the spatial entropy of tags to adjust the corresponding probabilities. For the visual-based location estimation, we attempted to build *visual location models*, though with limited success. All models were built solely on the training data provided by the organizers (i.e. no external gazetteers or Internet data were used).

2. APPROACHES

2.1 Tag-based location estimation

Baseline approach: The baseline method relies on an offline step, in which a complex geographical-tag model is built from the tags and locations of the approximately 5M images of the training set. The metadata used to build the model and the estimation of a query image are the tags, the title and the description. A pre-processing step was first applied to remove all punctuation and symbols and to transform all characters to lower case. After the pre-processing, all training images left with empty tags and title are removed, resulting in a training set of approximately 4.1M images. Note that the same pre-processing is applied on the test images before the actual location estimation process.

In contrast to last year's clustering [3], we divide the earth surface in rectangular cells with a side length of 0.01° for both latitude and longitude (approximately 1km near the equator). Consequently, a grid of cells is created, which we use to build our language model using the approach described in [4]. More specifically, we estimate the most probable cell for a query (test) image based on the respective tag probabilities. A tag probability in a particular cell is calculated as the total number of different Flickr users that used the tag inside the cell, divided with the the total count of different users in all cells. Note that in that way a user can be counted in the total count of all cells more than once.

In order to assign a query image in a cell, we calculate the probability of each cell summing up the contributions of individual tags and title words. The cell with the greatest probability is selected as the image cell. If during this process there is no outcome (i.e. the probability for all cells is zero), we use the description of the query image. For the test images where there is no result (e.g. complete lack of text), we set their location equal to the center of the most populated cell, of a coarse granularity grid (100km×100km), a kind of maximum likelihood estimation.

Extensions: We devised the following extensions:

Similarity Search: Having assigned a query image to a cell, we then employ the location estimation technique of [6]: we first determine the k most similar training images (using Jaccard similarity on the corresponding sets of tags) and use their center-of-gravity (weighted by the similarity values) as the location estimate for the test image.

Internal Grid: In order to ensure more reliable prediction in finer granularities, we built the language model using a finer grid (cell side length of 0.001° for both latitude and longitude, corresponding to a square of ≈100m×100m). Having computed the result from both the coarse and fine granularity, we use an internal grid technique. According to this, for a query image, if the estimate based on the finer granularity falls within the borders of the estimated cell of the coarser granularity, then we consider the fine granularity trustworthy and apply similarity search inside the fine cell. Otherwise, we perform similarity search inside the coarser granularity cell, since coarser granularity language models are by default more trustworthy (due to the use of more data for building them).

Spatial Entropy: In order to adjust the original language model tag probabilities for each cell, we built a Gaussian weight function based on the values of the *spatial tag entropy*. The spatial entropy for each tag t_k is calculated based on its probabilities over all m cells of the grid.

$$e(t_k) = -\sum_{i=1}^{m} p(t_k|c_i) \log p(t_k|c_i) \quad (1)$$

MediaEval 2014 Workshop, October 16-17, 2014, Barcelona, Spain

We chose a Gaussian model because the tags with either too high or too low entropy values typically carry no geographic cues, and we would therefore need to suppress their influence on the location estimation process. Equation 2 presents the entropy-based cell estimation equation.

$$p(c_i|j) = \sum_{k=1}^{T} P(t_k|c_i) * \mathcal{N}(e(t_k), \mu, \sigma) \quad (2)$$

where $p(c_i|j)$ is the probability of cell c_i for image j, T is the number of tags for image j, $P(t_k|c_i)$ is the probability of tag k for cell c_i and e_k is the value of the entropy of tag k. $\mathcal{N}$ is the Gaussian function, and the parameters μ, σ are estimated using the distribution over the training set.

2.2 Visual-based location estimation

To build the visual location models, we relied on two features, SURF+VLAD and CS-LBP+VLAD, concatenating them in a single vector. In particular, we first calculated the interest points of each image, and then extracted both SURF and CS-LBP descriptors corresponding to them. The parameters used for CS-LBP [2] were $P = 8$, $R = 2$, and the number of bins $N = 16$. $L2$ normalization was applied for SURF and $L1$ for CS-LBP. For both features we used distinct multiple vocabularies learned on independent collections (four visual vocabularies with $k = 128$ centroids each) and applied dimensionality reduction using PCA separately to each VLAD vector, keeping more principal components for the SURF+VLAD vector to a factor of 3-1 (due to the correspondingly higher dimensionality of the non-reduced SURF+VLAD). The final VLAD vectors had a concatenated length of 1024 and were $L2$ normalized. For VLAD, we used the implementation of [5].

The main part of the model building included the training of linear SVM to separate the samples in a predefined number of spatial clusters and subclusters (we used 50 clusters and up to 50 subclusters corresponding to each cluster). The clusters/subclusters were created using k-means on the coordinates of the training images, while the number of subclusters was determined by the number of samples N assigned to each cluster ($\min(\text{round}(N/3000), 50)$).

Subcluster Selection: For each cluster a one-vs-rest approach was applied resulting in 50-d prediction score vectors, while for the subclusters a similar approach was used but only for intra-cluster samples (resulting in better performance than using both intra- and inter-cluster samples). The decision about the cluster membership for each sample was a combination of the estimation scores provided by the cluster prediction score vectors and also the scores corresponding to the best subcluster score in each cluster. Finally, priors (based on number of images per cluster/subcluster) were applied to the respective scores, since they were found to lead to some improvement.

Similarity Search: To achieve location estimations of finer granularity, we applied a similarity search step at the subcluster level. In particular, the query image was compared to 1000 samples from the selected subcluster (sampling was necessary for efficiency reasons), and the location of the most similar of those was returned. Similarity was computed based on a low-dimensional concept-based representation, using the 94 concepts of ImageCLEF 2012 (i.e. each image was represented by 94 prediction scores coming from a set of corresponding pre-trained concept models).

measure	**Run 1**	**Run 2**	**Run 3**	**Run 4**	**Run 5**
acc(10m)	*0.5*	0	0	0.03	0.31
acc(100m)	*5.85*	0	0.01	0.65	4.36
acc(1km)	*23.02*	0.03	0.16	21.87	22.24
acc(10km)	*39.92*	0.76	1.27	38.96	38.98
acc(100km)	*46.87*	2.18	3.00	46.13	46.13
acc(1000km)	*60.11*	17.35	17.72	59.87	59.87
median error	*230*	6232	6086	258	259

Table 1: Geotagging accuracy (%) and median error (km) for five ranges. Runs 1, 4 and 5 used text metadata, while Runs 2 and 3 relied on visual features.

3. RUNS AND RESULTS

As described above, we prepared three tag-based runs and two visual runs. The tag-based runs are the Run 1, using the language model, similarity search, internal grid and spatial entropy, Run 4, using the language model and the center of cells as estimated location, and Run 5, using the language model and similarity search. Run 2 was based on the Subcluster Selection step of subsection 2.2 using the center of the subcluster as location estimate. Run 3 was based on the combination of Subcluster Selection with Similarity Search (according to subsection 2.2). For Run 3, we used a subset of 25,500 images due to lack of time. For the rest of the runs we used the full test set of 510K images.

According to Table 1, the best performance in terms of both median error and accuracy in all ranges was attained by Run 1. Comparing Run 4 and 5, it can be seen that similarity search had considerable impact on the low range accuracy results. Also the combination of all features in Run 1 improves further the overall performance (reaching a 5.85% accuracy for the $< 100m$ range), but the median error is still quite high (230km), which means further improvements can be achieved. The visual runs yielded very poor results.

In the future, we plan to look into utilizing external data for training, in particular the Flickr 100M Creative Commons dataset and gazetteers. Furthermore, we will look into alternative ways to utilize visual information for geotagging.

Acknowledgements: This work is supported by the SocialSensor FP7 project, partially funded by the EC under contract number 287975.

4. REFERENCES

[1] J. Choi and al. The placing task: A large-scale geo-estimation challenge for social-media videos and images. In *Proceedings of the 3rd ACM GeoMM Workshop*, 2014.

[2] M. Heikkilä, M. Pietikäinen, and C. Schmid. Description of interest regions with local binary patterns. *Pattern Recognition*, 42(3):425–436, 2009.

[3] G. Kordopatis-Zilos, S. Papadopoulos, E. Spyromitros-Xioufis, A. L. Symeonidis, and Y. Kompatsiaris. CERTH at MediaEval Placing Task 2013. In *Proceedings of MediaEval 2013*.

[4] A. Popescu. CEA LIST's participation at MediaEval 2013 Placing Task. In *Proceedings of MediaEval 2013*.

[5] E. Spyromitros-Xioufis, S. Papadopoulos, Y. Kompatsiaris, G. Tsoumakas, and I. Vlahavas. A comprehensive study over VLAD and Product Quantization in large-scale image retrieval. *Trans. on Multimedia*, 16(6):1713–1728, 2014.

[6] O. Van Laere, S. Schockaert, and B. Dhoedt. Finding locations of Flickr resources using language models and similarity search. ICMR '11, pages 48:1–48:8, New York, NY, USA, 2011. ACM.