Placing Images with Refined Language Models and Similarity Search with PCA-reduced VGG Features

Giorgos Kordopatis-Zilos[1], Adrian Popescu[2], Symeon Papadopoulos[1], and Yiannis Kompatsiaris[1]

[1]Information Technologies Institute, CERTH, Greece. [georgekordopatis,papadop,ikom]@iti.gr

[2]CEA, LIST, 91190 Gif-sur-Yvette, France. adrian.popescu@cea.fr

ABSTRACT

We describe the participation of the CERTH/CEA-LIST team in the MediaEval 2016 Placing Task. We submitted five runs to the estimation-based sub-task: one based only on text by employing a Language Model-based approach with several refinements, one based on visual content, using geospatial clustering over the most visually similar images, and three based on a hybrid scheme exploiting both visual and textual cues from the multimedia items, trained on datasets of different size and origin. The best results were obtained by a hybrid approach trained with external training data and using two publicly available gazetteers.

1. INTRODUCTION

The goal of the task is to estimate the location of 1,497,464 photos and 29,934 videos using a set of ≈5M geotagged items and their metadata for training [1]. All submitted runs are built upon the scheme of our last year's participation [4], integrating several refinements. For the text-based runs, we focused on improving the pre-processing of metadata of the training set items and refining the feature selection method. For the visual-based runs, we built a more generic deep neural network model for enhanced visual image representation. For the hybrid scheme, we devised a score for selecting between the text and visual estimations based on the prediction confidence. To further improve performance, we built a model using all geotagged items of the YFCC dataset [8] (items uploaded by users in the test set are not included), and we leveraged structured information from open geographical resources such as Geonames[1] and OpenStreetMap[2].

2. APPROACH DESCRIPTION

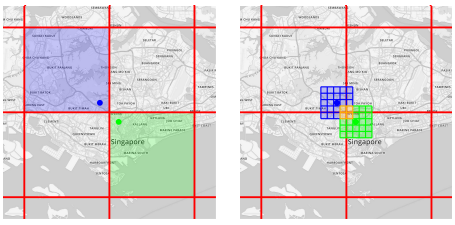
2.1 Text-based location estimation

In the first step, the tags and titles of the training set items were pre-processed. We applied URL decoding[3], lowercase transformation, tokenization and removed accents to generate a set of terms for every item. The multi-word tags were further split into their individual components, which were also included in the item's term set. Finally, symbols and punctuations in the terms were removed, and terms consisting of numerics or less than three characters were discarded.

The core of our approach is a probabilistic Language Model (LM) [5] built from the terms of the training set items. The earth surface was divided into (nearly) rectangular cells of size $0.01^\circ \times 0.01^\circ$ latitude/longitude, and the term-cell probabilities were computed based on the user count of each term in each cell. The most likely cell (*mlc*) of a query is derived from the summation of the respective term-cell probabilities. The estimated location of the query items with no textual information is the centre of the cell with the most users.

For feature selection, we used a refined version of the *locality* metric [4]: in our last participation, we computed *locality* based on the neighbor users that used the same term in the same cell. To this end, we utilized a coarse grid ($0.1^\circ \times 0.1^\circ$) for the calculation, based on which the neighbor users were assigned to a unique cell, as depicted in Figure 1(a). In that setting, it was possible that a pair of users were not assigned to the same cell even if the geodesic distance of their items was small. To tackle this issue, we now used a grid of $0.01^\circ \times 0.01^\circ$ and modified the assignment of the users to multiple cells: instead of assigning a user to a unique cell, we assigned a user to an entire neighborhood, as illustrated in Figure 1(b). The area highlighted in orange corresponds to the cells where both users were assigned. The terms with non-negative *locality* score form the selected term set T.

The contribution of each term was then weighted based on its *locality* and *spatial entropy* scores. *Spatial entropy* is a Gaussian weight function based on the term-cell entropy of the term [2]. The two measures are combined to generate a weight value for every term in T.



(a) (b)

Figure 1: Locality examples: (a) initial, (b) refined.

[1]http://www.geonames.org/

[2]https://www.openstreetmap.org/

[3]This was necessary because text in different languages was URL encoded in the released dataset.

MediaEval 2016 Workshop, Oct. 20-21, 2016, Hilversum, Netherlands.

To ensure more robust performance in fine granularity, we built an additional LM using a finer grid ($0.001^\circ \times 0.001^\circ$). Having computed the *mlc* for both coarse and fine granularities, we selected the most appropriate estimation: this is the *mlc* of the finer grid if it falls within the borders of the coarse grid, otherwise it is the *mlc* of the coarse one. Finally, we employed similarity search as in [9] to derive the location estimates from the the $k_t = 5$ most textually similar images inside the selected *mlc*, computing textual similarity using the Jaccard similarity between the corresponding term sets. Error case analysis of the text method is presented in [3].

2.2 Visual-based location estimation

The employed method is a refined version of the one employed in last year's participation [4]. The main objectives have been (1) to ensure that the visual features are generic and transferable from a training set independent of YFCC to the subset of the collection used for the task, and (2) to provide a compact representation of the features in order to scale up the visual search process. To meet the first objective, the VGG architecture [7] was fine-tuned with over 5000 diversified man-made and natural POIs, represented by over 7 million images. These were downloaded from Flickr using queries with (1) the POI name and a radius of 5km around its coordinates and (2) the POI name and the associated city name. Following the conclusions of [6] regarding the uselessness of manual annotation for POI representation, there was no manual validation of the training set. To meet the second objective, we used the same procedure as last year and compressed the initial features (VGG fc7, 4096 dimensions) to 128 dimensions using PCA. The PCA matrix was learned on a subset of 250,000 images of the training set.

Having calculated these similarities, we retrieved the top k_v most visually similar images (in our runs we set $k_v = 20$) and applied a simple spatial clustering scheme based on their geographical distance. We defined a confidence metric for our visual approach based on the size of the largest cluster:

$$\text{conf}_v(i) = \max((n(i) - n_t)/(k_v - n_t), 0) \quad (1)$$

where $n(i)$ is the number of neighbors in the largest cluster for query image i, n_t is the configuration parameter that determines the "strictness" of the confidence score. The confidence score gets values in the range [0,1]. We empirically set $n_t = 5$. Our visual approach is not designed for video analysis, thus all videos were placed in the centre of London, which is the densest geotagged region in the world.

2.3 Hybrid location estimation

The hybrid approach comprises a set of rules that determine the source of estimation between the text and visual approaches. First, for query images, for which no estimation could be produced by the text-based approach, the location was estimated based on the visual approach. Otherwise, in case the visual estimation fell inside the borders of the *mlc* calculated by the text-based approach, the visual estimation was selected. If not, the estimation was determined by comparing the confidence scores of the two approaches.

$$G_h(i) = \begin{cases} G_v(i) & if\ \text{conf}_t(i) \leq \text{conf}_v(i) \\ G_t(i) & otherwise \end{cases} \quad (2)$$

where G_h, G_t and G_v are the estimated locations for query item i of the hybrid, textual and visual approach, respectively, conf_t is the confidence score of the text-based estimation and is defined in [4], and conf_v is the confidence score of the visual-based estimation (Equation 1).

measure	RUN-1	RUN-2	RUN-3	RUN-4	RUN-5	RUN-E
P@10m	0.59	0.08	0.56	0.7	*0.72*	4.78
P@100m	6.42	1.84	6.58	7.96	*8.27*	8.41
P@1km	24.55	5.62	25.03	27.82	*28.54*	13.67
P@10km	43.32	8.16	43.73	*46.52*	46.45	16.6
P@100km	51.26	10.21	51.69	*53.96*	53.5	18.83
m. error	65	5031	56	*24*	27	3432

(a) Images

measure	RUN-1	RUN-2	RUN-3	RUN-4	RUN-5
P@10m	0.55	0.0	0.55	0.69	*0.71*
P@100m	6.86	0.06	6.86	7.89	*8.19*
P@1km	22.73	0.5	22.73	25.53	*26.16*
P@10km	40.6	2.48	40.6	*43.89*	43.62
P@100km	48.24	4.97	48.24	*51.2*	50.44
m. error	161	6211	161	*68*	85

(b) Videos

Table 1: Geotagging precision (%) and median error (km) for five runs (+RUN-E for images).

3. RUNS AND RESULTS

The submitted runs include one text-based (RUN-1), one visual-based (RUN-2) and three hybrid runs (RUN-3, RUN-4, RUN-5). For the first three runs, the system was trained on the set released by the organizers. In RUN-4 and RUN-5, the training set consisted of all YFCC items excluding those contributed by users appearing in the test set. Also, we report the results of an external run (RUN-E), based on the visual approach but using the full geotagged subset of YFCC. The results for RUN-E show that adding more training data significantly improves visual geolocation, especially for short ranges (10m and 100m), where this run outperforms even the best hybrid run.

To explore the impact of external data sources, in RUN-5, we further leveraged structured data from Geonames and OpenStreetMap. In particular, we used the geotagged entries of the two sources as additional training items for building the text-based LM: from Geonames we used a list of city names along with their alternative names, while from OpenStreetMap a list of *nodes* (points of interest) provided they were associated with an address. Since training items need to be associated with a contributor, we considered Geonames and OpenStreetMap as the two contributing users.

According to Table 1, the best performance at fine granularities ($\leq$1km) was attained by RUN-5 for both images and videos. RUN-4 reported the best results in terms of median distance error and precision at coarse granularities (>1km). Comparing the two runs, one may conclude that leveraging structured geographic information improves geolocation precision in short ranges (reaching 8.27% and 28.54% in $P@100m$ and $P@1km$ respectively), with a minor increase in median error. Moreover, the combination of visual and textual features (RUN-3) improved the overall performance of the system in case of images, but had no effect on video geotagging (since no visual information was used from videos).

4. ACKNOWLEDGMENTS

This work is supported by the REVEAL and USEMP projects, partially funded by the European Commission under contract numbers 610928 and 611596 respectively.

5. REFERENCES

[1] J. Choi, C. Hauff, O. Van Laere, and B. Thomee. The placing task at mediaeval 2016. In *MediaEval 2016 Placing Task*, 2016.

[2] G. Kordopatis-Zilos, S. Papadopoulos, and Y. Kompatsiaris. Geotagging social media content with a refined language modelling approach. In *PAISI 2015*, pages 21–40, 2015.

[3] G. Kordopatis-Zilos, S. Papadopoulos, and Y. Kompatsiaris. In-depth exploration of geotagging performance using sampling strategies on yfcc100m. In *Proceedings of the MMCommons 2016*. ACM, 2016.

[4] G. Kordopatis-Zilos, A. Popescu, S. Papadopoulos, and Y. Kompatsiaris. CERTH/CEA LIST at MediaEval placing task 2015. In *MediaEval 2015 Placing Task*, 2015.

[5] A. Popescu. Cea list's participation at mediaeval 2013 placing task. In *MediaEval 2013 Placing Task*, 2013.

[6] A. Popescu, E. Gadeski, and H. L. Borgne. Scalable domain adaptation of convolutional neural networks. *CoRR*, abs/1512.02013, 2015.

[7] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[8] B. Thomee et al. The new data and new challenges in multimedia research. *CoRR*, abs/1503.01817, 2015.

[9] 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.