CERTH/CEA LIST at MediaEval Placing Task 2015

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 Placing Task of MediaEval 2015. We submitted five runs in total to the Locale-based placing sub-task, providing the estimated locations for the test set released by the organisers. Out of five runs, two are based solely on textual information, using feature selection and weighting methods over an existing language model-based approach. One is based on visual content, using geo-spatial clustering over the most visually similar images, and two runs are based on hybrid approaches, using both visual and textual cues from the images. The best results (median error 22km, 27.5% at 1km) were obtained when both visual and textual features are combined, using external data for training.

1. INTRODUCTION

The goal of the task is to produce location estimates for a set of 931,573 photos and 18,316 videos using a set of 4.7M geotagged items and their metadata for training [1]. For the tag-based runs, we built upon the scheme of our 2014 participation [4] and a number of recent extensions on it [5], focusing on improved feature selection and feature weighting. For the visual-based location estimation, we use a geospatial clustering scheme of the most visually similar images for every query image. A hybrid scheme is composed by the combination of the textual and visual approaches. To further improve the model, we constructed it using all geotagged metadata from the YFCC dataset [9], after removing all images from the users contained in the test set.

2. APPROACH DESCRIPTION

2.1 Tag-based location estimation

According to our last year's approach [4] (baseline), the earth surface is divided in (nearly) rectangular cells of size 0.01° latitude/longitude (approximately $1km^2$ size near the equator). We construct a Language Model (LM) [6], i.e. a tag-cell probability map, by processing the tags and titles of the training set images. The tag-cell probabilities are computed based on the user count of each tag in each cell. Then, the Most Likely Cell (MLC) of a query (test) image is derived from the summation of the respective tag-cell probabilities. The contribution of each tag is weighted based on its spatial entropy through a Gaussian weight function [5], which is referred to as Spatial Entropy (SE) function.

To ensure more reliable prediction in finer granularities, we built an additional LM using a finer grid (cell side length of 0.001°). Having computed the MLCs for both the coarse and fine granularity, we apply an Internal Grid technique [4] as a means to produce more accurate, yet equally reliable location estimates. This is achieved by first selecting the most appropriate granularity (the finer grid cell if considered reliable, otherwise the coarser grid cell), and then producing the location estimate based on the center-of-gravity of the k most textually similar images inside the selected MLC ($k = 5$), by employing Similarity Search as in [10]. The textual similarity is computed using the Jaccard similarity of the corresponding sets of tags.

2.1.1 Feature Selection

To increase the robustness of the model and reduce its size, feature selection was performed based on two measures: the *accuracy* and the *locality* of the tags.

Accuracy is computed using the cross-validation scheme proposed in [5]. The training set is partitioned into p folds (here, $p = 10$). Subsequently, one partition at a time is withheld, and the rest $p-1$ partitions are used to build the LM. Having built the LM, the location of every item of the withheld partition is estimated. The accuracy of a tag is computed based on Equation 1.

$$\text{tgeo}(t) = \frac{N_r}{N_t}, \quad (1)$$

where $\text{tgeo}(t)$ is the accuracy score of each tag t, N_r is the total number of correctly geotagged items tagged with t and N_t is the total number of items tagged with t. The tags with non-zero accuracy score form a tag set denoted as T_a.

Locality captures the *spatial awareness* of tags. For every individual tag, the locality score is calculated based on the tag frequency and the neighbor users that have used it in the various cells. Every time that a user uses a given tag, he/she is assigned to the respective location cell. As a result, each cell has a set of users that have been assigned to it. All users assigned to the same cell are considered neighbors (for that particular cell). Then, the locality score can be computed by Equation 2.

$$\text{loc}(t) = N_t * \frac{\sum_{c \in C} \sum_{u \in U_{t,c}} |\{u' | u' \in U_{t,c}, u' \neq u\}|}{N_t^2}, \quad (2)$$

where $\text{loc}(t)$ is the locality score of tag t, N_t is the total occurrences of t, C denotes all cells and $U_{t,c}$ denotes the set

MediaEval 2015 Workshop, Sept. 14-15, 2015, Wurzen, Germany

of users that used tag t inside cell c. Since all users in $U_{t,c}$ are neighbors, Equation 2 can be simplified to:

$$\text{loc}(t) = \frac{\sum_{c \in C} \sum_{u \in U_{t,c}} |U_{t,c}| - 1}{N_t} = \frac{\sum_{c \in C} |U_{t,c}|(|U_{t,c}| - 1)}{N_t}$$

The tags with non-zero locality score are forming a tag set denoted as T_l. The final tag set T used by the approach is the intersection of the two tag sets: $T = T_a \cap T_l$.

2.1.2 Feature Weighting

Since the locality metric is sensitive to tag frequency, we consider it as an inappropriate for directly weighting tags. Alternatively, having computed the locality scores for every tag in T, we sort them based on their scores and calculate their weights using their position in the distribution.

$$w_l = \frac{|T| - (j-1)}{|T|} \tag{3}$$

where, w_l is the weight value of the tag t on the j-th position in the distribution and $|T|$ is the total number of tags contained in T. This weighting approach returns values in the range $(0, 1]$. To combine the two weighting functions, we normalize the values of the Spatial Entropy weighting function, denoted with w_{se}, and use Equation 4 to compute the final weights.

$$w = \omega * w_{se} + (1 - \omega) * w_l \tag{4}$$

The value of ω was set to 0.2 through empirical assessment on a sample of 10K images.

2.1.3 Confidence

To evaluate the confidence of the estimation of each query image, we use the confidence measure of Equation 5.

$$\text{conf}(i) = \frac{\sum_{c \in C} \{p(c|i) | \text{dist}(c, \text{mlc}) < l\}}{\sum_{c \in C} p(c|i)}, \tag{5}$$

where $\text{conf}(i)$ is the confidence for query image i, $p(c|i)$ is the cell probability of cell c for image i, $\text{dist}(c_1, c_2)$ is the distance between the centers of cells c_1 and c_2 and mlc stands for the Most Likely Cell.

2.2 Visual-based location estimation

We compute visual-based location estimations with CNN features adapted for the tourist domain using approximately 1000 Points Of Interest (POIs) for training, with approximately 1200 images per POI, that were fed directly to Caffe [3]. These features were computed by fine-tuning the VGG model proposed at ILSVRC 2014 [7]. The outputs of the $fc7$ layer (4096 dimensions) were compressed to 128 using a PCA matrix learned from a subset of $250,000$ images of the CNN training set and used to compute image similarities. CNN features were selected after a favorable comparison against compact VLAD features of similar size [8] and with SURF features of significantly larger size [2]. Having calculated these similarities, we retrieve the top k most visually similar images and use their location to perform the estimate. In the visual only run (`RUN-2`), $k = 20$ and we apply a simple incremental spatial clustering scheme, in which if the j-th image (out of the k most similar) is within 1km from the closest one of the previous $j-1$ images, it is assigned to its cluster, otherwise it forms its own cluster. In the end, the largest cluster (or the first in case of equal size) is selected and its centroid is used as the location estimate.

measure	RUN-1	RUN-2	RUN-3	RUN-4	RUN-5
acc(1m)	0.15	0.01	0.15	0.16	*0.16*
acc(10m)	0.61	0.08	0.62	0.75	*0.76*
acc(100m)	6.40	1.76	6.52	7.73	*7.83*
acc(1km)	24.33	5.19	24.61	27.30	*27.54*
acc(10km)	43.07	7.43	43.41	46.48	*46.77*
acc(100km)	51.08	9.07	51.45	54.02	*54.33*
acc(1000km)	63.81	23.98	64.18	65.81	*66.06*
m. error(km)	69	5663	61	24	*22*

Table 1: Geotagging accuracy (%) for different ranges and median error (km). `RUN-1` and `RUN-4` used only text, `RUN-2` relied on visual features, and `RUN-3` and `RUN-5` used both visual and text features.

2.3 Hybrid location estimation

For the hybrid approach, we build an LM using the scheme described in Section 2.1. To achieve further improvement in finer granularities wuith the use of the Similarity Search approach, the similarity between two images derives from the combination of the visual and textual similarities. To this end, we normalize the visual similarities to the range [0, 1]. The final similarity for a pair of images is computed as the arithmetic mean of the two similarities. We then retrieve the top $k = 5$ most similar images, within the borders specified by the Internal Grid technique [5], and we use their center-of-gravity as the final location estimate.

For those test images, where no estimate can be produced based on the LM or confidence is lower than 0.02 (which together amount to approximately 10% of the test set), we use the visual approach to produce the estimate.

3. RUNS AND RESULTS

We prepared two tag-based (`RUN-1`, `RUN-4`), one visual (`RUN-2`) and two hybrid runs (`RUN-3`, `RUN-5`). Runs 1-3 used the training set released by the organisers; in Runs 4-5, the entire YFCC dataset was used, excluding all images from users that appeared in the test set. All runs contained estimates for the full test set (949,889 items).

According to Table 1, the best performance in terms of both median error and accuracy in all ranges was attained by `RUN-5`. Comparing the corresponding runs with different training sets, one may conclude that the use of an extended training set (that does not contain user-specific information) had considerable impact on the accuracy results across all ranges. Furthermore, the combination of features (visual and textual) in `RUN-5` further improved the overall performance (reaching a 7.83% accuracy for the <100m range) and minimizing median error (22km). The visual-only run (`RUN-2`) obtained remarkable results (reaching a 5.19% accuracy for the <1km range).

In the future, we plan to look deeper into different weighting schemes trying to achieve further improvements. Moreover, we plan to develop more sophisticated clustering models for the visual-only runs.

4. ACKNOWLEDGEMENTS

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 2015. In *MediaEval 2014 Placing Task*, 2015.

[2] J. Choi and X. Li. The 2014 ICSI/TU delft location estimation system. In *Working Notes Proceedings of the MediaEval 2014 Workshop, Barcelona, Catalunya, Spain, October 16-17, 2014.*, 2014.

[3] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[4] G. Kordopatis-Zilos, G. Orfanidis, S. Papadopoulos, and Y. Kompatsiaris. Socialsensor at mediaeval placing task 2014. In *MediaEval 2014 Placing Task*, 2014.

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

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

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

[8] E. Spyromitros-Xioufis, S. Papadopoulos, I. Kompatsiaris, G. Tsoumakas, and I. Vlahavas. A comprehensive study over vlad and product quantization in large-scale image retrieval. *IEEE Transactions on Multimedia*, 2014.

[9] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L. Li. The new data and new challenges in multimedia research. *CoRR*, abs/1503.01817, 2015.

[10] 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.