SocialSensor: Finding Diverse Images at MediaEval 2014

Eleftherios Spyromitros-Xioufis[1,2], Symeon Papadopoulos[1],
Yiannis Kompatsiaris[1], Ioannis Vlahavas[2]
[1]Information Technologies Institute, CERTH, Thessaloniki, Greece
[2]Aristotle University of Thessaloniki, Thessaloniki, Greece
{espyromi,papadop,ikom}@iti.gr,vlahavas@csd.auth.gr

ABSTRACT

This paper describes the participation of the SosialSensor team in the Retrieving Diverse Social Images Task of MediaEval 2014. All our entries are produced by a different instantiation (set of features, parameter configuration) of the same diversification algorithm that optimizes a joint relevance-diversity criterion. All our runs are automated and use only resources given by the task organizers. Our best results in terms of the official ranking metric (F1@20 $\approx$ 0.59) came by the runs that combine visual and textual information, followed by the visual-only run.

1. INTRODUCTION

The Retrieving Diverse Social Images task of MediaEval 2014 deals with the problem of result diversification in social photo retrieval. Participants are given a list of images retrieved from Flickr in response to a query for a specific location e.g., "Eiffel Tower" and are asked to return a refined short-list that contains images which are at the same time relevant and diverse (see [4] for more details).

To deal with this problem we build upon the approach that we developed for the visual-only run of previous year's task [3], termed Relevance and Diversity (*ReDiv*) [1]. For this year's task, the *ReDiv* approach was refined and used to produce all our runs. Section 2 describes the *ReDiv* approach and Section 3 details the different instantiations of the approach used to produce each of the submitted runs. Finally, in Section 4 we briefly summarize and discuss our experimental results.

2. OVERVIEW OF OUR APPROACH

Let $I = \{im_1, \ldots, im_N\}$ be a set of images that have been retrieved from Flickr in response to a query q for a specific location. The goal of the diversification algorithm is to select a K-sized subset of images from I that are as relevant (to the query location) and as diverse (among each other) as possible. *ReDiv* formalizes this verbal description into the following optimization problem: $\arg\max_{S \subset I, |S|=k} U(S) = wR(S|q) + (1-w)D(S)$ where we want to identify the set S that has maximum utility $U(S)$, defined as a weighted combination of the relevance $R(S|q)$ and the diversity $D(S)$ of S. A similar formulation of the problem was used in [2]. In *ReDiv*, however, we use different definitions for $R(S|q)$ and $D(S)$ that we found more suitable for this task. These changes are described below.

Relevance: In [2], the authors define relevance as $R(S|q) = \sum_{im_i \in S} R(im_i|q)$, where $R(im_i|q) = 1 - d(im_i, im_q)$ and $d(im_i, im_q)$ denotes the dissimilarity between image im_i and the image that depicts the query location im_q. We observed that, in the context of this task, this definition can be problematic (especially when using only visual information) as there are several images that are visually dissimilar to the reference Wikipedia images of the location but are still considered relevant to the location e.g., inside views. Also, in many cases, images that are similar to the reference images are considered irrelevant to the location due to people being part of the image. Motivated by these shortcomings, we developed a more principled way for computing the relevance of each image to the query location. This is achieved by building a (distinct for each location) supervised classification model that is trained to distinguish relevant from irrelevant images. More specifically, we use the probabilistic output of this model in place of $R(im_i|q)$. To train this model, we use the relevance ground truth provided by the task organizers for the development set locations and use relevant/irrelevant images of other locations as positive/negative examples. Additionally, the Wikipedia images of each location are used as positive (relevant) examples and are assigned a large weight.

Diversity: Assuming a ranking $im_{r1}, \ldots, im_{rK}$ of the images in S, the authors in [2] define diversity as $D(S) = \sum_{i=1}^{K} \frac{1}{i} \sum_{j=1}^{i} d(im_{ri}, im_{rj})$, where $d(im_{ri}, im_{rj})$ is the dissimilarity between the images ranked at positions i and j. Thus, high diversity scores are given to image sets with a high average dissimilarity. We notice that this definition of diversity can assign relatively high diversity scores to image sets containing images with highly similar image pairs (probably belonging to the same cluster) and this results in a direct negative impact on the CR@20 measure and consequently to F1@20. Therefore, we adopt a more strict definition of diversity where the diversity of a set S is defined as the dissimilarity between the most similar pair of images in S: $D(S) = \min_{im_i, im_j \in S, i \neq j} d(im_i, im_j)$.

Optimization: An exact optimization of U comes with a high complexity as it would require computing the utility of all $\frac{N!}{K!(N-K)!}$ K-subsets of I. With $N \approx 300$ and $K = 20$ (in order to maximize F1@20) the computational cost of exact optimization becomes prohibitive. We therefore adopt the greedy, approximate optimization approach that was used in [2] with appropriate changes to reflect our new defini-

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

tions for relevance and diversity. This algorithm starts with an empty set S and sequentially expands it by adding at each step $J = 1, \ldots, K$ the image im^* that scores highest (among the unselected images), to the following criterion: $U(im^*) = wR(im^*) + (1-w) \min_{im_j \in S^{J-1}} d(im^*, im_j)$, where S^{J-1} represents S at step $J-1$. We also developed a less greedy version of this algorithm that in each step J keeps M highest scoring image subsets. Since the two algorithms coincide for $M = 1$ we used the less greedy version and tuned the M parameter.

Experimental Protocol: Depending on the type of the run (visual/textual/both) a variety of different (vector) representations of the images could be utilized for building the relevance detection models and computing pairwise image similarities in *ReDiv* (note that the algorithm allows using different representations for relevance and diversity). To reduce the complexity of the experiments, we first evaluated each representation in terms of its relevance detection ability and then evaluated combinations of only the top performing representations in the *ReDiv* algorithm. To judge the effectiveness of each representation in terms of relevance detection and to perform model selection we used a variant of leave-one(-location)-out cross-validation and measured performance via area under ROC (AUC). As classification algorithm we used L2-regularized logistic regression, as it led to near optimal results for a variety of representations in preliminary experiments.

Given an instantiation of the *ReDiv* approach (a specific combination of relevance detection model and diversity features) we performed leave-one(-location)-out cross-validation and evaluated the performance of each instantiation in terms of F1@20. The process was repeated for different values of w in the $[0, 1]$ range. We also noticed that using only the $n < N$ most relevant images (according to the relevance detection model) leads to improved performance. We therefore also performed a coarse search over the domain of $N = \{1, 2, \ldots, 300\}$ in order to find an optimal value. Finally, we tested the values $\{1, 2, 3, 4, 5\}$ for the M parameter.

3. INSTANTIATIONS

3.1 Visual (Run 1)

For this run we experimented with all the precomputed visual features made available by the task organizers and also extracted our own visual features. The best results were obtained using VLAD+CSURF [5] vectors (computed from a 128-dimensional visual vocabulary and projected to 128 dimensions with PCA and whitening) for both the relevance and the diversity component. Cosine distance was used as dissimilarity measure. The parameters used to produce the 1st run are: $w = 0.4$, $n = 75$, $M = 3$.

3.2 Textual (Run 2)

A bag-of-words representation with the 20K/7.5K most frequent words was used for the relevance/diversity component. Wikipedia images were represented using a parsed version of the corresponding Wikipedia page and Flickr images by a concatenation of the words in their titles (×3), description (×2) and tags (×1). Again, cosine distance was used as dissimilarity measure. The parameters used to produce the 2nd run are: $w = 0.95$, $n = 110$, $M = 1$.

Table 1: Performance of the submitted runs.

	Development Set			Test Set (official)		
Run	*P@20*	*CR@20*	***F1@20***	*P@20*	*CR@20*	***F1@20***
1	0.815	0.497	0.609	0.775	0.460	0.569
2	0.863	0.468	0.599	0.832	0.407	0.538
3	0.855	0.521	0.642	0.817	0.473	0.593
5	0.857	0.527	**0.647**	0.815	0.475	**0.594**

3.3 Visual+Textual (Runs 3 & 5)

An early fusion of the visual and textual features described above was used for the relevance component and the visual features described above were used for the diversity component. The parameters used to produce the 3rd run are: $w = 0.75$, $n = 90$, $M = 5$. The 5th run differs from the 3rd run only in the value used for n $(= 95)$.

4. RESULTS AND DISCUSSION

Table 1 shows the performance of the submitted runs in the test locations as well as their estimated performance using our internal evaluation procedure. We observe that in all cases we overestimated the final performance, nevertheless by a small and approximately constant (among different runs) margin. Most importantly, the relative ranking of the runs is the same, suggesting that our model selection procedure was appropriate. The best performance was obtained by the two variations of our visual+textual run, followed by the visual-only run. Despite the comparatively lower performance obtained using only textual features, we see that a significant performance boost was possible by combining them with visual features for relevance detection.

In the future we plan to develop a more principled approach for combining different types of features in the *ReDiv* algorithm, especially for the diversity component.

5. ACKNOWLEDGEMENTS

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

6. REFERENCES

[1] D. Corney, C. Martin, A. Göker, E. Spyromitros-Xioufis, S. Papadopoulos, Y. Kompatsiaris, L. Aiello, and B. Thomee. Socialsensor: Finding diverse images at mediaeval 2013. In *MediaEval*, 2013.

[2] T. Deselaers, T. Gass, P. Dreuw, and H. Ney. Jointly optimising relevance and diversity in image retrieval. In *ACM CIVR '09*, New York, USA, 2009.

[3] B. Ionescu, M. Menéndez, H. Müller, and A. Popescu. Retrieving diverse social images at MediaEval 2013: Objectives, dataset and evaluation. In *MediaEval*, 2013.

[4] B. Ionescu, A. Popescu, M. Lupu, A. Gînsca, and H. Müller. Retrieving diverse social images at MediaEval 2014: Challenge, dataset and evaluation. In *MediaEval*, 2014.

[5] 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.