USEMP: Finding Diverse Images at MediaEval 2015

E. Spyromitros-Xioufis[1], A. Popescu[2], S. Papadopoulos[1], I. Kompatsiaris[1]
[1]CERTH-ITI, Thermi-Thessaloniki, Greece, {espyromi,papadop,ikom}@iti.gr
[2]CEA, LIST, 91190 Gif-sur-Yvette, France, adrian.popescu@cea.fr

ABSTRACT

We describe the participation of the USEMP team in the Retrieving Diverse Social Images Task of MediaEval 2015. Our runs are produced based on a supervised diversification method that jointly optimizes relevance and diversity. All runs are automated and use only resources given by the task organizers. Our best results in terms of the official ranking metric on the one-topic part of the test set came by the runs that combine visual and textual information while the textual-only run performed better on the multi-topic part.

1. INTRODUCTION

The Retrieving Diverse Social Images task of MediaEval 2015 [7] deals with the problem of result diversification in social image retrieval. This year there are two notable differences with previous editions: (1) a larger development set is available and (2) "multi-concept" queries which are mainly related to events rather than specific places or landmarks were introduced.

We deal with the task using *supervised Maximal Marginal Relevance* (sMMR) [10], a refined version of the supervised diversification method that we developed in [12]. sMMR and earlier versions of the method are discussed in Section 2. Section 3 gives further details about our methodology and describes modifications compared to [10]. Section 4 provides descriptions of the employed features and Section 5 describes the submitted runs. Finally, Section 6 presents and discusses the experimental results.

2. PREVIOUS WORK

In the 2013 edition of the task, the SocialSensor team developed a supervised diversification method [3] and applied it for producing the visual-only run that achieved the best performance among runs of this type. Similarly to previous diversification methods [2, 5], that method greedily optimized a utility function that jointly accounts for relevance and diversity. The main difference compared to earlier approaches was the replacement of the unsupervised definition of relevance with a task-specific definition that is learned directly from the ground truth. More specifically, instead of computing an image's relevance score by measuring its similarity to a reference image (e.g., the Wikipedia image of a query topic), the approach exploited the available relevance annotations (for the images of the development set) to train a classifier which was then used to assign relevance scores to images of unseen queries. Although that classifier was not adapted to any particular query, it could accurately estimate relevance by capturing its task-specific notion.

In 2014, the SocialSensor team refined their approach [12], topping the scoreboard in several categories: best visual-only run, third textual run and best visual+textual run that was also ranked second overall, slightly surpassed by a run that used specialized filters (face/blur detectors) and user credibility information [4]. The main addition compared to [3] was that a different relevance classifier was trained for each query, using the query's Wikipedia images as additional positive examples. These examples were assigned a larger weight to have increased influence on the learned model. Thus the query-specific notion of relevance was captured in addition to the task-specific notion captured in [3].

The approach was evaluated in [10] in more detail identifying a link between relevance detection accuracy and diversification performance. Furthermore, a multimodal ensemble classifier, called Multi-Modal Stacking (MMS), was proposed for combining different types of features for relevance detection in a principled manner. Due to the addition of this multimodal scheme and to the use of state-of-the-art convolutional neural network features for relevance detection, [10] managed to achieve a 5.7% relative increase over the best result obtained in the 2014 [4] edition of the task.

3. METHOD

Given the effectiveness of sMMR in previous editions of the task, we opted for applying it in this year's task as well. In particular, we applied the sMMR_{aq} variant: the relevance detection model for each query was trained using relevant and irrelevant examples from other queries, combined with representative examples of the query itself (in the form of either the corresponding Wikipedia images or the Wikipedia page). For multi-topic queries, which were not accompanied by representative Wikipedia images, the visual relevance models were trained using only examples from other queries (sMMR_a variant).

To further improve the relevance detection models compared to [10], we performed careful tuning of two parameters: a) the number n_o of examples from other queries employed by each model and b) the ratio $r = \frac{n_e}{n_o}$ defined as the number of examples of this query[1] divided by the

[1] n_e is modified by repeating each representative example $\frac{n_e}{n*_e}$ times, where $n*_e$ is the actual number of representative examples.

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

Table 1: Estimated (one-topic/multi-topic) and final performance of the submitted runs.

Run	Development Set				Test Set (One-Topic)			Test Set (Multi-Topic)			Test Set (overall)		
	AUC	P@20	CR@20	F1@20	P@20	CR@20	F1@20	P@20	CR@20	F1@20	P@20	CR@20	F1@20
1	0.821/0.773	0.860/0.763	0.489/0.468	0.616/0.573	0.805	0.478	0.587	0.598	0.453	0.499	0.701	0.465	0. 542
2	0.688	0.836	0.459	0.586	0.824	0.455	0.569	0.734	0.442	**0.530**	0.779	0.448	**0.549**
3	0.857/0.816	0.893/0.840	0.515/0.489	0.646/0.609	0.833	0.504	**0.618**	0.617	0.408	0.471	0.724	0.456	0.544
5	0.857/0.816	0.877/0.823	0.526/0.499	0.650/0.613	0.802	0.509	0.611	0.608	0.417	0.474	0.704	0.462	0.542

number of examples from other queries, by performing a grid search over the values $n_o = \{1K, 5K, 10K, 20K\}$ and $r = \{0.0, 0.1, \ldots, 1.0\}$. Model selection was based on area under ROC (AUC), computed using a modified leave-one(-query)-out cross-validation procedure. For each query of the development set, n_o examples were randomly selected from other queries and combined (if possible) with n_e representative (Wikipedia) examples of that query to build a model that was evaluated on the remaining (Flickr) examples of that query. The per-query AUC scores were then averaged to obtain a single estimate. As in [10], an L2-regularized Logistic Regression classifier was used [6] with appropriate tuning of the c parameter. Besides the parameters of the relevance model, we also tuned the w and N parameters of the sMMR approach, so as to maximize F1@20 on the development set, as done in [12].

4. FEATURES

VLAD: VLAD+CSURF [11] vectors were computed from a 128-dimensional visual vocabulary and projected to 128 dimensions with PCA and whitening. Both the visual vocabulary and the PCA projection matrix are learned using the images of the development set.

CNN: Convolutional neural network features were adapted for the tourism use case using ≈1,000 Points Of Interest (POIs) instead of ImageNet classes. These features were computed by fine-tuning the VGG model proposed by [9]. Approximately 1,200 images were collected for each POI and fed directly to Caffe [8] for training. This change of training classes was inspired by recent domain adaptation work presented in [1] which shows that the feature transfer is more efficient when the training classes are conceptually close to the target dataset. The features are constituted by the outputs of the *fc7* layer and include $4,096$ dimensions.

BOW: To generate textual features, we transformed each query and each Flickr image into a text document. For queries, we used a parsed version of the corresponding Wikipedia page and for Flickr images we used a concatenation of the words in their titles, descriptions and tags. Bag-of-words features (BOW) were then computed for each document using all terms that appear at least twice in the collection to form the dictionary, and word frequencies as term weights. This led to an 80K-dimensional representation.

META: The following one-dimensional features were also computed from textual metadata and used as additional features in the meta input space of the MMS algorithm: distance from POI (only for one-topic queries) and Flickr rank.

5. RUNS

Run 1: *CNN* features were used for relevance and *VLAD* features for diversity. n_o was set to 20K in both instantiations and r was set to 0.5 for the one-topic instantiation. $\{w = 0.55, N = 170\}$ and $\{w = 0.00, N = 120\}$ were used for the one-topic and the multi-topic instantiation respectively.

Run 2: *BOW* features were used for both relevance and diversity. The following parameters were used: $n_o = 20K$, $r = 0.4$, $N = 80$ and $w = 0.8$.

Run 3: A different instantiation was used for each part of the collection. MMS was used to combine the outputs of relevance detection models built using *CNN* and *BOW* features with one-dimensional *META* features, and *VLAD* features were used for diversity in both instantiations. $\{w = 0.50, N = 220\}$ and $\{w = 0.55, N = 170\}$ were used for the one-topic and the multi-topic instantiation respectively.

Run 5: This is a variation of run 3 where we use the same relevance detection models but also tune the M parameter of the approach in addition to N and w, as done in [12]. This resulted into setting $\{w = 0.4, N = 210, M = 2\}$ for the one-topic instantiation and $\{w = 0.4, N = 300, M = 5\}$ for the mutli-topic instantiation.

6. RESULTS AND DISCUSSION

Table 1 shows the performance of the submitted runs on each part of the test collection and estimates of their performance obtained from the development set. The best overall performance on the test set is obtained with run 2. We observe that the performance is much better on the one-topic part. This was expected given the fact that model and parameter tuning was performed on the development set which did not contain examples of multi-topic queries. The best performance in terms of F1@20 on the one-topic part is obtained by runs 3 and 5, followed by run 1 and then run 2. We see that although being slightly over-optimistic (≈5% on average) our F1@20 estimates for the one-topic part are strongly correlated with the final results and are indicative of the relative run strength.

On the multi-topic part, the best performance is obtained by run 2, followed by run 1 and then runs 3 and 5. The superiority of run 2 over 1 on this part of the collection is attributed to the fact that representative examples of multi-topic queries were available only in textual form. Comparing the final results on this part with our estimates we see very poor correlation[2]. This suggests that this part of the collection has significantly different characteristics from the development set and that performing model selection and parameter tuning on the development set was not helpful. We expect that better results could have been achieved on the multi-topic part, provided that the development set contained queries of this type.

7. ACKNOWLEDGEMENTS

This work is supported by the USEMP FP7 project, partially funded by the EC under contract number 611596.

[2]Nevertheless, these estimates could serve as an indication of the performance of a system that has no access to the Wikipedia images of the one-topic queries.

8. REFERENCES

[1] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky. Neural codes for image retrieval. In *ECCV*, 2014.

[2] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *ACM SIGIR*, 1998.

[3] 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.

[4] D.-T. Dang-Nguyen, L. Piras, G. Giacinto, G. Boato, and F. De Natale. Retrieval of diverse images by pre-filtering and hierarchical clustering. In *MediaEval*, 2014.

[5] T. Deselaers, T. Gass, P. Dreuw, and H. Ney. Jointly optimising relevance and diversity in image retrieval. In *ACM CIVR*, 2009.

[6] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[7] B. Ionescu, A. Gînsca, B. Boteanu, A. Popescu, M. Lupu, and H. Müller. Retrieving diverse social images at MediaEval 2015: Challenge, dataset and evaluation. In *MediaEval*, 2015.

[8] 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.

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

[10] E. Spyromitros-Xioufis, A. Ginsca, A. Popescu, S. Papadopoulos, Y. Kompatsiaris, and I. Vlahavas. Improving diversity in image search via supervised relevance scoring. In *International Conference on Multimedia Retrieval (ICMR)*, 2015.

[11] 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.

[12] E. Spyromitros-Xioufis, S. Papadopoulos, Y. Kompatsiaris, and I. Vlahavas. Socialsensor: Finding diverse images at mediaeval 2014. In *MediaEval*, 2014.