UPC-UB-STP @ MediaEval 2015 Diversity Task: Iterative Reranking of Relevant Images

Aniol Lidon
Xavier Giró-i-Nieto
Universitat Politècnica de Catalunya
Barcelona, Catalonia/Spain
xavier.giro@upc.edu

Marc Bolaños
Petia Radeva
Universitat de Barcelona
Barcelona, Catalonia/Spain
marc.bolanos@ub.edu

Markus Seidl
Matthias Zeppelzauer
St. Pölten University of Applied Sciences
St. Pölten, Austria
m.zeppelzauer@fhstp.ac.at

ABSTRACT

This paper presents the results of the UPC-UB-STP team in the 2015 MediaEval Retrieving Diverse Images Task. The goal of the challenge is to provide a ranked list of Flickr photos for a predefined set of queries. Our approach firstly generates a ranking of images based on a query-independent estimation of its relevance. Only top results are kept and iteratively re-ranked based on their intra-similarity to introduce diversity.

1. INTRODUCTION

The diversification of search results is an important factor to improve the usability of visual retrieval engines. This motivates the 2015 MediaEval Retrieving Diverse Images Task [8], which defines the scientific benchmark targeted in this paper. The proposed methodology solves the trade-off between relevance and diversity by firstly filtering results based on a learned relevance classifier, and secondly building a diverse reranked list following an iterative scheme.

The first challenge in our system is filtering irrelevant images, as suggested in [2]. Relevance is a very abstract concept with a high subjectivity involved. Similar problems have been addressed in the visual domain, as for memorability [10] or interestingness [16]. In both cases, a crowdsourced task was organised to collect a large amount of human annotations used to train a classifier based on visual features.

The second challenge to address is the diversity in the ranked list. A seminar work from 1998 [1] introduced diversity in addition to relevance for text retrieval, a concept that was later ported to image [17, 4, 19] and video retrieval [7, 6]. Different features have been used for this purpose, both textual (e.g. tags [20]), visual (e.g. convolutional neural networks [18]), or multimodal fusion [5].

2. METHODOLOGY

A generic and easily extensible methodology of four steps has been applied in all our submitted runs. While steps 2 and 4 apply to all runs, steps 1 and 3 contain particularities for visual and textual processing.

1) Ranking by relevance: A relevance score for each image is estimated by either using visual or textual information (see details in Section 2.1 and 2.2 respectively).

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

2) Filtering of irrelevant images: Only a percentage of the top ranked images by relevance are considered in later steps. In the multimodal runs, the relevance scores for the visual and textual modalities are linearly normalized and fused by averaging.

3) Feature and distance computation: Visual and/or textual features are extracted for each image, and the similarity between each pair computed.

4) Reranking by diversity: An iterative algorithm selects the most different image with respect to all previously selected ones. The similarity is always assessed by averaging the considered visual and textual features. Iterations start by adding the most relevant image as the first element of the reranked list.

2.1 Visual data

The visual information was analyzed with Convolutional Neural Networks (CNN) [13, 12] with two different approaches:

1) Ranking by relevance: A Relevance CNN was created based on HybridNet [22], a CNN trained with objects from the ImageNet dataset [3] and locations from the Places dataset [22]. HybridNet was fine-tuned in two classes: *relevant* and *irrelevant*, as labeled by human annotators.

3) Feature and distance computation: The fully connected layers *fc7* from a CNN trained on ImageNet [11], and the fully connected layer *fc8* from HybridNet [22] were used as feature vectors [14].

2.2 Textual data

1) Ranking by relevance: For each query, we generate a textual term model in an unsupervised manner from all images returned for this query. We first remove stopwords, words with numeric and special characters and words of length ≤ 4. Next, we select the most representative terms by retaining only those terms where the term frequency (TF_q) is higher than the document frequency (DF_q) for the query q. For each term in the model we store the TF_q as a weight. Once this model has been established, we map the textual descriptions of the images to the model of the query. For each image only terms that appear also in the query model are retained. For each remaining term we retrieve the TF_i for the corresponding ith image and build a feature vector. To compute a relevance score s_i for an image, we compute the cosine similarity sim_i between the query model and a given image feature vector. Additionally, we add the inverse original Flickr rank r_i of the image to the score, yielding a final textual relevance score of $s_i = sim_i + (1/r_i)$ for im-

age i. This computation is inspired by that of [21] with the difference that we use TF instead of TFIDF in the scoring function which showed to be more expressive in our experiments.

3) Feature and distance computation: Diversity re-ranking requires the similarity comparison of all relevant images for a query. For a given image, we first align its terms to the query model. Next, we compute their TFIDF weights (TF_i/DF_i) [15, 23]. Terms from the query model that do not occur in the image's descriptions get a weight of zero. The resulting feature vectors are compared with the cosine metric in diversity re-ranking.

3. EXPERIMENTAL SETUP

The experimental setup is mostly defined by the 2015 MediaEval Retrieving Diverse Images Task, which provides a dataset partitioned into development (devset) and test (testset), two types of queries (single and multi-topic), and standardized and complementary evaluation metrics: Precision at 20 (*P@20*), Cluster Recall at 20 (*CR@20*) and F1-score at 20 (*F1@20*). The reader is referred to the task overview paper [8] to learn the details of the problem.

The Relevance CNN described in Section 2.1 was trained with a 2-fold cross validation, each split containing one half of the devset queries. For both splits we stopped after 2,000 iterations, when the validation accuracy was the highest one (76% and 75% respectively). When applying the best methods' parameters on the testset, we used all the dev data and fine-tuned the network stopping after 4,500 iterations, when the training loss was minimum.

The portion of images to be filtered in Step 2 was learned by measuring the evolution of the final F1-score for different percentages. From *Runs 1* to *3* the best results where obtained by keeping the top 20% of images, while for *Run 5* the best value was 15%.

4. RESULTS

Table 1 presents the results obtained in four different configurations: using visual information only (*Run 1*), using textual data only (*Run 2*), and using the best combination of textual and visual data (*Run 3*). An additional *Run 5* considers multimodal information only for relevance filtering (Step 2) and purely visual information for diversity reranking (Step 4). Rows 2 to 5 presents results on the devset for single-topic queries, while rows 6 to row 13 include the results on the testset for the single-topic and multi-topic queries. The overall results can be found in Rows 14 to 17.

Figure 1 plots the Precision, Cluster Recall and F1-Score curves depending on the amount of N top ranked images considered in the evaluation, averaged over all queries on our best run (Run 3).

Modality	Visual	Text	Multi	Multi
devset	**Run 1**	**Run 2**	**Run 3**	**Run 5**
P@20	0.756	0.802	0.836	**0.847**
CR@20	0.416	0.419	**0.452**	0.447
F1@20	0.530	0.543	**0.578**	0.577
testset (single)	**Run 1**	**Run 2**	**Run 3**	**Run 5**
P@20	0.705	0.6819	**0.749**	0.733
CR@20	0.423	0.383	**0.431**	0.412
F1@20	0.519	0.478	**0.533**	0.513
testset (multi)	**Run 1**	**Run 2**	**Run 3**	**Run 5**
P@20	0.593	**0.724**	0.627	0.621
CR@20	0.403	0.372	**0.414**	0.397
F1@20	0.463	0.47	**0.482**	0.464
testset (overall)	**Run 1**	**Run 2**	**Run 3**	**Run 5**
P@20	0.649	**0.703**	0.688	0.677
CR@20	0.413	0.378	**0.422**	0.405
F1@20	0.491	0.474	**0.508**	0.489

Table 1: Precision, Recall and F1-Scores obtained on each run with $N = 20$ on the devset, and the testset (single-topic, multi-topic and overall).

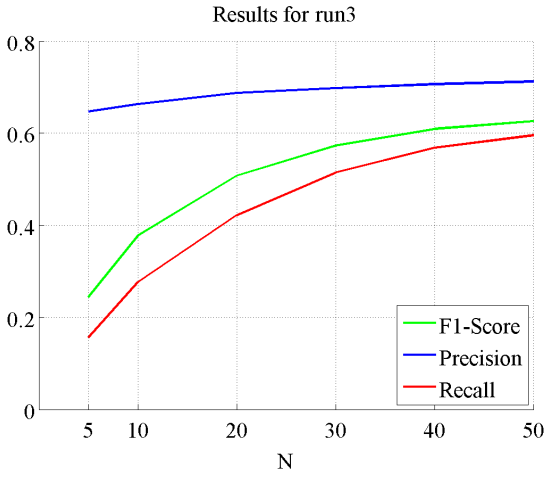

Figure 1: Overall Precision, Recall and F1-score curves for different cutoffs N of top ranked images on all testset queries.

5. CONCLUSIONS

The trade-off between relevance and diversity has been targeted in this work with relevance-based filtering and a posterior iterative process to introduce diversity. The final results, presented in Table 1, are comparable to the state of the art on the devset [9], and achieve up to a F1@20 of 0.508 on the testset.

Multi-topic queries seem to be more difficult to diversify than single-topic queries. A reason may be that multi-topic queries are more general and contain more heterogeneous content. Considering the fact that our method was trained on single-topic queries only, the results for the multi-topic queries are, however, still promising.

It is remarkable that increasing the number of N of retrieved images increases both, recall and precision (and not only recall as one would expect in a typical retrieval scenario), as shown in Figure 1. This indicates that the relevance ranking obtained by our method is accurate (at least for $N \leq 50$).

There is no clear winner between textual and visual information (*Runs 1* and *2*). The multimodal combination, however, clearly improves performance (*Runs 3* and *5*). Additionally, results indicate that using multimodal processing at all stages (*Run 3*) is better than using multimodal processing only during the relevance ranking (*Run 5*).

6. REFERENCES

[1] J. Carbonell and J. Goldstein. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336. ACM, 1998.

[2] D.-T. Dang-Nguyen, L. Piras, G. Giacinto, G. Boato, and F. G. De Natale. A hybrid approach for retrieving diverse social images of landmarks. In *Multimedia and Expo (ICME), 2015 IEEE International Conference on*, pages 1–6. IEEE, 2015.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[4] T. Deselaers, T. Gass, P. Dreuw, and H. Ney. Jointly optimising relevance and diversity in image retrieval. In *Proceedings of the ACM international conference on image and video retrieval*, page 39. ACM, 2009.

[5] Y. Gao, M. Wang, Z.-J. Zha, J. Shen, X. Li, and X. Wu. Visual-textual joint relevance learning for tag-based social image search. *Image Processing, IEEE Transactions on*, 22(1):363–376, 2013.

[6] X. Giro-i Nieto, M. Alfaro, and F. Marques. Diversity ranking for video retrieval from a broadcaster archive. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, page 56. ACM, 2011.

[7] M. Halvey, P. Punitha, D. Hannah, R. Villa, F. Hopfgartner, A. Goyal, and J. M. Jose. Diversity, assortment, dissimilarity, variety: A study of diversity measures using low level features for video retrieval. In *Advances in Information Retrieval*, pages 126–137. Springer, 2009.

[8] B. Ionescu, A. L. Gınsca, B. Boteanu, A. Popescu, M. Lupu, and H. Müller. Retrieving diverse social images at mediaeval 2015: Challenge, dataset and evaluation. In *MediaEval 2015 Workshop, Wurzen, Germany*, 2015.

[9] B. Ionescu, A. Popescu, M. Lupu, A. L. Gînsca, B. Boteanu, and H. Müller. Div150cred: A social image retrieval result diversification with user tagging credibility dataset. *ACM Multimedia Systems-MMSys, Portland, Oregon, USA*, 2015.

[10] P. Isola, J. Xiao, D. Parikh, A. Torralba, and A. Oliva. What makes a photograph memorable? *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(7):1469–1482, 2014.

[11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[13] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[14] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, pages 512–519. IEEE, 2014.

[15] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[16] M. Soleymani. The quest for visual interest. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 2015.

[17] K. Song, Y. Tian, W. Gao, and T. Huang. Diversifying the image retrieval results. In *Proceedings of the 14th annual ACM international conference on Multimedia*, pages 707–710. ACM, 2006.

[18] E. Spyromitros-Xioufis, S. Papadopoulos, A. L. Ginsca, A. Popescu, Y. Kompatsiaris, and I. Vlahavas. Improving diversity in image search via supervised relevance scoring. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 323–330. ACM, 2015.

[19] R. H. van Leuken, L. Garcia, X. Olivares, and R. van Zwol. Visual diversification of image search results. In *Proceedings of the 18th international conference on World wide web*, pages 341–350. ACM, 2009.

[20] R. Van Zwol, V. Murdock, L. Garcia Pueyo, and G. Ramirez. Diversifying image search with user generated content. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 67–74. ACM, 2008.

[21] B. Vandersmissen, A. Tomar, F. Godin, W. De Neve, and R. Van de Walle. Ghent university-iminds at mediaeval 2014 diverse images: Adaptive clustering with deep features. In *MediaEval 2014, Workshop*, 2014.

[22] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*, pages 487–495, 2014.

[23] J. Zobel and A. Moffat. Exploring the Similarity Space. *ACM SIGIR Forum*, 32(1):18–34, 1998.