ETH-CVL @ MediaEval 2015: Learning Objective Functions for Improved Image Retrieval

Sai Srivatsa R
Indian Institute of Technology, Kharagpur
saisrivatsan12@gmail.com

Michael Gygli
Computer Vision Laboratory, ETH Zurich
gygli@vision.ee.ethz.ch

Luc Van Gool
Computer Vision Laboratory, ETH Zurich
vangool@vision.ee.ethz.ch

ABSTRACT

In this paper, we present a method to select a refined subset of images, given an initial list of retrieved images. The goal of any image retrieval system is to present results that are maximally relevant as well as diverse. We formulate this as a subset selection problem and we address it using submodularity. In order to select the best subset, we learn an objective function as a linear combination of submodular functions. This objective quantifies how relevant and representative a selected subset it. Using this method we obtain promising results at MediaEval 2015.

1. INTRODUCTION

Image retrieval using text queries is a central topic in Multimedia retrieval. While early approaches relied solely on text associated with images, more recent approaches combine textual and visual cues to return more relevant results [12, 6]. Nonetheless, search engines of photo sharing sites such as Flickr still retrieve results that are often irrelevant and redundant. The MediaEval 2015 Retrieving Diverse Social Images Task fosters research to improve results retrieved by Flickr. It asks the participants to develop algorithms to refine a ranked list of photos retrieved from Flickr using the photo's visual, textual and meta information. An overview of the task is presented in [4].

2. METHODOLOGY

We formulate the task of diversifying Image retrieval results as a subset selection problem. Given a set of retrieved images, $\mathcal{I} = (I_1, I_2, \ldots, I_n)$ and a budget B, the task is to find a subset $S \subseteq \mathcal{I}, |S| = B$ such that S is maximally relevant as well as diverse. Such problems are usually solved by using a scoring function $\mathcal{F} : 2^n \rightarrow \mathbf{R}$ that assigns a higher score for diverse and relevant subsets. Let $\mathcal{V}$ be the power set of $\mathcal{I}$, we obtain the best subset S^* by computing:

$$S^* = \underset{S \subset \mathcal{V}, |S|=B}{\operatorname{argmax}} \mathcal{F}(S). \quad (1)$$

Evaluating the scores for all possible subsets (2^n) is intractable. We address this issue with submodularity.

A set function f(.) is said to be submodular if

$$f(A \cup v) - f(A) \geq f(B \cup v) - f(B), \quad (2)$$

where $A \subseteq B \subseteq V \backslash v$, V being the ground set of elements [9]. Submodular functions naturally model properties such as representativeness and relevance as they exhibit a diminishing returns property.

If the scoring function is monotone submodular, we can find a near optimal solution for equation 1 using greedy submodular maximization methods [10, 5]. A linear combination of submodular functions with non-negative weights is still submodular. Thus we define our scoring function as

$$\mathcal{F}(S) = \mathbf{w}^T \mathbf{f}(\mathbf{S}), \quad (3)$$

where $\mathbf{f}(\mathbf{S}) = [f_1(S), f_2(S) \ldots f_k(S)]^T$ are normalized submodular monotone functions and $\mathbf{w} \in \mathbb{R}_+^k$ is a weight vector. We learn these weights with sub-gradient descent[1] [7].

2.1 Submodular Scoring Functions

We use several submodular functions, aimed at quantifying how relevant or diverse the selected subset is.

Visual Representativeness We define the representativeness score as 1 - k-Medoid Loss. The k-Medoid loss for a subset is obtained by computing the sum of euclidean distance between images in the query and the nearest selected medoid (images in the selected subset) in the feature space [3] (using CNN features [1]). Thus k-Medoid loss is minimum when the selected subset is representative thereby resulting in a higher representativeness score.

Visual Relevance We use the relevance ground truth provided for the devset topics to train a generic SVM on CNN features with relevance ground truth as labels. The relevance score of a subset is the number of images in the subset that are predicted as relevant.

Text Relevance In order to obtain a text-based score for an image, given a query, we use a Bag-of-Words model. We represent the wikipage associated with the query as a vector. Similarly, each image is represented as vector obtained encoding its title, tags and description (with the same relative weighting as [13]). The text relevance of an image is computed as its cosine similarity to the wikipedia page, using tf-idf weighting[2]. Finally, the text relevance score of a set of image is simply the sum over the relevance of its individual elements.

Flickr Ranks For an image having Flickr rank i belonging to a topic having n images, its Flickr score is given by $\frac{n-i}{n}$. The sum of flickr scores of images in the subset is the flickr score of the subset.

[1]We use the implementation of [3] for submodular maximization and learning weights.

[2]Using the implementation provided in scikit-learn [11].

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

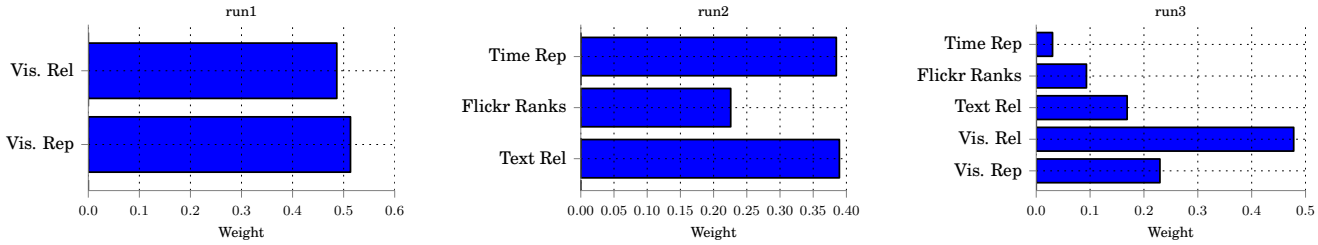

Figure 1: Weights learnt for normalized submodular objectives for various configurations (See Sec. 3).

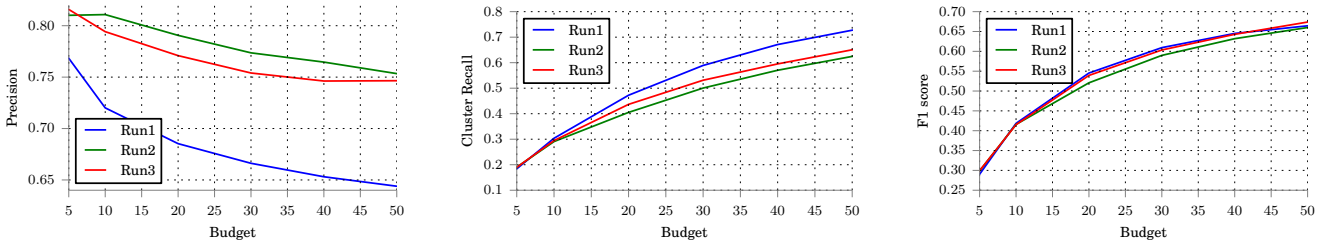

Figure 2: Precision, Cluster Recall and F1 scores for the official runs on the dataset of [4].

Time Representativeness This function quantifies how diverse the images are with respect to time taken. Photos taken during different times of the day, or taken during different seasons can also lead to increase in diversity. This score is computed using the same k-medoid loss as in Visual representativeness, but using the timestamp as the feature representation.

2.2 Learning

Using the relevance and cluster ground truth, for a given query and a budget B, we construct a ground truth subset (S_t^{gt}) for each query t in the devset. To learn the weights, we optimize the following large-margin formulation [7]

$$\min_{w \geq 0} \frac{1}{T} \sum_{t=1}^{T} \hat{L}_t(w) + \frac{\lambda}{2} ||w||^2 \tag{4}$$

where T is the total number of queries in the devset and $\hat{L}_t(w)$, the hinge loss of for training examples t is given by

$$\hat{L}_t(w) = \max_{S_t \in V_t} (\mathcal{F}(S_t) + \ell(S_t)) - \mathcal{F}(S_t^{gt}) \tag{5}$$

where $\ell(.)$ is the loss function. We use F1-loss ($\ell(S_t) = |S_t| - F1(S_t)$) as the loss function. As F1-loss is not submodular, we use its (pointwise) modular approximation [9]. We perform the optimization using sub-gradient descent [7] with an adaptive learning rate [2].

3. RESULTS AND DISCUSSION

We evaluated our method on the MediaEval 2015 diverse social images task [4]. The test data consists of 139 queries with more than 40,000 images. It includes single-topic (location) as well as multi-topic queries (events associated with locations). In Fig. 2 we show performance for different configurations and varying budgets. The configurations are: (i) **Run 1** - Visual only, i.e. relevance prediction and representativeness. (ii) **Run 2** - Meta only: In this run we only use information associated with the image, but not the image itself, i.e text relevance, Flickr rank and time representativeness. (iii) **Run 3** - we use a combination of the above mentioned objectives. In Tab. 1 we provide the results using the official performance metrics computed by [4]. The distribution of weights learnt for each shell is as shown in Fig. 1.

Run Type	Run Description	P@20	CR@20	F1@20
Run 1	all	0.6853	**0.4724**	**0.5453**
	single-topic	0.6877	**0.4829**	0.5575
	multi-topic	0.6829	**0.4622**	**0.5333**
Run 2	all	**0.7906**	0.4051	0.5207
	Single-topic	0.8290	0.4145	0.5406
	Multi-topic	**0.7529**	0.3958	0.5010
Run 3	All	0.7709	0.4366	0.5393
	Single-topic	**0.8420**	0.4420	**0.5674**
	Multi-topic	0.7007	0.4312	0.5116

Table 1: *Official Results.* We report performance metrics according to [4]. Best results are highlighted in bold.

The visual run yields higher cluster recall while the textual run yields a better value of precision. This suggests that using visual information is effective for diversifying the retrieval results while textual information is more effective for retrieving relevant images. The lower precision of the visual run is not surprising, as it only uses a generic relevance prediction. While this allows to filter out images of people and several non-landmarks, it does not score relevance in a query-specific way. In order to improve our visual approach it is thus necessary to compute similarities between text queries and images. This could be done by learning a joint embedding of text and images, similar to e.g. [8]. We also note that the method that we use performs better on the single-topic sets than the multi-topic sets.

4. REFERENCES

[1] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International Conference on Machine Learning (ICML)*, 2014.

[2] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, 2011.

[3] M. Gygli, H. Grabner, and L. Van Gool. Video Summarization by Learning Submodular Mixtures of Objectives. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[4] B. Ionescu, A. L. Ginsca, B. Boteanu, A. Popescu, M. Lupu, and H. Müller. Retrieving diverse social images at mediaeval 2015: Challenge, dataset and evaluation. In *Proceedings of MediaEval Benchmarking Initiative for Multimedia Evaluation*, 2015.

[5] A. Krause and D. Golovin. Submodular function maximization. *Tractability: Practical Approaches to Hard Problems*, 2012.

[6] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 2006.

[7] H. Lin and J. Bilmes. Learning mixtures of submodular shells with application to document summarization. In *Uncertainty in Artificial Intelligence (UAI)*, 2012.

[8] W. Liu, T. Mei, Y. Zhang, C. Che, and J. Luo. Multi-task deep visual-semantic embedding for video thumbnail selection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[9] M. Narasimhan and J. Bilmes. A submodular-supermodular procedure with applications to discriminative structure learning. In *Uncertainty in Artificial Intelligence (UAI)*, 2005.

[10] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions-I. *Mathematical Programming*, 1978.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011.

[12] Y. Rui, T. S. Huang, and S.-F. Chang. Image retrieval: Current techniques, promising directions, and open issues. *Journal of Visual Communication and Image Representation*, 1999.

[13] 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 *ACM on International Conference on Multimedia Retrieval*. ACM, 2015.