Ghent University-iMinds at MediaEval 2014 Diverse Images: Adaptive Clustering with Deep Features

Baptist Vandersmissen[1]
baptist.vandersmissen@ugent.be

Abhineshwar Tomar[1]
abhineshwar.tomar@ugent.be

Fréderic Godin[1]
frederic.godin@ugent.be

Wesley De Neve[1,2]
wesley.deneve@ugent.be

Rik Van de Walle[1]
rik.vandewalle@ugent.be

[1] Multimedia Lab, ELIS, Ghent University – iMinds, Ghent, Belgium
[2] Image and Video Systems Lab, KAIST, Daejeon, South Korea

ABSTRACT

In this paper, we attempt to tackle the MediaEval 2014 Retrieving Diverse Social Images challenge, a filter and refinement problem defined for a Flickr-based ranked set of social images. We build upon solutions proposed in [5] and mainly focus on exploiting the joint use of all modalities. The use of image features extracted from a deep convolutional neural network, combined with the use of distributed word representations, forms the basis of our approach.

1. INTRODUCTION

In this paper, we describe our approach for tackling the MediaEval 2014 Retrieving Diverse Social Images Task [1]. This task focuses on result diversification in the context of image retrieval. We refer to [1] for a complete task overview.

2. METHODOLOGY

This section describes four different approaches created to solve the aforementioned challenge. The approach used in the last run uses external data sources; all other approaches exclusively use data provided by the task organizers. We focused on two parts: relevance estimation of an image with respect to a specific location and similarity estimation between a pair of images. Particularly, run 2, 3 and 5 build upon these parts.

2.1 Run 1: Visual-only

We propose a hierarchical clustering-based approach for the ranking of images in accordance with their relevance and diversity for a specific location. We used the approach proposed in [5] (cf. "Visual run").

2.2 Run 2: Textual-only

The textual run makes use of information derived from the provided tags and other textual metadata. This approach aims at diversifying the results by optimizing an adapted performance metric. We modified both the relevance and diversity estimation of the algorithm proposed in [5] (cf. "Textual run") as presented in the following sections.

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

2.2.1 Relevance Estimation

The relevance of an image is estimated by making use of textual metadata. Let $\mathscr{T}_x$ denote the set of tags assigned to image x. Next formula predicts the relevance of image x:

$$Rel(x) = \alpha \times tags(x) \ + \ \beta \times \frac{1}{flickr(x)}, \qquad (1)$$

with α and β representing scalars,

$$tags(x) = \frac{|\{t \mid t \in \mathscr{T}_x, \ tf\_idf_t > \lambda\}|}{|\mathscr{T}_x|} \times \sum_{t \in \mathscr{T}_x} tf\_idf_t, \quad (2)$$

and $flickr(x)$ denoting the original Flickr ranking of image x. The TF-IDF score of tag t is denoted by tf_idf_t. The tag score (cf. Equation 2) is the sum of each tag's normalized TF-IDF score multiplied by the relative number of high scoring tags. In our approach, λ is set to the average TF-IDF score. This benefits images with a higher number of more relevant tags.

2.2.2 Diversity Estimation

Estimating the semantic difference between two images is based on the amount of shared tags. Let x and y denote two images with $\mathscr{T}_x$ and $\mathscr{T}_y$ denoting their set of tags, respectively. The diversity is then calculated as follows:

$$Div(x, y) = 1 - \frac{|\mathscr{T}_x \cap \mathscr{T}_y|}{max(|\mathscr{T}_x|, |\mathscr{T}_y|)}. \qquad (3)$$

2.3 Run 3: Visual and Textual

The fusion of both visual and textual information results in a relevance-based clustering approach (cf. "Combined run" in [5]). We modified the clustering technique to adaptive hierarchical clustering. The optimal distance to form clusters is determined by finding the "knee" point in the plot of number of clusters versus the inter-cluster distance (similar to [3]). To estimate the relevance of an image, we use our textual-only method (cf. Section 2.2.1). The diversity between two images is estimated based on the Euclidean distance between their visual descriptor, which is represented by a CN3x3 and LBP3x3 vector [1].

2.4 Run 5: External Sources

The algorithm used to produce the fifth run is based on the one used in Section 2.3. Both the relevance and diversity

Table 1: Results on development set.

	Flickr	Run 1	Run 2	Run 3	Run 5
P@20	0.8333	0.7083	0.7500	0.7700	**0.8567**
CR@20	0.3455	0.3967	**0.4441**	0.4043	0.4289
F1@20	0.4885	0.5086	0.5579	0.5302	**0.5716**

Table 2: Results on test set.

	Run 1	Run 2	Run 3	Run 5
P@20	0.6232	0.7480	0.7557	**0.8008**
CR@20	0.3600	**0.4279**	0.4035	0.4252
F1@20	0.4503	0.5369	0.5181	**0.5455**

estimation components are adapted and described below.

2.4.1 Relevance Estimation

In order to accurately estimate the relevance of an image, a well-defined target location is necessary. Thus, each location is first described in both a textual and visual manner.

To create this textual identity, related information of each location is extracted from DBpedia[1]. From this information textual keywords are extracted and combined with the top k most frequently occurring tags in the set of images of a location. The visual identity is formed on the basis of a set of representative photos, retrieved via Wikipedia. The relevance of an image is calculated based on a linear combination of the following three factors: textual relevance, visual relevance, and Flickr relevance.

The textual relevance of an image is entirely based on its tags. Again, assume that $\mathscr{T}_x$ denotes the set of tags of image x and that $\mathscr{T}_a$ denotes the set of tags depicting location a (i.e., textual identity):

$$Rel(x) = \frac{\sum_{t \in \mathscr{T}_x} e^{max_{k \in \mathscr{T}_a}\{sim(t,k)\}}}{|\mathscr{T}_x|}, \tag{4}$$

We propose a new method to compute the similarity between tags and omit the use of the ubiquitous TF-IDF. Therefore, we make use of distributed word representations, namely word2vec[2]. A pretrained model (the Google News Dataset-based dictionary defined as $\mathscr{T}_w$) is used to convert words to vectors. Such vectors preserve the semantic and linguistic regularities among words [2]. The following formula describes this approach:

$$sim(t_a, t_b) = \begin{cases} cos(\Theta) & \text{if } t_a \in \mathscr{T}_w \wedge t_b \in \mathscr{T}_w \\ 1 & \text{if } t_a \notin \mathscr{T}_w \vee t_b \notin \mathscr{T}_w, t_a = t_b \\ 0 & \text{else} \end{cases}, \tag{5}$$

with t_a and t_b depicting a tag, and $cos(\Theta)$ the cosine similarity between their representative vectors. With this technique, semantically similar and spelling-wise different tags can still have an influence on the eventual relevance score.

Visual relevance is calculated based on the maximum similarity between the image and the representative Wikipedia images. Finally, Flickr relevance is the inverse of the original Flickr ranking of the image.

2.4.2 Diversity Estimation

To improve the similarity estimation and thus dissimilarity estimation between two images, we attempt to find more effective visual descriptors. Therefore, we make use of a deep convolutional neural network, trained on 1.2 million images from ImageNet, named OverFeat[3], to extract high-level features [4]. Each image is resized and cropped to a size of 231 pixels by 231 pixels, then for each image a representative vector is extracted from a convolutional network. This is done by feed-forward propagation through the network and omitting the fully connected layers, which results in a vector of size 4096 for each image. Thus, we assume that the numerous filters in the convolutional layers extract high-level and representative features. The diversity between two images is then again estimated based on the Euclidean distance between their descriptors.

3. EXPERIMENTS

In Table 1, we can see the results of the original Flickr ranking together with the results of all algorithms on the development set. Table 2 shows the results on the test set. Clearly, run 5 outperforms the other approaches when observing the F1-measure. Run 5 reaches an F1-score of 57.16% on the development set and 54.55% on the test set.

4. CONCLUSIONS

We observe that run 5, using distributed word representations for the relevance estimation and OverFeat features for the diversity assessment, outperforms all others. Particularly, the use of advanced image features positively influences the F1-score. For future work, the influence of more focused distributed word representations will be investigated.

5. REFERENCES

[1] B. Ionescu, A. Popescu, M. Lupu, A. L. Ginsca, and H. Müller. Retrieving diverse social images at mediaeval 2014: Challenge, dataset and evaluation. In *MultimediaEval working Notes*, 2014.

[2] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*. NIPS, 2013.

[3] S. Salvador and P. Chan. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In *Tools with Artificial Intelligence*, pages 576–584, Nov 2004.

[4] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*, 2013.

[5] B. Vandersmissen, A. Tomar, F. Godin, W. De Neve, and R. Van de Walle. Ghent University-iMinds at MediaEval 2013 Diverse Images: Relevance-Based Hierarchical Clustering. *Working Notes Proceedings of the MediaEval 2013 Workshop, Barcelona, Spain, October 18-19, CEUR-WS, 1043*, 2013.

[1] http://dbpedia.org/

[2] https://code.google.com/p/word2vec/

[3] http://cilvr.nyu.edu/doku.php?id=code:start