NLE @ MediaEval'17: Combining Cross-Media Similarity and Embeddings for Retrieving Diverse Social Images

Jean-Michel Renders and Gabriela Csurka
Naver Labs Europe, Meylan, France
firstname.lastname@naverlabs.com

ABSTRACT

In this working note we briefly describe the methods we used in the MediaEval17, Retrieving Diverse Social Images Task and give details on the submitted runs.

1 INTRODUCTION

One of the main motivations for participating in the MediaEval'17 Retrieving Diverse Social Images Task [22] was to evaluate the cross-media similarity measure we proposed in [3, 4] , which has proven to give top-ranked retrieval results on several ImageCLEF multimedia search tasks between 2007 and 2011[1] [12].

The cross-media similarity we adopted this year differs from the one we used previously in the features used for both visual and textual modalities. Indeed, previously we used Fisher Vectors [15] for image representation and standard Dirichlet language model (LM) or Lexical Entailment [2] for text representation. However, recent progresses have shown that using activation layers of deep Convolutional Networks pre-trained on ImageNet as image representation performs better than Fisher Vectors [15] on visual task due to the large amount of knowledge learned from ImageNet. Similarly, word embedding-based representations such as *word2vec* relying on the information learned from large textual collections outperform standard *tfidf*-based and previous LM-based representations. Therefore, in our cross-media similarity model we used features extracted from deep models pre-trained on ImageNet and word embeddings learned from a large corpus of queries[2].

A second motivation was to compare this simple approach with more recent image and text combination strategies, such as joint image and text embedding [5, 6, 8, 20, 21]. These methods, in contrast to our fully unsupervised cross-media similarity, exploit labels or relevance scores to learn the embedding.

Finally, our third motivation was to evaluate several methods to make the top ranked images more diverse. In particular, we explored a clustering-based method, with several visual, textual and joint similarity measures: images were re-ranked based on the number of times a document shared clusters with documents already present in the upper ranked images (the lower, the better). While this family of methods allowed us to significantly increase the cluster recall, these methods turned out to perform below the classical Maximum Margin Relevance method (MMR) proposed in [1], at least for the development set.

[1]For more details, please visit www.imageclef.org
[2]The models used to get these representations were built prior and independently from the challenge.

MediaEval'17, 13-15 September 2017, Dublin, Ireland

2 INCREASING THE TOP RELEVANCE

2.1 Cross-media and Mono-media Relevance

First, we describe our cross-media similarity measure, which we already proposed in [3, 4]. This cross-media similarity measure is a relatively simple extension of pseudo-relevance feedback and, can be applied to a single media as well (text or image). It can be considered as a two-step similarity measure, where the final similarity between a query and a document[3] is nothing else but the average visual similarity between the document and the top-K documents most "textually"-similar (*i.e.* relevant) to the query.

More formally, if we denote by $S_V(d, d')$ the normalized visual similarity measure between documents d and d', and by $S_T(d, q)$ the textual relevance score of document d with respect to query q, the new relevance score of a document d is defined as a weighted average of its similarity with the top retrieved documents based on the textual relevance scores:

$$S_{T,V}(d,q) = \frac{\sum_{d_i \in NN_T^K(q)} S_T(d_i,q) S_V(d,d_i)}{\sum_{d_i \in NN_T^K(q)} S_T(d_i,q)} \quad (1)$$

where $NN_T^K(q)$ denotes the top-K documents most similar to the query q using only the textual modality. We called it cross-media similarity, because it represents in some sense the similarity between a textual query and the visual part of a document.

From our experiments, we observe better performances if we recombine this score with the initial relevance scores as a convex linear combination: $\tilde{S}_{T,V}(\mathbf{d},\mathbf{q}) = (1-\alpha) S_{T,V}(\mathbf{d},\mathbf{q}) + \alpha S_T(\mathbf{d_i},\mathbf{q})$. This algorithm is the core of *NLE-RUN3*.

Note that we can apply a similar two-step similarity measure, using only visual (or textual, resp.) features in both steps. Concretely, we obtain a pure text-based retrieval model (more or less equivalent to classical pseudo relevance feedback) by replacing in (1) $S_V(d, d')$ with $S_T(d, d')$, the normalized textual similarity measure between documents d and d'; this results in a purely textual relevance score $\tilde{S}_{T,T}(\mathbf{d},\mathbf{q})$. This method corresponds to *NLE-RUN2*.

By analogy, assuming (abusively) that the Flickr ranking is based on the image only, we can replace the term $S_T(d_i, q)$ in (1) with $S_F(d_i, q)$, the normalized "Flickr" relevance score, defined as $(n - r)/r$, where n is the number of images returned by Flickr and r is the provided Flickr rank of document d_i; this results in a purely visual relevance score $\tilde{S}_{F,V}(\mathbf{d},\mathbf{q})$. This method corresponds to *NLE-RUN1*.

2.2 Joint visual and textual embedding

We considered the joint textual and visual embedding model proposed in [20], where the idea is to use a two-view neural network with two layers of non-linearities on top of any representation of

[3]Here a document refers to a Flickr image with its textual and visual representations.

the image and text views. To train this network, in a way which is reminiscent of some "learning to rank" strategies, we use 4 different triplet losses (visual-visual, textual-textual, visual-textual and textual-visual). The aim is to enforce that two documents relevant to the same query should have both textual and visual embeddings close in the new common (*i.e.* joint) latent space, while a document relevant to a query q should be far from documents non-relevant to the same query or from documents relevant to other queries. More formally, given a set of triplets (d_i, d_j, d_k) built from the set of queries and their associated documents, the method amounts to minimizing the following loss function:

$$\begin{aligned}\mathcal{L}(d_i,d_j,d_k) &= max[0, m + d(p_i^V, p_j^V) - d(p_i^V, p_k^V)] \\ &+ max[0, m + d(p_i^T, p_j^T) - d(p_i^T, p_k^T)] \\ &+ max[0, m + d(p_i^V, p_j^T) - d(p_i^V, p_k^T)] \\ &+ max[0, m + d(p_i^T, p_j^V) - d(p_i^T, p_k^V)]\end{aligned}$$

where p_i^V and p_i^T are the projections of the visual respectively textual representation of document d_i into the common embedded space. To select such triplets for training, we experimented with using the ground-truth relevance scores provided with the development set but we have observed that they do not generalize for unseen topics. Therefore, instead, we used the pseudo-relevance scores (using our cross-media similarity scores) by considering the top-ranked documents[4] as relevant to the query; the bottom-ranked documents as well as all documents associated to the other queries were assumed to be non-relevant.

After the model was trained, we computed embeddings both for the textual queries and documents. For the documents – which have two embeddings –, we considered the centroid of their visual and textual embeddings and ranked them according to their distance to the query in the embedding space. This approach was used to build our *NLE-RUN4* and *NLE-RUN5* runs.

3 PROMOTING DIVERSITY

Note that in general promoting diversity comes with a risk of decreased precision as we discard in general relevant elements from the top that are similar to other elements on the top. Our aim therefore was to find a good trade-off between keeping the relevance as high as possible while introducing diversity. The best performance on the development set was obtained with the Maximum Margin Relevance method (MMR) proposed in [1]. The main idea of the method is that we re-rank documents by considering new scores which corresponds to their initial relevance scores diminished with the maximum similarity score compared to the documents already selected weighted by a penalty factor β.

4 RESULTS AND ANALYSIS

The methods presented here above, based on pseudo-relevance feedback, heavily depends on the choice of the mono-modal similarity measures and, consequently, on a good textual/visual representation of the query and the documents.

[4]We considered as relevant documents with scores $> mean + std$ and non-relevant scores $< mean + std$, where $mean$ and std are the mean and standard deviation of the the scores within the topic.

Table 1: The retrieval results for our main runs

Results (@20)	P	CR	F1	ERR-IA	α-nDCG
Run1 (V)	73.2	59.4	63.3	66.0	62.3
Run2 (T)	72.7	61.7	64.3	66.3	62.8
Run3 (VT)	78.2	67.9	70.5	73.3	68.9
Run4 (VT)	79.3	66.3	69.8	72.3	67.9
Run5 (VT)	78.1	66.4	69.4	73.0	68.6

For the textual facet, after trying *word2vec* and Glove [14] embeddings, we finally decided to adopt the Dual Embedding Space Model for Document Ranking [13], pre-trained on the Bing query corpus[5]. This choice was motivated by the fact that this embedding specifically designed for IR applications experimentally turned out to give better performance on the development set. Document and query embeddings are simply computed as the average of the embeddings of their constitutive words; we then use a simple mixture of the Dirichlet-smoothed LM relevance score with the cosine similarity of the textual embeddings as the $S_T(d,q)$ textual relevance score.

As visual representation, we considered several deep CNN models pretrained on ImageNet. We experimented with AlexNet [10], GoogleNet Inception V3 [18], Inception-ResNet [17] and RMAC[6] [7, 19] deep models. The pretrained models were used as such, without any fine tuning on the task collection. We used as visual representation the activations of the last fully connected layer preceding the class prediction one. The features were L2-normalized and the dot product used as similarity.

We used the provided ground truth on the development set and considered the P@50 to select the best visual similarity and to set the parameters. Best results were found with the features extracted from the Inception-ResNet [17] model. As best choice for the parameters in (1), we found $K = 25$ and $\alpha = 0.15$.

To promote diversity, we used for all runs the classical MMR applied to the initial relevance scored computed by the methods described above. The metrics used in MMR to penalize documents similar to higher rank documents was the RMAC visual similarity between images, except for *NLE-RUN2*, where we used the cosine similarity between text embeddings to keep the run purely textual. The weight factor β that penalizes a too high similarity with higher rank documents was tuned using the development set.

Our runs are summarized in Table 1. We can see that our visual only and textual only runs have similar performances, the visual one having slightly higher precision and the text higher diversity. Using the cross-media similarity allowed us to obtain a much better ranking both in terms of precision and also diversity. Learning joint visual and textual embedding using the relevance scores did not help, or even slightly degraded the results. The main reason is that the embedding only learned from information already captured by the cross-media similarity[7].

Acknowledgement: We would like to thank Jon Almazan, who provided us with RMAC representations for the images.

[5]See http://research.microsoft.com/projects/DESM

[6]The RMAC model [7, 19] is trained with a triplet loss instead of a classification loss, to make the distance between images from the same class smaller than the distance to images from other classes plus a margin.

[7]Originally, we intended to use external data such as Visual Genome [9] or Flickr30K Entities [16] to learn embeddings such as relationships between objects and persons, etc. Due to time constraint we will investigate this in the future.

REFERENCES

[1] Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.

[2] Stéphane Clinchant, Cyril Goutte, and Éric Gaussier. 2006. Lexical Entailment for Information Retrieval. In *European Conference on Information Retrieval Research*.

[3] Stéphane Clinchant, Jean-Michel Renders, and Gabriela Csurka. 2007. XRCE's participation to ImageCLEF. In *CLEF online Working Notes*.

[4] Stéphane Clinchant, Jean-Michel Renders, and Gabriela Csurka. 2008. Trans–Media Pseudo–Relevance Feedback Methods in Multimedia Retrieval. In *Advances in Multilingual and Multimodal Information Retrieval*. Vol. LNCS 5152. Springer, 569–576.

[5] Andrea Frome, Greg S. Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. DeViSE: A Deep Visual-Semantic Embedding Model. In *Annual Conference on Neural Information Processing Systems (NIPS)*.

[6] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. 2014. Improving Image-Sentence Embeddings Using Large Weakly Annotated Photo Collections. In *European Conference on Computer Vision (ECCV)*.

[7] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. 2016. Deep Image Retrieval: Learning global representations for image search. In *European Conference on Computer Vision (ECCV)*.

[8] Albert Gordo and Diane Larlus. 2017. Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[9] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. 2017. Visual Genome: Connecting language and vision using crowdsourced. 123 (2017), 32–73.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Annual Conference on Neural Information Processing Systems (NIPS)*.

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C.Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*.

[12] Henning Müller, Paul Clough, Theo Deselaers, and Barbara Caputo (Eds.). 2010. *ImageCLEF- Experimental Evaluation in Visual Information Retrieval*. Vol. INRE. Springer.

[13] Eric Nalisnick, Bhaskar Mitra, Nick Craswell, and Rich Caruana. 2016. Improving Document Ranking with Dual Word Embeddings.

[14] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543.

[15] Florent Perronnin and Chris Dance. 2007. Fisher Kernels on Visual Vocabularies for Image Categorization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[16] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models.

[17] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. 2016. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *CoRR* arXiv:1602.07261 (2016).

[18] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[19] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. 2016. Particular object retrieval with integral max-pooling of CNN activations.. In *International Conference on Machine Learning (ICML)*.

[20] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning Deep Structure-Preserving Image-Text Embeddings. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[21] jason Weston, bengio Bengio, and Nicolas Usunier. 2011. WSABIE: Scaling Up To Large Vocabulary Image Annotation. In *AAAI International Joint Conference on Artificial Intelligence (IJCAI)*.

[22] Maia Zaharieva, Bogdan Ionescu, Alexandru Lucian Gînscă, Rodrygo L.T. Santos, and Henning Müller. 2017. Retrieving Diverse Social Images at MediaEval 2017: Challenges, Dataset and Evaluation. In *Medieval 2017, Multimedia Benchmark Workshop*.