Large Scale Image Indexing Using Online Non-negative Semantic Embedding

Jorge A. Vanegas and Fabio A. González

MindLab Research Group, Universidad Nacional de Colombia, Bogotá, Colombia
{javanegasr,fagonzalezo}@unal.edu.co

Abstract. This paper presents a novel method to address the problem of indexing a large set of images taking advantage of associated multimodal content such as text or tags. The method finds relationships between the visual and text modalities enriching the image content representation to improve the performance of content-based image search.

This method finds a mapping that connects visual and text information that allows to project new (annotated and unannotated) images to the space defined by semantic annotations, this new representation can be used to search into the collection using a query-by-example strategy and to annotate new unannotated images. The principal advantage of the proposed method is its formulation as an online learning algorithm, which can scale to deal with large image collections. The experimental evaluation shows that the proposed method, in comparison with several baseline methods, is faster and consumes less memory, keeping a competitive performance in content-based image search.

1 Introduction

Large online collections of images are becoming common, thanks to the fast advance in acquisition, storage and communication technology. These collections are potential source of knowledge, but an effective and efficient access to them is fundamental to harness this potential. The classic way to search for images is by typing keywords on a search engine, but in many cases it is desirable to search by providing an example image. This approach, called content-based image retrieval, has been studied during the last two decades resulting in important progress . However, it is well known that matching visual features alone may lead to results with lack of semantic validity [16]. In this paper we address the problem of indexing the visual content of an image collection, enriching it with the semantic information provided by text annotations. The method presented in this papers learns relationships between visual features and text keywords co-occurring in images. A successful strategy to find these relationships is to build a common semantic representation space where both image and text content are embedded. This has been previously approached using different methods: Latent Semantic Analysis (LSA) [8], Latent Dirichlet Allocation (LDA) [1], Non-negative Matrix Factorization (NMF) [4] and Non-negative Semantic Embedding (NSE) [19], among others. The main drawback of most semantic learning strategies is that the algorithms are memory and computation intensive [7]. In order to address this drawback, it is proposed

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part I, LNCS 8258, pp. 367–374, 2013.

a reformulation of the NSE algorithm as an online learning process, which scales up to data collections with a vast amount of samples.

This work presents two main contributions: first, a reformulation of the NSE algorithm to make it scalable to large image collections, and second, an experimental evaluation of the algorithm performance in a content-based image retrieval task. The rest of this paper is organized as follows: Section 2 discusses the related work; Section 3 introduces the proposed method called Online Non-negative Semantic Embedding (ONSE); Section 4 presents the experimental evaluation; and, finally, Section 5 presents some concluding remarks.

2 Related Work

The strategy of finding relationships between visual and text representations has been extensively studied in the last years, specially focused in the task of image annotation. However many of the proposed algorithms have been designed without considering a large scale setup [15,10,11]. In some cases, these algorithms can be scaled up by relying on parallelized implementations and assuming the availability of abundant computational resources. However, this can be expensive, tricky and hard to accomplish.

There are some works that try to make semantic embedding approaches suitable for large scale collections. For example, Hsan et al. [18] propose to utilize multi-modality cues by incorporating visual and textual information as embedded objects, by using a simple linear projection to approximate the embedding functions, solving a non-smooth convex optimization problem. Their goal is to make the method (called Modified Multistage Convex Relaxation, MMCR) suitable for large scale image collections by reformulating the basic algorithm in some way that is possible to reduce the time complexity and the amount of storage, achieving a significant reduction in time complexity. Also, Jason Weston et al. [20] present a scalable architecture, proposing methods that learn to represent images and annotations jointly in a low dimension embedding space. To make training time efficient, they propose a loss function based in stochastic gradient descent (SGD) approach. Likewise, Juan Caicedo et al. [6] propose multimodal matrix factorization algorithms based on SGD to decompose a training data set, and find correspondences between visual patterns and text terms in large image collection.

The proposed algorithm in this work is based on a stochastic gradient descent approach, which, according to the work of Bottou [2], requires very little time to reach a predefined expected risk. This makes the strategy suitable for large scale learning problems, providing guarantees about convergence and scalability [2,3].

3 Online Non-negative Semantic Embedding Model

When the image associated text has a rich and clean semantic interpretation (e.g. tags provided by experts), the text representation may be used directly as the semantic space. So the problem of finding a common semantic representation for both visual and text content is reduced to map the visual content to the semantic space defined by the tags. A method that follows this strategy is the Non-negative Semantic Embedding (NSE) [19].

3.1 Non-negative Semantic Embedding

If the visual and semantic representations are vectors, a database of images can be represented with two matrices by joining the corresponding vectors of visual and semantic features as columns of the matrices. Let $V \in \mathbb{R}^{n\times l}$ be the matrix of visual features, where n is the number of visual patterns in the bag of features representations and l the number of images in the collection, and let $T \in \mathbb{R}^{m\times l}$ be the matrix of text terms, with m the number of keywords in the terms dictionary. NSE is used when we assume that the semantic encoding is already known, and we use it to index and represent all images in the collection. We formulate this problem as finding a linear transformation of the visual data imposing a non negativity constraint on the solution: $V \approx ST; S \geq 0$. Where, $S \in \mathbb{R}^{n\times m}$ is the transformation matrix representing the relationships between the visual and text modalities. The non-negativity constraint in this case enforces an additive reconstruction of visual features, since vectors in the matrix S can be thought of as parts of images that are combined according to the presence of associated labels. Notice that the vectors in S can be interpreted as the visual features related to each text term. Our purpose is to solve the problem under an online formulation using stochastic gradient descent, which is a gradient descent optimization method for minimizing an objective function that is written as a sum of differentiable functions. In this context, we can formulate the problem of semantic embedding as the optimization problem of $\min_{S\geq 0} d(V, ST)$. Where, $d(.,.)$ is a function that measures the difference between V and ST. The purpose is to find S that minimize this difference.

3.2 Kullback-Leibler Divergence Optimization

A popular measure function for NMF is the generalized Kullback-Leibler divergence between V and ST [14], Although the KL-divergence equation is not symmetric, and therefore, it is not strictly a distance metric. This allows to take advantage of the normalized visual and text representation that can be interpreted as probability distributions. Zhirong Yang et. al [21] show that projected gradient methods based in for KL-divergence runs faster and yields better approximation than others widely used NMF algorithms. The updating rule for gradient descent approach with τ as the index of iterations and γ as the step size is:

$$S_{\tau+1} = S_\tau + \gamma \left[\left(\frac{V}{ST} - [1]_{n\times l} \right) T^{\mathrm{T}} \right] \quad . \tag{1}$$

This algorithm requires a non-negativity restriction that can be incorporated by using a projected gradient strategy. The projection function maps a point back to the feasible region in each iteration [13], updating the current solution S_τ to $S_{\tau+1}$ by the following rule:

$$S_{\tau+1} = P\left[S_\tau - \gamma \nabla f(S_\tau)\right]; \quad P[s_{ij}] = \begin{cases} s_{ij} & if\, s_{ij} \geq 0, \\ 0 & if\, s_{ij} < 0, \end{cases} \quad . \tag{2}$$

3.3 Online Formulation

The idea of online learning using stochastic approximations is to compute the new solution for each unknown in the problem using a single data sample at a time.

Algoritmo 1. Online Non-negative Semantic Embedding

```
input $S^0$: Initial transformation matrix, $\gamma_0$: initial step size, $N$: number of iterations
for $k = 1$ to $N$ do
    1. Step size calculation: $\gamma_k = \gamma_0/(1+\gamma_0\lambda k)$
    2. Update transformation matrix: $S_{\tau+1} = P\left[S_\tau - \gamma_k\left[\left(\frac{v_\tau}{S_\tau t_\tau} - [1]_{n\times m}\right)t_\tau^\top\right]\right]$
end for
return $S_{\tau+1}$
```

Then, we can scan large data sets without memory restrictions. The updating rule has to be reformulated in such a way that it only depends on the τ-th sample (v_t , t_t, visual and text features for the τ-th image). The updating rule is reformulated as follows:

$$S_{\tau+1} = S_\tau + \gamma\left[\left(\frac{v_\tau}{S_\tau t_\tau} - [1]_{n\times 1}\right)t_\tau^\top\right] \ . \tag{3}$$

The resulting algorithm (Algorithm 1) starts by randomly initialization of the transformation matrix. Each iteration consists on updating the transformation matrix from an observed pair of visual and text features randomly obtained. The step size used in this algorithm is a decreasing rate [2] that depends on the number of iterations and an initial learning rate γ_0. A small variation of this algorithm is obtained by using several samples at each iteration instead of using only one. Experimental results show faster execution when using mini-batches instead of single examples, and also a better numerical stability for the solution.

3.4 Image Indexing and Search

A special indexing case is when images do not have attached text. An example of this situation is when users are interested in searching the database using example images as queries. A new image without text can be projected to the semantic space by finding the pseudo-inverse of the transformation matrix (S^+) .

$$t = S^+v; \quad S^+ = \left(S^\top S + \beta I\right)^{-1} S^\top \ . \tag{4}$$

where, v is the visual representation of the new image, t is the semantic representation and β is a regularization parameter. In this way we can searching the database using an inferred text representation based in its visual features. This pseudo-inverse matrix has to be preprocessed only once and storing in memory, making very efficient the process of projection for a new image. Finally, the ranking function for semantic search is based on the histogram intersection similarity[17].

4 Experiments and Results

4.1 Datasets

The performance of the proposed algorithm was evaluated using three different datasets with different sizes:

Carcinoma dataset. The Carcinoma dataset is a histopathology image collection that has been used to diagnose a special kind of skin cancer known as basal-cell carcinoma

[5]. It is composed of 1,502 images that were studied and annotated by pathologists to highlight various tissue structures and relevant diagnostic information, elaborating a list with 18 terms. These images were acquired at various magnification levels, including 8X, 10X and 20X, and stored at 1280 × 1024 pixels. The list of keywords includes terms like micro-nodules, elastosis, and fibrosis, among others.

Histology Dataset. The Histology dataset is composed of 2,641 images extracted from an atlas of histology for the study of the four fundamental tissues [19]. The collection includes photographs of histology in different magnification factors (10X, 20X and 40X). The resolution of these images is about 800 × 500 pixels. Each of these images was annotated by an expert, indicating the biological system and organs that can be observed. The total number of different keywords in this data set is 46.

MIRFlickr 25000 Dataset. The MIRFlickr-25000 image dataset is composed of 25,000 pictures downloaded from the popular online photo sharing service Flickr. These photos were collected directly from the web, to provide a realistic dataset for image retrieval research, with high-resolution images and associated metadata [12]. This image collection has been manually annotated using a set of 38 semantic terms.

4.2 Experimental Setup

We conducted retrieval experiments under the query-by-example paradigm. In all datasets 20% of images were randomly selected as queries and the remaining images were used as the target collection to find relevant images. We performed automatic experiments by sending a query to the system and evaluating the relevance of the results. A ranked image in the results list is considered relevant if it shares at least one keyword with the query. The evaluation was done using traditional measures of image retrieval, including precision at 10 and mean average precision (MAP).

Image Features. In all datasets we build a bag-of-features representation, with the following characteristics: Patches of 8 × 8 pixels are extracted from a set of training images with an overlap of 4 pixels along the x and y axes. The DCT (Discrete Cosine Transform) transform is applied in each of the 3 RGB channels to extract the largest 21 coefficients. (DCT-based visual codewords has been found to be an effective representation for microscopy image analysis [9]). A k-means clustering is applied to build a dictionary. For Carcinoma and Histology datasets we use 500 visual terms and for MIRFlickr we select a dictionary of 2000 features (larger dictionaries do not provide significant improvements, but just more computational load). Once the vocabulary has been built, every image in the collection goes through the patch extraction process. Each patch from an image is linked to one visual term in the dictionary using a nearest neighbor criterion. Finally, the histogram of frequencies is constructed for each image.

Text Annotations. In these data sets the text annotations are clean and clearly defined terms from a technical vocabulary and these represent directly the semantic space. We build semantic vectors following a boolean approach, assigning 1 to the terms attached to an image and 0 otherwise. This leads to 46-dimensional binary vectors, for text representation in the Histology dataset, 18-dimensional binary vectors for Carcinoma dataset and 39-dimensional binary vectors for Flickr.

4.3 Retrieval Performance

In order to evaluate the performance of the proposed algorithm, we compare the proposed online algorithm with the classical NSE and the MMCR (Modified Multi-stage Convex Relaxation) proposed by Hsan et. al [18]. Although the MMCR algorithm was proposed mainly for annotation, it is possible to use its semantic score vector as a new representation for retrieval task.

Parameter Tuning. The proposed algorithm has a set of parameters that can impact the quality of the resulting model. Improper settings of these parameters may cause the algorithm converge slowly or diverge. So, as preliminary evaluation, we perform an exploration of these parameters by retrieval experiments using cross-validation 10 fold in the subset of 80% of the images that were not selected as queries. And, we select the configuration that perform better in average in all folds (Table 1).

Table 1. Results of parameter tuning for Online Non-negative Semantic Embedding (ONSE)

Carcinoma				Histology				MIRFlickr			
λ_0	γ	β	**Mini-batch size**	λ_0	γ	β	**Mini-batch size**	λ_0	γ	β	**Mini-batch size**
2^{-5}	2^{-2}	2^4	16	2^{-6}	2^{-3}	2	16	2^{-8}	2^{-10}	2	32

Once, we had found the better configuration, we evaluate the proposed algorithm with the remaining 20% of images as test. So we use this 20% of images as queries and the 80% as finding objective. Table 2 summarizes the findings of our experimental results. In all cases, a general improvement over visual baseline (direct visual matching using visual representation) is shown in MAP measure. And, with the exception of the Histology dataset NSE, ONSE-KL and MMCR algorithms, present a very similar performance.

Table 2. Image retrieval performance. Reported measures are Mean Average Precision (MAP) and Precision at the first 10 results (P@10).

Algorithm	Carcinoma		Histology		MIRFlickr	
	MAP	P@10	MAP	P@10	MAP	P@10
Visual	0.2236	0.3503	0.2107	0.6104	0.2505	0.4931
MMCR [18]	0.3146	0.3322	0.5346	0.6030	0.3670	0.5063
NSE [19]	0.3265	0.3249	0.4025	0.4148	0.3672	0.5079
ONSE	0.3171	0.3651	0.3594	0.4439	0.3674	0.5065

4.4 Computational Load

Table 3 shows the average time consumption for the training phase. Reported times are the result of running all algorithms 5 times in a computer with 4 GB of ram memory and a CPU at 2.4Ghz using only one core. The size of each dataset is also reported to observe how the algorithm complexity grows. NSE algorithm take about 5 seconds to process the Carcinoma dataset, 9 to process the Histology collection and finally increases to

Table 3. Time consumption in training phase: Time required for each epoch (Epoch Avg. Time) and the total average time required until convergence (Total Avg. Time). The algorithm presented in this paper (ONSE) is compared against MMCR [18] and NSE [19].

Dataset	Size	Algorithm	Epochs	Epoch Avg. Time (sec)	Total Avg. Time (sec)
Carcinoma	1502	MMCR	8	0.2854	2.1878
		NSE	130	0.0411	5.3442
		ONSE	**4**	**0.0836**	**0.3345**
Histology	2641	MMCR	10	1.5351	14.2029
		NSE	90	0.1009	9.0869
		ONSE	**4**	**0.3027**	**1.2086**
MIRFlickr	25000	MMCR	10	283.4327	2834,3278
		NSE	200	2.4701	494.017
		ONSE	**2**	**13.755497**	**27.2188**

494 seconds for MIRFlickr. MMCR have the most time consuming, requiring about 2 seconds for Carcinoma 14 for Histology and 2834 for MIRFlickr. In contrast, the ONSE algorithm only requires 0.3 seconds for Carcinoma, 1.2 for Histology and 27 for MIRFlickr. Thus for MIRFlickr dataset, ONSE algorithm is 18 times faster than NSE and 104 times faster than MMCR.

The main reason for the reduction of training time, is, that the number of required epochs until the ONSE algorithm converges is reduced drastically (convergence in all algorithms is verified by means of a minimum threshold required to improve the error in each epoch). For instance, in the carcinoma dataset the NSE algorithm required 130 full scans to the training set and the online version only needed 4. In general, Bottou [3] shows that for a small collection, it is necessary to use very few epochs and for large collections, one full scan is enough. Furthermore, the proposed algorithm reduces the memory requirements, since the only element necessary to keep in memory is the transformation matrix, since visual and textual samples used in each update can be discarded,.

5 Conclusions

We presented an approach for large image indexing that takes advantage of text annotations to provide a semantic representation space where the visual content of images is embedded. This approach is a reformulation of NSE as an online learning algorithm allowing to deal with large collections of data, achieving a significantly reduction in memory requirements and computational load, but keeping a competitive retrieval performance.

Acknowledgements. This work was partially funded by projects Anotación Automática y Recuperación por Contenido de Imágenes Radiológicas Usando Semántica Latente, No. 110152128803 by Colciencias 0521-2010, Diseño e implementación de un sistema de cómputo sobre recursos heterogéneos para la identificación de estructuras atmosféricas en predicción climatológica, No. 1225-569-34920 by Colciencias 0213-2013 and Multimodal Image Retrieval to Support Medical Case-Based Scientific Literature Search, ID R1212LAC006 by Microsoft Research LACCIR.

References

1. Barnard, K., Duygulu, P., Forsyth, D., De Freitas, N., Blei, D.M., Kandola, J., Hofmann, T., Poggio, T., Shawe-Taylor, J.: Matching words and pictures. JMLR 3, 1107–1135 (2003)
2. Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: COMPSTAT 2010, Paris, France, pp. 177–187. Springer (August 2010)
3. Bottou, L., LeCun, Y.: Large scale online learning. In: NIPS (2003)
4. Caicedo, J.C., BenAbdallah, J., González, F.A., Nasraoui, O.: Multimodal representation, indexing, automated annotation and retrieval of image collections via non-negative matrix factorization. Neurocomput. 76(1), 50–60 (2012)
5. Caicedo, J.C., Cruz, A., Gonzalez, F.A.: Histopathology image classification using bag of features and kernel functions. In: Combi, C., Shahar, Y., Abu-Hanna, A. (eds.) AIME 2009. LNCS, vol. 5651, pp. 126–135. Springer, Heidelberg (2009)
6. Caicedo, J.C., González, F.A.: Online matrix factorization for multimodal image retrieval. In: Alvarez, L., Mejail, M., Gomez, L., Jacobo, J. (eds.) CIARP 2012. LNCS, vol. 7441, pp. 340–347. Springer, Heidelberg (2012)
7. Chandrika, P., Jawahar, C.V.: Multi modal semantic indexing for image retrieval. In: CIVR 2010, pp. 342–349. ACM, New York (2010)
8. Chen, Q., Tai, X., Jiang, B., Li, G., Zhao, J.: Medical image retrieval based on latent semantic indexing. In: CSSE 2008, pp. 561–564. IEEE Computer Society, Washington, DC (2008)
9. Cruz-Roa, A., Caicedo, J.C., González, F.A.: Visual pattern mining in histology image collections using bag of features. AIME 52(2), 91–106 (2011)
10. Fang, C., Torresani, L.: Measuring image distances via embedding in a semantic manifold. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part IV. LNCS, vol. 7575, pp. 402–415. Springer, Heidelberg (2012)
11. Guillaumin, M., Mensink, T., Verbeek, J., Schmid, C.: Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In: ICCV (2009)
12. Huiskes, M.J., Lew, M.S.: The mir flickr retrieval evaluation. In: MIR 2008. ACM, New York (2008)
13. Jen Lin, C.: Projected gradient methods for non-negative matrix factorization. Raport Instytutowy, Neural Computation (2007)
14. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: NIPS, pp. 556–562. MIT Press (2000)
15. Makadia, A., Pavlovic, V., Kumar, S.: A new baseline for image annotation. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 316–329. Springer, Heidelberg (2008)
16. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. TPAMI 22(12), 1349–1380 (2000)
17. Swain, M.J., Ballard, D.H.: Color indexing. IJCV 7, 11–32 (1991)
18. Tsai, M.-H., Wang, J., Zhang, T., Gong, Y., Huang, T.S.: Learning semantic embedding at a large scale. In: ICIP, pp. 2497–2500 (2011)
19. Vanegas, J.A., Caicedo, J.C., González, F.A., Romero, E.: Histology image indexing using a non-negative semantic embedding. In: Müller, H., Greenspan, H., Syeda-Mahmood, T. (eds.) MCBR-CDS 2011. LNCS, vol. 7075, pp. 80–91. Springer, Heidelberg (2012)
20. Weston, J., Bengio, S., Usunier, N.: Large scale image annotation: Learning to rank with joint word-image embeddings. In: ECML (2010)
21. Yang, Z., Zhang, H., Yuan, Z., Oja, E.: Kullback-leibler divergence for nonnegative matrix factorization. In: Honkela, T. (ed.) ICANN 2011, Part I. LNCS, vol. 6791, pp. 250–257. Springer, Heidelberg (2011)