Combining generic and specific information for cross-modal retrieval

Thi Quynh Nhi Tran, Hervé Le Borgne, M. Crucianu

Combining Generic and Specific Information for Cross-modal Retrieval

Thi Quynh Nhi Tran
CEA-LIST,
Vision & Content Engineering Laboratory
Gif-sur-Yvette, France
thiquynhnhi.tran@cea.fr

Hervé Le Borgne
CEA-LIST,
Vision & Content Engineering Laboratory
Gif-sur-Yvette, France
herve.le-borgne@cea.fr

Michel Crucianu
CEDRIC-CNAM
Paris, France
michel.crucianu@cnam.fr

ABSTRACT

Cross-modal retrieval increasingly relies on joint statistical models built from large amounts of data represented according to several modalities. However, some information that is poorly represented by these models can be very significant for a retrieval task. We show that, by appropriately identifying and taking such information into account, the results of cross-modal retrieval can be strongly improved. We apply our model to three benchmarks for the text illustration task and find that the more data has misrepresented information, the more our model is comparatively effective.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Retrieval Models*

General Terms

Algorithms, Experimentation

Keywords

Cross-modal retrieval; text illustration; canonical correlation analysis

1. INTRODUCTION

This paper focuses on cross-modal retrieval, where queries from one (or several) modalities can be matched to entries from one (or several) others. Recent work addressed this issue by attempting to develop models where all the involved modalities are mapped to a latent space that can be directly employed in several retrieval scenarios including text illustration or image annotation. Canonical correlation analysis (CCA), introduced by [6], allows to find such a latent space and was successfully employed or extended in many recent proposals. Let us consider a sample, e.g. a set of images, represented by two sets of variables (or two "views"), e.g. the visual features and the associated keywords. CCA will find linear combinations of variables from each set that are maximally correlated to each other. These linear combinations form a latent space where data issued from any of the two sets of variables (views) can be represented.

CCA was first applied to cross-modal retrieval in [5], where a kernel extension (Kernel CCA, KCCA) was also introduced in order to allow for more general, nonlinear latent spaces. Since not all the words (or tags) annotating an image have equal importance, [7] proposed a method taking advantage of their importance when building the KCCA representation. The importance of words for an image is obtained from the order of words in the annotations provided by users for that image. Gong et al. [4] put forward a multi-view (K)CCA method: a third view, explicitly representing image's high-level semantics, is taken into account when searching for the latent space. This "semantic" view corresponds to ground-truth labels, search keywords or semantic topics obtained by clustering tags. In [1], Costa Pereira et al. proposed semantic correlation matching (SCM), where the mappings of images and text by CCA are transformed into semantic vectors produced by supervised classifiers with respect to predefined semantic classes. As an alternative to CCA-based approaches to cross-modal retrieval, [8] introduced a regularized manifold alignment method. Recently, a deep architecture combining representation learning and correlation learning supporting cross-modal retrieval was proposed in [2].

The development of a latent joint representation for two or more views relies on extracting statistical regularities from a preferably large amount of data. Any piece of data having very few occurrences in the training set or very weak relations with other data is ignored in the resulting joint model. However, in a *retrieval* context, pieces of data that are rare in the training set or even new in the test set, such as names or trademarks, can be very significant in selecting the relevant results. It is then necessary to extend the retrieval framework beyond the joint model in order to be able to include "non regular but likely to be relevant" information. One should be able to identify such information (distinguish it from noise) and find ways to combine it with the evidence provided by the joint model. We consider these issues in the specific application context of text illustration tasks.

The proposed approach is described in Section 2. Section 3 presents the experimental results, followed by a discussion. One of the experiments considers domain transfer,

emphasizing the need to make use of information that may be absent from the training data.

2. PROPOSED APPROACH

2.1 Canonical Correlation Analysis

For data simultaneously represented in two different vector spaces, CCA [6, 5] finds maximally correlated linear subspaces of the initial spaces. Consider two random variables, X and Y, taking values in $\mathbb{R}^{d_X}$ and respectively $\mathbb{R}^{d_Y}$. Consider n samples $\{(x_i, y_i)\}_{i=1}^{n} \subset \mathbb{R}^{d_X} \times \mathbb{R}^{d_Y}$. CCA simultaneously seeks directions $w_x \in \mathbb{R}^{d_X}$ and $w_y \in \mathbb{R}^{d_Y}$ that maximize the correlation between the projections of x onto w_x and y onto w_y,

$$w_x^*, w_y^* = \underset{w_x, w_y}{argmax} \frac{{w_x}^T C_{XY}\, w_y}{\sqrt{{w_x}^T C_{XX}\, w_x\, {w_y}^T C_{YY}\, w_y}} \quad (1)$$

where C_{XX}, C_{YY} denote the autocovariance matrices of X and Y respectively, while C_{XY} is the cross-covariance matrix. The solutions w_x^* and w_y^* are eigenvectors of $C_{XX}^{-1} C_{XY} C_{YY}^{-1} C_{YX}$ and respectively $C_{YY}^{-1} C_{YX} C_{XX}^{-1} C_{XY}$. The eigenvectors corresponding to the d largest eigenvalues yield basis $W_x = \left[w_x^1, w_x^2, .., w_x^d\right]$ in $\mathbb{R}^{d \times d_X}$ and $W_y = \left[w_y^1, w_y^2, .., w_y^d\right]$ in $\mathbb{R}^{d \times d_Y}$ that define projections onto the desired "common" space.

2.2 Specific information and cross-modal retrieval

Consider a reference base of multi-modal documents $\{(x_j, y_j)\}_{j=1}^{p}$ that is *a priori* different from the learning base used to compute the CCA model. Cross-modal retrieval essentially consists in estimating the similarity between a query q and all the documents of this reference database. The query can be represented by one modality only (x_q or y_q). In CCA-based cross-modal retrieval, the similarity is computed after projection onto the "common" subspace.

Let us now investigate the consequence of poorly represented data is such a scheme. For instance, consider that the modality $Y = (V, S)$ is a textual feature vector that is composed of subspace V of the vocabulary well represented by the training data and a subspace S of *specific* vocabulary that is *infrequent* in training data, thus poorly represented by the CCA model. Since data in S is infrequent, we assume the cross-covariance matrix is null ($C_{VS} = C_{XS} \approx 0$) and the autocovariance of S is the identity ($C_{SS} = I_S$). With such approximations, the projection matrix becomes

$$W_y = \begin{bmatrix} W_v & 0 \\ 0 & I_s \end{bmatrix} \quad (2)$$

where W_v is the projection matrix obtained by CCA on (X, V) and I_S the identity matrix on S. Hence, the retrieval process within the CCA subspace relies on the (cosine) similarity between a query $y_q = \binom{v_q}{s_q}$ and a document described by $y_d = \binom{v_d}{s_d}$:

$$\underset{CCA(X,Y)}{Sim}(y_q, y_d) = \frac{< W_y^T y_q, W_y^T y_d >}{\left\|W_y^T y_q\right\| \left\|W_y^T y_d\right\|} = \frac{y_q^T W_y W_y^T y_d}{\left\|W_y^T y_q\right\| \left\|W_y^T y_d\right\|} \quad (3)$$

Combining (2) and (3) results into:

$$\underset{CCA(X,Y)}{Sim}(y_q, y_d) = \underset{CCA(X,V)}{Sim}(v_q, v_d) + \frac{s_q^T \cdot s_d}{\left\|W_v^T v_q\right\| \left\|W_v^T v_d\right\|} \quad (4)$$

On the right hand side, the first term is the similarity according to a CCA model computed on well-represented data. In the second term, the impact of S data is biased by the CCA-based denominator. To fix this, we propose to remove the CCA-based weighting from the second term and use a boolean model for this specific information. In other words, we keep only $s_q^T \cdot s_d$ with s_i a binary vector. So the second term simply reflects the number of common specific dimensions (corresponding to infrequent words). However, with such a model, when two documents share the same number of specific dimensions ($s_q^T \cdot s_{d1} = s_q^T \cdot s_{d2}$) their relative similarity to the query only depends on the first term (CCA similarity), that may be inaccurate in this case since the documents have specific dimensions. Hence, we propose to weight the second term by a better adapted measure of similarity, given by the well-known TF-IDF model. Finally, the model we propose (denoted by CCA*) can be written:

$$\underset{CCA^*}{Sim}(y_q, y_d) = \underset{CCA(X,V)}{Sim}(v_q, v_d) + s_q^T \cdot s_d \cdot \underset{TF-IDF}{Sim}(v_q, v_d) \quad (5)$$

When $y = \binom{v}{s}$ does not contain any specific dimension ($s = 0$) our model is equivalent to the classic CCA-based retrieval model. On the contrary, when $s \neq 0$ the second term may become dominant in the similarity estimation. Note that our model supports cross-modal retrieval since the similarity in the CCA space can be estimated from the projection of any feature (x or y). For instance:

$$\underset{CCA^*}{Sim}(y_q, x_d) = \underset{CCA(X,V)}{Sim}(v_q, x_d) + s_q^T \cdot s_d \cdot \underset{TF-IDF}{Sim}(v_q, v_d) \quad (6)$$

3. EXPERIMENTAL EVALUATION

The proposed method for automatic text illustration is evaluated on the BBC News dataset [3] and on a larger Wikipedia collection. We start by comparing our method with several baselines on the BBC News illustration task. Then, we specifically study the impact of information that is absent from the training data by considering a *domain transfer* context: training is performed on an independent dataset and testing on the BBC news dataset. We end up by a comparison on the Wikipedia collection, where the proposed method makes use of a large amount of specific information in comparison with the first two experiments.

3.1 Experimental Setting

BBC News dataset was introduced in [3] for image annotation and text illustration. It consists of 3121 articles for training and 240 for testing, downloaded from the BBC News website. Each article is accompanied by an image and associated caption. This dataset is especially challenging for text illustration because of its small size and the quite indirect relation between textual and visual content in most news articles. For many articles, other images in the collection can be objectively considered more relevant to the article's content than the image selected by the author.

Wikipedia 2010. We use the set of 106,822 articles in English from the ImageCLEF2010 Wikipedia dataset[1]. Each article is accompanied by an image and associated caption. We take 106,582 samples for training and 240 for testing to be consistent with the former benchmark.

Evaluation method. We employ the evaluation methodology based on top-1 accuracy proposed in [3]. For a query

[1]http://imageclef.org/2010/wiki

article, the system is expected to rank first the image that was selected by the author of the article. The reported accuracy is thus the percentage of successfully matched image-article pairs in the test set. We discuss the strictness of this evaluation in Section 3.5.

Content representation. To represent visual content we use OverFeat [9], widely known to provide powerful features for several classification tasks. More precisely, we employ 3072-dimensional vectors which are the layer-18 outputs at the stage 6 of the fast OverFeat network and further L2-normalize them. To represent texts, we construct the dictionary by removing stop words, stemming the remaining words and filtering the stems by their frequency. The BBC News vocabulary $\mathcal{V}_{bbc}$ has 23,617 words that are either stems appearing at least twice in the training set or proper nouns from this set. The Wikipedia vocabulary $\mathcal{V}_{wp}$ has 19,653 words, each appearing at least 5 and at most 1500 times. Texts have a TF-IDF representation, L2-normalized.

Baselines. The proposed method is compared with three text illustration methods presented in [3]. The first two baselines disregard visual content. *Overlap* selects the image whose caption has the largest number of words in common with the test document. In *Vector Space Model (VSM)*, articles and captions are represented by TF-IDF vectors and the cosine similarity measure is used to find the image whose caption is most similar to the test article. A vocabulary of about 6,300 words (as in [3]) or larger is used for texts. *mixLDA*[3] considers both visual and textual content in defining a latent space. The method consists in computing the probability of each visual term in the visual vocabulary to a given text query through hidden topics and delivering a ranked list of relevant visual terms for the query. The image having the highest overlap with the top 30 visual terms in the list is considered as the best image to illustrate the text.

Notations. We denote by CCA_{cap} (CCA_{img}) the basic CCA illustration model in which document-to-caption (document-to-image) nearest neighbor search with cosine similarity measure is applied on document and caption (image) projections. The models in Eq.(5) and Eq.(6) are denoted by CCA^*_{cap} and CCA^*_{img} respectively. In our experiments, CCA spaces are constructed from images and text that cumulates documents (articles) and captions.

3.2 Results on BBC News

We first compare the basic CCA model to the three baselines. CCA dimension selection is performed via 10-fold cross-validation on the 3121-articles *training* set of BBC News. We retain 2500 as the projection dimension since it corresponds to the highest score, see Fig. 1. As shown in Table 1, the corresponding accuracy of CCA_{cap} on the test set is 76.7%, higher than the accuracy of all the baselines including ones presented in [3] as well as the Vector Space $\text{VSM}_{\mathcal{V}_{bbc}}$ model with the larger $\mathcal{V}_{bbc}$ vocabulary. The large improvement of $\text{VSM}_{\mathcal{V}_{bbc}}$ (73.8%) over $\text{VSM}_{\mathcal{V}_{[3]}}$ (38.7%) highlights the importance of an appropriate text representation for this task. The fact that the textual baseline $\text{VSM}_{\mathcal{V}_{bbc}}$ yields a high illustration performance, together with the weak score of CCA_{img} (5.4%), are indications that in the BBC News dataset the visual and textual contents are rather poorly related, so the latent representation learned by CCA is not so reliable. Meanwhile, the VSM model can easily take advantage of the connection between documents and captions, that appears to be strong.

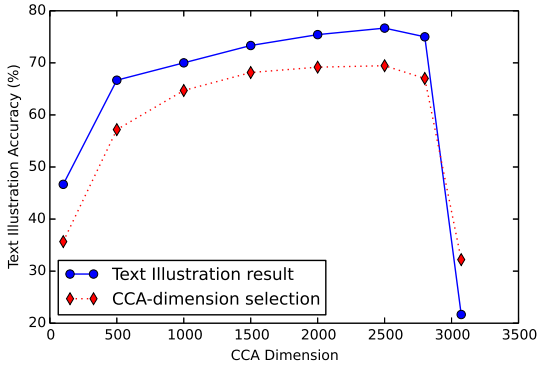

Figure 1: CCA_{cap} result and dimension selection

Table 1: Results on BBC News

Model	Accuracy(%)	10-fold CV Accuracy(%)
Overlap [3]	31.3	-
$\text{VSM}_{\mathcal{V}_{[3]}}$ [3]	38.7	-
mixLDA [3]	57.3	-
$\text{VSM}_{\mathcal{V}_{bbc}}$	73.8	72.8 ± 2.1
CCA_{img}	5.4	8.9 ± 1.7
CCA_{cap}	76.7	74.7 ± 2.5
CCA^*_{img}	15.8	19.0 ± 2.0
CCA^*_{cap}	80.8	78.9 ± 2.4

To apply the CCA* model we proposed in Section 2, we identify the vocabulary $\mathcal{S}_{bbc}$ of 41 specific words that are present in 240 testing captions but not in the initial vocabulary $\mathcal{V}_{bbc}$. As shown in Table 1, our CCA^*_{cap} model achieves the best score with 80.8% accuracy. The third column in Table 1 reports the results of proposed models using a 10-fold cross-validation on 3,361 BBC News data. The results show that the CCA^*_{cap} model outperforms the others.

3.3 Results in a domain transfer context

For the second experiment we develop the latent CCA space using the large ImageCLEF 2013 Photo Annotation and Retrieval dataset[2] and we evaluate text illustration on the same BBC News test set of 240 documents. Our aim is to study the impact of having a larger difference between the vocabularies of the test set and of the training set, which is an important issue in practical applications. The ImageCLEF 2013 collection includes 250,000 images downloaded from the Internet. Each image has at most 100 tags (words) extracted from the content of the web page where the image appears. The textual vocabulary $\mathcal{V}_{ic}$ for the ImageCLEF 2013 training dataset contains 18,003 words occurring each between 5 and 1500 times. The CCA space has 3072 dimensions here and is learned from images and tags using OverFeat features of size 3072 and TF-IDF textual representations. In this domain transfer context, there are 208 specific words $\mathcal{S}_{ic}$ present in the BBC News test set captions but not in $\mathcal{V}_{ic}$. Table 2 reports the performance of three models in this context: $\text{VSM}_{\mathcal{V}_{ic}}$, CCA_{cap} and CCA^*_{cap}. By taking the specific information into account, our model significantly improves the result of basic CCA, showing that it can be quite effective in a domain transfer context. However, while CCA^*_{cap} with the latent space obtained on Image-

[2]http://www.imageclef.org/2013/photo/annotation

Table 2: Results on BBC News with domain transfer

Model	Accuracy(%)
$\text{VSM}_{\mathcal{V}_{ic}}$	53.8
CCA_{cap}	43.3
CCA^*_{cap}	61.3

Table 3: Results on Wikipedia 2010

Model	Accuracy(%)
$\text{VSM}_{\mathcal{V}_{wp}}$	20.8
CCA_{img}	9.2
CCA_{cap}	16.3
CCA^*_{img}	58.3
CCA^*_{cap}	55.4

CLEF 2013 works better than the state-of-the-art mixLDA on BBC News (57.3%, see Table 1), it is weaker than the CCA^*_{cap} with the model learned on the proper BBC News training corpus (80.8%). This reveals that the two datasets, ImageCLEF 2013 and BBC News, present rather different relations between images and words.

3.4 Results on Wikipedia 2010

For the third experiment we consider the Wikipedia 2010 collection that can be directly employed for text illustration while being larger than the BBC News dataset. The results of the Vector Space model baseline, of the basic CCA model and of the proposed CCA* model are shown in Table 3. As for the first experiment, the CCA space is obtained from images and texts that cumulate documents and captions, using the features described in Section 3.1. From the testing set captions, we determined 2868 specific words out of the training vocabulary $\mathcal{V}_{wp}$. Our approach CCA^*_{cap} improves the text illustration accuracy over basic CCA from 16.25% to 55.4% and significantly outperforms the Vector Space model (20.8%). The cross-modal model CCA^*_{img} of Eq. (6) with visual features obtains even better results (58.3%).

3.5 Discussion

The method we proposed aims to make better use of specific information that is poorly represented by a learned cross-modal model but nevertheless likely to be relevant for retrieval. In the text illustration experiments presented above, this information corresponds to words that are present in the test data but absent from the training data (or below the filtering threshold). This selection condition may appear weak, but for large datasets the training vocabulary covers most of the common words, so the words that are new in the test set are mostly names, trademarks or other very informative tags.

An evaluation based on the top-1 result alone is quite strict for text illustration. The system must return as first relevant result the very image chosen by the author, but it may not be the best one to illustrate the document. Actually, in many cases, for both BBC News and Wikipedia 2010, other images in the collection are at least as relevant for the document. It is then important to also evaluate the methods based on the accuracy of top-k results, with $k > 1$. As shown in Table 4 for $k = 10$, the proposed method compares well with the Vector Space model on both datasets.

Table 4: Results with top-10 evaluation

Model	BBC News	Wikipedia
VSM	94.2	27.1
CCA^*_{cap}	95.0	70.8

4. CONCLUSION

We proposed a new approach for CCA-based cross-modal retrieval that takes advantage of information that is poorly represented in the training data but likely to be relevant for the task. We have shown its interest in the context of a challenging text illustration task formulated as top-1 cross-modal retrieval. The new approach was compared to others on a previously published benchmark and shown to produce better results. We also proposed two new benchmarks that are more realistic in the sense that they contain more data that is new in the test set with respect to the training set. The results show that the proposed method improves even more effectively over the performance of CCA in these cases.

5. AKNOWLEDGEMENT

This work is partially supported by the Datascale project, funded by the French Ministère de l'economie, des finances et de l'industrie and the USEMP FP7 project, funded by the EC under contract number 611596.

6. REFERENCES

[1] J. Costa Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. Lanckriet, R. Levy, and N. Vasconcelos. On the role of correlation and abstraction in cross-modal multimedia retrieval. *TPAMI*, 36(3):521–535, 2014.

[2] F. Feng, X. Wang, and R. Li. Cross-modal retrieval with correspondence autoencoder. In *Proc. of ACM Intl. Conf. on Multimedia*, MM '14, 2014.

[3] Y. Feng and M. Lapata. Topic models for image annotation and text illustration. In *Human Language Technologies: 2010 Annual Conf. of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 831–839, 2010.

[4] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV*, 106(2):210–233, Jan. 2014.

[5] D. R. Hardoon, S. R. Szedmak, and J. R. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Comput.*, 16(12):2639–2664, Dec. 2004.

[6] H. Hotelling. Relations between two sets of variables. *Biometrika*, 28:312–377, 1936.

[7] S. J. Hwang and K. Grauman. Learning the relative importance of objects from tagged images for retrieval and cross-modal search. *IJCV*, 100(2):134–153, Nov. 2012.

[8] X. Mao, B. Lin, D. Cai, X. He, and J. Pei. Parallel field alignment for cross media retrieval. In *Proc. of ACM Intl. Conf. on Multimedia*, MM '13, 2013.

[9] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *CoRR*, abs/1312.6229, 2013.