Aggregating image and text quantized correlated components

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

Aggregating Image and Text Quantized Correlated Components

Thi Quynh Nhi Tran
CEA, LIST and CEDRIC-CNAM
thiquynhnhi.tran@cea.fr

Hervé Le Borgne
CEA, LIST
Gif-sur-Yvette, France
herve.le-borgne@cea.fr

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

Abstract

Cross-modal tasks occur naturally for multimedia content that can be described along two or more modalities like visual content and text. Such tasks require to "translate" information from one modality to another. Methods like kernelized canonical correlation analysis (KCCA) attempt to solve such tasks by finding aligned subspaces in the description spaces of different modalities. Since they favor correlations against modality-specific information, these methods have shown some success in both cross-modal and bi-modal tasks. However, we show that a direct use of the subspace alignment obtained by KCCA only leads to coarse translation abilities. To address this problem, we first put forward a new representation method that aggregates information provided by the projections of both modalities on their aligned subspaces. We further suggest a method relying on neighborhoods in these subspaces to complete uni-modal information. Our proposal exhibits state-of-the-art results for bi-modal classification on Pascal VOC07 and improves it by over 60% *for cross-modal retrieval on FlickR 8K/30K.*

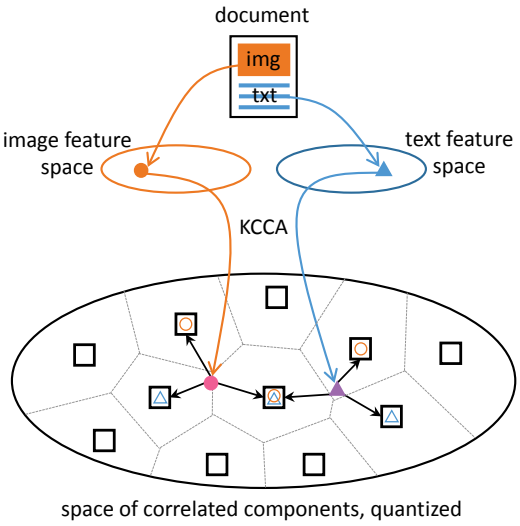

Figure 1. Visual and textual contents of a document are projected onto a common space that has been previously quantized. Both points, corresponding to the same document, are encoded according to a common vocabulary before their aggregation.

1. Introduction

An increasing number of multimedia documents are described along two or more modalities that convey partly common and partly complementary information. This gives the opportunity to devise rich multimedia representations that support both multi-modal and cross-modal tasks. For example, images have a visual content but may also have associated textual data (keywords or sentences). In bi-modal image classification, visual and textual content are employed together for solving the task. Cross-modal tasks like text illustration or image annotation require instead to "translate" information from one modality to another.

Simple fusion approaches such as early fusion (i.e. concatenation of visual and textual features) have been extensively employed, with some success, in bi-modal tasks. But to address cross-modal tasks it was necessary to devise methods that are able to link the two modalities more closely. This is accomplished through the development of a common, latent representation space resulting from a maximization of the relatedness between the different modalities. The methods typically rely on Canonical Correlation Analysis or its kernel extension [10, 14, 5, 9] and on deep learning [20, 26, 8, 7, 18, 27].

Given a set of documents described along two different modalities like image and text, Kernel Canonical Correlation Analysis (KCCA) aims to find maximally correlated manifolds in the feature spaces associated to the two modalities. These manifolds are seen as a common space, where each multimedia document is represented by the two projections of its visual and respectively textual features. This approach was shown to be quite effective [10, 14, 12, 5, 9] and

some recently proposed neural networks for cross-modal tasks [7, 27] also make use of KCCA.

While mainly considered for cross-modal tasks, a common representation space also has the potential to improve the results obtained in bi-modal tasks. For images described by both a visual and a textual content, bi-modal tasks typically focus on semantics. The common representation space favors inter-related information that usually highlights semantics and discounts modality-specific information. Using features from the common representation space instead of early fusion features can then be seen as a form of regularization that reduces the risk of overfitting.

However, by observing the distribution of data projected on the common representation space obtained with KCCA, we found that this space only provides a very coarse association between modalities. For any given document, the projections of its visual and respectively textual features fall far apart. A direct use of these projections results in limited quality "translation" between modalities.

To deal with this problem, we put forward a new representation method for the projections on the common space, called Multimedia Aggregated Correlated Components (MACC). It aims to reduce the gap between the projections of visual and textual features by embedding them in a local context reflecting the data distribution in the common space. Given a database of multimedia documents, we first perform KCCA and build a codebook from all the projections of visual and textual features on the KCCA common space. Subsequently, for each multimedia document, visual and textual features are projected on this common space, then coded using the codebook and eventually aggregated into a single MACC vector that is the multimedia representation of the document (see Figure 1).

When one modality is missing from the initial description of a document, we further propose a method for completing its MACC representation using data from an auxiliary dataset. First, its nearest neighbors in KCCA space, according to the *available* modality, are found in this auxiliary data. Then, the descriptions of the nearest neighbors according to the *other* modality are combined and complete the MACC representation of the target document.

We show that MACC representations allow to reach state-of-the-art performance in a bi-modal task (image classification on Pascal VOC07) and in a cross-modal task (cross-modal retrieval on FlickR 8K and FlickR 30K). We also find that the representation completion method supports an interesting novel usage consisting in training classifiers on data from one modality and testing them on data from the other.

The remainder of this paper is organized as follows. The related work section reviews the usage of KCCA in the recent literature for addressing either cross-modal or bi-modal tasks. After a brief reminder of KCCA, we focus on the construction of MACC representations, involving an aggregation of the projections of visual and textual content represented on a common vocabulary. The method we devised for completing the MACC representation when data is missing for one of the modalities is also presented. The evaluation in Section 4, conducted on three datasets, concerns both image classification and cross-modal retrieval. We briefly highlight the deficiency of KCCA-based representations that motivated our introduction of MACC representations. Several experiments are then presented, supporting comparisons to the state of the art and to several baselines, but also allowing to explore the impact of some key parameters. We end up with the conclusion and a few directions for future work.

2. Related Work

In the recent literature, various (K)CCA-based approaches have been proposed to deal with either cross-modal or bi-modal tasks. CCA was first applied to cross-modal retrieval in [10], where its kernel extension KCCA was also introduced in order to allow for more general, non-linear latent spaces. Since not all the words (or tags) annotating an image have equal importance, [14, 15] proposed a method taking advantage of their importance when building the KCCA representation. The importance of a word for an image is obtained from the order of words in the annotations provided by users for that image. Gong *et al*. [9] 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. This first group of approaches focus on investigating complete representations of data for building a robust common space. Nevertheless, they directly use the projections of the textual and visual descriptors the KCCA common space in order to perform cross-modal tasks.

Approaches in a second group aim to build upon these direct projections on the KCCA common space. Specifically, Costa Pereira *et al*. [5] proposed semantic correlation matching (SCM), where the projections of image and text features by (K)CCA are first transformed into semantic vectors produced by supervised classifiers with respect to predefined semantic classes. These vectors are then used for cross-modal retrieval. Ahsan *et al*. [1] employed the concatenation of textual and visual KCCA-descriptors as inputs of a clustering algorithm to perform a bi-modal task, social event detection. Our proposal follows this second group of approaches. The novelty of our work compared to existing methods is to build a common vocabulary for image and text on the KCCA space and to represent multimedia documents by aggregating their visual and textual descriptors defined on this common vocabulary.

3. MACC: Multimedia Aggregated Correlated Components Representation

We briefly remind in Section 3.1 the theoretical foundations of CCA and its kernelized version KCCA. In Section 3.2, we describe a new representation of multimedia documents relying on an aggregation of the projections of visual and textual content defined on a common vocabulary. Since (K)CCA aims to find a projection space where the correlation between modalities is maximized, we named this new representation "Multimedia Aggregated Correlated Components" (MACC). In Section 3.3 we propose an extension for completing the MACC representations of documents for which only one modality is available. While MACC addresses problems with the representation of bimodal documents, this extension focuses on actual cross-modal cases.

3.1. Kernel Canonical Correlation Analysis

For data simultaneously represented in two different vector spaces, CCA [10] finds maximally correlated linear subspaces of these spaces. Let X^T and X^I be two random variables, taking values in $\mathbb{R}^{d_T}$ and respectively $\mathbb{R}^{d_I}$. Consider N samples $\{(x_i^T, x_i^I)\}_{i=1}^N \subset \mathbb{R}^{d_T} \times \mathbb{R}^{d_I}$. CCA simultaneously seeks directions $w_T \in \mathbb{R}^{d_T}$ and $w_I \in \mathbb{R}^{d_I}$ that maximize the correlation between the projections of x^T onto w_T and of x^I onto w_I,

$$w_T^*, w_I^* = \arg\max_{w_T, w_I} \frac{{w_T}' C_{TI} w_I}{\sqrt{{w_T}' C_{TT} w_T {w_I}' C_{II} w_I}} \quad (1)$$

where C_{TT}, C_{II} denote the autocovariance matrices of X^T and X^I respectively, while C_{TI} is the cross-covariance matrix. The solutions w_T^* and w_I^* are eigenvectors of $C_{TT}^{-1} C_{TI} C_{II}^{-1} C_{IT}$ and respectively $C_{II}^{-1} C_{IT} C_{TT}^{-1} C_{TI}$. The d eigenvectors associated to the d largest eigenvalues define maximally correlated d-dimensional subspaces in $\mathbb{R}^{d_T}$ and respectively $\mathbb{R}^{d_I}$. Even though these are linear subspaces of two different spaces, they are often referred to as "common" representation space.

Kernel CCA (KCCA, see *e.g.* [10]) aims to remove the linearity constraint by using the "kernel trick" to first map the data from each initial space to the reproducing kernel Hilbert space (RKHS) associated to a selected kernel and then looking for correlated subspaces in these RKHS. KCCA seeks vectors of coefficients $\alpha_T, \alpha_I \in \mathbb{R}^N$ that allow to define these maximally correlated subspaces. α_T, α_I are solutions of

$$\alpha_T^*, \alpha_I^* = \arg\max_{\alpha_T, \alpha_I} \frac{{\alpha_T}' K_T K_I \alpha_I}{V(\alpha_T, K_T)\, V(\alpha_I, K_I)} \quad (2)$$

where $V(\alpha, K) = \sqrt{\alpha^t \left(K^2 + \kappa K\right) \alpha}$, $\kappa \in [0, 1]$ is a regularization parameter and K_T, K_I denote the $N \times N$ kernel matrices obtained from $\{x_i^T\}_{i=1}^N$ and $\{x_i^I\}_{i=1}^N$.

3.2. Aggregation of textual and visual information in the projection space

Let us consider a document with a textual and a visual (image) content. A feature vector x^T is extracted from its textual content and another feature vector x^I from the visual one. In what follows, we assimilate a document to a couple of feature vectors (x^T, x^I). A set of such data is a set of couples $\mathbf{X} = \{(x_i^T, x_i^I), i = 1 \dots N\}$. By applying KCCA to this data, as explained in Section 3.1, we obtain $2N$ points (vectors) belonging to a "common" vector space where the two modalities are maximally correlated. In this space, a document (x^T, x^I) is represented by two points, p^T that is the projection of x^T and p^I the projection of x^I. Ideally, since they represent the same document, p^T and p^I should be closer to each other than to any other point in the projection space. However, in practice, this is far from being the case as shown in Section 4.3. It is thus quite problematic for a given document to be represented by two (very) distinct points for multimedia recognition tasks. We propose to create a unified representation for each document, by the following process:

1. define a unifying vocabulary in the projection space,
2. describe both p^T and p^I according to this vocabulary,
3. aggregate both descriptions into a unique representative vector of the document.

Simply said, the "unified vocabulary" is obtained by quantizing the projection space, then p^T and p^I are projected to this codebook and sum pooled to get the final representation. Since it is well known that in computer vision devil is in the details [2, 3], this is further explained below.

Codebook learning. As for the bag of words (BoW) model, we learn a codebook $\mathcal{C} = \{c_1, .., c_k\}$ of k codewords with k-means directly in the projection space. A crucial point is that *all* the projected points, coming from both textual and visual modalities, are employed as input to the k-means algorithm. Hence, the clustering potentially results into three types of codewords (that are centers of the clusters). Some are representative of textual data only, others of visual data only, while some clusters contain both textual and visual projection points. The codebook is thus intrinsically cross-modal and can serve as "common vocabulary" for all the points in the projection space, whether they result from the projection of a textual content or of a visual one.

MACC representation. A bi-modal document (x^T, x^I) is projected on the KCCA projection space of dimension d into (p^T, p^I). Each of these points is then encoded by its *differences* with respect to its nearest codewords:

$$v_i^T = p^T - c_i; \quad c_i \in NN^n(p^T) \quad (3)$$

$$v_i^I = p^I - c_i; \quad c_i \in NN^n(p^I) \quad (4)$$

where $i = 1, .., k$ and $NN^n(p)$ denotes the set of the n nearest codewords of p. The modality-specific representations v^T and v^I result from the concatenation of the d-dimensional vectors v_i^T and respectively v_i^I. The MACC representation v is then obtained by aggregating the visual and textual descriptors v^I, v^T by sum pooling, leading to:

$$\begin{aligned} v &= [v_1, v_2, \ldots, v_i, \ldots, v_k] \quad s.t. \\ v_i &= (p^T - c_i)\mathbb{1}_{NN^n(p^T)}(c_i) + (p^I - c_i)\mathbb{1}_{NN^n(p^I)}(c_i) \end{aligned} \tag{5}$$

where $\mathbb{1}_A(.)$ is the indicator function. Vector v is subsequently L2-normalized. The projection space obtained with KCCA has dimension d, so the *modality-specific encoded vectors* v^T and v^I, as well as the MACC vector v, have a size of $D = d \times k$, where k is the size of the codebook $\mathcal{C}$.

The vectors v^T and v^I are component-wise differences of p^T and p^I with some codewords. When $n = 1$, such a gradient can be seen as a simplified non-probabilistic version of a Fisher Vector (FV) representation. The FV representation is itself an extension of the BoW model resulting from a Maximum Likelihood estimation of the gradient with respect to the parameters of a Gaussian Mixture that models the log-likelihood of data used to learn the codebook [16]. However, in our case we show in the experimental Section 4 that choosing $n > 1$ is advantageous. In some cases, the best results are even obtained with $n = k$. With respect to the vocabulary of a BoW model [2], we could say that [16] uses a *hard coding* ($n = 1$) while we prefer *soft coding* ($n = k$) or possibly *local soft coding* ($1 < n < k$). The benefits of soft coding are well known in the BoW context [13] but have not been proven in the context of FV-like signatures (*i.e.* when one uses component-wise gradients with respect to the codebook).

There is also another advantage in our context, where some codewords may be representative of "modality-specific" Voronoi cells, *i.e.* clusters that contain projected points of only one modality after k-means (see Section 4.3). Therefore, by encoding p^T and p^I according to several codewords, it is more likely to include information from both modalities. Hence, the "modality vectors" v^T and v^I are not exactly modality-specific since they benefit from a sort of "modality regularization" with the multimodal codebook. Yet another advantage is that if p^T and p^I are close enough then they certainly share one or several nearest codewords. These codewords will then be enforced by Eq. (5) in the final vector v.

All this indicates that the MACC representation is a *soft synthesis* of the contributions of both modalities that compensates for the imperfection of the KCCA projection space in the context of bi-modal tasks.

3.3. MACC completion with the missing modality

The MACC representation proposed in the previous section is defined when the multimedia document it describes has both a visual and a textual content. But this condition does not hold for several important multimedia tasks. In particular, for cross-modal tasks, data in the reference base and/or the query usually come from only one modality. In such a case, we estimate MACC representations by completing uni-modal data with suitable information that concerns the missing modality and is obtained from an *auxiliary dataset*.

Modality completion. Consider an auxiliary dataset containing m documents where both visual and textual contents are present. Let $\mathcal{A}$ be the set of pairs of KCCA projections of the visual and textual features of these documents on the common space, with $\mathcal{A} = \{(q^T, q^I)\}$, $q^T \in \mathcal{A}^T$, $q^I \in \mathcal{A}^I$, $|\mathcal{A}| = m$. In practice, the auxiliary dataset could be the training data used to obtain the KCCA space.

To explain the completion process, let us consider a document with textual content only, described by a feature vector x^T that is projected as p^T on KCCA space. The same development could be symmetrically applied to a document having only visual content. A "naive" choice would be to combine p^T with a vector obtained from its μ nearest neighbors among the points projected from the other modality (visual modality in this case), $NN^\mu_{\mathcal{A}^I}(p^T)$. Preliminary experiments (not reported here) have shown that such a strategy is far from being optimal. We propose instead to find the auxiliary documents having similar projected content in the *available* modality (textual modality in this case) and to use the projections of the visual content of these documents to complete p^T. Formally, the set of contributors to the "modality complement" of p^T is defined as

$$\mathcal{M}_c(p^T) = \{q_j^I\} \quad \text{such that} \quad \begin{cases} q_j^T \in NN^\mu_{\mathcal{A}^T}(p^T) \\ (q_j^T, q_j^I) \in \mathcal{A} \end{cases} \tag{6}$$

where the condition $(q_j^T, q_j^I) \in \mathcal{A}$ means that q_j^T and q_j^I are the projections of two feature vectors extracted from the *same* multimedia document. Note that $\left|\mathcal{M}_c(p^T)\right| = \mu$.

MACC representation with the completed modality. Once the complementary information regarding the missing modality has been collected on the KCCA space as $\mathcal{M}_c(p^T)$, the MACC representation of the initially textual-only document is obtained as

$$\begin{aligned} v =& [v_1, v_2, \ldots, v_i, \ldots, v_k] \quad s.t \\ v_i =& (p^T - c_i)\,\mathbb{1}_{NN^n(p^T)}(c_i) \\ &+ \frac{1}{\mu} \sum_{q_j^I \in \mathcal{M}_c(p^T)} (q_j^I - c_i)\mathbb{1}_{NN^n(q_j^I)}(c_i) \end{aligned} \tag{7}$$

4. Experimental Evaluation

We first describe the datasets and the visual/textual features employed. Then we highlight the limits of the KCCA projection, justifying the need of MACC representations.

Our contribution is then evaluated for bi-modal and also *cross-modal* classification on PascalVOC 07. Finally, we show that MACC establishes a new state of the art in cross-modal retrieval, improving former results on FlickR 8K (+11 pts R@1) and FlickR 30K (+15.4 pts R@1).

4.1. Datasets and evaluation metrics

Pascal VOC07 [15]. This dataset includes 5011 training and 4952 testing images collected from Flickr without their original user tags. Twenty class labels were defined and each image receives between 1 and 6 positive labels. Using Amazon Mechanical Turk, in [15] each image also received several tags. The classification results are evaluated using mean Average Precision (mAP), following the literature.

FlickR 8K [22] and **FlickR30K** [29]. These datasets contain 8000 and 31783 images respectively. Each image was annotated by 5 sentences using Amazon Mechanical Turk. These datasets have the same 1000 images for validation and 1000 images for testing. While the training set of FlickR 8K contains 6000 images, the one of FlickR 30K is much larger containing 29783 images. We employ the evaluation metric proposed in [4] for image retrieval with textual query: the five sentences annotating a given image are used together to retrieve images and we report the Recall@K, *i.e.* the fraction of times the ground-truth image is found among the top K images.

4.2. Feature extraction

To represent visual content we use the 4096-dimensional features of the Oxford VGG-Net [24], L2-normalized. This representation was shown to provide very good results in several classification and retrieval tasks.

To represent texts (sets of tags or sentences, respectively) we employ the features built from Word2Vec [19], an efficient method for learning vector representations of words from large amounts of unstructured text data. In our experiments, textual features are 300-dimensional L2-normalized vector representations.

4.3. Limitations of KCCA projections

As previously mentioned in Section 1, the common representation space obtained with KCCA only provides a coarse association between modalities. Several data analysis results shown here highlight this problem.

Table 1 reports several average distances between KCCA projections of the training data (10022 points in Pascal VOC07 and 12000 points in FlickR 8K). We denote by $d_{\text{intramodality}}(I)$ and $d_{\text{intramodality}}(T)$ the average within-modality distances between image and respectively text projected points. Next, $d_{\text{intermodality}}(\text{sample})$ is the average distance between visual projection and associated textual projection on the KCCA space of a training sample, while $d_{\text{intermodality}}(\text{overall})$ is the average distance between visual and textual projections over all training data. The values obtained in Table 1 show that projected points are closer to their within-modality neighbors than to their corresponding points in the other modality.

Average Distance	Pascal VOC07	FlickR 8K
$d_{\text{intramodality}}(I)$	1.18 ± 0.16	1.17 ± 0.13
$d_{\text{intramodality}}(T)$	1.11 ± 0.19	0.75 ± 0.13
$d_{\text{intermodality}}(\text{sample})$	1.39 ± 0.07	1.02 ± 0.12
$d_{\text{intermodality}}(\text{overall})$	1.42 ± 0.06	1.28 ± 0.10

Table 1. Average distances between projections on KCCA space.

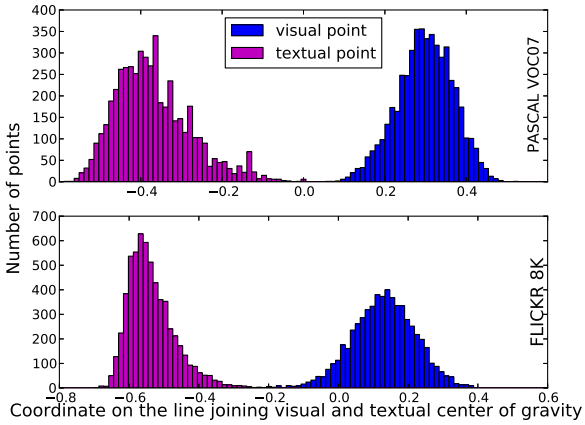

Figure 2. Separation between modalities on the KCCA space.

Dataset	# projected points	k	# visual clusters	# textual clusters
Pascal VOC07	10022	16	12	4
FlickR 8K	12000	8	6	2

Table 2. Distribution of textual and visual KCCA-projected points into clusters.

For a better visualization, we computed the centers of gravity of the visual and respectively textual points, then projected all the points onto the line that joins these two centers. In Figure 2, we report the distribution of these projected points. The separation in the KCCA space between data points from the two modalities appears very clearly, for both Pascal VOC07 and FlickR 8K datasets.

Given this separation between modalities on the common space, the clusters we obtain with k-means contain mostly data from a single modality (image or text). Table 2 shows the number of clusters associated to each modality in Pascal VOC07 and FlickR 8K. They are qualified as "visual" or "textual" according to the majority of points they contain, but each cluster can have both visual and textual points. The clusters are used for codebooks in the following experiments. The value of k is chosen on a validation set.

Approach	mAP (%)
BoVW	54.5
FV [23]	63.9
improved FV[3]	68.0
BoMW [30]	67.8
AGS [6]	71.1
FV+CNN [21]	76.2
He *et al.* [11]	82.4
Chatfield *et al.* [3]	82.4
HCP-2000C [28]	85.2
VGG NetD&NetE [24]	89.7
MACC	**90.12**

Table 3. Pascal VOC07: comparison with published results.

4.4. Image classification on Pascal VOC07

The KCCA is learnt on the 5011 training data, with both visual and textual content. We used the seminal KCCA implementation [10], with a regularization parameter $\kappa = 0.1$ and a Gaussian kernel with standard deviation $\sigma = 0.2$. The dimension of the "common" projected space is set to $d = 150$. All 5011 training data are then projected on this common space and a codebook $\mathcal{C}$ is learnt with k-means from this set ($2 \times 5011 = 10022$ points) for $k \in \{8, 16, 32\}$.

Classification of bi-modal documents. The first evaluation considers the classification of documents having both a visual and a textual content, such that a MACC representation (of size $d \times k$) of each document is directly obtained from Eq. 5, using the previously built codebook. The parameter n in Eq. 5 varies in $\{1, 2, 5, 16, 32\}$. For each category, we learn a SVM classifier with linear kernel, following a one-versus-all strategy.

With such settings, the best result we obtain on the testing set is a mAP of 90.37, with $(k = 16, n = 5)$, resulting into a 2400-dimensional MACC representation. However, when a full cross-validation is conducted on the training set, we obtain a mAP of **90.12** with $(n = 5, k = 32)$. Table 3 compares this performance to other results in the literature. We report superior performance with respect to methods that use only the original (visual) data of the Pascal VOC07 challenge, such as BoVW and Fisher Vectors (FV) [23, 3]. Our approach also outperforms methods employing additional information sources for training, such as text [30], ground-truth bounding box information [6], or based on deep learning [21, 11, 3, 28, 24].

We also compare our image classification result to several baselines that uses the same features as MACC in Table 4. For the VGG-Net (respectively Word2Vec) baseline, classifiers are trained and tested on VGG-Net (respectively Word2Vec) features only, *i.e.* using the visual (respectively textual) content alone. For the VGG-Net+Word2Vec baseline, representations for both training and testing data are obtained by early fusion, *i.e.* by concatenating VGG-Net features and Word2Vec features.

Baseline	Size of representation	mAP (%)
VGG-Net	4096	86.10
Word2Vec	300	82.50
VGG-Net+Word2Vec	4396	86.16
KCCA_{img}	150	84.84
KCCA_{img}	2400	85.29
KCCA_{txt}	150	82.01
KCCA_{txt}	2400	82.60
MACC	2400	**90.12**

Table 4. Pascal VOC07: comparison with baselines.

	k=8	k=16	k= 32
n=1	88.75	87.73	86.33
n=2	90.1	89.71	89.18
n=5	89.96	**90.37**	90.10
n=16	-	89.68	90.33
n=32	-	-	89.68

Table 5. Pascal VOC07: mAP (%) for different values of k and n.

For the KCCA_{img} (respectively KCCA_{txt}) baseline, the visual (respectively textual) features are first projected on the KCCA common space for both training and testing data and then used for classifiers learning. We consider two different sizes of the KCCA common space, 150 and 2400, so that the results can be compared to our 2400-dimensional MACC representation (built from a 150-dimensional common space, with 16 codewords). The results in Table 4 show that the MACC approach outperforms all the mentioned baselines.

We report in Table 5 the results obtained with the MACC approach for different values of k and n (for $d = 150$). We note that the results are quite stable and consistently above the performance of the previously mentioned baselines for this entire range of parameters. Furthermore, these results show that (local) soft coding ($n > 1$) is more effective than hard coding ($n = 1$) to build the MACC representations.

Classification in a cross-modal context. Let us now consider a scenario where a global resource is available, consisting of a projection space obtained by KCCA and a codebook built on this space. One may wish to train classifiers on new classes, using new data for which only *one* modality is available, and then run these classifiers on other data that may also have only one modality available (and maybe not the same as the one used for training). Thanks to the completion mechanism (Eq. 7), the MACC representation addresses not only classical cross-modal tasks but also such a scenario, that is tested in the following.

Specifically, we use the same 150-dimensional projection space obtained by KCCA from the bi-modal training

data of Pascal VOC07 and the codebook learnt on this space ($k = 16$). We then define two new symmetric tasks. In the Text-Image task, the SVM classifiers are trained with documents from the training set of Pascal VOC07 but the visual content was removed. Each document, originally described by its textual content alone, has its MACC representation completed with a visual part following the procedure described in Section 3.3, with the training set of PascalVOC 07 chosen as auxiliary dataset $\mathcal{A}$. Hence, the visual part of the signature is not computed from the original visual content of that document but results from combining the contributions of the visual parts of its nearest neighbors according to the textual modality (the document itself is *not* considered among its μ nearest neighbors). The resulting classifiers are then evaluated on the testing documents of Pascal VOC07 but where the textual content was removed and the MACC representations completed following the procedure in Section 3.3. The Image-Text task is symmetric to the Text-Image task: classifiers are trained with documents without textual content and tested on documents without visual content, all being completed according to Eq. 7.

The results obtained on these two novel tasks are shown in Table 6 for several values of the parameter μ and compared to two baselines. All representations are 2400-dimensional vectors. For the "Random" baseline, the MACC representation is completed with randomly selected data point along the missing modality. For the KCCA$_{inc}$ baseline, classifiers are trained with the projections of one modality on the common KCCA space and tested with the projections of the other modality on this space.

Without completion ($\mu = 0$), the performance of MACC representations is very low. However, as soon as the completion is considered, the performance is significantly above that of the baseline. To our knowledge, no previous work has investigated this type of cross-modal classification scenario on Pascal VOC07, thus there is no other comparison in Table 6. It is not surprising that the results obtained in this cross-modal scenario are not as good as those obtained in the bi-modal task (90.12%, see Table 3). However, the difference is not so large and the improvement with respect to the baselines is significant.

4.5. Image retrieval on FlickR 8K and FlickR 30K

KCCA is learnt on the 6000 training documents with both visual and textual content. To select the parameters, a grid search is performed employing the validation set of 1000 documents. This leads to use a Gaussian kernel with a standard deviation $\sigma = 2$, a regularization parameter $\kappa = 1$ and only $d = 50$ dimensions for the projected space. The visual and textual features of the training documents are then all projected on this common space and a codebook is learnt from this set of 12000 ($= 2 \times 6000$) points.

FlickR 8K image retrieval. For the text-to-image retrieval task the training dataset of FlickR 8K is used as auxiliary dataset $\mathcal{A}$. Parameters being cross-validated on the training data, we get **R@1=27.6%** for $k = n = 32; \mu = 64$. As shown in Table 7, the proposed approach has higher R@1, R@5 and R@10 than the other image retrieval methods in the recent literature on the FlickR 8K dataset. Hodosh *et al*. [12] work was also based on cross retrieval in the KCCA space but their visual and textual representations are simply described by several specific kernels on classical features such as color, texture or GIST descriptors for images, and bag of words for texts. In the KCCA$_{(VGG+W2V)}$ baseline, we apply the image retrieval method of in [12] with our KCCA space built from VGG-Net and Word2Vec features, leading to much better performance (26.1%) than [12]. Our method also significantly outperforms several recent deep learning approaches [25, 17, 4] that use content representation similar to ours. Furthermore, the MACC representation achieves better results than [4], the current state-of-the-art on both FlickR 8K and FlickR 30K image retrieval, in which the VGG-Net features are also employed.

We studied the impact of different coding parameters on the effectiveness of MACC representations. Codebook size being fixed to $k = 64$, Figure 3 reports the performance with hard coding ($n = 1$), local soft coding ($1 < n < k$) and soft coding ($n = k$). Since soft coding provides a better location of data points in feature space (with respect to all k codewords, not only to one or to a few of them), it usually performs better for retrieval. The most important result is nevertheless that our method achieves better per-

	mAP (%) Text-Image	mAP (%) Image-Text
Random	7.76	7.33
KCCA$_{inc}$	71.21	51.20
MACC($\mu = 0$)	12.03	10.04
MACC($\mu = 1$)	79.00	76.88
MACC($\mu = 3$)	81.72	79.18
MACC($\mu = 5$)	82.17	78.82
MACC($\mu = 8$)	82.18	78.65
MACC($\mu = 10$)	82.09	77.97

Table 6. Pascal VOC07: classification in a cross-modal context using the completion mechanism for MACC representations.

Approach	R@1	R@5	R@10
Socher *et al*. [25]	6.1	18.5	29
Hodosh *et al*. [12]	7.6	20.7	30.1
Karpathy *et al*. [17]	11.8	32.1	44.7
Chen *et al*. [4]	17.3	42.5	57.4
KCCA(VGG+W2V)	26.1	53.7	65.6
MACC	**27.6**	**55.6**	**69.4**

Table 7. Image retrieval results on FlickR 8K.

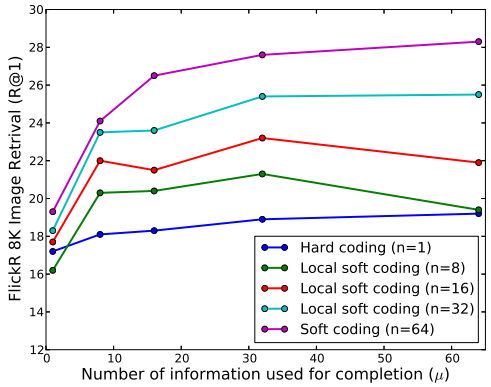

Figure 3. Coding methods comparison for MACC representation.

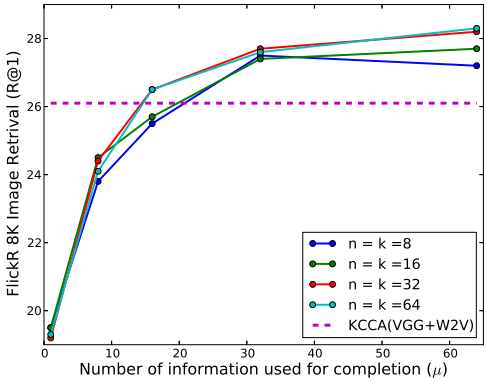

Figure 4. FlickR 8K image retrieval: stability of MACC representations over a large range of parameters.

formance than the state-of-the-art [4] as soon as $\mu > 10$. On the FlickR 8K benchmark, the performances are quite stable with any given coding scheme for $\mu > 20$.

In a third experiment we study the stability of our approach with regard to k, n and μ. Figure 4 reports performance on FlickR 8K while varying these parameters. Following the conclusion of the second experiment, soft coding ($n = k$) is employed in this experiment for its effectiveness. The results firstly show that even when the size of codebook and resulting MACC representations is very small, we consistently achieve better performance than the other methods in Table 7. For instance, our approach has a first rank recall (R@1) of 18.5% with k as low as 8 (the corresponding MACC representation is only 400-dimensional). Besides, an interesting observation is that with a sufficiently large number μ of contributors to MACC completion, the proposed approach yields superior performance over the robust KCCA$_{(VGG+W2V)}$ baseline regardless of the size of the codebook. These results show the stability of our approach over a large range of parameters.

FlickR 30K: benefit of auxiliary dataset. To study the impact of the auxiliary dataset $\mathcal{A}$ used for MACC completion in cross-modal tasks, we conducted an experiment on FlickR 30K that has the same validation and testing sets as FlickR 8K but a larger training set. The experimental protocol was the same as for FlickR 8K (same KCCA space and codebook) except for the choice of $\mathcal{A}$, where we used the full training set of FlickR 30K (29783 images) instead of the training set of FlickR 8K (6000 images). The results in Table 8 show a significant improvement of 5 points for the MACC approach, which is thus due to the larger auxiliary dataset. This improvement is higher that those obtained in previous publications. It increases when the parameters are cross-validated on FlickR30k training set. While the improvement in the previous state-of-the-art [4] is from 17.3% on FlickR 8k to 18.5% on FlickR 30K, in our case it is from 28.3% to 33.9%, showing a better use of the extended training dataset, at a limited cost (KCCA and the codebook are always computed on FlickR 8K).

Approach	R@1	R@5	R@10
Socher *et al*. [25]	8.9	29.8	41.1
Karpathy *et al*. [17]	15.2	37.7	50.5
Chen *et al*. [4]	18.5	45.7	58.1
MACC (F8k)	**33.9**	**65.6**	**77.5**
MACC (F30k)	**35.3**	**66.0**	**78.2**

Table 8. Image retrieval results on FlickR 30K. MACC parameters are cross-validated on FlickR 8k (F8k) or FlickR 30k (F30k)

5. Conclusion and Discussion

We proposed a new representation of a multimedia document that aggregates information provided by the projections of both modalities on their aligned subspaces. We also suggested a method to complete uni-modal information relying on neighborhoods in these subspaces. The interest of our approach was demonstrated in bi-modal classification, cross-modal classification and cross-modal retrieval, where our method provides state-of-the-art performance.

We believe that this approach should also be relevant for other types of joint text-image representations built using *e.g*. Latent Dirichlet Allocation, Partial Least Squares or deep neural networks. However, the scheme we propose already relies on very effective methods (VGG and Word2Vec) to produce uni-modal representations from raw content. The choice of an algorithm to compute the joint representation should be made in compliance with the characteristics of the uni-modal representations employed.

Acknowledgement: this work is supported by the USEMP FP7 project, partially funded by the European Commission under contract number 611596.

References

[1] U. Ahsan and I. Essa. Clustering social event images using kernel canonical correlation analysis. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, CVPRW '14, pages 814–819, Washington, DC, USA, 2014. IEEE Computer Society.

[2] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *British Machine Vision Conference*, 2011.

[3] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv:1405.3531*, 2014.

[4] X. Chen and C. Lawrence Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In *CVPR*, June 2015.

[5] 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.

[6] J. Dong, W. Xia, Q. Chen, J. Feng, Z. Huang, and S. Yan. Subcategory-aware object classification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 827–834. IEEE, 2013.

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

[8] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, pages 2121–2129, 2013.

[9] 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.

[10] 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, 2004.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *TPAMI*, 37(9):1904–1916, 2015.

[12] M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, pages 853–899, 2013.

[13] Y. Huang, Z. Wu, L. Wang, and T. Tan. Feature coding in image classification: A comprehensive study. *TPAMI*, 36(3):493–506, 2014.

[14] 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.

[15] S. J. Hwang and K. Grauman. Reading between the lines: Object localization using implicit cues from image tags. *TPAMI*, 34(6):1145–1158, June 2012.

[16] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid. Aggregating local image descriptors into compact codes. *TPAMI*, 34(9):1704–1716, 2012.

[17] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[18] A. Karpathy, A. Joulin, and F. F. F. Li. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*, pages 1889–1897, 2014.

[19] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546, 2013.

[20] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

[21] F. Perronnin and D. Larlus. Fisher vectors meet neural networks: A hybrid classification architecture. In *CVPR*, June 2015.

[22] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, CSLDAMT '10, pages 139–147, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[23] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *IJCV*, 105(3):222–245, 2013.

[24] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[25] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014.

[26] N. Srivastava and R. R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230, 2012.

[27] W. Wang, R. Arora, K. Livescu, and J. Bilmes. On deep multi-view representation learning. In *International Conference on Machine Learning*, Lille, France, 2015.

[28] Y. Wei, W. Xia, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan. CNN: single-label to multi-label. *CoRR*, abs/1406.5726, 2014.

[29] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014.

[30] A. Znaidia, A. Shabou, H. Le Borgne, C. Hudelot, and N. Paragios. Bag-of-multimedia-words for image classification. In *ICPR*, pages 1509–1512. IEEE, 2012.