Multi-Label Cross-modal Retrieval

Viresh Ranjan[1]* Nikhil Rasiwasia[2] C. V. Jawahar[3]

[1]Virginia Tech [2]Snapdeal.com [3]CVIT, IIIT Hyderabad

Abstract

In this work, we address the problem of cross-modal retrieval in presence of multi-label annotations. In particular, we introduce multi-label Canonical Correlation Analysis (ml-CCA), an extension of CCA, for learning shared subspaces taking into account the high level semantic information in the form of multi-label annotations. Unlike CCA, ml-CCA does not rely on explicit pairings between the modalities, instead it uses the multi-label information to establish correspondences. This results in a discriminative subspace which is better suited for cross-modal retrieval tasks. We also present Fast ml-CCA, a computationally efficient version of ml-CCA, which is able to handle large scale datasets. We show the efficacy of our approach by conducting extensive cross-modal retrieval experiments on three standard benchmark datasets. The results show that the proposed approach achieves state-of-the-art retrieval performance on the three datasets.

1. Introduction

With the huge surge in multimedia content over the internet, coupled by factors such as reduction in the cost of storage devices and high speed internet connectivity, most of the websites now contain rich content pertaining to various modalities such as text, image, video, audio, animation etc. This has led to the development of cross-modal systems, where the goal is to build systems that enable interactivity across content modalities. For example, the design of cross-modal retrieval systems [5, 9, 22, 23, 27, 35, 14, 29, 34, 38, 36, 20], where queries from one modality (e.g., image) can be matched to database entries from another (e.g., text articles).

In recent years, a lot of work has been done in the computer vision community towards the development of cross-modal systems. One popular solution is to learn a common latent space by learning modality-specific projections using paired samples across the two modalities, followed by classification/retrieval operation. This latent space is cross-modal as both the modalities can be represented in this space without any distinction. Several approaches have been proposed to learn such cross-modal spaces [5, 9, 22, 23, 27, 35, 14]. Of these, Canonical Correlation Analysis (CCA) is fast becoming the de-facto standard where the common subspace is learned by maximizing the correlation between the projections of the two modalities.

Figure 1: Figure shows two images from Caltech-256 dataset [6]. The images belong to the classes "camel" (L) and "laptop" (R) respectively. Clearly, both the images can also be labeled with the class "people".

While CCA has been popular for its simplicity and efficiency, it has several drawbacks. First and foremost is the inability of the classic CCA to account for additional high-level semantic information such as the class label of the datapoints. Since class information is not utilized while learning the projection directions, the learned subspace is less discriminative in nature [23, 27]. In the past few years, several works have successfully addressed the above shortcoming by proposing alternatives and extensions of CCA to account for label information [27, 23, 5, 30]. However, most of these strategies assume that the data is annotated with a *single label*. Although the simplest case of additional semantic information is that of single-label, often it is not sufficient to talk about multimedia data belonging to a single class. In fact, even in standard single-label single-modality datasets, it is often the case that a natural image can be assigned to multiple classes. For example, Figure 1, shows two images from popular single-label 'Caltech256' dataset [6]. Although they were labeled with "camel" and "laptop" classes respectively, they can also be labeled with

*This work was done while VR was a student at CVIT, IIIT Hyderabad.

"people" class. Moreover, many large scale datasets such as Imagenet [3] are pre-organized into class hierarchies where an image labeled with the leaf node class can also be labeled with any of its parent nodes. Thus, it is important to design cross-modal systems that naturally account for *multi-label* images. In recent years, several multi-label image datasets have been introduced [4, 24, 2]. Multi-label datasets have also been prevalent in domains as diverse as protein function classification [33, 18], document classification [25], music classification [31] etc.

In this work, we address the problem of cross-modal retrieval in presence of multi-label data. In particular, we propose an extension of CCA to account for additional semantic information which is multi-label in nature. A naive strategy for handling multi-label data would be to ignore multi-label information and to use only the pairwise correspondence (if available) to learn the common subspace. However, as we shall see, discarding multi-label information leads to a drop in performance for cross-modal retrieval tasks. Another alternative is to convert the multi-label dataset into single-label dataset by retaining one label per data item and then relying on various cross-modal techniques suited for single-label datasets. Such an approach is often the choice for tackling multi-label datasets [32]. Again, in addition to the problem of choosing which label to retain, such an operation creates artificial separation between the classes leading to poor performance (see Table 1).

In this work, we propose *Multi-Label Canonical Correlation Analysis* (ml-CCA), a cross-modal retrieval framework for multi-label datasets. ml-CCA utilizes multi-label information while learning a common semantic space for the two modalities, and hence, is able to learn a discriminative semantic space which is more suitable for cross-modal tasks. The highlights of our current work are

1. We present ml-CCA, which utilizes semantic information in the form of multi-label information. ml-CCA does not require explicit pairings across the different modalities. Instead, it uses the semantic information to establish correspondences across the modalities.

2. We also present Fast ml-CCA, a computationally efficient version of ml-CCA, which shows minimal degradation in performance over ml-CCA.

3. We validate our approach on three multi-label datasets, and show significant improvement over related approaches for the cross-modal retrieval task. The proposed approach achieves state-of-the-art results on these benchmark datasets.

Although the fundamental ideas are applicable to any combination of content modalities, we restrict the discussion to multi-label documents containing images and text.

The paper is organized as follows: Section 2 discusses previous work in cross-modal multimedia modeling and multi-label approaches. Section 3 presents a mathematical formulation for the proposed ml-CCA framework. Section 4 summarizes an extensive experimental evaluation designed to test the efficacy of the proposed approach for cross-modal retrieval tasks on benchmark multi-label datasets. Results and conclusions are presented in Section 4 and Section 5 respectively.

2. Related Work

The problem of cross-modal retrieval, for image and text modalities, has been the subject of extensive research in the recent past [5, 9, 22, 23, 27, 35, 14, 29, 34, 38, 36, 20], Cross-modal retrieval has also been studied in other contexts such as corpora of images and audio [16, 37], text and audio [28], or even other sources of data like EEG and fMRI [17].

Although, several techniques have been proposed for the task of cross-modal retrieval, such as, bilinear model, partial least squares [27], etc., in the recent years, CCA has become the workhorse of many of the cross-modal retrieval approaches [23, 9, 5, 7]. These approaches employ CCA / kernel CCA for learning a common subspace where cross-modal queries could be performed. However, most of these approaches do not utilize label information while learning the common subspace, resulting in a subspace which is not discriminative in nature [22]. To the best of our knowledge, use of class labels for cross-modal retrieval tasks was first proposed in [23], where semantic correlation matching was introduced. An extension of CCA incorporating single-label semantic information was proposed in [22], where label information was used to establish correspondences between all possible datapoints within a class, across the two modalities. Several other approaches have been proposed to incorporate single-label class information for the task of cross-modal retrieval [27, 13, 15]. However, it is not clear how any of these approaches can be used for datasets with multiple labels. In fact, since a large portion of multimedia data is multi-label in nature, few of the cross-modal approaches ignore the multi-label information, and use the classic CCA for handling multi-label data [9].

An extension of CCA, which is able to handle multi-label annotations, was proposed in [5]. The high-level semantics, represented either by a single category or multiple non-mutually-exclusive concepts was incorporated as the third view and the three-view formulation of CCA [7] (3-view CCA) was employed to learn the cross-modal subspace. However, as we shall see, multi-label datasets introduce implicit many to many relationships between the data objects. The 3-view CCA is not able to incorporate such relationships and utilizes only a single three way pairing of the data.

3. Multi-Label Cross-Modal Retrieval

In this section we present the proposed multi-label cross-modal retrieval system. As discussed in Section 1, it is often the case that multimedia data (images, text, etc.) cannot be sufficiently represented by single semantic label. Earlier, cross-modal retrieval systems relied on single-label multimodal datasets [22, 27], however in the recent past, several multi-label datasets (see Section 4.1) with multiple modalities have been proposed [10, 2]. As shown in Figure 2, multi-label datasets introduce a natural many-to-many correspondence across different modalities, i.e. each data point from one modality is related to several other data points from the other modality. Any multi-label cross-modal retrieval system should be able to incorporate such relationships while learning the common cross-modal subspace.

In this work, as is popular for cross-modal retrieval systems, we rely on Canonical Correlation Analysis (CCA) as the fundamental approach to learn the shared cross-modal subspace. We propose Multi-Label Canonical Correlation Analysis (ml-CCA) for the task of cross-modal retrieval. ml-CCA takes into account the semantic information available in the form of multiple labels for each data point to learn the cross-modal subspace. Note that although the proposed multi-label cross-modal retrieval system leverages the CCA framework, the fundamental ideas can easily be adapted to other frameworks for learning shared subspaces. Next, we present a brief review of CCA, followed by the detailed description of the proposed ml-CCA framework for multi-label cross-modal retrieval.

3.1. Background

Canonical correlation analysis, first proposed by Hotelling [8], is a strategy for finding basis vectors for two sets of data vectors so that the correlation between the projection of the data vectors along the basis vectors are mutually maximized. Let $x \in \mathbb{R}^p$, and $y \in \mathbb{R}^q$ be two random multivariate vectors. Also, let $S_x = \{x_1, x_2, \ldots, x_n\}$ and $S_y = \{y_1, y_2, \ldots, y_n\}$ be two sets of paired vectors, i.e. x_i in S_x is paired with y_i in S_y. Let $w \in \mathbb{R}^p$ and $v \in \mathbb{R}^q$ be two projection vectors, and the projection of the two sets along w and v be $S_{wx} = (\langle w, x_1\rangle, \langle w, x_2\rangle, \ldots, \langle w, x_n\rangle)$, and $S_{vy} = (\langle v, y_1\rangle, \langle v, y_2\rangle, \ldots, \langle v, y_n\rangle)$. CCA aims at finding projection vectors w and v so that the correlation between S_{wx} and S_{vy} is maximized, i.e.,

$$\rho^* = \max_{w,v} corr(S_{wx}, S_{vy}), \tag{1}$$

where $corr(S_{wx}, S_{vy})$ is the correlation between S_{wx} and S_{wy}, and ρ^* is the maximum correlation. The above optimization problem can be expressed in terms of covariance matrices of the data points as

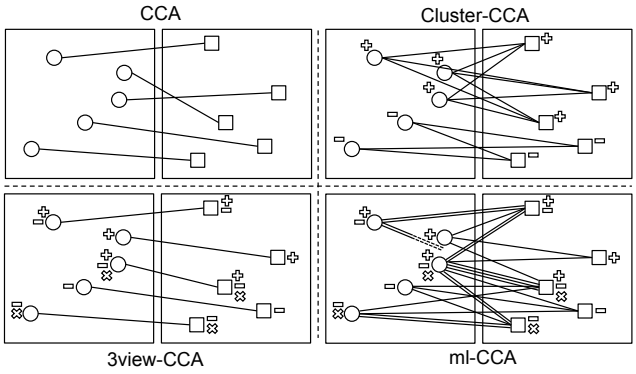


Figure 2: shows the nature of correspondence in CCA, cluster-CCA, 3-view CCA and ml-CCA. The circles and squares denote the datapoint in the two modalities respectively, and '+', '-', 'x' denote the class labels. In CCA, each sample in first set is paired with a single sample in the second set. In cluster-CCA, each point in one modality is paired with all the same class points in the other modality. In 3-view CCA, each sample in first modality is paired with a single sample from the second modality, and both the samples are paired with the underlying label. In ml-CCA, a sample in one set can be paired with multiple samples in the second set. Pairs having similar labels are given more preference in ml-CCA, shown in the figure by multiple connections.

$$\rho^* = \max_{w,v} \frac{w' C_{xy} v}{\sqrt{w' C_{xx} w}\sqrt{v' C_{yy} v}} \tag{2}$$

where $C_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i^T$, $C_{xx} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$ and $C_{yy} = \frac{1}{n}\sum_{i=1}^{n} y_i y_i^T$. The problem is reduced to a generalized eigenvalue problem (see [7] for more details), and can be solved easily. In its original form, CCA does not utilize label information. As a result, the projection directions w, v are not able to incorporate class information and are not discriminative enough for retrieval operation [23].

3.2. Multi-Label Canonical Correlation Analysis

Extensions of CCA have been proposed to incorporate class label [23, 27] information for cross-modal retrieval tasks, nevertheless, they do not fully address the multi-label setting. In this section, we propose multi-label Canonical Correlation Analysis (ml-CCA) for building cross-modal retrieval system in presence of multi-label data. In ml-CCA, each data point in the first modality is paired with multiple data points in the second modality and the contribution of each pairing is decided by the similarity between the corresponding multi-label vectors.

Let x_i be the ith vector in first modality, and z_i be its label vector, similarly (y_j, z_j) be the jth vector and its corresponding label vector in the second modality. X be a

$p \times n_x$ dimensional matrix whose columns are the observation in the first modality, i.e. $X = [x_1, x_2, ..., x_{n_x}]$, Z_x be the $C \times n_x$ dimensional label matrix whose columns are the label vectors, i.e. $Z_x = [z_{x_1}, z_{x_2}, ..., z_{x_{n_x}}]$. Similarly, $Y = [y_1, y_2, ..., y_{n_y}]$ be a $q \times n_y$ dimensional data matrix and $Z_y = [z_{y_1}, z_{y_2}, ...z_{y_{n_y}}]$ be the $C \times n_y$ dimensional label matrix of the second modality. Since the two modalities can have multiple labels, multiple elements in each column of Z_x and Z_y could be nonzero. Let f be a function which gives similarity between any two label vectors (columns of Z_x and Z_y) (z_i, z_j). To learn the common cross-modal subspace under the above setting, ml-CCA is formulated as:

$$\rho = \max_{w,v} \frac{w' \tau_{xy} v}{\sqrt{w' \tau_{xx} w} \sqrt{v' \tau_{yy} v}}, \tag{3}$$

where,

$$\tau_{xy} = \frac{1}{N} \sum_{i=1}^{n_x} \sum_{j=1}^{n_y} f(z_i, z_j) x_i y_j^T \tag{4}$$

$$\tau_{xx} = \frac{1}{N} \sum_{i=1}^{n_x} \alpha_i x_i x_i^T \tag{5}$$

$$\tau_{yy} = \frac{1}{N} \sum_{j=1}^{n_y} \beta_j y_j y_j^T \tag{6}$$

are the weighted covariance matrices, $N = n_x \times n_y$ is the total number of pairs across the two modalities, $\alpha_i = \sum_{j=1}^{n_y} f(z_i, z_j)$ and $\beta_j = \sum_{i=1}^{n_x} f(z_i, z_j)$. Once the weighted covariance matrices τ_{xy}, τ_{xx} and τ_{yy} are computed, the projection directions w and v can be obtained in a similar manner as that in the case of CCA as

$$\tau_{xx}^{-1} \tau_{xy} \tau_{yy}^{-1} \tau_{yx} w = \lambda^2 w, \tag{7}$$

$$v = \frac{\tau_{yy}^{-1} \tau_{yx} w}{\lambda}. \tag{8}$$

It can be seen from the above equations that unlike CCA and 3-view CCA [5], ml-CCA does not require correspondence across the different modalities.

3.2.1 Similarity Function

We want the similarity function f to assign a high value to the label pair (z_i, z_j) when the labels are similar, and to assign a low value when the two labels are not similar. We consider two types of similarity functions in our current work

- **Dot-product based similarity:** Given two multi-label vectors z_i, z_j, we define the dot-product based similarity function f as

$$f(z_i, z_j) = \frac{< z_i, z_j >}{\|z_i\| \|z_j\|}. \tag{9}$$

- **Squared exponential distance based similarity:** Given two multi-label vectors z_i, z_j, we define the squared exponential based similarity function f as,

$$f(z_i, z_j) = e^{(-\|z_i - z_j\|_2^2 / \sigma)}, \tag{10}$$

where σ is a constant factor.

3.3. Computational Complexity & Fast ml-CCA

Assuming both the modalities have n data points, i.e. $n_x = n_y = n$, the asymptotic time complexity for ml-CCA is $O(n^2 d^2) + O(d^3)$, where $d = max(p, q)$; $O(n^2 d^2)$ for computing the covariance matrices and $O(d^3)$ for matrix multiplication, inverse and eigenvalue decomposition. Clearly, applying ml-CCA to large datasets becomes computationally infeasible because of the quadratic complexity in n.

In order to handle large datasets, we propose an efficient version of ml-CCA, referred to as Fast Multi-Label Canonical Correlation Analysis (Fast ml-CCA). Fast ml-CCA relies on the fact that not all pairs (x_i, y_j) contribute equally to learning of the common cross-modal space. Pairs of (x_i, y_j) for which the labels z_i and z_j are significantly different, the similarity score $f(z_i, z_j)$ will be low and therefore the contribution towards the computation of the weighted covariance matrices will be negligible. Fast ml-CCA builds on this observation and only considers pairs for which the similarity score is high. To this end, an approximate nearest neighbor search [19] is performed over the label space in order to determine (x_i, y_j) pairs with high similarity score $f(z_i, z_j)$. Thus, by using only the k nearest neighbors to compute the weighted covariance matrices, Fast ml-CCA ensues a time complexity of $O(nCKI(\frac{\log n}{\log K})) + O(nd^2 LC(\frac{\log n}{\log K})) + O(d^3)$, where C is the label vector dimension, K is the branching factor (each node at pth level is divided into K clusters at next level), I is the number of iterations in the k-means clustering loop, assuming the tree is balanced, $\frac{\log n}{\log K}$ is the height of the tree, L is the number of data points after which the search for approximate neighbors stop; $O(nCKI(\frac{\log n}{\log K}))$ goes into the building the hierarchical k-means tree, $O(nd^2 LC(\frac{\log n}{\log K}))$ goes into finding nearest neighbors and computing the weighted covariance matrices and $O(d^3)$ for matrix multiplication, inverse and eigenvalue decomposition. Please see [19] for a more detailed treatment of the hierarchical k-means complexity.

As we shall see in Section 5.3, in practice, Fast ml-CCA results in significant reduction in the computational time without any significant loss in retrieval performance, even on large scale datasets such as NUS [2].

3.4. Comparison of ml-CCA with other CCA extensions

3.4.1 Cluster Canonical Correlation Analysis

An extension of CCA for single-label datasets, referred to as cluster-CCA, was proposed in [23]. Cluster-CCA introduces correspondence between each sample from any class in the first modality to all the same class samples in the second modality. Once this correspondence is established, regular CCA is applied to obtain basis vectors for the two modalities (see [23] for more information).

The way it is designed, cluster-CCA can be effective only for datasets which can be separated into distinct clusters (a manifestation of single class labels, see Figure 2). In a multi-label scenario, there is no natural separation of the data into distinct clusters. However, distinct clusters can be forced for multi-label data in several possible ways. We present two different approaches for using cluster-CCA for multi-label datasets, i.e. *single-label cluster-CCA* and *multi-label cluster-CCA*. Single label cluster-CCA is a naive approach where labels can be discarded to retain only one label per data point. For multi-label cluster-CCA, we form distinct clusters by an unsupervised clustering of the data using the label information as features. This provides a mapping between a multi-label vector and a cluster. These discovered clusters are then used as a proxy for supervised single class label. Both these approaches introduce artificial constrains on the dataset, i.e. separation of the dataset into distinct clusters. In Table 1 and Table 2, we show that both these approaches lead to loss in performance, which is sometimes even worse than the classic CCA.

Although cluster-CCA can not be used for multi-label datasets, the proposed ml-CCA can readily be used for datasets with single labels. In fact, in presence of single labels, ml-CCA can be shown to reduce to cluster-CCA. If z_i and z_j are two single label vectors, using the indicator function $\mathbb{1}z_i = z_j$, which takes a value of 1 when the two vectors are equal, as the similarity function f, reduces ml-CCA to cluster-CCA.

3.4.2 3-View Canonical Correlation Analysis

CCA can be extended to handle multiple views [7, 1]. Its three view extension (3-view CCA) has been used for cross-modal retrieval task in a multi-label setting [5], where multi-label vector serves as the third view. However, 3-view CCA requires correspondence information across the three modalities. It cannot be applied for such datasets where correspondence information is not available. Moreover, as shown in Figure 2, in 3-view CCA, each data point in one modality is only paired with a single data point from the other modalities. As pointed out in Section 3, there could be a natural many-to-many correspondence in multi-label datasets. Clearly, 3-view CCA overlooks this correspondence while learning the cross-modal subspace. ml-CCA is free from both of these limitations. It avoids the need for explicitly paired data across the three modalities, and it also establishes multiple correspondences across the modalities, thus utilizing the natural many-to-many correspondence of multi-label datasets.

4. Experiments

In this section, we discuss the different datasets used for the experiments, the feature representation used for representing the images as well as the text/tags modalities and the performance measures used for evaluation. Although the modalities being considered in this work are image and text, ml-CCA can very well be used for other modalities, for instance image and sketch, provided the data points are annotated with multiple concepts.

4.1. Datasets

Evaluation of cross-modal retrieval systems in a multi-label setting requires datasets with two modalities that are also annotated with multiple labels. In this work, we consider datasets with image and text/tags as the two modalities, where semantic labels corresponding to image and text/tags modalities are also available. In particular we present results on three benchmark datasets, viz., NUS [2], Pascal [4] and LabelMe [24]. Next we briefly describe the three datasets.

NUS Wide [2] is a large-scale dataset consisting of $269,648$ data objects. Each datapoint consists of an image, its textual tags and is labeled with 81 underlying semantic concepts (which serve as the class labels). Note that each image could have multiple annotations, i.e. each image could be annotated with several of the 81 semantic concepts. Various low level features, viz., color histogram in LAB color space, color correlogram in HSV color space, block-wise LAB-based color moments, bag of visual words, edge distribution histogram, wavelet texture features, are provided for each image. We use the combination of the first four features for our experiments. For text representation, 1000 dimensional tag features are used. We use the original train-test split provided in the dataset for training and testing, i.e., $161,789$ image-text-label triplets are used for training and remaining $107,859$ for testing.

Pascal VOC 2007 [4] consists of 5011 train and 4952 test images. We use the publicly available bag of visual words, gist [21] and color histogram features provided by [9] for image representation. For tag/text representation, we use the 399 dimensional absolute tag rank features provided by [9]. For label representation, we use the groundtruth annotation of the images. We use the original train-test split provided in the dataset for training and testing.

LabelMe dataset consists of a total of 3825 images collected by [10]. We use the publicly available bag of visual words, gist and color histogram features provided by [9] for image representation. For tag/text representation, we use the 209 dimensional absolute tag rank features provided by [9]. For label representation, we use the groundtruth annotation of the images. We perform a random $50-50$ split of the dataset for creating training and testing sets.

In addition to the original image features provided by the corresponding dataset curators, we also perform experiments using the popular convolutional neural net (CNN) based image features. To the best of our knowledge, CNN based image features have not been evaluated for cross-modal retrieval tasks[1]. CNN features are computed using Caffe [12], using the pre-trained architecture learned on ImageNet. We use the central crop of the image (no mirroring). The 4096 dimensional output from the 'fc7' layers serves as the image feature.

4.2. Performance Measures

The proposed multi-label cross-modal retrieval system is benchmarked using different performance metrics, viz., normalized discounted cumulative gain (NDCG) [11], Precision and mean average precision (MAP). These measures have been popular to measure performance of cross-modal retrieval systems [10, 5, 27, 23, 22, 35, 14]. Precision@K (P@K) measures precision at fixed low levels of retrieved results. It does not take into account the rank order within the top-K retrieved items. NDCG gives graded relevance to retrieved results instead of binary relevance, giving more importance to the top results. Finally MAP score is used to predict the overall performance of the retrieval system.

4.3. Implementation Details

For all the common subspace learning approaches considered in this work, i.e. CCA, cluster-CCA, 3-view CCA and ml-CCA, we fix the dimensionality of the common subspace to be 10. While testing, the query and the test points are projected to the common subspace, and the retrieval performance is measured by comparing the label vector of the query with the label vectors of retrieved test points. Also, for regularization in case of CCA, cluster-CCA and ml-CCA, we add a constant value of 0.2 to the diagonal elements of the covariance matrices before solving the eigenvalue problem. For 3-view CCA, we use the implementation provided by [5]. For cluster-CCA, since the optimal number of clusters over the label space is not known apriori, we vary the number of clusters from 10 to 100 for LabelMe and Pascal, and 10 to 10000 for NUS Wise, and report the best results. The scale parameter(Eqn 10) and optimal number of neighbors in Fast ml-CCA are set by doing a three-fold cross validation on the training set(see supplementary for details). For ml-CCA as well as Fast ml-CCA, we experiment with both the similarity functions given in Section 3.2.1. We find that the squared exponential similarity function gives better results in most scenarios, and hence, all the reported results use this similarity function (see supplementary section for a detailed comparison).

[1]Given the improvements for other vision related tasks [26], we expect similar improvement for cross-modal retrieval systems.

Method	T2I (a)	T2I (b)
CCA [7]	**59.95**	**43.40**
cluster-CCA(single-label) [23]	54.43	39.72
cluster-CCA(multi-label) [23]	55.61	41.36

Table 1: Comparison of CCA (which does not use any label information) with two different adaptations of cluster-CCA, (1) single-label, which converts the multi-label dataset into single-label dataset by retaining one label randomly, and (2) multi-label, where single label is enforced via clustering. T2I stands for text to image retrieval. (a) NDCG@30 and (b) Precision@10 are used as the performance measures. Experiment is performed on the LabelMe dataset [10].

5. Results

In this section, we present the results for ml-CCA and other cross modal retrieval approaches. First, we compare ml-CCA with CCA and its extensions. Next, we compare ml-CCA with other existing cross modal retrieval approaches. Finally, we show results for large scale experiments on the NUS Wide dataset [2].

5.1. Comparison of ml-CCA with CCA and its Extensions

In this section we present a comparison of CCA with its published extensions for multi-label cross-modal retrieval. First, we compare different adaptations of cluster-CCA for the multi-label task as discussed in Section 3.4.1. Table 1 reports the NDCG@30 and Precision@10 scores for CCA and the two adaptations of cluster-CCA [22]. It is clear from the table that when adapted to multi-label setting, both variants of cluster-CCA perform significantly worse than classic CCA, suggesting that the artificial separation introduced by cluster-CCA is detrimental to the task of cross-modal retrieval. Also note that multi-label cluster-CCA performs better than single-label cluster-CCA. Given that single-label cluster-CCA discards most of the label information, this result is expected. Hence, for further experiments, we report results for the multi-label variant of cluster-CCA.

The comparison of ml-CCA with CCA, cluster-CCA and the 3-view CCA is reported in Table 2, for both text-to-image and image-to-text cross-modal tasks. For both these tasks, the relevance of any retrieved object is decided based

Method	Dataset	Image-to-text					Text-to-image				
		Bow	Color	Gist	**Comb.**	**CNN**	Bow	Color	Gist	**Comb.**	**CNN**
CCA	LabelMe	55.22	47.65	53.52	57.91	55.86	55.09	51.29	55.19	59.95	56.85
3-view CCA	LabelMe	42.23	45.73	47.73	56.18	54.82	58.28	**56.15**	54.61	63.22	61.71
cluster-CCA	LabelMe	50.21	42.25	43.68	50.97	57.48	53.19	48.25	47.23	55.61	56.95
ml-CCA	LabelMe	**58.13**	**47.89**	**54.04**	**62.12**	**58.85**	**61.60**	55.36	**58.19**	**65.34**	**61.85**
CCA	Pascal	30.15	23.89	31.62	36.73	63.10	43.62	29.97	42.72	51.10	77.20
3-view CCA	Pascal	24.26	**29.00**	30.34	24.65	59.67	35.20	26.93	39.82	45.55	69.85
cluster-CCA	Pascal	27.30	24.29	27.71	29.41	59.31	37.63	28.86	36.30	41.23	65.96
ml-CCA	Pascal	**32.13**	26.18	**34.00**	**39.42**	**64.74**	**47.23**	**32.36**	**43.44**	**55.68**	**80.04**

Table 2: Performance of CCA, cluster-CCA and ml-CCA is compared for cross-modal retrieval task. NDCG@30 is used as the performance measure.

on the similarity between the labels of query and retrieved item. We use NDCG as the performance measure and consider the top 30 retrieved results for computing the NDCG score. Several observations can be made from the table. First, as observed above, cluster-CCA again is not able to outperform CCA (which does not use any label information) for almost all features and datasets. Second, even 3-view CCA is not able to out-perform standard CCA, probably because it does not account for all possible relationships introduced in a multi-label setting. This again suggests that its better to not use the semantic information in a multi-label setting than to use existing supervised extensions of CCA. Finally, it is clear that ml-CCA outperforms all the other approaches across most features in both datasets. Thus, ml-CCA effectively utilizes the multi-label information to learn the cross-modal subspace. Overall, ml-CCA achieves an average relative gain of 8.1% over CCA using combination of the publicly available features. Table 2 also reports the NDCG score using CNN based image features. ml-CCA again shows an average relative improvement of 5.1% over CCA. In Fig 3, we show few text-to-image query results on Pascal the dataset.

5.2. Comparison with Existing Cross-modal Approaches

In this section, we compare ml-CCA with other cross-modal approaches which account for label information. Most of the other cross-modal approaches can only handle single label data. Such approaches extract the single label image-text pairs from the multi-label datasets, and show results on this single label subset. Table 3 shows the performance of several such approaches where cross-modal retrieval was performed on the Pascal dataset, using publicly available image and text features provided by [10]. For their inability to account for multi-label data, these approaches use only the single label pairs from the Pascal dataset, resulting in a 2808/2841 train/test split.

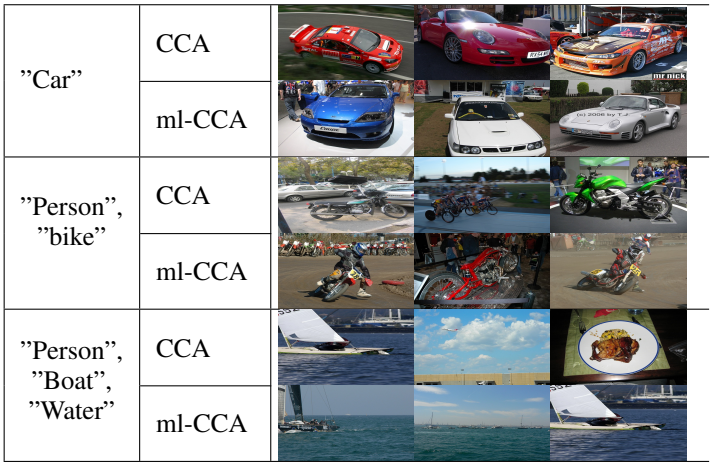
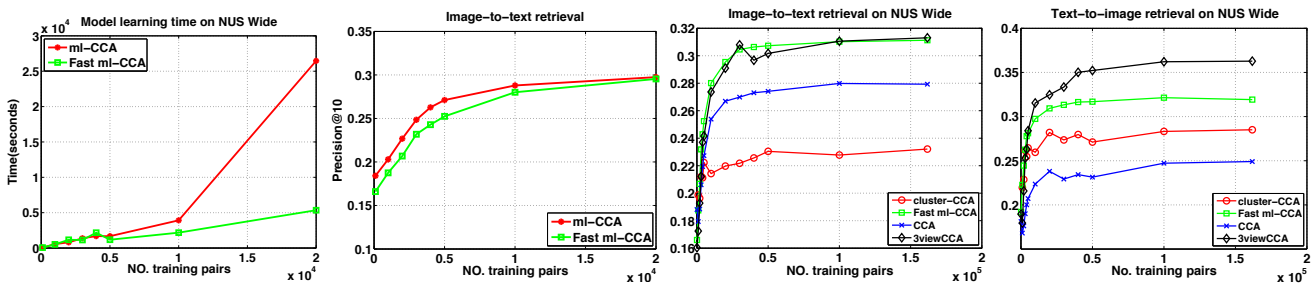


Figure 3: Text queries and top retrieved images obtained using CCA and ml-CCA are shown.

Method	I2T	T2I	Average (MAP)
CCA	42.1	30.6	36.4
3-view CCA	42.3	33.6	37.9
cluster-CCA	41.2	31.6	36.4
KGMMFA [27]	42.1	32.8	37.5
KGMLDA [27]	42.7	33.9	38.3
LCFS [35]	34.4	26.7	30.6
LGCFL [14]	37.8	32.9	35.3
ml-CCA	**48.4**	**38.0**	**43.2**

Table 3: Comparison of ml-CCA with various recent approaches. I2T stand for image to text retreival, T2I stands for text to image retrieval. Mean average precision is used as the performance measure.

Following the same protocol as in [27, 35, 14], Table 3



(a) (b) (c) (d)

Figure 4: Results on NUS Wide. (a,b) compares the time requirement and retrieval performance for ml-CCA and Fast ml-CCA with increase in number of training pairs. (c,d) comparison of ml-CCA with CCA, cluster-CCA and 3-view CCA for image-to-text retrieval and text-to-image retrieval tasks. Precision@10 is used as the performance measure.

also reports the performance of ml-CCA. The test set is exactly the same as used by the rest of the approaches, i.e. 2841 single image-text pairs. However, since ml-CCA can learn from multi-labeled pairs, we also add the multi-labeled image-text pairs to the training set. It is clear from the table that ml-CCA with its multi-label capabilities is able to significantly outperform the rest of the approaches. Note that these multi-label pairs *cannot be added* during training for other approaches because they are designed to work with only single labels. Also note that although KGMMFA and KGMLDA [27] are kernelized projection strategies and ml-CCA is a linear projection approach, still it is able to outperform them[2].

5.3. Fast ml-CCA

In this section, we show results on the NUS Wide dataset. Given the large dataset size of NUS Wide, learning the projection directions using ml-CCA would be computationally expensive. As discussed in Section 3.3, Fast ml-CCA is a faster version of ml-CCA, and hence, is better suited for handling large datasets. Fig 4(a) compares the computation time of ml-CCA and Fast ml-CCA as the number of training samples is increased. For large number of training pairs, it can be clearly observed that the computation time for ml-CCA becomes much larger than the time required for Fast ml-CCA. Fig 4(b) shows the retrieval performance of both the approaches for the NUS dataset. It is clear that Fast ml-CCA is able to match the retrieval performance of ml-CCA at a fraction of the training time.

Similar to the previous datasets, we conduct cross-modal experiments on the NUS Wide. Results corresponding to image-to-text and text-to-image retrieval are presented in Fig 4(c) and (d) respectively. For both cases, Fast ml-CCA outperforms CCA, as well as cluster-CCA. With sufficient number of training pairs, 3-view CCA performs comparably with ml-CCA for image-to-text retrieval task, and slightly better than ml-CCA for text-to-image retrieval task. Both ml-CCA as well as 3-view CCA learn discriminative subspaces. In case of relatively smaller datasets such as Pascal and LabelMe, ml-CCA outperforms 3-view CCA as the latter does not have sufficient number of training pairs to learn sufficiently discriminative subspaces. ml-CCA takes care of this scarcity of training pairs by introducing multiple pairings across the two modalities. In case of large datasets such as NUS Wide, because of the presence of large number of training pairs, 3-view CCA is able to learn more discriminative subspaces leading to competitive results.

[2]Note that we could not compare with a recent multi-label approach [5] since the features are no longer publicly available.

6. Conclusions

In this work, we addressed the problem of cross-modal retrieval problem for multi-label data. We proposed Multi-Label Canonical Correlation Analysis (ml-CCA), a novel extension of CCA, which is able to effectively incorporate multi-label information while learning the cross-modal subspaces. We conducted extensive experimental evaluation of the proposed approach and the results show that ml-CCA achieves state-of-the-art retrieval performance for all three benchmark datasets considered.

References

[1] F. R. Bach and M. I. Jordan. Kernel independent component analysis. *The Journal of Machine Learning Research*, 2003. 5

[2] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, 2009. 2, 3, 4, 5, 6

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009. 2

[4] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 2010. 2, 5

[5] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *International journal of computer vision*, 2014. 1, 2, 4, 5, 6, 8

[6] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. 2007. 1

[7] D. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 2004. 2, 3, 5, 6

[8] H. Hotelling. Relations between two sets of variates. *Biometrika*, 1936. 3

[9] S. J. Hwang and K. Grauman. Accounting for the relative importance of objects in image retrieval. In *BMVC*, 2010. 1, 2, 5, 6

[10] S. J. Hwang and K. Grauman. Reading between the lines: Object localization using implicit cues from image tags. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2012. 3, 6, 7

[11] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 2002. 6

[12] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv*, 2014. 6

[13] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen. Multi-view discriminant analysis. In *Computer Vision–ECCV*. 2012. 2

[14] C. Kang, S. Xiang, S. Liao, C. Xu, and C. Pan. Learning consistent feature representation for cross-modal multimedia retrieval. *IEEE Transactions on Multimedia*, 2015. 1, 2, 6, 7

[15] C. H. Lampert and O. Krömer. Weakly-paired maximum covariance analysis for multimodal dimensionality reduction and transfer learning. In *Computer Vision–ECCV 2010*. 2010. 2

[16] D. Li, N. Dimitrova, M. Li, and I. K. Sethi. Multimedia content processing through cross-modal association. In *Proceedings of the eleventh ACM international conference on Multimedia*, 2003. 2

[17] V. Mahadevan, C. W. Wong, J. C. Pereira, T. Liu, N. Vasconcelos, and L. K. Saul. Maximum covariance unfolding: Manifold learning for bimodal data. In *Advances in Neural Information Processing Systems*, 2011. 2

[18] S. Mostafavi and Q. Morris. Fast integration of heterogeneous data sources for predicting gene function with limited annotation. *Bioinformatics*, 2010. 2

[19] M. Muja and D. G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. *VISAPP (1)*, 2009. 4

[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*, 2011. 1, 2

[21] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 2006. 5

[22] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the international conference on Multimedia*, 2010. 1, 2, 3, 6

[23] N. Rasiwasia, D. Mahajan, V. Mahadevan, and G. Aggarwal. Cluster canonical correlation analysis. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, 2014. 1, 2, 3, 5, 6

[24] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 2008. 2, 5

[25] E. Sandhaus. The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 2008. 2

[26] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv*, 2013. 6

[27] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs. Generalized multiview analysis: A discriminative latent space. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, 2012. 1, 2, 3, 6, 7, 8

[28] M. Slaney. Semantic-audio retrieval. In *Acoustics, Speech, and Signal Processing (ICASSP), IEEE International Conference on*, 2002. 2

[29] N. Srivastava and R. R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, 2012. 1, 2

[30] T. Sun, S. Chen, J. Yang, and P. Shi. A novel method of combined feature extraction for recognition. In *Data Mining, Eighth IEEE International Conference on(ICDM)*, 2008. 1

[31] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. P. Vlahavas. Multi-label classification of music into emotions. In *ISMIR*, 2008. 2

[32] G. Tsoumakas, I. Katakis, and I. Vlahavas. Mining multi-label data. In *Data mining and knowledge discovery handbook*. 2010. 2

[33] K. Tsuda, H. Shin, and B. Schölkopf. Fast protein classification with multiple networks. *Bioinformatics*, 2005. 2

[34] Y. Verma and C. Jawahar. Im2text and text2im: Associating images and texts for cross-modal retrieval. In *BMVC*, 2014. 1, 2

[35] K. Wang, R. He, W. Wang, L. Wang, and T. Tan. Learning coupled feature spaces for cross-modal matching. In *Computer Vision (ICCV), IEEE International Conference on*, 2013. 1, 2, 6, 7

[36] X. Zhai, Y. Peng, and J. Xiao. Learning cross-media joint representation with sparse and semisupervised regularization. 2014. 1, 2

[37] H. Zhang, Y. Zhuang, and F. Wu. Cross-modal correlation learning for clustering on image-audio dataset. In *Proceedings of the 15th international conference on Multimedia*, 2007. 2

[38] X. Zhu, Z. Huang, H. T. Shen, and X. Zhao. Linear cross-modal hashing for efficient multimedia search. In *Proceedings of the 21st ACM international conference on Multimedia*, 2013. 1, 2