This CVPR workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Multi-view Multi-label Canonical Correlation Analysis for Cross-modal Matching and Retrieval

Rushil Sanghavi
IIT Jodhpur, India
sanghavi.1@iitj.ac.in

Yashaswi Verma
IIT Jodhpur, India
yashaswi@iitj.ac.in

Abstract

In this paper, we address the problem of cross-modal retrieval in presence of multi-view and multi-label data. For this, we present Multi-view Multi-label Canonical Correlation Analysis (or MVMLCCA), which is a generalization of CCA for multi-view data that also makes use of high-level semantic information available in the form of multi-label annotations in each view. While CCA relies on explicit pairings/associations of samples between two views (or modalities), MVMLCCA uses the available multi-label annotations to establish correspondence across multiple (two or more) views without the need of explicit pairing of multi-view samples. Extensive experiments on two multi-modal datasets demonstrate that the proposed approach offers much more flexibility than the related approaches without compromising on scalability and cross-modal retrieval performance. Our code and precomputed features are available at https://github.com/Rushil231100/MVMLCCA.

1. Introduction

During the last decade, there has been a significant increase in the amount of digital data in diverse modalities such as images, videos, audio, text, etc. Because of this, cross-modal matching and retrieval has emerged as a promising research problem where given a sample in one modality (or view), the objective is to retrieve semantically similar samples from another modality. E.g., given a query text (a set of words or a caption), retrieve images that are semantically relevant to it. One popular idea to address this problem is to project samples (features) from diverse modalities into a learned common embedding space where similar samples are close to each other, and then perform cross-modal retrieval by comparing them using some conventional similarity measure such as cosine similarity.

Canonical Correlation Analysis (or CCA) [18] is one of the earliest and most popular methods based on this idea. It learns a common embedding space by maximizing the correlation between the projections of features from two modalities. However, CCA has three major limitations: (1) its application is limited to data with only two modalities, (2) it requires explicit pairing/association between cross-modal samples at the time of training, and (3) it is an unsupervised approach and cannot make use of additional semantic information that is generally available in the form of multi-label annotations along with samples. Lately, some of the papers have attempted to address these limitations of CCA. Two promising attempts in this direction are: (1) Multi-view Canonical Correlation Analysis (or MVCCA) [15], which is a generalization of CCA to multi-view data and learns the common embedding space using explicitly associated samples from multiple (two or more) modalities, and (2) Multi-label Canonical Correlation Analysis (or MLCCA) [26], which uses multi-label annotations to determine associations between samples from two modalities rather than requiring explicit associations as in CCA. Both MVCCA and MLCCA have shown compelling results on a variety of cross-modal retrieval tasks compared to CCA. Specifically, in MLCCA, independence from the requirement of explicit associations between samples from two modalities at the time of training and the flexibility to integrate multi-label semantics make it an appealing alternate to CCA. However, one major limitation of MLCCA is that it can be applied when the input data contains samples from only two modalities, and its generalization to multi-modal data with more than two modalities is non-trivial and non-existent in the literature as per our knowledge. On the other hand, while MVCCA can be applied to data with two or more modalities, it requires explicit associations among samples from different modalities at the time of training and cannot make use of multi-label annotations.

To address these limitations, we present Multi-view Multi-label Canonical Correlation Analysis (or MVMLCCA), which is a generalization of CCA to multi-view data (data with two or more modalities), does not require explicit associations among samples from different modalities, and can make use of semantic information available in the form of multi-label annotations at the time of train-

Method	Number of modalities	Samples associated across modalities	Multi-label semantics
CCA [18]	2	✓	✗
MLCCA [26]	2	✗	✓
MVCCA [15]	≥ 2	✓	✗
MVMLCCA (Ours)	≥ 2	✗	✓

Table 1. Comparison of the proposed MVMLCCA with CCA [18] and its extensions MLCCA [26] and MVCCA [15] in terms of (a) how many modalities (or views) they can work with, (b) whether they require supervision in the form of explicit pairing or association among samples from diverse modalities at the time of training, and (c) whether they can use semantic information available in the form of multi-label annotations to compute associations among samples from diverse modalities at the time of training.

ing to determine associations among multi-modal samples (Table 1). Further, our approach is also applicable to the real-world scenarios where the vocabularies of labels are non-overlapping across modalities. In such cases, we make use of real-valued feature representations of labels (e.g., using Word2Vec [23]) to compute similarities between multi-label annotations of samples in different modalities[1].

We validate our approach on two popular multi-modal datasets (IAPRTC-12 [8] and MS-COCO [20]), and demonstrate that it offers scalability and cross-modal retrieval performance that is comparable to the competing methods. It should be noted that while we limit the scope of our discussion and analyses to datasets containing samples from image and text modalities, the proposed technique is applicable to any set of content modalities.

2. Related Work

Cross-modal matching and retrieval is a long-standing research problem, which was first introduced in [18] along with a technique to address this - Canonical Correlation Analysis or CCA. Particularly during the last two decades, there has been a surge of multimedia content on the internet and thus cross-modal matching and retrieval has gained significant attention in various domains such as image-text [13, 15, 21, 26, 28], image-audio [14], text-text [40], etc. Being the first approach to facilitate cross-modal retrieval, a large number of recent techniques are inspired from CCA, including both non-deep learning based as well as deep learning based methods. Non-deep learning based methods [13, 15, 26–28, 31] assume the availability of pre-computed features of samples in different modalities, and then learn a common embedding space using them by introducing novel learning formulations [31, 36], additional information in the form of meta-data/tags [13, 28], or both [5, 26, 27, 32]. While CCA has inspired a large number of cross-modal matching algorithms, another category of methods adopt ranking based optimizations that not only pull positive (semantically similar) pairs closer like CCA-based methods, but also push negative (semantically dissimilar) pairs farther [17, 33]. However, their training stage is generally computationally intensive as they utilize both positive as well as negative pairs. Unlike non-deep learning based methods, deep learning based methods learn both features as well as a common embedding space simultaneously starting with raw data. While the initial deep learning method worked with global features [2, 9–11, 22, 30, 35, 39, 40], several recent methods use local features and align them across diverse modalities using attention mechanism [6, 19, 29, 37]. Some of the recent papers [1, 25] have also explored self-supervised learning to learn a common multi-modal embedding space. While learning features from raw data is not the focus of our work, we believe that any progress in this direction will lead to an improvement in the retrieval accuracy of non-deep methods, as also validated in [26].

From the above discussion, we can observe that both non-deep as well as deep learning based methods have been actively explored in parallel for cross-modal matching and retrieval tasks by the research community, which indicates the utility of both the directions given the widespread applicability of such methods in diverse domains and experimental conditions. Ours is a non-deep learning based approach and is closely related to two popular extensions of CCA: MVCCA [15] and MLCCA [26]. MVCCA [15] is a generalization of CCA to multi-view (two or more views) data, however it cannot use multi-label annotations and additionally requires explicit associations among samples from different modalities. In [13], multi-label annotations were used in an indirect way by considering them as an additional view, and then the original formulation of MVCCA was employed to learn the common embedding space. However, this results in a single multi-way association of samples and thus fails to utilize the many-to-many relationships implicit in multi-label data. In MLCCA [26], rather than considering multi-label annotations as another view, these are considered as a common ground to establish many-to-many associations among samples from diverse modalities, and thus does not require explicit pairing among cross-modal samples at the time of training. However, one major limitation of MLCCA is that it is applicable to only two views and cannot learn a single common embedding space when data contains three or more views. We address this limitation of MLCCA by presenting a generalization of of MVCCA that does not require explicit associations among samples from different modalities and uses multi-label annotations to es-

[1]While label features can also be incorporated in the MLCCA [26] approach to compute similarity between two sets of labels, this was not explored in the original paper. For comparisons in our experiments, we update the implementation of MLCCA provided by its authors.

tablish such associations analogous to MLCCA. However, unlike MLCCA, it is capable to learn a common embedding space using data containing any number of (two or more) views. As our method carries the desired characteristics of both MVCCA and MLCCA, we call it Multi-view Multi-label CCA (or MVMLCCA).

3. Preliminaries

Below, we first discuss the notation and problem set-up that we use in the subsequent parts of the paper, and then we present an overview of MVCCA.

3.1. Notation and Problem set-up

Suppose we are given data $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_n\}$ in n modalities (e.g., audio, video, text, etc.). In the p^{th} modality ($p \in \{1, \ldots, n\}$), the dataset $\mathcal{D}_p = \{(S_p^a, Y_p^a)\}_{a=1}^{m_p}$ consists of m_p pairs of samples and corresponding label-(sub)sets respectively. Each sample S_p^a is represented by a real-valued feature vector $\mathbf{x}_p^a$ in a d_p-dimensional space $\mathbb{X}_p = \mathbb{R}^{d_p}$, and each Y_p^a denotes a set of labels associated with the sample S_p^a such that $Y_p^a \subseteq \mathcal{Y}_p$, where $\mathcal{Y}_p$ denotes the vocabulary of discrete labels that describe the semantic space of the p^{th} modality.

We assume that the semantic labels in all the vocabularies are represented by unique feature vectors in a single/common feature space. Precisely, each label $y \in \mathcal{Y}_p$ ($\forall p$) is represented by a feature vector z in an l-dimensional feature space $\mathbb{Z} = \mathbb{R}^l$, which is common across all the n vocabularies. For example, this space could be a simple one-hot encoding space in case of a common vocabulary across all the modalities, or a Word2Vec [23] or GloVe [24] embedding space in case of partially or non-overlapping vocabularies across modalities. This way, a label-set $Y_p^a = \{y_p^{a1}, \ldots, y_p^{ak}\}$ of k labels (the value of k may be different for different samples) maps to a set of corresponding feature vectors $Z_p^a = \{\mathbf{z}_p^{a1}, \ldots, \mathbf{z}_p^{ak}\}$ that represent those labels. Thus, another way to represent the dataset in the p^{th} modality can be $\mathcal{D}_p = \{(\mathbf{x}_p^a, Z_p^a)\}_{a=1}^{m_p}$.

For each modality p, our objective is to learn a projection function $F_p(\mathbf{x}_p^a; \mathbf{W}_p) : \mathbb{X}_p \rightarrow \mathbb{X}$, that projects an input feature vector $\mathbf{x}_p^a$ into another d-dimensional space $\mathbb{X} = \mathbb{R}^d$ such that the semantic correlations among all the modalities are jointly maximized in this space. In this paper, we assume that each F_p is a linear function of the input feature vector, *i.e.*, $F_p = \mathbf{W}_p^\top \mathbf{x}_p$, where $\mathbf{W}_p \in \mathbb{R}^{d_p \times d}$ denotes the learned projection matrix.

3.2. Overview of Multi-view CCA

As discussed earlier, MVCCA [15] does not make use of semantic information in the form of multi-label annotations. Further, it requires explicit associations/pairings among samples from all the modalities, thus restricting the number of samples in all the modalities to be the same; i.e., $m_1 = m_2 = \ldots = m_n$. Let us assume this to be m. The objective of MVCCA is to learn the projection matrices $\mathbf{W}_p$ that project the input vectors into the common embedding space $\mathbb{X}$ such that after projection, the total distance among all pairs of associated samples is minimized; *i.e.*,

$$\min_{\mathbf{W}_1, \ldots, \mathbf{W}_n} \sum_{\substack{p,q=1 \\ p \neq q}}^{n} \frac{1}{m} \left(\sum_{a=1}^{m} ||\mathbf{W}_p^\top \mathbf{x}_p^a - \mathbf{W}_q^\top \mathbf{x}_q^a||_{\mathrm{F}}^2 \right) \quad (1)$$

$$\text{s.t.} \quad \mathbf{W}_p^\top \mathbf{\Sigma}_{pp} \mathbf{W}_p = \mathbf{I}, \;\; \mathbf{w}_{pu}^\top \mathbf{\Sigma}_{pq} \mathbf{w}_{qv} = 0,$$

$$p, q = 1, .., n, \;\; p \neq q, \;\; u, v = 1, .., d, \;\; u \neq v,$$

where $\mathbf{\Sigma}_{pq} = \frac{1}{m} \sum_{a=1}^{m} \mathbf{x}_p^a \mathbf{x}_q^{a\top}$ denotes the covariance matrix between the (paired) samples in the p^{th} and q^{th} modality, and $\mathbf{w}_{pu}$ is the u^{th} column of $\mathbf{W}_p$. Note that the number of columns in each $\mathbf{W}_p$ is equal to d, which is the dimensionality of the resulting common embedding space. The solution of the above optimization problem is given by the following generalized eigenvalue problem:

$$\begin{pmatrix} \mathbf{\Sigma}_{11} & \ldots & \mathbf{\Sigma}_{1n} \\ \vdots & \ddots & \vdots \\ \mathbf{\Sigma}_{n1} & \ldots & \mathbf{\Sigma}_{nn} \end{pmatrix} \begin{pmatrix} \mathbf{w}_1 \\ \vdots \\ \mathbf{w}_n \end{pmatrix} = \lambda \begin{pmatrix} \mathbf{\Sigma}_{11} & \ldots & \mathbf{\Sigma}_{1n} \\ \vdots & \ddots & \vdots \\ \mathbf{\Sigma}_{n1} & \ldots & \mathbf{\Sigma}_{nn} \end{pmatrix} \begin{pmatrix} \mathbf{w}_1 \\ \vdots \\ \mathbf{w}_n \end{pmatrix}, \quad (2)$$

where $\mathbf{w}_p$ denotes a column of $\mathbf{W}_p$ ($p = 1 \ldots n$).

3.2.1 Retrieval

The output of the above equation is a matrix $\mathbf{W}$ of size $d' \times d'$, where $d' = d_1 + d_2 + \ldots + d_n$. In this matrix, the first column denotes the concatenation of the first/topmost eigenvector (the one with the maximum eigenvalue) corresponding to each modality, the second column denotes the concatenation of the second-top eigenvector corresponding to each modality, and so on. Thus, $\mathbf{W}$ can be thought of as a vertical concatenation of the matrices $\mathbf{W}_p^{'}$ ($p = 1, \ldots, n$) of size $d_p \times d'$. For the p^{th} modality, we obtain a d-dimensional embedding using the projection matrix $\mathbf{W}_p \in \mathbb{R}^{d_p \times d}$, which is obtained by picking the first d columns from $\mathbf{W}_p^{'}$. Thus, the projection of a sample $\mathbf{x}_p^a$ into the common embedding space $\mathbb{X}$ is given by $\mathbf{W}_p^\top \mathbf{x}_p^a$. In this space, since the projected data points are directly comparable, we can perform retrieval by nearest-neighbour search using some similarity measure such as the cosine similarity.

4. Multi-view Multi-label CCA

Inspired by MVCCA, in this section we present our approach for leaning a common embedding space using multi-view and multi-label data. However, rather than requiring explicit associations among samples from different modalities, we make use of multi-label semantics to compute these

associations. Analogous to MVCCA, our objective is to learn the projection matrices $\mathbf{W}_p, \forall p \in \{1, \ldots, n\}$ that project the input vectors into $\mathbb{X}$ such that after projection, the total distance among all pairs of semantically similar samples are minimized. We formulate this as below:

$$\min_{\mathbf{W}_1,\ldots,\mathbf{W}_n} \sum_{\substack{p,q=1 \\ p\neq q}}^{n} \frac{1}{m_p \times m_q} \left(\sum_{a=1}^{m_p} \sum_{b=1}^{m_q} \Gamma_{pq}^{ab} \right), \quad (3)$$

$$\text{where} \quad \Gamma_{pq}^{ab} = g(Z_p^a, Z_q^b) \, ||\mathbf{W}_p^\top \mathbf{x}_p^a - \mathbf{W}_q^\top \mathbf{x}_q^b||_{\text{F}}^2,$$
$$\text{s.t.} \quad \mathbf{W}_p^\top \mathbf{\Psi}_{pp} \mathbf{W}_p = \mathbf{I}, \quad \mathbf{w}_{pu}^\top \mathbf{\Psi}_{pq} \mathbf{w}_{qv} = 0,$$
$$p, q = 1, .., n, \quad p \neq q, \quad u, v = 1, .., d, \quad u \neq v,$$

where $\mathbf{w}_{pu}$ is the u^{th} column of $\mathbf{W}_p$ as in Eq. 1, $g(Z_p^a, Z_q^b)$ denotes the label-set similarity function which gives the degree of similarity between two sets of labels (where each label is denoted by a feature vector), and $\mathbf{\Psi}_{pq}$ denotes the *weighted* covariance matrix between the p^{th} and q^{th} modalities and is computed as below:

$$\mathbf{\Psi}_{pq} = \frac{1}{m_p \times m_q} \sum_{a=1}^{m_p} \sum_{b=1}^{m_q} g(Z_p^a, Z_q^b) \mathbf{x}_p^a \mathbf{x}_q^{b\top} \quad (4)$$

The solution of the optimization problem in Eq. 3 is given by a generalized eigenvalue problem similar to Eq. 2 obtained by replacing $\mathbf{\Sigma}_{pq}$ with $\mathbf{\Psi}_{pq}$. The output will be a matrix $\mathbf{W}$ of size $d' \times d'$, where $d' = d_1 + d_2 + \ldots + d_n$, using which retrieval can be performed similar to that in MVCCA. Note that analogous to MLCCA [26], the semantic information available in the form of discrete multi-label annotations is required by MVMLCCA only at the time of training, and only data points are needed to perform matching during testing/retrieval.

From the optimization problem in Eq. 3, we can observe that if we assume $g(\cdot, \cdot)$ to be binary function such that it is 1 only when the corresponding samples are explicitly paired in the training data and 0 otherwise, we get the optimization problem of MVCCA as in Eq. 1 (scaled by a positive factor). Based on these observations, MVMLCCA can be considered as a generalization of MVCCA [15]. Also, the procedure of retrieval in MVMLCCA during testing is the same as that in MVCCA (Section 3.2.1). Specifically, we first project a given feature vector into the common embedding space using the corresponding projection matrix, and then rank the samples in the retrieval set based on their cosine similarity scores.

Remarks: Here, we should note two points about MVMLCCA (these also apply to MVCCA): (a) The projection matrix for a sample from a particular modality is applied to it individually, which means we can compute the embeddings of samples in a given modality independent of samples in other modalities, thus allowing any-to-any modality matching. (b) The projection matrices for multiple modalities can be concatenated vertically and applied to the concatenated feature vectors (in the same order) from the corresponding modalities to obtain the joint multi-modal embedding in the same common embedding space, thus allowing many-to-many modality (*i.e.*, uni-modal, cross-modal and multi-modal) matching.

4.1. Label-set Similarity

Recall that our approach does not require the vocabularies across different modalities to be the same; instead we assume that each label is represented by a unique feature vector which encodes the semantics of that label in a common feature space $\mathbb{Z}$[2]. This is a relaxed and practically feasible constraint; *e.g.*, using this, we can learn a unified multi-modal embedding space of audio, image and text samples, each tagged with labels from non-overlapping vocabularies.

During training, for a given pair of samples, we are given two corresponding labels-sets Y_p^a and Y_q^b. These are denoted by the corresponding sets of label feature vectors Z_p^a and Z_q^B. We consider two scenarios where either we have a common vocabulary or have distinct vocabularies across different modalities. (1) In the first scenario, we assume the label vocabularies of all the modalities are the same, *i.e.*, $\mathcal{Y}_1 = \mathcal{Y}_2 = \ldots = \mathcal{Y}_n$. To represent labels in this case, we use a simple one-hot encoding. Specifically, we assume that each label is represented by a binary (0/1) one-hot encoding vector in an l-dimensional feature space $\mathbb{Z} = \{0, 1\}^l$. Using this, a label-set $Y_p^a = \{y_p^{a1}, \ldots, y_p^{ak}\}$ of k labels maps to a set of corresponding one-hot encoding vectors $Z_p^a = \{\mathbf{z}_p^{a1}, \ldots, \mathbf{z}_p^{ak}\}$ that represent those labels. (2) In the second scenario, we assume that the label vocabularies of all the modalities are different, though in the same language. In this case, each label $y \in \mathcal{Y}_p$ $(\forall p)$ is represented by a real-valued feature vector z in an l-dimensional feature space $\mathbb{Z} = \mathbb{R}^l$. This feature space is common across all the n vocabularies; *e.g.*, this space can be a Word2Vec [23] or GloVe [24] embedding space. This way, a label-set $Y_p^a = \{y_p^{a1}, \ldots, y_p^{ak}\}$ of k labels is mapped to a set of corresponding real-valued feature vectors $Z_p^a = \{\mathbf{z}_p^{a1}, \ldots, \mathbf{z}_p^{ak}\}$ that represent those labels.

In both the above scenarios, to get a single feature representation $\mathbf{z}_p^a$ of a label-set Y_p^a, we take summation of vector representations of all the labels within that set. To compute similarity between two label-sets Y_p^a and Y_q^b, we want the similarity function g to return a high value when they are similar and a low value when they are not. With this objective, we compute similarity using a squared exponential

[2] In general, this space may also support multi-lingual labels (*e.g.*, labels in English and French), thus allowing comparison of labels across languages. However, in our experiments, we use datasets with vocabularies from a single language; *i.e.* English.

based function as below:

$$g(Z_p^a, Z_q^b) = \exp\left(\frac{-||\mathbf{z}_p^a - \mathbf{z}_q^b||_2^2}{2\sigma}\right), \tag{5}$$

where σ denotes the bandwidth parameter and is set to 1 in our experiments. We note that g can also be defined as a multi-instance matching function [34] where the objective would be to compute similarity between two *bags* of features Z_p^a and Z_q^b, and can be explored in the future.

4.2. Computational Complexity

Let us assume that the maximum number of data points in all the modalities is M (*i.e.*, $M = \max(m_1, m_2, \ldots, m_n)$), the maximum input feature dimensionality of data points in all the modalities is Δ (*i.e.*, $\Delta = \max(d_1, d_2, \ldots, d_n)$), and the total input feature dimensionality of data points in all the modalities is D (*i.e.*, $D = d_1 + d_2 + \ldots + d_n$). Then, the asymptotic computational complexity of computing a cross-modal covariance matrix is $O(M^2\Delta^2)$. Next, since the dimensionality of each label feature vector is l, the asymptotic computational complexity of computing the function $g(\cdot, \cdot)$ for one pair of label-sets will be approximately $O(l)$. Finally, since the size of the generalized eigenvalue problem is $D \times D$ (which is independent of the number of training data points in individual modalities), the asymptotic computational complexity of the eigenvalue decomposition problem is $O(D^3)$. By combining these, the asymptotic computational complexity of MVMLCCA is $O(n^2l^2M^2\Delta^2 + D^3)$. It is clear that a naive implementation of MVMLCCA will not scale to large datasets. However, our block-wise distributed implementation, inspired by [3, 12], makes it scalable to large datasets.

5. Experiments

5.1. Datasets and Features

We use two multi-label and multi-modal datasets in our experiments: IAPRTC-12 and MS-COCO.

IAPRTC-12 [8]: It is a multi-label dataset comprising 19,627 image-text pairs, labeled by a vocabulary of 291 semantic labels. Each text corresponds to captions in English, which are also translated in German and Spanish languages. Thus, it can be considered as a multi-modal dataset comprising associated samples from four modalities. The train/test split is of 17,665/1,962 samples respectively. For representing images, we use the output of the penultimate layer of the ResNet101 [16] model pre-trained on the ImageNet dataset [7]. The captions are represented using summation of 300-dimensional vector representation of each word in the caption. For English captions, we used the Word2Vec [23] model pre-trained on the Google News data. For German and Spanish captions, we use the corresponding pre-trained models from fastText [4].

MS-COCO [20] : This dataset contains 123,287 images, out of which 82,783 comprise the training set and the remaining 40,504 comprise the validation set. Each image is described using five different captions in English. Each pair of an image and its five captions is annotated with a subset of labels from a vocabulary of 80 labels. Similar to earlier works, we report results on the validation set. For representing images, we use the output of the penultimate layer of the pre-trained ResNet101 model as before. For representing text, we take average of summation of word vectors in all the five captions corresponding to each image.

5.2. Evaluation Metrics

We use two evaluation metrics in our quantitative analyses depending on whether the label vocabularies across modalities are same or different: (a) Mean Average Precision or mAP in case of common vocabulary, and (b) Weighted Mean Average Precision or Weighted mAP in case of non-overlapping vocabularies.

Mean Average Precision: The mAP is a widely used evaluation metric in information retrieval, object detection, object segmentation, etc. It is calculated as the average of average precision (or AP) over all the queries. The AP is calculated by finding the area under the Precision-Recall curve. In our task, we consider a retrieved sample as a true positive if at least one label in its ground-truth label-set matches with a label from the query's ground-truth label-set following [26]. As only the top few ranked samples are relevant in a retrieval task, we calculate the AP only for the top 50 retrieved samples for each query q as below:

$$AP_q = \frac{1}{N_{Rt}} \sum_{k=1}^{N_{Rlv}} P_q(k) \times mask_q(k) \tag{6}$$

Here, k denotes the rank in the retrieved samples, N_{Rt} denotes the number of retrieved samples (50 in our case), N_{Rlv} denotes the number of relevant samples, $P_q(k)$ denotes the precision at cut-off k, and $mask_q(k)$ is a function that returns 1 if the retrieved sample at rank k is relevant to the query and 0 otherwise. Finally, the mAP is calculated as the average of AP over all the queries:

$$mAP = \frac{1}{Q} \sum_q AP_q \ , \tag{7}$$

where Q denotes the total number of queries.

Weighted mAP: Inspired by weighted precision introduced in [38], weighted mAP is a modification of mAP and is used when the label vocabularies across the query and retrieval sets are non-overlapping. In such cases, we use label representations (*e.g.*, Word2Vec [23]) to quantify similarity in

IAPRTC-12

Task	CCA	MLCCA	MVCCA	MVMLCCA
Image→Text	0.4080	0.4358	0.4259	0.4247
Text→Image	0.4108	0.4407	0.4301	0.4273
Average	0.4096	0.4383	0.4280	0.4260

MS-COCO

Task	CCA	MLCCA	MVCCA	MVMLCCA
Image→Text	0.9112	0.9158	0.7335	0.9031
Text→Image	0.9395	0.6868	0.7739	0.9275
Average	0.9254	0.8013	0.7537	0.9153

Table 2. Cross-modal retrieval results (mAP) on the IAPRTC-12 and MS-COCO datasets (Section 5.3.1).

the semantic space. We denote the query's label-set as Y_q, and the label-set of the i^{th} retrieved sample as Y_{r_i}. For a given query, all the retrieved samples are considered as a hit with a similarity h. Then, h is calculated as the maximum cosine similarity value between all pairs of labels in Y_q and Y_{r_i}. Using this, we compute weighted AP as:

$$AP_q^w = \frac{1}{N_{Rlv}} \sum_{k=1}^{N_{Rt}} h(Y_q, Y_{r_k}) \times P_q(k) \times mask_q(k) \quad (8)$$

Finally, weighted mAP is calculated by averaging weighted AP over all the queries.

5.3. Results and Discussion

We evaluate the proposed approach and compare it with three baseline methods (CCA [18], MVCCA [15] and MLCCA [26]) under different set-ups. Please refer to the respective papers for more details on these methods.

5.3.1 Cross-modal Retrieval

This is the most popular set-up which is being followed by the papers addressing the problem of cross-modal retrieval, where there are samples from two modalities and they are annotated with labels from a common/single vocabulary. In this experiment, we consider images from both IAPRTC-12 and MS-COCO as one modality. In case of IAPRTC-12 dataset, we consider English captions as the second modality. In case of MS-COCO dataset, we merge all the five captions corresponding to each image into a single paragraph and compute the text representation as described in Section 5.1. The results of this experiment are shown in Table 2. From the results, we can make the following observations: (a) On the smaller IAPRTC-12 dataset, the results obtained using different methods do not vary much. Specifically, MLCCA achieves the best performance, while MVCCA and MVMLCCA are slightly inferior to it. (b) On the larger MS-COCO dataset, the results using CCA and MVMLCCA are comparable, while those using MLCCA and MVCCA are significantly low. Interestingly, MVCCA does not outperform the standard CCA. As also noted by Ranjan *et al*. [26], this suggests that it is better to use multi-label semantic information in an explicit way rather than using it as another modality as in [13]. (c) While the experimental set-up of both MLCCA and MVMLCCA is exactly the same, the difference in their results signifies the conceptual difference between the two approaches. (d) Overall, we can observe that rather than relying on strong supervision in the form of explicit associations between cross-modal samples as in CCA and MVCCA, MVMLCCA is able to achieve competitive results on both the datasets by effectively utilizing weak supervision available in the form of multi-label semantics.

Dataset	CCA	MLCCA	MVCCA	MVMLCCA
IAPRTC-12	9	251	12	353
MS-COCO	47	7600	68	7500

Table 3. Training time (in seconds) comparisons on the IAPRTC-12 and MS-COCO datasets.

Training time comparison: In Table 3, we compare the training time of all the four methods on both the datasets. It is important to note that in all of our experiments, we use the fast implementation of MLCCA (Fast MLCCA) provided by its authors, as the original MLCCA is not scalable to large datasets such as MS-COCO. This scalability in MLCCA is achieved by computing an approximate covariance matrix that considers only the 50 nearest neighbours of a given sample. Rather than adopting a similar approximation, we use a block-wise distributed implementation of MVMLCCA that makes it memory efficient and also helps in achieving training time comparable to that of MLCCA as evident from Table 3. As expected, this is significantly higher than that of CCA/MVCCA as both MLCCA and MVMLCCA need to compute associations between samples using multi-label semantics during the training phase. After training, since all the four approaches return one feature embedding matrix per modality, the testing/retrieval time is the same for all.

5.3.2 Multi-modality Cross-modal Retrieval

This is an extension of the previous set-up where during the training phase, we assume availability of samples from multiple (more than two) modalities which are annotated with labels from a single vocabulary, and during the testing phase, we perform cross-modal retrieval between dif-

Task	CCA	MLCCA	MVCCA	MVMLCCA
Image→English	0.6104	0.6982	0.5754	0.5912
Image→German	0.5366	0.5350	0.5471	0.4157
Image→Spanish	0.5504	0.5413	0.5511	0.4278
English→Image	0.6526	0.5887	0.6437	0.6537
English→German	0.5653	0.5428	0.5658	0.4251
English→Spanish	0.5755	0.5525	0.5704	0.4411
German→Image	0.5786	0.4428	0.4838	0.4250
German→English	0.5642	0.3090	0.4090	0.3022
German→Spanish	0.5441	0.5222	0.4961	0.3738
Spanish→Image	0.5831	0.4352	0.4815	0.4267
Spanish→English	0.5679	0.3150	0.4029	0.3005
Spanish→German	0.5387	0.5079	0.4892	0.4127
Average	0.5721	0.4976	0.5180	0.4329

Table 4. Cross-modal retrieval results (mAP) using different pairs of modalities of the IAPRTC-12 dataset (Section 5.3.2).

ferent pairs of modalities. For this experiment, we consider all the four modalities available in the IAPRTC-12 dataset (*i.e.*, RGB images, English captions, German captions and Spanish captions), which leads to availability of four associated views for each data point. This results in twelve cross-modal retrieval tasks (two for each pair of modalities). It is important to note that in this set-up, both MVCCA and the proposed MVMLCCA involve one-time training and learn a single model using samples from *all* the modalities simultaneously (each model contains one feature embedding per modality). However, since CCA and MLCCA are applicable to only two-modality datasets, we need to learn their embeddings for *each* pair of modalities *separately*. This results in learning six models - one for each pair of modalities. This is quite likely to reduce the amount of noise introduced because of the presence of multiple modalities, thus giving both CCA and MLCCA an advantage over MVCCA and MVMLCCA. This also means that the results of CCA and MLCCA are not directly comparable to those of MVCCA and MVMLCCA, however we include them for completeness and also to demonstrate the effectiveness of multi-modal learning in one-pass compared to learning individually from bi-modal data.

In Table 4, we show the results for all the twelve cross-modal retrieval tasks. We can observe that: (a) CCA achieves the best results on an average as it learns one model at a time using a pair of modalities and strong supervision in the form of explicit associations between samples. (b) MVCCA, which learns using all the four modalities simultaneously, however still uses strong supervision in the form of explicit associations among samples from all the modalities, achieves the second best results on an average. This reduction in performance compared to CCA validates our hypothesis that learning with a large number of modalities simultaneously introduces noise in the learning process, as each sample is now being pulled by multiple samples from other modalities. (c) The performance of MLCCA, which learns using a pair of modalities at a time analogous to CCA, however uses weak supervision in the form of multi-label annotations for computing associations between samples, is significantly inferior to CCA on an average. This result indicates that though MLCCA is practically more feasible than CCA (since in the real-world scenarios, it is easier to obtain weak supervision in the form multi-label semantics compared to strong supervision in the form of explicit pairings of cross-modal samples), the sample associations computed using multi-label annotations are not as good as those obtained using explicit pairings of samples. (c) By comparing the results of MLCCA with MVCCA, we can infer that the reduction in performance due to non-availability of explicit associations of samples across all the modalities can be controlled by reducing the amount of noise in the training process by considering only a pair of modalities at a time rather than considering all of them. (d) The performance of MVMLCCA is the least among all the compared methods on an average. This is as expected since MVMLCCA does not make use of explicit associations among samples as in CCA and MVCCA, and learns a model using samples from all the modalities simultaneously and not from a pair of modalities at a time as in CCA and MLCCA. However, it is practically the most feasible algorithm among all for the real-world scenarios, where we have abundant samples in different modalities tagged with a variety of semantic labels, which we will examine further in the next section.

Task	CCA	MLCCA	MVCCA	MVMLCCA
Image→Image	0.5403	0.7678	0.7179	0.7175
English→English	0.5497	0.6615	0.6531	0.6647
German→German	0.5552	0.5804	0.5824	0.5418
Spanish→Spanish	0.5701	0.5940	0.5896	0.5376
Average	0.5538	0.6509	0.6358	0.6154

Table 5. Uni-modal retrieval results (mAP) using different modalities of the IAPRTC-12 dataset (Section 5.3.2).

For completeness, we also show the results for uni-modal retrieval in Table 5. In this case, both CCA and MLCCA learn a feature embedding separately for each modality assuming both the input and output modalities to be the same, whereas MVCCA nd MVMLCCA use the previously learned embeddings (as in Table 4). These results show that MVMLCCA performs favourably compared to other methods, and further strengthen our understanding of the practical advantages of MVMLCCA compared to other methods.

5.3.3 Multi-Vocabulary Multi-modality Cross-modal Retrieval

In this set-up, we consider the multi-vocabulary situation where the samples in different modalities are annotated with labels from different (non-overlapping) vocabularies, and also there are no associations among them. For this, we use a fusion of the IAPRTC-12 and MS-COCO datasets, by considering only the RGB images and English captions from the MS-COCO dataset, and German and Spanish captions from the IAPRTC-12 dataset. This leads to sixteen cross-modal retrieval tasks by considering all possible pairs of input and output modalities (including uni-modal retrievals). Note that since now we do not have explicit pairings among samples from different pairs of modalities, CCA and MVCCA will not be applicable in this set-up. Also, in this case, since we need to use a real-valued feature representation (*e.g.*, Word2Vec) for representing and comparing labels, which was not explored in the original implementation of MLCCA, we modify it appropriately for our comparisons. Similar to the set-up in Section 5.3.2, while MVMLCCA learns a a single embedding for each modality by using samples from all the modalities simultaneously, MLCCA learns a separate model for each pair of modalities which leads to learning separate (and thus better optimized) embeddings for each given modality when it is paired with different modalities including uni-modal retrieval tasks.

In Table 6, we show the results for different tasks in terms of weighted mAP. From these results, we can observe that in this challenging set-up, MVMLCCA marginally outperforms MLCCA on an average, though it is much more relaxed than MLCCA. These results, along with the non-applicability of CCA and MVCCA in this set-up, indicate that MVMLCCA not only relaxes all the constraints/assumptions required by the competing methods but also achieves good empirical results, and thus can be preferably adopted for all practical cross-modal retrieval tasks.

Task	MLCCA	MVMLCCA
COCO_Image→COCO_Image	0.9508	0.9570
COCO_Image→COCO_Caption	0.9560	0.9632
COCO_Image→IAPR_German	0.6753	0.6759
COCO_Image→IAPR_Spanish	0.6682	0.6747
COCO_Caption→COCO_Image	0.9655	0.8691
COCO_Caption→COCO_Caption	0.9783	0.9801
COCO_Caption→IAPR_German	0.6770	0.6762
COCO_Caption→IAPR_Spanish	0.6670	0.6742
IAPR_German→COCO_Image	0.6827	0.6929
IAPR_German→COCO_Caption	0.6899	0.6871
IAPR_German→IAPR_German	0.8200	0.8359
IAPR_German→IAPR_Spanish	0.7762	0.8230
IAPR_Spanish→COCO_Image	0.6826	0.6925
IAPR_Spanish→COCO_Caption	0.6899	0.6882
IAPR_Spanish→IAPR_German	0.7804	0.8224
IAPR_Spanish→IAPR_Spanish	0.8111	0.8392
Average	0.7794	0.7845

Table 6. Cross-modal retrieval results (weighted mAP) using multi-vocabulary and multi-modality data (Section 5.3.3).

6. Summary and Conclusion

Although CCA [18] was introduce more than eight decades back, it is still considered as the first baseline in a variety of cross-modal matching and retrieval tasks. This widespread popularity of CCA has led to the development of several extensions of CCA in the past that attempt to address some of its limitations. In this paper, motivated by the limitations of CCA and two of its popular extensions MLCCA [26] and MVCCA [15], we have presented MVMLCCA which can be considered as a true generalization of CCA for learning multi-modal embeddings using multi-label data. MVMLCCA can accommodate any number of modalities and does not require explicit associations among samples from diverse modalities during the training phase. Rather, it relies on weak supervision available in the form of multi-label annotations to compute such associations. Extensive experiments demonstrate that the proposed MVMLCCA approach successfully captures the semantics of large-scale multi-modal datasets, and thus can be an attractive solution for building flexible and scalable cross-modal retrieval systems.

Limitations and Potential Negative Social Impact:

One limitation of our method is that it depends on additional meta-data in the form of multi-label annotations. While we do not have any empirical evidence that compares the efforts between annotating a database with labels and making pairwise correspondence with samples from another modality, we believe that the former is both easier and cheaper than the latter. Another limitation is that in all the experiments, we assume the multi-label annotations to be available in the English language, which may not hold for a large amount of digital content. However, this limitation may be overcome by using an appropriate machine translation or multi-lingual word-embedding system. Also, as true with all computational learning based techniques, our experiments consumed substantial energy generated by burning of fossil fuels and thus warmed our planet. However, we hope that our non-deep learning based approach managed to keep it low compared to computationally intensive deep learning based techniques.

Acknowledgement: YV would like to thank the Department of Science and Technology (India) for the INSPIRE Faculty award 2017.

References

[1] Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelovic, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. In *NeurIPS 2020*, 2020. 2

[2] Galen Andrew, Raman Arora, Jeff A. Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *ICML*, volume 28, pages 1247–1255, 2013. 2

[3] Rohit Babbar and Bernhard Schölkopf. DiSMEC: Distributed sparse machines for extreme multi-label classification. In *ACM International Conference on Web Search and Data Mining*, pages 721–729, 2017. 5

[4] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *TACL*, 5:135–146, 2017. 5

[5] Wei Chen, Yu Liu, Erwin M. Bakker, and Michael S. Lew. Integrating information theory and adversarial learning for cross-modal retrieval. *Pattern Recognition*, 117:107983, 2021. 2

[6] Yanbei Chen and Loris Bazzani. Learning joint visual semantic matching embeddings for language-guided retrieval. In *ECCV*, 2020. 2

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 5

[8] Hugo Jair Escalante, Carlos A. Hernández, Jesús A. González, Aurelio López-López, Manuel Montes y Gómez, Eduardo F. Morales, Luis Enrique Sucar, Luis Villaseñor Pineda, and Michael Grubinger. The segmented and annotated IAPR TC-12 benchmark. *Comput. Vis. Image Underst.*, 114(4):419–428, 2010. 2, 5

[9] Fartash Faghri, David J. Fleet, J. Kiros, and S. Fidler. VSE++: Improving visual-semantic embeddings with hard negatives. In *BMVC*, 2018. 2

[10] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. Cross-modal retrieval with correspondence autoencoder. In *ACM Multimedia*, pages 7–16, 2014. 2

[11] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc' Aurelio Ranzato, and Tomas Mikolov. DeViSE: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*. 2013. 2

[12] Xiao Fu, Kejun Huang, Evangelos E. Papalexakis, Hyun Ah Song, Partha P. Talukdar, Nicholas D. Sidiropoulos, Christos Faloutsos, and Tom M. Mitchell. Efficient and distributed generalized canonical correlation analysis for big multiview data. *IEEE Trans. Knowl. Data Eng.*, 31(12):2304–2318, 2019. 5

[13] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *IJCV*, 106(2):210–233, 2013. 2, 6

[14] Wang-Li Hao, Zhaoxiang Zhang, and He Guan. CMCGAN: A uniform framework for cross-modal visual-audio mutual generation. In *AAAI*, pages 6886–6893, 2018. 2

[15] David R. Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Comput.*, 16(12):2639–2664, 2004. 1, 2, 3, 4, 6, 8

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[17] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Intell. Res.*, 47:853–899, 2013. 2

[18] Harold Hotelling. Relations Between Two Sets of Variates. *Biometrika*, 28(3-4):321–377, 12 1936. 1, 2, 6, 8

[19] Andrej Karpathy and Fei Fei Li. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2, 5

[21] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *ICCV*, 2015. 2

[22] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). In *ICLR*, 2015. 2

[23] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, pages 3111–3119. 2013. 2, 3, 4, 5

[24] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014. 3, 4

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 2

[26] Viresh Ranjan, Nikhil Rasiwasia, and C. V. Jawahar. Multi-label cross-modal retrieval. In *ICCV*, 2015. 1, 2, 4, 5, 6, 8

[27] Nikhil Rasiwasia, Dhruv Mahajan, Vijay Mahadevan, and Gaurav Aggarwal. Cluster canonical correlation analysis. In *AISTATS*, 2014. 2

[28] N. Rasiwasia, J. C. Pereira, E. Coviello, G. Doyle, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *ACM MM*, 2010. 2

[29] Scott E. Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *CVPR*, 2016. 2

[30] N. Sarafianos, X. Xu, and I. Kakadiaris. Adversarial representation learning for text-to-image matching. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[31] A. Sharma, A. Kumar, H. D. III, and D. W. Jacobs. Generalized multiview analysis: A discriminative latent space. In *CVPR*, 2012. 2

[32] Xin Shu and Guoying Zhao. Scalable multi-label canonical correlation analysis for cross-modal retrieval. *Pattern Recognit.*, 115:107905, 2021. 2

[33] Yashaswi Verma and C. V. Jawahar. A support vector approach for cross-modal search of images and texts. *Comput. Vis. Image Underst.*, 154:48–63, 2017. 2

[34] J. Wang and J. D. Zucker. Solving the multiple-instance problem: A lazy learning approach. In *ICML*, 2000. 5

[35] L. Wang, Y. Li, and S. Lazebnik. Learning deep structure-preserving image-text embeddings. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[36] Shuo Wang, Dan Guo, Xin Xu, Li Zhuo, and Meng Wang. Cross-modality retrieval by joint correlation learning. *ACM Trans. Multim. Comput. Commun. Appl.*, 15(2s):56:1–56:16, 2019. 2

[37] Xin Wang, Yuan-Fang Wang, and William Yang Wang. Watch, Listen, and Describe: Globally and locally aligned cross-modal attentions for video captioning. In *NAACL-HLT*, 2018. 2

[38] Jason Weston, Samy Bengio, and Nicolas Usunier. WSABIE: Scaling up to large vocabulary image annotation. In *IJCAI*, 2011. 5

[39] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[40] Will Y Zou, Richard Socher, Daniel M Cer, and Christopher D Manning. Bilingual word embeddings for phrase-based machine translation. In *EMNLP*, pages 1393–1398, 2013. 2