Cross-Modal Retrieval: A Pairwise Classification Approach

Aditya Krishna Menon* Didi Surian† Sanjay Chawla‡

Abstract

Content is increasingly available in multiple modalities (such as images, text, and video), each of which provides a different representation of some entity. The cross-modal retrieval problem is: given the representation of an entity in one modality, find its best representation in all other modalities. We propose a novel approach to this problem based on pairwise classification. The approach seamlessly applies to both the settings where ground-truth annotations for the entities are absent and present. In the latter case, the approach considers both positive and *unlabelled* links that arise in standard cross-modal retrieval datasets. Empirical comparisons show improvements over state-of-the-art methods for cross-modal retrieval.

1 Introduction

Suppose we have a database comprising a number of texts (e.g. poems) and images (e.g. of natural scenes). The *cross-modal retrieval* problem is: given a poem, what is the corresponding natural scene that best accompanies it, and vice versa? To facilitate such retrieval, we are given a few examples of suitable pairings – for example, a poem and image both about a daffodil. As semantically related content typically exhibits a similar form of *feature correlation* in its representation, this can be used to facilitate retrieval.

The de-facto approach to cross-modal retrieval has been *canonical correlation analysis (CCA)* or variants thereof. Given two modalities of the data represented by feature matrices, the objective of CCA is to project these matrices onto a *latent subspace* where the modalities exhibit maximum correlation. One then uses the latent subspace to perform retrieval with standard distance based querying. Recently, a distinct approach termed *semantic matching (SM)* has been explored [17]. This approach relies on *ground-truth annotations* for the representations, describing for example the underlying concepts or categories that they represent. One projects both modalities into a rich *supervised* subspace, where distance based retrieval is performed as usual.

*National ICT Australia and Australian National University.
†The University of Sydney and National ICT Australia.
‡Qatar Computing Research Institute and The University of Sydney.

CCA and semantic matching represent prototypical approaches for two distinct settings, where annotations are absent and present respectively. The contribution of this paper is the development of a framework which lets us reason about *both* settings in a unified way. Our approach is based on a novel reduction of cross-modal retrieval to binary classification over pairs, and opens the problem to the rich set of supervised learning methods. The framework also suggests a variant of semantic matching with superior retrieval performance.

We now review prior work on the problem (§2). We then proceed to describe the basis for our approach (§3), detailing how it can adapt to settings without (§4) and with (§5) ground-truth annotation, and comparing to previous work (§6). Experiments in §7 demonstrate superior performance to CCA and SM.

2 Previous Work

The problem of content-based multimedia retrieval has been studied extensively in the fields of multimedia and computer vision [6, 14, 5]. Classically, the focus is on *uni-modal retrieval* problems i.e. settings where we are interested in mapping the first modality (typically text) to the second modality (typically images), but not the other way around [18, 21, 3].

Recently, there has been an increasing interest in *cross-modal retrieval problems* that learn the mappings between two objects from different modalities such as text and images. *Canonical Correlation Analysis* (CCA) [7] is a generic approach to this problem which has enjoyed wide success. The power of CCA is that it maps all objects in different modalities into a common latent space, where comparisons between elements can be made based on standard distance metrics. (A related idea is the use of hashing functions [27], which provide an additional benefit of fast indexing.) Extensions to CCA which consider both positive and negative matches between pairs of elements have also been considered [11].

Recently, [17, 15] explored the idea of *semantic matching*, which exploits ground-truth annotations together with CCA. This is shown to significantly improve performance compared to CCA. Similar approaches have also been explored in [24, 28].

An alternate strategy is to represent one modality by any available *metadata*, thus reducing the problem

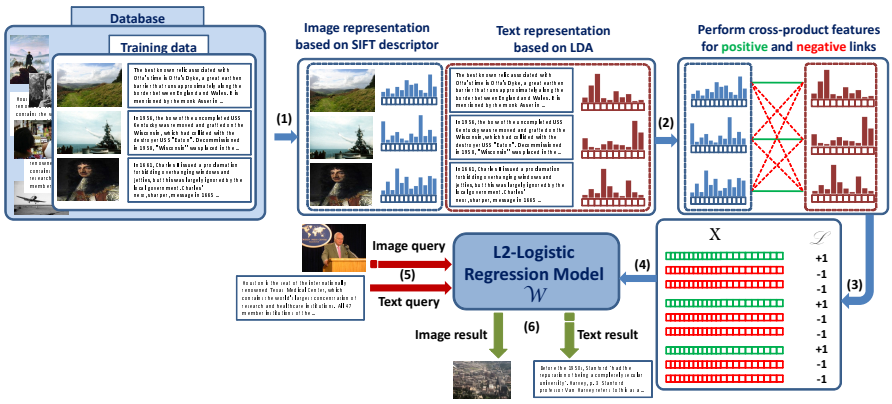

Figure 1: Workflow for the supervised learning approach for cross-modal retrieval. The key novel idea is to create positive and negative links and associate each link with a cross-product feature. This is in contrast to the traditional approaches like CCA which operate on positive links only. Annotation can be incorporated; see §5.

to one of uni-modal retrieval. For example, in image retrieval [22, 13] one can associate an image with any accompanying text (e.g. its caption). One can then match the query text with this accompanying text. However, the quality of retrieval strongly depends on the availability, quantity, and quality of the metadata. An alternate strategy is to represent an image by "visual words", formed by appropriate segmentation of the image, and learn the joint distribution over visual and textual words [1, 12]. Recent work on applying neural networks to automatically learn mappings between images and text is another promising direction [10].

3 Formalising Cross-Modal Retrieval

Let $\mathcal{A}, \mathcal{B}$ be two sets, being the possible representations of the two modalities. Typically, $\mathcal{A} \subseteq \mathbb{R}^{d_a}$ and $\mathcal{B} \subseteq \mathbb{R}^{d_b}$ for $d_a, d_b \in \mathbb{N}_+$. We are given training examples $\mathcal{D} = \{(a^{(i)}, b^{(i)})\}_{i=1}^n$, where each $(a^{(i)}, b^{(i)}) \in \mathcal{A} \times \mathcal{B}$ represents a suitable pairing of objects from each modality, such as an (image, text) pair. The cross-modal retrieval problem is to learn mappings $f : \mathcal{A} \to \mathcal{B}$ and $g : \mathcal{B} \to \mathcal{A}$ that map each representation to its best pairing in the other modality. We now review the standard perspective on learning f, g.

3.1 The Latent Map Perspective

One approach to cross-modal retrieval is to learn *latent maps* $\psi_\mathcal{A} : \mathcal{A} \to \mathbb{R}^k$, $\psi_\mathcal{B} : \mathcal{B} \to \mathbb{R}^k$, where $k \in \mathbb{N}_+$ is the dimensionality of some *latent subspace* and compute:

$$f : a \mapsto \operatorname*{argmin}_{b \in \mathcal{B}} d(\psi_\mathcal{A}(a), \psi_\mathcal{B}(b)) \qquad g : b \mapsto \operatorname*{argmin}_{a \in \mathcal{A}} d(\psi_\mathcal{A}(a), \psi_\mathcal{B}(b)), \tag{3.1}$$

where $d(\cdot, \cdot)$ is some suitable distance function (e.g. Euclidean distance). For example, CCA uses $\psi_\mathcal{A} : a \mapsto Ua$, $\psi_\mathcal{B} : b \mapsto Vb$, where $U \in \mathbb{R}^{k \times d_a}, V \in \mathbb{R}^{k \times d_b}$ are chosen so as to maximise the correlation between $\psi_\mathcal{A}(a)$ and $\psi_\mathcal{B}(b)$ for each $(a, b) \in \mathcal{D}$.

3.2 The Similarity Function Perspective

This paper proposes a more general approach to reason about cross-modal retrieval: we cast the problem as one of learning a *joint similarity function*, $s : \mathcal{A} \times \mathcal{B} \to \mathbb{R}$, which assigns a high score whenever the given pair of representations are related to each other. One then constructs mappings:

$$f : a \mapsto \operatorname*{argmax}_{b \in \mathcal{B}} s(a, b) \ , \ g : b \mapsto \operatorname*{argmax}_{a \in \mathcal{A}} s(a, b). \tag{3.2}$$

The latent feature approach may be seen as a special case where $s(a, b) = e^{-d(\psi_\mathcal{A}(a), \psi_\mathcal{B}(b))}$.

To learn a good similarity function, we must define a measure of success. Following [17], we shall do this in terms of *ground-truth annotations* for representations of an entity. An annotation is simply an element from some set $\mathcal{Y}$, which is typically $\{1, 2, \ldots, K\}$ for some $K \in \mathbb{N}_+$. For example, $\mathcal{Y}$ may represent the category of a news article, such as sports, finance, or arts. Intuitively, a good similarity function assigns a high score whenever the given pair of representations *possesses the same ground-truth annotation*. We now formalise this, and relate it to pairwise classification.

3.3 Similarities and Classification

We formalise our performance measure as follows. Let A be a random variable over $\mathcal{A}$, being a representation in the first modality. Let Y be a random variable over $\mathcal{Y}$, being the annotation for A. Similarly, let B be a random variable over $\mathcal{B}$, with annotation Y'. We can define a random variable $\mathsf{Z} := 2[\![\mathsf{Y} = \mathsf{Y}']\!] - 1 \in \{\pm 1\}$, which

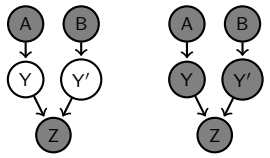

Figure 2: Graphical model for cross-modal retrieval, without (**Left**) and with (**Right**) ground-truth annotation.

measures whether or not the two annotations agree. A good similarity function s is one that minimises:

$$\mathbb{L}(s) = \mathbb{E}_{\mathsf{A},\mathsf{B},\mathsf{Z}}\left[\ell(\mathsf{Z}, s(\mathsf{A},\mathsf{B}))\right], \tag{3.3}$$

where $\ell(z, v) = [\![zv < 0]\!]$ is the 0-1 loss.

We make two comments on this setup. First, our performance measure allows Y and Y' to be *unobserved*; we simply require that Z is observed, for which observing Y and Y' is sufficient but not necessary. Thus, we can seamlessly capture *both* the cases of learning without and with ground-truth annotations. (See Figure 2.) Second, the random variable Z can be viewed as a binary label on the instance pair (A,B). Thus, *learning a similarity function is equivalent to a binary classification problem on the instance space* $\mathcal{A}\times\mathcal{B}$. In particular, given samples $\{((a^{(i)}, b^{(i)}), z^{(i)})\}_{i=1}^n$, we can minimise an empirical binary classification surrogate risk:

$$L_{\text{emp}}(s) = \sum_{i=1}^{n} \ell(z^{(i)}, s(a^{(i)}, b^{(i)})), \tag{3.4}$$

where $\ell : \{\pm 1\} \times \mathbb{R} \to \mathbb{R}_+$ is some convex surrogate to the 0-1 loss, e.g. the logistic loss $\ell(y, v) = \log(1 + e^{-yv})$.

We now explore the details of making this approach feasible in the cases where ground-truth annotations are absent and present respectively.

4 Pairwise Classification Without Annotations

Recall that the training set for a cross-modal problem in the absence of ground-truth annotation is $\mathcal{D} = \{(a^{(i)}, b^{(i)})\}$, being pairings of representations from each modality. From the perspective of the previous section, we can associate with each training example a label $z^{(i)} = +1$, denoting the corresponding representations are suitably matched i.e. the training set is effectively $\mathcal{D}' = \{((a^{(i)}, b^{(i)}), z^{(i)})\}$. We now show how to use this dataset to learn a similarity function that approximately minimises Equation 3.3, building up from a simple baseline based on learning from positive only data (§4.1), to two methods that reduce the problem to learning from positive and unlabelled samples (§4.2).

4.1 Learning from positive only data

Naïvely minimising the empirical risk (Equation 3.4) with $\mathcal{D}'$ is problematic as the training set comprises *only positive* instances: this will lead to trivial solutions. Instead, we can use techniques for learning from positive only samples, such as the *one-class SVM* [19]. Given the perspective of the previous section, this seems a natural candidate for cross-modal retrieval, but to our knowledge has not previously been evaluated.

However, learning from positive only data in this setting has two important limitations. First, if we have two pairs $(a^{(1)}, b^{(1)})$ and $(a^{(2)}, b^{(2)})$, then we do not utilise the fact that $(a^{(1)}, b^{(2)})$ are somewhat *less* similar than the given positive pairs. Second, one-class methods were designed for the setting where outliers form a minority of instances. However, in our case the positive pairs form a minority. We thus consider other approaches of augmenting positive and unlabelled data.

4.2 Learning from positive and unlabelled data

At first glance, it is not clear how to improve upon the one-class SVM: we inherently appear to lack negative examples. Our simple observation is that progress can be made as follows. We can augment the dataset with *all pairs* of instances from each modality, giving:

$$\mathcal{D}'' = \{((a^{(i)}, b^{(j)}), z^{(i,j)})\}_{i,j=1}^n.$$

As in the previous section, we have $z^{(i,i)} = +1$. However, for $i \neq j$, we do not *a priori* know whether the pair of instances $(a^{(i)}, b^{(j)})$ match (we do expect that *most* pairs will be negative, assuming that there is an equal representation of all categories in $\mathcal{D}'$). Formally, these comprise *unlabelled* examples, i.e. we have $z^{(i,i} = 1$, and $z^{(i,j)} =?$ for $i \neq j$. The problem of learning from the dataset $\mathcal{D}''$ is thus one of learning from *positive and unlabelled examples* [4]. We now look to import techniques for this problem.

4.2.1 Reduction to binary classification

Following [4], our first approach to learning with $\mathcal{D}''$ is to simply learn a model to *distinguish the positive from unlabelled examples*. While apparently naïve, [4] shows this is in fact *optimal* when the goal is simply to learn a good ranking over instances. The proviso is that an *unlabelled at random* assumption must hold. This assumption is simply that the observed positive examples are a representative sample of all positive examples. In a cross-modal retrieval dataset, this means that we believe there is no bias towards having observed certain positive matches over others; whether this is justified depends on the nature of the data collection process.

Formally, for this approach, we define $z^{(i,j)} = 2[\![i = j]\!] - 1$, giving us a completely labelled dataset $\mathcal{D}''$. We can now minimise Equation 3.4 with two further restrictions. First, we use s that are linear models on a

feature transformation Φ of the pair a, b:

$$s(a, b; w) = \langle w, \Phi(a, b) \rangle. \tag{4.5}$$

We will discuss the choice of Φ in the next section. Second, we employ ℓ_2 regularisation to prevent overfitting. This yields the (strongly convex) objective:

$$\frac{1}{n \cdot n} \sum_{i=1}^{n} \sum_{j=1}^{n} \ell(z^{(i,j)}, \langle w, \Phi(a^{(i)}, b^{(j)}) \rangle) + \frac{\lambda}{2} ||w||_2^2, \tag{4.6}$$

which is simply a regularised empirical risk for a binary classification problem over pairs. Given the learned weights w, we perform retrieval via Equation 3.2 using the scoring function Equation 4.5.

The apparent limitation of this reduction is that the given positive labeled instances and the *hypothesised* negative labeled instances are placed on an equal footing. We now attempt to relax this.

4.2.2 Reduction to bipartite ranking An alternate strategy to dealing with positive and unlabelled data is to reduce the problem to one of bipartite ranking: we simply look to score the known positive instances above the unlabelled instances [26, 20]. As above, we set $z^{(i,j)} = 2[\![i = j]\!] - 1$, but we instead minimise:

$$\frac{1}{n \cdot n} \sum_{i=1}^{n} \sum_{j \neq i} \ell(z^{(i,j)}, \langle w, \Phi(a^{(i)}, b^{(i)}) - \Phi(a^{(i)}, b^{(j)}) \rangle) + \frac{\lambda}{2} ||w||_2^2. \tag{4.7}$$

Intuitively, if the pairs in $\mathcal{D}'$ represent the top matches for each entity (e.g. the best pairing of poem and natural scene), the assumption that positives must score higher than unlabelled instances is reasonable.

4.3 Choice of feature mapping An important ingredient in either of the above reductions is to design a feature mapping Φ. The obvious first choice is to use the concatenated representation $\Phi(a, b) = \begin{bmatrix} a & b \end{bmatrix}$. However, this has a subtle problem. For any $a \in \mathcal{A}$, the highest ranked instance from the second modality is:

$$\operatorname*{argmax}_{b \in \mathcal{B}} \langle \begin{bmatrix} w & v \end{bmatrix}, \begin{bmatrix} a & b \end{bmatrix} \rangle = \operatorname*{argmax}_{b \in \mathcal{B}} \langle v, b \rangle.$$

Crucially, this does not depend on a – that is, it induces the same ranking over instances from the second modality *regardless* of the choice for a.[1] This is clearly not desirable. Therefore, we need a feature mapping that explicitly models the interaction between a and b.

Perhaps the simplest form of nonlinearity is the cross-product representation $\Phi(a, b) = \begin{bmatrix} a_d \cdot b_{d'} \end{bmatrix}_{d,d'}$, so that the similarity function is equivalently:

$$s(a, b) = \langle w, \Phi(a, b) \rangle = \langle a, Wb \rangle, \tag{4.8}$$

where $W \in \mathbb{R}^{d_a \times d_b}$. Clearly, the choice of Φ is equivalent to a choice of suitable kernel over the pair-space $\mathcal{A} \times \mathcal{B}$, where $K((a, b), (a', b')) = \langle \Phi((a, b)), \Phi((a', b')) \rangle$. Other choices of kernel beyond that implied by the cross-product features [8] are of course feasible. However, we shall pursue the cross-product approach as it gives favourable empirical performance, and lets us draw connections to existing methods (§6.2, §6.1).

4.4 Computational complexity Both Equations 4.6 and 4.7 are regularised risk minimisation problems. Ostensibly, a gradient following method to minimise these will require $O(md)$ time, where $m = n^2$ is the number of pairs in the objective, and $d = d_a \cdot d_b$ is the number of cross-product features. Both terms may be reduced significantly by subsampling: for the m term, this corresponds to only considering a *subset* of all possible pairs across the modalities, and for the d term, a subset of all possible feature pairs. In §9.3 (Supplementary Material), we empirically investigate the sensitivity of our method to subsampling along both dimensions.

5 Pairwise Classification With Annotations

We now show how our approach seamlessly carries over the setting where each pair of representations possesses ground-truth annotations. The training set in this case is $\mathcal{D}' = \{((a^{(i)}, b^{(i)}), y^{(i)})\}_{i=1}^{n}$, where $y^{(i)} \in \mathcal{Y}$ is the category for the ith entity. We now detail two approaches to learning a similarity function from $\mathcal{D}'$.

5.1 Pairwise classification approach Recall that our basic strategy in the previous section was to approximately minimise Equation 3.3. We may follow the same strategy here, albeit with a different choice of labels. The presence of annotations $y^{(i)}$ immediately implies whether or not each pair (i, j) of instances are compatible: we can simply set $z^{(i,j)} = 2[\![y^{(i)} = y^{(j)}]\!] - 1$, and minimise the objective of Equation 4.6 with this.

This illustrates the conceptual simplicity of our framework – all that varies in the settings without and with annotations in the choice of labels $z^{(i,j)}$ on pairs. In fact, our framework suggests a simpler approach in this setting, which we now detail.

5.2 The marginal approach As it represents the risk for a binary classification problem, Equation 3.3 is minimised by $\operatorname{sign}(2 \Pr[\mathsf{Z} = 1 | \mathsf{A}, \mathsf{B}] - 1)$. That is, the key quantity to estimate is $\Pr[\mathsf{Z} = 1 | \mathsf{A}, \mathsf{B}]$. The structure of this quantity suggests a simpler approach than directly learning over pairs. Let $\eta_{\mathcal{A}} : \mathcal{A} \to \Delta_{|\mathcal{Y}|-1}$ and $\eta_{\mathcal{B}} :$

[1] This problem has been noted in slightly different contexts that also involve prediction with pairs of entities [23].

$\mathcal{B} \to \Delta_{|\mathcal{Y}|-1}$ be the instance-conditional distributions over labels, e.g. $(\eta_{\mathcal{A}})_y(a) = \Pr[\mathsf{Y} = y|\mathsf{A} = a]$. As shown in §9.1 (supplementary material), we have:

$$\Pr[\mathsf{Z} = 1|\mathsf{A}, \mathsf{B}] = \langle \eta_{\mathcal{A}}(\mathsf{A}), \eta_{\mathcal{B}}(\mathsf{B}) \rangle.$$

This suggests the following approach to learning s: using the datasets $\{(a^{(i)}, y^{(i)})\}$ and $\{(b^{(i)}, y^{(i)})\}$, learn *independent* models for $\eta_{\mathcal{A}}$ and $\eta_{\mathcal{B}}$, and use their dot-product as our joint similarity function. For example, we can use multi-class logistic regression on both datasets, learning $w \in \mathbb{R}^{d_a \times |\mathcal{Y}|}, v \in \mathbb{R}^{d_b \times |\mathcal{Y}|}$ where e.g.

$$\eta_{\mathcal{A}}(a; w) = \left[\frac{\exp(\langle w^{(y)}, a \rangle)}{\sum_{y' \in \mathcal{Y}} \exp\langle w^{(y')}, a \rangle} \right]_{y \in \mathcal{Y}}, \tag{5.9}$$

and then compute the similarity function:

$$s(a, b; w, v) = \langle \eta_{\mathcal{A}}(a; w), \eta_{\mathcal{B}}(b; v) \rangle. \tag{5.10}$$

We dub this the "marginal" approach to learning with annotations. This has two advantages over minimising Equation 4.6. First, the latter has quadratic complexity, while the marginal approach has linear complexity (involving the solution of two independent logistic regression models). Second, with a base logistic regression model, the form of $\eta_{\mathcal{A}}, \eta_{\mathcal{B}}$ are given by non-linear softmax functions (Equation 5.9). Thus, the joint similarity function $s(a, b; w, v)$ is also a nonlinear function of the inputs a and b, even though in learning both $\eta_{\mathcal{A}}, \eta_{\mathcal{B}}$, we can resort to *linear* models. By contrast, the pairwise approach requires explicit nonlinearity in Φ.

6 Relation to Existing Approaches

We now carry out a deeper analysis between our proposed approaches and related work from the literature.

6.1 Comparison to CCA Given data matrices $A \in \mathbb{R}^{n \times d_a}, B \in \mathbb{R}^{n \times d_b}$, CCA can be shown to solve (see §9.2 of Supplementary Material):

$$\min_{U,V} \sum_{i=1}^{n} ||Ua^{(i)} - Vb^{(i)}||_2^2 : UA^TAU^T = I, VB^TBV^T = I.$$

Thus, CCA can be seen to employ the similarity

$$s(a, b; U, V) = \exp(-||Ua - Vb||_2^2) = \exp(a^T(U^TV)b) \tag{6.11}$$

trained to minimise the loss $\ell(1, v) = -\log v$. Comparing this to our approach (Equation 4.6), we see that both methods rely on bilinear scoring functions. A crucial difference is that we explicitly model the unmatched pairs, seeking to push them apart. Two other important differences are worth noting. First, CCA explicitly constraints the weights on the cross-features to be low rank, while we consider a full rank W, employing ℓ_2 regularisation to prevent overfitting. Second, our formulation is in terms of a generic loss function, which may be adapted to best fit the nature of a real-world dataset.

6.2 Comparison to CMML CMML [11] operates in the setting where there are explicitly labelled positive and negative pairs. Like CCA, it uses a scoring function of the form of Equation 6.11. Compared to our approach, an advantage of CMML is that the scoring function directly induces a metric over the latent subspace, which may be used e.g. for nearest neighbour queries. An advantage of our approach is that the objective is convex in the parameters. Further, the perspective of our approach is that one can create unlabelled samples from a generic cross-modal dataset, which allows us to use techniques suited to this problem, such as the bipartite ranking approach we discussed in §4.2.2.

The recent work of [9] is similar to CMML, but involves a convex objective with trace norm regularisation. Again, this method operates in the setting where there are explicitly labelled positive and negative pairs. Exploring the use of trace norm regularisation in our framework is an interesting direction for future work.

6.3 Comparison to hashing methods Hashing techniques such as [27] typically comprise two steps: from each modality one extracts a rich representation, which is then hashed to a lower-dimensional space. The second step is similar in spirit to CCA, albeit with potential computational benefits for indexing. The first step can be applied to any cross-modal retrieval method, including our own – that is, one can apply our supervised learning techniques after first clustering each modality – but leave exploring this to future work.

6.4 Comparison to SM Our "marginal approach" in the setting with annotations follows the semantic matching (SM) approach of [17], with one crucial difference: we compute the (unnormalised) inner product between the probability distributions for each modality representation. By contrast, [17] experiment with the cosine similarity and other metrics. Our analysis indicates that computing the inner product is theoretically optimal, and our experiments will illustrate this.

The idea of learning the marginal probabilities has been explored previously [25]. However, we have attempted to justify the optimality of the approach, and shown how it follows from a more general graphical model interpretation of cross-modal retrieval.

7 Experiments

We now compare our approaches to existing methods.

7.1 Datasets We evaluate all methods on three benchmark datasets[2] as used in [15, 17, 16]:

- WIKIPEDIA, containing 2,173 training and 693 test pairs, with a label from 10 semantic categories.
- PASCAL, containing 700 training and 300 test pairs, with labels from 20 categories.
- TVGRAZ, containing 1,558 training and 500 testing pairs with labels from 10 categories.

In all datasets, the text is represented by topic assignment distributions derived from Latent Dirichlet Allocation (LDA) [2] and the images are based by the scale invariant feature transformation (SIFT). For each dataset, we perform PCA to reduce the dimensions for text and image features to 10 and 128 respectively.

7.2 Methods compared We compare our methods against several baselines: Canonical correlation analysis (CCA), also known as content matching (CM) in [17]; Cross-modal metric learning (CMML) [11]; Semantic matching (SM) [17]; and Semantic correlation matching (SCM) [17]. We further explore the performance of random guessing, and multivariate linear regression between each modality. For each of these methods, following [17], we use negative cosine similarity as the distance metric to perform retrieval as per Equation 3.1.

We evaluate the following methods proposed in this paper: One-class SVM (OC-SVM), trained on both the concatenation and cross-product features; ℓ_2 regularised logistic regression (Equation 4.6), trained on both the concatenation and cross-product features; ℓ_2 regularised logistic regression over pairs (Equation 4.7), trained on only the cross-product features; SM with the cross-product features learned against the agreement of the ground-truth annotations (§5.1); and SM, SCM with the dot product used to provide the final joint similarity function (Equation 5.10).

We implemented all methods in MATLAB, with the following exceptions: for the methods in [17], we use the code provided by the authors; for OC-SVM, we use the LIBSVM implementation. Our code is available at http://users.cecs.anu.edu.au/~amenon/.

7.3 Evaluation Following [17], we evaluate the performance of all methods based on their mean average precision (MAP) scores. The MAP score is computed based on agreement of the corresponding ground truth annotations $y^{(i)}$; A MAP of 0.2 means that 1 in every 5 queries retrieves the correct result. We also report the training times for all methods, excluding the time required for pre-processing of the raw feature matrices (e.g. computation of the cross-product features). All times reported are wall-clock time on a 2.4GHz Intel iCore 7 Macbook Pro with 8GB of RAM.

7.4 Parameter tuning For each dataset, we further create a random holdout set comprising 25% of the training set. For each hyperparameter setting, we train the appropriate learner on the remaining 75% of the data, and evaluate performance on the holdout set. We repeat this procedure 5 times, and compute the average MAP score across each of these random holdout sets. We use this to select the best hyper parameter, which is then used for learning on the full training set.

For CCA, we tune the latent dimensionality $k \in \{1, 2, \ldots, 9\}$. For the OC-SVM, we select the parameter $\nu \in \{2^{-4}, 2^{-3}, \ldots, 2^{-1}\}$. For all methods using ℓ_2 regularisation, we select the strength of regularisation by $\lambda = 1/nC$, where n is the number of training instances, and C is selected from $\{10^{-8}, 10^{-3}, \ldots, 10^{4}\}$. For the methods relying on computing unlabelled pairs, we select the ratio of unlabelled to positive pairs, S, from $\{1, 5, 10\}$ (in supplementary material (§9.3), we assess the sensitivity of our method to S). Finally, for each option, we compare performance on the raw and ℓ_2 normalised versions of the input features.

7.5 Results Our results are presented in Tables 1 – 3. We separate the methods into those that rely on ground-truth annotations, and those that do not. Clearly, we expect the former to have consistently superior performance over the latter. Overall, the experiments consistently demonstrate the following:

- Our supervised method which learns with cross-product features is generally superior to other approaches that also do not use ground-truth annotation. In particular, we see MAP score improvements over CCA that are ∼2% on WIKIPEDIA, ∼10% on PASCAL, and ∼3% on TVGRAZ. We similarly outperform the CMML method on all datasets in terms of the average image and text MAP score.
- Our method of exploiting ground-truth annotations via a dot-product joint similarity outperforms the SM method based on cosine similarity, with MAP score improvements of ∼10% on WIKIPEDIA, ∼10% on PASCAL, and ∼5% on TVGRAZ.
- The runtimes of our methods, while higher than that of CCA, are comparable to that of CMML.
- Methods that exploit ground-truth annotation significantly outperform those that do not exploit this information. Interestingly, on WIKIPEDIA, learning with the cross-product features manages to outperform the SM and SCM methods.
- We do not find considerable difference between the

[2] See http://www.svcl.ucsd.edu/projects/crossmodal/ and http://www.svcl.ucsd.edu/~josecp/files/ris_cvpr12.zip.

logistic regression and bipartite ranking approaches to dealing with unlabelled data. This indicates that the assumptions underlying [4] are reasonable here.

Figure 3 shows the full precision-recall curves on the WIKIPEDIA dataset. (See supplementary material for other datasets.) We see that the marginal SM method generally dominates all others, except at very high levels of recall (i.e. at the bottom of the ranked list). Further, the cross-product method convincingly dominates CCA for image queries.

7.6 Case study We perform a case study on the WIKIPEDIA dataset, to assess the image and text retrieval performance of the Cross-Product and CCA methods. Figure 4 shows an example of an image retrieval query given text pertaining to a military aviator. We find that while the top two retrieved images by CCA are reasonable, the next two are less so compared to the Cross-Product method. Similarly, for the text retrieval query given an image of a mountain peak, the Cross-Product method returns several results relating to the Columbia river, which begins in the Rocky Mountains. CCA again tends to return some apparently irrelevant results, such as an article pertaining to primates.

8 Conclusion

We have proposed an approach to address the cross-modal retrieval problem using pairwise classification, both in the presence and absence of ground truth annotations. Beginning with a simple reduction to learning from positive-only instances, we show how to approach the problem using techniques from learning from positive and unlabelled samples, and bipartite ranking methods. On three large benchmark data sets, we have demonstrated that supervised learning approaches consistently outperform methods based on Canonical Correlation Analaysis (CCA) and its variants. We believe that supervised learning approaches provide a novel perspective for cross-modal retrieval.

Acknowledgements

This work was supported by NICTA. NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program.

References

[1] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David M. Blei, and Michael I. Jordan. Matching words and pictures. *JMLR*, 3:1107–1135, 2003.

[2] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[3] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):1–60, 2008.

[4] Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *SIGKDD '08*, pages 213–220, 2008.

[5] Hugo Jair Escalante, Carlos A. Hérnadez, Luis Enrique Sucar, and Manuel Montes. Late fusion of heterogeneous methods for multimedia image retrieval. In *ACM MIR*, pages 172–179, 2008.

[6] Xiaofei He, Wei-Ying Ma, and Hong-Jiang Zhang. Learning an image manifold for retrieval. In *ACM International Conference on Multimedia*, pages 17–23, 2004.

[7] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–377, 1936.

[8] Martial Hue and Jean-Philippe Vert. On learning with kernels for unordered pairs. In *ICML*, pages 463–470, 2010.

[9] Cuicui Kang, ShengCai Liao, Yonghao He, Jian Wang, Shiming Xiang, and Chunhong Pan. Cross-modal similarity learning : A low rank bilinear formulation. *CoRR*, abs/1411.4738, 2014.

[10] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. Multimodal neural language models. In Tony Jebara and Eric P. Xing, editors, *ICML '14*, pages 595–603. JMLR Workshop and Conference Proceedings, 2014.

[11] Alexis Mignon and Frédéric Jurie. CMML : a new metric learning approach for cross modal matching. In *Asian Conference on Computer Vision (ACCV)*, 2012.

[12] F. Monay and D. Gatica-Perez. Modeling semantic aspects for cross-media image indexing. *IEEE Transactions On Pattern Analysis and Machine Intelligence*, 29(10):1802–1817, 2007.

[13] M. Paramita, M. Sanderson, and P. Clough. Diversity in photo retrieval: Overview of the ImageCLEF 2009 photo task. *Multilingual Information Access Evaluation: Multimedia Experiments*, pages 45–59, 2010.

[14] Yuxin Peng, Zhiguo Yang, and Jianguo Xiao. Audio retrieval by segment-based manifold-ranking. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 686–689, 2009.

[15] Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Nikhil Rasiwasia, Gert R.G. Lanckriet, Roger Levy, and Nuno Vasconcelos. On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE Transactions On Pattern Analysis and Machine Intelligence*, 36(3):521–535, 2014.

[16] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147, 2010.

[17] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele

Regime	Method	Image query	Text query	Training time (secs)	Parameters
No ground truth	Random	0.1177	0.1178	0.02	NA
	Linear regression	0.2028	0.1459	0.12	$\lambda = 10^{-8}$
	CCA	0.2757	0.2002	0.04	$k = 5$
	CMML	0.2699	0.1964	188.36	$k = 10, S = 10$
	OC-SVM Concatenation	0.1321	0.1221	0.13	$\nu = 2^{-4}$
	OC-SVM Cross-Product	0.1664	0.1330	0.94	$\nu = 2^{-4}$
	Logistic Concatenation	0.1689	0.1207	0.35	$C = 10^{-7}, S = 10$
	Logistic Cross-Product	**0.2760**	**0.2118**	16.74	$C = 1, S = 10$
	Bipartite Cross-Product	0.2700	0.2068	70.19	$C = 10^{-1}, S = 10$
Ground truth	SM	0.2935	0.2278	0.28	$C = 10$
	SCM	0.2769	0.2200	0.10	$k = 9, C = 10000$
	Pairwise SM	0.2923	0.2131	10.41	$S = 10, C = 1$
	Marginal SM	**0.3328**	**0.2411**	0.20	$C = 1$
	Marginal SCM	0.3324	0.2257	0.08	$k = 9, C = 1000$

Table 1: MAP scores of various methods on Wikipedia dataset.

Regime	Method	Image query	Text query	Training time (secs)	Parameters
No ground truth	Random	0.0652	0.0652	0.00	NA
	Linear regression	0.1317	0.1069	0.03	$\lambda = 10^{-6}$
	CCA	0.1681	0.1422	0.01	$k = 3$
	CMML	0.1802	0.1431	18.12	$k = 10, S = 5$
	OC-SVM Concatenation	0.0744	0.0690	0.08	$\nu = 2^{-1}$
	OC-SVM Cross-Product	0.0960	0.0787	0.18	$\nu = 2^{-3}$
	Logistic Concatenation	0.0987	0.0681	0.09	$C = 10^{-1}, S = 1$
	Logistic Cross-Product	0.1797	**0.1688**	3.12	$C = 1, S = 10$
	Bipartite Cross-Product	**0.1818**	0.1610	3.98	$C = 1, S = 5$
Ground truth	SM	0.1984	0.1608	0.14	$C = 10$
	SCM	0.2054	0.1692	0.04	$k = 9, C = 10$
	Pairwise SM	0.1889	0.1630	0.91	$S = 5, C = 10$
	Marginal SM	**0.2229**	**0.1738**	0.14	$C = 10$
	Marginal SCM	0.2156	0.1671	0.03	$k = 7, C = 1$

Table 2: MAP scores of various methods on Pascal dataset.

Regime	Method	Image query	Text query	Training time (secs)	Parameters
No ground truth	Random	0.1129	0.1129	0.01	NA
	Linear regression	0.2535	0.2082	0.19	$\lambda = 10^{-7}$
	CCA	0.3230	**0.3041**	0.02	$k = 6$
	CMML	0.3252	0.2621	551.56	$k = 40, S = 10$
	OC-SVM Concatenation	0.1232	0.1217	0.36	$\nu = 2^{-2}$
	OC-SVM Cross-Product	0.1616	0.1697	1.12	$\nu = 2^{-3}$
	Logistic Concatenation	0.1318	0.1186	0.05	$C = 10^{-5}, S = 1$
	Logistic Cross-Product	**0.3333**	0.2925	17.71	$C = 1, S = 10$
	Bipartite Cross-Product	0.3299	0.2812	66.62	$C = 1, S = 10$
Ground truth	SM	0.4144	0.3928	0.21	$C = 10$
	SCM	0.3727	0.3710	0.05	$k = 9, C = 10^2$
	Pairwise SM	0.3453	0.3182	4.60	$S = 10, C = 1$
	Marginal SM	**0.4319**	**0.4209**	0.21	$C = 10$
	Marginal SCM	0.3999	0.3890	0.05	$k = 9, C = 10^4$

Table 3: MAP scores of various methods on TVGraz dataset.

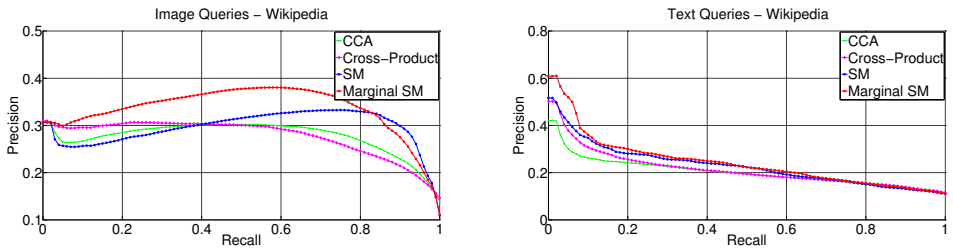

Figure 3: Precision-recall curves on WIKIPEDIA dataset. Our "Marginal SM" approach generally outperforms all other methods; our "Cross Product" approach similarly outperforms CCA. (See text for details.)

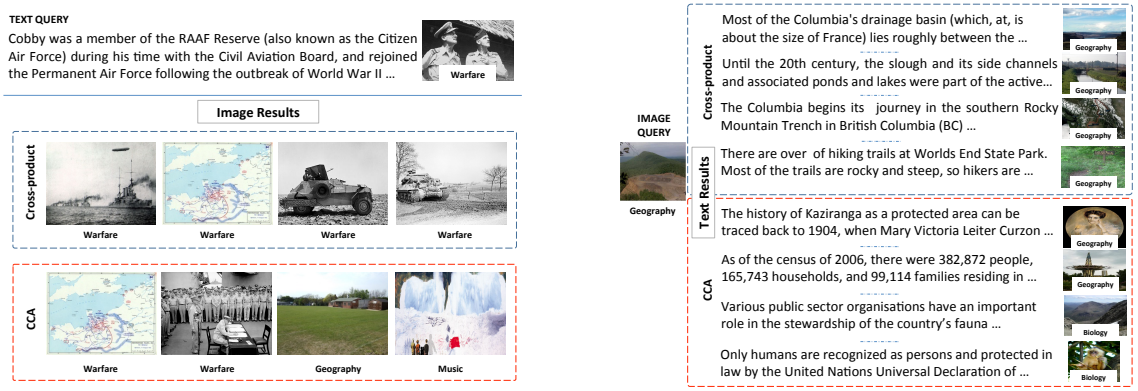

Figure 4: Cross-Product and CCA retrieval results on WIKIPEDIA dataset. **(Left)** Given a text (top), find images. The ground truth image is shown on the top right. **(Right)** Given an image (top), find texts. Here we present images that corresponding to the top retrieved texts.

Coviello, Gabriel Doyle, Gert R.G. Lanckriet, Roger Levy, and Nuno Vasconcelos. A new approach to cross-modal multimedia retrieval. In *MM*, pages 251–260, 2010.

[18] Gerard Salton. *The SMART Retrieval System-Experiments in Automatic Document Processing*. Prentice Hall, 1971.

[19] Bernhard Schölkopf, Robert C. Williamson, Alexander J. Smola, John Shawe-Taylor, and John Platt. Support vector method for novelty detection. In *NIPS*, pages 582–588, 2000.

[20] Sundararajan Sellamanickam, Priyanka Garg, and Sathiya Keerthi Selvaraj. A pairwise ranking based approach to learning with positive and unlabeled examples. In *CIKM '11*, pages 663–672, 2011.

[21] Arnold W. M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions On Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[22] T. Tsikrika and J. Kludas. Overview of the Wikipedia multimedia task at ImageCLEF 2008. *Evaluating Systems for Multilingual and Multimodal Information Access*, pages 539–550, 2009.

[23] J.-P. Vert and L. Jacob. Machine learning for in silico virtual screening and chemical genomics: New strategies. *Combinatorial Chemistry & High Throughput Screening*, 11(8):677–685, 2008.

[24] Yi Yang, Dong Xu, Feiping Nie, Jiebo Luo, and Yueting Zhuang. Ranking with local regression and global alignment for cross media retrieval. In *ACM International Conference on Multimedia*, pages 175–184, 2009.

[25] Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao. Effective heterogeneous similarity measure with nearest neighbors for cross-media retrieval. In *MMM*, pages 312–322, 2012.

[26] D. Zhang and Wee Sun Lee. Learning classifiers without negative examples: A reduction approach. In *International Conference on Digital Information Management (ICDIM)*, pages 638–643, 2008.

[27] Xiaofeng Zhu, Zi Huang, Heng Tao Shen, and Xin Zhao. Linear cross-modal hashing for efficient multimedia search. In *MM '13*, pages 143–152, 2013.

[28] Yue-Ting Zhuang, Yi Yang, and Fei Wu. Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval. *IEEE Transactions on Multimedia*, 10(2):221–229, 2008.

9 Supplementary Material

9.1 Marginalisation $\Pr[\mathsf{Z}=1|\mathsf{A},\mathsf{B}]$ is

(9.12)

$$\sum_{y,y'=1}^{K} \Pr[\mathsf{Z}=1|\mathsf{Y}=y,\mathsf{Y}'=y'] \cdot \Pr[\mathsf{Y}=y,\mathsf{Y}'=y'|\mathsf{A},\mathsf{B}]$$
$$= \sum_{y,y'=1}^{K} [\![y=y']\!] \cdot \eta_{\mathcal{A}}(\mathsf{A})_y \cdot \eta_{\mathcal{B}}(\mathsf{B})_{y'}$$
$$= \sum_{k=1}^{K} \eta_{\mathcal{A}}(\mathsf{A})_k \cdot \eta_{\mathcal{B}}(\mathsf{B})_k$$
$$= \langle \eta_{\mathcal{A}}(\mathsf{A}), \eta_{\mathcal{B}}(\mathsf{B}) \rangle$$

9.2 CCA objective CCA seeks the solution to

$$\max_{U,V} \sum_{i=1}^{n} \langle Ua^{(i)}, Vb^{(i)} \rangle : UA^TAU^T = I, VB^TBV^T = I,$$

where $U \in \mathbb{R}^{k \times d_a}, V \in \mathbb{R}^{k \times d_b}$. Observe that

$$||AU^T - BV^T||_F^2 = \text{tr}[UA^TAU^T] + \text{tr}[VB^TBV^T] - 2\text{tr}[UA^TBV^T],$$

where the first two terms are constants by virtue of the constraints, and the third term is the negation of the objective above. That is, we can rewrite CCA as

$$\min_{U,V} \sum_{i=1}^{n} ||Ua^{(i)} - Vb^{(i)}||_2^2 : UA^TAU^T = I, VB^TBV^T = I.$$

9.3 Subsampling sensitivity We now show the sensitivity of our methods to two subsampling options: the ratio of negative to positive pairs, S, and the fraction of cross-product features. Recall that our experiments tuned the parameter S, but used the entire set of cross-product features.

9.3.1 Ratio of positive to negative pairs Figure 5 shows the test MAP score of the cross-product method on the WIKIPEDIA dataset, as S is varied. While performance is suboptimal for $S = 1$ – corresponding to a single negative pair for each positive pair – for larger S, we see that the MAP score is quite stable, indicating low sensitivity to the parameter S.

9.3.2 Fraction of cross-product features We assess the effect of subsampling cross-product features to the MAP scores using the WIKIPEDIA dataset. Specifically, we randomly take $X\%$ of the resulted cross-product features to form the training and test datasets, where $X\% = \{100\%, 90\%, \ldots, 1\%\}$. Figure 6 shows that even until 80% reduction of the cross-product features, the method still possesses good performance.

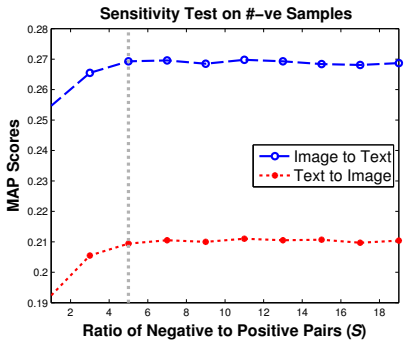

Figure 5: Sensitivity of cross-product method to ratio S on WIKIPEDIA dataset.

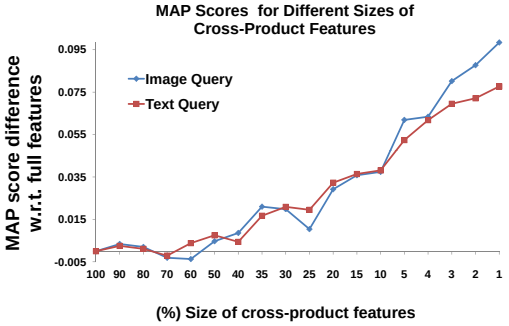

Figure 6: MAP scores changes on different sizes of cross-product features compared to using full cross-product features on WIKIPEDIA dataset.

9.4 Precision-recall curves Figures 7 and 8 show the precision-recall curves on the PASCAL and TVGRAZ datasets. We generally observe similar trends as on the WIKIPEDIA dataset.

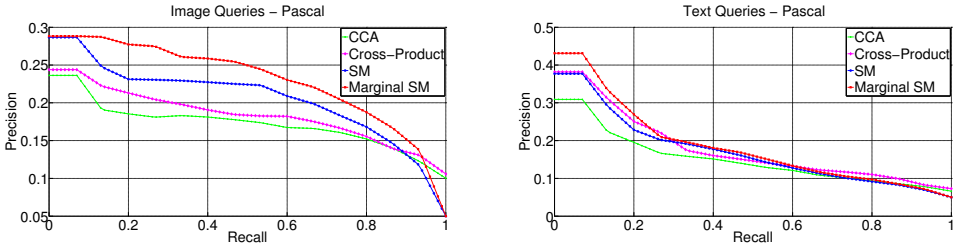

Figure 7: Precision-recall curves on PASCAL dataset.

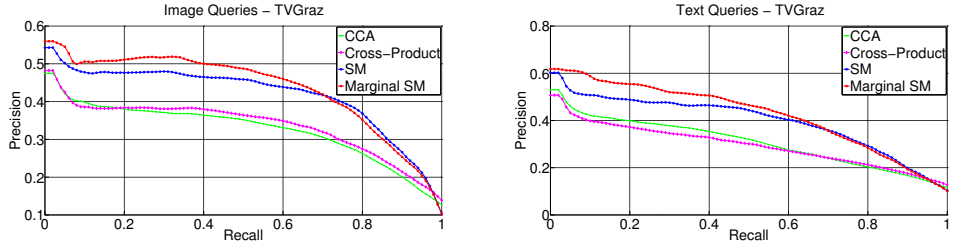

Figure 8: Precision-recall curves on TVGRAZ dataset.