Chapter 5

Cross Modal Multimedia Retrieval

5.1 Introduction

Over the last decade there has been a massive explosion of multimedia content on the web. This explosion has not been matched by an equivalent increase in the sophistication of multimedia content modeling technology. Today, the prevailing tools for searching multimedia repositories are still *uni-modal* in nature. Text repositories are searched with text queries, image databases with image queries, and so forth. To address this problem, the academic community has devoted itself to the design of models that can account for *multi-modal* data, i.e. data with multiple content modalities. Recently, there has been a surge of interest in multi-modal modeling, representation, and retrieval [106, 148, 132, 138, 28, 60, 31]. Multi-modal retrieval relies on queries combining multiple content modalities (*e.g.* the images and sound of a music video-clip) to retrieve database entries with the same combination of modalities (*e.g.* other music video-clips). These efforts have, in part, been spurred by a variety of large-scale research and evaluation experiments, such as TRECVID [132] and ImageCLEF [106, 148], involving datasets that span multiple data modalities. However, much of this work has focused on the straightforward extension of methods shown successful in the uni-modal scenario. Typically, the different modalities are fused into a representation that does not allow individual access to any of them, *e.g.* some form of dimensionality reduction of a large feature vector that concatenates measurements from images and text. Classical uni-modal techniques are then applied to the low-dimensional representation. This limits the applicability of the resulting multimedia models and retrieval systems.

An important requirement for further progress in these areas is the development of sophisticated joint models for multiple content modalities. In this chapter, we consider a richer interaction paradigm, which is denoted *cross-modal* retrieval. The goal is to build multi-modal content models that enable interactivity with content *across* modalities. Such models can then be used to design *cross-modal retrieval systems,* where queries from one modality (*e.g.* video) can be matched to database entries from another (*e.g.*, the best accompanying audio-track). This form of retrieval can be seen as a generalization of current content labeling systems, where one dominant modality is augmented with simple information from another, which can be subsequently searched. Examples include keyword-based image [4, 97, 21] and song [151, 149, 89, 36] retrieval systems. One property of cross-modal retrieval is that, by definition, it requires *representations that generalize across content modalities*. This implies the ability to establish cross-modal links between the attributes (of different modalities) characteristic of each document, or document class. Detecting these links requires much deeper content understanding than the classical matching of uni-modal attributes. For example, while an image retrieval system can retrieve images of roses by matching red blobs, and a text retrieval system can retrieve texts about roses by matching the "rose" word, a cross-modal retrieval system must *abstract* that the word "rose" matches the visual attribute "red blob". This is much closer to what humans do than simple color or word matching. Hence, cross-modal retrieval is a better context than uni-modal retrieval for the study of fundamental hypotheses on multimedia modeling.

We exploit this property to study two hypotheses on the joint modeling of images and text. The first, denoted the *correlation hypothesis*, is that explicit modeling of low-level correlations between the different modalities is of importance for the success of the joint models. The second, denoted the *abstraction hypothesis*, is that the modeling benefits from semantic abstraction, *i.e.*, the representation of images and text in terms of semantic (rather than low-level) descriptors. These hypotheses are partly motivated by previous evidence that correlation, *e.g.*, correlation analysis on fMRI [55], and abstraction, *e.g.*, hierarchical topic models for text clustering [14] or semantic representations for image retrieval(see Chapter 3), improve performance on uni-modal retrieval tasks. Three joint image-text models that exploit low-level correlation, denoted *correlation matching*, semantic abstraction, denoted *semantic matching*, and both, denoted *semantic correlation matching*, are introduced. Both semantic matching and semantic correlation matching build upon the proposed semantic image representation (see Chapter 2).

The hypotheses are tested by measuring the retrieval performance of these models on two reciprocal cross-modal retrieval tasks: 1) the retrieval of text documents in response to a query image, and 2) the retrieval of images in response to a query text. These are basic cross-modal retrieval problems, central to many applications of practical interest, such as finding pictures that effectively illustrate a given text (*e.g.*, to illustrate a page of a story book), finding the texts that best match a given picture (*e.g.*, a set of vacation accounts about a given landmark), or searching using a combination of text and images. Model performance on these tasks is evaluated with two datasets: TVGraz [66] and a novel dataset based on Wikipedia's featured articles. These experiments show independent benefits to both correlation modeling and abstraction. In particular, best results are obtained by a model that accounts for both low-level correlations — by performing a kernel canonical correlation analysis (KCCA) [127, 163] — and semantic abstraction — by projecting images and texts into a common semantic space (see Chapter 2) designed with logistic regression. This suggests that the abstraction and correlation hypotheses are complementary, each improving the modeling in a different manner. Individually, the gains of abstraction are larger than those of correlation modeling.

This chapter is organized as follows. Section 5.2 discusses previous work in multi-modal and cross-modal multimedia modeling. Section 5.3 presents a mathematical formulation for cross-modal modeling and discusses the two fundamental hypotheses analyzed in this work. Section 5.4 introduces the models underlying correlation, semantic, and semantic correlation matching. Section 5.5 discusses the experimental setup used to evaluate the hypotheses. Model validation and parameter tuning are detailed in Section 5.6. The hypotheses are finally tested in Section 5.7.

5.2 Previous Work

The problems of image and text retrieval have been the subject of extensive research in the fields of information retrieval, computer vision, and multimedia [28, 133, 132, 106, 93]. In all these areas, the emphasis has been on *uni-modal* approaches, where query and retrieved documents share a single modality [125, 124, 156, 28, 133]. For example, in [124], a query text and in [156], a query image is used to retrieve similar text documents and images, based on low-level text (e.g., words) and image (e.g., DCTs) representations, respectively. However, this is not effective for all problems. For example, the existence of a well known *semantic gap,* (see Chapter 1) between current image representations and those adopted by humans, severely limits the performance of uni-modal image retrieval systems [133](see Chapter 3).

In general, successful retrieval from large-scale image collections requires that the latter be augmented with text metadata provided by human annotators. These manual annotations are typically in the form of a few keywords, a small caption, or a brief image description [106, 148, 132]. When this metadata is available, the retrieval operation tends to be uni-modal and ignore the images — the text metadata of the query image is simply matched to the text metadata available for images in the database. Because manual image labeling is labor-intensive, recent research has addressed the problem of automatic image labeling [1] [21, 63, 41, 73, 96, 4]. As we saw in Chapter 2, rather than labeling images with a small set of most relevant semantic concepts, images can be represented as a weighted combination of all concepts in the vocabulary, by projecting them into a *semantic space*, where each dimension is a semantic concept. Semantic space was used for uni-modal image retrieval in Chapter 3, which enabled retrieval of images using *semantic similarity* — by combining the semantic space with a suitable similarity function.

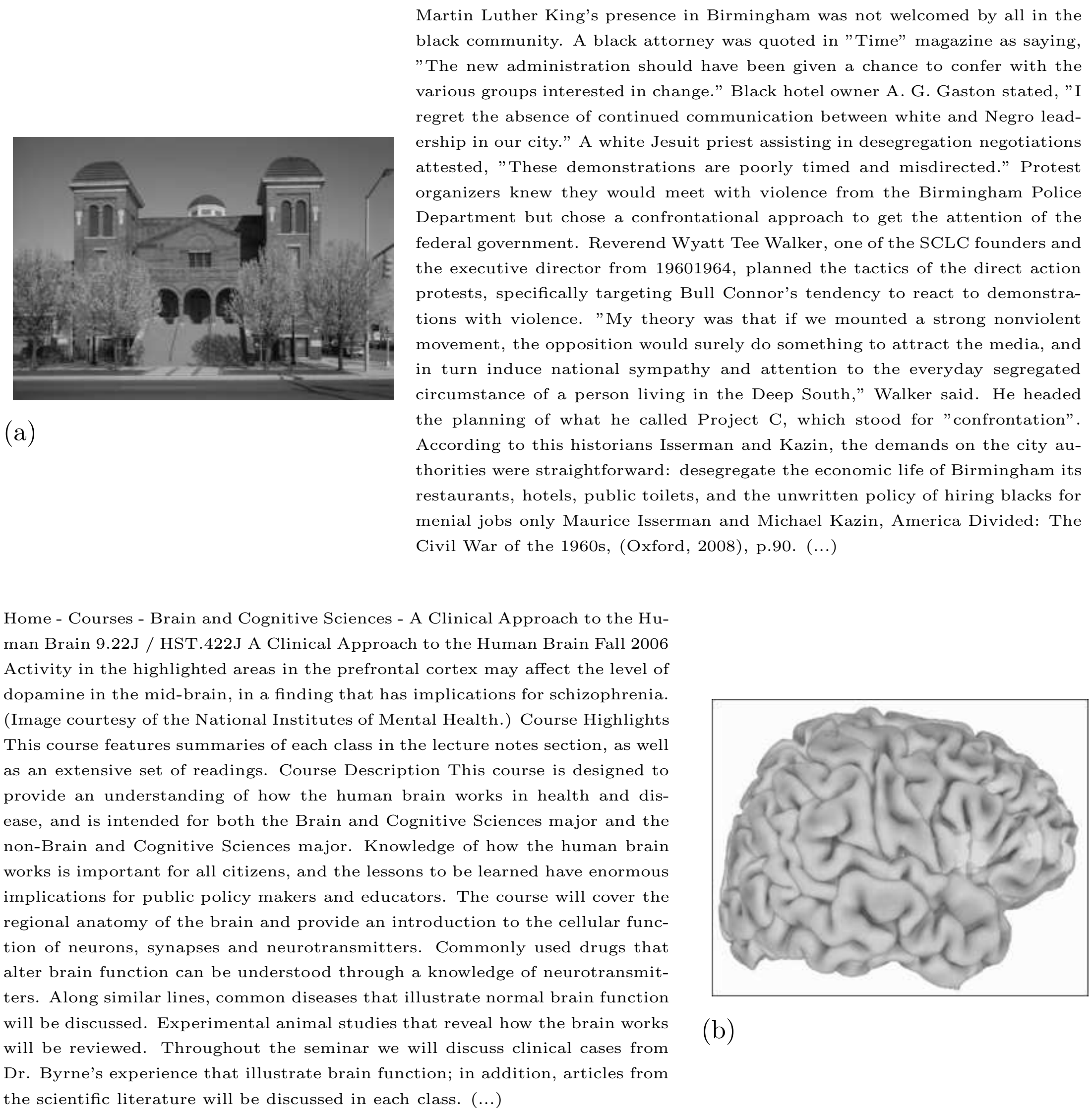
In parallel, advances have been reported in the area of *multi-modal* retrieval systems [106, 148, 132, 138, 28, 60, 31]. These are extensions of the classic uni-modal systems, where a common retrieval system integrates information from various modalities. This can be done by fusing features from different modalities into a single vector [171, 108, 37], or by learning different models for different modalities and fusing their predictions [168, 69]. One popular approach is to concatenate features from different modalities into a common vector and rely on unsupervised structure discovery algorithms, such as latent semantic analysis (LSA), to find statistical patterns that span the different modalities. A good overview of these methods is given in [37], which also discusses the combination of uni-modal and

[1] Although not commonly perceived as being *cross-modal*, these systems support cross-modal retrieval, e.g., by returning images in response to explicit text queries.

multi-modal retrieval systems. Multi-modal integration has also been applied to retrieval tasks including audio-visual content [99, 44]. In general, the inability to access each data modality individually (after the fusion of modalities) limits the applicability of these systems to cross-modal retrieval.

Recently, there has been progress towards multi-modal systems that do not suffer from this limitation. These include retrieval methods for corpora of images and text [31], images and audio [178, 76], text and audio [131], or images, text, and audio [175, 178, 182, 181, 176]. One popular approach is to rely on graph-based manifold learning techniques [175, 178, 182, 181, 176]. These methods learn a manifold from a matrix of distances between multi-modal objects. The multi-modal distances are formulated as a function of the distances between individual modalities, which allows to single out particular modalities or ignore missing ones. Retrieval then consists of finding the nearest document, on the manifold, to a multimedia query (which can be composed of any subset of modalities). The main limitation of methods in this class is the lack of out-of-sample generalization. Since there is no computationally efficient way to project the query into the manifold, queries are restricted to the training set used to learn the latter. Hence, all unseen queries must be mapped to their nearest neighbors in this training set, defeating the purpose of manifold learning. An alternative solution is to learn correlations between different modalities [76, 178, 164]. For example, [76] compares canonical correlation analysis (CCA) and cross-modal factor analysis (CFA) in the context of audio-image retrieval. Both CCA and CFA perform a joint dimensionality reduction that extracts highly correlated features in the two data modalities. A kernelized version of CCA was also proposed in [164] to extract translation invariant semantics of text documents written in multiple languages. It was later used to model correlations between web images and corresponding captions, in [55].

Despite these advances in multi-modal modeling, current approaches tend to rely on a limited textual representation, in the form of keywords, captions, or small text snippets. We refer to all of these as forms of *light annotation*. This is at odds with the ongoing explosion of multimedia content on the web, where it is now possible to collect large sets of extensively annotated data. Examples



Martin Luther King's presence in Birmingham was not welcomed by all in the black community. A black attorney was quoted in "Time" magazine as saying, "The new administration should have been given a chance to confer with the various groups interested in change." Black hotel owner A. G. Gaston stated, "I regret the absence of continued communication between white and Negro leadership in our city." A white Jesuit priest assisting in desegregation negotiations attested, "These demonstrations are poorly timed and misdirected." Protest organizers knew they would meet with violence from the Birmingham Police Department but chose a confrontational approach to get the attention of the federal government. Reverend Wyatt Tee Walker, one of the SCLC founders and the executive director from 19601964, planned the tactics of the direct action protests, specifically targeting Bull Connor's tendency to react to demonstrations with violence. "My theory was that if we mounted a strong nonviolent movement, the opposition would surely do something to attract the media, and in turn induce national sympathy and attention to the everyday segregated circumstance of a person living in the Deep South," Walker said. He headed the planning of what he called Project C, which stood for "confrontation". According to this historians Isserman and Kazin, the demands on the city authorities were straightforward: desegregate the economic life of Birmingham its restaurants, hotels, public toilets, and the unwritten policy of hiring blacks for menial jobs only Maurice Isserman and Michael Kazin, America Divided: The Civil War of the 1960s, (Oxford, 2008), p.90. (...)

(a)

Home - Courses - Brain and Cognitive Sciences - A Clinical Approach to the Human Brain 9.22J / HST.422J A Clinical Approach to the Human Brain Fall 2006 Activity in the highlighted areas in the prefrontal cortex may affect the level of dopamine in the mid-brain, in a finding that has implications for schizophrenia. (Image courtesy of the National Institutes of Mental Health.) Course Highlights This course features summaries of each class in the lecture notes section, as well as an extensive set of readings. Course Description This course is designed to provide an understanding of how the human brain works in health and disease, and is intended for both the Brain and Cognitive Sciences major and the non-Brain and Cognitive Sciences major. Knowledge of how the human brain works is important for all citizens, and the lessons to be learned have enormous implications for public policy makers and educators. The course will cover the regional anatomy of the brain and provide an introduction to the cellular function of neurons, synapses and neurotransmitters. Commonly used drugs that alter brain function can be understood through a knowledge of neurotransmitters. Along similar lines, common diseases that illustrate normal brain function will be discussed. Experimental animal studies that reveal how the brain works will be reviewed. Throughout the seminar we will discuss clinical cases from Dr. Byrne's experience that illustrate brain function; in addition, articles from the scientific literature will be discussed in each class. (...)

(b)

Figure 5.1: Two examples of image-text pairs: (a) section from the Wikipedia article on the Birmingham campaign ("History" category), (b) part of a Cognitive Science class syllabus from the TVGraz dataset ("Brain" category).

include news archives, blog posts, or Wikipedia pages, where pictures are related to complete text articles, not just a few keywords. We refer to these datasets as *richly annotated*. While potentially more informative, rich annotation establishes a much more nuanced connection between images and text than that of *light annotation*. Indeed, keywords usually are explicit image labels and, therefore, clearly relate to it, while many of the words in rich text may be unrelated to the image used to illustrate it. For example, Figure 5.1a shows a section of the Wikipedia article on the "Birmingham campaign", along with the associated image. Notice that, although related to the text, the image is clearly not representative of all the words in the article. The same is true for the web-page in Figure 5.1b, from the TVGraz dataset [66] (see Appendix A for more details on both Wikipedia and TVGraz datasets). This is a course syllabus that, beyond the pictured brain, includes course information and other unrelated matters. A major long-term goal of modeling richly annotated data is to recover this *latent* relationship between the text and image components of a document, and exploit it in benefit of practical applications.

5.3 Fundamental Hypotheses

In this section, we present a novel multi-modal content modeling framework, which is flexible and applicable to rich content modalities. Although the fundamental ideas are applicable to any combination of modalities we restrict the discussion to documents containing images and text.

5.3.1 The problem

We consider the problem of information retrieval from a database $\mathcal{B} = \{D_1, \ldots, D_{|\mathcal{B}|}\}$ of *documents* comprising *image* and *text* components. In practice, these documents can be quite diverse: from documents where a single text is complemented by one or more images (*e.g.*, a newspaper article) to documents containing multiple pictures and text sections (*e.g.*, a Wikipedia page). For simplicity, we consider the case where each document consists of a single *image* and its

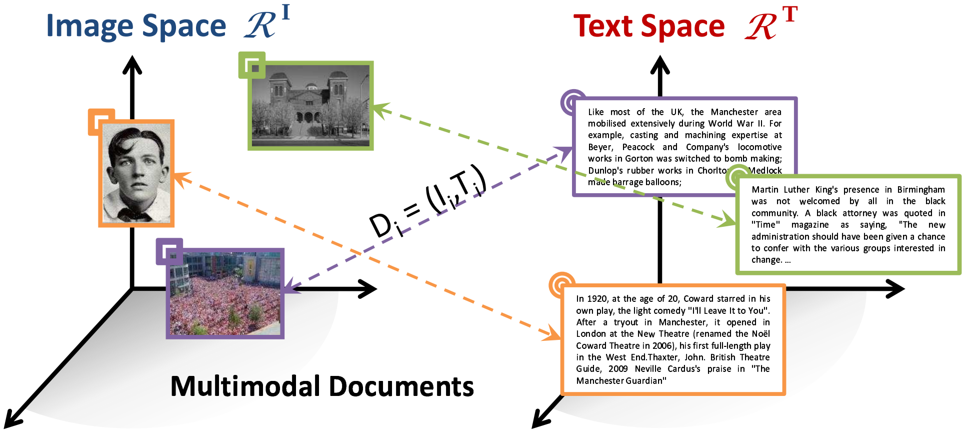


Image Space $\mathcal{R}^{\mathrm{I}}$
Text Space $\mathcal{R}^{\mathrm{T}}$
$D_i = (I_i, T_i)$
Multimodal Documents

Like most of the UK, the Manchester area mobilised extensively during World War II. For example, casting and machining expertise at Beyer, Peacock and Company's locomotive works in Gorton was switched to bomb making; Dunlop's rubber works in Chorlton-on-Medlock made barrage balloons;

Martin Luther King's presence in Birmingham was not welcomed by all in the black community. A black attorney was quoted in "Time" magazine as saying, "The new administration should have been given a chance to confer with the various groups interested in change. ...

In 1920, at the age of 20, Coward starred in his own play, the light comedy "I'll Leave It to You". After a tryout in Manchester, it opened in London at the New Theatre (renamed the Noël Coward Theatre in 2006), his first full-length play in the West End.Thaxter, John. British Theatre Guide, 2009 Neville Cardus's praise in "The Manchester Guardian"

Figure 5.2: Each document (D_i) consists of an *image* (I_i) and accompanying *text* (T_i), *i.e.*, $D_i = (I_i, T_i)$, which are represented as vectors in feature spaces $\Re^I$ and $\Re^T$, respectively. Documents establish a one-to-one mapping between points in $\Re^I$ and $\Re^T$.

accompanying *text*, *i.e.*, $D_i = (I_i, T_i)$. Images and text are represented as vectors in feature spaces $\Re^I$ and $\Re^T$ respectively[2], as illustrated in Figure 5.2, documents establish a one-to-one mapping between points in $\Re^I$ and $\Re^T$. Given a text (image) query $T_q \in \Re^T$ ($I_q \in \Re^I$), the goal of *cross-modal retrieval* is to return the closest match in the image (text) space $\Re^I$ ($\Re^T$).

5.3.2 Multi-modal modeling

Whenever the image and text spaces have a natural correspondence, cross-modal retrieval reduces to a classical retrieval problem. Let

$$\mathcal{M} : \Re^T \to \Re^I$$

be an invertible mapping between the two spaces. Given a query T_q in $\Re^T$, it suffices to find the nearest neighbor to $\mathcal{M}(T_q)$ in $\Re^I$. Similarly, given a query I_q in $\Re^I$, it suffices to find the nearest neighbor to $\mathcal{M}^{-1}(I_q)$ in $\Re^T$. In this case,

[2] Note that, in this chapter we deviate from the standard representation of an image (adopted in this work) as a bag of N feature vectors, $\mathcal{I} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}, \mathbf{x}_i \in \mathcal{X}$, to one where an image is represented as a vector in $\Re^I$. The motivation is to maintain a simple and consistent representation across all different modalities. See Section 2.1.1 for a brief description on mapping images from $\mathcal{X}^N$ to $\Re^I$

the design of a cross-modal retrieval system reduces to the design of an effective similarity function for determining the nearest neighbors.

In general, however, different representations are adopted for images and text, and there is no natural correspondence between $\Re^I$ and $\Re^T$. In this case, the mapping $\mathcal{M}$ has to be learned from examples. In this work, we map the two representations into intermediate spaces, $\mathcal{V}^I$ and $\mathcal{V}^T$, that have a natural correspondence. First, consider learning invertible mappings

$$\mathcal{M}_I : \Re^I \rightarrow \mathcal{V}^I \qquad\qquad \mathcal{M}_T : \Re^T \rightarrow \mathcal{V}^T$$

from each of the image and text spaces to two *isomorphic* spaces $\mathcal{V}^I$ and $\mathcal{V}^T$, such that there is an invertible mapping

$$\mathcal{M} : \mathcal{V}^T \rightarrow \mathcal{V}^I$$

between these two spaces. In this case, given a text query T_q in $\Re^T$, cross-modal retrieval reduces to finding the nearest neighbor of

$$\mathcal{M}_I^{-1} \circ \mathcal{M} \circ \mathcal{M}_T(T_q)$$

in $\Re^I$. Similarly, given an image query I_q in $\Re^I$, the goal is to find the nearest neighbor of

$$\mathcal{M}_T^{-1} \circ \mathcal{M}^{-1} \circ \mathcal{M}_I(I_q)$$

in $\Re^T$. This formulation can be generalized to learning non-invertible mappings $\mathcal{M}_I$ and $\mathcal{M}_T$ by seeking the nearest neighbors of $\mathcal{M} \circ \mathcal{M}_T(T_q)$ and $\mathcal{M}^{-1} \circ \mathcal{M}_I(I_q)$ in the intermediate spaces $\mathcal{V}^I$ and $\mathcal{V}^T$, respectively, and matching them up with the corresponding image and text, in $\Re^I$ and $\Re^T$. Under this formulation, followed in this work, the main problem in the design of a cross-modal retrieval system is the design of the intermediate spaces $\mathcal{V}^I$ and $\mathcal{V}^T$ (and the corresponding mappings $\mathcal{M}_I$ and $\mathcal{M}_T$).

5.3.3 The fundamental hypotheses

Since the goal is to design *representations that generalize across content modalities,* the solution of this problem requires some ability to derive a more

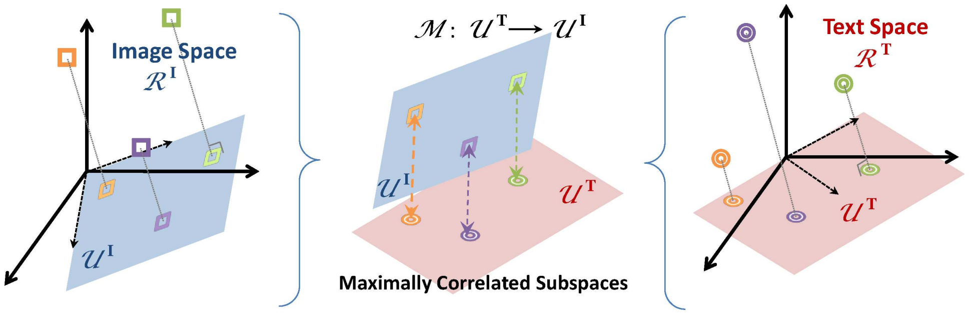


Image Space
$\mathcal{R}^{\mathrm{I}}$
$\mathcal{U}^{\mathrm{I}}$
$\mathcal{M}: \mathcal{U}^{\mathrm{T}} \longrightarrow \mathcal{U}^{\mathrm{I}}$
$\mathcal{U}^{\mathrm{I}}$
$\mathcal{U}^{\mathrm{T}}$
Maximally Correlated Subspaces
Text Space
$\mathcal{R}^{\mathrm{T}}$
$\mathcal{U}^{\mathrm{T}}$

Figure 5.3: Correlation matching (CM) performs joint feature selection in the text and image spaces, projecting them onto two maximally correlated subspaces $\mathcal{U}_T$ and $\mathcal{U}_I$.

abstract representation than the sum of the parts (low-level features) extracted from each content modality. Given that such abstraction is the hallmark of true image or text *understanding,* this problem enables the exploration of some central questions in multimedia modeling. Considering a query for "swan" 1) a uni-modal image retrieval system can successfully retrieve images of "swans" in that they are the only white objects in a database, 2) a text retrieval system can successfully retrieve documents about "swans" because they are the only documents containing the word "swan", and 3) a multi-modal retrieval system can just match "white" to "white" and "swan" to "swan", a cross-modal retrieval system cannot solve the task without *abstracting* that "white is a visual attribute of swan". Hence, cross-modal retrieval is a more effective paradigm for testing fundamental hypotheses in multimedia representation than uni-modal or multi-modal retrieval. In this work, we exploit the cross-modal retrieval problem to test two such hypotheses regarding the joint modeling of images and text.

- $\mathcal{H}_1$ (**correlation** hypothesis): low-level cross-modal correlations are important for joint image-text modeling.

- $\mathcal{H}_2$ (**abstraction** hypothesis): semantic abstraction is important for joint image-text modeling.

The hypotheses are tested by comparing three possibilities for the design of the intermediate spaces $\mathcal{V}^I$ and $\mathcal{V}^T$ of cross-modal retrieval. In the first case, two

Table 5.1: Taxonomy of the proposed approaches to cross-modal retrieval.

	correlation hypothesis	abstraction hypothesis
CM	$\surd$	
SM		$\surd$
SCM	$\surd$	$\surd$

feature transformations map $\Re^I$ and $\Re^T$ onto *correlated* d-dimensional *subspaces* denoted as $\mathcal{U}^I$ and $\mathcal{U}^T$, respectively, which act as $\mathcal{V}^I$ and $\mathcal{V}^T$. This maintains the level of semantic abstraction of the representation while maximizing the correlation between the two spaces. We refer to this matching technique as *correlation matching* (CM). In the second case, a pair of transformations are used to map the image and text spaces into a pair of *semantic spaces* $\mathcal{S}^I$ and $\mathcal{S}^T$, which then act as $\mathcal{V}^I$ and $\mathcal{V}^T$. This increases the semantic abstraction of the representation without directly seeking correlation maximization. The spaces $\mathcal{S}^I$ and $\mathcal{S}^T$ are made isomorphic by using the same set of semantic concepts for both modalities. We refer to this as *semantic matching* (SM). Finally, a third approach combines the previous two techniques: project onto maximally correlated subspaces $\mathcal{U}^I$ and $\mathcal{U}^T$, and then project again onto a pair of semantic spaces $\mathcal{S}^I$ and $\mathcal{S}^T$, which act as $\mathcal{V}^I$ and $\mathcal{V}^T$. We refer to this as *semantic correlation matching* (SCM).

5.1 summarizes which hypotheses hold for each of the three approaches. The comparative evaluation of the performance of these approaches on cross-modal retrieval experiments provides indirect evidence for the importance of the above hypotheses to the joint modeling of images and text. The intuition is that a better cross-modal retrieval performance results from a more effective joint modeling.

5.4 Cross-modal Retrieval

In this section, we present each of the three approaches in detail.

5.4.1 Correlation matching (CM)

The design of a mapping from $\Re^T$ and $\Re^I$ to the correlated spaces $\mathcal{U}^T$ and $\mathcal{U}^I$ requires a combination of dimensionality reduction and some measure of correlation between the text and image modalities. In both text and vision literatures, dimensionality reduction is frequently accomplished with methods such as latent semantic indexing (LSI) [29] and principal component analysis (PCA) [64]. These are members of a broader class of learning algorithms, denoted subspace learning, which are computationally efficient, and produce linear transformations that are easy to conceptualize, implement, and deploy. Furthermore, because subspace learning is usually based on second order statistics, such as correlation, it can be easily extended to the multi-modal setting and kernelized. This has motivated the introduction of a number of multi-modal subspace methods in the literature. In this work, we consider *cross-modal factor analysis* (CFA), *canonical correlation analysis* (CCA), and *kernel canonical correlation analysis* (KCCA). All these methods include a training stage, where the subspaces $\mathcal{U}^I$ and $\mathcal{U}^T$ are learned, followed by a projection stage, where images and text are projected into these spaces. Figure 5.3 illustrates this process. Cross-modal retrieval is finally performed within the low-dimensional subspaces.

Linear subspace learning

CFA seeks transformations that best represent coupled patterns between different subsets of features (e.g., different modalities) describing the same objects [76]. It finds the orthonormal transformations Ω_I and Ω_T that project the two modalities onto a shared space, $\mathcal{U}^I = \mathcal{U}^T = \mathcal{U}$, where the projections have minimum distance

$$\left\|X_I\Omega_I - X_T\Omega_T\right\|_F^2. \tag{5.1}$$

X_I and X_T are matrices containing corresponding features from the image and text domains, and $||\cdot||_F^2$ is the Frobenius norm. It can be shown that this is equivalent to maximizing

$$trace(X_I\Omega_I\Omega_T'X_T'), \tag{5.2}$$

and the optimal matrices Ω_I, Ω_T can be obtained by a singular value decomposition of the matrix $X_I' X_T$, *i.e.*,

$$X_I' X_T = \Omega_I \Lambda \Omega_T, \tag{5.3}$$

where Λ is the matrix of singular values of $X_I' X_T$ [76].

CCA [59] learns the d-dimensional subspaces $\mathcal{U}^I \subset \Re^I$ (image) and $\mathcal{U}^T \subset \Re^T$ (text) where the correlation between the two data modalities is maximal. It is similar to principal components analysis (PCA), in the sense that it learns a basis of canonical components, directions $w_i \in \Re^I$ and $w_t \in \Re^T$, but seeks directions along which the data is maximally correlated

$$\max_{w_i \neq 0,\, w_t \neq 0} \frac{w_i' \Sigma_{IT} w_t}{\sqrt{w_i' \Sigma_I w_i}\sqrt{w_t' \Sigma_T w_t}} \tag{5.4}$$

where Σ_I and Σ_T are the empirical covariance matrices for images $\{I_1, \ldots, I_{|D|}\}$ and text $\{T_1, \ldots, T_{|D|}\}$ respectively, and $\Sigma_{IT} = \Sigma_{TI}'$ the cross-covariance between them. Repetitively solving (5.4), for directions that are orthogonal to all previously obtained solutions, provides a series of canonical components. It can be shown that the canonical components in the image space can be found as the eigenvectors of $\Sigma_I^{-1/2} \Sigma_{IT} \Sigma_T^{-1} \Sigma_{TI} \Sigma_I^{-1/2}$, and in the text space as the eigenvectors of $\Sigma_T^{-1/2} \Sigma_{TI} \Sigma_I^{-1} \Sigma_{IT} \Sigma_T^{-1/2}$. The first d eigenvectors $\{w_{i,k}\}_{k=1}^d$ and $\{w_{t,k}\}_{k=1}^d$ define a basis of the subspaces $\mathcal{U}^I$ and $\mathcal{U}^T$.

Non-linear subspace learning

CCA and CFA can only model linear dependencies between image and text features. This limitation can be avoided by mapping these features into high-dimensional spaces, with a pair of non-linear transformations $\phi_T : \Re^T \rightarrow \mathcal{F}^T$ and $\phi_I : \Re^I \rightarrow \mathcal{F}^I$. Application of CFA or CCA in these spaces can then recover complex patterns of dependency in the original feature space. As is common in machine learning, the transformations $\phi_T(\cdot)$ and $\phi_I(\cdot)$ are computed only implicitly, by the introduction of two kernel functions $\mathcal{K}_T(\cdot, \cdot)$ and $\mathcal{K}_I(\cdot, \cdot)$, specifying the inner products in $\mathcal{F}^T$ and $\mathcal{F}^I$, i.e., $\mathcal{K}_T(T_m, T_n) = \langle \phi_T(T_m), \phi_T(T_n) \rangle$ and $\mathcal{K}_I(I_m, I_n) = \langle \phi_I(I_m), \phi_I(I_n) \rangle$, respectively.

KCCA [127, 163] implements this type of extension for CCA, seeking directions $w_i \in \mathcal{F}^I$ and $w_t \in \mathcal{F}^T$, along which the two modalities are maximally correlated in the transformed spaces. The canonical components can be found by solving

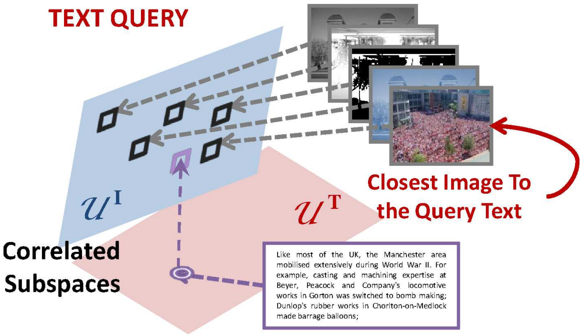
$$\max_{\alpha_i \neq 0,\, \alpha_t \neq 0} \frac{\alpha_i' K_I K_T \alpha_t}{V(\alpha_i, K_I) V(\alpha_t, K_T)}, \tag{5.5}$$

where $V(\alpha, K) = \sqrt{(1-\kappa)\alpha' K^2 \alpha + \kappa \alpha' K \alpha}$, $\kappa \in [0, 1]$ is a regularization parameter, and K_I and K_T are the kernel matrices of the image and text representations, *e.g.*, $(K_I)_{mn} = \mathcal{K}_I(I_m, I_n)$. Given optimal α_i and α_t for (5.5), w_i and w_t are obtained as linear combinations of the training examples $\{\phi_I(I_k)\}_{k=1}^{|\mathcal{B}|}$, and $\{\phi_T(T_k)\}_{k=1}^{|\mathcal{B}|}$, with α_i and α_t as weight vectors, i.e., $w_i = \Phi_I(X_I)^T \alpha_i$ and $w_t = \Phi_T(X_T)^T \alpha_t$, where $\Phi_I(X_I)$ $(\Phi_T(X_T))$ is the matrix whose rows contain the high-dimensional representation of the image (text) features. To optimize (5.5), we solve a generalized eigenvalue problem using the software package of [163]. The first d generalized eigenvectors provide us with d weight vectors $\{\alpha_{i,k}\}_{k=1}^d$ and $\{\alpha_{t,k}\}_{k=1}^d$, from which bases, $\{w_{i,k}\}_{k=1}^d$ and $\{w_{t,k}\}_{k=1}^d$, of the two maximally correlated d-dimensional subspaces $\mathcal{U}^I \subset \mathcal{F}^I$ and $\mathcal{U}^T \subset \mathcal{F}^T$ can be derived, with $1 \leq d \leq |\mathcal{B}|$.

Image and text projections

Images and text are represented by their projections p_I and p_T onto the subspaces $\mathcal{U}^I$ and $\mathcal{U}^T$, respectively. p_I (p_T) is obtained by computing the dot-products between the vector representing the image (text) $I \in \Re^I$ $(T \in \Re^T)$ and the image (text) basis vectors spanning $\mathcal{U}^I$ $(\mathcal{U}^T)$. For CFA, the basis vectors are the columns of Ω_I and Ω_T, respectively. For CCA, they are $\{w_{i,k}\}_{k=1}^d$ and $\{w_{t,k}\}_{k=1}^d$. In the case of KCCA, an image $I \in \Re^I$ is first mapped into $\mathcal{F}^I$ and subsequently projected onto $\{w_{i,k}\}_{k=1}^d$, *i.e.*, $p_I = \mathcal{P}_I(\phi_I(I))$ with

$$\begin{aligned} p_{I,k} &= \langle \phi_I(I), w_{i,k} \rangle \\ &= \langle \phi_I(I), \left[\phi_I(I_1), \ldots, \phi_I(I_{|\mathcal{B}|})\right] \alpha_{i,k} \rangle \\ &= \left[\mathcal{K}_I(I, I_1), \ldots, \mathcal{K}_I\left(I, I_{|\mathcal{B}|}\right)\right] \alpha_{i,k}, \end{aligned} \tag{5.6}$$


TEXT QUERY
$\mathcal{U}^{\mathrm{I}}$
$\mathcal{U}^{\mathrm{T}}$
Correlated Subspaces
Closest Image To the Query Text

Like most of the UK, the Manchester area mobilised extensively during World War II. For example, casting and machining expertise at Beyer, Peacock and Company's locomotive works in Gorton was switched to bomb making; Dunlop's rubber works in Chorlton-on-Medlock made barrage balloons;

Figure 5.4: Cross-modal retrieval using CM. Here, CM is used to find the images that best match a query text.

where $k = 1, \ldots, d$. Analogously, a text $T \in \Re^T$ is mapped into $\mathcal{F}^T$ and then projected onto $\{w_{t,k}\}_{k=1}^d$, *i.e.*, $p_T = \mathcal{P}_T(\phi_T(T))$, using $\mathcal{K}_T\,(.\,,\,.)$.

Correlation matching

For all methods, a natural invertible mapping between the projections onto $\mathcal{U}^I$ and $\mathcal{U}^T$ follows from the correspondence between the d-dimensional bases of the subspaces, as $w_{i,1} \leftrightarrow w_{t,1}$, ..., $w_{i,d} \leftrightarrow w_{t,d}$. This results in a compact, efficient representation of both modalities, where vectors p_I and p_T are coordinates in two isomorphic d-dimensional subspaces, as shown in Figure 5.3. Given an image query I with projection p_I, the text $T \in \Re^T$ that most closely matches it is that for which p_T minimizes

$$D(I, T) = d(p_I, p_T), \tag{5.7}$$

for some suitable distance measure $d(\cdot, \cdot)$ in a d-dimensional vector space. Similarly, given a query text T with projection p_T, the closest image match $I \in \Re^I$ is that for which p_I minimizes $d(p_I, p_T)$. An illustration of cross-modal retrieval using CM is given in Figure 5.4.

5.4.2 Semantic matching (SM)

An alternative to subspace learning is to map images and text to representations at a higher level of abstraction, where a natural correspondence can be established. This is obtained by augmenting the database $\mathcal{B}$ with a vocabulary

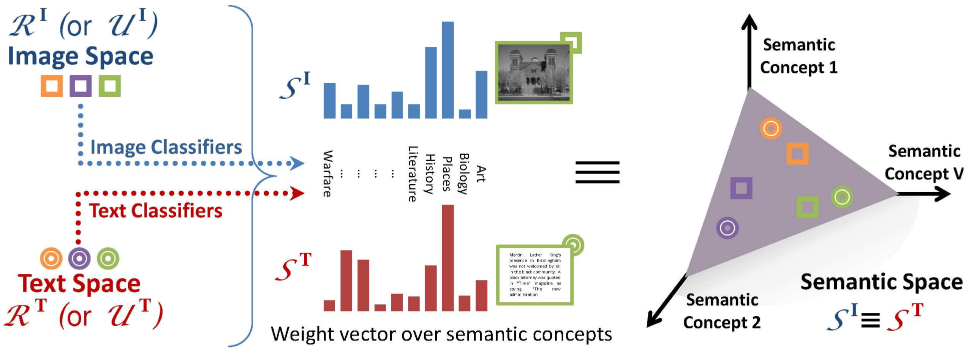

$\mathcal{R}^{\mathrm{I}}$ (or $\mathcal{U}^{\mathrm{I}}$)
Image Space
Image Classifiers
Text Classifiers
Text Space
$\mathcal{R}^{\mathrm{T}}$ (or $\mathcal{U}^{\mathrm{T}}$)
$\mathcal{S}^{\mathrm{I}}$
Art
Biology
Places
History
Literature
Warfare
$\mathcal{S}^{\mathrm{T}}$
Martin Luther King's presence in Birmingham was not welcomed by all in the black community. A black attorney was quoted in "Time" magazine as saying, "The new administration
Weight vector over semantic concepts
Semantic Concept 1
Semantic Concept V
Semantic Concept 2
Semantic Space
$\mathcal{S}^{\mathrm{I}} \equiv \mathcal{S}^{\mathrm{T}}$

Figure 5.5: Semantic matching (SM) maps text and images into a semantic space. For each modality, classifiers are used to obtain a semantic representation, *i.e.*, a weight vector over semantic concepts.

$\mathcal{L} = \{1, \ldots, L\}$ of semantic concepts, such as "History" or "Biology". Individual documents are grouped into these classes. Two mappings $\mathbf{\Pi}_T$ and $\mathbf{\Pi}_I$ are then implemented using classifiers of text and images, respectively. $\mathbf{\Pi}_T$ maps a text $T \in \Re^T$ into a vector π_T of posterior probabilities $P_{W|T}(w|T), w \in \{1, \ldots, L\}$ with respect to each of the classes in $\mathcal{L}$. The space $\mathcal{S}^T$ of these vectors is referred to as the *semantic space for text*, and the probabilities $P_{W|T}(w|T)$ as *semantic text features*. Similarly, $\mathbf{\Pi}_I$ maps an image I into a vector π_I of *semantic image features* $P_{W|I}(w|I), w \in \{1, \ldots, L\}$ in a *semantic space for images* $\mathcal{S}^I$.

Semantic representations have two advantages for cross-modal retrieval. First, they provide a higher level of abstraction. While standard features in $\Re^T$ and $\Re^I$ are the result of unsupervised learning, and frequently have no obvious interpretation (*e.g.*, image features tend to be edges, edge orientations or frequency bases), the features in $\mathcal{S}^T$ and $\mathcal{S}^I$ are semantic concept probabilities (*e.g.*, the probability that the image belongs to the "History" or "Biology" document classes). In Chapter 3, it was shown that this increased semantic abstraction can lead to substantially better generalization for tasks such as image retrieval. Second, the semantic spaces $\mathcal{S}^T$ and $\mathcal{S}^I$ are isomorphic, since both images and text are represented as vectors of posterior probabilities with respect to the *same* document classes. Hence, the spaces can be treated as being the same, *i.e.*, $\mathcal{S}^T = \mathcal{S}^I$, leading to the schematic representation in Figure 5.5.

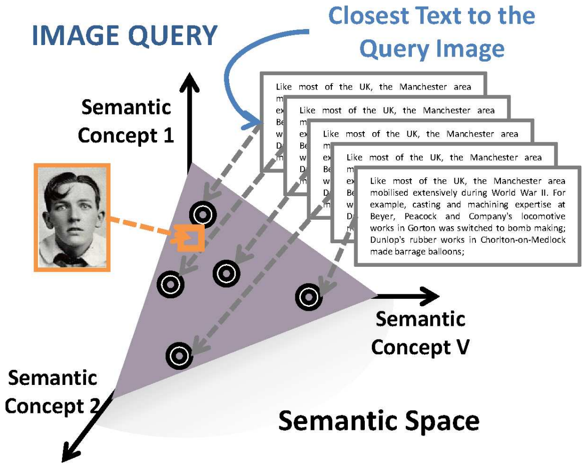


IMAGE QUERY
Closest Text to the Query Image
Semantic Concept 1
Like most of the UK, the Manchester area mobilised extensively during World War II. For example, casting and machining expertise at Beyer, Peacock and Company's locomotive works in Gorton was switched to bomb making; Dunlop's rubber works in Chorlton-on-Medlock made barrage balloons;
Semantic Concept V
Semantic Concept 2
Semantic Space

Figure 5.6: Cross-modal retrieval using SM used to find the text that best matches a query image.

In Chapter 2, it was highlighted that it is not necessary to model each class explicitly and any system that computes posterior probabilities can be employed to obtain the semantic representation. For the evaluation of cross-modal retrieval systems, the posterior probability distributions are computed through multi-class logistic regression which produces linear classifiers with a probabilistic interpretation. Logistic regression based classification is chosen due to its simplicity. Under this, the posterior probability of class w is computed, by fitting the image (text) features to a logistic function,

$$P_{W|X}(w|x;\boldsymbol{\beta}) = \frac{1}{Z(x,\boldsymbol{\beta})}\exp\left(\beta_w^T x\right), \tag{5.8}$$

where $Z(x,\boldsymbol{\beta}) = \sum_w \exp\left(\beta_w^T x\right)$ is a normalization constant, W the class label, X the feature vector in the input space, and $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_L\}$ with β_w a vector of parameters for class w. A multi-class logistic regression is learned for the image and text modality, by making X the image and text representation, $I \in \Re^I$ and $T \in \Re^T$, respectively. In our implementation we use the software package Liblinear [38]. Given a query image I (text T), represented by $\pi_I \in \mathcal{S}^I$ ($\pi_T \in \mathcal{S}^T$), cross-modal retrieval will find the text T (image I), represented by $\pi_T \in \mathcal{S}^T$ ($\pi_I \in \mathcal{S}^I$), that minimizes

$$D(I,T) = d(\pi_I, \pi_T), \tag{5.9}$$

for some suitable distance measure d between probability distributions. An illustration of cross-modal retrieval using SM is given in Figure 5.6.

5.4.3 Semantic Correlation Matching (SCM)

CM and SM are not mutually exclusive. In fact, a corollary to the two hypotheses discussed above is that there may be a benefit in combining CM and SM. CM extracts maximally correlated features from $\Re^T$ and $\Re^I$. SM builds semantic spaces using original features to gain semantic abstraction. When the two are combined, by building semantic spaces using the feature representation produced by correlation maximization, it may be possible to improve on the individual performances of both CM and SM. To combine the two approaches, the maximally correlated subspaces $\mathcal{U}^I$ and $\mathcal{U}^T$ are first learned with correlation modeling. Logistic regressors $\mathbf{\Pi}_I$ and $\mathbf{\Pi}_T$ are then learned in each of these subspaces to produce the semantic spaces $\mathcal{S}^I$ and $\mathcal{S}^T$, respectively. Retrieval is finally based on the image-text distance $D(I, T)$ of (5.9), based on the semantic mappings $\pi_I = \mathbf{\Pi}_I(p_I)$ and $\pi_T = \mathbf{\Pi}_T(p_T)$ after projecting them onto $\mathcal{U}^I$ and $\mathcal{U}^T$, respectively.

5.5 Experimental Setup

In this section, we describe an extensive experimental evaluation of the proposed framework. Two tasks were considered: text retrieval from an image query, and image retrieval from a text query. The cross-modal retrieval performance is measured with *precision-recall* (PR) curves and *mean average precision* (MAP) scores. The standard 11-point interpolated PR curves [91] are used. The MAP score is the average precision at the ranks where recall changes. Both metrics are evaluated at the level of *in-* or *out-of-category*, which is a popular choice in the information retrieval literature [119].

Dataset

For the evaluation of the cross-modal retrieval system we use two different datasets, viz. TVGraz and Wikipedia. The TVGraz dataset is a collection of web-pages compiled by Khan *et al.* [66] and contains $2,058$ image-text pairs divided into 10 categories (see Appendix A.1.5 for more details). Wikipedia is novel dataset assembled from the "Wikipedia featured articles", a continually updated collection of Wikipedia articles, and contains a total of $2,866$ image-text pairs again divided into 10 categories (see Appendix A.1.6 for more details)

The two datasets have important differences. TVGraz images are archetypal members of the categories, due to the collection procedure [66]. The dataset is eminently visual, since its categories (*e.g.*, "Harp", "Dolphin") are specific objects or animals, and the classes are semantically well-separated, with little or no semantic overlap. For example, the syllabus of a Neuroscience class can be attached to a picture of a brain. However, the texts are small and can be less representative of the categories to which they are associated. In Wikipedia, on the other hand, the category membership is assessed based on text content. Hence, texts are mostly of good quality and representative of the category, while the image categorization is more ambiguous. For example, a portrait of a historical figure can appear in the class "War". The Wikipedia categories (*e.g.*, "History", "Biology") are more abstract concepts, and have much broader scope. Frequently, documents could be classified into one or more categories. Individually, the images can be difficult to classify, even for a human. Together, the two datasets represent an important subset of the diversity of practical cross-modal retrieval scenarios: applications where there is more uniformity of text than images, and vice-versa.

5.5.1 Image and text representation

For both modalities, the base representation is a bag-of-words (BOW) representation. Text words were obtained by stemming the text with the Python Natural Language Toolkit[3]. Direct word histograms were not suitable for text because the large lexicon made the correlation analysis intractable. Instead, a latent Dirichlet allocation (LDA) [14] model was learned from the text features, using the implementation of [32]. LDA summarizes a text as a mixture of topics. More precisely, a text is modeled as a multinomial distribution over topics, each of which

[3] http://www.nltk.org/

Table 5.2: Cross-modal retrieval performance (MAP) on the validation set using different distance metrics for TVGraz. μ_p and μ_q are the sample averages for p and q, respectively.

			TVGraz		
Experiment	measure	$d(p,q)$	img query	txt query	avg
CM	ℓ_1	$\sum_i \|p_i - q_i\|$	0.376	0.418	0.397
	ℓ_2	$\sum_i (p_i - q_i)^2$	0.391	0.444	0.417
	NC	$\frac{p^T q}{\|\|p\|\| \|\|q\|\|}$	0.498	0.476	**0.487**
	NC_c	$\frac{(p-\mu_p)^T (q-\mu_q)}{\|\|p-\mu_p\|\| \|\|q-\mu_q\|\|}$	0.486	0.462	0.474
SM	KL	$\sum_i p_i \log \frac{p_i}{q_i}$	0.296	0.546	0.421
	ℓ_1	$\sum_i \|p_i - q_i\|$	0.412	0.548	0.480
	ℓ_2	$\sum_i (p_i - q_i)^2$	0.380	0.550	0.465
	NC	$\frac{p^T q}{\|\|p\|\| \|\|q\|\|}$	0.533	0.560	0.546
	NC_c	$\frac{(p-\mu_p)^T (q-\mu_q)}{\|\|p-\mu_p\|\| \|\|q-\mu_q\|\|}$	0.579	0.556	**0.568**
SCM	KL	$\sum_i p_i \log \frac{p_i}{q_i}$	0.576	0.636	0.606
	ℓ_1	$\sum_i \|p_i - q_i\|$	0.637	0.645	0.641
	ℓ_2	$\sum_i (p_i - q_i)^2$	0.614	0.63	0.622
	NC	$\frac{p^T q}{\|\|p\|\| \|\|q\|\|}$	0.669	0.646	0.658
	NC_c	$\frac{(p-\mu_p)^T (q-\mu_q)}{\|\|p-\mu_p\|\| \|\|q-\mu_q\|\|}$	0.678	0.641	**0.660**

Table 5.3: Cross-modal retrieval performance (MAP) on the validation set using different distance metrics for Wikipedia. μ_p and μ_q are the sample averages for p and q, respectively.

			Wikipedia		
Experiment	measure	$d(p,q)$	img query	txt query	avg
CM	ℓ_1	$\sum_i \lvert p_i - q_i \rvert$	0.193	0.234	0.214
	ℓ_2	$\sum_i (p_i - q_i)^2$	0.199	0.243	0.221
	NC	$\frac{p^T q}{\lVert p \rVert \lVert q \rVert}$	0.288	0.239	**0.263**
	NC_c	$\frac{(p-\mu_p)^T (q-\mu_q)}{\lVert p-\mu_p \rVert \lVert q-\mu_q \rVert}$	0.287	0.239	**0.263**
SM	KL	$\sum_i p_i \log \frac{p_i}{q_i}$	0.188	0.276	0.232
	ℓ_1	$\sum_i \lvert p_i - q_i \rvert$	0.232	0.276	0.254
	ℓ_2	$\sum_i (p_i - q_i)^2$	0.211	0.278	0.245
	NC	$\frac{p^T q}{\lVert p \rVert \lVert q \rVert}$	0.315	0.278	0.296
	NC_c	$\frac{(p-\mu_p)^T (q-\mu_q)}{\lVert p-\mu_p \rVert \lVert q-\mu_q \rVert}$	0.354	0.272	**0.313**
SCM	KL	$\sum_i p_i \log \frac{p_i}{q_i}$	0.287	0.282	0.285
	ℓ_1	$\sum_i \lvert p_i - q_i \rvert$	0.329	0.286	0.308
	ℓ_2	$\sum_i (p_i - q_i)^2$	0.307	0.286	0.296
	NC	$\frac{p^T q}{\lVert p \rVert \lVert q \rVert}$	0.375	0.288	0.330
	NC_c	$\frac{(p-\mu_p)^T (q-\mu_q)}{\lVert p-\mu_p \rVert \lVert q-\mu_q \rVert}$	0.388	0.285	**0.337**

is in turn modeled as a multinomial distribution over words. Each word in a text is generated by first sampling a topic from the text-specific topic distribution, and then sampling a word from that topic's multinomial. This serves two purposes: it reduces dimensionality and increases feature abstraction, by representing text as a distribution over topics instead of a distribution over words. In text modeling the number of topics in LDA ranged from 5 to 800.

Image words were learned with the scale invariant feature transformation (SIFT-GRID) [85] computed on a grid of image patches. A bag of SIFT descriptors was first extracted from each image in the training set, using the SIFT implementation of LEAR[4]. A codebook, or dictionary of visual words was then learned with the K-means clustering algorithm. The SIFT descriptors extracted from each image were vector quantized with this codebook, producing a vector of visual word counts per image. Besides this BOW representation, we also use a lower-dimensional representation for images, similar to that for text, by fitting an LDA model to visual word histograms and representing images as a distribution over topics. Preliminary experiments indicated that this outperformed an image representation of reduced dimensionality through principal component analysis (PCA). In image modeling for LDA representation the number of topics ranged from 5 to $4,000$, for BOW the number of visual words ranged from 128 to $8,192$.

5.6 Parameter selection

The combination of three retrieval modes (CM, SM, and SCM), three correlation matching approaches (CFA, CCA, KCCA), two image representations (BOW, LDA), and various distance measures d generates a large number of possibilities for the implementation of cross-modal retrieval. Since each configuration has a number of parameters to tune, it is difficult to perform an exhaustive comparison of all possibilities. Instead, we pursued a sequence of preliminary comparisons to prune the configuration space, using a random 80/20 split of the training set, for training and validation respectively (splitting TVGraz' training set into $1,245$

[4] https://lear.inrialpes.fr/people/dorko/downloads.html

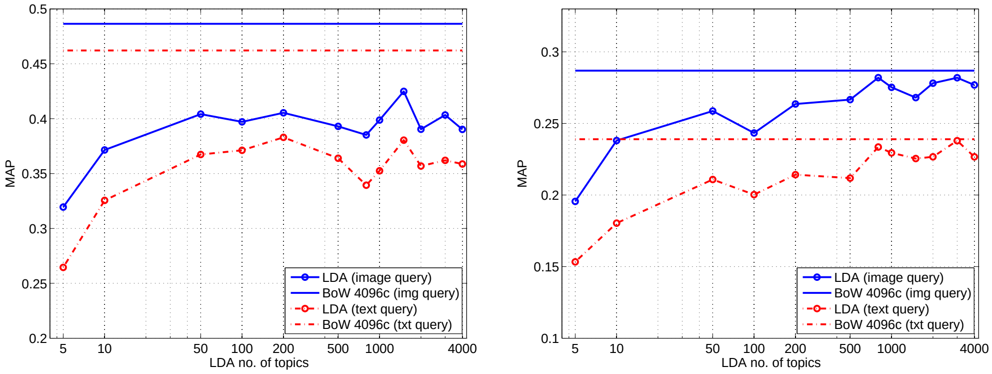
training and 313 validation examples, and Wikipedia's into $1,738$ training and 435 validation documents). This suggested a cross-modal retrieval architecture that combines i) the centered normalized correlation (for distances d), ii) a BOW (rather than LDA) representation for images, and iii) KCCA to learn correlation subspaces. Supporting experiments are presented below. For each retrieval mode – CM, SM, SCM for image queries or text queries – and each dataset – TVGraz, Wikipedia –, the codebook size (for image representation), the number of topics (for text representation) and/or the number of KCCA components were determined, where applicable, by performing a grid search and adopting the settings with maximum retrieval performance on the validation set, unless indicated otherwise. In the following section, the top performing approaches are compared on the test set.

Distance Measures

We started by comparing a number of distance measures d, for the evaluation of (5.7) and (5.9), in CM, SM, and SCM retrieval experiments (using KCCA to produce the subspaces for CM and SCM, and BOW to represent images). The distance measures are listed in 5.2 and 5.3 for TVGraz and Wikipedia respectively, and include the Kullback-Leibler divergence (KL), ℓ_1 and ℓ_2 norms, normalized correlation (NC), and centered normalized correlation (NC_c). The KL divergence was not used with CM because this technique does not produce a probability simplex. 5.2 and 5.3 present the MAP scores achieved with each measure, on the validation set. NC_c achieved the best average performance in all experiments other than CM-based retrieval on TVGraz, where it was outperformed by NC. Since the difference was small even in this case, NC_c was adopted as distance measure in all remaining experiments.

Text and image representation

Due to the intractability of word counts, we considered only the LDA representation for text. In the image domain, we compared the performance of the BOW and LDA representations, using an SCM system based on KCCA subspaces



(a) TVGraz

(b) Wikipedia

Figure 5.7: MAP performance (cross-modal retrieval, validation set) of SCM using two image models: BOW (flat lines) and LDA, for (a) TVGraz and (b) Wikipedia.

and $4,096$ codewords for BOW (an optimal setting, as evidenced in Section 5.6). Figure 5.7 presents the results for both text and image queries. Since the retrieval performance of LDA was inferior to that of BOW, for all topic cardinalities, BOW was adopted as the image representation for all remaining experiments.

Correlation matching

The next set of experiments was designed to compare the different CM methods. These methods have different degrees of freedom and thus require different amounts of parameter tuning. The most flexible representation is KCCA, whose performance varies with the choice of kernel and regularization parameter κ of (5.5). We started by comparing various combinations of text and image kernels. Best results were achieved for a *chi-square radial basis function* kernel[5] for images combined with a *histogram intersection* kernel [141, 18] for text. Combinations involving other kernels (*e.g.*, linear, Gaussian, exponential) achieved inferior validation set performance. Regarding regularization, best results were obtained with $\kappa = 10\%$ on TVGraz and $\kappa = 50\%$ on Wikipedia. The need for a stronger regular-

[5] $\mathcal{K}(x,\, y) = \exp\left(\frac{d_{\chi^2}(x,\, y)}{\gamma}\right)$ where $d_{\chi^2}(x,\, y)$ is the chi-square distance between x and y and γ the average chi-square distance among training points.

Table 5.4: MAP for CM hypothesis (validation sets).

Experiment	Image Query	Text Query	Average	Average Gain	Dataset
KCCA	**0.486**	**0.462**	**0.474**	-	TVGraz
CCA	0.284	0.254	0.269	76%	
CFA	0.195	0.179	0.187	153%	
KCCA	**0.287**	**0.239**	**0.263**	-	Wiki.
CCA	0.210	0.174	0.192	37%	
CFA	0.195	0.156	0.176	50%	

izer in Wikipedia suggests that there are more spurious correlations on this dataset, which could lead to over-fitting. This is sensible, given the greater diversity and abstraction of the concepts in this dataset.

For CCA (CFA), the only free parameter is the number of canonical components (dimensionality of the shared space) used for both image and text representation. This parameter also remains to be tuned for KCCA. For each experiment and data set, a grid search was performed and the parameter of best retrieval performance was adapted under each method (CFA, CCA, KCCA). 5.4 presents best CM performances achieved with each method. In all cases, KCCA yields top performance. On TVGraz, the average gain (for text and image queries) is 153% over CFA and 76% over CCA. On Wikipedia, the gain over CFA is 50% and over CCA 37%. KCCA was chosen to implement the correlation hypothesis in the remaining experiments.

Parameter Tuning

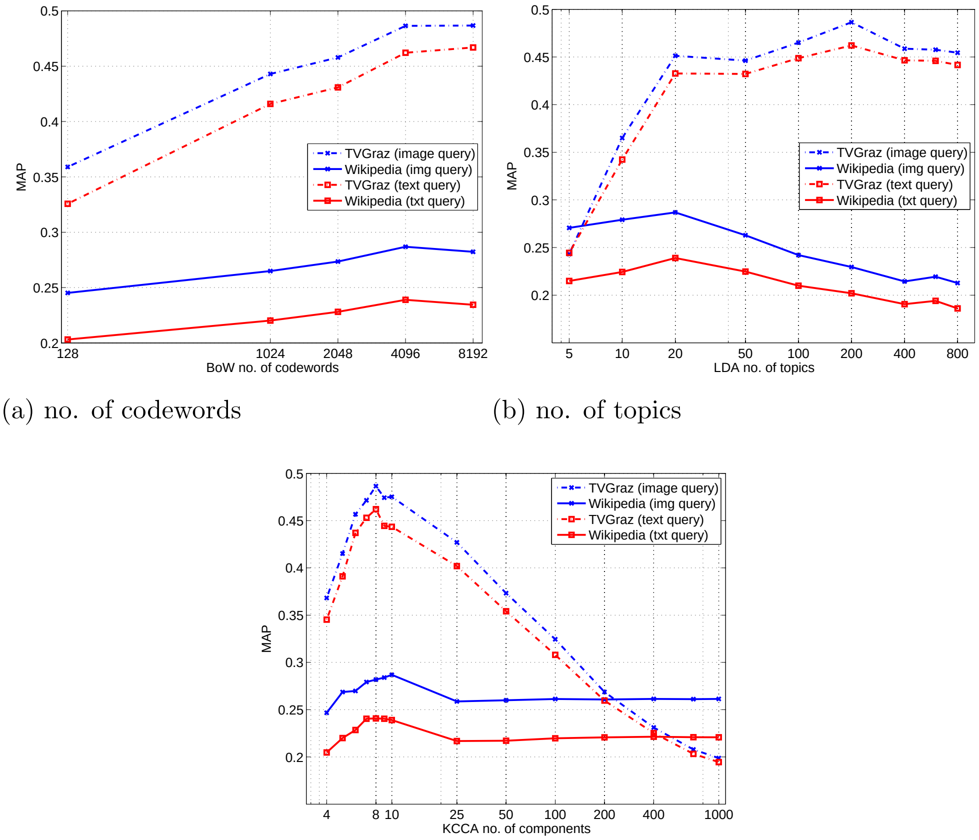
For a cross-modal retrieval architecture combining the best of the above, i.e., KCCA (to learn correlation subspaces), NC_c (as distance measure), and the BOW representation for images, we take a closer look at the codebook size for image (BOW) representation, the number of topics for text (LDA) representation and the number of KCCA components. Figure 5.5 summarizes the optimal parameter

Table 5.5: Best parameter settings for CM, SM and SCM, on both TVGraz and Wikipedia (validation sets).

	CM	SM	SCM	
MAP image / text query	0.49 / 0.46	0.59 / 0.56	0.68 / 0.64	TVGraz
BOW codewords	4096			
LDA topics	200	100	400	
KCCA components	8	-	1125	
MAP image / text query	0.29 / 0.24	0.35 / 0.27	0.39 / 0.29	Wikipedia
BOW codewords	4096			
LDA topics	20	600	200	
KCCA components	10	-	38	

settings (after performing a grid search with cross-validation) and corresponding retrieval performance on the validation set, for CM, SM and SCM experiments. 5.8 provides more detail on how varying each parameter individually affects the performance, for CM. Note that the best MAP scores are obtained with a small number of KCCA components (< 10). For the image representation, best performance was achieved with codebooks of $4,096$ visual words, on both datasets. For text, 200 topics performed best on TVGraz and 20 on Wikipedia. Note that in the test set experiments of Section 5.7, the number of KCCA components of 5.5 is scaled by the ratio of the number of training points of the test experiments and that of the validation experiments (see A.4 and A.5 in Appendix A), so that a comparable fraction of correlation is preserved after dimensionality reduction[6].

[6] KCCA seeks directions of maximum correlation in Span$\{\phi_I(I_1), \ldots, \phi_I(I_{|\mathcal{B}|})\}$ and Span$\{\phi_T(T_1), \ldots, \phi_T(T_{|\mathcal{B}|})\}$, where $|\mathcal{B}|$ is the training set size. This is larger for test than for validation experiments ($2,173$ v.s. $1,738$ on Wikipedia and $1,558$ v.s. $1,245$ on TVGraz). Hence, on average, a KCCA component will explain less correlation in the test than in the validation experiments. It follows that a larger number of KCCA components are needed to capture the same fraction of the total correlation.



(a) no. of codewords

(b) no. of topics

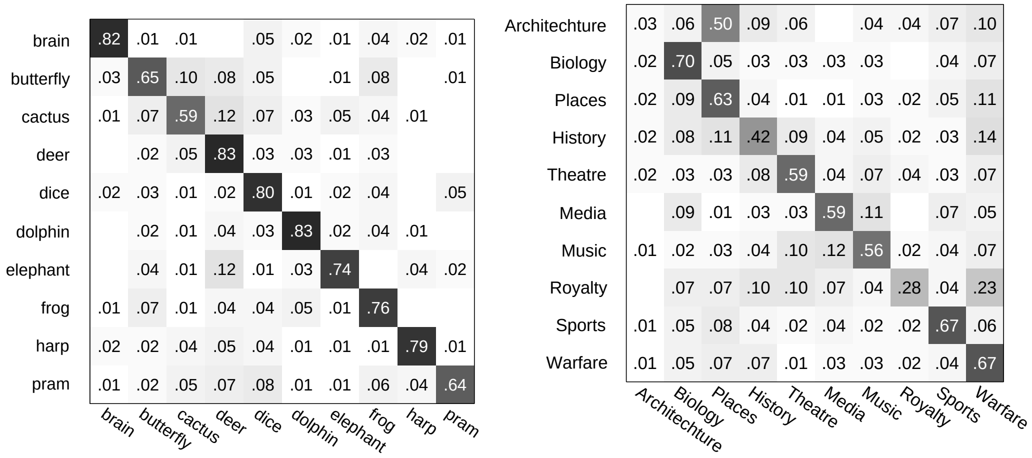
(c) no. of KCCA components

Figure 5.8: Cross-modal MAP for CM on TVGraz and Wikipedia (validation sets), as a function of (a) the number of image codewords, (b) the number of text LDA topics, and (c) the number of KCCA components (while keeping the other two parameters fixed at the values reported in 5.5).

5.7 Testing the fundamental hypotheses

In this section, we compare the performance of CM, SM, and SCM on the test set. In all cases, the parameter configurations are those that achieved best cross-validation performance in the previous section. 5.6 compares the MAP scores of cross-modal retrieval — text-to-image, image-to-text, and their average — using CM, SM and SCM, to chance-level performance[7]. Two distinct observations can be made from this table with regards to TVGraz. First, it provides evidence in

[7] Random images (text) returned in response to a text (image) query.



	brain	butterfly	cactus	deer	dice	dolphin	elephant	frog	harp	pram
brain	.82	.01	.01		.05	.02	.01	.04	.02	.01
butterfly	.03	.65	.10	.08	.05		.01	.08		.01
cactus	.01	.07	.59	.12	.07	.03	.05	.04	.01	
deer		.02	.05	.83	.03	.03	.01	.03		
dice	.02	.03	.01	.02	.80	.01	.02	.04		.05
dolphin		.02	.01	.04	.03	.83	.02	.04	.01	
elephant		.04	.01	.12	.01	.03	.74		.04	.02
frog	.01	.07	.01	.04	.04	.05	.01	.76		
harp	.02	.02	.04	.05	.04	.01	.01	.01	.79	.01
pram	.01	.02	.05	.07	.08	.01	.01	.06	.04	.64

	Architechture	Biology	Places	History	Theatre	Media	Music	Royalty	Sports	Warfare
Architechture	.03	.06	.50	.09	.06		.04	.04	.07	.10
Biology	.02	.70	.05	.03	.03	.03	.03		.04	.07
Places	.02	.09	.63	.04	.01	.01	.03	.02	.05	.11
History	.02	.08	.11	.42	.09	.04	.05	.02	.03	.14
Theatre	.02	.03	.03	.08	.59	.04	.07	.04	.03	.07
Media		.09	.01	.03	.03	.59	.11		.07	.05
Music	.01	.02	.03	.04	.10	.12	.56	.02	.04	.07
Royalty		.07	.07	.10	.10	.07	.04	.28	.04	.23
Sports	.01	.05	.08	.04	.02	.04	.02	.02	.67	.06
Warfare	.01	.05	.07	.07	.01	.03	.03	.02	.04	.67

Figure 5.9: Confusion matrices on the test set, for both TVGraz (left) and Wikipedia (right). Rows refer to true categories, and columns to category predictions. The more confusion on Wikipedia motivates the lower retrieval performance.

support of the two hypotheses of Section 5.3.3. Both joint dimensionality reduction (CM) and semantic abstraction (SM) are beneficial for multi-modal modeling, leading to a non-trivial improvement over chance-level performance. For example, in TVGraz, CM achieves an average MAP score of 0.497, over four times the random retrieval performance of 0.114. SM yields an even greater improvement, attaining a MAP score of 0.622. Second, combining correlation modeling with semantic abstraction (SCM) is desirable, leading to higher MAP scores. On TVGraz, SCM improves about 12% over SM and 40% over CM, achieving an average MAP score of 0.694. This suggests that the contributions of cross-modal correlation and semantic abstraction are *complementary*: not only is there an independent benefit to both correlation modeling and abstraction, but the *best performance is achieved when the approaches underlying the two hypotheses are combined*. The gains hold for both cross-modal retrieval tasks, *i.e.*, image and text queries.

Similar conclusions can be drawn for Wikipedia. However, the improvement of SCM over SM is less substantial than in TVGraz. In fact, the retrieval performances on Wikipedia are generally lower than those on TVGraz. As discussed in Section 5.5, this is likely due to the broader scope of the Wikipedia categories. In

Table 5.6: Cross-modal MAP on TVGraz and Wikipedia (test sets).

Experiment	Image Query	Text Query	Average	Average Gain	
SCM	**0.693**	**0.696**	**0.694**	-	TVGraz
SM	0.625	0.618	0.622	11.6%	
CM	0.507	0.486	0.497	39.6%	
Random	0.114	0.114	0.114	509%	
SCM	**0.372**	**0.268**	**0.320**	-	Wiki.
SM	0.362	0.252	0.307	4.2%	
CM	0.282	0.225	0.253	26.5%	
Random	0.119	0.119	0.119	170%	

this dataset, a significant fraction of documents could be classified into multiple categories, making the data harder to model. This explanation is supported by the confusion matrices of Figure 5.9. These were built by assigning each text and image query to the class of highest MAP in the ranking produced by SCM[8]. Note, for example, the significant confusion between the categories "Architecture" and "Places", or "Royalty" and "Warfare". Figure 5.10 and 5.11 presents PR curves and precision at N curves, of cross-modal retrieval with CM, SM and SCM for TVGraz and Wikipedia respectively. All methods yield non-trivial precision improvements, at all levels of recall, when compared to the random baseline. On TVGraz, SM has higher precision than CM, and SCM has higher precision than SM, at all levels of recall. On Wikipedia, SCM improves over CM, at all levels of recall, but the improvement over SM is small. Figure 5.12 shows the MAP scores achieved per category by all approaches. SCM has a significantly higher MAP than CM and SM on all classes of TVGraz, and is either comparable or better than CM and SM on the majority of classes of Wikipedia.

Few examples of text queries and corresponding retrieval results, using the SCM methodology, are shown in Figure 5.13, 5.14, Figure 5.15, and 5.16. The text

[8] Note that this is not ideal for classification, since the MAP is computed over a ranking of the test set.

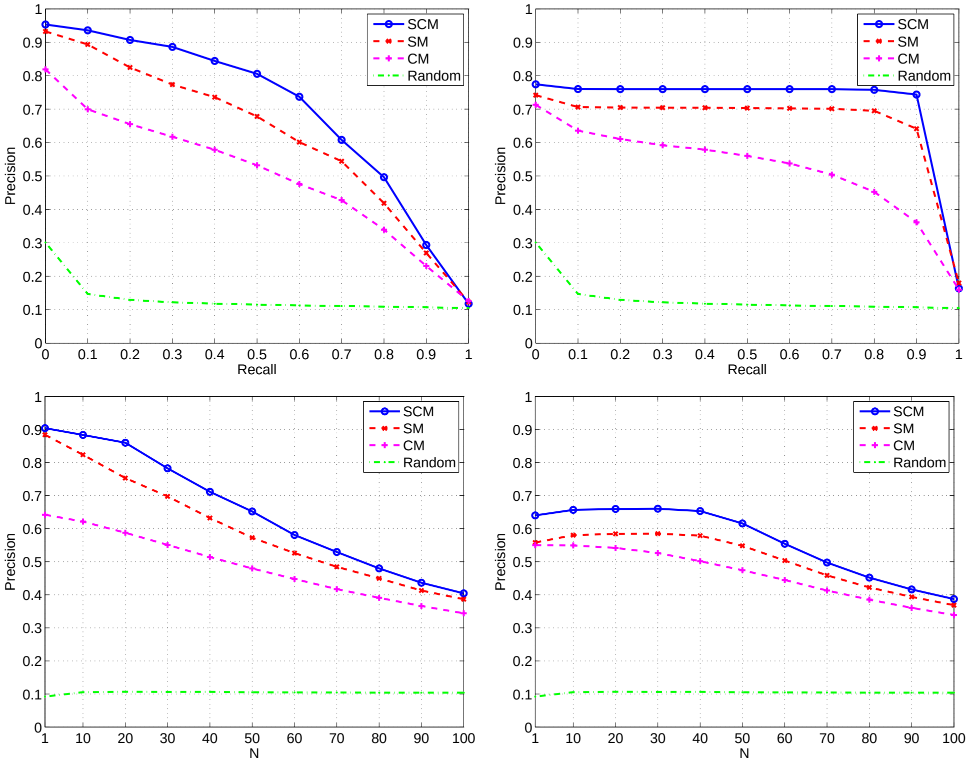


Figure 5.10: top) Precision recall curves, bottom) Precision at N curves for left) Text query, right) Image query for TVGraz

query is presented along with its probability vector π_T and the ground truth image. The top five image matches are shown below the text, along with their probability vectors π_I. Note that SCM assigns these images the highest ranks in the retrieved list because their semantic vectors (π_I) most closely match that of the text (π_T). For the TVGraz example (Figure 5.16) this can be verified by noting the common concentration of probability mass around the “Butterfly” bin. In the Wikipedia example (Figure 5.14) the probability is concentrated around the “Warfare” bin. Finally, Figure 5.17 shows some examples of image-to-text retrieval. The query images are shown on the top row, and the images associated with the four best text matches are shown on the bottom.

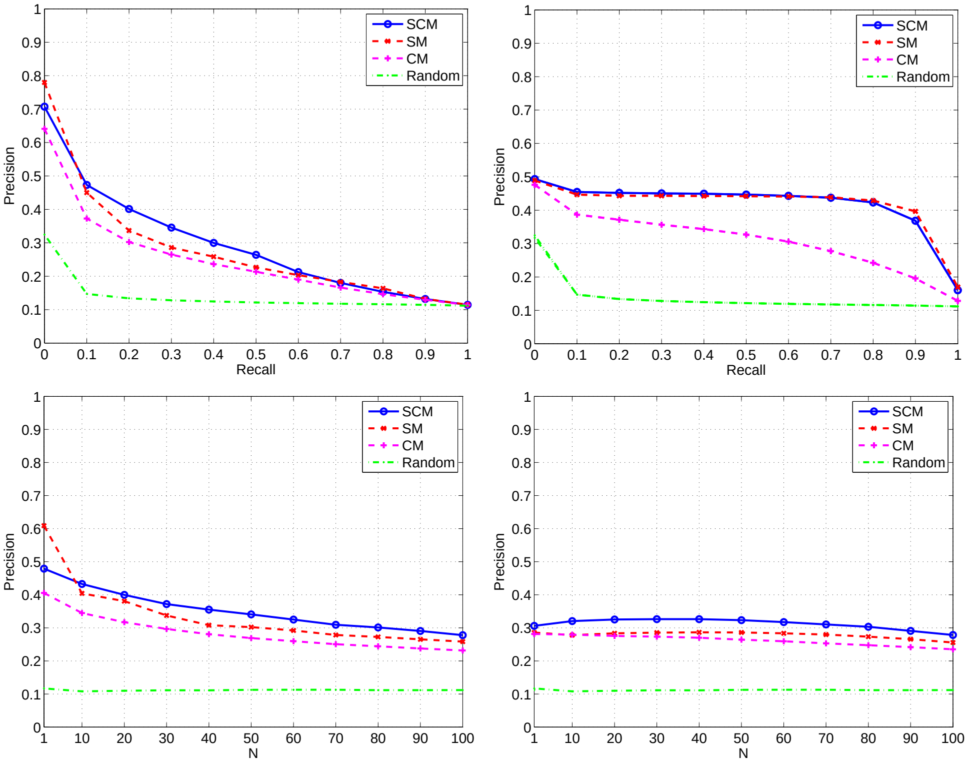
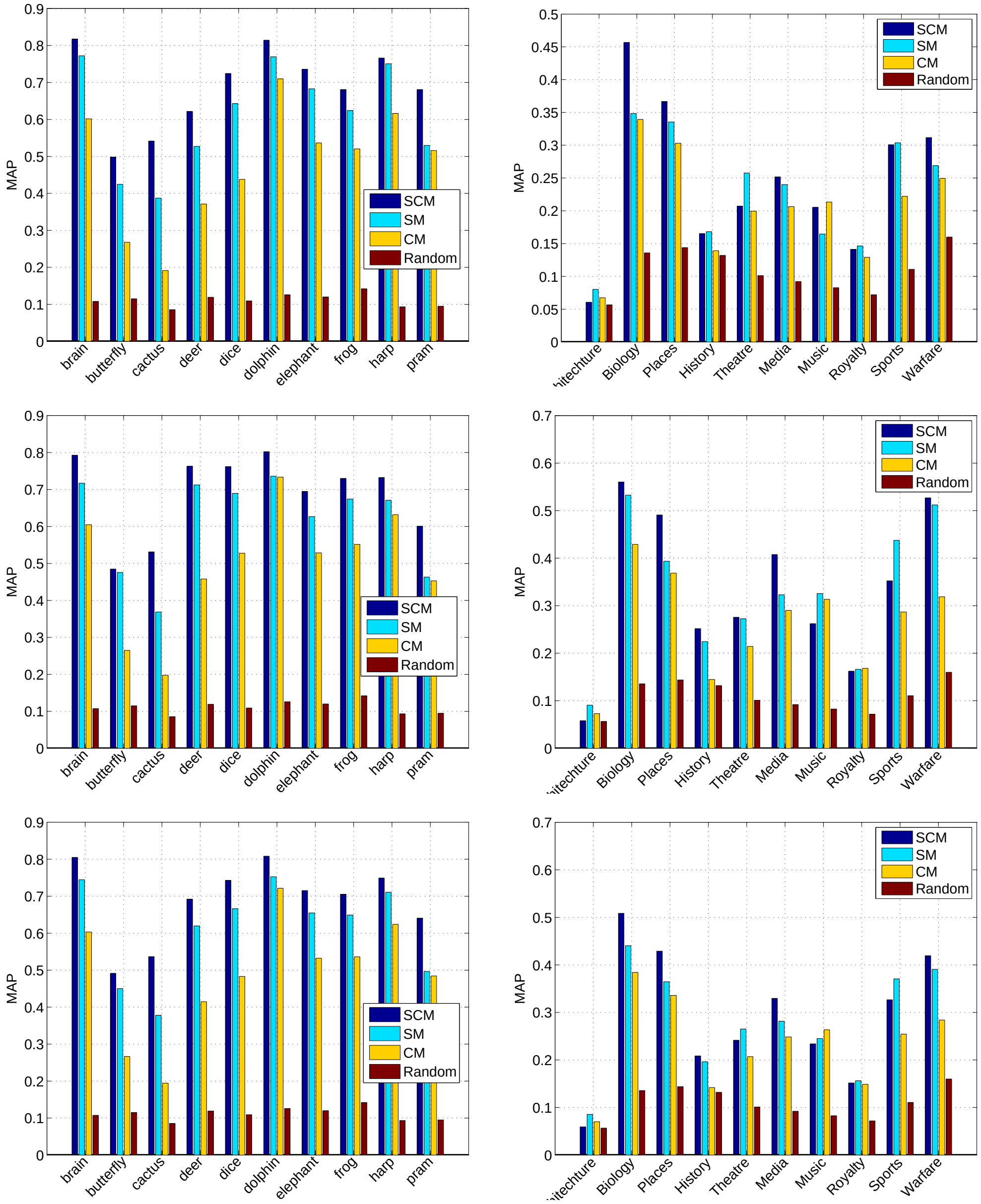


Figure 5.11: top) Precision recall curves, bottom) Precision at N curves for left) Text query, right) Image query for Wikipedia

5.8 Acknowledgments

The author would like to thank Jose Costa Pereira, Emanuele Coviello, Gabe Doyle, Gert Lanckriet and Roger Levy, for their help and contribution in developing the cross-model multimedia system.

The text of Chapter 5, in part, is based on the material as it appears in: N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G.R.G. Lanckriet, R. Levy, N. Vasconcelos “A New Approach to Cross-Modal Multimedia Retrieval”, ACM Proceedings of the 15th international conference on Multimedia, Florence, Italy, Oct 2010. The dissertation author was a primary researcher and an author of the cited material.

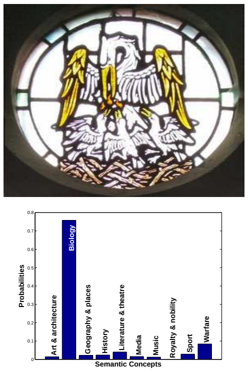


TVGraz

Wikipedia

Figure 5.12: Per-class MAP for the cross-modal retrieval tasks on TVGraz (left) and Wikipedia (right): text queries (top); image queries (middle); and average performance over both types of queries (bottom).

Many seabirds are little studied and poorly known, due to living far out to sea and breeding in isolated colonies. However, some seabirds, particularly, the albatrosses and gulls, have broken into popular consciousness. The albatrosses have been described as ”the most legendary of birds”, Carboneras, C. (1992) ”Family Diomedeidae (Albatrosses)” in ”Handbook of Birds of the World” Vol 1. Barcelona:Lynx Edicions, ISBN 84-87334-10-5 and have a variety of myths and legends associated with them, and today it is widely considered unlucky to harm them, although the notion that sailors believed that is a mythCocker, M., & Mabey, R., (2005) ”Birds Britannica” London:Chatto & Windus, ISBN 0-7011-6907-9 which derives from Samuel Taylor Coleridge's famous poem, ”The Rime of the Ancient Mariner”, in which a sailor is punished for killing an albatross by having to wear its corpse around his neck. ”Instead of the Cross the Albatross” ”About my neck was hung” Sailors did, however, consider it unlucky to touch a storm-petrel, especially one that has landed on the ship. Carboneras, C. (1992) ”Family Hydrobatidae (Storm-petrels)” in ”Handbook of Birds of the World” Vol 1. Barcelona:Lynx Edicions, ISBN 84-87334-10-5 Gulls are one of the most commonly seen seabirds, given their use of human-made habitats (such as cities an d dumps) and their often fearless nature. They therefore also have made it into the popular consciousness - they have been used metaphorically, as in ”Jonathan Livingston Seagull” by Richard Bach, or to denote a closeness to the sea, such as their use in the ”The Lord of the Rings” both in the insignia of Gondor and therefore Númenor (used in the design of the films), and to call Legolas to (and across) the sea.



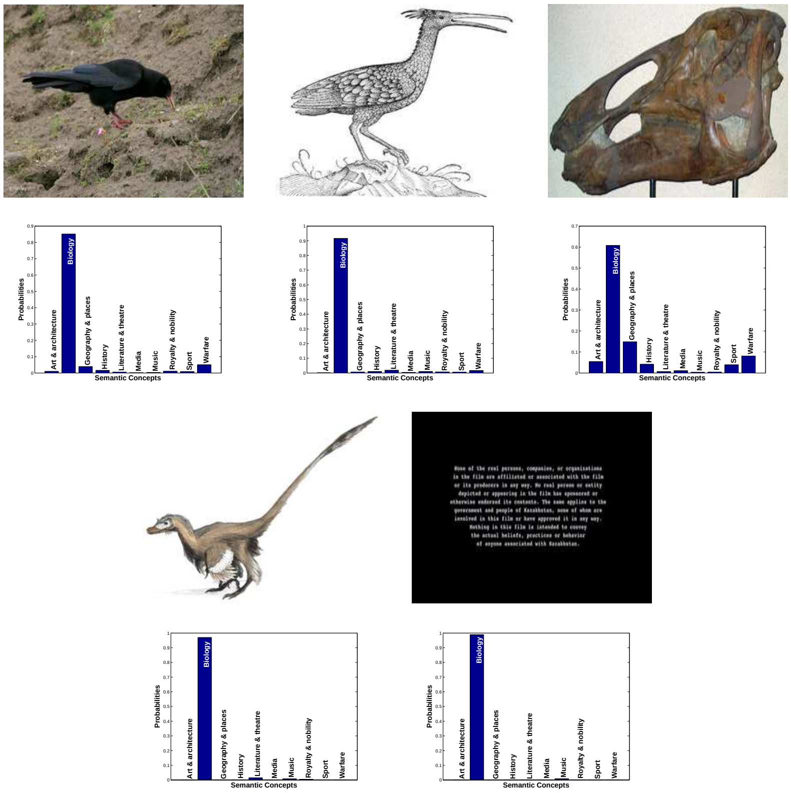
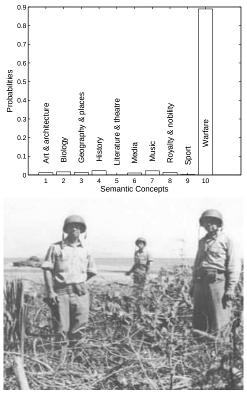


Figure 5.13: Text query from Biology class of Wikipedia and the top 5 retrieved images retrieved using SCM. The query text, associated probability vector, and ground truth image are shown on the top; retrieved images are presented at the bottom.

Between October 1 and October 17, the Japanese delivered 15,000 troops to Guadalcanal, giving Hyakutake 20,000 total troops to employ for his planned offensive. Because of the loss of their positions on the east side of the Matanikau, the Japanese decided that an attack on the U.S. defenses along the coast would be prohibitively difficult. Therefore, Hyakutake decided that the main thrust of his planned attack would be from south of Henderson Field. His 2nd Division (augmented by troops from the 38th Infantry Division), under Lieutenant General Masao Maruyama and comprising 7,000 soldiers in three infantry regiments of three battalions each was ordered to march through the jungle and attack the American defences from the south near the east bank of the Lunga River.Shaw, “First Offensive”, p. 34, and Rottman, “Japanese Army”, p. 63. (...)



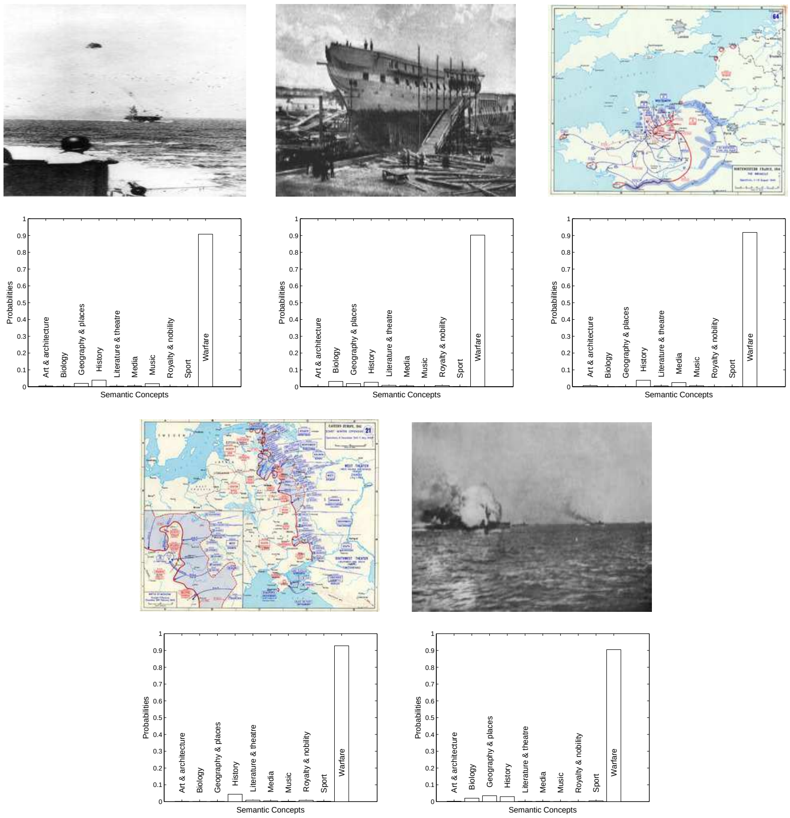
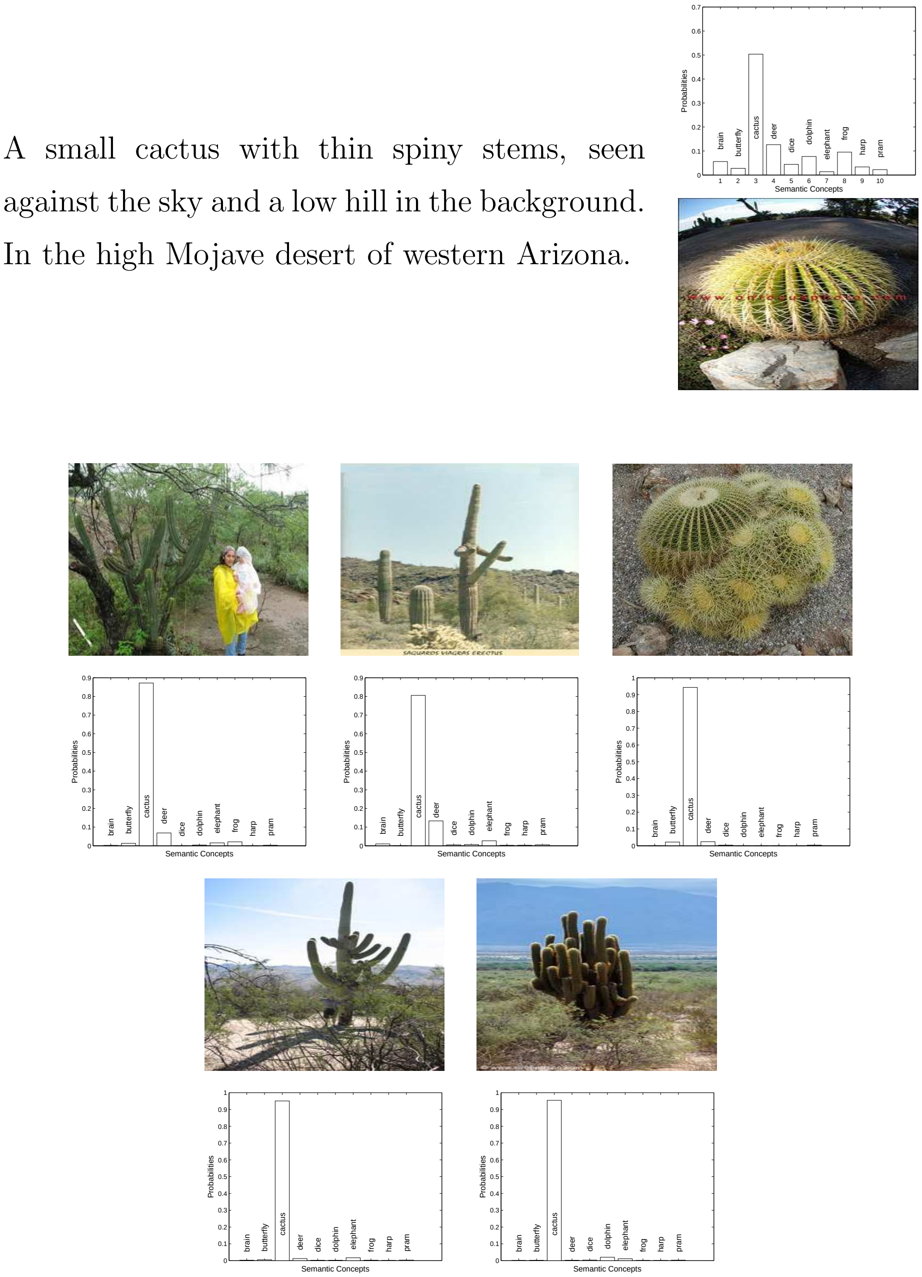


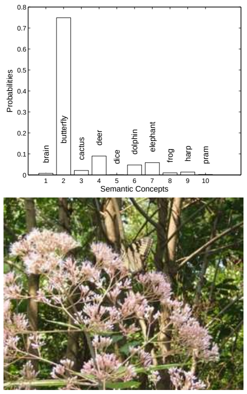
Figure 5.14: Text query from ’Warfare’ class of Wikipedia and the top 5 retrieved images retrieved using SCM. The query text, associated probability vector, and ground truth image are shown on the top; retrieved images are presented at the bottom.



A small cactus with thin spiny stems, seen against the sky and a low hill in the background. In the high Mojave desert of western Arizona.

Figure 5.15: Text query from 'Cactus' class of TVGraz and the top 5 retrieved images retrieved using SCM. The query text, associated probability vector, and ground truth image are shown on the top; retrieved images are presented at the bottom.

On the Nature Trail behind the Bathabara Church, there are numerous wild flowers and plants blooming, that attract a variety of insects, bees and birds. Here a beautiful Butterfly is attracted to the blooms of the Joe Pye Weed.



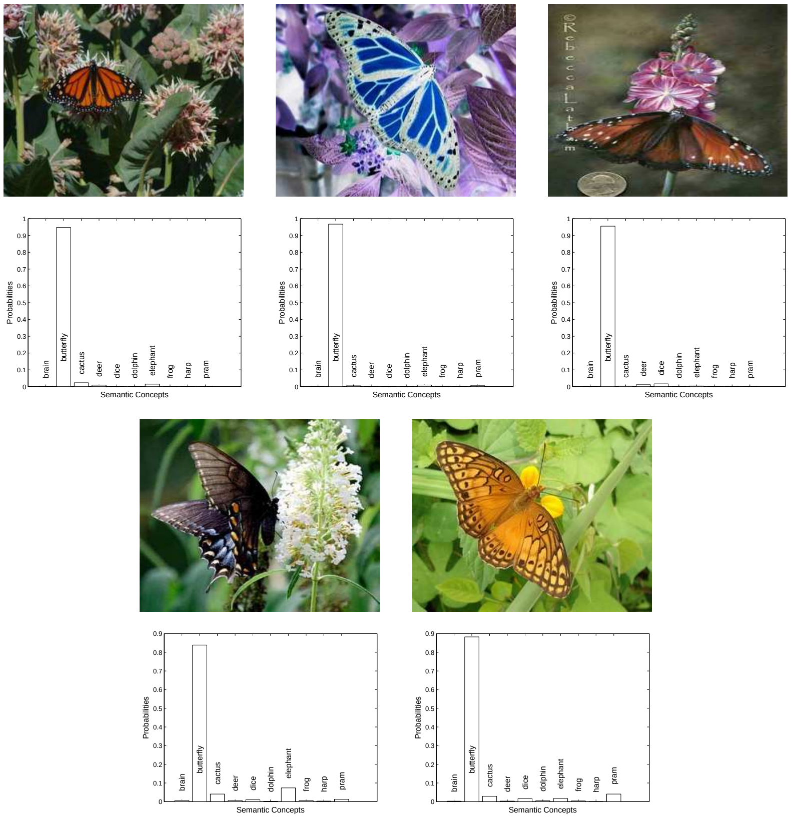


Figure 5.16: Text query from 'Butterfly' class of TVGraz and the top 5 retrieved images retrieved using SCM. The query text, associated probability vector, and ground truth image are shown on the top; retrieved images are presented at the bottom.

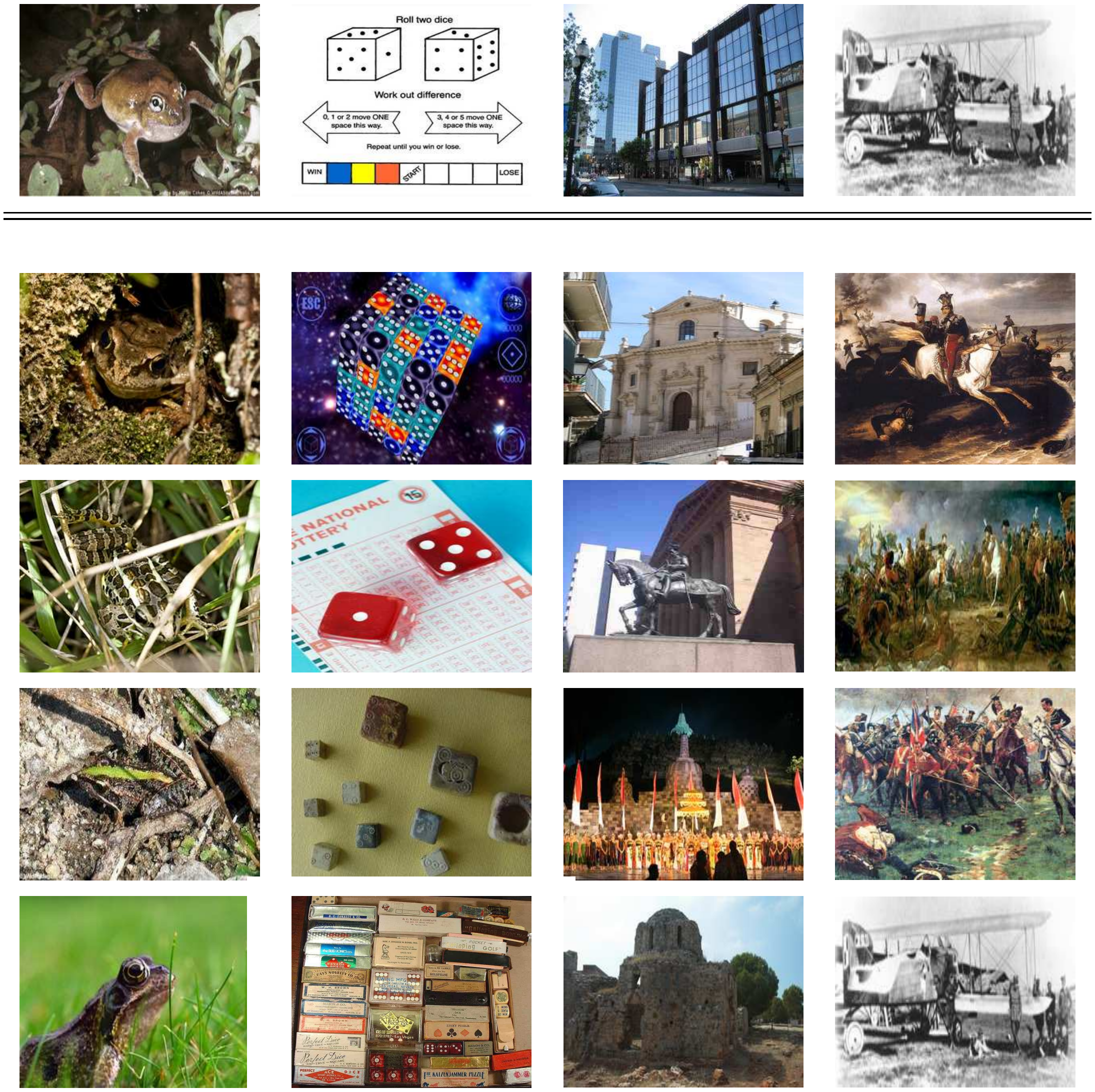

Roll two dice
Work out difference
0, 1 or 2 move ONE space this way.
3, 4 or 5 move ONE space this way.
Repeat until you win or lose.
WIN
START
LOSE

Figure 5.17: Image-to-text retrieval on TVGraz (first two columns) and Wikipedia (last two columns). Query images are shown on the top row. The four most relevant texts, represented by their ground truth images, are shown in the remaining columns.