Cross-modal Classification by Completing Unimodal Representations

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

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

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

ABSTRACT

We argue that *cross-modal* classification, where models are trained on data from one modality (*e.g.* text) and applied to data from another (*e.g.* image), is a relevant problem in multimedia retrieval. We propose a method that addresses this specific problem, related to but different from cross-modal retrieval and bimodal classification. This method relies on a common latent space where both modalities have comparable representations and on an auxiliary dataset from which we build a more complete bimodal representation of any unimodal data. Evaluations on Pascal VOC07 and NUS-WIDE show that the novel representation method significantly improves the results compared to the use of a latent space alone. The level of performance achieved makes cross-modal classification a convincing choice for real applications.

1. INTRODUCTION

We are interested into cross-modal classification, where models are trained on data from *one* modality and applied to data from *another* modality. More specifically, we consider here cases where training is on labelled textual-only data and testing on visual data or, symmetrically, training on labelled visual-only data and testing on textual data.

This task has not been extensively investigated in the literature, first and foremost because text and images are normally not described with the same features, and usually not even in the same vector space, making the task quite incongruous. However, beyond an academic interest, we believe this task also has an increasing practical interest. Suppose, for example, that classifiers for many concepts have been learned from textual data because of the massive availability of such labeled data. One could wish to detect these concepts on content corresponding to another modality, *e.g.* images, even if class labels are not (yet) available for this content. Such a situation may become more common with the current evolution of micro-blogging, that changes from purely textual content (historical Twitter) to multi-modal content (current Twitter) or purely visual content (Instagram). Furthermore, the study of the cross-modal classification task allows to explore in a more clear setting methods that aim to make the best use of the many datasets that *mix* unimodal and bimodal data.

Cross-modal classification can be seen as a step beyond bimodal image classification that usually considers images associated to keywords or sentences (*e.g.* captions) as input data and uses both visual and textual content to solve the task. Cross-modal classification is further related to cross-modal tasks such as text illustration or image annotation that require to match the information from one modality to the other. Various fusion approaches have been extensively employed with some success to address bimodal classification, as in *e.g.* [23]. However, for cross-modal retrieval it was necessary to devise methods that are able to relate the two modalities more closely. The development of a common latent representation space, resulting from a maximization of the "relatedness" between the different modalities, is a generally adopted solution. These methods typically rely on Canonical Correlation Analysis or its kernel extension [10, 12, 6, 9] and on deep learning [17, 22, 8, 7, 15, 24].

In the common latent space, visual and textual information have similar representations and become directly comparable. Hence, it is perfectly conceivable to train a classifier on vectors of the common space that are projections of textual features and predict an output for a vector that is a projection of a visual feature. Such a common space was employed for cross-modal retrieval, *i.e.* information retrieval with both unimodal and multimodal queries [17, 22, 12, 22, 9, 7]. However, to the best of our knowledge, no attempt was made to employ classifiers as we suggested. This is precisely the question we investigate in this paper.

As we show in Section 4, the direct approach described above is not very effective for cross-modal classification. This may explain why corresponding results have not been reported in the literature. We propose instead to "complete" the projection of a unimodal feature on the common space with information coming from the other modality. For this, we suggest to rely on an auxiliary multimodal dataset that acts as a set of connections between the modalities within the common latent space. Such a dataset is always available, as it is required for obtaining the common space. However, we further consider the case where the auxiliary dataset is totally different from the one employed to learn the common space. While we also mention a "naive" approach based on the auxiliary dataset, we propose a more sophisticated one to identify the complementary information, leading to significantly better results. Last, our method includes a step

iV&L-MM'16, October 16 2016, Amsterdam, Netherlands

DOI: http://dx.doi.org/10.1145/2983563.2983570

that aggregates the original vector coming from the projection of a unimodal feature with the identified complementary information to synthesize a unique representative vector of the document. This new representation thus embeds both modalities. Consequently, learning a classifier with such a representation and applying it to unimodal documents naturally leads to much better results than the "direct" approaches.

The rest of the paper is organized as follows. In Section 2 we briefly review recent work having more direct connections with cross-modal classification. The approach we propose, including the projection on a common space, the "completion" of the missing modality and the construction of an aggregated representation, is presented in Section 3. Section 4 reports the evaluation results on Pascal VOC07 and NUS-WIDE. Comparisons are performed with two baselines, showing that the proposed method leads to significant performance improvements. We eventually compare the cross-modal classification results we obtained to state-of-the-art results concerning cross-modal retrieval, as well as unimodal and bimodal classification, showing that the performance level attained in cross-modal classification makes it a convincing choice for real applications.

2. RELATED WORK

While different, cross-modal classification relates to cross-modal retrieval. Recent work in this area is based on the projection of visual and textual features onto a common latent space. Canonical Correlation Analysis (CCA) and its kernelised version (KCCA) were applied to cross modal retrieval in the seminal work of [10]. The principle is to compute a common latent space from both visual and textual features such that the correlation between the projections of both modalities for a given bi-modal dataset is maximized. All the documents of a reference database are then projected onto the common latent space. When a query is processed, it is also projected onto this space and its nearest neighbors can be found directly, independently of their original modality (visual or textual), according to their similarity in the latent space.

A refinement was proposed by [12] to take into account the objects present in a scene with their relative significance within that scene. This is modelled by the rank of the tags used by an "ordinary user" to describe the scene. Then, KCCA is employed with an average kernel over three features to describe the visual aspect and three other features for the textual aspect, including the relative and absolute tag rank.

An important extension to the method of [10] was put forward in [9], where a third view was added to the traditional two-view algorithm. Above the visual and textual view, semantic classes are also considered. They are derived from the ground-truth annotations, the search keyworks used to download the images or, in an "unsupervised scenario", from a set of clusters obtained from the tags. KCCA is reformulated as a linear CCA applied to the kernel space, following the idea of approximate kernel maps [19].

In Semantic Correlation Matching [6] the image and text features projected onto a KCCA space are used to build semantic features, *i.e.* a document is represented as a set of supervised classifiers learned from projections on the common latent space. These classifiers are only employed to represent a unimodal document and the contribution addresses cross-modal retrieval alone. Ahsan et al. [1] employed the concatenation of textual and visual KCCA-descriptors as inputs of a clustering algorithm to perform a bi-modal task, social event detection.

As an alternative to KCCA for obtaining a common latent space, several recent publications investigated deep networks [17, 22, 8, 7, 15, 24]. Among them, Ngiam *et al.* focuses on the the use of a deep autoencoder to learn features over multiple modalities. Similarly to the present paper, they are interested in cross-modal learning, although they focus on leaning representations of speech audio coupled with videos of the lips. Their approach explicitly learns the absence of a modality, by feeding the network with incomplete data during training. Results are nevertheless much better when CCA is first applied to learn a common representation. To handle a unimodal document, [22] propose to complete it by clamping the observed modality at the inputs of their Restricted Boltzman Machine and sampling the hidden modalities from the conditional distribution by running the standard alternating Gibbs sampler.

More recently, [7] addresses cross-modal retrieval by training a "correspondence" autoencoder between visual and textual features. It allows to learn a common latent space in which cross-modal retrieval is performed directly, as in [10]. However, cross-modal learning is not investigated in [7], nor modality completion.

The recent literature has also addressed the question of "completing" a missing modality only. In [23], visual classifiers are "enhanced" with textual ones. However, contrary to our approach, they use the original visual and textual feature without projecting them onto a common space. Hence, the textual classifiers can not employ visual data as input and similarly for visual classifiers and textual inputs.

3. PROPOSED APPROACH

Cross-modal classification consists in training models on data from one modality (*e.g.* text) and applying them to data from another modality (*e.g.* image). This requires unimodal labelled training data from the first modality and unimodal testing data from the other modality. However, to *relate* the two modalities, an *auxiliary bi*modal dataset should also be available. There is no need for this dataset to have class labels. As in many cross-modal retrieval methods, this auxiliary dataset is employed for learning a "common" latent space for the two modalities. The projection of unimodal data on this common space makes data representations for the two modalities directly comparable but, as shown in Section 4, this is not sufficient for obtaining good results in cross-modal classification.

What we suggest here is to employ the auxiliary bimodal dataset not only for learning the common latent space but also for *building a bimodal representation* for any unimodal data. Indeed, given a unimodal example, we expect the auxiliary dataset to provide relevant information concerning the *missing* modality. To train the classifiers, such a bimodal representation is first obtained for each unimodal labelled training example and then learning is performed with these synthetic bimodal representations. For each unimodal testing example (in the other modality than the one used for training), first the bimodal representation is built with the help of the auxiliary dataset and then the available classifiers are applied to this representation. Figure 1 illustrates this approach to cross-modal classification.

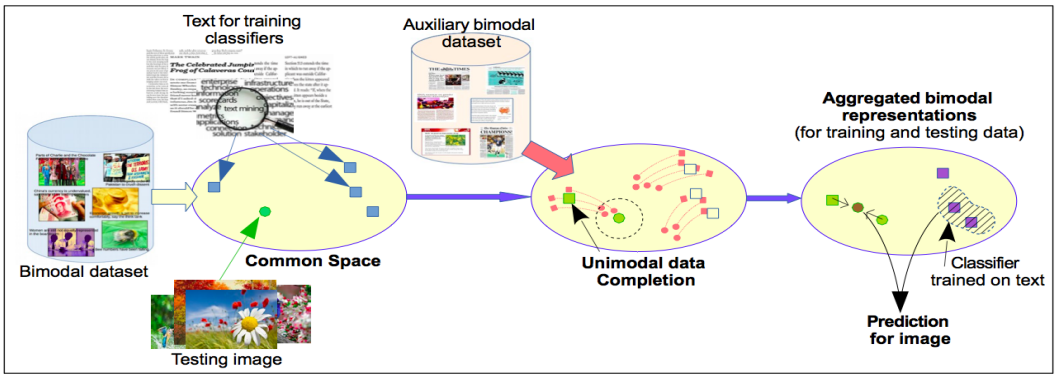

Figure 1: Illustration of the proposed approach for cross-modal classification

To obtain the common latent space we employ here Kernel Canonical Correlation Analysis [10] because it has a clear theoretical foundation and was extensively used in the cross-modal retrieval literature (*e.g.* [10, 12, 6, 9]). However, we believe that the novel later stages we propose, for representation completion and aggregation, can be employed with alternative solutions for building the common latent space. Joint text-image representations could be obtained by using *e.g.* Latent Dirichlet Allocation, Partial Least Squares or deep neural networks. The choice of an algorithm to compute the joint representations should nevertheless be made in compliance with the characteristics of the base unimodal representations employed. The evaluation of our completion approach with alternative latent representations is beyond the scope of this paper.

3.1 KCCA for "common" space learning

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

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

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

In Kernel CCA (KCCA, [10]) the linearity constraint is removed by using the "kernel trick" to first map the data from each initial space to the reproducing kernel Hilbert space (RKHS) associated to the selected kernel and then looking for correlated subspaces in these RKHS. KCCA seeks $2d$ vectors of coefficients $\{\alpha_{T,k}\}_{k=1}^d, \{\alpha_{I,k}\}_{k=1}^d \in \mathbb{R}^N$ defining the maximally correlated subspaces. They are solutions of

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

with $V(\alpha, K) = \sqrt{\alpha^t \, (K^2 + \kappa \, K) \, \alpha}$, $\kappa \in [0, 1]$ is the regularization parameter and K_T, K_I denote the $N \times N$ kernel matrices issued from $\{x_i^T\}_{i=1}^N$ and $\{x_i^I\}_{i=1}^N$. Finding the solutions amounts to solving a generalized eigenvalue problem and keeping the d highest eigenvalues together with their associated eigenvectors, $\{\alpha_{T,k}\}_{k=1}^d$ and $\{\alpha_{I,k}\}_{k=1}^d$.

The projections p^T of x^T onto $\mathcal{U}^T$ and p^I of x^I onto $\mathcal{U}^I$ are obtained as $p_k^T = [\mathcal{K}_T(x^T, x_1^T) \ldots \mathcal{K}_\mathcal{T}(x^T, x_N^T)]\alpha_{T,k}$ and respectively $p_k^I = [\mathcal{K}_I(x^I, x_1^I) \ldots \mathcal{K}_\mathcal{I}(x^I, x_N^I)]\alpha_{I,k}$, for $k \in \{1, .., d\}$. The bijection between projection spaces for p^T and p^I defined by pairing $\{\alpha_{T,k}, \alpha_{I,k}\}_{k=1}^d$ makes possible to train models on projections of data from one modality and apply them to projections of data from the other modality. This simple approach, used as baseline in Section 4, does not lead to very good results. We instead suggest to *complete* any unimodal data with data for the missing modality estimated from an auxiliary bimodal dataset.

3.2 Finding relevant completion information

Let us consider an auxiliary dataset of m documents, each having both visual and textual contents. Let $\mathcal{A}$ be the set of pairs of KCCA projections of the visual and textual features of these documents on the common space, with $\mathcal{A} = \{(q^T, q^I)\}$, $q^T \in \mathcal{A}^T$, $q^I \in \mathcal{A}^I$, $|\mathcal{A}| = m$. Dataset $\mathcal{A}$ can be seen as a sample of pairs of "linked" points, each concerning one modality. If the points are considered in the original spaces of visual and textual features, these links may be loose because part of the visual content of a document is unrelated to its textual content and conversely. The links should be stronger between the *projections* of the visual and textual features of the documents on the common space. The sample $\mathcal{A}$ provides information regarding the relations between the two modalities. The more representative this sample is, the more reliable is the information.

In practice, the auxiliary dataset $\mathcal{A}$ can be the training data employed to obtain the KCCA space and denoted by $\mathcal{T}$ in Section 4. However, we also consider and evaluate the use of *different* datasets for building the common space and for completion of the missing modality. This can have a practical interest when, for example, the common space is an open resource but the dataset employed to build it is private or no longer available. Alternately, the dataset used to obtain the common space may be too large and generic, thus a smaller but "better focused" auxiliary dataset would be preferable to better model the characteristics of a narrow target domain.

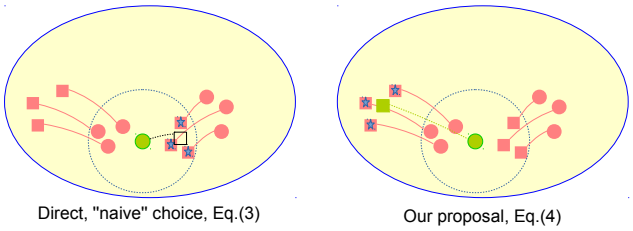

Figure 2: Proposed *vs.* naive completion. Squares and circles are text and resp. image projections. Connected red squares and circles represent bimodal documents in $\mathcal{A}$. The green circle is the projection of an image-only document. The naive approach seeks neighbors in the missing modality directly, while our proposal looks for them in the *available* modality.

To explain the completion process, let us consider a document D with textual content only, described by a feature vector x^T that is projected as p^T on the KCCA space. The method described here (and in Sections 3.3 and 3.4) for a textual-only document can be symmetrically applied to a document having only visual content.

A direct but "naive" choice would be to complete p^T with a vector obtained from its μ nearest neighbors among the points of the auxiliary dataset projected from the *other* modality (visual one here), $NN^{\mu}_{\mathcal{A}^I}(p^T)$, because this is the missing modality for D. This naive choice, considered in Section 4 as a second baseline, can be expressed as

$$\mathcal{M}_c(p^T) = \{q_j^I\} \quad \text{such that} \quad q_j^I \in NN^{\mu}_{\mathcal{A}^I}(p^T) \tag{3}$$

To go further to such an approach, we need to consider the properties of the common space. While it results from an overall maximization of the relatedness between the two modalities, the projections of the textual and of the visual content of a same document on this space are not necessarily very close. So, given the unimodal representation of a document D, its direct nearest neighbors within the other modality are not the best source for "filling in" the missing modality of D. However, we expect that documents having similar content according to one modality are likely to have quite similar content according to the other modality.

So, we propose to find the auxiliary documents having similar projected content with D in the *available* modality for D (textual modality in this case) and to use the projections of the visual content of these documents to complete p^T. Formally, we define the set of contributors to the "modality complement" of p^T as

$$\mathcal{M}_c(p^T) = \{q_j^I\} \quad \text{such that} \quad \begin{cases} q_j^T \in NN^{\mu}_{\mathcal{A}^T}(p^T) \\ (q_j^T, q_j^I) \in \mathcal{A} \end{cases} \tag{4}$$

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

3.3 Completion of the missing modality

Once the relevant complementary information regarding the missing modality of a document D has been collected on the common space as $\mathcal{M}_c(p^T)$, we employ it for building a representation for the missing modality of D.

Let $\hat{p}^I$ be the representation of this missing modality (the visual modality here) on the common space. A simple solution is to obtain $\hat{p}^I$ as the centroid of the q_j^I in $\mathcal{M}_c(p^T) = \{q_j^I\}$, *i.e.*

$$\hat{p}^I = \frac{1}{\mu} \sum_{q_j^I \in \mathcal{M}_c(p^T)} q_j^I \tag{5}$$

With several neighbors ($\mu > 1$), the neighborhood of p^T is better sampled, making the representation more robust. This is confirmed by experiments in Section 4.

The use of the centroid gives equal importance to all the μ neighbors. However, the similarity between p^T and each point $q_j^T \in NN^{\mu}_{\mathcal{A}^T}(p^T)$ should have an impact on the construction of the representation $\hat{p}^I$ of the missing modality. If, within the available modality (textual modality here), p^T is closer to a textual point $q_{j_1}^T$ than to $q_{j_2}^T$ (with $q_{j_1}^T, q_{j_2}^T \in NN^{\mu}_{\mathcal{A}^T}(p^T)$), then within the missing modality (visual modality here) the representation $\hat{p}^I$ should be closer to the corresponding visual point $q_{j_1}^I$ than to $q_{j_2}^I$ (with $q_{j_1}^I, q_{j_2}^I \in \mathcal{M}_c(p^T)$). Consequently, we prefer to define the representation $\hat{p}^I$ of the missing modality for p^T as a *weighted* centroid:

$$\hat{p}^I = \sum_{q_j^I \in \mathcal{M}_c(p^T)} \omega_j q_j^I \tag{6}$$

where ω_j is the weight of q_j^T. Among the μ nearest neighbors of p^T considered, some may be very close to p^T and others comparatively far away. The weighting method should allow to take into account the close neighbors and ignore the others, so the weight should quickly drop when the distance increases. We consequently define the weights as:

$$\omega_j = \frac{\sigma(p^T, q_j^T)}{\sum_{q_j^I \in \mathcal{M}_c(p^T)} \sigma(p^T, q_j^T)} \tag{7}$$

with $\sigma(p^T, q_j^T) = 1/(\epsilon + \|p^T - q_j^T\|)$. Here, $\|p^T - q_j^T\|$ is the Euclidean distance between p^T and each $q_j^T \in \mathcal{M}_c(p^T)$. Also, ϵ is set to 10^{-16} to avoid marginal singularities for the points that may actually belong to the auxiliary dataset. The representative point into the missing modality is thus built from the complementary points of the neighbors found in the available modality and weighted according to the similarity computed in the available modality as well.

To complete the missing modality of a document D with the help of the auxiliary dataset $\mathcal{A}$ according to Eq. 6, it is necessary to retrieve $NN^{\mu}_{\mathcal{A}^T}(p^T)$, the μ nearest neighbors of p^T among the points in $\mathcal{A}^T$. If the auxiliary dataset is relatively small ($\|\mathcal{A}\| \leq 10^6$), exact exhaustive search is fast enough. For larger $\mathcal{A}$ a sublinear approximate retrieval method can be employed, *e.g.* [14, 18].

3.4 Aggregated representation construction

For any unimodal document D originally described by p^T alone, after building the representation $\hat{p}^I$ of the missing modality we aggregate p^T and $\hat{p}^I$ to obtain a unique descriptor p of D. Various aggregation methods can be used and several are compared in Section 4.

A widely employed method is the concatenation of the components, in this case of p^T and $\hat{p}^I$, resulting in a vector of size $2d$. This "unfolded" representation allows the classifier to process the textual and visual components separately but doubles the dimension of the description space.

Max-pooling consists in building a descriptor where the i^{th} element is the maximum between p_i^T and $\hat{p}_i^I$. This method has already been used with good results for bag-of-visual-words (BoVW) representations, see *e.g.* [3]. We also evaluate max-pooling here, even though quantization is not employed for p^T and $\hat{p}^I$.

Averaging is also considered in Section 4. It obtains the aggregated description as the element-wise average of the two components p^T and $\hat{p}^I$:

$$p = (p^T + \hat{p}^I)/2 \quad (8)$$

The approach presented in sections 3.2, 3.3 and 3.4 for a textual-only document can be symmetrically employed for a visual-only document.

In what follows, we call "Weighted Completion with Averaging" (**WCA**) the proposed method with completion following Eq. (4), weights given by Eq. (7) and aggregation by averaging.

4. EXPERIMENTS

We conduct several experiments on publicly available datasets according to standard experimental protocols. Beyond the raw performance of the proposed WCA method and its comparison to baselines, we study the influence of the main components of WCA, namely the completion process, the aggregation method and the relation between the auxiliary data $\mathcal{A}$ and the $\mathcal{T}$ dataset used for KCCA. We then compare the cross-modal classification results of WCA to state-of-the-art results concerning cross-modal retrieval. To better situate the performances attained by WCA on cross-modal classification, we eventually compare them to unimodal and bimodal classification results of the state of the art.

4.1 Datasets, tasks and evaluation metrics

Pascal VOC07 [13]. This dataset includes $5,011$ training and $4,952$ testing images collected from Flickr without their original user tags. Each image has between 1 and 6 labels from a set of 20 labels. Using Amazon Mechanical Turk, several tags were also made available for each image [13]. Each image is associated to 1 to 75 tags for training (6.9 on average) and between 1 and 18 tags for testing (3.7 on average).

NUS-WIDE [5]. This dataset contains $161,789$ training and $107,859$ testing Flickr images with both tags and "ground truth" labels according to 81 concepts. The tag list has 5,018 unique tags. In what follows, we denote the full NUS-WIDE training set by NW160K. We also selected two smaller subsets of NW160K, NW12K of 12K images and NW23K of 23K images, both containing training images for each of the 81 concepts.

NUS-WIDE 10K. We follow the protocol proposed in [7] to collect this NUS-WIDE subset. Thus, only the following ten concepts are chosen: *animal*, *clouds*, *flowers*, *food*, *grass*, *person*, *sky*, *toy*, *water* and *window*. For each of these concepts we select 1000 image-text pairs (800 for training, 100 for validation and 100 for testing) that only belong to this single concept.

Cross-modal classification tasks. We consider here two cross-modal classification tasks: *Text-Image* (T-I) and *Image-Text* (I-T). In the *Text-Image* task, the classifiers are trained with documents that have only textual content and then evaluated on documents in which only the visual modality is available. Symmetrically, in the *Image-Text* task, classifiers are trained with visual-only documents and tested on textual-only documents.

Evaluation metric. The classification results are evaluated using mean Average Precision (mAP), following the literature. Unless otherwise stated, the results shown correspond to the mAP (in %) on the test set and parameters are beforehand chosen on the validation set.

4.2 Experimental settings

To represent visual content we use the 4096-dimensional features of the Oxford VGG-Net [21], L2-normalized. These VGG features were shown to provide very good results in several classification and retrieval tasks [20].

For texts (sets of tags or sentences) we employ Word2Vec [16], an efficient method for learning vector representations of words from large amounts of unstructured text. In our experiments, textual features are 300-dimensional, L2-normalized vectors. User-provided tags are quite noisy for both Pascal VOC07 and NUS-WIDE datasets. For instance, some tags are concatenations of words (*e.g. sunsetoverthesea*), naturally absent from the Word2Vec vocabulary. To improve the quality of textual features, we automatically separate the words (producing *e.g. sunset over the sea*) before employing Word2Vec. For this, each tag is matched to the tag dictionary and we retain only the valid largest substrings. Following [16], a single vector is obtained from several tags or a sentence associated to a given image, by summing the vectors of the individual words.

In all the experiments we use the KCCA implementation in [10] to build the common space, with a regularization parameter $\kappa = 0.1$ and a Gaussian kernel with standard deviation set to $\sigma = 0.2$. These are the default values, also employed in other references [11].

For each category, an SVM classifier with a linear kernel is trained, following a one-vs-all strategy. In practice, we use the implementation proposed by Bottou [2].

4.3 Baselines

We compare WCA to two cross-modal classification baselines. The first, denoted by KCCA_0, is simply the direct use of the projections on the KCCA space. The common space is learned from the dataset $\mathcal{T}$ and the two cross-modal tasks are performed without any completion, both for training and testing. More explicitly, classifiers are trained with the projections of one modality on the common KCCA space and tested with projections of the other modality on this space.

The second baseline, denoted by KCCA_{nc} (*nc* stands for "naive completion"), employs the "naive" completion method following Eq.(3). For either training or testing, the available modality is projected on the KCCA space and this projection is then completed, according to Eq. (3), with a vector obtained by the centroid method (Eq. 5) from its μ nearest neighbors among the points in $\mathcal{A}$ projected from the other modality. The averaging aggregation method of Eq. (8) is employed.

4.4 Proposed completion *vs.* naive completion *vs.* no completion

We first study the effectiveness of the completion mechanism for cross-modal classification on all the datasets. In the *Text-Image* task, the classifiers are trained with documents (of the training set) from which the visual content

was removed and then evaluated on testing documents from which the textual content was removed. Symmetrically, in the *Image-Text* task, the classifiers are trained with image-only documents and then evaluated on text-only documents. Table 1 reports the results obtained on these tasks by WCA and compares them to the KCCA_0 and KCCA_{nc} baselines.

On Pascal VOC07, we employ the training examples (5011 image-text pairs) both as training data $\mathcal{T}$ for learning the KCCA space and as auxiliary data $\mathcal{A}$ for the modality completion stage. The best performances of the KCCA_0 baseline (78.98% for Text-Image and 59.88% for Image-Text) are obtained with $d = 4000$ dimensions. For the sake of comparison, the results of the KCCA_{nc} baseline and of WCA are reported in Table 1 for this 4000-dimensional common space. With $\mu = 15$, WCA yields a better performance than the two cross-modal classification baselines (+21.6% and +17.4% on average compared to KCCA_0 and KCCA_{nc} respectively).

On NUS-WIDE and NUS-WIDE 10K, the common space is learnt from the data in NW23K. Subsequently, the 161,789 training and 107,859 testing examples of NUS-WIDE (respectively the 8,000 training and 1,000 testing data of NUS-WIDE 10k) are projected onto the common space to perform cross-modal classification tasks. We use NW23K as auxiliary data $\mathcal{A}$ to complete unimodal data in the NUS-WIDE benchmark. The 8,000 training plus 1,000 validation data in NUS-WIDE 10K are employed together as auxiliary data $\mathcal{A}$ for the NUS-WIDE 10K benchmark. In this experiment, the number of neighbors μ used for completion is set to 10 both for the KCCA_{nc} baseline and for WCA. The best performances of KCCA_0 and KCCA_{nc} are obtained with $d = 10$ for the two datasets. In this 10-dimensional common space, WCA (with $\mu = 10$) outperforms these two baselines by reaching a mAP of 18.81% for the Text-Image task and 17.9% for the Image-Text task on the NUS-WIDE dataset, and respectively 58.62% and 52.77% on NUS-WIDE 10K. WCA further improves these results for higher values of d. The WCA results in Table 1 are obtained with $d = 1000$.

4.5 Influence of the completion and aggregation methods

We study the influence of the different completion and aggregation methods described in Section 3 on the performance obtained on Pascal VOC07, with the same parameters as in Section 4.4. The training examples in Pascal VOC07 were employed here both for KCCA learning ($d = 4000$) and as auxiliary dataset $\mathcal{A}$. WCA uses the weighted centroid for completion, and aggregation by averaging. "Weighted+Concatenation" combines the weighted centroid for completion with aggregation by concatenation. "Weighted+Max" also employs the weighted centroid for completion but max-pooling for aggregation. The "Centroid+Average" method uses the unweighted centroid (Eq. 5) for completion and average-pooling for aggregation. For each method, we report in Figure 3 the average of the mAP values obtained for the Text-Image and Image-Text tasks with $\mu \in \{1, 5, 10, 15\}$.

Both WCA and "Centroid+Average" perform significantly better than KCCA_{nc}, showing the interest of the proposed completion method of Eq. (4) in comparison to the naive completion of Eq. (3). Averaging is consistently better than max-pooling but the difference is small. Both averaging and max-pooling are significantly better than concatenation.

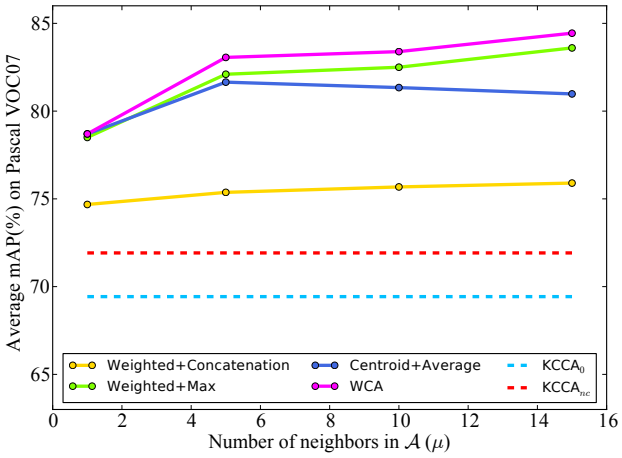

Figure 3: Results of different completion and aggregation methods on Pascal VOC07, showing the mAP(%) with respect to the number of neighbor points μ used in the auxiliary dataset $\mathcal{A}$. For each method, the curves are the average of the Text-Image and Image-Text tasks.

Performance increases with the number of neighbors μ for the three aggregation methods if the weighted centroid is employed, while with the unweighted centroid mAP slightly diminishes beyond $\mu = 5$. Indeed, for higher values of μ some neighbors that contribute to the completion of the missing modality are not near enough to be representative. With the weighting function in Eq. (7), the near neighbors are taken into account while those that are not "near enough" are ignored when present.

4.6 Impact of the auxiliary dataset $\mathcal{A}$ and of the common space

One of our motivations is to develop a common representation space (using here KCCA) as a generic "resource", from a large and general bimodal dataset $\mathcal{T}$, then address different cross-modal classification problems with this resource. This allows to avoid re-learning a common space for each problem, using a specific problem-related dataset. Projections onto this space benefit from the generic text-image relations learned from $\mathcal{T}$. A different, potentially more problem-related dataset $\mathcal{A}$ can then be employed for representation completion, taking thus into account problem-specific text-image links in the aggregated data representation.

To explore this idea, we study in this section the impact of using different datasets for obtaining the common KCCA space (dataset $\mathcal{T}$) and for completing the unimodal representations (dataset $\mathcal{A}$) on Pascal VOC07.

Impact of the auxiliary dataset $\mathcal{A}$. We fix the dataset $\mathcal{T}$ employed for learning the KCCA space as the bimodal training set of Pascal VOC07. The dimension of the common space is set to $d = 4000$ because the baseline KCCA_0 reaches its best performance for this value. While in the previous experiments the auxiliary dataset $\mathcal{A}$ was the same as $\mathcal{T}$, here we successively evaluate as auxiliary dataset NW12K, NW23K, NW160K and eventually $\mathcal{T}$. The number μ of nearest neighbors in $\mathcal{A}$ used for data completion is set to 5. The cross-modal classification results on the Pascal VOC07 test set are reported in Table 2.

The results show that the performance of WCA depends

Method	Pascal VOC07			NUS-WIDE			NUS-WIDE 10K		
	T-I	I-T	Average	T-I	I-T	Average	T-I	I-T	Average
$KCCA_0$	78.98	59.88	69.43	16.97	11.69	14.33	53.34	43.69	48.51
$KCCA_{nc}$	75.07	68.77	71.92	14.87	11.61	13.24	46.41	39.28	42.85
WCA	**85.49**	**83.38**	**84.44**	**37.80**	**34.02**	**35.91**	**79.53**	**79.15**	**79.34**

Table 1: Cross-modal classification results (mAP%) on Pascal VOC07, NUS-WIDE and NUS-WIDE 10K.

Method	$\mathcal{A}$	T-I	I-T	Average
$KCCA_0$	-	78.98	59.88	69.43
$KCCA_{nc}$	VOC07	75.07	68.77	71.92
WCA	NW12K	65.06	57.30	61.18
	NW23K	69.34	60.49	64.92
	NW160K	73.77	64.47	69.12
	VOC07	83.79	81.33	82.56

Table 2: Results on Pascal VOC07 with the common space obtained from the Pascal VOC07 training set. Different auxiliary datasets $\mathcal{A}$ are used for WCA (with $d = 4000, \mu = 5$).

Method	$\mathcal{T}$	$\mathcal{A}$	T-I	I-T	Avg.
$KCCA_0$	NW12K	-	11.03	15.62	13.33
WCA	NW12K	NW12K	21.10	37.12	29.11
		NW23K	23.05	40.33	31.69
		VOC07	59.92	75.48	67.70
$KCCA_0$	NW23K	-	14.93	19.99	17.46
WCA	NW23K	NW12K	26.37	41.84	34.11
		NW23K	29.32	43.11	36.22
		VOC07	67.82	75.24	71.53
$KCCA_0$	VOC07	-	11.68	11.82	11.75
WCA	VOC07	NW12K	45.67	44.63	45.15
		NW23K	50.31	45.76	48.04
		NW160K	56.50	52.49	54.50
		VOC07	80.70	76.23	78.47

Table 3: Results on Pascal VOC07 with different datasets $\mathcal{T}$ to learn the common space and different auxiliary datasets $\mathcal{A}$ (with $d = 100$, $\mu = 5$) to connect the modalities in the common space.

both on the size of the auxiliary dataset $\mathcal{A}$ and on the "agreement" between $\mathcal{A}$ and the specific classification problem considered. As expected, with NW12K, NW23K and NW160K as auxiliary datasets, the larger $\mathcal{A}$, the better the performance. Nevertheless, the mAP value obtained when $\mathcal{A}$ is the comparatively small (5011 bimodal documents) Pascal VOC07 training dataset is 82.56%, significantly higher than the one obtained when $\mathcal{A}$ is the much larger NW160K dataset (only 69.1%). A first potential explanation is that NUS-WIDE does not sample well the domain in the common space covered by Pascal VOC07. Consequently, given a projection of a unimodal document in Pascal VOC07, its μ nearest neighbors in NW12K, NW23K or NW160K are not as close as the ones in the Pascal VOC07 training set, so completion is less reliable with NUS-WIDE data.

A second potential explanation is that NUS-WIDE is not so well represented by the projections on the common space obtained with KCCA performed on the small Pascal VOC07 training set, because text-image relations may differ between the two datasets. As yet another explanation, we note that the NUS-WIDE data remains noisy even after separating the concatenated tags (Section 4.2). This is shown by the fact that the cross-modal classification results obtained on NUS-WIDE are significantly lower than those attained on Pascal VOC07, see Table 1.

Impact of the common space. In this experiment, we still consider cross-modal classification tasks on Pascal VOC07, but we vary both $\mathcal{T}$ and $\mathcal{A}$. When $\mathcal{T}$ is NW12K or NW23K, the baseline $KCCA_0$ reaches its best performance for $d = 100$. To support comparisons, we consider $d = 100$ and $\mu = 5$ for the entire experiment. Table 3 reports the cross-modal classification results on the Pascal VOC07 dataset for each common space learned from $\mathcal{T} \in$ {NW12K, NW23K, VOC07} and $\mathcal{A} \in$ {NW12K, NW23K, VOC07}. The results for $KCCA_{nc}$ were omitted from Table 3 because they are very close to those of $KCCA_0$.

As seen from Table 3, performance improves when the data in $\mathcal{T}$ is problem-related rather than some other dataset (Pascal VOC07 training set instead of NW23K). This is true even though NW23K is more than four times larger than the training set of Pascal VOC07. Also, with a same $\mathcal{A}$, cross-modal classification results improve for larger $\mathcal{T}$ sampled from the NUS-WIDE data (NW23K instead of NW12K). Using more data for obtaining the common space does improve performance, even if this data (NW12K, NW23K) is not related to the specific problem to be solved (in this case, cross-modal classification on Pascal VOC07).

An interesting observation is that the results are significantly better when $\mathcal{T}$ is NW23K (respectively NW12K) and $\mathcal{A}$ is the training set of Pascal VOC07 than when $\mathcal{T}$ is the training set of Pascal VOC07 and $\mathcal{A}$ is NW23K (respectively NW12K). Using problem-related data as auxiliary dataset $\mathcal{A}$, *i.e.* for completing the unimodal representations, has a much larger positive impact than using problem-related data for obtaining the common space. Together with the fact that the increase in performance from $\mathcal{T}$ = NW12K to $\mathcal{T}$ = NW23K is relatively high, this makes us optimistic about the possibility that, with a much larger but generic $\mathcal{T}$, results can improve beyond the level attained when $\mathcal{T}$ is problem-related.

Another observation is that regardless of the dataset $\mathcal{T}$ used for learning the common space, the highest performance is always obtained with Pascal VOC07 training data as auxiliary dataset $\mathcal{A}$. The result obtained in Table 2 with problem-related $\mathcal{T}$ is thus extended to the use of a $\mathcal{T}$ that is not related to the problem. A smaller but "better focused" auxiliary dataset supports more reliable completion of unimodal representations, with a significant positive impact on cross-modal classification performance. This is also important from a complexity perspective. Indeed, our completion mechanism requires nearest-neighbor retrieval from the projections of the points in $\mathcal{A}$, according to the available modality. If good results can be obtained with a relatively

Classification type	Method	PascalVOC07	NUS-WIDE	NUS-WIDE10K
Uni-modal	VGG [21]	86.10	50.38	78.53
	W2V [16]	82.50	46.57	70.20
Bi-modal	VGG+W2V	86.16	50.87	82.89
	LSMP [4]	n/a	19.30	n/a
Cross-modal	WCA (T-I)	85.49	37.80	79.53
	WCA (I-T)	83.38	34.02	79.15

Table 4: Comparison in terms of mAP(%) with unimodal and bimodal classification results.

Cross-modal Task	Method	I-T	T-I	Avg.
Retrieval	Bimodal AE [17, 7]	25.0	29.7	27.4
	Corr-Full-AE [7]	33.1	37.9	35.5
Classification	WCA	89.2	89.7	89.5

Table 5: mAP@50 for cross-modal *retrieval* and for cross-modal *classification* on NUS-WIDE 10K.

small $\mathcal{A}$ then retrieval can be very fast and sublinear solutions may not be needed.

4.7 Comparison to the state of the art

To our knowledge, cross-modal classification for text and image data was not previously investigated. It is *not* directly comparable, in principle, to the more classical unimodal and bimodal classification scenarios where classifiers are trained and tested with information of a same nature (same modality for the unimodal case, both modalities together for the bimodal case). Since it is nevertheless useful to have an idea of the relative levels of performance attained in these different scenarios, we compare in Table 4 the performance of WCA on cross-modal tasks to state-of-the-art results obtained on unimodal and bimodal classification.

In unimodal classification, for the visual-only (denoted by VGG) and respectively textual-only (W2V) case, classifiers are trained and tested on VGG (resp. W2V) features alone. For bimodal classification, in the VGG+W2V case of Table 4, representations for both training and testing data are produced by concatenating VGG and W2V features. The good results obtained in unimodal classification, also very close to those of bimodal classification with VGG+W2V, show the high effectiveness of the features employed.

On Pascal VOC07, the results of both cross-modal classification tasks are lower but quite close to those of unimodal classification with VGG or bimodal classification with VGG+W2V. On NUS-WIDE the difference is larger and we suspect that this may be due to a comparatively weaker link between the two modalities on this dataset. WCA provides slightly better results than unimodal classification and weaker performance than bimodal classification on NUS-WIDE 10K, we believe that the protocol put forward in [7] selects data where the visual and textual modalities are better related. The mechanism we proposed for completing the unimodal features with complementary information in the missing modality appears to have a very significant contribution in bringing the performance of cross-modal classification closer to the state-of-the-art in unimodal and bimodal classification. We also compare WCA to [4] that reported previous state-of-the-art results for bimodal classification on NUS-WIDE. WCA significantly outperforms this method.

Cross-modal *retrieval* is another well-known task and it may be interesting to see how the cross-modal *classfication* approach proposed here compares to this task. For cross-modal retrieval, the query is an item described along one modality and the ranked answers belong to the other modality. In [17, 7], the cross-modal retrieval results reported on NUS-WIDE 10K employed the available concepts (our class labels) as ground-truth for computing the mAP@50. For our cross-modal classification, the "query" is a decision boundary learned in one modality and the ranked answers are items described along the other modality. Table 5 shows both the mAP@50 results of cross-modal retrieval and of cross-modal classification on NUS-WIDE 10K. Note that [17, 7] employed "classical" low or medium-dimensional features such as color histograms or bag of SIFT descriptors for images and bag of words for text, while we made use of VGG and W2V. The reader should however keep in mind that these two tasks are different, so Table 5 should be interpreted with care.

5. CONCLUSION

We put forward an approach that addresses *cross-modal classification* for visual and textual data, *i.e.* where training is performed with data from one modality and testing with data from the other modality. In line with recent literature on cross-modal *retrieval*, this approach relies on the development of a common latent representation space. The novelty of our approach lies in the use of an auxiliary dataset to systematically *complete* unimodal data, both for training and testing, resulting in more comprehensive bimodal representations. The completion method we propose goes beyond a more direct completion solution that we also mention.

We provide an in-depth study of several aspects of our approach and compare it to recent work in the literature. It outperforms two cross-modal classification baselines and provides interesting results compared to recent cross-modal retrievalmethods. Furthermore, the performance level we attain on cross-modal classification also compares well to state-of-the-art unimodal and bimodal classification results. Such a performance level makes our approach to cross-modal classification a convincing choice for real applications, such as learning classifiers from an existing large amount of annotated textual data and applying them to visual content.

We believe that our representation completion and aggregation approach is not limited to latent spaces obtained by KCCA but can be employed with alternative methods. We intend to investigate this direction in future work.

6. AKNOWLEDGEMENT

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

7. REFERENCES

[1] U. Ahsan and I. Essa. Clustering social event images using kernel canonical correlation analysis. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, pages 814–819, June 2014.

[2] L. Bottou. Large-scale machine learning with stochastic gradient descent. In Y. Lechevallier and G. Saporta, editors, *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010)*, pages 177–187, Paris, France, August 2010. Springer.

[3] Y.-L. Boureau, J. Ponce, and Y. Lecun. A theoretical analysis of feature pooling in visual recognition. In *ICML*, Haifa, Israel, 2010.

[4] X. Chen, Y. Mu, S. Yan, and T.-S. Chua. Efficient large-scale image annotation by probabilistic collaborative multi-label propagation. In *Proceedings of the 18th ACM International Conference on Multimedia*, pages 35–44, NY, USA, 2010.

[5] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng. NUS-WIDE: A real-world web image database from National University of Singapore. In *Proc. of ACM Conference on Image and Video Retrieval (CIVR'09)*, Santorini, Greece., July 8-10, 2009.

[6] J. Costa Pereira, E. Coviello, G. Doyle, N. Rasiwasia, G. Lanckriet, R. Levy, and N. Vasconcelos. On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):521–535, 2014.

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

[8] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, and T. Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, pages 2121–2129, 2013.

[9] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *International Journal of Computer Vision*, 106(2):210–233, Jan. 2014.

[10] D. R. Hardoon, S. R. Szedmak, and J. R. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.

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

[12] S. J. Hwang and K. Grauman. Learning the relative importance of objects from tagged images for retrieval and cross-modal search. *International Journal of Computer Vision*, 100(2):134–153, 2012.

[13] S. J. Hwang and K. Grauman. Reading between the lines: Object localization using implicit cues from image tags. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(6):1145–1158, June 2012.

[14] A. Joly and O. Buisson. Random maximum margin hashing. In *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*, pages 873–880. IEEE, 2011.

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

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

[17] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *International Conference on Machine Learning (ICML)*, pages 689–696, 2011.

[18] D. Novak, M. Batko, and P. Zezula. Large-scale image retrieval using neural net descriptors. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*, pages 1039–1040, 2015.

[19] F. Perronnin, J. Sánchez, and Y. Liu. Large-scale image categorization with explicit data embedding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2297–2304, June 2010.

[20] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. *CoRR*, abs/1403.6382, 2014.

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

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

[23] G. Wang, D. Hoiem, and D. A. Forsyth. Building text features for object image classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1367–1374. IEEE Computer Society, 2009.

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