Discovering Semantic Vocabularies for Cross-Media Retrieval

Amirhossein Habibian[†], Thomas Mensink[†], Cees G. M. Snoek[†‡]
[†]University of Amsterdam [‡]Qualcomm Research Netherlands

ABSTRACT

This paper proposes a data-driven approach for cross-media retrieval by automatically learning its underlying semantic vocabulary. Different from the existing semantic vocabularies, which are manually pre-defined and annotated, we automatically discover the vocabulary concepts and their annotations from multimedia collections. To this end, we apply a probabilistic topic model on the text available in the collection to extract its semantic structure. Moreover, we propose a learning to rank framework, to effectively learn the concept classifiers from the extracted annotations. We evaluate the discovered semantic vocabulary for cross-media retrieval on three datasets of image/text and video/text pairs. Our experiments demonstrate that the discovered vocabulary does not require *any* manual labeling to outperform three recent alternatives for cross-media retrieval.

1. INTRODUCTION

We consider the problem of cross-media retrieval, where for an image query we search for the relevant text or vice versa. Initially, cross-media retrieval emphasized on simple queries made of few keywords or tags [6, 14], but recently they have addressed more complex retrieval problems like searching for an image based on a long article [22, 27, 7, 2, 9], or automatically finding the best sentences as the caption to describe a video [12].

The major challenge in cross-media retrieval is that the query and the retrieval set instances belong to different domains, so they are not directly comparable. In other words, the images are represented by visual feature vectors which have a different intrinsic dimensionality, meaning, and distribution than the textual feature vectors used for the sentences. As a solution, many works aim to *align* the two feature spaces so they become comparable. We discuss below related work based on the level of alignment of feature spaces used, as also illustrated in Figure 1.

Low-level Alignment Works based on low-level alignment aim to align the images and texts directly from the low-level feature spaces, *i.e.,* the low-level visual descriptors and the individual words. Cross-media hashing [25, 28], canonical correlation analysis (CCA) [13] and its variants, such as cross-modal factor analysis [17] and Kernel-CCA [2, 7], are the main techniques for learning the low-level alignments. In a nutshell, these techniques extract the highly correlated features in the two feature spaces and use them to make a correlated representation of images and text. While encouraging results for cross-media retrieval using the low-level alignments have been reported, they suffer from two limitations. First, a low-level feature space is not the most effective place to find correlations, as the semantic gap is maximized. Second, the learned alignments are difficult to interpret, making it hard for a user to explain why a certain result was retrieved.

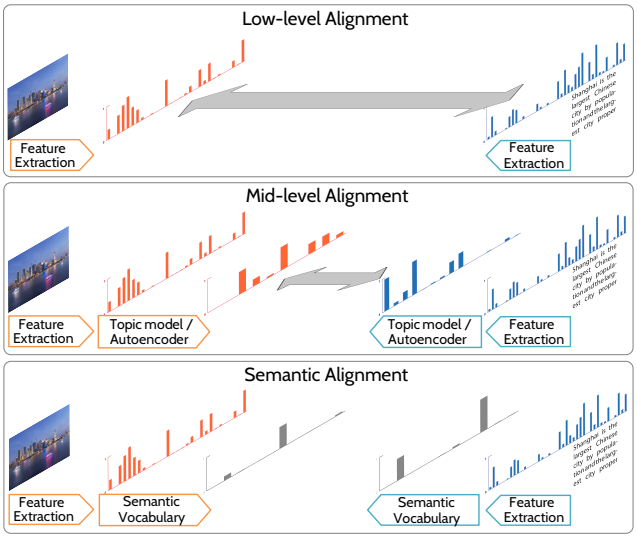

Figure 1: Level of alignments (gray) for cross-media retrieval using visual (orange) and textual (blue) features. This paper contributes to semantic alignment.

Mid-level Alignment Works based on mid-level alignment first extract mid-level features from each modality. Then the alignment is learned between the two mid-level feature spaces. Multi-modal topic models [6, 26] and multi-modal autoencoders [9, 19, 24] are the two major trends for learning the mid-level alignments

Blei and Jordan [6] were the first to extend the latent Dirichlet allocation (LDA) to align text and images at the

ICMR'15, June 23-26, 2015, Shanghai, China.

http://dx.doi.org/10.1145/2671188.2749403.

topic level. Despite their effectiveness in aligning the discrete visual features, such as the traditional bag-of-visual-words, applying the multi-modal topic models on modern visual features, such as Fisher vectors [23] or deep learned visual features [20], is not straightforward. Since probabilistic topic models rely on assumptions about the prior distribution of features, which are not well known for state-of-the-art visual descriptors, the multi-modal topic models cannot benefit from them.

Recently, there has been a trend of developing multi-modal autoencoders with deep architectures for learning the mid-level alignments [9, 19, 24]. These works learn the mid-level representation of images and texts by training autoencoders. Then they align the learned representation by CCA [19] or regression [24]. Recently, Feng *et al.* [9] propose to learn the alignments jointly with the autoencoders. These techniques are shown to be effective for learning the mid-level alignments from large training collections. However, when there is not enough training data available, the deep autoencoders might be overfitted due to their large number of parameters. Moreover, similar to the low-level alignments, the mid-level alignments are incapable of recounting why an image and text are similar.

Semantic Alignment Instead of explicitly learning the correspondences between the images and texts, either at low-level or mid-level features, Rasiswasia *et al.* [22, 7] propose to embed the images and texts into a mutual *semantic space*. In the semantic space, each image or text is represented in terms of the probabilities of being relevant to a pre-defined *vocabulary* of semantic concepts. By representing the images and texts as their concept probabilities, they have aligned representations which are directly comparable.

The semantic representations are obtained by following three consecutive steps: First, the vocabulary is defined by specifying its concepts. The vocabulary concepts should be diverse and comprehensive enough to provide a descriptive representation of images and texts. As the second step, a train set of image/text pairs are labeled as relevant or irrelevant to each vocabulary concept. Finally using the labeled train data, a set of visual and textual concept classifiers are trained to predict the vocabulary concepts on images and texts. Each visual concept classifier is trained on the image parts and the concept labels as a binary classifier *i.e.,* binary SVM or logistic regression. The textual concept classifiers are trained in a similar way, but on the textual data of the train set. After training the concept classifiers, each image or text is embedded into the semantic space by simply applying the visual and textual concept classifiers.

In [22, 7], Rasiswasia *et al.* rely on manual annotations for learning the vocabulary concept classifiers. More specifically, they manually specify the vocabulary concepts and also manually annotate each image/text pair as positive or negative with respect to each vocabulary concept. This requires a substantial amount of annotation effort, which is restrictive in creating descriptive vocabularies with a comprehensive set of concepts. Moreover, for each new dataset a new vocabulary, which is relevant and descriptive to the data at hand need to be defined manually. To overcome these problems, we propose a data-driven approach to automatically discover the vocabulary concepts and their annotations from the textual data, which are available in the dataset.

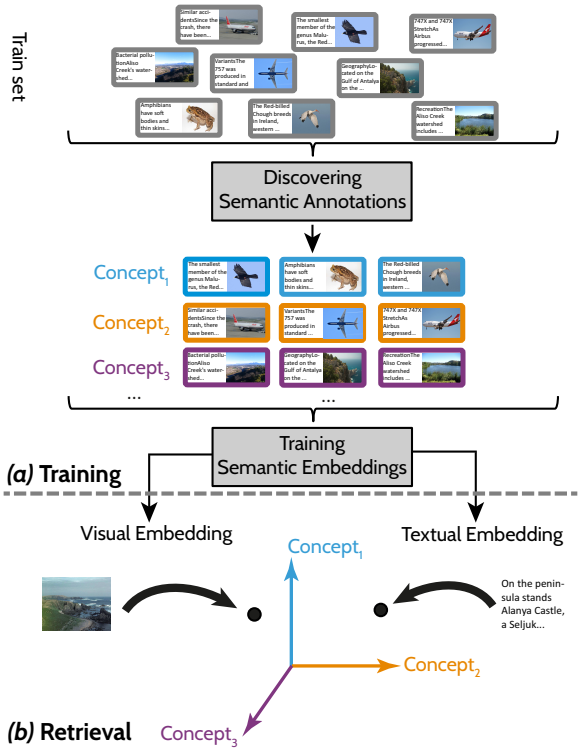

Figure 2: The overall pipeline, which we follow for cross-media retrieval. The "discovering semantic annotations" and "training semantic embeddings", as our main contributions, are detailed in Section 2.1 and Section 2.2

Semantic alignment is related to an attribute-based representation, commonly used for image and video classification from few training examples [8, 3, 11, 10]. In these works, the images or videos are represented as the outputs of attribute classifiers. However, the attribute classifier are applied with the purpose of enriching the image representations by transferring the knowledge from the attributes' training source. This is different from the intention behind the semantic alignment, where the concept classifiers are applied to align the textual and visual instances.

Contributions We propose a data-driven approach to cross-media retrieval using semantic alignment that automatically discovers the vocabulary concepts and their annotations from multimedia collections. Different from [22, 7], we do not pre-define the vocabulary and we do not require *any* human annotation effort to learn the concept classifiers. Moreover, we propose a learning framework for training the vocabulary concept classifiers from the discovered annotations. We experimentally show that our discovered vocabulary outperforms the state-of-the-art low-level, mid-level, and semantic alignments in three datasets of image/text and video/text pairs.

2. OUR PROPOSAL

The overall pipeline, which we follow for cross-media retrieval is shown in Figure 2. In the training phase, we learn two sets of visual and textual embeddings to map each visual or textual instance into a mutual semantic space. We learn the embeddings from a training collection of multimedia in-

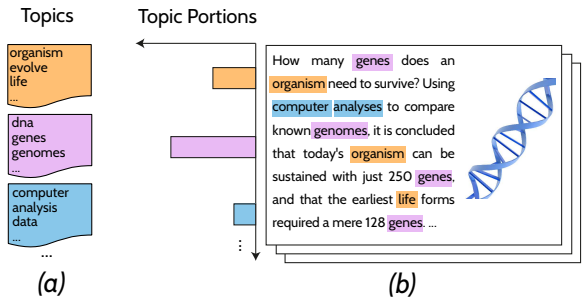

Figure 3: Illustration of topics and topic portions. (a) Three examples of topics extracted from a text collection. (b) Topic portions estimated by LDA for a multimedia instance. Figure inspired from [5].

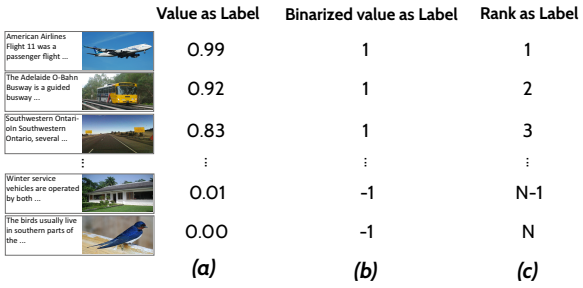

Figure 4: Strategies for extracting concept labels from estimated topic portions.

stances including images or videos and their corresponding textual descriptions. Wikipedia articles, which are made of textual articles and their illustrative images, and captioned videos, which are made of videos and their textual captions, are examples of the multimedia instances used in this paper.

As the first step for learning the embeddings, we automatically specify the vocabulary concepts by discovering a set of concepts, which can effectively describe the training collection. Moreover, we automatically annotate each multimedia instance in the train set as relevant or irrelevant with respect to the vocabulary concepts. Our method for automatically discovering the semantic annotations is detailed in Section 2.1. After extracting the semantic annotations we use them for training visual and textual embeddings, as detailed in Section 2.2.

In the retrieval phase (Figure 2-b), we use the learned visual and textual embeddings to embed the query and the test set instances into the semantic space. Then the cross-media retrieval is cast into a single media retrieval problem, where the traditional similarity measurements, such as cosine similarity, are applicable for the retrieval.

2.1 Discovering Semantic Annotations

Instead of manually specifying the vocabulary concepts and their annotations we propose to automatically discover them by applying a probabilistic topic model.

Probabilistic topic models are statistical models for discovering the hidden semantic structure in a collection of text [4]. They automatically discover sets of interrelated terms, as *topics*, based on co-occurrences of terms in the collection. Figure 3-a illustrates three examples topics extracted from a collection of text.

In addition to discovering the topics, probabilistic topic models also extract *topic portions* for each textual instance in the collection. For each text, the topic portions determine the relevance of the text to each topic, as a probability value between 0 and 1. In the example in Figure 3-b, the text includes the terms from all three topics but with different portions. The more terms from a topic occur in a text, the higher the topic portion value assigned to the topic.

We observe that topics and topic portions are suitable to serve as the vocabulary concepts and their annotations. The topics provide a comprehensive summary of the collection, so a vocabulary composed of topics has a high descriptiveness. Moreover, as experimentally validated in Section 4.1, combining the interrelated concepts into a more abstract concept, as performed by topic modeling, generally improves learning the concept classifiers. More specifically, the abstract concepts, such as "animal", are in general more accurately predicted in images/videos than the specific concepts, such as "goat", "elephant", and "tiger", partly due to the amount of train data available per concept. Hence, a vocabulary composed of discovered topics serves as good concepts for the semantic representation.

To summarize, we apply the latent Dirichlet allocation (LDA) [4], as a widely used probabilistic topic model, on the textual instances in the train set. The number of topics is a parameter which is determined by cross-validation, as is detailed in Section 4.1. Then the discovered topics are considered as the vocabulary concepts. Moreover, the discovered topic portions are considered as the concept annotations and are used for training the visual and textual embeddings.

2.2 Training Semantic Embeddings

The semantic embeddings are defined as projections from visual or textual feature spaces into a mutual semantic space, by predicting the relevance of each instance to the vocabulary concepts. We train the visual and textual embeddings in a similar way. Hence, we first explain our learning algorithm for the visual embedding. Then we explain how the proposed algorithm can be applied for training the textual embedding.

Training Visual Embedding. We denote the visual embedding by $\Phi_{\boldsymbol{W}}(\boldsymbol{x}) : \mathbb{R}^{D_v} \rightarrow \mathbb{R}^k$ as the projection of each visual instance $\boldsymbol{x} \in \mathbb{R}^{D_v}$ into its k-dimensional representation in the semantic space. The visual embedding is defined as:

$$\Phi_{\boldsymbol{W}}(\boldsymbol{x}) = \boldsymbol{W}^\top \boldsymbol{x},$$

where $\boldsymbol{W} \in \mathbb{R}^{D_v \times k}$ is the visual prediction matrix stacking the weight vectors corresponding to each vocabulary concept classifier denoted as $\boldsymbol{w}_c \in \mathbb{R}^{D_v}$.

Each vocabulary concept classifier $\boldsymbol{w}_c$ is trained on its train set $\mathcal{D}_c = \{(\boldsymbol{x}_i, y_i), i = 1 \dots N\}$, where $\boldsymbol{x}_i \in \mathbb{R}^{D_v}$ denotes the visual instances in the train set. For example, in a train set of Wikipedia articles, $\boldsymbol{x}_i$ refer to the visual features extracted from the article's image. Moreover, y_i is the concept label denoting the relevance of the visual instance $\boldsymbol{x}_i$ to the concept c.

As proposed in Section 2.1, instead of manually annotating the concept labels, we automatically extract them from the topic portions estimated by LDA. More specifically, for each visual instance x_i we use the estimated topic portion of its corresponding text to extract the concept label y_i. However, the topic portions are estimated as continuous prob-

ability values between 0 and 1, so using them as labels for training the classifiers is not trivial. We consider three possible strategies for defining the concept labels based on the topic portions and discuss their applicability as follows:

- *Topic portion values as labels*, where the topic portion values are directly used as the concept labels (Figure 4-a). In this case, the concept labels are continuous, so the concept classifier is trained as a regressor. In other words, the $\boldsymbol{w}_c$ is trained to predict the exact value of the topic portions from the visual instances.

- *Binarized topic portion values as labels*, where the topic portion values are first quantized into binary labels, based on a threshold parameter. More specifically, the binary label is equal to 1, if the topic portion is higher than the threshold, otherwise the binary label is set to -1 (Figure 4-b). Then, the concept classifier $\boldsymbol{w}_c$ can be trained as a binary classifier using the binarized labels.

- *Topic portion ranks as labels*, where the visual instances are ordered based on their topic portion values and their rank is considered as the label, (Figure 4-c). In this case, the concept classifier is trained as a ranking function. In other words, the $\boldsymbol{w}_c$ is trained to rank the visual instances by predicting their relative concept relevancies.

Neither continuous values nor binarized labels are appropriate for training the concept classifiers from the estimated topic portions. Predicting the continuous label values from the images is difficult and ineffective, as we will show in the experiments. Although predicting binary labels is in general more simple than predicting continuous labels, a substantial amount of information in the topic portions can be lost by the binarization. Hence, we speculate that defining the topic portion ranks as labels and learning the concept classifiers as a learning to rank problem is the most effective solution for training the concept classifiers from the automatically discovered topic portions.

More formally, we first define each concept label $y_i \in \mathcal{D}_c$ by measuring its topic portion rank as follows:

$$y_i = \sum_{j=1}^{N} \mathbb{1}(\theta_i^c < \theta_j^c), \quad (1)$$

where $\mathbb{1}(.)$ is the indicator function. Moreover, θ_i^c is the estimated topic portion value for instance i for topic c. After determining the labels, the concept classifier $\boldsymbol{w}_c$ can be learned by any learning to rank method. In this paper, we follow the rankSVM formulation [15], which learns the ranking function by minimizing the following objective function:

$$\frac{\lambda}{2}\|\boldsymbol{w}_c\|^2 + \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1, y_i<y_j}^{N} \max(0, 1 - \boldsymbol{w}_c^\top(\boldsymbol{x}_j - \boldsymbol{x}_i)), \quad (2)$$

where λ is the regularizer penalty parameter. By minimizing this objective we learn to predict a higher value for the j^{th} visual instance than the i^{th} visual instance, if its label has a higher rank. We minimize the Eq. (2) for each concept c independently using stochastic gradient descent [1], note that the ranking of the documents will be different for each topic.

Training Textual Embedding. While we could have used the LDA model directly as the textual embedding, we propose to yield a textual embedding which is more aligned with the visual embedding. The textual embedding are trained in a similar fashion as the visual embeddings, however instead of using the visual features the concept classifiers are now trained on the textual features.

Figure 5: Two examples of the videos and their captions from the Captioned Videos dataset [21].

For each vocabulary concept c, the train set $\mathcal{D}_c = \{(\boldsymbol{x}_i, y_i), i = 1 \dots N\}$ is defined, where $\boldsymbol{x}_i \in \mathbb{R}^{D_t}$ denotes the textual instances in the train set. For example, in a train set of Wikipedia articles, $\boldsymbol{x}_i$ refer to the textual features extracted from the article's text. Moreover, y_i are the concept labels, which are defined as the topic portion ranks as determined by Eq. (1). Note, that for each concept c the labels y_i are identical for both the visual embedding and the textual embedding, yielding a desired alignment between the two embeddings. Based on the train set $\mathcal{D}_c$, the textual classifier of the concept $\boldsymbol{w}_c \in \mathbb{R}^{D_t}$ is trained by minimizing the objective function of Eq. (2).

3. EXPERIMENTAL SETUP

3.1 Datasets

We use three multimedia datasets in our experiments, as summarized in Table 1. The first two datasets are made of image/text pairs, while the third set includes video/text pairs.

1. Wikipedia [22]. This dataset is provided by Rasiswasia *et al.* [22] and includes 2,866 Wikipedia articles from 10 categories. Each article is made of few paragraphs in text explaining the article, as well as an illustrative image. We adopt the protocol in [9] and use 2,173 articles as train set, 231 articles as validation set, and the remaining 462 articles as test set.

2. Wikipedia++. We expand the Wikipedia dataset [22] by collecting more articles from more categories. Our dataset, which we name as Wikipedia++, is collected by exactly following the same procedure as used for creating the Wikipedia dataset [22]. More specifically, we collected the articles from the Wikipedia "featured articles". Each featured article is categorized by Wikipedia into one of 30 categories. Excluding scarce categories, with less than 50 articles, we end up with the 20 article categories listed in Table 1. We split each article into sections, based on its section headings, and assign each image in the article to the section in which it was placed. This leads to a set of short and focused articles, containing a single image. Then the dataset is pruned by excluding the sections without any image. The final corpus contains 12,617 articles. We randomly split the articles into three partitions: 50% of the articles as train set, 25% as validation set, and 25% as test set. We make our collected

Table 1: Statistics of the Wikipedia++ and the Captioned Videos datasets used in our experiments. The statistics of the Wikipedia dataset are available in [22]. Our created Wikipedia++ dataset is publicly available at http://www.mediamill.nl.

Wikipedia++		Captioned Videos [21]	
Category	Size	Category	Size
Animal	1,454	Attempting board trick	160
Art & architecture & archaeology	695	Feeding animal	161
Biology	477	Landing fish	119
Business & economics & finance	234	Wedding ceremony	123
Culture & society	326	Working wood working project	141
Education	271	Birthday party	343
Engineering & technology	231	Changing vehicle tire	221
Geography & places	1,758	Flash mob gathering	303
History	793	Getting vehicle unstuck	211
Literature & theatre	453	Grooming animal	218
Media	566	Making sandwich	255
Meteorology	345	Parade	322
Music	400	Parkour	213
Physics & astronomy	573	Repairing appliance	196
Religion & mysticism & mythology	212	Working sewing project	196
Royalty & nobility & heraldry	583	Attempting bike trick	130
Sport & recreation	700	Cleaning appliance	130
Transport	638	Dog show	130
Video gaming	401	Giving directions location	130
Warfare	1,507	Marriage proposal	129
		Renovating home	130
		Rock climbing	130
		Town hall meeting	130
		Winning race without vehicle	130
		Working metal crafts project	130
Total	12,617		4,481

Wikipedia++, which is four times larger than the Wikipedia dataset [22], publicly available.

3. Captioned Videos [21]. This dataset is part of the TRECVID Multimedia Event Detection corpus [21]. It consists of user-generated videos accompanied with human provided textual captions. The dataset contains videos depicting 25 complex events, including life events, instructional events, and sport events. The textual caption of each video summarizes what happens in the content, as shown in Figure 5. We use the videos, which are assigned to one of the 25 event categories. It leads to 4,481 videos and their captions, where 50% of them are used as train set, 25% as validation set and the remaining 25% as test set.

3.2 Evaluation Protocol

We perform our cross-media retrieval experiments by exactly following [22, 9]. The multimedia instances in the train set and validation set are used for learning the visual and textual embeddings. The test set instances are used to perform two cross-media retrieval experiments: *i) Visual Query* experiments, where each image or video in the test set is considered as a query, for which we rank the textual instances in the test set by measuring their similarities. *ii) Textual Query* experiments, where each textual instance in the test set is considered as a query, for which we rank the images or videos in the test set by measuring their similarities. The similarities are measured by first projecting the instances into the semantic space, by the trained visual and textual embeddings, then we use the normalized correlation as the similarity metric [22].

The retrieval performance is evaluated using the mean average precision (mAP) metric, as in [22, 9]. More specifically, for each query and its R top retrieved instances, the

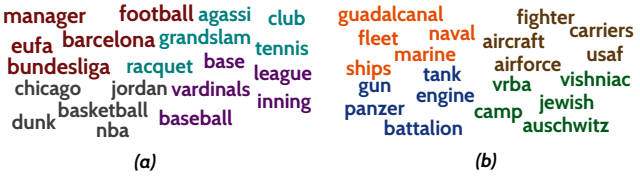

Figure 7: Eight examples of the topics extracted from the Wikipedia++ dataset. Each topic is represented by its five most relevant terms in the same color. The topics in (a) and (b) are more specific than the predefined concepts "sport & recreating" and "warfare", respectively.

AP is measured as:

$$\frac{1}{M}\sum_{r=1}^{R} p(r).rel(r),$$

where M is the total number of relevant instances in the retrieved set, $p(r)$ is the precision at r, and $rel(r)$ is a binary value determining the relevance of the r^{th} ranked instance. If the retrieved instance and the query has the same category label the $rel(.)$ is one, otherwise it is zero. Finally, mAP is obtained by averaging the measured AP values over all the queries. Similar to [9] we report the mAP@50 (R = 50) in all the experiments.

3.3 Implementation Details

Features. We use the deep learned visual descriptors [20] as ***image features***. Each image is fed into a pre-trained convolutional neural network and the output of the second fully connected layer is considered as a 4,096 dimensional feature vector. The convolutional neural network has an AlexNet architecture [16]. It is pre-trained on all the 15,293 categories in ImageNet dataset, for which there are at least 50 positive examples available. The ***video features*** are obtained by first extracting the video frames by uniformly sampling the frames every two seconds. Then each video frame is represented by deep learning features, which are extracted in the same way as the image features. Afterward, each video is represented by the average pooling of its frames as a 4,096 dimensional feature vector. As the ***text features*** we use the term histograms. More specifically, a dictionary of 3,000 high-frequency terms are extracted per dataset. Then each textual instance is represented as the histogram of its terms with respect to the dictionary [9].

Learning parameters. We learn all the vocabulary concept classifiers by the stochastic gradient descent solver [1]. The learning rate parameter η, the regularization penalty parameter λ, and the number of epochs are empirically set to 0.01, 0.001, and 100, respectively.

Topic model parameters. We extract the topics by MALLET implementation of LDA [18]. The Dirichlet prior parameters α and β are empirically set to 1 and 0.1. Moreover, the optimal number of topics for each dataset is obtained by cross-validation over 5, 10, 20, 50, 100, 500, and 1000 topics.

3.4 Experiments

1. Discovering Semantic Vocabulary. We evaluate the effectiveness of our automatically discovered concept labels, as proposed in Section 2.1, for training semantic vocab-

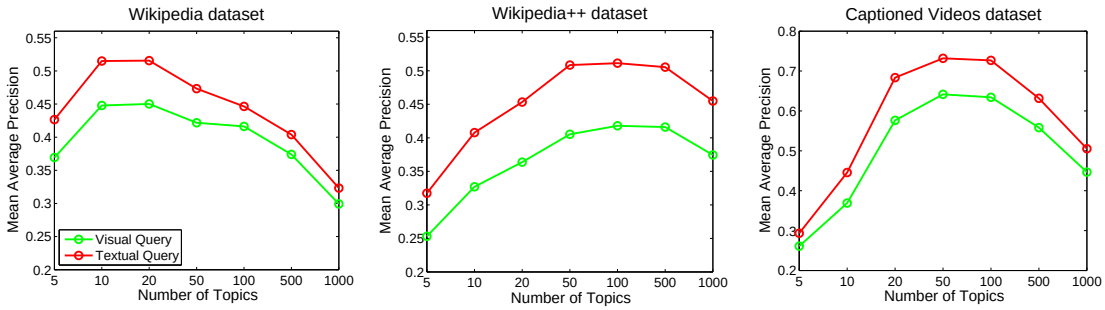

Figure 6: Effect of increasing the number of vocabulary concepts by extracting more topics from text. Extracting more topics generally improves the vocabulary by making it more descriptive. However, extracting too many topics is prone to overfitting, which degrades the effectiveness of the trained vocabulary .

ularies. We compare two concept vocabularies: First, our proposed ***discovered vocabulary***, which is trained on the automatically extracted concept labels, as detailed in Section 2. Second, the ***pre-defined vocabulary*** baseline [7], which is trained on manually provided concept labels. In this baseline, each category in the train set is considered as a vocabulary concept, and the concept labels come from the category labels, which are manually provided per instance. Please note that our proposed discovered vocabulary does not need *any* manual category labels in the train set, so we ignore all of them when training the discovered vocabulary. Each vocabulary is evaluated by its cross-media retrieval accuracy.

2. Training Semantic Embeddings. We evaluate our proposed strategy for training the semantic embeddings from the discovered vocabulary labels, as detailed in Section 2.2, by comparing three concept vocabularies: First, our proposed vocabulary of ***concept rankers***, where the concept classifiers are trained as rankers on the topic portion ranks as labels. The rankers are trained based on the rankSVM formulation [15]. Second, a baseline vocabulary of ***concept binary classifiers***, where the concept classifiers are trained as binary classifiers on the binarized topic portion as labels. The binary classifiers are trained based on the binary SVM formulation. Third, a baseline vocabulary of ***concept regressors***, where the concept classifiers are trained as regressors on the topic portion values as labels. The regressors are trained based on the ridge regression formulation. To be consistent, we use the stochastic gradient descent solver for training all the rankers, binary classifiers, and regressors with the same parameter settings as detailed in Section 3.3.

3. Comparison to Others Alignments. We investigate the effectiveness of our discovered vocabulary for cross-media retrieval by comparing it with three state-of-the-art baselines: ***i) CCA*** [22], as a low-level alignment, which uses CCA to align the two modalities. ***ii) Correspondence AE*** [9], as a recent mid-level alignment, which uses correspondence autoencoders with deep architectures to align the textual and visual features. ***iii) pre-defined vocabulary*** [7], as a semantic alignment, which uses a pre-defined vocabulary of concept classifiers to embed the textual and visual instances in the mutual semantic space. For the CCA and pre-defined vocabulary baselines, we run the author's implementation on our features. For the correspondence AE baseline the numbers are exactly reported from [9].

Table 2: Experiment 1. Evaluating the discovered vocabulary. The discovered vocabulary is more effective and does not require manual supervision during training of concept classifiers.

	Visual Query		*Textual Query*	
Dataset	Pre-defined [7]	Discovered	Pre-defined [7]	Discovered
Wikipedia	0.431	**0.450**	0.491	**0.516**
Wikipedia++	0.377	**0.418**	0.493	**0.511**
Captioned Videos	0.528	**0.642**	0.627	**0.732**

4. RESULTS

4.1 Discovering Semantic Vocabulary

The results are shown in Table 2. The proposed discovered vocabulary outperforms the pre-defined vocabulary consistently for both cross-media retrieval tasks on all the three datasets. It demonstrates that the vocabulary concept labels can be effectively extracted from the textual instances in the train set without any manual supervision.

We provide two reasons to explain the better performance of the discovered vocabularies over the predefined vocabularies: First, many of the manually defined vocabulary concepts are very general, *i.e.*, the concepts "warfare" and "sport & recreation" in the *Wikipedia++*. These general concepts have a large diversity in their training examples, which undermines the accuracy of their concept classifiers. In contrast, the discovered vocabulary defines the vocabulary concepts by topic modeling, so is able to define more specific concepts by extracting an appropriate number of topics. Figure 7 shows some examples of the specific concepts, which are discovered by topic modeling.

As the seconds reason, the pre-defined vocabularies include a limited number of concepts, which are not enough to effectively represent all of the instances. More specifically, for the *Wikipedia*, *Wikipedia++*, and *Captioned Videos* datasets, the pre-defined vocabulary includes 10, 20, and 25 concepts as pre-defined in the dataset. In contrast, our discovered vocabulary is able to discover higher number of vocabulary concepts by extracting more topics from the textual instances. We further investigate the impact of increasing the number of vocabulary concepts by extracting more topics in Figure 6. It demonstrates that extracting more topics generally leads to more comprehensive and descriptive concept vocabulary, which is more effective for cross-media retrieval. However, after a certain vocabulary size,

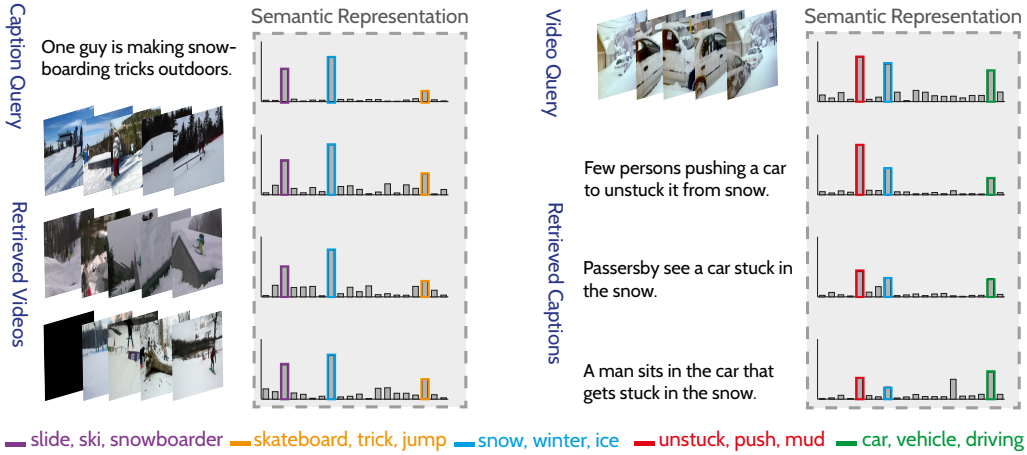

Figure 8: Top three retrievals for a caption query (left) and a video query (right) in the Captioned Videos dataset. The semantic representations are shown in the gray box. The common concepts between the query and the retrieved instances, as highlighted by colors, explain why the query and the retrieved instances are similar. Each highlighted concept is represented by its three most relevant terms in the bottom.

Table 3: Experiment 2. Evaluation of the three strategies for training vocabulary classifiers from the discovered concept labels. The most effective vocabulary is obtained by training ranking functions on topic portion ranks as labels.

	Visual Query			*Textual Query*		
Dataset	Concept Binary Classifiers	Concept Regressors	Concept Rankers	Concept Binary Classifiers	Concept Regressors	Concept Rankers
Wikipedia	0.416	0.412	**0.450**	0.497	0.487	**0.516**
Wikipedia++	0.342	0.401	**0.418**	0.406	0.495	**0.511**
Captioned Videos	0.453	0.594	**0.642**	0.522	0.706	**0.732**

extracting more topics has a negative effect on the vocabulary. We explain it by the fact that when we extract too many topics from a small collection of texts, the extracted topics might be overfitted to the random co-occurrences of terms. It is also observable in Figure 6, where there is a relation between the dataset size and the optimal number of topics. The optimal number of topics for *Wikipedia++*, as the largest dataset in our experiments, is 100 that is higher than for the other two datasets.

As the conclusion, the results demonstrate that by discovering the concept labels from the textual instances, we not only alleviate the manual labeling but also train a more effective vocabulary of concepts for cross-media retrieval.

4.2 Training Semantic Embeddings

The results are shown in Table 3. The vocabulary of concept rankers consistently outperforms the vocabularies of concept binary classifiers and concepts regressors for the both cross-media retrieval tasks on all the three datasets. It demonstrates that defining the concept labels as the topic portion ranks and training ranking functions on them is the most effective strategy for learning from the automatically discovered concept labels.

The lowest performing vocabulary is obtained with the concept binary classifiers, which are trained as binary SVMs on the binarized topic portions as labels. We explain it by the fact that lots of information in the topic portions are lost by binarization. More specifically, all the instances above the binarization threshold are equally considered as positive examples without consideration of their relevancy degrees encoded in topic portions. The low performance of the concept binary classifiers implies the importance of the relevancy degrees for training the vocabulary.

The concept regressors vocabulary performs better than the binary classifiers vocabulary, but it is still outperformed by the vocabulary of concept rankers. As an explanation, we speculate that the concept regressors does not binarize the topic portions, so are not suffered by the information lost as in the binary classifiers. However, predicting the exact value of the topic portions from the visual features, as targeted by the concept regressors, is generally very hard. Hence the trained concept regressors might be inaccurate. This drawback is relaxed by the concept rankers by predicting the relative orders between the topic portions instead of their exact values. As a conclusion, training the concept rankers is the most effective strategy, compared to the alternatives, for training concept classifiers from the automatically discovered labels.

4.3 Comparison to Other Alignments

The results are shown in Table 4. Our discovered vocabulary consistently outperforms the state-of-the-art alternatives for the both cross-media retrieval tasks on all the three datasets. It demonstrates the effectiveness of our proposed vocabulary for cross-media retrieval.

Furthermore, the results demonstrate that the both predefined and discovered vocabularies substantially outperform the CCA and correspondence autoencoders baselines. We explain the relatively low performance of the correspondence autoencoders by the fact that this baseline uses a deep

Table 4: Experiment 3. Comparison of our discovered vocabulary with three state-of-the-art alignments. The discovered vocabulary consistently outperforms the alternatives.

	Visual Query				*Textual Query*			
Dataset	CCA [22]	Correspondence AE [9]	Pre-defined [7]	Discovered	CCA [22]	Correspondence AE [9]	Pre-defined [7]	Discovered
Wikipedia	0.348	0.335	0.431	**0.450**	0.359	0.368	0.491	**0.516**
Wikipedia++	0.352	N.A.	0.377	**0.418**	0.412	N.A.	0.493	**0.511**
Captioned Videos	0.475	N.A.	0.528	**0.642**	0.545	N.A.	0.627	**0.732**

architecture, which generally requires large amount of training data to be effectively learnt. However, the largest train set in our experiments includes 6K examples, which seems to be not large enough. Moreover, the low performance of CCA baseline, as a low-level alignment, validates that learning the correspondences between the two modalities directly from the low-level features is not effective, since the semantic gap is maximized.

Besides their higher retrieval accuracies, the semantic vocabularies provide interpretable representation of instances, which make it possible to explain why two visual and textual instances are similar, as shown in Figure 8. In summary, the discovered vocabulary is not only effective for cross-media retrieval, but also recounts why the instances are retrieved.

5. CONCLUSION

We propose a data-driven approach for cross-media retrieval by automatically learning its underlying semantic vocabulary, rather than specifying and annotating the vocabulary as commonly done. We demonstrate that the textual instances in cross-media collections are a rich source of semantics, which can be utilized to (weakly) supervise the concept classifier training. More specifically, we demonstrate that probabilistic topic models are effective tools to extract the underlying semantic structures from a collection of text. Moreover, we experimentally show that learning to rank is an effective strategy for learning the classifiers from the text-driven annotations. Our experiments show that the discovered vocabulary outperform the state-of-the-art alternatives for cross-media retrieval. These conclusions may generalize for any other problem, where the textual descriptions are served as labels for concept classifier training.

Acknowledgments This research is supported by the STW STORY project and the Dutch national program COMMIT.

6. REFERENCES

[1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Good practice in large-scale learning for image classification. *IEEE TPAMI*, 36:507–520, 2014.

[2] L. Ballan, T. Uricchio, L. Seidenari, and A. Del Bimbo. A cross-media model for automatic image annotation. In *ICMR*, 2014.

[3] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*, 2010.

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

[5] D. M. Blei. Probabilistic topic models. *Communications of the ACM*, 55:77–84, 2012.

[6] D. M. Blei and M. I. Jordan. Modeling annotated data. In *ACM SIGIR*, 2003.

[7] 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 TPAMI*, 36:521–535, 2014.

[8] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009.

[9] F. Feng, X. Wang, and R. Li. Cross-modal retrieval with correspondence autoencoder. In *ACM MM*, 2014.

[10] A. Habibian, T. Mensink, and C. G. Snoek. Videostory: A new multimedia embedding for few-example recognition and translation of events. In *ACM MM*, 2014.

[11] A. Habibian and C. Snoek. Recommendations for recognizing video events by concept vocabularies. *CVIU*, 124, 2014.

[12] A. Habibian and C. G. Snoek. Video2sentence and vice versa. In *ACM MM*, 2013.

[13] D. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16:2639–2664, 2004.

[14] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. In *ACM SIGIR*, 2003.

[15] T. Joachims. Optimizing search engines using clickthrough data. In *ACM SIGKDD*, 2002.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[17] D. Li, N. Dimitrova, M. Li, and I. K. Sethi. Multimedia content processing through cross-modal association. In *ACM MM*, 2003.

[18] A. K. McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[19] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *ICML*, 2011.

[20] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 2014.

[21] P. Over, J. Fiscus, G. Sanders, et al. TRECVID 2013-an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *TRECVID*, 2013.

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

[23] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *IJCV*, 105, 2013.

[24] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *NIPS*, 2013.

[25] J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen. Inter-media hashing for large-scale retrieval from heterogeneous data sources. In *ACM SIGMOD*, 2013.

[26] Y. Wang, F. Wu, J. Song, X. Li, and Y. Zhuang. Multi-modal mutual topic reinforce modeling for cross-media retrieval. In *ACM MM*, 2014.

[27] F. Wu, X. Lu, Z. Zhang, S. Yan, Y. Rui, and Y. Zhuang. Cross-media semantic representation via bi-directional learning to rank. In *ACM MM*. ACM, 2013.

[28] J. Zhou, G. Ding, and Y. Guo. Latent semantic sparse hashing for cross-modal similarity search. In *ACM SIGIR*, 2014.