# Image2song: Song Retrieval via Bridging Image Content and Lyric Words

Xuelong Li[*],   Di Hu[†],   Xiaoqiang Lu[*]

[*]Xi'an Institute of Optics and Precision Mechanics,
Chinese Academy of Sciences, Xi'an 710119, P. R. China

[†]School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL),
Northwestern Polytechnical University, Xi'an 710072, P. R. China

`xuelong_li@opt.ac.cn, hdui831@mail.nwpu.edu.cn, luxiaoqiang@opt.ac.cn`

## Abstract

*Image is usually taken for expressing some kinds of e-motions or purposes, such as love, celebrating Christmas. There is another better way that combines the image and relevant song to amplify the expression, which has drawn much attention in the social network recently. Hence, the automatic selection of songs should be expected. In this paper, we propose to retrieve semantic relevant songs just by an image query, which is named as the image2song problem. Motivated by the requirements of establishing correlation in semantic/content, we build a semantic-based song retrieval framework, which learns the correlation between image content and lyric words. This model uses a convolutional neural network to generate rich tags from image regions, a recurrent neural network to model lyric, and then establishes correlation via a multi-layer perceptron. To reduce the content gap between image and lyric, we propose to make the lyric modeling focus on the main image content via a tag attention. We collect a dataset from the social-sharing multimodal data to study the proposed problem, which consists of (image, music clip, lyric) triplets. We demonstrate that our proposed model shows noticeable results in the image2song retrieval task and provides suitable songs. Besides, the song2image task is also performed.*

## 1. Introduction

Images are usually taken for the purpose of memorizing, which could contain some specific contents and convey some kinds of emotions. For example, when celebrating the Christmas day, the captured pictures commonly contain dressed up people and Christmas trees covered with gifts, which is used to remember the happiness time. However, images appear to exist in only visual modality, which could be weak in expressing the above purpose. Inspired by visual music and musical vision [1], we consider that



object: tree, flower, ball, dress, stock…
attribute: smiling, green, happy, red…

Figure 1. A Christmas-relevant image and coupled song (lyric). There are several words (in different color) appearing in both image tags and lyric words. Best viewed in color.

the stimulates come from different senses (e.g. vision, auditory, tactile) may share similar performance. Hence, if the captured image is combined with a relevant song that expresses similar purpose, the expression will be obviously enhanced, which results from the more useful merged information from multimodal data [39, 22]. For example, showing the Christmas image while playing the song *Jingle Bells* is easier to touch viewers than single image. Hence, this kind of combination has attracted much attention in nowadays, which is simpler than video but richer than photo. But existing approaches of song selection for a given image are almost based on manual manner. Such methods often cost users a lot of time to decide but could still suffer from the small song library of users and lack of song comprehension. Hence, the technique of automatic image-based song recommendation should be expected, which is named as image2song in this paper.

*Music Information Retrieval* (MIR) is a traditional research field, which focuses on indexing proper music according to specific criteria. In this paper, the proposed image2song retrieval task aims to find the semantic-related songs for images, which therefore relies on the specific image content analysis. Hence, this task belongs to the

semantic/tag-based music retrieval category [31]. Such retrieval tasks utilize multiple textual data sources as the music modality for meeting the semantic requirements, such as music blog in *SerachSounds* [4, 37] and web pages in *Gedoodle* [19]. In this paper, we focus on retrieving songs (not instrumental music) for images, which contain sufficient textual data in lyric. More importantly, these textual data contains multiple common words in image tags (as shown in Fig. 1), which are considered as the related content across the modalities. Hence, lyric is adopted as the textual data source for retrieval. However, there still remains another two problems. First, image and lyric are different, where the former is non-temporal and the latter is temporal-based. More importantly, lyrics are not specifically created for images, which results in the content description gap between them [31]. Second, there is barely dataset providing corresponding images and songs, which makes it difficult to learn the correlation via a data-driven fashion.

In this paper, to overcome the above problems, our contributions are threefold:

- We leverage the lyric as the textual data modality for semantic-based song retrieval task, which provides an effective way to establish the correlation between image and song in semantic.
- We develop a multimodal model based on neural network, which learns the latent correspondence by embedding the image and lyric representations into a common space. To reduce the content gap, we introduce a tag attention approach that makes the bidirectional *Recurrent Neural Network* (RNN) of lyric focus on the main content of image regions, especially for the related lyric words.
- We build a dataset that consists of (image, music clip, lyric) triplets, which are collected from the social-sharing multimodal data on the Internet[1].

Experimental results verify that our model can provide noticeable retrieval results. In addition, we also perform the song2image task and our model has an improvement over state-of-the-art method in this task.

## 2. Related Work

**Image description.** An explicit and sufficient description of image is necessary for establishing the correlation with lyric in content. There have been many attempts over the years to provide detailed and high level description of images effectively and efficiently. Sivic and Zisserman [35] represented the image by integrating the low-level visual features into bag-of-visual-words, which has been widely applied in the scene classification [21] and object recognition [34]. Recently, in view of the advantages of *Convolutional Neural Network* (CNN) in producing high-level

semantic representation, many approaches based on it have shown great success in image understanding and description [30, 24]. However, these methods just focus on describing the image in specific fixed labels, while our work aims at providing a richer description of images, which points out the detailed image contents. Wu *et al.* [45] proposed an attribute predictor with the same purpose, which viewed the prediction as a multi-label regression task but without focusing on the image contents in specific regions.

**Lyric modeling.** A number of approaches have been proposed to extract the semantic information of lyrics. Most of these works viewed lyric as a kind of text, therefore *Bag-of-Words* (BoW) approach was usually used to describe the lyrics [5, 13, 16], which accounted for the frequency of word across the corpus, *e.g. Term Frequency-Inverse Document Frequency* (TF-IDF) [43]. However, almost all these works aimed at emotion recognition [16] or sentiment classification [46]. Recently, Schwarz *et al.* [32] proposed to embed the lyric words into vector space and extract relevant semantic representation. But, this work just focused on sentences rather than the whole lyric. What is more, Schwarz *et al.* [32] just performed pooling operation over the sentence words, which ignored to model the contextual information of lyric. Compared with the previous works, our work focuses on the lyric content analysis for providing sufficient information to learn the correlation with image, which takes consideration of both the semantic and context information of words.

**Multimodal learning across image and text.** Several works have studied the problem of annotating image with text. Srivastava and Salakhutdinov [38] proposed to use *Multimodal Deep Boltzmann Machine* (MDBM) to jointly model the images and corresponding tags, which could be used to infer the missing text (tags) from image queries, or inverse. Sohn *et al.* [36] and Hu *et al.* [12] extended such framework to explore and enhance the similarity across image and text. Recently, amounts of works focused on using natural language to annotate image instead of tags, which is usually named as image caption. These works [17, 18, 25] commonly utilized deep CNN and RNN to encode the image and corresponding sentences into a common embedding space, respectively. The aforementioned works aimed to generate relevant tags or sentences to describe the content of images. In contrast, our work focuses on learning to maximize the correlation between image and lyric, where the lyric is not specially generated for describing the concrete image content but a text containing several related words.

**Music visualization.** Music visualization has an opposite goal of our work, which aims at generating imagery to depict the music characteristic. Yoshii and Goto [49] proposed to transform musical pieces into a color and gradation visual effects, *i.e.*, a visual thumbnail image. Differently, Miyazaki and Matsuda [27] utilized a moving icon to visualize the

---

[1]The dataset is available at https://dtaoo.github.io/

music. However, these methods are weak in conveying semantic information in music, such as content and emotion. A possible way to deal with such defects is to learn the semantic from lyrics and establish correlation with real images. Cai *et al.* [3] manually extracted the salient-words from lyrics, *e.g.* location and non phrases, then took them as the key words to retrieve images. Similar frameworks could be found in [44, 48], while Schwarz *et al.* [32] focused on the sentences of lyrics and organized the retrieved images of each sentence into a video. In addition, Chu *et al.* [6] attempted to sing the generated description of an image for the first time, but it was too limited for various music style and lack of natural melody. In contrast to these methods which are based on rough tag description [3, 44, 48, 32] and direct similarity measure between features of different modalities [32], we propose a multimodal learning framework to jointly learn the correlation between lyric and image in a straightforward way while analyzing the specific content for both of them.

## 3. The Shuttersong Dataset

In order to explore the correlation between image and song (lyric), we collect a dataset that contains amounts of pairwise images and songs from the Shuttersong application. Shuttersong is a social sharing software, just like Instagram[2]. However, the shared content contains not only an image, but also a corresponding song clip selected by users, which is for strengthening the expression purpose. A relevant mood can also be appended by users. We collect almost the entire updated data from Shuttersong, which consists of 36,646 pairs of images and song clips. Some optional mood and favorite count information are also included.

In this paper, the lyric is viewed as a bridge connecting image and song, but it is not contained in the collected data from Shuttersong. To acquire the lyrics, we develop a software to automatically search and download them from the Internet based on song title and artist. However, there exist some abnormal ones in the collected lyrics, such as non-English songs, undetected ones, etc. Hence, we ask twenty participants to refine the lyrics. Specifically, the participants first exclude the non-English song, then judge whether the lyric matches the song clip. For the incorrect or undetected ones, the participants manually search the lyrics via the Internet. Then, the refined lyrics update the original ones in the dataset. A detailed explanation of the collected data can be found in the supplementary material.

**Statistics and analysis.** The full dataset consists of 16,973 available triplets after excluding the abnormal ones, where each triplet contains corresponding music clip[3], lyric, and image. As shown in Table 1, there are totally 6,373 differ-

---

2www.shuttersong.com, www.instagram.com

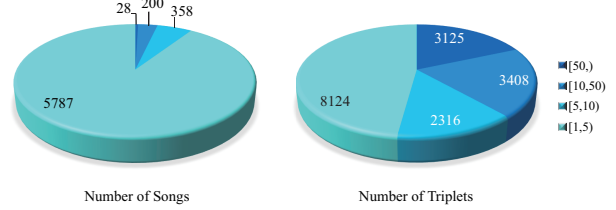[3]Due to the relevant legal, it is difficult to obtain the complete audio.



Figure 2. The statistics of the frequency of song occurrence. For example, there are 5,787 songs appearing less than 5 times, which results in 8,124 triplets.

| Triplet | Song | Available Mood | Favorite Count |
|---------|------|----------------|----------------|
| 16,973 | 6,373 | 3,158 | [1, 8964] |

Table 1. Aggregated statistics of the Shuttersong dataset.

ent songs among these triplets, but only 3,158 (18.6%) ones have available moods. The favorite counts of the shared image-song pairs created by users vary from 1 to 8,964, which could be used as a reference for estimating the quality of the pairs.

We also perform statistical analysis about the frequency of song occurrence, as shown in Fig. 2. Although there are 6,373 different songs in the dataset, 586 songs that appear
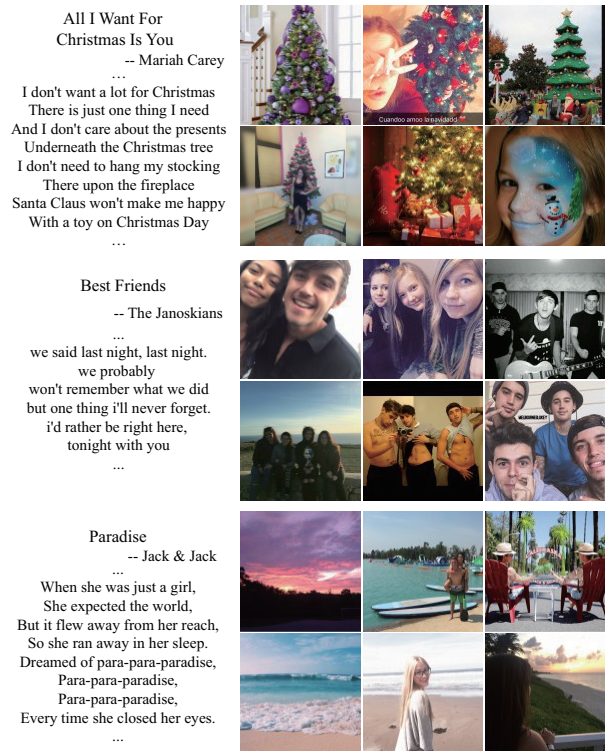


Figure 3. Examples of songs and corresponding images in the dataset. One song could relate to multiple images. It is easy to find out the images belonging to the same song have similar content that expresses the song to some extent.

at least 5 times take up more than half the triplets, which means each of these songs relates to at least 5 images. For example, Fig. 3 shows some examples of image-song pairs, where these songs appear more than 5 times in the built dataset. It is obvious that the images belong to the song *All I Want For Christmas Is You* are Christmas relevant, where trees, ribbons, and lights commonly appear among them. Meanwhile, these objects or attributes are related to some words of the corresponding lyric to some extent and provide a similar expression with the song. We can also find the similar situations in the other two groups, as shown in Fig. 3. These relevant images of the same song could provide an efficient way to explore the valuable information in lyrics and establish correlation with songs. Therefore, we conduct our experiments based on these songs with high occurrence frequency.

## 4. Our Model

In this paper, our proposed model aims at learning the correlation between images and lyrics, which can be used for song retrieval by image query, and vice versa. We first introduce a CNN-based model for fully representing image with amounts of tags, then encode the lyric sequence into a vector representation via a bi-directional LSTM model. A proposed multimodal model finally embeds the encoded lyric representation into the image tag space to correlate them under tag attention.

### 4.1. Image representation

We aim to generate a rich description of image, which could point out the specific content in certain text, *i.e.*, image tags. A common strategy is to detect objects and their attributes via extracting proposal regions and feeding them into specific classifier [20]. Hence, we propose and classify the regions of each image via a common *Region CNN* (RCNN) [29]. However, different from most of the previous works that just make use of the CNN features in the top K regions [15, 47], we use the whole prediction results to
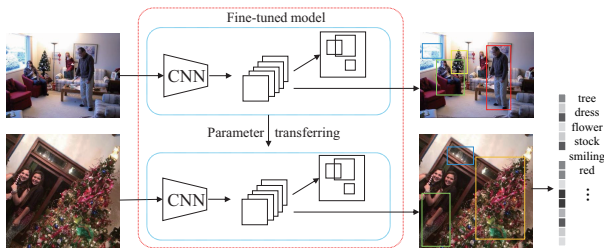


Figure 4. Overview of the image tag prediction network. The model is developed based on faster-rcnn, which is firstly fine-tuned on the Scene Graph Dataset [14], then used to generate the tag prediction result for images in the Shuttersong dataset via parameter transferring.

provide a richer description. We adopt the powerful VGG net [33] as the CNN architecture. As shown in Fig. 4, the CNN is first initialized from ImageNet, then fine-tuned on the Real-World Scene Graphs Dataset [14] that provides more object and attribute classes compared with COCO 2014 [23] and Pascal VOC [7], where 266 object classes and 249 attribute types[4] are employed for fine-tuning, respectively. The images in the Shuttersong dataset are then fed into the network. The top prediction probabilities of each class constitute the final image representation as a 515-dimensional vector $v$.

### 4.2. Lyric representation

Considering the song lyric is a kind of text containing dozens of words, we expect to generate a sufficient and efficient representation to establish inner-modalities correlation with the image representation. RNN-like architectures have recently shown advantages in encoding sequence while containing enough information for a range of natural language processing tasks, such as language modeling [40] and translation [8]. In our work, in view of the remarkable ability of LSTM [11] in encoding long sequence, we employ it to embed the lyric into a vector representation but with minor modification [9].

Words of lyric encoded in one-hot representations are first embedded into a continuous vector space, where the nearby points share similar semantic,

$$\mathrm{x}_t = \mathrm{W}_e \mathrm{l}_t, \tag{1}$$

where $\mathrm{l}_t$ is the $t$-th word in the song lyric, and the embedding matrix $\mathrm{W}_e$ is pre-trained based on the part of Google News dataset (about 100 billion words) [26]. The weight $\mathrm{W}_e$ is kept during the training due to overfitting concerns. Then the word vectors constitute the lyric matrix representation and are fed into the LSTM network,

$$
\begin{aligned}
i_t &= \sigma\left(\mathrm{W}_i \mathrm{x}_t + \mathrm{U}_i \mathrm{h}_{-1} + \mathrm{b}_i\right) & (2) \\
f_t &= \sigma\left(\mathrm{W}_f \mathrm{x}_t + \mathrm{U}_f \mathrm{h}_{-1} + \mathrm{b}_f\right) & (3) \\
\tilde{C}_t &= tanh\left(\mathrm{W}_c \mathrm{x}_t + \mathrm{U}_c \mathrm{h}_{-1} + \mathrm{b}_c\right) & (4) \\
C_t &= i_t * \tilde{C}_t + f_t * C_{t-1} & (5) \\
o_t &= \sigma\left(\mathrm{W}_o \mathrm{x}_t + \mathrm{U}_o \mathrm{h}_{-1} + \mathrm{b}_o\right) & (6) \\
\mathrm{h}_t &= o_t * tanh\left(C_t\right). & (7)
\end{aligned}
$$

Three gates (input $i$, forget $f$, and output $o$) and one cell memory $C$ constitute a LSTM cell and work in cooperation to determine whether remembering or not. $\sigma$ is the sigmoid function. And the network parameters $\mathrm{W}_*$, $\mathrm{U}_*$, and $\mathrm{b}_*$ will be learned during the training procedure.

Considering the relevant images of a given lyric have high variance in the content, such as the examples in Fig. 3,

---

[4]Different from the settings in [14], we select the attributes appear at least 30 times to provide a more detailed description.
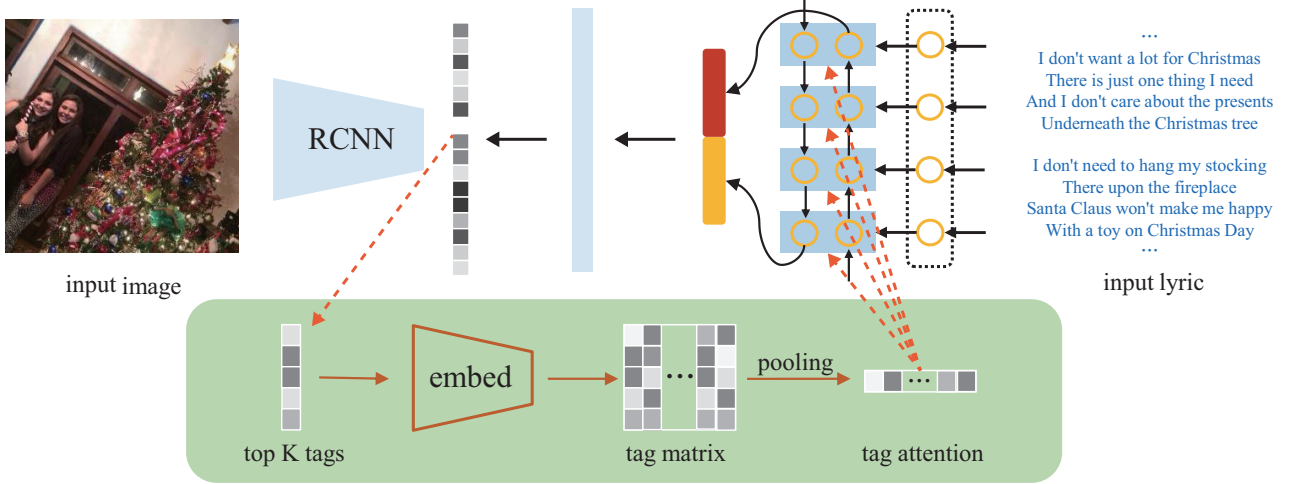
Figure 5. The diagram of the proposed model. The image content tags are first predicted via the R-CNN, meanwhile, a bi-directional LSTM is utilized to model the corresponding lyric. Then the generated lyric representation is mapped into the space of the image tags via a MLP. To reduce the content gap between image and lyric, the top K image tags are embedded into a tag matrix and represented as tag attention for the lyric modeling by performing max/average pooling.

it is difficult to directly correlate the image tags with each embedded lyric word, especially for the longer lyric compared with normal sentence [10, 41]. Hence, we just take the final output of the LSTM, which remains the context of single word and provides efficient information of the whole lyric. And experiments (Sec. 5.2) indicate its effectiveness in establishing the correlation with images. Besides, both the forward and backward LSTM are employed to simultaneously model the history and future information of the lyric, and the final lyric representation can be denoted as $l = \vec{h}_{final} \parallel \overleftarrow{h}_1$, where $\parallel$ indicates concatenation.

### 4.3. Multimodal learning across image and lyric

The image and lyric representation have been extracted via respective model, it is intuitive to embed them into a common space to establish their correlation via a *Multi-Layer Perceptron* (MLP) model, as shown in the upper part of Fig. 5. The effectiveness of such model has been verified in the image-text retrieval [28], this is because the text information is specially written to describe the image content, hence there are several common features representing the same content across both modalities. However, the MLP model is not entirely suitable for our task. As lyric is not specially generated for images, it can be found that there is few lyric words directly used to describe the corresponding image but multiple related words sharing the similar meaning with image content, which could result in a content description gap. For example, we can find such situations in the *Paradise* group in Fig. 3. There exists few specific lyric words could be used for describing the *beach*, *wave*, or *sky* in the images, but the words *paradise*, *flew*, and *girl* are related to some image contents.

To address the aforementioned problem, we propose to use the tag attention to make the lyric words focus on the image content, as shown in Fig. 5. We first sort the image prediction results of all the tag classes, then choose the top-K tags that are assumed as the correct prediction of corresponding image content[5]. To share the representation space with lyric words, the selected tags are also embedded into a vector of continuous value via Eq. 1, which results in a tag matrix $T \in R^{K \times 300}$. Then we perform the pooling operation over the matrix $T$ into a tag attention vector $\tilde{v}$. Inspired by the work of Hermann *et al.* [10] and Tan *et al.* [41], the tag attention is designed to modify the output vector of the LSTM and make the lyric words focus on the concrete image content. To be specific, the output vector $h_t$ at time step $t$ is updated as follows,

$$m_t = \sigma\left(W_{hm}h_t + W_{vm}\tilde{v}\right) \quad (8)$$
$$s_t \propto \exp\left(w_{ms}^T m_t\right) \quad (9)$$
$$\tilde{h}_t = h_t s_t, \quad (10)$$

where the weights $W_{hm}$, $W_{vm}$, and $w_{ms}$ are considered as the attention degree of the lyric words given the image content (tags). During modeling the lyrics, the related words are paid more attention via the attention weights, which acts like the TF-IDF in document retrieval based on key words. However, different from the pooling operation over the entire sequence in the previous works [10, 41], the output vector with attention, $\tilde{h}_t$, just flow through the entire lyric words, which results in a refined lyric representation.

During training, we aims at minimizing the difference between the image and lyric pair in the tag space, which is

---

[5]The top 5 tags are validated and employed in this paper.

essentially the *Mean Squared Error* (MSE),

$$l_{mse} = \sum_{i=1}^{T} \left\| v_i - \tilde{l}_i \right\|_2^2, \qquad (11)$$

where $v_i$ and $\tilde{l}_i$ are the generated image and projected lyric representation, and $T$ is the number of training pairs. Except the MSE loss, we also employ the cosine proximity and marginal ranking loss. The relevant experiment results are reported in the materials. For the retrieval, both the query and retrieved items are fed into the proposed model, then the cosine similarity is computed as the relevance.

## 4.4. Optimization

The proposed model is optimized by employing stochastic gradient descent with RMSprop [42]. The algorithm adaptively rescales the step size for updating trainable weights according to the corresponding gradient history, which achieves the best result when faced with the word frequency disparity in lyric modeling. The parameters are empirically set as: the learning rate $l = 0.001$, the weight decay $\rho = 0.9$, the tiny constant $\varepsilon = 10^{-8}$, and The model is trained with mini-batches of 100 image-lyric pairs.

## 5. Experiments

### 5.1. Data preprocessing and setup

**Dataset.** In this paper, we choose the triplets whose lyrics appear at least 5 times, which results in 8,849 triplets (586 songs). Such operation is a common preprocessing method and also better for learning the correlation across modalities. To reduce the influence of the imbalanced number of images, we choose five triplets with top favorite counts for each song, which are considered to have more reliable correlation. Within these filtered triplets, 100 songs and corresponding images are randomly selected for testing, and the rest for training, which forms dataset†. Note that, we also employ another kind of train/test partition to constitute dataset§: we randomly select one from five images of each song for testing, and the rest for training. In the dataset§, the train and test set share the same 586 songs but with different images, which is developed to exactly evaluate the models in shrinking the content gap, when faced with variable image content and lack of related lyric words.

**Preprocessing.** For the lyric data preprocessing, we remove all the non-alphanumeric and stop words.

**Metric.** We employ the rank-based evaluation metrics. R@K is Recall@K that computes the percentage of a correct result found in the top-K retrieved items (higher is better), and Med r is the median rank of the closest correct retrieved item (lower is better).

**Baselines.** In our experiments, we compare with the following models: **Bag-of-Words** [2]: The image features

and lyric BoW representation are mapped into a common subspace. **CONSE** [28]: The lyric representation is obtained by performing pooling operation over all words and then established correlation with image features. **Attentive-Reader** [10]: This method is mainly developed for Question-Answering task, which performs a weighted pooling over the encoded words via LSTM with question attention. Here, the question attention is replaced with the image tag attention, and a non-linear combination is used to measure the correlation.

**Our Models.** We evaluate two variants of our models: **Our** baseline model: Our proposed model except the tag attention, as shown in the upper part of Fig. 5. **Our-attention** model: The proposed complete model. Note that, average pooling is employed for obtaining tag attention, which could remain more image content information compared with max-manner.

### 5.2. Image2song retrieval

This experiment aims to evaluate the song retrieval performance by given image query. There are two kinds of tags used for representing images, *i.e.*, object and attribute. It is expected to explore which kind of them influences the image description most and provides more valuable information to establish the correlation across the two modalities. In view of this, we group the image tags into three categories, *i.e.*, object, attribute, and both of them. Table. 2 shows the comparison results on both datasets, where both the model variants and other methods are considered. And we also provide example retrieval results in Fig. 6.

For each tag category, there are three points we should pay attention to. First, the bi-directional LSTM model provides better lyric representations than direct pooling (CONSE [28]) and BoW [2], as the LSTM takes consideration of both word semantic and context information. Second, the proposed complete model shows the best result in most conditions. When we employ the tag attention for lyric modeling, more related words in the lyrics will be emphasized, which shrinks the content gap and improves the relevant performance, especially on the dataset§. Although Attentive-Reader [10] also employ similar manner, it takes the attentive pooling result as the lyric representation. The direct pooling operation in Attentive-Reader will make it suffer from the difficulty in establishing correlation, this is because the variable image contents change the attention weight. While our model takes the final output of LSTM, where the attention weight is not immediately utilized but conveyed by the updated output vector $\tilde{h}_t$. Third, although our methods show the best performance on both datasets, overall performance is not excellent as expected. But this is a common situation in the semantic-based music retrieval [31]. This is because the song retrieval based on textual data has to estimate the semantic labels from

| Image tags | obj-tags | | | | attr-tags | | | | obj-attr-tags | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dataset† | R@1 | R@5 | R@10 | Med r | R@1 | R@5 | R@10 | Med r | R@1 | R@5 | R@10 | Med r |
| BoW [2] | 1.4 | 7.6 | 13.4 | 45.06 | 1.6 | 7.0 | 13.0 | 46.11 | 1.4 | 7.2 | 13.2 | 45.01 |
| CONSE [28] | 1.8 | 8.0 | 13.0 | 46.11 | 1.8 | 6.6 | 12.2 | 47.38 | 2.0 | 6.8 | 14.8 | 46.13 |
| Attentive-Reader [10] | 1.8 | 7.4 | 14.0 | 44.45 | 1.8 | 7.6 | 13.0 | 46.97 | 1.8 | 7.2 | 15.0 | 43.78 |
| Our | **2.2** | 7.2 | 14.2 | 43.13 | 1.6 | 7.8 | 14.0 | 46.32 | 2.2 | 7.6 | 14.8 | 43.30 |
| Our-attention | 2.0 | **9.2** | **17.6** | **41.36** | **2.2** | **8.4** | **14.4** | **45.21** | **2.6** | **9.4** | **16.8** | **41.50** |
| Dataset§ | R@10 | R@50 | R@100 | Med r | R@10 | R@50 | R@100 | Med r | R@10 | R@50 | R@100 | Med r |
| BoW [2] | 4.27 | 13.48 | 24.23 | 251.77 | 3.07 | 12.12 | 23.55 | 260.17 | 4.27 | 14.33 | 23.72 | 250.63 |
| CONSE [28] | 3.41 | 14.33 | 25.27 | 248.88 | 3.24 | 12.97 | 23.38 | 262.17 | 3.58 | 14.51 | 25.09 | 245.32 |
| Attentive-Reader [10] | 4.10 | 14.51 | 24.57 | 254.53 | 3.75 | 13.14 | 23.04 | 267.12 | 3.92 | 14.51 | 25.43 | 243.53 |
| Our | **4.61** | 15.02 | 26.28 | 246.51 | 4.27 | 13.65 | 23.89 | 254.17 | 4.94 | 15.36 | 28.33 | 240.00 |
| Our-attention | 4.47 | **15.36** | **26.86** | **245.61** | **4.61** | **13.99** | **25.76** | **249.42** | **5.63** | **17.58** | **29.35** | **233.82** |

Table 2. Image2song retrieval experiment result in R@K and Med r on dataset† and dataset§. Three kinds of image representation are considered, *e.g.*, object (obj), attribute (attr), and both them (obj-attr).
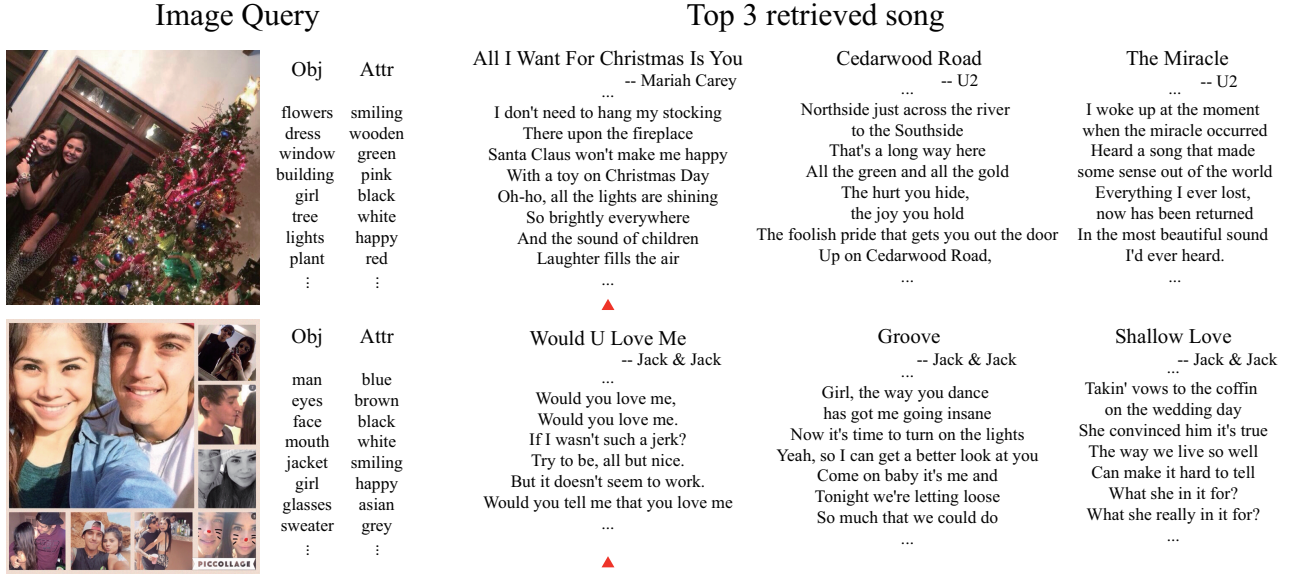


Figure 6. Image2song retrieval examples generated by our model. The generated object and attribute tags are shown next to each image query, and the songs with red triangle are the ground truth in the dataset.

lyric, which is charactered as a low specificity and long-term granularity [31]. Even so, our proposed models still enjoy a relatively big improvement, and the retrieved examples show the effectiveness of the models in Fig. 6.

Across the three groups of image tags, we could find that the attribute tags almost always perform worse than object ones. We consider there are mainly three reasons to explain this phenomenon. First, the object words are usually employed for image description, which is more powerful in identifying the image content compared with attribute ones. Second, as the Shuttersong is actually a kind of sharing application, most of the updated images are self-photography, which makes it difficult to correlate images with songs, especially for the attribute tags. Third, most of the images

share similar attribute tags with high prediction scores (*e.g.*, *black*, *blue*, *white*). This is actually a long-tailed distribution[6] and therefore it becomes difficult to establish the correlation between image and specific lyric. However, the group with both tags nearly shows the best performance across different models, which results from more detailed description for images.

Apart from the influence of image tag property, the lyric also impacts the performance significantly. We show the specific results of 28 songs with more 50 times occurrence in Fig. 7 and Fig. 8. As shown, some song lyrics are of remarkable performance, while some fail to establish the correlation with relevant images. The potential reason is

---

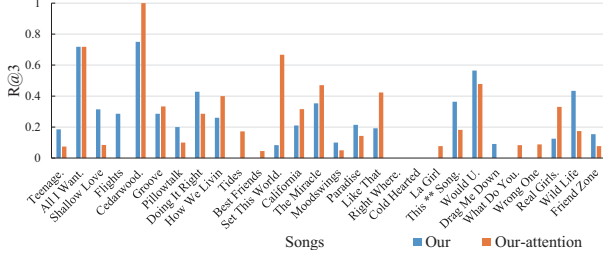[6]A detailed illustration can be found in the supplementary material.

Figure 7. Detailed comparison results among song examples in R@3. The complete model with attention improves the average performance, especially for the songs with **zero** score.
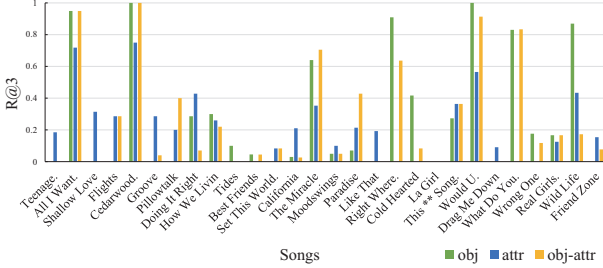


Figure 8. Detailed comparison results among the three groups of tags. All the experiments are conducted with the proposed baseline model and evaluated by R@3.

| Models | R@1 | R@5 | R@10 | R@20 | Med r |
|--------|-----|-----|------|------|-------|
| Schwarz [32] | 0.10 | 0.27 | 0.45 | 0.71 | 16 |
| Our | 0.19 | 0.34 | 0.52 | 0.74 | 15 |

Table 3. The image retrieval results given lyric query. Here, the image tags in the method of Schwarz *et al.* [32] are generated by the R-CNN approach [29] for a fair comparison.

that parts of the lyrics are weak in providing obvious or even related words. For example, the lyric words of *Best Friends* are about *forget*, *tonight*, *remember*, etc, which are not specifically related to the image content, as shown in Fig. 3. Even so, our proposed tag attention mechanism can still reduce the content gap and improve the performance, as show in Fig. 7. While for the song *Paradise* and *All I Want For Christmas Is You*, the lyric words and image content are closely related, hence this case achieves better results.

### 5.3. Song2image retrieval

In this experiment, we aims to retrieve relevant images for a given song (lyric) query, which is similar to the music visualization task. And we employ the proposed baseline model, which could perform more efficiently without the attention of each image. Table. 3 shows the comparison results on dataset†. First, our proposed model outperforms the method of Schwarz *et al.* [32]. In our model, the image content information are employed to supervise the lyric modeling during training, while Schwarz *et al.* [32] directly use the pooling results of lyric word vector for similarity
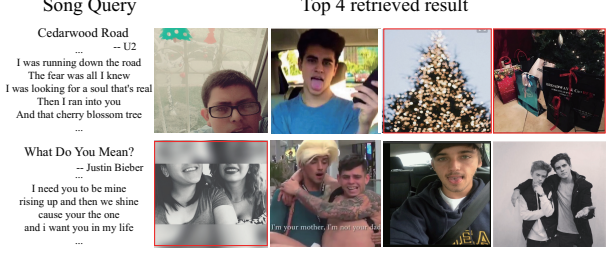


Figure 9. Example retrieval results by our model. The images in red bounding box are the corresponding ones in the Dataset.

retrieval without the inner-model interaction like ours. Second, although our model get better performance, it still suffers from the lack of related words in some songs, just like the image2song task.

In addition, we also show some examples of the song query, and the top 4 retrieved results are illustrated in Fig. 9. Although some retrieved results are not correct, they share similar content conveyed by the lyric. For example, in terms of the song *Cedarwood Road* about tree and Christmas, the top two retrieved images indeed have tree-relevant content and the left two are about Christmas.

## 6. Discussion

In this paper, we introduce a novel problem that retrieves semantic-related songs based on given images, which is named as the image2song task. We collect a dataset that consists of pairwise images and songs to study this problem. We propose a semantic-based song retrieval framework, which employs the lyric as the textual data source for estimating the semantic label of songs, then a deep neural network based multimodal framework is proposed to learn the correlation, where the lyric modeling is proposed to focus on the main image content to reduce the content gap between them. The experiment results show that our model can recommend suitable songs for a given image. In addition, the proposed approach can also retrieve relevant images according to a song query with a better performance than other methods.

There still remains a direction that should be explored in the future. Song is about several minutes long, which is too long for just showing an image. A possible way could be expected is to correlate the image with only parts of the corresponding song, which is more natural for expression. Furthermore, we also hope to perform other kinds of cross-modal retrieval task, which essentially attempts to establish the correlation among different senses of human.

## 7. Acknowledgement

# References

[1] X. Li, D. Tao, S. J. Maybank, and Y. Yuan. Visual music and musical vision. *Neurocomputing*, 71(10):2023–2028, 2008. 1

[2] B. Bai, J. Weston, D. Grangier, R. Collobert, K. Sadamasa, Y. Qi, C. Cortes, and M. Mohri. Polynomial semantic indexing. In *Advances in Neural Information Processing Systems*, pages 64–72, 2009. 6, 7

[3] R. Cai, L. Zhang, F. Jing, W. Lai, and W.-Y. Ma. Automated music video generation using web image resource. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 2, pages II–737. IEEE, 2007. 3

[4] Ò. Celma, P. Cano, and P. Herrera. Search sounds an audio crawler focused on weblogs. In *7th International Conference on Music Information Retrieval (ISMIR)*, 2006. 2

[5] R. Chen, Z. Xu, Z. Zhang, and F. Luo. Content based music emotion analysis and recognition. In *Proc. of 2006 International Workshop on Computer Music and Audio Technology*, volume 68275, 2006. 2

[6] H. Chu, R. Urtasun, and S. Fidler. Song from pi: A musically plausible network for pop music generation. *arXiv preprint arXiv:1611.03477*, 2016. 3

[7] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 4

[8] A. Graves. Neural networks. In *Supervised Sequence Labelling with Recurrent Neural Networks*, pages 15–35. Springer, 2012. 4

[9] A. Graves, A.-r. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE, 2013. 4

[10] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701, 2015. 5, 6, 7

[11] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 4

[12] D. Hu, X. Lu, and X. Li. Multimodal learning via exploring deep semantic similarity. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 342–346. ACM, 2016. 2

[13] X. Hu, J. S. Downie, and A. F. Ehmann. Lyric text mining in music mood classification. *American music*, 183(5,049):2–209, 2009. 2

[14] J. Johnson, R. Krishna, M. Stark, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei. Image retrieval using scene graphs. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3668–3678. IEEE, 2015. 4

[15] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015. 4

[16] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull. Music emotion recognition: A state of the art review. In *Proc. ISMIR*, pages 255–266. Citeseer, 2010. 2

[17] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Multimodal neural language models. In *ICML*, volume 14, pages 595–603, 2014. 2

[18] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 2

[19] P. Knees, T. Pohle, M. Schedl, and G. Widmer. A music search engine built upon audio-based and web-based similarity measures. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 447–454. ACM, 2007. 2

[20] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903, 2013. 4

[21] L.-J. Li and L. Fei-Fei. What, where and who? classifying events by scene and object recognition. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007. 2

[22] C. C. S. Liem, M. Larson, and A. Hanjalic. When music makes a scene. *IJMIR*, 2:15–30, 2012. 1

[23] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 4

[24] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *CoRR*, abs/1411.4038, 2015. 2

[25] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014. 2

[26] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013. 4

[27] R. Miyazaki and K. Matsuda. Dynamicicon: A visualizing technique for musical pieces in moving icons based on acoustic features. *Journal of Information Processing*, 51(5):1283–1293, 2010. 2

[28] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013. 5, 6, 7

[29] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 4, 8

[30] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 2

[31] M. Schedl, E. Gómez, J. Urbano, et al. Music information retrieval: Recent developments and applications. *Foundations and Trends® in Information Retrieval*, 8(2-3):127–261, 2014. 2, 6, 7

[32] K. Schwarz, T. L. Berg, and H. P. A. Lensch. Auto-illustrating poems and songs with style. In *Asian Conference on Computer Vision (ACCV)*, 2016. 2, 3, 8

[33] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[34] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering object categories in image collections. 2005. 2

[35] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1470–1477. IEEE, 2003. 2

[36] K. Sohn, W. Shang, and H. Lee. Improved multimodal deep learning with variation of information. In *Advances in Neural Information Processing Systems*, pages 2141–2149, 2014. 2

[37] M. Sordo, Ò. Celma, and C. Laurier. Querybag: Using different sources for querying large music collections. In *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, 2009. 2

[38] N. Srivastava and R. R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230, 2012. 2

[39] B. E. Stein and M. A. Meredith. *The merging of the senses.* The MIT Press, 1993. 1

[40] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014. 4

[41] M. Tan, B. Xiang, and B. Zhou. Lstm-based deep learning models for non-factoid answer selection. *arXiv preprint arXiv:1511.04108*, 2015. 5

[42] T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 4(2), 2012. 6

[43] M. Van Zaanen and P. Kanters. Automatic mood classification using tf* idf based on lyrics. In *ISMIR*, pages 75–80, 2010. 2

[44] Z.-K. Wang, R. Cai, L. Zhang, Y. Zheng, and J.-M. Li. Retrieving web images to enrich music representation. In *ICME*, 2007. 3

[45] Q. Wu, C. Shen, L. Liu, A. Dick, and A. v. d. Hengel. What value do explicit high level concepts have in vision to language problems? *arXiv preprint arXiv:1506.01144*, 2015. 2

[46] Y. Xia, L. Wang, and K.-F. Wong. Sentiment vector space model for lyric-based song sentiment classification. *International Journal of Computer Processing Of Languages*, 21(04):309–330, 2008. 2

[47] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2(3):5, 2015. 4

[48] S. Xu, T. Jin, and F. C.-M. Lau. Automatic generation of music slide show using personal photos. In *ISM*, 2008. 3

[49] K. Yoshii and M. Goto. Music thumbnailer: Visualizing musical pieces in thumbnail images based on acoustic features. In *ISMIR*, pages 211–216, 2008. 2