Representation Learning for Image-based Music Recommendation

Chih-Chun Hsia,* Kwei-Herng Lai,† Yian Chen,§ Chuan-Ju Wang,† Ming-Feng Tsai*
*National Chengchi University †Academia Sinica §KKBOX Inc.

arXiv Institution

Abstract

Image perception is one of the most direct ways to **provide contextual information** about a user concerning his/her surrounding environment; hence images are a suitable proxy for contextual recommendation. We propose **a novel representation learning framework for image-based music recommendation** that bridges the heterogeneity gap between music and image data; the proposed method is a key component for various contextual recommendation tasks. Preliminary experiments show that for an image-to-song retrieval task, the proposed method retrieves relevant or conceptually similar songs for input images.

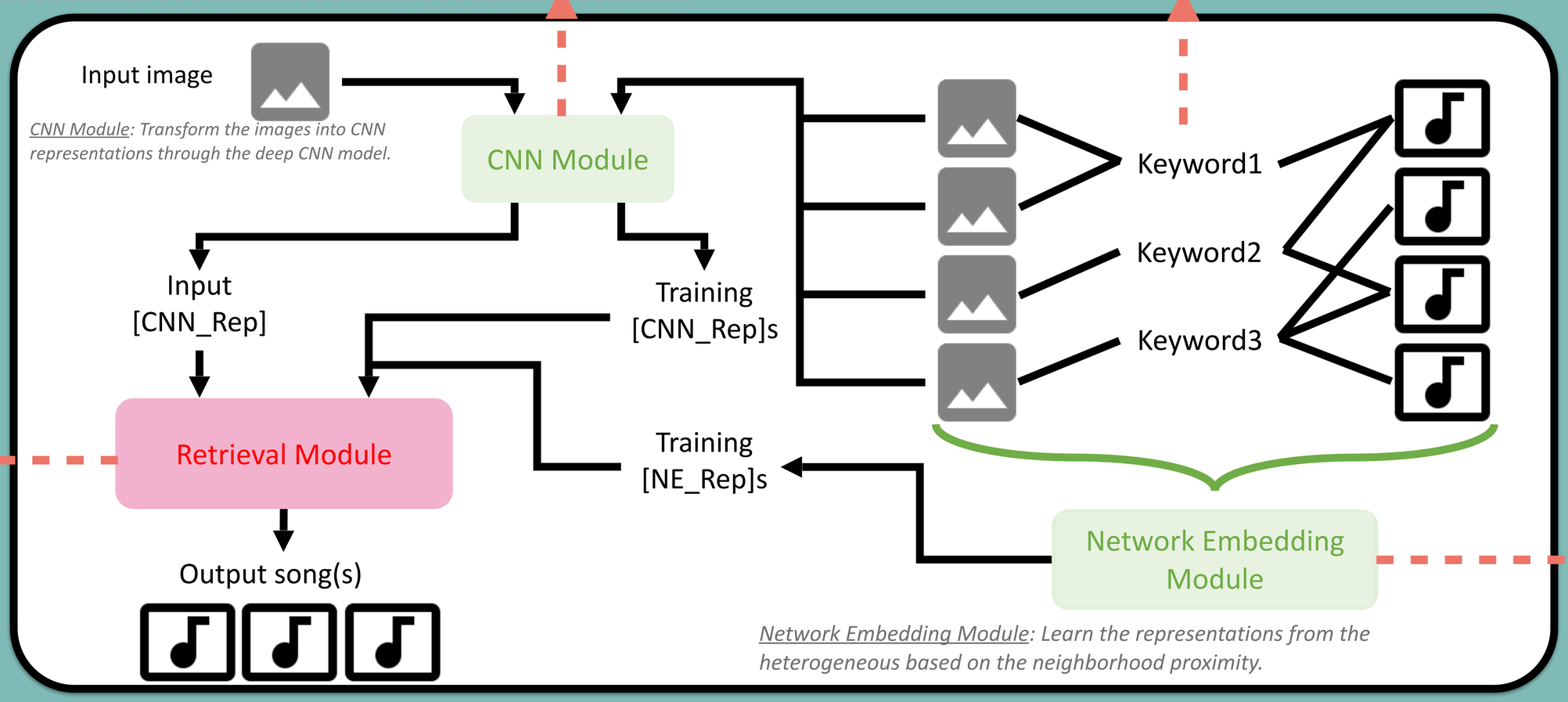
Image-Based Music Recommendation

Image Representation Learning

We apply the **VGG-19 pre-trained model**, to obtain the image representations. The network structure of VGG-19 includes 16 convolutional layers and 3 fully-connected layers, with the use of 3 × 3 lters. To generate the 4096-dimensional representation for each image, **we extract the representations from the second rather than the third fully-connected layer.**

Data Preparation

We obtained the music dataset from KKBOX; and crawled and segmented the lyrics by Jieba toolkit, **extracted 72 frequent keyword from the titles and song lyrics.** Moreover, we constructed our image dataset by using search engines to collect images given the 72 selected keywords, and also included in the experimental dataset that each song contains at least one of the keywords, **yielding a total of 62,316 songs, 72 keywords, and 33,459 images.**



Making a Recommendation

Three stages for recommendation:

- **Image Transformation:** Transforming the input image into a 4096-dimensional representation by CNN module.
- **Image Retrieval:** Retrieving the most relevant images by calculating the Euclidean distance between input image representation and pre-trained image representations.
- **Music Recommendation:** Recommending the most relevant songs for each relevant image based on the Euclidean distances between network representations of songs and the images.

Crossing Heterogeneity Gap

We connect the two types of multimedia data with corresponding keywords; and construct a heterogeneous network with two type of edges:

- **Song-Keyword:** connecting each song with the keywords in its lyrics; the weight indicates the relevance between the song and the keyword.
- **Image-Keyword:** connecting each image with its corresponding keyword; the weight indicates the relevance between the image and the keyword

The vertex representations are learned based on their neighborhood proximity using SGD with edge sampling and negative sampling.

Experiment

For each input image, we recommended the top 10 songs obtained from top-2 relevant songs for the top 5 relevant images. **As for ground truth, we collected the top-n conceptually similar words of extracted keywords from ConceptNet to construct the ground truth.**

Table 1: Image-to-song retrieval (hit rate@10)

n	Proposed model	KM	POP
5	**0.913**	0.902	0.124
10	**0.918**	0.917	0.157
50	**0.943**	0.941	0.356
100	**0.943**	0.943	0.378

User Study

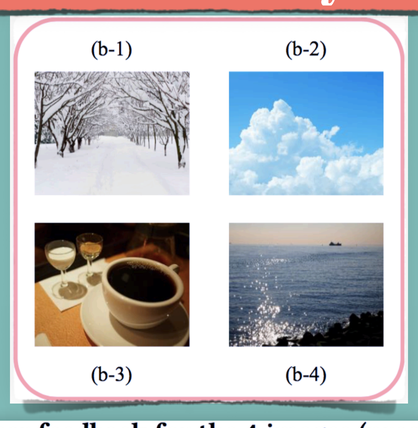


Table 2: User feedback for the 4 images (precision@10)

Theme	Snow forest (b-1)	Sky with clouds (b-2)	Coffee (b-3)	Ocean (b-4)
Proposed model	**0.776**	**0.623**	**0.655**	**0.709**
KM	0.531	0.569	0.546	0.531
POP	0.414	0.514	0.371	0.586

Case Study

Input Images

Relevant Images

Recommended Songs

花與蝶	雪在飛
(Flower and Butterfly)	(Dancing Snowflake)
花有情花有愛	胭脂雪
(Flower Affectionate Love)	(Rouge Snow)
你是我的花朵	今年沒聖誕
(You Are My flower)	(This year without Christmas)
花季未了	未來
(By the End of Flora Season)	(Future)