Enhancing Music Features by Knowledge Transfer from User-item Log data

Donmoon Lee[1,2,3], Jaejun Lee[1], Jeongsoo Park[2], Kyogu Lee[1,3]

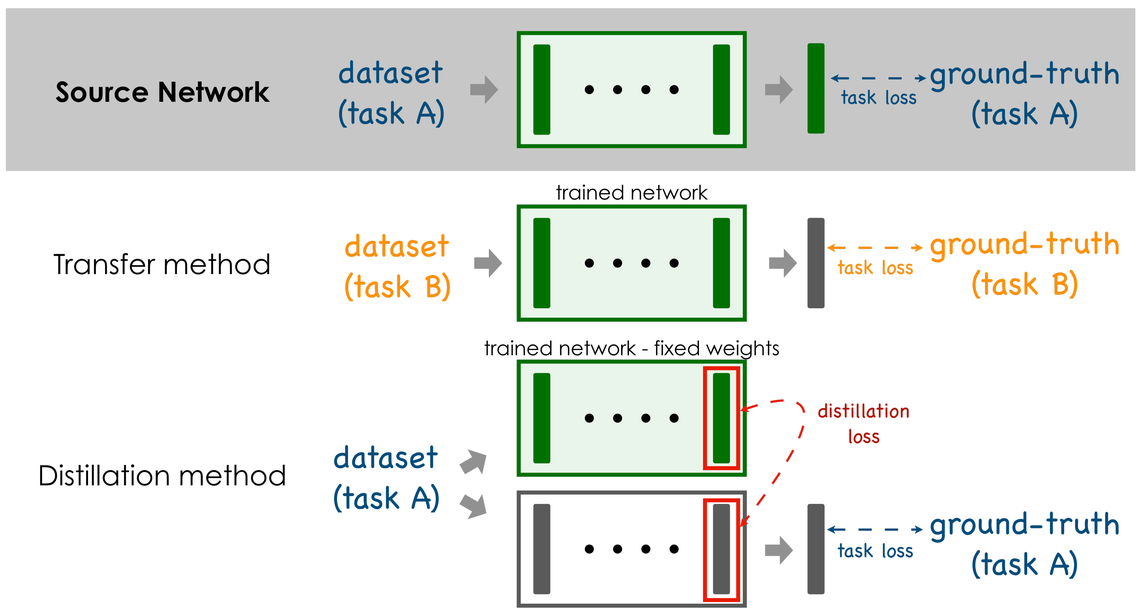
[1)] Music and Audio Research Group, Seoul National University, Korea
[2)] Cochelar.ai, Korea
[3)] Center for Super Intelligence, Seoul National University, Korea

lunideal@snu.ac.kr

Introduction

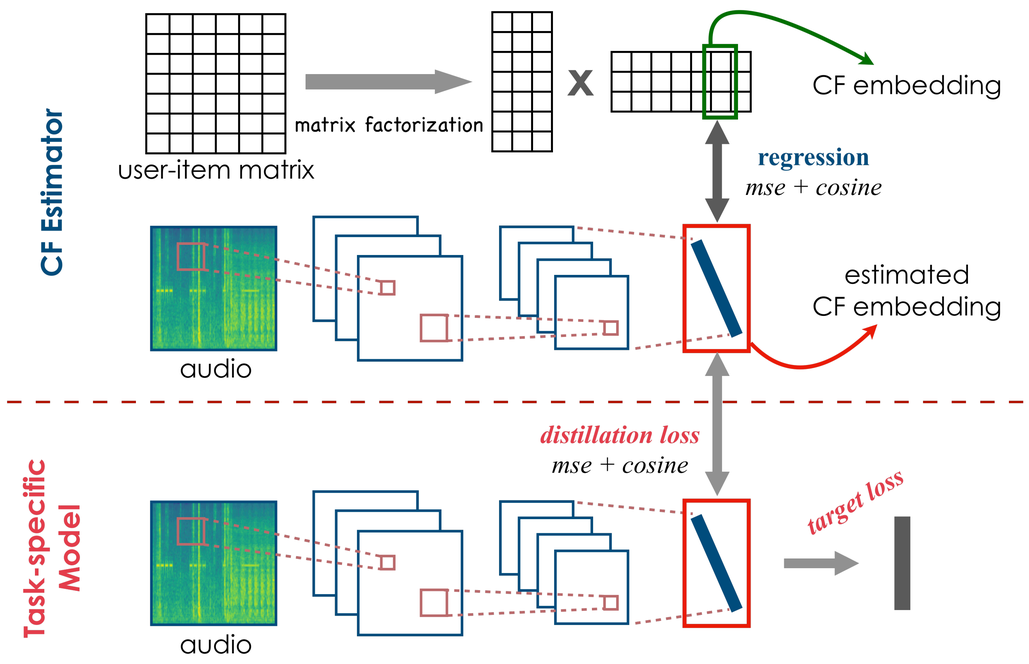
- There is not enough annotated data on music-related tasks, because the audio labeling is difficult and time-consuming.
- It can be a problem because the performance of a deep learning-based approach depends heavily on the amount of labeled data.
- We address this problem by using the training approach inspired by knowledge transfer along with the user listening log data that can be autonomously obtained and can be estimated from audio contents.

Knowledge Transfer



- Types of knowledge transfer
  - The *transfer method*: It directly uses the outputs of the learned networks or their intermediate activation as a feature for different tasks.
  - The *distillation method:* It utilizes a posterior distribution of the teacher network itself as the ground-truth to train other models and is often referred to as the *teacher-student method*.

Proposed Method

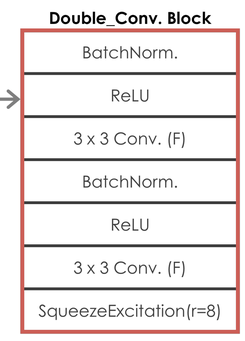


- The knowledge to transfer is trained from a large-scale user log dataset.
  - Raw user log is sparse and high-dimensional, so they are usually converted to a compressed representation called CF embedding.
- The proposed model constrains the embedding of the penultimate layer of task-specific models.
  - Task-specific models are trained to make their embedding similar to CF embedding during a given task.

Training CF Estimator

- Dataset
  - Music and user listening log data from Melon, which is streaming service in Korea. (5M songs and 2.5M users)
  - Each song is compressed into a 40-dimensional CF vector through the ALS algorithm.
  - A total of 244,975 songs are randomly chosen for the experiment and divided into training set (187,404 songs), validation set (20,831 songs), and test set (36,740 songs).

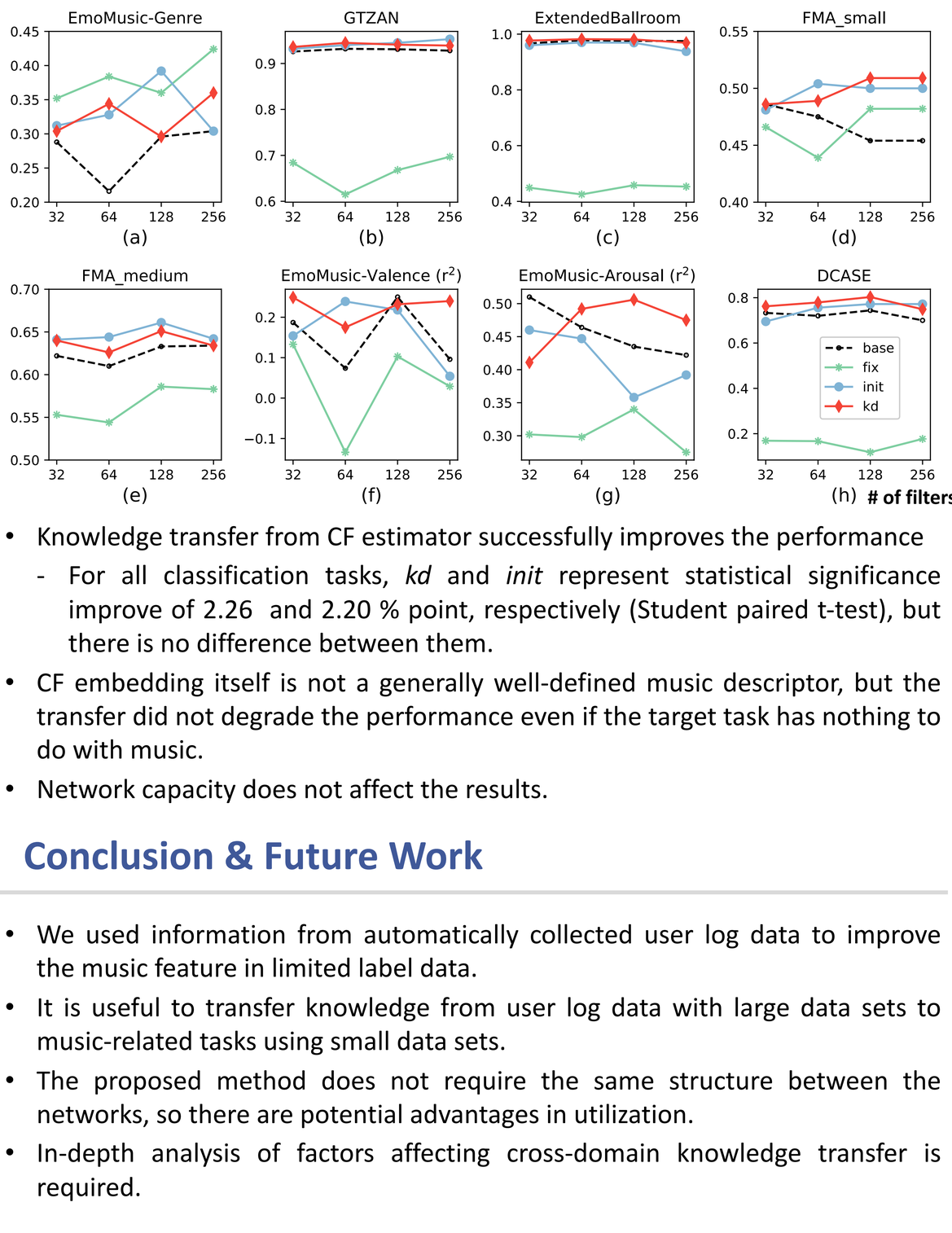
Layer	Output shape
Audio input	(1,480000)
Mel spectrogram	(96, 1280, 1)
Double_conv block	
4 x 5 Max-pooling	(24, 256, F)
Double_conv block	
3 x 4 Max-pooling	(8, 64, F)
Double_conv block	
2 x 4 Max-pooling	(4, 16, F)
Double_conv block	
2 x 4 Max-pooling	(2, 4, F)
Global average pooling	(F)
Fully-connected layer	(40)



Transfer to Other Tasks

- List of tasks.
  - Music genre classification: EmoMusic, GTZAN, ExtendedBallRoom, FMA
  - Music emotion regression task: EmoMusic
  - Acoustic scene classification task: DCASE2016 dataset (completely different domain from music)
- Training methods
  - *base:* training model from scratch
  - *fix:* only training one layer classifier from CF estimator
  - *init:* training model from CF estimator initialization
  - *kd:* distillation loss is added to the training procedure

Experimental Result



- Knowledge transfer from CF estimator successfully improves the performance
  - For all classification tasks, *kd* and *init* represent statistical significance improve of 2.26 and 2.20 % point, respectively (Student paired t-test), but there is no difference between them.
- CF embedding itself is not a generally well-defined music descriptor, but the transfer did not degrade the performance even if the target task has nothing to do with music.
- Network capacity does not affect the results.

Conclusion & Future Work

- We used information from automatically collected user log data to improve the music feature in limited label data.
- It is useful to transfer knowledge from user log data with large data sets to music-related tasks using small data sets.
- The proposed method does not require the same structure between the networks, so there are potential advantages in utilization.
- In-depth analysis of factors affecting cross-domain knowledge transfer is required.