DEEP LEARNING AND MULTIMODAL MODELS FOR MUSIC INFORMATION RETRIEVAL

Ilaria Manco, Emmanouil Benetos, George Fazekas
Centre for Digital Music, Queen Mary University of London

AI + MUSIC

Overview

Research on music-related tasks focusses on techniques to analyse audio content. However, music is experienced in a multimodal way and information about music is often conveyed through non-audio modalities (**images**, **text**, **video**, **metadata**). These can be exploited to enhance the performance of existing **music information retrieval (MIR) tasks** or solve **new multimodal challenges** (mapping, retrieval, etc.).

Deep multimodal learning (DML) extends the ability of deep neural network to automatically learn **hierarchica**l and increasingly more **abstract representations** of the input data by leveraging **supplementary** and **complementary information** provided by different data modalities with the aim of building a richer representation.

Deep multimodal architectures have successfully been employed to improve performance in **speech recognition**, **emotion detection** in videos, **automatic image captioning**, **activity recognition**, multimedia **content indexing and retrieval** [1], but have only rarely been exploited to enhance **machine intelligence** in music-related tasks.

Related Work

- Music **genre classification** using audio tracks, text reviews and cover art images [2].
- Cold-start **music recommendation** by combining text and audio with user feedback data [3].
- Music **emotion recognition** using audio with tags or images [4].
- Cross-modal **music retrieval** by embedding lyrics, song audio and artist IDs into the same vector space [5].

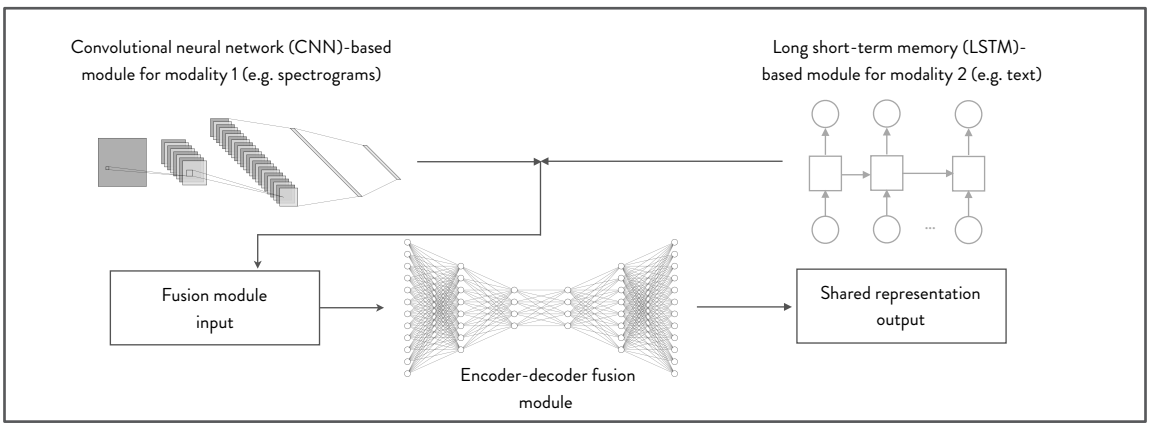
Research Directions

- Identifying **modality-specific modules** that preserve inter- and intra-modality **correlations**.
- Investigating **fusion strategies** which employ an **attention mechanism** to learn useful shared modality representations by extracting salient features [6].
- Exploring **deep transfer learning** in a multimodal setting, especially when one of the domains is characterised by noisy or missing data [7].

Two main **challenges**:

- **Multimodal representation**: joint (vectors which encode modality-invariant semantics) or coordinated (vectors which preserve inter-modality correlations)?
- Devising a **fusion strategy**: early or late fusion?

Example of a Multimodal Architecture



References

[1] Baltrušaitis T. et al. "Multimodal machine learning: A survey and taxonomy." *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 41.2: 423-443, 2018.
[2] Oramas S. et al. "Multimodal deep learning for music genre classification." *Transactions of the International Society for Music Information Retrieval*. 1 (1): 4-21, 2018.
[3] Oramas S. et al. "A deep multimodal approach for cold-start music recommendation." *Proceedings of the 2nd Workshop on Deep Learning for Recommender Systems*. ACM, 2017.
[4] Kim YE. et al. "Music emotion recognition: A state of the art review." *Proceedings of ISMIR*: 937-952, 2010.
[5] Watanabe K., Goto M. "Query-by-Blending: a Music Exploration System Blending Latent Vector Representations of Lyric Word, Song Audio, and Artist." Proceedings of ISMIR: 144-151, 2019.
[6] Huang F. et al. "Learning joint multimodal representation with adversarial attention networks." *2018 ACM Multimedia Conference on Multimedia Conference*. ACM 2018.
[7] Kim, Jaehun, et al. "One deep music representation to rule them all? A comparative analysis of different representation learning strategies." *Neural Computing and Applications*: 1-27, 2018

UK Research and Innovation
UNIVERSAL MUSIC GROUP
Queen Mary University of London