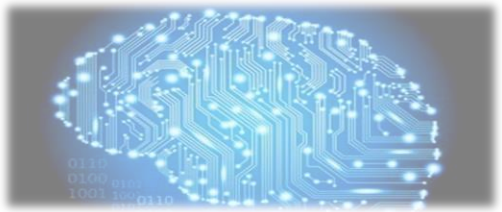
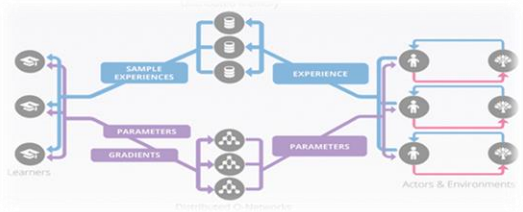



Multimodal Research Papers

AutoML 연구실
석사과정 문아성
21.10.05

Outline

- 개인 연구

  Multimodal Research Papers

- 효율적인 음악 장르 분류 시스템

  새로운 모델 제안 (channel compression 사용하기 힘든 이유를 설명해야 함)

Multimodal Research Papers

- MMTM: Multimodal transfer module for CNN fusion

Joze, Hamid Reza Vaezi, et al. "MMTM: Multimodal transfer module for CNN fusion."

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

- End-to-end multimodal emotion recognition using deep neural networks

Tzirakis, Panagiotis, et al. "End-to-end multimodal emotion recognition using deep neural networks."

IEEE Journal of Selected Topics in Signal Processing 11.8 (2017): 1301-1309.

- CNN-based sensor fusion techniques for multimodal human activity recognition

Münzner, Sebastian, et al. "CNN-based sensor fusion techniques for multimodal human activity recognition."

Proceedings of the 2017 ACM International Symposium on Wearable Computers. 2017.

MMTM: Multimodal transfer module for CNN fusion

- **Multimodal Transfer Module**
  - *Squeeze and Multimodal Excitation*

  To recalibrate channel-wise features in each modality

- **Applications**
  - *Hand Gesture Recognition: RGB, Depth Image*
    EgoGesture, NVGestures Datasets
  - *Audio-Visual Speech Enhancement: RGB Image, Sound*
    VoxCeleb2 Dataset
  - *Human Action Recognition: RGB, Skeleton Image*
    NTU-RGBD Dataset

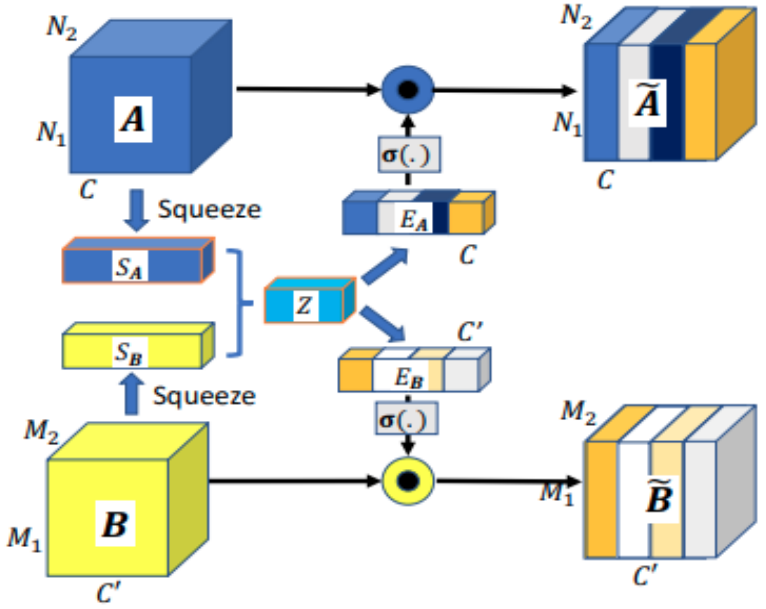


Figure 2. Architecture of MMTM for two modalities. **A** and **B**, that represent the features at a given layer of two unimodal CNNs, are the inputs to the module. For better visualization we limit the number of their spatial dimensions to 2. MMTM uses squeeze operations to generate global feature descriptor from each tensor. Both tensors are map into a joint representation Z by using concatenation and fully-connected layer. The excitation signals $E_{\mathbf{A}}$ and $E_{\mathbf{B}}$ are generated using the joint representation. Finally the excitation signals are used to gate the channel-wise features in each modality.

End-to-end multimodal emotion recognition using deep neural networks

- **Feature Extraction**
  - *Visual Network*
  - *Speech Network*

- **Dataset: RECOLA Database**
  - 4 Modalities
    - *Audio*
    - *Video*
    - *Electro-cardiogram*
    - *Electro-dermal activity*

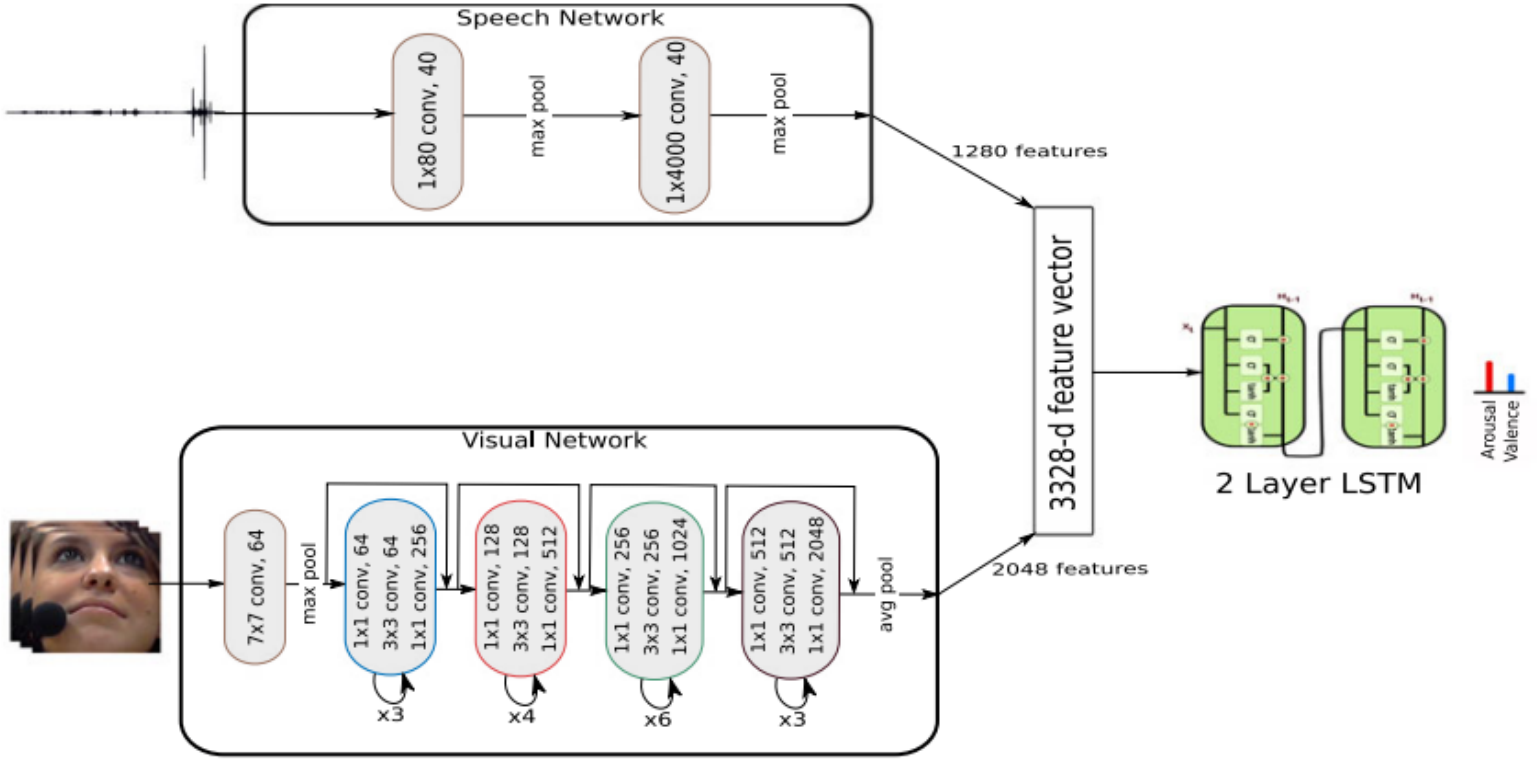


Fig. 1. The network comprises of two parts: the multimodal feature extraction part and the RNN part. The multimodal part extracts features from raw speech and visual signals. The extracted features are concatenated and used to feed 2 LSTM layers. These are used to capture the contextual information in the data.

CNN-based sensor fusion techniques for multimodal human activity recognition

- **Multimodal Sensor Fusion**
  - *Early fusion, Sensor-based late fusion, Channel-based late fusion, Shared filters hybrid fusion*

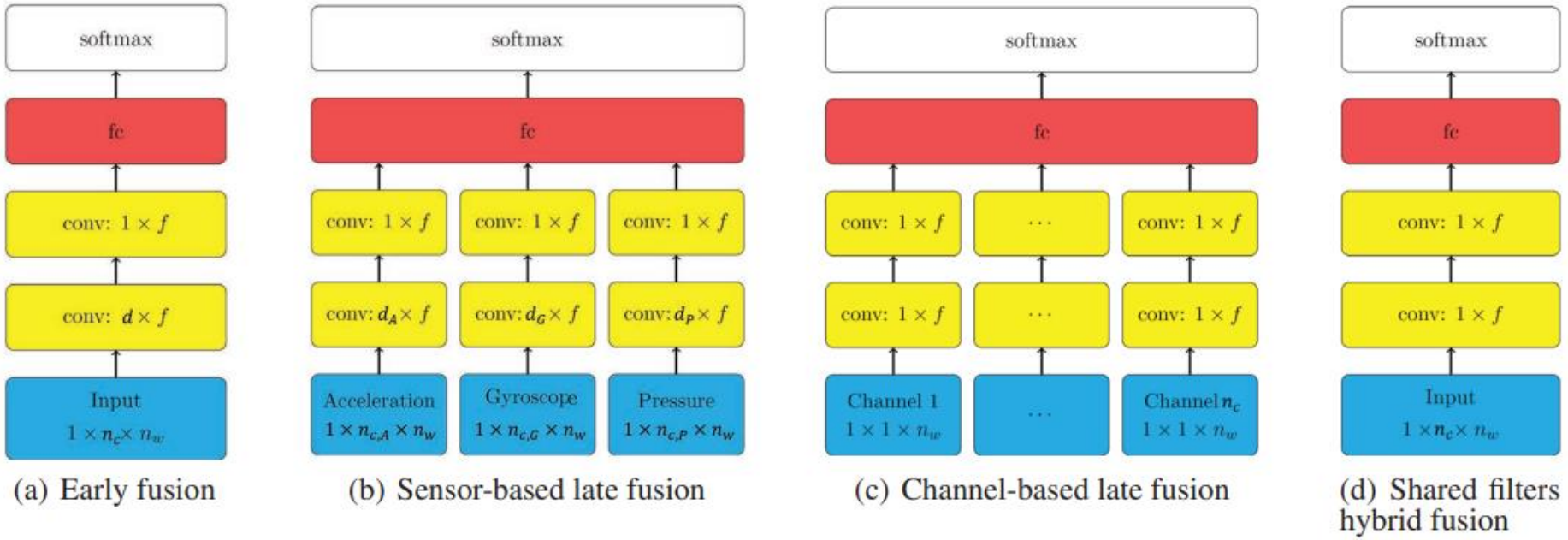
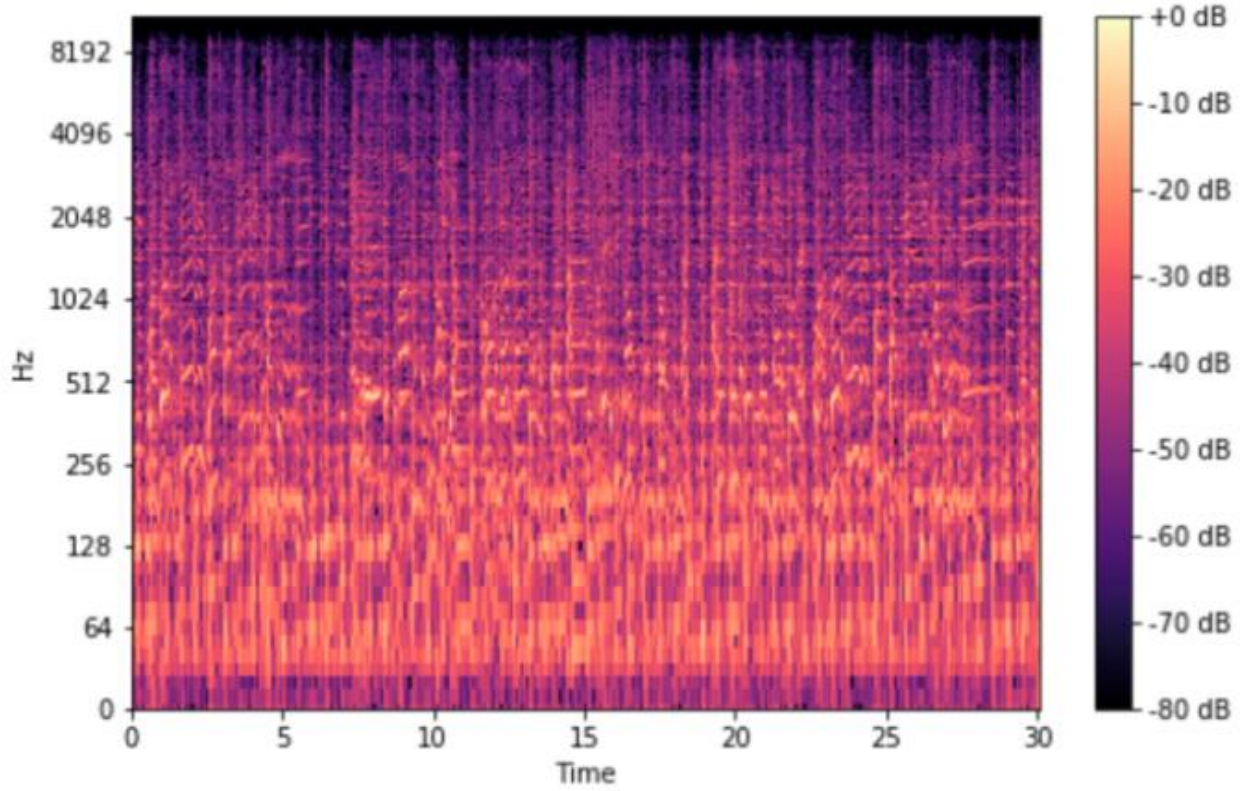


Figure 1. Schematic illustration of the four investigated sensor fusion techniques. Networks consist of two convolutional (conv) layers with filter size f in time dimension, one fully connected (fc), and a soft-max layer.

효율적인 음악 장르 분류 시스템

Input Image: ***Mel Spectrogram***



Mel Spectrogram

- Height: 주파수(Hz)
- Width: 시간(Time)
- Channel: 크기(dB)

- **Image shape: (H, W, C)**
- **30초 기준: (240, 960, 3)**

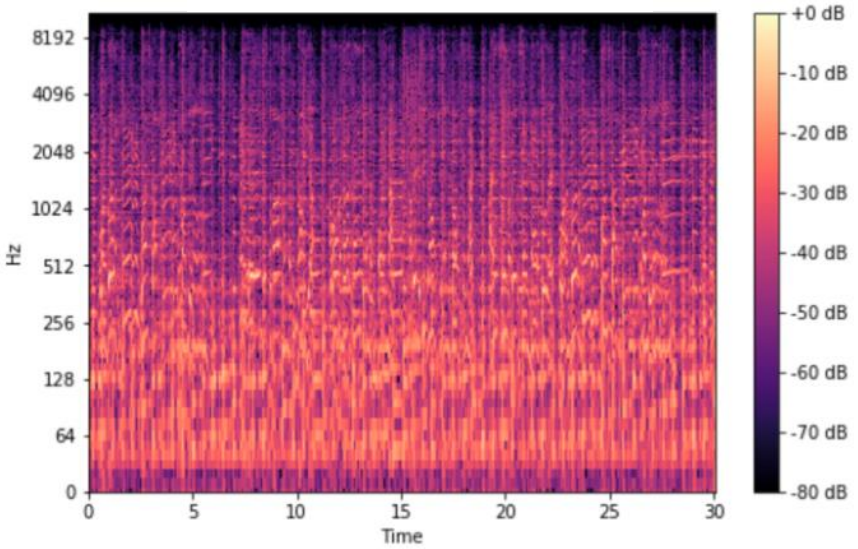
효율적인 음악 장르 분류 시스템

기존 Early Compression 모델을 적용하기 어려운 점

- 기존 CT Slice Image 에서의 채널간 불필요한 정보 부재

- Early Compression에 사용한 채널간 계산 량을 줄이기 위한 PointWise Convolution 사용이 힘듬

- Image shape: (H, W, C)
- 30초 기준: (240, 960, 3)



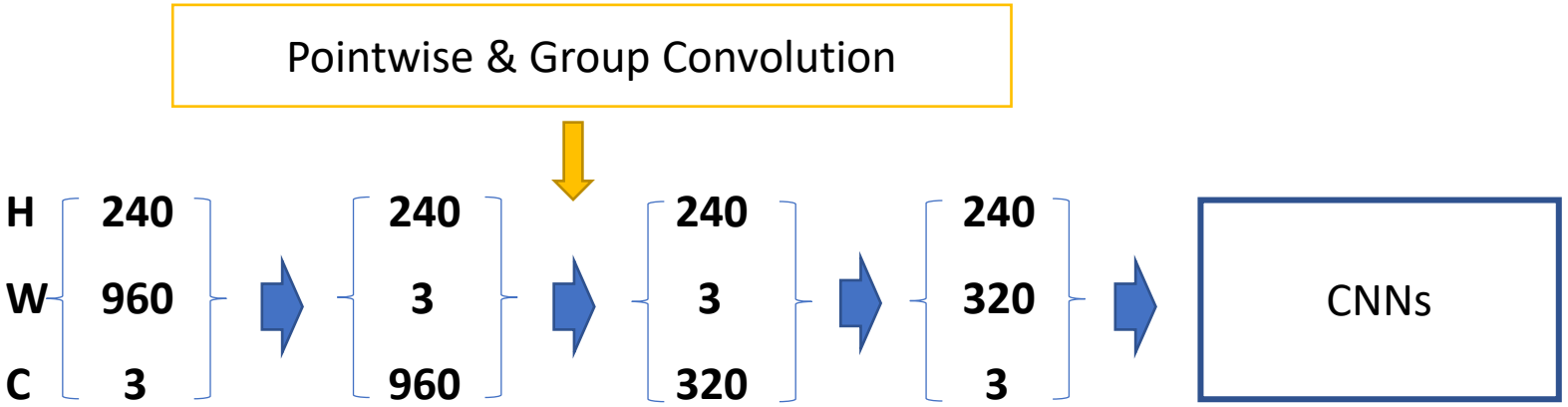
Mel Spectrogram

효율적인 음악 장르 분류 시스템

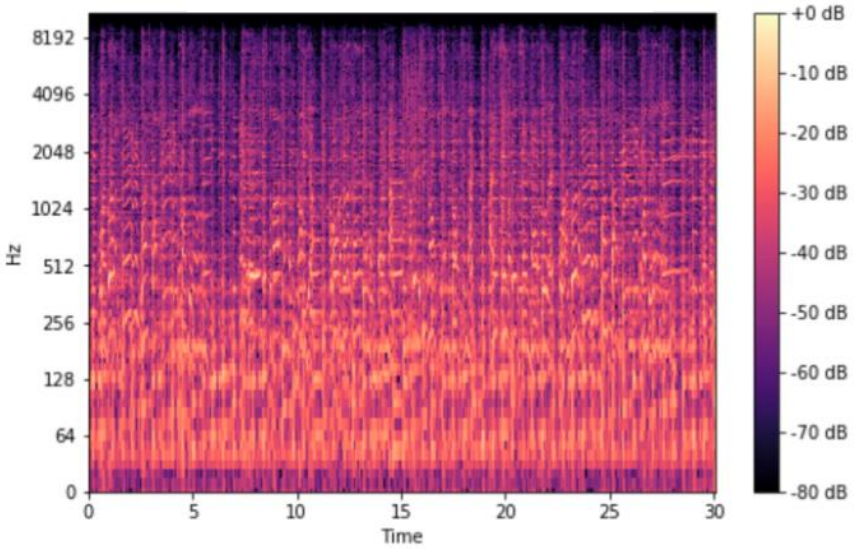
현 문제점을 개선하기 위한 새로운 모델 제안

- 시간(Time)을 채널방향으로 변형하여 PointWise Convolution 수행

- Time Sequence를 최대한 보존하기 위해 Group Convolution 수행
  (일정 시간 간격으로 그룹을 나눠 Convolution 수행)



- Image shape: (H, W, C)
- 30초 기준: (240, 960, 3)



Mel Spectrogram