Enhancing Affective Representations of Music-Induced EEG through Multimodal Supervision and Latent Domain Adaptation

K. Avramidis, C. Garoufis, A. Zlatintsi, P. Maragos

avramidi@usc.edu cgaroufis@mail.ntua.gr nzlat@cs.ntua.gr maragos@cs.ntua.gr

Signal Analysis and Interpretation Lab, USC

School of ECE, National Technical University of Athens

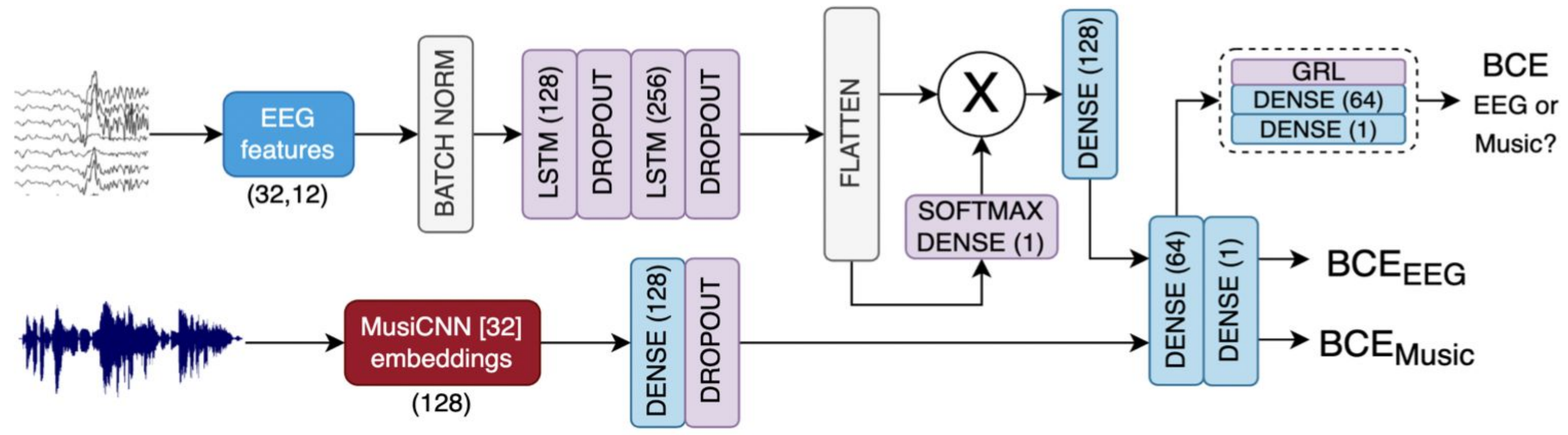
icassp 2022 Singapore, China, Online

#6829

Introduction: Music Perception

- A powerful form of emotion induction
- Greatly influences brain and body function
- Efficient tool to study human emotions
- Easy to see the affective impact of music
- **Challenge**: Map this effect to informative brain activity features of affect
- EEG over fMRI for a time-series analysis

Framework Architecture



- EEG Stream: Double LSTM + Dropout
- Lightweight attention head for EEG
- Music Stream: Pre-trained Embeddings
- Multimodal Supervision through BCE Losses against affect labels of both modalities
- Gradient Reversal Layer [1] as adversarial discriminator to reduce distribution shift
- Concurrent training of all modules

The Multimodal Approach

Concept: Utilize stimulus info to drive feature extraction from EEG

- Fusion using neural signals incorporates additional goals
  - Disentangle noisy signals from artifacts other than the stimulus
  - Enable dynamic (temporal) modeling of emotion induction
- Mapping of both representations onto a **common latent space**
- **Multimodal Supervision** instead of directly contrasting embeddings
- Inverse **Domain Discriminator** to reduce the distribution shift
  a. Gradient Reverse Layer [1] shifts the gradient to opposite direction
  b. By reversing the gradients of produced modality predictions (EEG or Music), we can extract modality-invariant features.

Metric	$\mathcal{J}$	ℓ_a only	$\neg\,\ell_{dd}$
$\mathrm{Acc}_{\mathrm{EEG}}$	**70.4%** – **68.9%**	67.8% – 68.0%	67.9% – 63.4%
P@10	**63.8%** – 65.0%	57.3% – 53.1%	63.4% – **66.7%**
mAP	59.1% – 67.8%	51.9% – 55.8%	**59.8%** – **68.1%**

Table 3. Ablation on the Objective Function for (Valence – Arousal). Here we solely consider mean aggregated scores over 32 subjects.

- ➔ Ablation study on the optimization objective: $\mathcal{J} = \lambda_{11}\ell_a + \lambda_{12}\ell_b + \lambda_2\ell_{dd}$
- ➔ Higher overall recognition performance for the joint objective
- ➔ Conditioning the common space on music crucial for retrieval
- ➔ GRL breaks modality-specific clusters but skews retrieval metrics

Results

Dimension	Non-Aggregated	Aggregated
Valence	62.9% – 71.5%	**70.4% – 78.7%**
Arousal	63.3% – 88.0%	**68.9% – 91.9%**

Table 1. Emotion Accuracy Scores for (EEG – Music) modalities, reporting mean values over 32 subject-specific models.

Dimension	Precision@10	mAvg. Precision
Valence	**19.4% – 63.8%**	18.8% – 59.1%
Arousal	18.4% – 65.0%	**19.9% – 67.8%**

Table 2. (Track – Emotion) Retrieval Scores on EEG input queries, reporting mean aggregated scores over 32 subjects.

→ Music discriminates better in both dimensions, as expected
→ Both modalities benefit from aggregation → **temporal variance**
→ Valence improves more → less uniform emotion alignment

→ Exact track retrieval emerges a challenge – not effective
→ Again signs of uniformity for arousal in the common space
→ Valence provides much lower mAP – **fragmented space**

- 32 personalized subject-specific models within 5-fold cross-validation
- Binary Classification – Binarizing VA labels at median, 5
- Aggregation: Classification is correct when for at least *half* samples

Feature Extraction

- Utilized Dataset: **DEAP [2]**
  - 32 participants, 34 stimuli music videos of 1min
  - Single global annotations of valence and arousal
  - EEG from 32 channels sampled at 128Hz
- EEG Features: **Differential Entropy**
  - Assuming an EEG sample X with (gaussian) distribution f(x), its differential entropy DE is defined as

$$h(X) = -\int_X f(x)\log(f(x))dx = \frac{1}{2}\log 2\pi e\sigma^2.$$

  - We compute variance using STFT for the major freq. bands
- Music Features: Pre-trained popular model **MusiCNN [3]**

Prospects – References

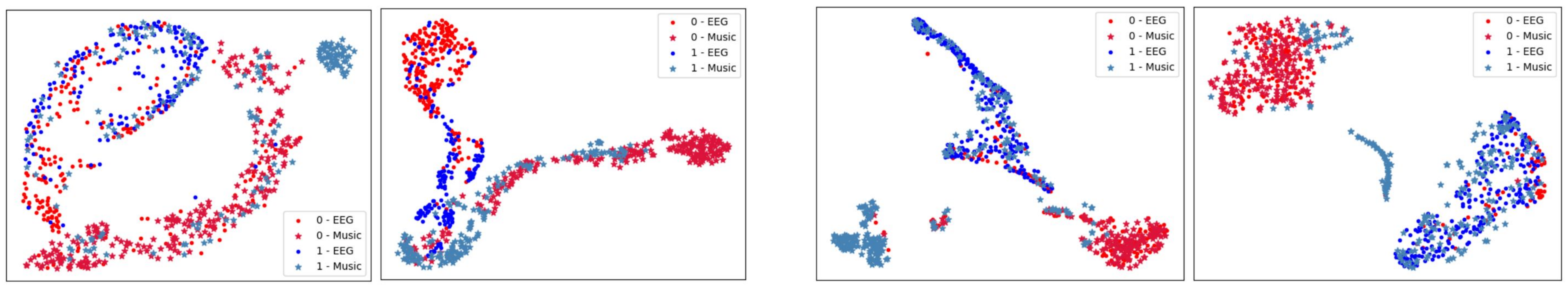
Concrete baseline for future work on dynamic modeling of music affect

- **Direction**: Emotion as a condition to the latent space – alternate labels
- **Direction**: Exact stimulus retrieval – poor outcomes so far

[1] Y. Ganin and V. Lempitsky, "Unsupervised Domain Adaptation by Backpropagation," in Proc. ICML 2015, Lille, France, 2015 **[2]** S. Koelstra et al., "DEAP: A Database for Emotion Analysis Using Physiological Signals," IEEE Trans. Affect. Computing, 2011 **[3]** J. Pons et al., "End-to-End Learning for Music Audio Tagging at Scale," in Proc. ISMIR 2018, Paris, France, 2018.

Looking at the Common Latent Space

Valence **Arousal**



Temporal Modeling of Emotion

- Retrieval scores for arousal across time
- Each score is averaged on all participants for the corresponding music clip sample
- Significant variability despite averaging
- Indicates salient moments in a music track
- Each track elicits emotion differently
- Similar findings for valence scores