A Multimodal BiMamba Network with Test-Time Adaptation for Emotion Recognition Based on Physiological Signals

Ziyu Jia[1], Tingyu Du[2], Zhengyu Tian[3], Hongkai Li[3], Yong Zhang[4], Chenyu Liu[5]

[1]Institute of Automation, Chinese Academy of Sciences, [2]Institute of Computing Technology, Chinese Academy of Sciences,
[3]Beijing Jiaotong University, [4]Huzhou University, [5]Nanyang Technological University

Introduction

Emotion Recognition based on Physiological Signals:

◆**Physiological signals** reflect the true emotional state of the human body objectively. Sleep staging is important for assessing sleep quality and diagnosing sleep disorders.

Related Work:

a) Traditional machine learning methods:

- Support Vector Machine.
- Rely heavily on expert knowledge, are limited by feature design and selection.

b) Deep learning methods:

- **Unimodal**: SST-EmotionNet, TSception.
- **Multimodal**: the residual LSTM network, the Neurophysiological Transformer, GJFusion, etc.
- Existing emotion recognition methods **rely on traditional backbones**, which struggle to model long-range dependencies and inter-modal correlations, and rarely address missing modality issues.

Challenge

C1: **How to effectively model both the intra-modal long-range dependencies and inter-modal correlations of physiological signals.**

C1.1: The intra-modal long-range dependencies:

- Physiological signals exhibit long-range dependencies reflecting the gradual accumulation of emotional changes.
- CNN-based networks struggle to capture essential long-range temporal information. Transformer-based networks lacks an explicit temporal-filtering capability.

C1.2: The inter-modal correlations:

- The inter-modal correlations are evident in the way different modalities respond.
- Transformer-based networks cannot be fully captured through pairwise channel-wise interactions, making it difficult to model high-order correlations.

C2: **How to mitigate the impact of missing multimodal emotion-related physiological data on model.**

- When acquiring emotion-related physiological signals, uncontrollable factors may result in incomplete multimodal signal acquisition to varying degrees.
- The unavoidable presence of missing data amplify the distribution shifts in physiological data.

Methods

C1: **How to effectively model both the intra-modal long-range dependencies and inter-modal correlations of physiological signals.**

S1: **We design a multimodal BiMamba network, where the intra-modal BiMamba module models long-range dependencies within modalities, and the inter-modal BiMamba module captures inter-modal correlations.**

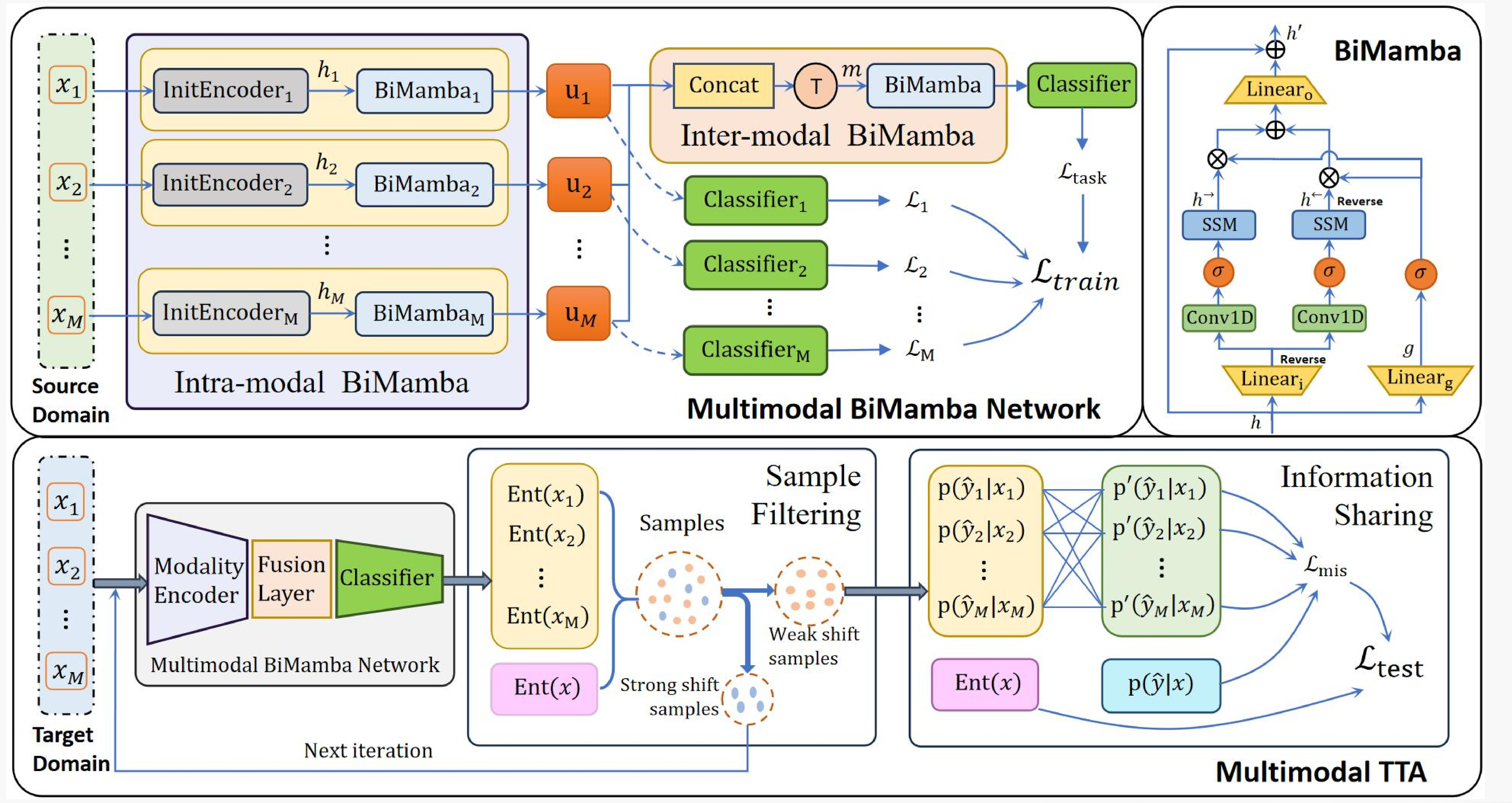
S1.1: The intra-modal BiMamba module.

- $\boldsymbol{InitEncoder_i}$ **:** Extract shallow features from each modality input from a specific modality.
- **BiMamba:** Model the temporal dimension further.

S1.2: The inter-modal correlations:

- **Concatenate** features of each modality along channel dimension, and **swap** time and channel dimensions.
- **BiMamba**: perform bidirectional state modeling of the features from different modalities along the channel dimension.

BiM-TTA: A Multimodal BiMamba Network with Test-TimeAdaptation



BiMamba:

- **Gating mechanism**: That adaptively weights features to highlight emotion-relevant information while suppressing.
- **State-space modeling**: That models temporal dependencies from both forward and backward perspectives, thereby capturing richer contextual dynamics.
- **Linear projection and residual connection**: Those integrate bidirectional features and stabilize training.

C2: **How to mitigate the impact of missing multimodal emotion-related physiological data on model.**

S2: **We propose the multimodal TTA method consisting of two key steps: 1) two-level entropy-based sample filtering and 2) mutual information sharing across modalities.**

Step1: Two-level entropy-based sample filtering:

- Selectively retains samples with low multimodal entropy and high unimodal entropy.
- Next, we employ an iterative entropy-based sample selection strategy to progressively expand the range of target-domain samples.
- Multimodal entropy reflects the model's certainty about its prediction. Unimodal entropy measures the extent to which a sample relies on multimodal information.

Step2: Mutual information sharing across modalities:

- Mutual information sharing across modalities leverages more informative modalities to guide those with significant missing information, alleviating amplified inter-modal distribution shifts.
- To improve the consistency of predictions across different modalities, we can minimize the KL divergence between the probability and its complementary probability.

Results

◆BiM-TTA is evaluated on two publicly available multimodal datasets: DEAP and MAHNOB-HCI.

	DEAP		MAHNOB-HCI	
	Valence	Arousal	Valence	Arousal
SVM	0.552	0.584	0.564	0.573
EEGNet	0.566	0.593	0.609	0.612
ACRNN	0.609	0.638	0.610	0.613
HetEmotionNet	0.625	0.633	0.607	0.601
TSception	0.613	0.635	0.633	0.599
LGGNet	0.618	0.636	0.632	0.609
VSGT	0.631	0.628	0.613	0.599
EEG-Deformer	0.609	0.630	0.587	0.595
MambaFormer	0.621	0.587	0.588	0.619
SST	0.613	0.623	0.606	0.616
BiM-TTA(ours)	**0.673**	**0.641**	**0.650**	**0.635**

	Valence(%)					Arousal(%)				
Mask ratio	0.2	0.4	0.6	0.8	Avg	0.2	0.4	0.6	0.8	Avg
Tent	-0.162	0.000	-0.314	-0.162	-0.162	0.471	-0.469	0.167	0.232	0.101
EATA	-0.315	0.157	-0.154	0.076	-0.061	0.701	-0.705	0.305	0.234	0.134
READ	**1.562**	0.625	0.234	0.703	0.781	0.312	-0.070	0.391	0.859	0.372
2LTTA	0.937	0.937	0.546	1.010	0.858	0.390	0.156	0.937	0.234	0.429
BiM-TTA(ours)	1.406	**1.250**	**0.859**	**1.172**	**1.172**	**1.016**	**1.719**	**1.094**	**1.406**	**1.309**

	Valence(%)					Arousal(%)				
Mask ratio	0.2	0.4	0.6	0.8	Avg	0.2	0.4	0.6	0.8	Avg
Tent	-0.185	0.120	-0.291	0.183	-0.043	0.529	0.046	-0.370	0.139	0.086
EATA	0.556	-0.185	0.046	0.265	0.171	0.635	0.185	-0.139	**0.185**	0.217
READ	0.523	0.370	0.079	-0.741	0.058	0.529	-0.741	-0.556	0.079	-0.146
2LTTA	0.450	-0.218	0.079	0.185	0.124	0.575	-0.324	-0.185	0.000	0.017
BiM-TTA(ours)	**2.413**	**0.787**	**0.787**	**0.370**	**1.089**	**0.866**	**0.417**	**0.556**	**0.185**	**0.506**

◆Under the missing-data setting, the two Tables report improvements over "No Adapt," which evaluates a source-trained model directly on a missing-data target without adaptation. Across both datasets and mask ratios, BiM-TTA consistently achieves larger gains than other TTA methods and sets the state of the art.

◆Under the complete-data setting, as shown in Table, we compare BiM-TTA with four families of methods: traditional machine learning, CNN-based and unidirectional RNN-based models, GNN-based models, and Transformer or Transformer plus SSM models. BiM-TTA achieves state-of-the-art performance on both datasets.

Conclusion

Contribution:

- This paper proposes **a multimodal BiMamba network with TTA** for emotion recognition.
- The multimodal BiMamba network effectively captures **intra-modal dependencies** and **inter-modal correlations** of multimodal physiological signals.
- The TTA **alleviates the negative impact** of amplified distribution shifts caused by missing multimodal data.
- Experiment results demonstrate that our model achieves **state-of-the-art** performance.

Prospect:

- We will further extend BiM-TTA to broader physiological analysis tasks, including sleep stage classification and motor imagery.

Paper: https://openreview.net/forum?id=3vLp3J7540; code: https://anonymous.4open.science/r/BiM-TTA-B604.