Multimodal Learning for Detecting Stress under Missing Modalities

Julie Mordacq[1,2,4] Leo Milecki[1,3] Maria Vakalopoulou[1,3] Steve Oudot[1,2]
Vicky Kalogeiton[2,4]
[1] Inria Saclay [2] Ecole Polytechnique [3] CentraleSupelec, Paris-Saclay University [4] LIX, CNRS, IP Paris

Abstract

Dealing with missing modalities is critical for many real-life applications. In this work, we propose a scalable framework for detecting stress induced by specific triggers in multimodal data with missing modalities. Our method has two key components: (i) aligning all modalities in the space of the strongest modality (the video) for learning a joint embedding space and (ii) a Masked Multimodal Transformer, leveraging inter- and intra-modality correlations while handling missing modalities. We validate our method through experiments on the `StressID` *dataset, where we set the new state of the art while demonstrating its robustness across various modality scenarios and its high potential for real-life applications.*

1. Introduction

Monitoring physiological changes is crucial for assessing individuals' well-being, especially in safety-critical contexts. Examples include stress, a response to emotional and physical challenges [14], and a triggering or aggravating factor for various pathological conditions [4]. Physiological changes may be detected visually (videos), acoustically (audio), or via biomedical signals (e.g., electrocardiograms). Yet, specific modalities may be unavailable during testing and sporadically absent during training. Thus, a method that can handle missing modalities during training and testing while balancing modalities' contributions for robustness is pivotal. A few existing methods address the challenge of handling missing modalities. However, they may suffer from notable limitations, including (a) requiring all modalities during training [10], (b) not being easily generalizable to more than two modalities [8, 9], or (c) considering the same input dimension for all modalities [17, 19]. In this work, we propose (i) a framework that aligns multimodal representations to a common rich feature space, (ii) a fusion strategy to handle missing modalities during training and inference, (iii) we set the new state of the art on `StressID` [1] in various modality settings.

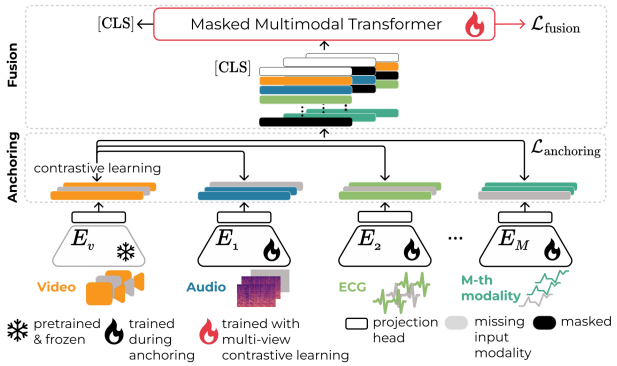

Figure 1. **Overview of our method.** (i) Contrastive learning aligns unimodal representations to the video. (ii) Fusion models multimodal interactions and handles missing modalities.

2. Anchored Multimodal Transformer

This study addresses stress detection using multimodal data, including video, audio, and biomedical signals. Real-world scenarios often involve missing modalities, motivating our goal to develop a modality-agnostic representation. Let $\mathcal{D} = \{(x_m^i)_{m=1}^M, y^i\}_{i=1}^N$ be our dataset, with M modalities and N labeled observations, where $x^i = (x_m^i)_{m=1}^M$ is the i-th observation (i.e., a family of m modality values) and $y^i \in \mathcal{Y} = \{0, 1\}$ the corresponding label (i.e., *stressed* or *not*). An overview is presented in Figure 1.

Anchoring. We train modality-specific encoders with a contrastive learning objective to align their representations to the one of the *video* (the strongest modality). Let us consider a pair of modalities with aligned observation $(\mathcal{V}, \mathcal{M}_m)$, where $\mathcal{V}$ the video, and $\mathcal{M}_m$ another modality. The video x_v^i and its corresponding observation x_m^i are encoded using $z_v^i{=}E_v(x_v^i)$ and $z_m^i{=}E_m(x_m^i)$, respectively, where E_v is a pre-trained and frozen video encoder and E_m a DNN. Projection heads map the embeddings to $f_a^i, f_m^i{\in}\mathbb{R}^d$. The loss is computed on f_a^i and f_m^i [5]: $\mathcal{L}_{\mathcal{V},\mathcal{M}_m} = -\sum_{i=1}^B \log \frac{\exp(\cos(f_v^i, f_m^i)/\tau)}{\sum_{k=1}^B \exp(\cos(f_v^i, f_m^k)/\tau)}$, $\tau \in \mathbb{R}^+$ the temperature parameter, $\cos(.,.)$ the cosine similarity, and

B the batch size. In practice, we use a symmetric loss: $\mathcal{L}_{\mathcal{V},\mathcal{M}_m} + \mathcal{L}_{\mathcal{M}_m,\mathcal{V}}$. and define the anchoring loss for M modalities: $\mathcal{L}_{\text{anchoring}} = \sum_{m=1,\mathcal{M}_m \neq \mathcal{A}}^{M} (\mathcal{L}_{\mathcal{V},\mathcal{M}_m} + \mathcal{L}_{\mathcal{M}_m,\mathcal{V}})$.

Masked Multimodal Transformer. To effectively build modality-agnostic representations, we resort to the transformer architecture [16]. For each sample, we stack the modality-specific representations, $f_m^i \in \mathbb{R}^d, \forall m \in [1, M]$, into a single matrix and prepend a special token [CLS], yielding a matrix $F \in \mathbb{R}^{(M+1)\times d}$. Similarly to Liu et al. [7], the query, key and value are derived from F via: $Q = W^Q F$, $K = W^K F$ and $V = W^V F$ where $Q, K \in \mathbb{R}^{(M+1)\times d_k}$ and $V \in \mathbb{R}^{(M+1)\times d_v}$. Our modelization of inter-modal interactions differs from the usual cross-attention [2, 6], which asymmetrically combines two separate embedding sequences of same dimension. Using stacked features F allows to generalize to any number of modalities, with linear scalability in the number of modalities instead of quadratic.

Handling missing modalities. Inspired by Milecki et al. [11], we apply our strategy to the scaled dot-product, core of each multi-head self-attention sub-layer. We use a masking binary matrix Z that specifies which modalities are missing: $z_{ij} = 1$ if i and j are available, else $z_{ij} = 0$. The output O of the attention mechanism, for O_i each line of O is :

$$O_i = \sum_j z_{ij} \frac{\exp(Q_i^T K_j / \sqrt{d_k})}{\sum_{\{j', z_{ij'}=1\}} \exp(Q_i^T K_{j'} / \sqrt{d_k})} V_j \quad .$$

We train the Masked Multimodal Transformer with a multi-view contrastive objective [3]. Inspired by Shi et al. [15], we mitigate the model's over-reliance on a single modality while enhancing its robustness in the absence of modalities through the *modality dropout* augmentation technique.

3. Results & Discussion

Dataset. `StressID` [1] for stress identification contains physiological responses via electrocardiogram, electrodermal activity, respiration, audio, and videos. We denote $X_{\text{train}}, X_{\text{test}}$ the entire train and test set (considering samples with and without missing modalities), and $X^*_{\text{train}}, X^*_{\text{test}}$ for the train and test sets where all modalities are available. We follow the train, validation, and test splits provided in [1] and report the same metrics: balanced accuracy (ACC), weighted F1-Score (F1) in format mean(std).

Implementation details. The Anchoring and Masked Multimodal Transformer steps are trained separately on X^*_{train} and X_{train}, respectively. A linear classifier is trained using the [CLS] token output for the final task. At each step, we train for 70 epochs using AdamW optimizer, with a learning rate of 1e-4 and a batch size of 128. For the *Video* modality, we use the Hiera [13] pre-trained encoder. For *audio*, each sample is encoded into a mel-spectrogram and fed to BYOL-A [12]. *Biomedical signals* are processed using 1D CNNs as suggested by [18]. Projection heads encode each representation as a $d = 64$ dimensional vector.

	ACC	F1
Video	62(4)‡	67(3)‡
Biomedical signals	58(4)‡	66(5)‡
Audio	62(4)‡	67(4)‡
Feature Fusion [1]	61(3)‡	66(4)‡
Decision Fusion† [1]	65(5)‡	72(5)‡
Ours	**69.5(3.7)**	**75.9(4.3)**

Table 1. **Comparison to SOTA on X^*_{test}.** **Bold**, underlined indicate the top **1**, 2 performing, respectively. ‡Results from [1]. For decision fusion†, we report the best out of all 4 decision rules.

	ACC	F1
	69.5(2.9)	69.6(3.1)
no video	61.2(4.6)	63.0(4.2)
no audio	68.3(2.9)	68.4(3.0)

Table 2. **Evaluation on two modality scenarios on X_{test}.** For each scenario, we systematically remove one modality from the test set.

Comparison to SOTA. Table 1 presents the comparison to the SOTA in the presence of all modalities. We compare against 'feature fusion' and 'decision fusion' [1] (rows 4 & 5) and unimodal baselines for Video, audio, and the stacked biomedical signals (rows 1, 2 & 3). Our method outperforms all other methods by a notable margin. For instance, it outperforms 'decision fusion' by 4% in ACC and 6% in F1.

Robustness to missing modalities. Table 2 reports the performances under missing modalities during inference. We remove audio or Video, the most cumbersome modalities to acquire, from X_{test} and compare the results and differences (Δ) to the ones obtained on X_{test}, the default set (row 1). The results remain competitive for both *no-audio* and *no-video*: $|\Delta| < 8.6\%$, even-though these modalities individually perform the best on X^*_{test} (Table 1). Additionally, Figure 2 shows ACC and F1 under different missing modality ratios η among $M * N$ with M the number of modalities and N the number of observations for training and inference. Our approach can successfully handle high ratios of missing modalities. More precisely, we report a delta of 2% between $\eta = 12.5\%$ (i.e., η inherent to the dataset) and $\eta = 30\%$ for ACC and F1.

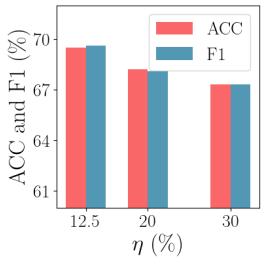

Figure 2. **Robustness to missing modality on X_{test}.**

Conclusion. We propose a modality-agnostic representation learning framework tailored to operate under missing modalities during training and testing. Applied to stress detection, our approach outperforms current SOTA and showcases robustness to missing modalities.

Acknowledgments

This work was partially supported by Inria Action Exploratoire PREMEDIT (Precision Medicine using Topology) and the ANR-22-CE39-0016 APATE. Additionally, it was partly performed using HPC resources from GENCI-IDRIS (Grant 2023-AD011014747).

References

[1] Hava Chaptoukaev, Valeriya Strizhkova, Michele Panariello, Bianca Dalpaos, Aglind Reka, Valeria Manera, Susanne Thümmler, Esma Ismailova, Massimiliano Todisco, Maria A Zuluaga, et al. Stressid: a multimodal dataset for stress identification. *NeurIPS*, 36, 2023. 1, 2

[2] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *ICCV*, pages 357–366, 2021. 2

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607, 2020. 2

[4] Joel E Dimsdale. Psychological stress and cardiovascular disease. *J. Am. Coll. Cardiol.*, 51(13):1237–1246, 2008. 1

[5] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, pages 15180–15190, 2023. 1

[6] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *ICML*, pages 4651–4664, 2021. 2

[7] Zhisong Liu, Robin Courant, and Vicky Kalogeiton. Funnynet: Audiovisual learning of funny moments in videos. In *ACCV*, pages 3308–3325, 2022. 2

[8] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *AAAI*, pages 2302–2310, 2021. 1

[9] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *CVPR*, pages 18177–18186, 2022. 1

[10] Mayur Mallya and Ghassan Hamarneh. Deep multimodal guidance for medical image classification. In *MICCAI*, pages 298–308, 2022. 1

[11] Leo Milecki, Vicky Kalogeiton, Sylvain Bodard, Dany Anglicheau, Jean-Michel Correas, Marc-Olivier Timsit, and Maria Vakalopoulou. Contrastive masked transformers for forecasting renal transplant function. In *MICCAI*, pages 244–254, 2022. 2

[12] Daisuke Niizumi, Daiki Takeuchi, Yasunori Ohishi, Noboru Harada, and Kunio Kashino. Byol for audio: Self-supervised learning for general-purpose audio representation. In *IJCNN*, pages 1–8, 2021. 2

[13] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *ICML*, pages 29441–29454, 2023. 2

[14] Neil Schneiderman, Gail Ironson, and Scott D Siegel. Stress and health: psychological, behavioral, and biological determinants. *Annu. Rev. Clin. Psychol.*, 1:607–628, 2005. 1

[15] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*, 2022. 2

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 2

[17] Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *CVPR*, pages 15878–15887, 2023. 1

[18] Yihe Wang, Yu Han, Haishuai Wang, and Xiang Zhang. Contrast everything: A hierarchical contrastive framework for medical time-series. *NeurIPS*, 36, 2023. 2

[19] Yao Zhang, Nanjun He, Jiawei Yang, Yuexiang Li, Dong Wei, Yawen Huang, Yang Zhang, Zhiqiang He, and Yefeng Zheng. mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In *MICCAI*, pages 107–117, 2022. 1