A Dual-Branch Transformer for Affective Computing: Tackling Missing Modalities and Imbalance via Safe Attention and Focal Loss

Jun Yu[1]; Naixiang Zheng[1]; Guoyuan Wang[1]; Yunxiang Zhang[1]; Lingsi Zhu[1]; Jiaen Liang[2]; Wei Huang[2]; Shengping Liu[2]
[1]University of Science and Technology of China, [2]Unisound AI Technology Co., Ltd.

Abstract

Emotion recognition in real-world environments is hindered by partial occlusions, missing modalities, and severe class imbalance. To address these issues, we propose a multimodal framework that dynamically fuses visual and audio representations.

Our approach uses a dual-branch Transformer architecture featuring a safe cross-attention mechanism and a modality dropout strategy. This design allows the network to rely on audio-based predictions when visual cues are absent. To mitigate the long-tail distribution of the Aff-Wild2 dataset, we apply focal loss optimization, combined with a sliding-window soft voting strategy to capture dynamic emotional transitions and reduce frame-level classification jitter.

Main Contributions

•**Dual-Branch Transformer:** Extracts audio-visual features independently, utilizing cross-attention and a learnable gating mechanism for adaptive cross-modal fusion.

•**Missing Modality Handling:** Features a specialized attention module with dynamic masking. If visual data is heavily occluded or missing, the system gracefully degrades to rely solely on the audio branch via residual connections.

•**Temporal Modeling:** Employs a sliding window approach and **soft voting inference** to capture dynamic emotional transitions and eliminate frame-level prediction jitter.

•**Long-tail Optimization:** Integrates focal loss to mitigate dataset bias (Aff-Wild2), forcing the model to focus on challenging minority classes and boosting overall generalization.

Method Details

Feature Extraction

- Visual (BEiT): Two-stage pipeline. Pre-trained on mixed static datasets (Raf-DB, FERPlus, AffectNet) for facial expressions, then domain-adaptive fine-tuning on target videos (Aff-Wild2) for temporal features V.
- Audio (WavLM): Extracts acoustic prosody and subtle emotional fluctuations from raw audio streams yielding features A.

Multimodal Attention Network

- Alignment: Projects V and A to a shared dimension with sinusoidal positional encodings.
- Cross-Attention: Bidirectional inter-modal interactions ($V \rightarrow A$, $A \rightarrow V$) via Transformer encoders.
- Dynamic Gating Fusion: Learnable gates (G_v, G_a) dynamically regulate information flow: $F_v = G_v \odot H_v + (1 - G_v) \odot H_{v \rightarrow a}$
- Classification: Concatenated representations are fed to an MLP for 8-class prediction.

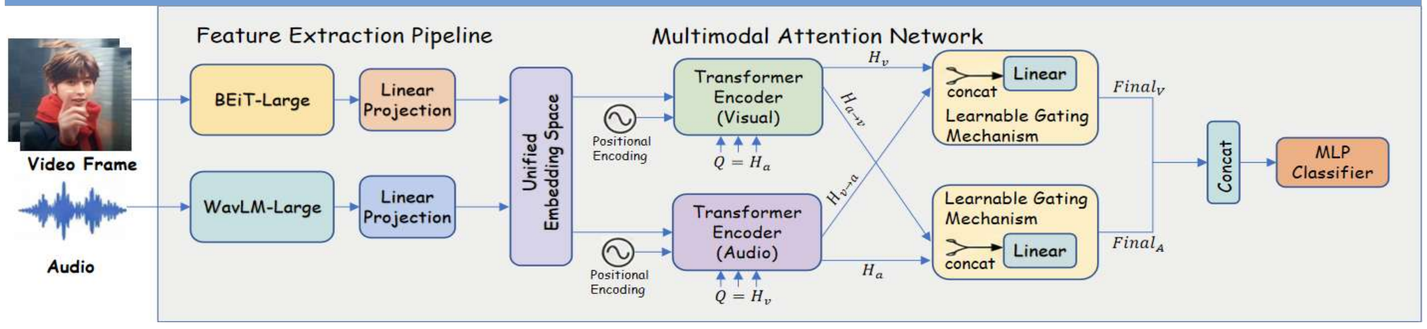
Robustness: Modality Dropout & Safe Attention

- Modality Dropout: Randomly masks visual inputs within a batch to simulate signal loss, preventing modality over-reliance.
- Safe Attention: If a visual window is entirely missing, the network forces Attn_output $= 0$. Combined with residual connections, the layer simplifies to $Q = \text{LayerNorm}(Q)$, ensuring the model purely relies on audio cues without breaking the forward pass.

Optimization & Inference

- Focal Loss: Addresses severe long-tail class imbalance ($\gamma = 2.0$), forcing focus on difficult minority classes.
- Sliding Windows & Soft Voting: Extracts overlapping sequences (Window=64, Stride=8) and averages predicted logits for continuous sequences.
- Post-processing: Median filtering (kernel k=11) smooths transient classification jitter while preserving emotional boundaries.

Overall Framework



Experiment

Ablation studies on Validation Set. Modality dropout probability (p), dimension (d), and the number of self-attention layers (l).

p	d	l	Accuracy	F1-Score
0.0	256	2	0.5677	0.4628
0.0	512	2	0.5820	0.4739
0.0	256	3	0.5824	0.4764
0.0	512	3	0.5730	0.4626
0.10	256	3	**0.6079**	**0.5029**
0.10	256	4	0.5981	0.4814
0.15	256	3	0.5815	0.4819
0.20	256	3	0.5935	0.4734

Conclusion

We presented a multimodal emotion recognition framework to address missing modalities, noisy data, and class imbalance. Our experiments on the Aff-Wild2 dataset provide practical insights into multimodal affective computing. While vision is the dominant modality, audio offers essential supplementary cues. Simulating visual loss through moderate modality dropout improves fault tolerance.
Experiments demonstrate that our framework effectively handles missing modalities and complex spatiotemporal dependencies, achieving third place in the 10th ABAW Expression challenge.