Towards Emotionally Aware AI:
Challenges and Opportunities in the Evolution of Multimodal Generative Models

Matteo Spanio

Centro di Sonologia Computazionale, University of Padova. Contact: spanio@dei.unipd.it

Introduction

Modern multimodal generative models have demonstrated impressive capabilities, such as generating text-based images or soundscapes. These systems rely on the fusion of modalities in order to be able to transform a modality into a different one. Although powerful, they lack the ability to esplicitally incorporate the emotional context [2] that is essential to human communication. In fact, at the core of human brain there is the *limbic system*, where are located, among others, the **amygdala** and the **hypothalamus**, which are responsible for sensory perception and emotional regulation.

Emotion-aware AI opens opportunities for applications in Human-Computer Interaction where empathy and context matter, and would allow to achieve better explainability as AI should be more aligned with psychological models.

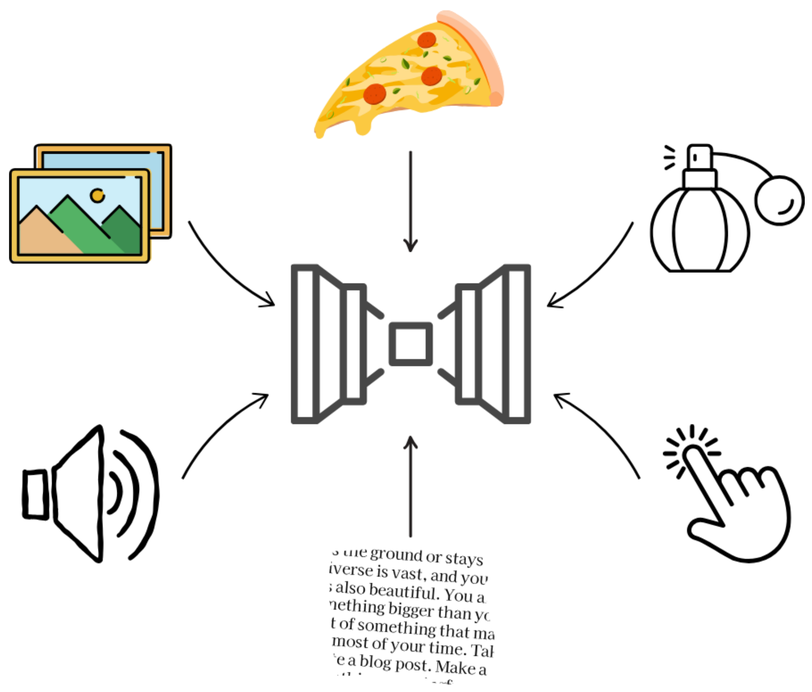

Figure 1: AI model designed with an emotional latent space. This model encodes diverse sensory modalities into a unified emotional latent space. By leveraging many encoders-decoders, the model can seamlessly transform inputs from one modality to another, enabling emotionally coherent content generation.

Challenges of Integrating Emotional Awareness

1. **Lack of multimodal emotional datasets**: While large datasets exist for individual modalities, few integrate diverse sensory data. High-quality data collection is costly and labor-intensive, limiting progress in this area.
2. **Neglected modalities**: Sensory inputs like smell and taste are critical to human emotion but lack computational representations and standardized datasets.
3. **Computational complexity**: Building emotion-aware models requires significant computational resources. Aligning different modalities within an emotional latent space demands advanced architectures, which are expensive and energy-intensive.

Proposed solutions

- **Dataset normalization**: Existing research in psychology and neuroscience offers a wealth of high-quality data on emotion and perception. By aggregating and standardizing these disparate datasets, researchers can create comprehensive multimodal emotional datasets.
- **Leveraging transfer learning**: Instead of building models from scratch, transfer learning techniques can fine-tune existing multimodal encoders to incorporate emotional understanding.
- **Emotion alignment in latent space**: Recent methods like contrastive learning have demonstrated the potential to align modalities within an emotional latent space [1]. This approach provides nuanced, dynamic representations of human emotions.

Conclusions

By embedding emotions into multimodal AI, we can create systems that:

- Interact naturally with humans.
- Deliver richer, more immersive experiences in areas like virtual reality, gaming, and media.
- Support empathetic technologies for therapeutic and assistive applications.

However, progress depends on addressing challenges in data availability and computational scalability. Multimodal AI must evolve to represent less-explored sensory domains.

Acknowledgements

Special thanks to Prof. Antonio Rodà (University of Padova) and Prof. Massimiliano Zampini (University of Trento) for their precious guidance and support.

References

[1] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: 2103.00020.

[2] Sicheng Zhao et al. "Emotion-Based End-to-End Matching Between Image and Music in Valence-Arousal Space". In: *Proceedings of the 28th ACM International Conference on Multimedia*. MM '20. Seattle, WA, USA: Association for Computing Machinery, 2020, pp. 2945–2954.

Università degli Studi di Padova

Associazione Italiana per l'Intelligenza Artificiale

CSC Centro di Sonologia Computazionale