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

Matteo Spanio[1,*]

[1]Centro di Sonologia Computazionale (CSC), Department of Information Engineering, University of Padova, Via Giovanni Gardenigo, 6b, 35131 Padova (PD), Italy

Abstract

The evolution of generative models in artificial intelligence (AI) has significantly expanded the capacity of machines to process and generate complex multimodal data such as text, images, audio, and video. Despite these advancements, the integration of emotional awareness remains an underexplored dimension. This paper examines the state of the art in multimodal generative AI, with a focus on existing models developed by major technology companies. It then proposes an approach to incorporate emotional awareness into AI models, which would enhance human-machine interaction by improving the interpretability and explainability of AI-generated decisions. The paper also addresses the challenges associated with building emotion-aware models, including the need for comprehensive multimodal datasets and the computational complexity of incorporating less-explored sensory modalities like olfaction and gustation. Finally, potential solutions are discussed, including the normalization of existing research data and the application of transfer learning to reduce resource demands. These steps are essential for advancing the field and unlocking the potential of emotion-aware multimodal AI in applications such as healthcare, robotics, and virtual assistants.

Keywords

Multimodal AI, Emotion-aware AI, Generative models, Human-computer interaction, Multisensory integration

1. Introduction

Generative Artificial Intelligence (AI) has experienced rapid advancements, fundamentally transforming how machines interact with data and create new content. Generative AI models, particularly those based on deep neural networks, have revolutionized traditional data processing by autonomously learning and generating complex patterns from raw data. This shift is especially significant in unsupervised learning, where machines produce coherent and meaningful outputs without explicit guidance. These models can now generate text, images, audio, and video, opening vast possibilities across industries such as creative design, healthcare, and robotics [1]. Among these advancements, the rise of multimodal generative AI has been particularly impactful. Multimodality refers to AI systems' ability to process and integrate various types of data, such as text, images, audio, and video, to perform tasks involving cross-modal generation or understanding. By bridging different sensory inputs and outputs, multimodal AI models can generate content spanning multiple domains, mimicking a more human-like understanding of the world. The significance of multimodal AI lies in its capacity to address the limitations of traditional AI models confined to single modalities. These systems enhance machine perception and understanding, thereby increasing their applicability in real-world scenarios. However, the rise of multimodal systems introduces unique challenges. As AI models expand to include more diverse data forms and sensory inputs, the need for scalable and interpretable models becomes more pressing. Integrating emotional understanding into generative AI could enrich human-computer interaction and create systems that better grasp the subtleties of human experience. Despite significant progress in multimodal AI, the

Doctoral Consortium at the 23rd International Conference of the Italian Association for Artificial Intelligence Bolzano, Italy, November 25-28, 2024.

*Corresponding author.

spanio@dei.unipd.it (M. Spanio)

https://matteospanio.github.io/ (M. Spanio)

0000-0002-2436-7208 (M. Spanio)

emotional dimension remains largely underexplored [2]. The author argue that developing emotion-aware models is crucial. Emotions represent a fundamental intersection of perceptual modalities in the human brain. Key regions such as the amygdala and the hypothalamus play pivotal roles in sensory perception and emotional regulation, underscoring the importance of emotions and sensations in our cognitive architecture. Consequently, an emotionally aware AI system should model data in a manner more aligned with human cognition. The author firmly believes that emotion-aware models could revolutionize human-machine interaction by improving the interpretability and explainability of model decisions, addressing a key limitation of current AI systems [3, 4].

The subsequent sections will explore various aspects of this domain in the following order:

1. Background and Related Work: Outlining the current state of multimodal models, both with and without emotional awareness, highlighting widely used models from major tech companies like Google, Meta, and OpenAI.
2. Benefits of Emotionally Aware Models: Examining existing research on emotion-aware models, focusing on how emotions mediate other perceptual modalities and the potential advantages of integrating emotional understanding into AI systems.
3. Challenges and Limitations: Providing a critical analysis of the challenges and potential objections to emotion-aware models, including the complexities in creating multimodal datasets and the computational demands of representing olfactory and gustatory modalities.
4. Conclusions: Summarizing the key points, reiterating the importance of developing emotion-aware multimodal generative models to enhance human-computer interaction and some proposals on how to achieve such results.

2. Background and Related Work

The landscape of multimodal generative artificial intelligence (AI) is currently shaped by significant investments from major technology companies such as Google, Meta, OpenAI, and Microsoft. These companies are at the forefront of developing cutting-edge models capable of processing and generating diverse types of data, including images, audio, and video. Their efforts encompass not only the creation of expansive datasets but also the release of models that are either open source or proprietary. This trend underscores the immense resources being allocated globally to develop autonomous systems proficient in generating rich multimedia content.

2.1. Key Models and Developments

Several notable multimodal generative models have been introduced, showcasing the field's breadth of capabilities. Text-to-image generation models, such as DALL·E [5] and Stable Diffusion [6], create detailed images from textual descriptions. Text-to-audio models, like MusicLM [7], generate music or soundscapes from text prompts, with promising applications in entertainment and virtual environments. Although in its early stages, text-to-video generation shows potential in media production and simulation environments [8]. In the other direction, models such as image-to-text [9, 10, 11] translate visual inputs into descriptive narratives, providing enhanced capabilities for tasks like automated captioning and assisting individuals with visual impairments. Audio-to-text models, commonly seen in speech-to-text systems, have long been applied in areas such as transcription and virtual assistants, but recent advances in generative models enable more nuanced and context-aware interpretations of spoken language. However, even the simplest models involving two non-textual modalities, like the one discussed in [12] are essentially concatenations of multiple models exchanging textual information. Recently, multimodal models such as Mirasol, Chameleon, and others (including gtp4o) [13, 14, 15] have adopted a different approach called early fusion [16, 17], where the modalities converge into a single latent space that mixes tokens from different domains. Although this approach has yielded better results than previously described models, it remains difficult to interpret.

2.2. Integration of Emotional Awareness

A relatively less explored but emerging area within multimodal AI is the integration of emotional awareness. While extensive efforts have been dedicated to recognizing emotions within a single modality [18], there has been an increasing interest in fusing information from multiple modalities [18, 19]. This multimodal approach is advantageous because the combined information from different modalities provides a complementary capability for emotion recognition. However, relatively few efforts have been made to understand the emotion-centric correlation between different modalities. Recent approaches, such as those presented in [2], have shown concrete possibilities for connecting images and sounds through an emotional valence-arousal latent space leveraging supervised contrastive learning techniques. This contribution allows for a more nuanced and dynamic representation of emotional states compared to the older theory of discrete emotional states. By capturing the subtleties and complexities of human emotions, these newer models offer a more sophisticated understanding of how emotions interplay across different sensory inputs.

3. Benefits of Emotionally Aware Models

While the experiments conducted by [2] yielded promising results, the potential of contrastive learning in emotional contexts remains largely underexplored. Utilizing supervised contrastive learning allows for the alignment of various encoders corresponding to different modalities within a shared emotional latent space. This methodology supports the development of models that align with established psychological research linking modalities to emotions, as evidenced by studies examining the relationships between audio and emotions [20], odors and emotions [21], and temperature and emotions [22]. By leveraging these insights, it becomes feasible to include often-overlooked modalities such as touch, taste, and smell, which also have emotional correlations. Integrating this knowledge could propel advancements toward Artificial General Intelligence (AGI). Current research in these areas is still nascent, primarily focusing on textual descriptions, as seen in [23], which utilize transformer-based models. Emotion-aware models represent a significant advancement in AI research, as mapping human emotions to the latent space of generative models can foster more natural interactions in applications like virtual assistants, therapeutic tools, and social robots. By embedding emotional understanding, these systems can engage users in a more human-like manner, leading to smoother and more relatable interactions. For instance, an emotion-aware virtual assistant could adapt its tone and suggestions based on the user's emotional state, enhancing the user experience. Additionally, imposing constraints on the latent space allows for the application of known psychological models to clarify AI behavior and decision-making, enhancing transparency and trustworthiness. Emotions play a crucial role in human perception, influencing how we interpret sensory inputs. Emotion-aware models can bridge multiple modalities—vision, hearing, smell, and taste—creating a richer understanding of the environment and leading to more immersive applications in virtual reality, gaming, and interactive media.

To realize the benefits of emotionally aware models, we propose a framework utilizing pretrained encoder/decoder architectures tailored for each modality. This approach efficiently encodes emotional information from sensory data while minimizing computational demands. A pretrained encoder first processes the input data, such as a computational description of food, transforming it into a high-dimensional embedding. This embedding is then input into a specialized middle encoder model designed to capture the emotional essence, producing an emotional embedding represented as a vector of valence and arousal values. Following this, a pretrained decoder model converts the emotional vector into an audio token, which is processed by the pretrained audio model to generate the corresponding output. This architecture effectively manages the computational load through pretrained models, requiring only the middle encoder/decoder model to be trained. This design streamlines the training process and enhances the model's capacity to translate emotional information across modalities, fostering a more integrated and emotionally aware AI system.

4. Challenges and Limitations

The development of multimodal AI models faces several significant challenges and limitations, particularly concerning the availability and quality of datasets. One of the primary obstacles is finding comprehensive datasets that integrate multiple modalities. While substantial progress has been made in creating large-scale datasets for individual modalities, the integration of diverse sensory data, such as combining visual, auditory, and textual information, remains a challenge. This paucity of integrated datasets hampers the ability to train and evaluate multimodal models effectively. Moreover, the collection of multimodal data is both costly and labor-intensive, requiring expert evaluations to ensure data quality and alignment across modalities. Web scraping, a common method for gathering large amounts of data, proves insufficient for creating high-quality multimodal datasets. For instance, aligning different types of data (e.g., synchronizing audio with visual inputs) necessitates precise and controlled conditions, often achievable only in well-equipped laboratories. Existing datasets predominantly rely on massive, web-scraped data, which, while voluminous, often lack the quality needed for advanced deep learning applications. The release of models like Microsoft's Phi [24] has underscored the importance of data quality, demonstrating how high-quality datasets can enhance model efficiency and reduce computational resource requirements. In addition to these general challenges, specific modalities such as olfaction and gustation present unique difficulties. In these research communities, there is no established practice of sharing data in formats suitable for use as training datasets. Furthermore, there are no widely adopted computational representations for olfactory and gustatory information, making it challenging to integrate these modalities into multimodal models. Before end-to-end models that encompass these senses can be developed, significant research is needed to establish standardized computational frameworks and methodologies for these less-explored sensory domains.

5. Conclusions

The integration of emotional awareness into multimodal generative AI represents a pivotal next step for advancing human-computer interaction. Emotion-aware AI models can interpret and respond more contextually to human emotions, which is particularly relevant in applications like virtual assistants and healthcare, where nuanced emotional understanding is critical. Furthermore, these models enhance interpretability and trustworthiness in AI decisions. However, key challenges remain, especially in dataset availability and computational complexity. Comprehensive multimodal datasets that incorporate emotions are scarce, and the difficulty in representing sensory modalities such as olfaction and gustation hinders progress. Addressing these issues is crucial to fully realize the potential of emotion-aware AI.

5.1. Future Research Directions

To advance research in emotion-aware multimodal AI, two key strategies are proposed:

1. Dataset aggregation and normalization: a considerable body of research, particularly in psychology and neuroscience, has already explored the correlation between emotions and various sensory modalities. Although these data are currently dispersed and non-standardized, they often come from high-quality studies. A concerted effort to systematically aggregate and normalize these existing datasets could form the basis for a comprehensive multimodal dataset. Such a resource would support AI models in learning emotional correlations across multiple sensory inputs, creating a foundation for more robust emotion-aware systems.
2. Leveraging transfer learning to reduce computational complexity: the computational demands of building models from scratch, as seen in large tech companies, are a significant barrier for many research initiatives. However, existing deep learning-based encoders have reached a high level of performance. By adopting a contrastive learning framework and leveraging transfer learning or fine-tuning techniques, researchers can utilize pre-existing models to enhance emotion-aware capabilities without requiring massive computational resources. This approach not only shortens

the time needed to achieve results but also reduces energy consumption, making research more sustainable and accessible.

In summary, advancing emotion-aware multimodal AI requires addressing the current challenges of dataset availability and computational demands. By capitalizing on existing research data and leveraging transfer learning, these obstacles can be overcome, enabling the development of AI systems that are more aligned with human emotions. Such systems will significantly enhance human-computer interaction and broaden the scope of AI applications in various fields.

Acknowledgments

I would like to express my sincere gratitude to Professor Antonio Rodà and Professor Massimiliano Zampini for their invaluable discussions and advice throughout the development of this paper. Professor Rodà, from the Centro di Sonologia Computazionale (CSC), Department of Information Engineering, University of Padova, provided essential guidance and support. Professor Zampini, from the Center for Mind/Brain Sciences (CIMeC), University of Trento, offered key insights that significantly shaped this research. Their expertise and encouragement have been crucial to this work.

References

[1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language Models are Few-Shot Learners, 2020. URL: http://arxiv.org/abs/2005.14165. doi:10.48550/arXiv.2005.14165, arXiv:2005.14165 [cs].

[2] S. Zhao, Y. Li, X. Yao, W. Nie, P. Xu, J. Yang, K. Keutzer, 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, Association for Computing Machinery, New York, NY, USA, 2020, pp. 2945–2954. URL: https://dl.acm.org/doi/10.1145/3394171.3413776. doi:10.1145/3394171.3413776.

[3] R. A. Calvo, S. D'Mello, Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications, IEEE Transactions on Affective Computing 1 (2010) 18–37. URL: https://ieeexplore.ieee.org/document/5520655. doi:10.1109/T-AFFC.2010.1, conference Name: IEEE Transactions on Affective Computing.

[4] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, C. Zhong, Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges, 2021. URL: http://arxiv.org/abs/2103.11251. doi:10.48550/arXiv.2103.11251, arXiv:2103.11251 [cs, stat].

[5] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever, Zero-Shot Text-to-Image Generation, 2021. URL: http://arxiv.org/abs/2102.12092. doi:10.48550/arXiv.2102.12092, arXiv:2102.12092 [cs].

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-Resolution Image Synthesis with Latent Diffusion Models, 2022. URL: http://arxiv.org/abs/2112.10752. doi:10.48550/arXiv.2112.10752, arXiv:2112.10752 [cs].

[7] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi, M. Sharifi, N. Zeghidour, C. Frank, MusicLM: Generating Music From Text, 2023. URL: http://arxiv.org/abs/2301.11325. doi:10.48550/arXiv.2301.11325, arXiv:2301.11325 [cs, eess].

[8] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, D. Parikh, S. Gupta, Y. Taigman, Make-A-Video: Text-to-Video Generation without Text-Video Data, 2022. URL: http://arxiv.org/abs/2209.14792. doi:10.48550/arXiv.2209.14792, arXiv:2209.14792 [cs].

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning Transferable Visual Models From Natural Language Supervision, 2021. URL: http://arxiv.org/abs/2103.00020. doi:10.48550/arXiv.2103.00020, arXiv:2103.00020 [cs].

[10] J. Li, D. Li, C. Xiong, S. Hoi, BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, 2022. URL: https://arxiv.org/abs/2201.12086v2.

[11] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, K. Simonyan, Flamingo: a Visual Language Model for Few-Shot Learning, 2022. URL: http://arxiv.org/abs/2204.14198. doi:10.48550/arXiv.2204.14198, arXiv:2204.14198 [cs].

[12] R. Sheffer, Y. Adi, I Hear Your True Colors: Image Guided Audio Generation, 2023. URL: http://arxiv.org/abs/2211.03089. doi:10.48550/arXiv.2211.03089, arXiv:2211.03089 [cs, eess].

[13] A. J. Piergiovanni, I. Noble, D. Kim, M. S. Ryoo, V. Gomes, A. Angelova, Mirasol3B: A Multimodal Autoregressive model for time-aligned and contextual modalities, 2024. URL: http://arxiv.org/abs/2311.05698. doi:10.48550/arXiv.2311.05698, arXiv:2311.05698 [cs].

[14] C. Team, Chameleon: Mixed-Modal Early-Fusion Foundation Models, 2024. URL: http://arxiv.org/abs/2405.09818. doi:10.48550/arXiv.2405.09818, arXiv:2405.09818 [cs].

[15] H. Laurençon, L. Tronchon, M. Cord, V. Sanh, What matters when building vision-language models?, 2024. URL: http://arxiv.org/abs/2405.02246. doi:10.48550/arXiv.2405.02246, arXiv:2405.02246 [cs].

[16] K. Gadzicki, R. Khamsehashari, C. Zetzsche, Early vs Late Fusion in Multimodal Convolutional Neural Networks, in: 2020 IEEE 23rd International Conference on Information Fusion (FUSION), 2020, pp. 1–6. URL: https://ieeexplore.ieee.org/document/9190246. doi:10.23919/FUSION45008.2020.9190246.

[17] F. Yang, B. Ning, H. Li, An Overview of Multimodal Fusion Learning, in: Y. Chenggang, W. Honggang, L. Yun (Eds.), Mobile Multimedia Communications, Springer Nature Switzerland, Cham, 2022, pp. 259–268. doi:10.1007/978-3-031-23902-1_20.

[18] S. Poria, E. Cambria, R. Bajpai, A. Hussain, A review of affective computing: From unimodal analysis to multimodal fusion, Information Fusion 37 (2017) 98–125. URL: https://www.sciencedirect.com/science/article/pii/S1566253517300738. doi:10.1016/j.inffus.2017.02.003.

[19] S. Zhao, S. Wang, M. Soleymani, D. Joshi, Q. Ji, Affective Computing for Large-scale Heterogeneous Multimedia Data: A Survey, ACM Trans. Multimedia Comput. Commun. Appl. 15 (2019) 93:1–93:32. URL: https://dl.acm.org/doi/10.1145/3363560. doi:10.1145/3363560.

[20] H.-T. Hung, J. Ching, S. Doh, N. Kim, J. Nam, Y.-H. Yang, EMOPIA: A Multi-Modal Pop Piano Dataset For Emotion Recognition and Emotion-based Music Generation, 2021. URL: http://arxiv.org/abs/2108.01374. doi:10.48550/arXiv.2108.01374, arXiv:2108.01374 [cs, eess].

[21] C. A. Levitan, S. Charney, K. B. Schloss, S. E. Palmer, The Smell of Jazz: Crossmodal Correspondences Between Music, Odor, and Emotion., in: CogSci, volume 1, 2015, pp. 1326–1331. URL: https://www.academia.edu/download/84424911/paper0233.pdf.

[22] F. B. Escobar, C. Velasco, K. Motoki, D. V. Byrne, Q. J. Wang, The temperature of emotions, PLOS ONE 16 (2021) e0252408. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0252408. doi:10.1371/journal.pone.0252408, publisher: Public Library of Science.

[23] C. Boscher, C. Largeron, V. Eglin, E. Egyed-Zsigmond, SENSE-LM : A Synergy between a Language Model and Sensorimotor Representations for Auditory and Olfactory Information Extraction, in: Y. Graham, M. Purver (Eds.), Findings of the Association for Computational Linguistics: EACL 2024, Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 1695–1711. URL: https://aclanthology.org/2024.findings-eacl.119.

[24] S. Gunasekar, Y. Zhang, J. Aneja, C. C. T. Mendes, A. Del Giorno, S. Gopi, M. Javaheripi, P. Kauffmann, G. de Rosa, O. Saarikivi, A. Salim, S. Shah, H. S. Behl, X. Wang, S. Bubeck, R. Eldan, A. T. Kalai, Y. T. Lee, Y. Li, Textbooks Are All You Need, 2023. URL: http://arxiv.org/abs/2306.11644. doi:10.48550/arXiv.2306.11644, arXiv:2306.11644 [cs].