Cultural-Aware AI Model for Emotion Recognition

Mehrdad Baradaran
Department of Computer and Data Sciences,
Shahid Beheshti University,
Tehran, Iran
meh.baradaran@mail.sbu.ac.ir

Payam Zohari
Department of Computer Engineering,
Khaje Nasir University Of Technology,
Tehran, Iran
m.zohari@email.kntu.ac.ir

Abtin Mahyar
High Performance Computing Laboratory,
School of Computer Science,
Institute for Research in Fundamental Sciences,
Tehran, Iran
abtinmahyar@gmail.com

Hossein Motamednia
High Performance Computing Laboratory,
School of Computer Science,
Institute for Research in Fundamental Sciences,
Tehran, Iran
h.motamednia@ipm.ir

Dara Rahmati
Faculty of Computer Science and Engineering,
Shahid Beheshti University,
Tehran, Iran
d_rahmati@sbu.ac.ir

Saeid Gorgin
Department of Computer Engineering,
Chosun University, Gwangju, 61452,
South Korea
gorgin@irost.ir

***Abstract*—Emotion AI is a research domain that aims to understand human emotions from visual or textual data. However, existing methods often ignore the influence of cultural diversity on emotional interpretation. In this paper, we propose a multimodal deep learning model that integrates cultural awareness into emotion recognition. Our model uses images as the primary data source and comments from individuals across different regions as the secondary data source. Our results show that our model achieves robust performance across various scenarios. Our contribution is to introduce a novel fusion approach that bridges cultural gaps and fosters a more nuanced understanding of emotions. Due to the best of our knowledge, few works are using this approach, for Emotion AI, combining different types of data sources and models. We evaluate our model on the ArtELingo dataset, which contains image-comment pairs with Chinese, Arabic, and English annotations. The experimental results in the evaluation phase demonstrate an impressive 80% recognition accuracy for the model that merges image-text features.**

***Index Terms*—Emotion Recognition, DeepFeature Extraction, Image Representation, Text Representation**

I. Introduction

Artificial Intelligence (AI) has made significant progress in recent years. AI-enabled emotion, known as Emotion AI, is a pivotal and formidable research domain, as it has essential applications in developing services for human interaction and analysis. Emotion AI encompasses two distinct fields: Human-Computer Interaction (HCI) and Affective Computing, which use different methodologies to determine the emotions of individuals from visual or textual data, such as images, actions, texts, or videos [1]. The ability to understand human emotions based on their expressions is vital for creating systems that can communicate and interact with humans effectively [2]. This goal aligns with the vision of endowing computers with human-like attributes, which has attracted substantial attention in recent years to enhance emotional awareness in human-computer communication [3].

Recent studies predominantly employ Machine Learning (ML) and Deep Learning (DL) models for emotion recognition, leveraging their ability to extract significant patterns and features from input data and learn to adeptly detect emotions. This task extends across various data types, encompassing images (primarily facial expressions), texts for sentiment analysis, videos, speech signals, and physiological signals pertinent to healthcare applications [4]. Previous studies have explored diverse sources of information, reflecting the multi-modal nature of human expression. Notably, some have adopted an integrative approach, combining information from different sources to enhance the accuracy and efficacy of emotion recognition systems [5]–[9].

Existing work on emotion recognition has overlooked the intricate influence of diverse cultures and backgrounds on the interpretation of emotions [10], leading to an oversimplification of the problem. For instance, consider a painting depicting a religious ritual specific to a particular region. Individuals from that cultural milieu might typically manifest emotions closely linked to the associated religious practices upon viewing the artwork. However, individuals from differing cultural backgrounds might exhibit varied emotional responses influenced by the content or artistic style of the piece. This example underscores the nuanced and multifaceted nature of emotional reactions, shaped significantly by an individual's history and cultural context. It highlights the necessity to move beyond the homogenization of emotional interpretation and emphasizes the crucial role of cultural diversity in shaping emotional perceptions and expressions across different stimuli.

To overcome the limitations inherent in current methodologies, we propose the integration of cultural awareness into deep learning models for emotion recognition. Our approach utilizes images as one primary data source and introduces cultural nuances through corresponding comments from individuals across diverse regions as a secondary data source. This innovative model demonstrates robust performance across various scenarios. Our experiments are conducted on the ArtELingo dataset [11], a rich repository comprising image-comment pairs encompassing Chinese, Arabic, and English annotations for each image within the dataset. This comprehensive dataset empowers our model to effectively address and surmount the challenges posed by multicultural differences in emotional interpretation.

The main contribution of this work is to introduce a multimodal deep learning model for emotion recognition, by seamlessly incorporating both visual images and textual sources, accounting for cultural variations, fostering a more nuanced understanding of emotions. This fusion approach allows us to bridge cultural gaps, offering a promising avenue for enhanced accuracy and generalization in emotion recognition systems. The rest of the paper is organized as follows: Section II explains related works, section III introduces the proposed method, section IV discusses our experiments and results, and Section V concludes the paper.

II. Related Work

A. Multimodality

Previous studies have addressed emotion detection across diverse modalities, including texts, images, and videos. Within this domain, studies can be categorized based on the specific source of data they have worked on. Some studies have focused on emotion detection from textual data, where natural language processing techniques are employed to analyze the sentiment and emotional content of written text. These studies propose various machine learning algorithms and deep learning models to accurately classify emotions such as happiness, sadness, anger, and fear.

On the other hand, some studies have explored emotion detection from visual sources, such as images and videos. These studies employ computer vision techniques to extract visual features and analyze facial expressions, body language, and other visual patterns to infer emotions. Convolutional neural networks (CNNs) [12], [13], are commonly used in this context to achieve high accuracy in emotion classification. Furthermore, recent studies aim to combine both textual and visual information for a more comprehensive understanding of emotions. These studies leverage multimodal fusion techniques to integrate textual and visual features, enabling a more robust and accurate emotion detection system.

1) Text-based Emotion Detection: Related studies have addressed emotion detection from textual data, employing natural language processing techniques to semantically analyze texts. [14]–[16] have specifically focused on the analysis of tweets collected during SemEval2018 challenge [17], recognizing that tweets often provide valuable patterns for classifying thoughts and emotions. In a similar vein, [18], [19] detected emotions within datasets comprising textual data. Their efforts aim to harness the inherent emotional expressions within textual materials, contributing to a comprehensive analysis in the field of emotion detection.

2) Multi-source based Emotion Detection: While earlier studies concentrated solely on text sources, recent related works have embraced a more comprehensive approach by incorporating images alongside textual data. This multifaceted methodology aims to attain more generalized results in the analysis of emotions, recognizing the complementary nature of visual and textual patterns in understanding and categorizing emotional expressions. Notably, two inspiring papers in this field are CLIP [13] and ArtEmis [11]. CLIP utilizes a vision-language transformer architecture, combining a vision transformer (ViT) and a language transformer with a contrastive learning objective. The model learns a shared representation space for images and text, enabling semantic understanding across modalities. Trained on a large-scale dataset, CLIP's embeddings allow it to generalize to diverse tasks without task-specific training. Besides CLIP, ArtEmis curated a dataset by gathering images from renowned artwork available on WikiArt [20] and annotating them with English descriptions, enabling the detection of emotions from these images. CLIP has also introduced a state-of-the-art model for emotion detection based on images and texts. Expanding beyond text and images, some related studies have explored the inclusion of voice or dialogues to achieve a more comprehensive understanding of multimodal data [21], [22].

B. Multilinguality

Another approach to categorize related work on emotion detection is based on the language diversity of the texts [23]. Initially, studies have concentrated on analyzing emotions solely with monolingual English texts [24], [25]. However, the recognition of biases stemming from cultural differences has proved the necessity of balanced datasets that encompass diverse cultures [26], [27]. To facilitate the development of multilingual emotion detection models, there has been employment of techniques such as machine translation and cross-lingual transfer learning. These methods enable the transfer of knowledge and insights gained from one language to another, thereby enhancing the performance and generalization of the models [28].

Moreover, the availability of large-scale multilingual datasets has significantly contributed to the advancement of research in this area. These datasets encompass texts from various languages, enabling researchers to train and evaluate their models on a diverse range of linguistic data [29], [30]. By focusing on multilingual datasets, [28] aims to create more robust and unbiased emotion detection AI models that can be applied across different languages and cultures. This approach not only improves the accuracy and reliability of emotion detection models but also ensures their applicability in real-world scenarios involving diverse populations.

1) Monolinguals: This language-centric approach was evident in benchmark studies such as ArtEmis, which primarily concentrated on emotion detection within the English language context and corresponding images. Similarly, [16], [31] also dedicated their research to a single language.

2) multilinguals: In the domain of multilingual approaches to emotion extraction, a notable contribution is highlighted in [28] titled *ArtELingo*. This research serves as a pivotal foundation and inspiration for the current study. ArtELingo extends the ArtEmis dataset by including annotations in Chinese and Arabic, in addition to English. This augmentation significantly enhances the generalization of the model, emphasizing the importance of incorporating multiple languages in emotion extraction research.

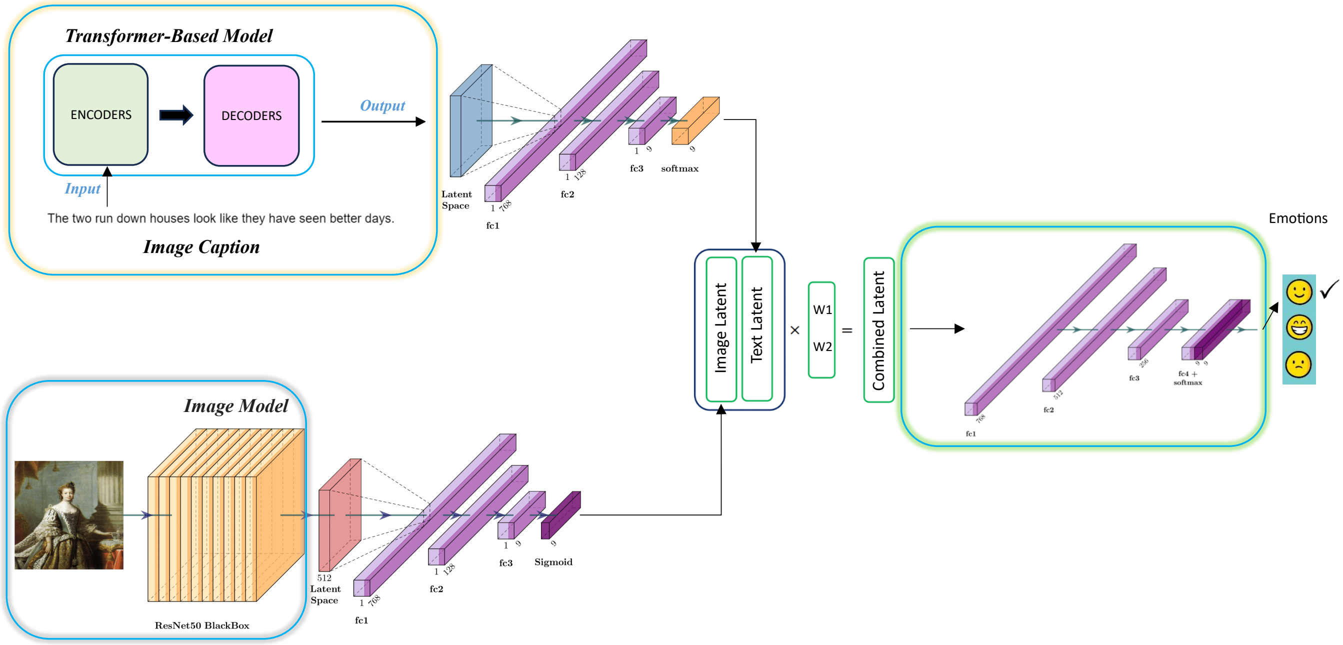


Fig. 1. The proposed architecture for culturally-aware emotion recognition.

This shift in focus aimed to address the diversity of languages and cultures, acknowledging that emotions are expressed differently across various linguistic contexts [19].

III. Proposed Method

Deep neural networks exhibit remarkable versatility, extending their utility beyond the scope of their primary training objectives. Within the realm of model development, a prevalent strategy involves leveraging pre-trained models from the same domain to facilitate the training of new models. To illustrate, a pre-trained model initially designed for an image classification task can be repurposed to detect objects in an image. Similarly, pre-trained text models find application in diverse natural language processing tasks such as sentiment analysis or text generation.

However, addressing multimodal tasks introduces a distinct challenge, necessitating data integration from multiple domains. In our research of emotion recognition within a multicultural context, the complexity arises from the interplay of images and captions. This scenario mandates the utilization of pre-trained models capable of effectively processing data from diverse domains, facilitating an integrated approach to address the intricacies inherent in multimodal tasks.

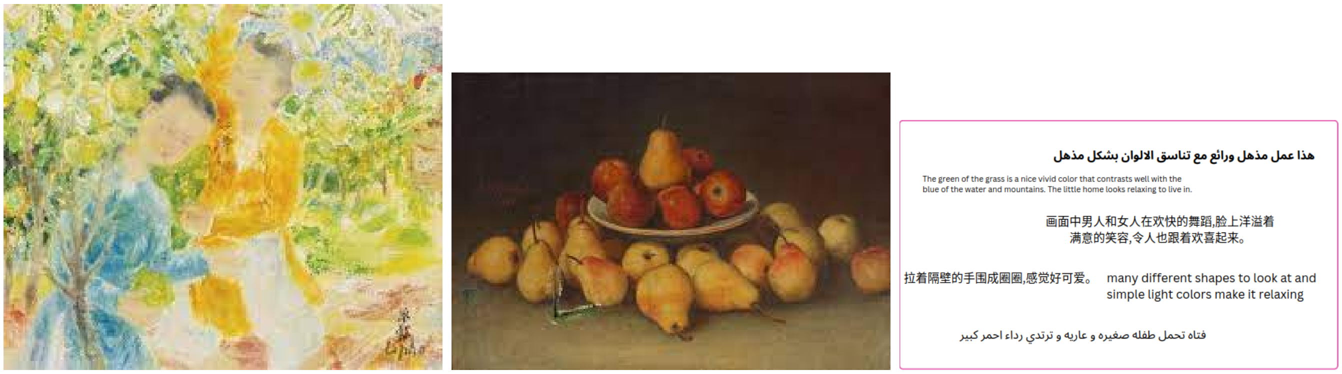
Effectively tackling the challenges inherent in this task necessitates a thorough comprehension of the inherent characteristics of both images and text, as well as the nuanced interplay of contextual features and concepts encapsulated in their respective captions. Mere reliance on features extracted from images and text proves insufficient; rather, the task mandates a capacity to seamlessly integrate these extracted features across both domains. Given the complexity of the problem, a multidimensional approach is essential, one that carefully considers the intricate nature of the data involved.

Similar to the model presented in [32], our proposed model adopts a similar architecture featuring two branches of encoders. These encoders are designed to effectively map both image and text data into their respective latent space. The methodology we propose involves training dedicated models for each data type (image or text) to ensure an accurate representation of their distinctive features.

The primary model we employ consists of three distinct sections, as depicted in Figure 1. The initial section, signified by the yellow box, embodies our text model, which is transformer-based. The prime objective of this section is to receive caption texts as input and extract latent features that represent the essential contextual information within the text. We train this model exclusively for emotion recognition, to acquire a representation that is effective for classification.

The second section, as illustrated by the blue box in Figure 1, refers to our image model, which is based on Convolutional Neural Networks (CNNs).

In the first step of our process, we attempt to adjust the latent spaces and domains. Adjusting features is crucial in determining the requisite amount of each latent for precise predictions regarding each input. To achieve this, we employ the linear combination technique, which yields a final latent representation suitable for our ultimate task. The primary advantage of this approach is that the combined latent captures content from textual and visual data. The process is illustrated in Figure 1, highlighted within the green box. This combination is defined as:



هذا عمل مذهل ورائع مع تناسق الالوان بشكل مذهل

The green of the grass is a nice vivid color that contrasts well with the blue of the water and mountains. The little home looks relaxing to live in.

画面中男人和女人在欢快的舞蹈,脸上洋溢着满意的笑容,令人也跟着欢喜起来。

拉着隔壁的手围成圈圈,感觉好可爱。

many different shapes to look at and simple light colors make it relaxing

فتاه تحمل طفله صغيره و عاريه و ترتدي رداء احمر كبير

Fig. 2. Some artworks and annotations from ArtELingo [20].

$$Combined\_latent = (w_1 \cdot Image_f) + (w_2 \cdot Text_f) \quad (1)$$

Where w_1 and w_2 are weighting factors determined through experimentation and validation, and $Image_f$ and $Text_f$ are features extracted from image and text, respectively. This linear combination ensures that the strengths of both modalities are appropriately weighed and contribute synergistically to the integrated representation. The whole combination process is presented in Figure 1.

It is crucial to understand that the entire process is trainable. The model can independently determine the weight allocation of each latent during the linear combination. This trainable aspect is essential because it permits the model to adjust and determine the importance of various attributes within the latent spaces of both textual and visual data. As a result, the model can accurately capture pertinent information for emotion recognition and enhance its performance based on the intricacies of the assigned task. This adaptability is a significant advantage that enables the model to perform well with varied inputs and tasks.

Finally, fully connected layers are utilized to distill emotional information from the input data. The model generates a probability score, ranging from one to zero, to signify the likelihood of specified emotions occurring in the input data.

IV. Experiments

The proposed framework evaluated multilingual annotated images. To this end, a publicly available dev set of the ArtELingo dataset, which contains 1.2 million annotations on 80k artworks of Wikiart, is employed. The Wikiart dataset included famous paintings classified into 27 different art styles, such as Impressionism, Romanticism, Symbolism, Minimalism, etc. For multilingualism, ArtELingo provided rich annotations in 3 languages, English, Arabic, and Chinese, adding almost 75k utterances to the monolingual and English-only Artemis dataset. Moreover, 4.8k annotations are provided in the test set for evaluation to measure how well-generalized models are based on their performance on unseen language formats. The tokens used in annotation have been extracted and provided by the dataset for further text processing. In addition to more than 15 annotations per painting on average, ArtELingo has a grounding emotion per annotation, all mapped to their corresponding key from 0 to 9. There are nine distinct emotional states, including positive emotions namely amusement, awe, contentment, and excitement, and negative emotions like anger, disgust, fear, and sadness. Additionally, there is a separate category for neutral or contradictory emotions, referred to as something else. Upon closer examination, the development set contained 1101841 annotations, of which 926055 were utilized for training, while the test and validation sets comprised 94083 and 46936 samples, respectively, some of the dataset items depicted in 2.

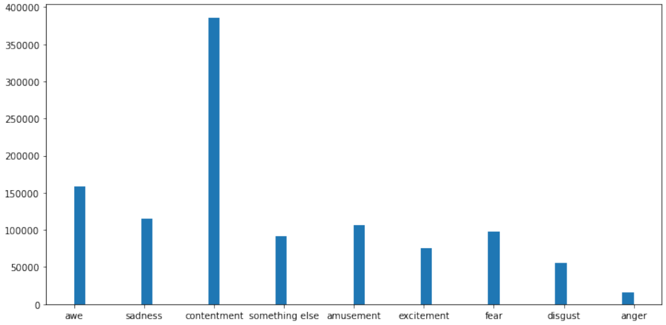


Fig. 3. The number of each emotion class over all annotations

On the other hand, the languages of annotations have a noticeable uniform distribution since each language has almost 30 percent of total annotations. English with 386412 and Chinese with 383266 annotations, followed by Arabic with 332163 annotations, are the languages used for annotating these pictures. Statistics from ArtELingo detail that positive emotions are the dominant cluster since they consist of 725069 artworks. In contrast, negative emotions were picked for 284756 paintings, and the rest were labeled with something else.

The proposed framework undergoes training via multi-label classification on the image model. Given that multiple labels may apply to each image, the model is trained correspondingly for multi-label classification. The image model backbone is

TABLE I
The evaluation of the Image models on the ArtELingo dataset

Model/Accuracy	Train ACC	Validation Acc
ResNet50	0.711386	0.698935688
VGG	0.717367	0.706483998
VIT	0.702534	0.687801932

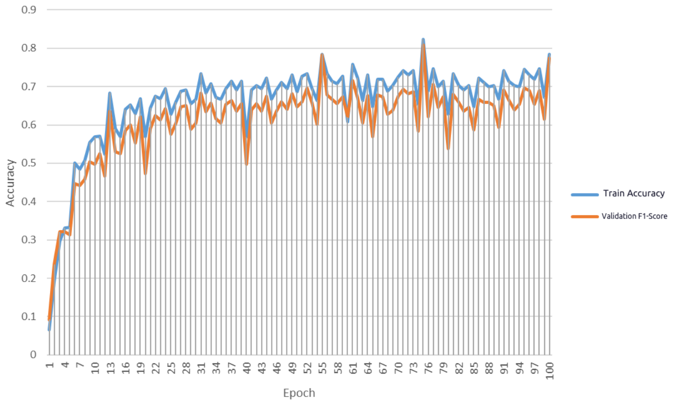


Fig. 4. The progress of training and evaluation of the proposed method

chosen from the VGG, ResNet50, and VIT models. The outcomes are detailed in table I. As illustrated in the table, the VGG model, pre-trained on the image-net dataset, achieved accurate convergence on the emotion recognition training set and received the highest evaluation score on the validation set. As a result, the VGG model trained on the emotion-recognition trainset is selected to extract visual features from images and adjusted within the final model.

The subsequent phase entails the development of a text classification model utilizing cutting-edge transformer models that deliver exceptional results in the field of Natural Language Processing. This model has been trained to understand emotions expressed in image captions. It works like a classifier, assigning each caption to a specific emotion. The number of emotions present in each data sample can be viewed in Figure 4. Following the model's training for emotion recognition captions, the feature extraction backbone is utilized to extract text features in the emotion recognition model.

The final step of the process involves combining the features obtained from both the image and text models. This combined feature set is then used to train a model for emotion recognition. The resulting model takes input from both the image and caption features to ensure accurate results. A diagram of the training and evaluation process can be found in Figure 5. The graph in the figure shows that the classifier model converged well on the merged image-text features and achieved an evaluation accuracy of nearly 80%.

V. Conclusion

In this paper, we aimed to develop a multi-modal deep learning model that integrates cultural awareness into emotion recognition from visual and textual data. We proposed a novel fusion approach that combines image and text features, accounting for cultural variations in emotional interpretation. We evaluated our model on the ArtELingo dataset, which contains image-comment pairs with Chinese, Arabic, and English annotations. Our results showed that our model achieved a promising accuracy of 80% on the test set in this domain. Our contribution is to introduce a more nuanced and comprehensive understanding of emotions, bridging cultural gaps and enhancing emotional awareness in human-computer communication. Our research has implications for developing services that can interact with humans effectively across different regions and languages. However, our research also has some limitations, such as the reliance on a single dataset, the lack of generalization to other data types, and the potential ethical issues of emotion recognition. Therefore, we suggest future research to explore other sources of data, such as speech, video, or physiological signals, to extend the applicability and robustness of our model.

References

[1] A. Saxena, A. Khanna, and D. Gupta, "Emotion recognition and detection methods: A comprehensive survey," *Journal of Artificial Intelligence and Systems*, vol. 2, no. 1, pp. 53–79, 2020.

[2] D. Goleman, "Emotional intelligence. why it can matter more than iq." *Learning*, vol. 24, no. 6, pp. 49–50, 1996.

[3] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE Access*, vol. 7, pp. 100 943–100 953, 2019.

[4] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Information Fusion*, vol. 59, pp. 103–126, 2020.

[5] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 112–118.

[6] K. Sailunaz, M. Dhaliwal, J. Rokne, and R. Alhajj, "Emotion detection from text and speech: a survey," *Social Network Analysis and Mining*, vol. 8, pp. 1–26, 2018.

[7] S. M. S. A. Abdullah, S. Y. A. Ameen, M. A. Sadeeq, and S. Zeebaree, "Multimodal emotion recognition using deep learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 02, pp. 52–58, 2021.

[8] W. Liu, J.-L. Qiu, W.-L. Zheng, and B.-L. Lu, "Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 2, pp. 715–729, 2021.

[9] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 02, 2020, pp. 1359–1367.

[10] E. Meyer, *The culture map: Breaking through the invisible boundaries of global business*. Public Affairs, 2014.

[11] P. Achlioptas, M. Ovsjanikov, K. Haydarov, M. Elhoseiny, and L. Guibas, "Artemis: Affective language for visual art," 2021.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, p. 84–90, may 2017. [Online]. Available: https://doi.org/10.1145/3065386

[13] C. Grover, I. D. Mastan, and D. Gupta, "Contextclip: Contextual alignment of image-text pairs on clip visual representations," in *Proceedings of the Thirteenth Indian Conference on Computer Vision, Graphics and Image Processing*, ser. ICVGIP'22. ACM, Dec. 2022. [Online]. Available: http://dx.doi.org/10.1145/3571600.3571653

[14] S. Hassan, S. Shaar, and K. Darwish, "Cross-lingual emotion detection," 2022.

[15] G. Chochlakis, G. Mahajan, S. Baruah, K. Burghardt, K. Lerman, and S. Narayanan, "Using emotion embeddings to transfer knowledge between emotions, languages, and annotation formats," 2023.

[16] ——, "Leveraging label correlations in a multi-label setting: A case study in emotion," 2023.

[17] F. Barbieri, J. Camacho-Collados, F. Ronzano, L. E. Anke, M. Ballesteros, V. Basile, V. Patti, and H. Saggion, "Semeval 2018 task 2: Multilingual emoji prediction," in *Proceedings of the 12th international workshop on semantic evaluation*, 2018, pp. 24–33.

[18] M. Li, F. Ding, D. Zhang, L. Cheng, H. Hu, and F. Luo, "Multi-level distillation of semantic knowledge for pre-training multilingual language model," 2022.

[19] E. Öhman, M. Pàmies, K. Kajava, and J. Tiedemann, "Xed: A multilingual dataset for sentiment analysis and emotion detection," 2020.

[20] S. Mohammad and S. Kiritchenko, "Wikiart emotions: An annotated dataset of emotions evoked by art," in *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*, 2018.

[21] V. Chudasama, P. Kar, A. Gudmalwar, N. Shah, P. Wasnik, and N. Onoe, "M2fnet: Multi-modal fusion network for emotion recognition in conversation," 2022.

[22] K. Haydarov, X. Shen, A. Madasu, M. Salem, L.-J. Li, G. Elsayed, and M. Elhoseiny, "Affective visual dialog: A large-scale benchmark for emotional reasoning based on visually grounded conversations," 2023.

[23] P. Heracleous and A. Yoneyama, "A comprehensive study on bilingual and multilingual speech emotion recognition using a two-pass classification scheme," *PLOS ONE*, vol. 14, no. 8, pp. 1–20, 08 2019. [Online]. Available: https://doi.org/10.1371/journal.pone.0220386

[24] S. Y. M. Lee and Z. Wang, "Multi-view learning for emotion detection in code-switching texts," in *2015 International Conference on Asian Language Processing (IALP)*, 2015, pp. 90–93.

[25] S. Peng, L. Cao, Y. Zhou, Z. Ouyang, A. Yang, X. Li, W. Jia, and S. Yu, "A survey on deep learning for textual emotion analysis in social networks," *Digital Communications and Networks*, vol. 8, no. 5, pp. 745–762, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352864821000833

[26] Y. Mohamed, F. F. Khan, K. Haydarov, and M. Elhoseiny, "It is okay to not be okay: Overcoming emotional bias in affective image captioning by contrastive data collection," 2022.

[27] A. Ye, S. Santy, J. D. Hwang, A. X. Zhang, and R. Krishna, "Cultural and linguistic diversity improves visual representations," 2023.

[28] Y. Mohamed, M. Abdelfattah, S. Alhuwaider, F. Li, X. Zhang, K. W. Church, and M. Elhoseiny, "Artelingo: A million emotion annotations of wikiart with emphasis on diversity over language and culture," 2022.

[29] M. Neumann and N. g. Thang Vu, "Cross-lingual and multilingual speech emotion recognition on english and french," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5769–5773.

[30] K. Becker, V. Moreira, and A. Santos, "Multilingual emotion classification using supervised learning: Comparative experiments," *Information Processing & Management*, vol. 53, pp. 684–704, 05 2017.

[31] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021.

[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017. [Online]. Available: https://arxiv.org/pdf/1706.03762.pdf