Transformers Beyond NLP

Cesar Bryam Rodriguez Aybar
Matrícula:A01795980
Instituto Tecnológico y de Estudios Superiores de Monterrey.
Monterrey, Mexico
A01795980@tec.mx

Christian Erick Mercado Flores
Matrícula:A00841954
Instituto Tecnológico y de Estudios Superiores de Monterrey.
Monterrey, Mexico
A00841954@tec.mx

Manuel Alejandro de Luis López
Matrícula:A00466615
Instituto Tecnológico y de Estudios Superiores de Monterrey.
State of Mexico, Mexico
A00466615@tec.mx

Carlos Ricardo Álvarez Pérez
Matrícula:A01796116
Instituto Tecnológico y de Estudios Superiores de Monterrey.
San Luis Potosí, Mexico
A01796116@tec.mx

Diego Andres Bernal Diaz
Matrícula:A01795975
Instituto Tecnológico y de Estudios Superiores de Monterrey.
Bogotá, Colombia
A01795975@tec.mx

***Abstract*—Transformers, initially conceived for natural language processing (NLP), have exhibited remarkable adaptability and efficacy across diverse domains, including computer vision, audio processing, and multi-modal analysis. In computer vision, models such as Vision Transformer (ViT) and Swin Transformer have surpassed traditional convolutional neural networks (CNNs) in tasks like image classification and object detection, leveraging self-attention mechanisms to capture intricate spatial relationships. Similarly, in audio processing, architectures like the Audio Spectrogram Transformer (AST) have achieved state-of-the-art performance in speech recognition and music information retrieval by effectively modeling complex acoustic patterns. Furthermore, Transformers have proven valuable in multi-modal applications, particularly in emotion recognition from audio-visual data, where their ability to integrate and process heterogeneous sensory inputs enhances classification robustness and accuracy. This paper reviews the transformative impact of Transformers beyond NLP, highlighting their versatility in addressing complex analytical challenges across multiple disciplines and solidifying their role as a cornerstone of modern artificial intelligence research.**

I. Introduction

The transformative potential of Transformers initially demonstrated in natural language processing (NLP), has spurred their widespread adoption across diverse fields, ranging from computer vision to audio processing and multi-modal learning. Departing from conventional sequential methods, Transformers leverage self-attention mechanisms to simultaneously capture intricate interdependencies within data sequences. This architectural innovation has enabled significant advancements in tasks such as image classification, object detection, semantic segmentation, speech recognition, audio classification, music generation, visual question answering (VQA), image captioning, and emotion recognition. Building on this foundation, the versatility and adaptability of Transformers have been further validated by models like BERT [1], which revolutionized language understanding, and Vision Transformers (ViTs) [2], which extended their utility to image analysis. Moreover, the development of efficient architectures like EfficientFormer [3], capable of achieving high performance with low latency on mobile devices, underscores the feasibility of deploying Transformers in resource-constrained environments. This paper aims to provide a comprehensive overview of the transformative impact of Transformers beyond NLP by focusing on three primary areas: (1) Computer Vision, encompassing image classification, object detection, and semantic segmentation; (2) Audio Processing, including speech recognition, audio classification, and music generation; and (3) Multi-Modal Learning, with an emphasis on VQA, image captioning, and emotion recognition.

The burgeoning interest in Transformers beyond NLP is reflected in the exponential growth of research publications in recent years. This surge in activity necessitates a comprehensive review to synthesize the diverse approaches, identify key trends, and highlight areas where further investigation is warranted. This paper aims to fill this gap by providing a structured and critical analysis of the existing literature, focusing on the specific adaptations and innovations that have enabled Transformers to excel in computer vision, audio processing, and multi-modal learning. By examining the strengths and limitations of current approaches, we hope to provide valuable insights for researchers and practitioners seeking to leverage the power of Transformers in these domains.

The subsequent sections will delve into each of these areas, providing a detailed analysis of the current state-of-the-art while highlighting key challenges and opportunities for future research. By exploring these diverse applications, we seek to provide a holistic understanding of the versatility and potential of Transformers in addressing complex analytical challenges across multiple disciplines.

II. Literature Review

A. Vision Transformers

The remarkable success of Transformers in natural language processing (NLP) has spurred a paradigm shift in computer vision, leading to the development of Vision Transformers (ViTs) and related architectures that challenge the dominance of convolutional neural networks (CNNs). As Jalil et al. [4] note in their comprehensive survey, ViTs offer a compelling alternative to CNNs for various computer vision applications, including image classification, object detection, image segmentation, and more as we can see in Figure 1. This transition is driven by the inherent limitations of CNNs in capturing long-range dependencies and global context, which are crucial for understanding complex visual scenes [5].

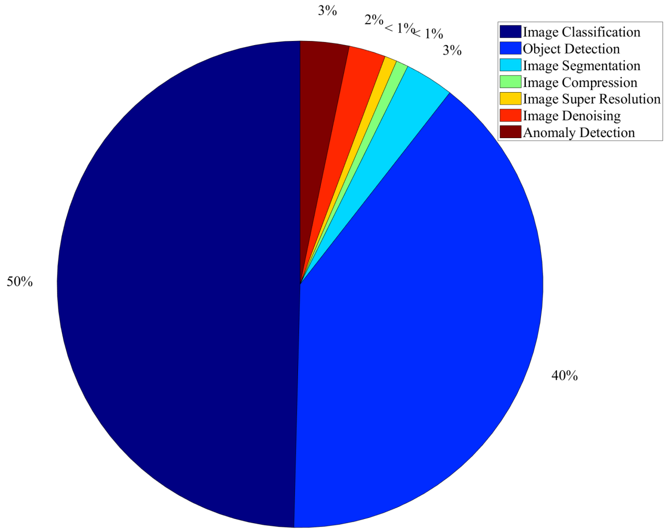

Fig. 1. Use of ViTs for CV applications. Source: [4]

Unlike CNNs, which rely on local receptive fields and hierarchical feature extraction, ViTs leverage self-attention mechanisms to simultaneously process all parts of an image, capturing intricate interdependencies between different regions [2]. This approach allows ViTs to model global context more effectively, leading to improved performance on tasks that require understanding the relationships between distant objects or image regions. As Sagri et al. [2] highlight in their work on applying Transformers to the game of Go, the ability to model long-range dependencies is crucial for strategic tasks where decisions depend on understanding the global state of the environment.

However, the transition to ViT-based architectures is not without its challenges. One significant limitation is the computational complexity of the self-attention mechanism, which scales quadratically with the number of input tokens. This can be particularly problematic for high-resolution images, where the number of tokens can be very large. To address this challenge, researchers have developed various techniques for reducing the computational cost of self-attention, such as patch embedding and efficient Transformer architectures like EfficientFormer [6]. Sagri et al. [2] observed that while the latency characteristics of Transformers and Residual Networks are similar, Transformers tend to have significantly higher memory usage, highlighting the need for further optimization.

Despite these challenges, ViTs have demonstrated state-of-the-art performance on a wide range of computer vision tasks, often surpassing the accuracy of traditional CNN-based models. Moreover, ViTs offer several advantages over CNNs, including improved robustness to adversarial attacks and greater interpretability [7]. Alahmadi highlights the potential of ViTs for computer-aided diagnosis (CAD) in the context of ovarian cancer, demonstrating that ViTs can achieve high accuracy in classifying ovarian cancer subtypes based on histopathological images. Furthermore, Alahmadi [7] emphasizes the importance of interpretability in medical applications, showcasing how techniques like LIME can be used to visualize the regions of the image that are most important for the ViT's decision-making process.

As ViTs automate tasks that previously required human intervention, such as image classification and object detection, they also create new opportunities for skilled professionals in areas such as deep learning, data engineering, and AI architecture. The key to mitigating the potential negative effects of automation lies in investing in education and training, preparing the workforce for the emerging roles that AI is creating.

In conclusion, Vision Transformers represent a significant advancement in computer vision, offering a compelling alternative to CNNs for a wide range of tasks. While challenges remain, ongoing research efforts are focused on addressing these limitations and unlocking the full potential of ViTs for real-world applications. The successful application of ViTs to complex problems like ovarian cancer diagnosis [7] underscores their transformative potential and highlights the importance of continued research in this area.

B. Transformers in Audio Processing

Recently, Transformers have been used in different architectures for various audio processing and detection tasks. Transformers have emerged as a powerful technique in audio detection and processing, offering several notable advantages. For instance, their self-attention mechanism allows Transformers to capture relationships between different parts of the audio signal, thereby understanding both local and global context [8].

These Transformers models have demonstrated exceptional performance in tasks such as speech recognition, speaker identification, emotion recognition, and audio signal classification. This way, for this last, audio-oriented Transformer models such as Wav2vec, or HUBERT are used [9]. Wav2vec operates by employing a multi-layered Transformer network, which consists of self-attention mechanisms that capture contextual dependencies in the input audio. In contrast, By incorporating self-attention mechanisms, HUBERT can effectively model long-range dependencies within audio signals, enabling a comprehensive understanding of the underlying acoustic features [10].

Regarding audio-emotion recognition, the Transformer framework, known for its ability to model long-range dependencies, has been introduced in the field [11]. For example, Wang et al. [12] propose a hierarchical Swin Transformer-based architecture to aggregate multi-scale emotional features, demonstrating the potential of Transformer frameworks in capturing emotion-related information.

As He et al. [13] demonstrate, combining multi-dimensional attention mechanisms, dual-stream fusion networks, and multi-scale Transformer frameworks effectively captures temporal, spatial, and channel dependencies, enhancing emotion recognition performance. Unlike most models that rely on a single type of input feature, the latter employs a cross-attention Transformer to integrate three acoustic features—raw waveforms, spectrograms, and MFCCs—showing superior performance compared to single-feature approaches.

In terms of practical applications, audio generation, and transfer learning have presented opportunities for leveraging the strengths of transformers in music creation tools. Xu proposed a semantic-based sequence-to-music transformer framework, which allows for an advanced understanding of musical structures by operating on elemental notes, facilitating more intentional compositions that resonate with specific themes [14]. This illustrates how transformers can do more than just analyze; they can create music that adheres to specified emotional and contextual guidelines.

Moreover, the synthesis and restoration of traditional ethnic musical instruments using transformer networks have advanced efforts to preserve cultural soundscapes. Mengmeng et al. detailed a timbre synthesis approach using transformers, enabling the reconstruction of intricate sound profiles associated with traditional instruments by analyzing time-frequency data [15]. This effort aids in the preservation of musical heritage and exemplifies the role of transformers in enhancing audio quality and clarity when representing complex sound forms.

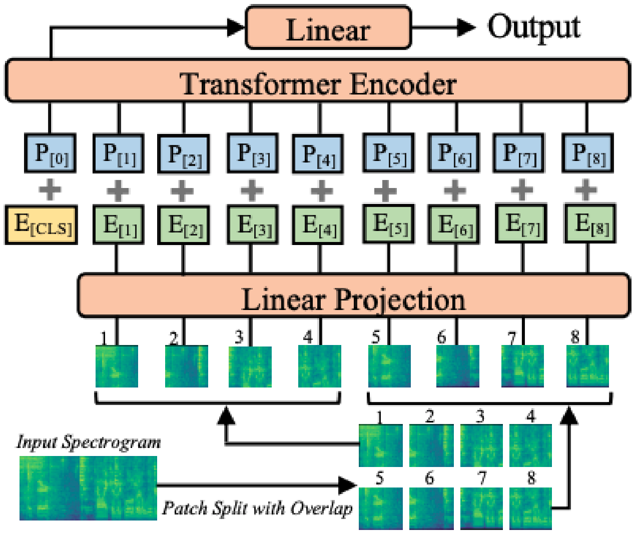

Fig. 2. The Audio Spectogram Transformer Mode. Source: [19]

The Audio Spectrogram Transformer (AST) is the model that made a shift in audio processing by leveraging the Transformer model instead of traditional convolutional neural networks (CNNs)[16]. Historically, CNNs have dominated audio classification due to their ability to learn spatially local and translationally equivariant features from spectrogram representations of audio signals [17]–[20]. However, CNNs have limitations in capturing long-range dependencies, leading to the development of hybrid models that incorporate self-attention mechanisms to enhance global feature extraction [21], [22]. To address this, AST is designed as an attention-based model, effectively eliminating the dependency on convolutions and instead using self-attention mechanisms throughout its architecture [23].

The proposed audio spectrogram transformer (AST) architecture in Figure 2 is formed as follows: The 2D audio spectrogram is split into a sequence of 16×16 patches with overlap, and then linearly projected to a sequence of 1-D patch embeddings. Each patch embedding is added with a learnable positional embedding. An additional classification token is prepended to the sequence. The output embedding is input to a Transformer, and the output of the classification token is used for classification with a linear layer.

The core of AST's application lies in its ability to process spectrograms as a sequence of patches, similar to how the Vision Transformer (ViT) handles images [20], [24]. The spectrogram is divided into overlapping patches, which are then projected onto embeddings and processed using a Transformer encoder. A critical innovation in AST is its ability to take advantage of the pre-training of ImageNet-trained Vision Transformer models [19]. This cross-modal transfer learning is possible because spectrograms and images share similar two-dimensional structures, allowing AST to inherit learned feature representations from visual data [20].

One of the primary advantages of AST is its ability to handle variable-length audio inputs, unlike CNN-based models that often require architectural modifications for different input sizes [24]. Furthermore, AST achieves state-of-the-art results in multiple audio classification benchmarks, including AudioSet, ESC-50, and Speech Commands [22], [23], [25]. For example, AST achieves 0.485 mean average precision (mAP) in AudioSet, outperforming previous CNN-attention models, and reaches 95. 6% precision in ESC-50, highlighting its robustness across different datasets [23].

The application of the Transformer model in AST provides a more streamlined and computationally efficient approach compared to CNN-attention hybrids. AST benefits from a simpler architecture with fewer parameters and faster convergence during training [23]. This efficiency is further enhanced by the ImageNet pretraining strategy, which significantly improves performance, especially when training data are limited [19], [20].

In terms of practical applications, the ability of AST to handle various types of audio data, including environmental sounds, speech commands, and general audio event classification, makes it a versatile model for tasks in speech recognition, automated audio tagging, and sound event detection [35]. Furthermore, its attention-based mechanism enables better interpretability in understanding which parts of the spectrogram contribute the most to classification decisions, a feature particularly useful in domains requiring explainability [19], [23].

Overall, the introduction of AST represents a shift in audio classification methodologies, demonstrating that purely attention-based architectures can surpass CNNs and CNN-attention hybrids in performance while offering greater flexibility and efficiency [19].

C. Transformers in Multi-Modal Learning

The introduction of the Transformer model has significantly impacted multimodal learning, particularly in integrating multiple data modalities such as audio, video, and text. One of the foundational works in multimodal deep learning was introduced by Ngiam et al. [26], who proposed a deep learning framework to learn features from multiple modalities, demonstrating how cross-modal learning can enhance single-modality representations. Their work focused on integrating audio and video inputs for speech classification tasks, leveraging deep autoencoders to extract shared representations. Prior to the adoption of Transformer-based models, deep learning approaches such as restricted Boltzmann machines (RBMs) and deep autoencoders were commonly used to learn shared multimodal representations [27].

The Transformer model, originally introduced by Vaswani et al. [28], has since revolutionized multimodal learning by providing an architecture that effectively captures long-range dependencies and facilitates cross-modal interactions. Unlike earlier deep learning methods that relied on unsupervised pretraining for feature extraction, Transformers allow for direct sequence-to-sequence modeling, making them highly effective in fusing information across modalities. Recent advancements have leveraged Transformers to integrate different sensory inputs, improving applications such as speech recognition, image captioning, and audiovisual understanding [29]. For instance, Dosovitskiy et al. [24] demonstrated that Vision Transformers (ViTs) could be successfully applied to image classification, inspiring subsequent multimodal adaptations.

A key advantage of using Transformers in multimodal learning is their ability to handle complex relationships between different modalities. Traditional deep networks struggled with learning representations that could generalize across varied inputs. However, Transformer-based models, such as multimodal Transformers, effectively bridge this gap by learning joint representations across modalities through self-attention mechanisms [30]. This approach has been applied in tasks such as visual question answering [31], video understanding [32], and speech translation, where the model simultaneously processes audio and visual data to generate highly accurate predictions [33]. Gong et al. [19] demonstrated the effectiveness of Transformer-based architectures for audio processing, highlighting their superiority over conventional convolutional neural networks in tasks like environmental sound classification.

The application of Transformers in multimodal learning extends beyond traditional classification tasks. They have enabled advancements in cross-modal retrieval, where a query in one modality (e.g., text) retrieves relevant content in another (e.g., image or video) [34]. Additionally, Transformers have improved multimodal sentiment analysis, where textual and visual cues are combined to infer emotional states, outperforming previous recurrent neural network-based approaches [35]. The recent work by Wang et al. [36] on hierarchical Transformer architectures for speech emotion recognition further underscores the effectiveness of Transformers in capturing cross-modal dependencies.

In conclusion, the transition from early multimodal deep learning frameworks to Transformer-based architectures marks a significant advancement in the field. Transformers have revolutionized feature fusion across modalities while enhancing the interpretability and robustness of multimodal models. Their applications continue to expand across various domains, consistently achieving state-of-the-art performance in multimodal tasks.

III. Future Directions

Future research in Vision Transformers (ViTs) is anticipated to focus on several key areas to address current challenges and enhance their effectiveness in various applications. A fundamental area of concern is the complexity of the self-attention mechanism, which exhibits quadratic complexity, limiting the application of transformers to high-resolution images [37]. Researchers are likely to explore more efficient attention mechanisms or hybrid models that synergistically integrate convolutional operations with attention layers. Such approaches not only aim to minimize computational costs but also strive to retain or improve performance metrics that define success within the realm of computer vision [38].

Another critical direction for future ViT research is the enhancement of interpretability and explainability. As ViTs gain traction in high-stakes fields such as medical imaging and autonomous systems, the ability to comprehend decision-making processes becomes increasingly urgent. Techniques like attention visualization and model-agnostic interpretation approaches such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (Shapley Additive explanations) will likely be refined to provide greater insight into how ViTs generate their predictions. This interpretability is essential to fostering trust and understanding between these models and their human operators, especially in applications that significantly impact health and safety [39].

In line with the operational demands of modern applications, research into the efficiency and scalability of audio transformers is essential. Current architectures, like Attention-based models, while powerful, tend to be highly computationally expensive, posing challenges for deployment on resource-constrained devices. Hence, the significance of model compression techniques—including knowledge distillation, quantization, and pruning—has gained traction. These methods aim to reduce the model size and computational burden while maintaining, if not improving, performance levels [40], [41]. Additionally, recent advancements in understanding and implementing deep learning model-compression strategies specifically for audio classification on edge devices have highlighted the necessity for adaptive methods that fit into real-world computational frameworks [40]. The importance of tailored architectures is further emphasized, providing the foundation for scalable implementations that can meet diverse application needs in real-time environments [41].

Regarding multimodal learning, we recognize that the exploration of innovative mechanisms for feature extraction and fusion across different modalities is crucial. Recent studies have indicated that networks employing attention mechanisms exhibit superior performance in extracting features from diverse inputs, supporting the notion that integrating self-attention can significantly boost the performance of multimodal systems [42]. Moreover, hybrid architectures that combine various modalities, such as visual data and textual information, can yield richer feature representations. This is the case, as we have previously discussed in multimodal sentiment analysis research, which asserts the efficacy of multi-layer attention mechanisms for capturing complex interrelations between modalities [43]. Additionally, the application of advanced task-specific learning strategies, such as multi-task learning frameworks, contributes to the robustness of models tasked with processing multimodal data.

IV. Conclusions

The Transformer model has completely reshaped artificial intelligence, driving breakthroughs in language understanding, computer vision, and multimodal learning. Unlike earlier AI models that struggled with long-range dependencies and required sequential processing like expert systems, Transformers introduced self-attention, enabling more efficient and powerful learning across vast datasets. This shift has fueled advancements in machine translation, speech recognition, and cross-modal applications like image and video processing, captioning and understanding.

One of the most exciting impacts of Transformers is their ability to integrate different types of information—text, images, and audio—creating AI that understands the world more holistically. However, their high computational demands have sparked ongoing efforts to make them more efficient. Researchers are developing lighter architectures and adaptive mechanisms to reduce energy consumption while maintaining high performance.

Looking ahead, Transformers will continue evolving, driving AI towards more intelligent, context-aware, and adaptable systems through increasingly sophisticated and flexible architectures. From healthcare to autonomous systems, their influence is only growing. While challenges like scalability and ethical concerns remain, the future of AI is being shaped by these powerful models, unlocking possibilities we are only beginning to explore.

References

[1] I. Tenney, D. Das, and E. Pavlick, *Bert rediscovers the classical nlp pipeline*, 2019. arXiv: 1905.05950 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1905.05950.

[2] A. Sagri, T. Cazenave, J. Arjonilla, and A. Saffidine, *Vision transformers for computer go*, 2023. arXiv: 2309.12675 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2309.12675.

[3] J. Breen, K. Allen, K. Zucker, G. Hall, N. Ravikumar, and N. M. Orsi, *Predicting ovarian cancer treatment response in histopathology using hierarchical vision transformers and multiple instance learning*, 2023. arXiv: 2310.12866 [eess.IV]. [Online]. Available: https://arxiv.org/abs/2310.12866.

[4] S. Jamil, J. P. Md, and K. Oh-Jin, "A comprehensive survey of transformers for computer vision," English, *Drones*, vol. 7, no. 5, p. 287, 2023, Copyright - © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/). Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License; Last updated - 2024-10-16. [Online]. Available: https://www.proquest.com/scholarly-journals/comprehensive-survey-transformers-computer-vision/docview/2819433642/se-2.

[5] B. Heo, S. Yun, D. Han, S. Chun, J. Choe, and S. J. Oh, *Rethinking spatial dimensions of vision transformers*, 2021. arXiv: 2103.16302 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2103.16302.

[6] Y. Li, G. Yuan, Y. Wen, *et al.*, *Efficientformer: Vision transformers at mobilenet speed*, 2022. arXiv: 2206.01191 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2206.01191.

[7] A. Alahmadi, "Towards ovarian cancer diagnostics: A vision transformer-based computer-aided diagnosis framework with enhanced interpretability," *Results in Engineering*, vol. 23, p. 102 651, 2024. DOI: 10.1016/j.rineng.2024.102651.

[8] K. Zaman, K. Li, M. Şah, C. Direkoğlu, S. Okada, and M. Unoki, "Transformers and audio detection tasks: An overview," *Digital Signal Processing*, vol. 158, p. 104 956, 2025. DOI: 10.1016/j.dsp.2024.104956.

[9] Ş. S. Çalık, A. Küçükmanişa, and Z. H. Kilimci, "A novel framework for mispronunciation detection of arabic phonemes using audio-oriented transformer models," *Applied Acoustics*, vol. 215, p. 109 711, 2024. DOI: 10.1016/j.apacoust.2023.109711.

[10] A. Chakhtouna, S. Sekkate, and A. Adib, "Unveiling embedded features in wav2vec2 and hubert msodels for speech emotion recognition," *Procedia Computer Science*, vol. 232, pp. 2560–2569, 2024. DOI: 10.1016/j.procs.2024.02.074.

[11] Z. Sun, H. Liu, H. Li, Y. Li, and W. Zhang, "Averformer: End-to-end audio-visual emotion recognition transformer framework with balanced modal contributions," *Digital Signal Processing*, vol. 161, p. 105 081, 2025. DOI: 10.1016/j.dsp.2025.105081.

[12] Y. Wang, C. Lu, H. Lian, *et al.*, *Speech swin-transformer: Exploring a hierarchical transformer with shifted windows for speech emotion recognition*, 2024. arXiv: 2401.10536 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2401.10536.

[13] Y. He, N. Minematsu, and D. Saito, "Multiple acoustic features speech emotion recognition using cross-attention transformer," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10095777.

[14] Y. Xu, "Enhancing music generation with a semantic-based sequence-to-music transformer framework," *International Journal on Semantic Web and Information Systems*, vol. 20, pp. 1–19, 1 2024. DOI: 10.4018/ijswis.343491.

[15] C. Mengmeng, Y. Xiang, and C. Xiong, "Synthesis and restoration of traditional ethnic musical instrument timbres based on time-frequency analysis," *Traitement Du Signal*, vol. 41, pp. 1063–1072, 2 2024. DOI: 10.18280/ts.410247.

[16] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," in *The Handbook of Brain Theory and Neural Networks*. Cambridge, MA, USA: MIT Press, 1998, pp. 255–258, ISBN: 0262511029.

[17] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, *Panns: Large-scale pretrained audio neural networks for audio pattern recognition*, 2020. arXiv: 1912.10211 [cs.SD]. [Online]. Available: https://arxiv.org/abs/1912.10211.

[18] Y. Gong, Y.-A. Chung, and J. Glass, "Psla: Improving audio tagging with pretraining, sampling, labeling, and aggregation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3292–3306, 2021, ISSN: 2329-9304. DOI: 10.1109/taslp.2021.3120633. [Online]. Available: http://dx.doi.org/10.1109/TASLP.2021.3120633.

[19] Y. Gong, Y.-A. Chung, and J. Glass, *Ast: Audio spectrogram transformer*, 2021. arXiv: 2104.01778 [cs.SD]. [Online]. Available: https://arxiv.org/abs/2104.01778.

[20] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, *Training data-efficient image transformers and distillation through attention*, 2021. arXiv: 2012.12877 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2012.12877.

[21] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[22] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780. DOI: 10.1109/ICASSP.2017.7952261.

[23] P. Warden, *Speech commands: A dataset for limited-vocabulary speech recognition*, 2018. arXiv: 1804.

03209 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1804.03209.

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, *An image is worth 16x16 words: Transformers for image recognition at scale*, 2021. arXiv: 2010.11929 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2010.11929.

[25] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM International Conference on Multimedia*, ser. MM '15, Brisbane, Australia: Association for Computing Machinery, 2015, pp. 1015–1018, ISBN: 9781450334594. DOI: 10.1145/2733373.2806390. [Online]. Available: https://doi.org/10.1145/2733373.2806390.

[26] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ser. ICML'11, Bellevue, Washington, USA: Omnipress, 2011, pp. 689–696, ISBN: 9781450306195.

[27] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006. DOI: 10.1162/neco.2006.18.7.1527.

[28] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, *Attention is all you need*, 2023. arXiv: 1706.03762 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1706.03762.

[29] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, L. Màrquez, C. Callison-Burch, and J. Su, Eds., Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 1412–1421. DOI: 10.18653/v1/D15-1166. [Online]. Available: https://aclanthology.org/D15-1166/.

[30] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "Untitled," *Proceedings of the 2019 Conference of the North*, 2019. DOI: 10.18653/v1/n19-1423.

[31] H. Tan and M. Bansal, *Lxmert: Learning cross-modality encoder representations from transformers*, 2019. arXiv: 1908.07490 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1908.07490.

[32] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, *Videobert: A joint model for video and language representation learning*, 2019. arXiv: 1904.01766 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1904.01766.

[33] D. Bahdanau, K. Cho, and Y. Bengio, *Neural machine translation by jointly learning to align and translate*, 2016. arXiv: 1409.0473 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1409.0473.

[34] X. Li, X. Yin, C. Li, *et al.*, *Oscar: Object-semantics aligned pre-training for vision-language tasks*, 2020. arXiv: 2004.06165 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2004.06165.

[35] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, *Multimodal sentiment analysis using hierarchical fusion with context modeling*, 2018. arXiv: 1806.06228 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1806.06228.

[36] Y. Wang, C. Lu, H. Lian, *et al.*, *Speech swin-transformer: Exploring a hierarchical transformer with shifted windows for speech emotion recognition*, 2024. arXiv: 2401.10536 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2401.10536.

[37] M. Zhang, R. Wang, J. Yang, and L. Xue, "Efficient vision transformer for dynamic embedding of multiscale features," *Third International Conference on Machine Vision, Automatic Identification, and Detection (MVAID 2024)*, p. 24, 2024. DOI: 10.1117/12.3035520.

[38] P. Wang, X. ZHANG, Y. Zhao, Y. LI, K. XU, and S. ZHAO, "Analysis of blood cell image recognition methods based on improved cnn and vision transformer," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E107.A, pp. 899–908, 6 2024. DOI: 10.1587/transfun.2023eap1056.

[39] M. L. Abimouloud, K. Bensid, M. Elleuch, O. Aiadi, and M. Kherallah, "Vision transformer-convolution for breast cancer classification using mammography images: A comparative study," *International Journal of Hybrid Intelligent Systems*, vol. 20, pp. 67–83, 2 2024. DOI: 10.3233/his-240002.

[40] A. Mou and M. Milanova, "Performance analysis of deep learning model-compression techniques for audio classification on edge devices," *Sci*, vol. 6, p. 21, 2 2024. DOI: 10.3390/sci6020021.

[41] Z. Wang, H. Liu, H. Coppock, B. W. Schuller, and M. D. Plumbley, "Neural compression augmentation for contrastive audio representation learning," *Interspeech 2024*, pp. 3335–3339, 2024. DOI: 10.21437/interspeech.2024-1156.

[42] E. Warner, J. Lee, W. Hsu, *et al.*, "Multimodal machine learning in image-based and clinical biomedicine: Survey and prospects," *International Journal of Computer Vision*, vol. 132, pp. 3753–3769, 9 2024. DOI: 10.1007/s11263-024-02032-8.

[43] L. Ma, J. Li, D. Shao, J. Yan, J. Wang, and Y. Yan, "Bcd-mm: Multimodal sentiment analysis model with dual-bias-aware feature learning and attention mechanisms," *IEEE Access*, vol. 12, pp. 74 888–74 902, 2024. DOI: 10.1109/access.2024.3405586.