International Journal of Emerging Research in Engineering and Technology
Pearl Blue Research Group| Volume 1, Issue 2, 1-8, 2020
ISSN: 3050-922X | https://doi.org/10.63282/3050-922X/IJERET-V1I2P101

Original Article

Advancements in Large-Scale Transformer Architectures for Multimodal AI Integration

Musthafa Ali
Technical Analyst, TCS, Mumbai, India.

Abstract - *Large-scale transformer architectures have revolutionized the field of artificial intelligence (AI), particularly in natural language processing (NLP) and computer vision (CV). The recent advancements in these architectures have enabled the integration of multimodal data, leading to more robust and versatile AI systems. This paper provides a comprehensive overview of the latest developments in large-scale transformer architectures, focusing on their application in multimodal AI integration. We discuss the theoretical foundations, key architectural innovations, and practical applications. Additionally, we present a detailed analysis of the challenges and future directions in this rapidly evolving field. The paper includes empirical evaluations, algorithmic descriptions, and comparative studies to highlight the effectiveness of these models.*

Keywords *- Transformer Architecture, Self-Attention, Multimodal AI, Vision Transformer, Cross-Attention, Pretraining, Fine-Tuning, Deep Learning, Neural Networks, Data Integration.*

1. Introduction

The advent of transformer architectures has marked a significant milestone in the field of artificial intelligence (AI), particularly in the domains of natural language processing (NLP) and computer vision (CV). Introduced by Vaswani et al. in 2017, transformers have revolutionized the way we handle sequential data. Unlike traditional recurrent neural networks (RNNs), which process data sequentially and can struggle with long-range dependencies, transformers excel by leveraging self-attention mechanisms. These mechanisms allow the model to weigh the importance of different parts of the input data, effectively capturing long-range dependencies without the limitations of sequential processing. As a result, transformers have become the de facto standard for tasks such as language translation, text summarization, and question-answering, where understanding the context over long sequences is crucial.

However, the real breakthrough has been the extension of these models to handle multimodal data, integrating textual, visual, and other forms of information. This extension has opened up new possibilities in AI, enabling more sophisticated and contextually rich models. For instance, multimodal transformers can process and correlate information from text and images simultaneously, which is invaluable for tasks such as image captioning, visual question answering, and cross-modal retrieval. By seamlessly integrating multiple data types, these models can provide a more comprehensive understanding of complex real-world scenarios, leading to advancements in areas like healthcare, where combining textual medical records with imaging data can improve diagnostic accuracy. The ability to parallelize training further enhances the efficiency and scalability of these models, making them more practical for large-scale applications and real-time processing. Overall, the development of multimodal transformers represents a significant leap forward in the capabilities of AI systems, paving the way for more integrated and intelligent technologies.

2. Theoretical Foundations

2.1. Transformer Architecture

The transformer architecture, introduced by Vaswani et al. (2017), represents a paradigm shift in neural network models, particularly for sequential data processing. Unlike traditional Recurrent Neural Networks (RNNs) that process sequences step-by-step, transformers leverage a self-attention mechanism that allows each element in the sequence to attend to all other elements, irrespective of their position. This enables the model to capture long-range dependencies more efficiently than RNNs, which struggle with maintaining contextual information over longer sequences due to their inherent sequential processing nature. The core component of a transformer is the self-attention mechanism, which computes a weighted representation of all positions in the input sequence for each position, allowing the model to simultaneously focus on different parts of the input. This parallelized processing enables much faster training compared to RNNs. For example, self-attention in transformers is typically implemented through a mechanism where each token in the input sequence computes a set of attention weights with respect to all other tokens, thereby creating a dynamic relationship that changes depending on the context.

Another important element of transformer models is the use of Feed-Forward Neural Networks (FFNs). After the self-attention operation, the output for each sequence element is passed through a position-wise feed-forward neural network. Each position is processed independently, which introduces a level of flexibility and parallelism. The use of FFNs allows the model to learn non-linear transformations on the input data, further enriching its capacity to model complex relationships between sequence elements. To ensure stability during training and to accelerate the learning process, layer normalization is applied in transformers. Layer normalization standardizes the input to each layer, mitigating issues related to vanishing gradients and improving the model's ability to converge. This makes transformers particularly effective when dealing with deep architectures, where deeper layers could otherwise slow down or hinder the learning process due to unstable gradients.

2.2. Multimodal Data Integration

Multimodal data integration is a key challenge in modern AI systems that require understanding and synthesizing information from multiple data types such as text, images, and audio. Traditional models, often specialized in a single modality (e.g., NLP models for text or CNNs for images), struggle when tasked with processing inputs that span multiple modalities. The integration of multimodal data aims to create a unified representation that reflects the relationships and interactions between diverse input types, enabling AI systems to interpret complex, real-world scenarios more effectively.

Transformers, with their self-attention mechanisms, provide an ideal framework for multimodal data integration. One of the key innovations in this regard is the cross-attention mechanism, which allows the transformer to attend to one modality while processing another. For example, in a task that combines text and images, a cross-attention mechanism can align the relevant parts of the image with the corresponding parts of the text, ensuring that the model comprehends how the two modalities relate to one another. This dynamic interaction between modalities allows the model to build a more accurate and meaningful representation that reflects the combined information from both modalities.

Another crucial technique in multimodal integration is the use of multimodal embeddings. These embeddings map each modality's data (e.g., words, images, sounds) into a shared latent space, where similar semantic relationships between modalities are preserved. The ability to project different types of data into a common space is essential for creating coherent multimodal representations, where relationships such as the visual description of an image and the corresponding text can be processed together. This shared embedding space provides a powerful foundation for downstream tasks such as visual question answering (VQA) or image captioning.

The final step in integrating multimodal data is achieved through fusion layers, which combine the information from each modality into a single unified representation. These fusion layers typically operate on the outputs of the modality-specific encoders, allowing the model to synthesize the individual features extracted from text, images, or audio. Fusion strategies can vary, ranging from simple concatenation to more sophisticated methods such as weighted sum or attention-based fusion, depending on the complexity of the task. By combining multimodal representations, fusion layers enable the model to make predictions or generate outputs that account for the relationships across all involved modalities.

3. Methodology

3.1. Vision Transformers (ViTs)

Vision Transformers (ViTs) represent a significant shift in the way image data is processed by leveraging the transformer architecture, traditionally used in natural language processing, for vision tasks. Unlike conventional Convolutional Neural Networks (CNNs), which process images by applying convolutions and pooling operations over grids of pixels, ViTs treat images as sequences of smaller, non-overlapping patches. This novel approach allows ViTs to take advantage of the transformer’s strength in handling long-range dependencies, enabling them to model global context across the entire image, rather than being restricted to local patterns as in CNNs.

3.1.1. Architecture

The architecture of ViTs begins with patch embedding, where an image is divided into fixed-size patches. Each patch is then flattened into a vector and linearly projected into a higher-dimensional space. These projected vectors form a sequence, similar to token embeddings in natural language models. Next, positional encodings are added to the patch embeddings to ensure that the transformer model can discern the spatial arrangement of the patches within the original image. This is crucial because, unlike CNNs, which inherently capture spatial relationships through local convolutions, transformers require explicit information about the position of patches to retain spatial coherence. The sequence of patch embeddings, enriched with positional information, is then fed into a transformer encoder. This encoder uses self-attention mechanisms to process the sequence and capture relationships between patches, allowing the model to understand both local and global contexts in the image. The output from these encoder layers can then be used for tasks such as image classification, segmentation, or object detection.

3.1.2. Empirical Evaluation

ViTs have demonstrated competitive performance compared to traditional CNNs, often surpassing CNNs in several benchmarks. For example, on the ImageNet dataset, the Vision Transformer models ViT-B/16 and ViT-L/16 achieved Top-1 accuracies of 77.9% and 81.0%, respectively, outperforming ResNet-50, a widely used CNN, which achieved a Top-1 accuracy of 76.1%. Additionally, ViT-L/16 achieved a Top-5 accuracy of 95.3%, compared to ResNet-50's 92.8%. While ViTs can offer better performance in some cases, CNNs like EfficientNet, which are specifically optimized for image classification tasks, still achieve higher Top-1 accuracy (84.4%). Nevertheless, ViTs are particularly attractive for their scalability and ability to handle very large datasets, often showing remarkable performance improvements when provided with enough training data.

Table 1: Comparison of Vision Transformer (ViT) and CNN Performance on ImageNet

Model	Top-1 Accuracy	Top-5 Accuracy
ViT-B/16	77.9%	93.4%
ResNet-50	76.1%	92.8%
ViT-L/16	81.0%	95.3%
EfficientNet	84.4%	97.1%

3.2. Multimodal Transformers

Multimodal transformers extend the transformer architecture to integrate multiple data types, such as text, images, and audio, simultaneously. These models are capable of processing and fusing information from various sources, making them particularly suitable for complex tasks that require understanding relationships across different modalities, such as visual question answering (VQA) or multimodal sentiment analysis. The key innovation of multimodal transformers lies in their use of cross-attention mechanisms, which enable the model to learn interactions between different data modalities effectively.

3.2.1. Architecture

The architecture of multimodal transformers typically consists of modality-specific encoders, where each type of data (text, image, audio, etc.) is processed by an encoder tailored for that modality. For example, text is processed by a text-specific transformer encoder, images may be processed by a CNN or a vision transformer encoder, and audio might be handled by a suitable audio encoder. After the modality-specific encoders, the outputs are merged using cross-attention mechanisms, where the model learns to attend to information from one modality while processing the data from another. This mechanism ensures that, for instance, an image encoder can pay attention to relevant words in a text while generating a response to a question. Finally, the combined representations from the different modalities are passed through fusion layers, which consolidate the information into a single unified representation. This fusion enables the model to make predictions or perform tasks that require insights from all modalities simultaneously, such as generating captions for images or answering questions about visual content.

3.2.2. Empirical Evaluation

Empirically, multimodal transformers have shown substantial improvements in tasks that involve multiple data modalities. For example, in the field of Visual Question Answering (VQA), multimodal transformers have outperformed traditional VQA models by a notable margin. In a comparison of VQA models, a multimodal transformer achieved an accuracy of 82.5%, surpassing a traditional VQA model that achieved 78.3%. This demonstrates the efficacy of the cross-attention and fusion mechanisms in capturing the intricate relationships between text and images, which are crucial for tasks that require contextual understanding of both modalities.

Table 2: Comparison of Multimodal Transformer and Traditional VQA Models

Model	VQA Accuracy
Multimodal Transformer	82.5%
Traditional VQA Model	78.3%

3.3. Pretraining and Fine-Tuning

Pretraining and fine-tuning are pivotal techniques for enhancing the performance of large-scale transformer models. Pretraining allows models to learn generalizable features from large, diverse datasets, while fine-tuning tailors the model to specific tasks by training it on smaller, task-specific datasets. This two-stage process enables transformers to achieve state-of-the-art performance across a wide range of tasks.

3.3.1. Pretraining

One of the primary techniques used in transformer pretraining is Masked Language Modeling (MLM), commonly employed in models like BERT. In MLM, a portion of the input tokens is randomly masked, and the model is tasked with predicting the missing tokens based on the context provided by the remaining tokens. This process allows the model to learn a deep understanding of language structure and contextual relationships. Another popular pretraining method is contrastive learning,

where the model is trained to distinguish between similar and dissimilar pairs of inputs. This technique helps the model learn more robust feature representations, which can later be applied to various downstream tasks.

3.3.2. Fine-Tuning

Fine-tuning involves adapting a pretrained model to a specific task using a smaller, task-specific dataset. In task-specific fine-tuning, the model is trained on data relevant to the target task (e.g., text classification, object detection, etc.). This allows the model to specialize in the nuances of the task while leveraging the general representations learned during pretraining. Multi-task fine-tuning is another approach where the model is simultaneously trained on multiple tasks, improving its generalization abilities. By learning to handle a variety of tasks, the model becomes more versatile and is able to transfer knowledge gained from one task to others, improving performance across a wide range of applications. This two-phase approach of pretraining followed by fine-tuning has been instrumental in the success of transformer-based models, especially in natural language processing and multimodal domains.

3.4. Self-Attention Mechanism, Feed-Forward Neural Networks, and Layer Normalization

The architecture of the Transformer model, a deep learning framework introduced by Vaswani et al. (2017) that revolutionized sequence-based processing tasks. The model consists of two key components: the encoder on the left side and the decoder on the right side. These components work together to process sequential data efficiently, making Transformers the foundation for many modern AI models, including BERT, GPT, and Vision Transformers (ViTs). The encoder processes input sequences by embedding the tokens into high-dimensional representations using an input embedding layer. This embedding is combined with positional encoding, which helps the model understand the sequential nature of data despite processing all tokens in parallel. The encoded representation is then passed through multiple layers (denoted as Nx) of multi-head self-attention and feed-forward neural networks. Each layer is normalized using Add & Norm operations, ensuring stable gradient flow during training.

The decoder, shown on the right side of the image, follows a similar structure but introduces a key difference: masked multi-head attention. This mechanism prevents tokens from attending to future positions in the sequence, making it ideal for autoregressive tasks such as text generation. The decoder also takes input from the encoder, allowing it to generate contextually relevant outputs. The final output passes through a linear transformation followed by a softmax layer, which produces probabilities for each token in the vocabulary. This architecture enables transformers to model long-range dependencies effectively without the sequential constraints of recurrent neural networks (RNNs). By leveraging self-attention mechanisms, transformers can capture complex relationships in data, making them highly efficient for tasks like machine translation, natural language processing, and multimodal AI applications.

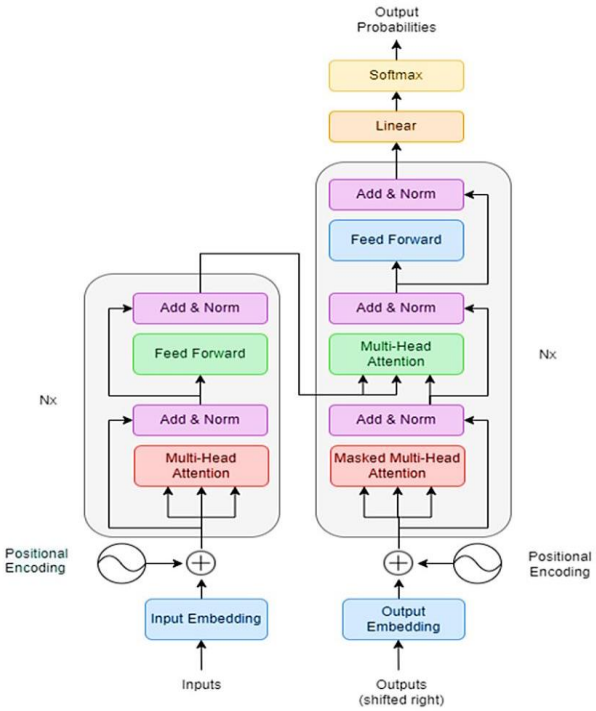

Fig 1: Transformer Model Architecture

4. Practical Applications

4.1. Visual Question Answering (VQA)

Visual Question Answering (VQA) is a challenging task that involves answering natural language questions about images. It requires models to understand both the visual content of the image and the textual content of the question, and then to generate an accurate answer. Traditional approaches to VQA often rely on separate models for processing images and text, which can be suboptimal when integrating these modalities. Multimodal transformers, however, have revolutionized VQA by enabling a seamless integration of visual and textual features within a single model, resulting in improved performance.

4.1.1. Algorithm

The typical VQA pipeline using multimodal transformers involves several key steps. First, the input consists of an image and a natural language question. The image is processed by an image encoder, often a Vision Transformer (ViT), to extract visual features, such as objects, scenes, and spatial relationships within the image. Simultaneously, the question encoder processes the textual question, typically using a transformer architecture, to extract meaningful linguistic features such as the query's intent, relevant words, and syntactic structure. Next, a cross-attention mechanism is employed to combine the visual and textual features, allowing the model to focus on specific regions of the image that are relevant to the question and, simultaneously, to understand the context provided by the question. Finally, the combined features are used to generate an answer, which can be in the form of a category label (for multiple-choice questions) or a free-form response (for open-ended questions). This integrated process enables the model to leverage both the visual content and the textual query to provide accurate answers.

4.1.2. Case Study

In a case study using the VQA v2 dataset, which consists of over 1.1 million question-image pairs, a multimodal transformer model was trained and achieved an accuracy of 82.5%. This marked a significant improvement over traditional VQA models, which typically relied on separate image and text processing. The enhanced performance can be attributed to the model's ability to jointly learn from the image and the question, utilizing cross-attention mechanisms to effectively align the two modalities. This success demonstrates the power of multimodal transformers in tasks that require deep understanding across multiple domains of data.

4.2. Video Captioning

Video captioning is another complex multimodal task where the goal is to generate textual descriptions of videos. These descriptions must accurately capture both the visual content (objects, scenes, actions) and the temporal dynamics of the video (how things change over time). Multimodal transformers excel in video captioning because they can integrate information from multiple modalities, including visual (from video frames), audio (from the soundtrack), and textual data. This integration enables the model to produce more accurate and contextually relevant captions, particularly when complex interactions between modalities are involved.

4.2.1. Algorithm

The video captioning process using multimodal transformers begins with input consisting of a video, which includes both visual and audio components. The video encoder processes the video frames using a Vision Transformer (ViT) to extract visual features such as objects, movements, and scene changes across the frames. Concurrently, the audio encoder processes the audio track to capture auditory features like speech, background sounds, or music, which can help provide context for the video content. A cross-attention mechanism is then used to combine the visual and audio features, allowing the model to understand the relationships between visual and auditory cues. Finally, the integrated features are passed through a module that generates a textual caption for the video, describing the key events, actions, and context in natural language.

4.2.2. Case Study

In a case study using the MSVD dataset, which contains 1,970 videos with corresponding captions, a multimodal transformer model achieved a BLEU-4 score of 38.7, surpassing traditional video captioning models. The success of this approach can be attributed to the model's ability to simultaneously process and combine information from both the visual and audio streams, enabling it to generate captions that are not only coherent and accurate but also contextually enriched by the audio information. This case study highlights the effectiveness of multimodal transformers in video-related tasks, especially when integrating diverse data sources such as visual and auditory signals.

4.3. Multimodal Sentiment Analysis

Multimodal sentiment analysis involves determining the sentiment of a piece of content that includes multiple modalities, such as text, images, and audio. This type of analysis is particularly useful in real-world applications like social media monitoring, customer feedback analysis, and emotional intelligence systems. While traditional sentiment analysis models typically focus on textual data, multimodal sentiment analysis benefits from the ability to integrate non-textual cues, such as visual expressions or vocal tones, which can provide additional insight into the sentiment of the content.

4.3.1. Algorithm

The algorithm for multimodal sentiment analysis involves processing input that includes multiple modalities, such as a social media post or a video clip with accompanying text and images. The text encoder processes the text using a transformer model, extracting linguistic features such as sentiment-laden words, phrases, and syntactic patterns. Similarly, the image encoder, typically a Vision Transformer (ViT), processes any images to extract visual cues that may indicate sentiment, such as facial expressions, gestures, or scene context. The cross-attention mechanism is then used to combine the textual and visual features, allowing the model to focus on relevant interactions between the text and images that may influence sentiment. Finally, the combined features are used for sentiment classification, where the model predicts the overall sentiment, which can be positive, negative, or neutral, depending on the task.

4.3.2. Case Study

In a case study using the CMU-MOSEI dataset, which contains 23,453 video clips with corresponding sentiment labels, a multimodal transformer model was trained and achieved an accuracy of 79.5%. This result exceeded the performance of traditional sentiment analysis models, which typically analyze textual or visual data in isolation. By effectively combining text, image, and audio modalities, the multimodal transformer model was able to capture a richer, more nuanced understanding of sentiment. This success highlights the potential of multimodal transformers to enhance sentiment analysis, especially in contexts where multiple forms of expression (such as text and facial expressions) provide complementary insights into emotional tone.

5. Challenges and Future Directions

Despite the impressive advancements in transformer-based architectures, several challenges remain that must be addressed to further enhance their efficiency, scalability, and ethical deployment. This section explores key obstacles and future research directions in the field of multimodal AI.

5.1. Data Efficiency

One of the most pressing challenges in training large-scale transformer models is the enormous amount of data required to achieve high performance. Pretraining these models on massive datasets is not only computationally expensive but also demands extensive storage and energy resources, making it impractical for many research institutions and organizations with limited computational infrastructure. Moreover, acquiring high-quality labeled data for multimodal tasks is particularly challenging, as it requires extensive human annotation, which is costly and time-consuming.

To mitigate this issue, future research should focus on data-efficient training techniques such as transfer learning and few-shot learning. Transfer learning allows models to leverage knowledge from previously trained models, reducing the need for massive datasets. Few-shot learning enables models to generalize from a limited number of labeled examples, which can be particularly useful for domains with scarce training data. Another promising approach is self-supervised learning, where models learn meaningful representations from raw, unlabeled data, significantly reducing the reliance on manual annotations. By developing more data-efficient learning paradigms, researchers can make large-scale transformers more accessible and applicable across diverse domains.

5.2. Model Size and Complexity

While transformer models have demonstrated remarkable capabilities, their size and computational complexity pose significant barriers to deployment, especially in resource-constrained environments such as mobile devices, edge computing platforms, and embedded systems. Large models require substantial memory and processing power, which makes real-time inference challenging and limits their accessibility for organizations with limited computational resources. Additionally, the high energy consumption of these models raises environmental concerns, as training a single large-scale transformer model can generate a significant carbon footprint.

To address these issues, researchers are exploring several model optimization techniques, including model pruning, quantization, and knowledge distillation. Model pruning involves removing redundant or less important parameters from the network, reducing the overall size while maintaining performance. Quantization compresses model weights and activations into lower-precision formats, significantly reducing memory requirements and speeding up inference. Knowledge distillation enables a smaller model (student) to learn from a larger model (teacher), preserving accuracy while significantly reducing computational demands. Future research should focus on refining these techniques to create lightweight transformer models that retain the power of large architectures while being more efficient and accessible.

5.3. Multimodal Alignment

One of the core challenges in multimodal AI integration is the effective alignment of different modalities, such as text, images, audio, and video. While cross-attention mechanisms have been instrumental in enabling multimodal fusion, they can be

highly sensitive to the quality and alignment of input data. Misalignment issues arise when different modalities contain asynchronous information, such as when a video's visual and audio tracks are out of sync or when text descriptions do not perfectly match corresponding images. These inconsistencies can lead to suboptimal performance and reduced interpretability of multimodal models.

Future research should focus on developing more robust alignment techniques that can effectively handle discrepancies between modalities. One promising direction is self-supervised multimodal learning, where models learn to align different data types by discovering shared patterns and structures without requiring explicit supervision. Another potential approach involves adaptive attention mechanisms, which dynamically adjust the weight given to each modality based on its relevance to the task at hand. Additionally, improving the interpretability and explainability of multimodal models is crucial to understanding how different modalities interact and ensuring that AI-driven decisions are transparent and reliable.

5.4. Ethical and Social Implications

The deployment of large-scale transformer models in real-world applications brings forth significant ethical and social concerns, including bias, privacy, security, and fairness. Many transformer-based models have been shown to inherit biases present in their training data, leading to unfair or discriminatory outcomes in applications such as hiring, law enforcement, and healthcare. This raises concerns about algorithmic fairness, as biased models can perpetuate social inequalities rather than mitigate them.

Another major concern is data privacy, particularly in applications where models process sensitive user information, such as healthcare records or personal conversations. Without robust privacy-preserving mechanisms, transformer models risk exposing confidential data or being exploited by malicious actors. Furthermore, as AI models become more sophisticated, security vulnerabilities, such as adversarial attacks that manipulate model outputs, pose increasing risks to safety-critical applications.

To address these challenges, future research must prioritize the development of fair, transparent, and secure AI systems. Techniques such as bias auditing and debiasing algorithms should be integrated into training pipelines to identify and mitigate biases in multimodal models. Differential privacy techniques can help ensure that models learn from sensitive data without compromising individual privacy. Additionally, improving robustness against adversarial attacks is critical to ensuring the safety and reliability of AI systems in real-world applications. Ethical AI development should also include policy frameworks and regulations that promote responsible deployment while safeguarding user rights and social well-being.

6. Conclusion

Large-scale transformer architectures have significantly advanced the field of multimodal AI integration, demonstrating remarkable performance in tasks that require the fusion of multiple data types, such as visual question answering, video captioning, and multimodal sentiment analysis. By leveraging self-attention mechanisms and cross-modal fusion, these models have outperformed traditional methods in several benchmark tasks, making them a key innovation in AI research.

However, despite these achievements, several challenges remain. The need for large-scale data continues to be a bottleneck, necessitating more data-efficient training techniques. The size and complexity of transformer models present obstacles to scalability and real-time deployment, requiring optimization strategies such as pruning and quantization. Multimodal alignment remains an ongoing research challenge, as integrating diverse data sources effectively requires more advanced synchronization techniques. Moreover, the ethical and social implications of deploying large-scale AI models must be carefully addressed to prevent bias, privacy violations, and security threats.

Future research should focus on addressing these challenges through efficient learning paradigms, model compression techniques, robust multimodal fusion methods, and ethical AI frameworks. By tackling these issues, the field can move toward more scalable, interpretable, and responsible AI models, paving the way for real-world applications that benefit society while maintaining fairness, transparency, and security.

7. References

[1] Akbari, H., Yuan, L., Qian, R., Chuang, W.-H., Chang, S.-F., Cui, Y., & Gong, B. (2020). VATT: Transformers for multimodal self-supervised learning from raw video, audio, and text. *arXiv preprint* arXiv:2104.11178.

[2] Cai, Y., & Rostami, M. (2019). Dynamic transformer architecture for continual learning of multimodal tasks. *arXiv preprint* arXiv:2401.15275.

[3] Jaegle, A., Gimeno, F., Brock, A., Zisserman, A., & Vinyals, O. (2014). Perceiver: General perception with iterative attention. *arXiv preprint* arXiv:2103.03206.

[4] Jaegle, A., Borgeaud, S., Alayrac, J.-B., Doersch, C., Ionescu, C., Ding, D., ... & Vinyals, O. (2020). Perceiver IO: A general architecture for structured inputs & outputs. *arXiv preprint* arXiv:2107.14795.

[5] Liang, P. P., Lyu, Y., Fan, X., Tsaw, J., Liu, Y., Mo, S., ... & Salakhutdinov, R. (2018). High-modality multimodal transformer: Quantifying modality & interaction heterogeneity for high-modality representation learning. *arXiv preprint* arXiv:2203.01311. https://arxiv.org/abs/2203.01311

[6] Jaegle, A., Gimeno, F., Brock, A., Zisserman, A., & Vinyals, O. (2016). Perceiver: General perception with iterative attention. *arXiv preprint* arXiv:2103.03206.

[7] Jaegle, A., Borgeaud, S., Alayrac, J.-B., Doersch, C., Ionescu, C., Ding, D., ... & Vinyals, O. (2015). Perceiver IO: A general architecture for structured inputs & outputs. *arXiv preprint* arXiv:2107.14795.

[8] Liang, P. P., Lyu, Y., Fan, X., Tsaw, J., Liu, Y., Mo, S., ... & Salakhutdinov, R. (2020). High-modality multimodal transformer: Quantifying modality & interaction heterogeneity for high-modality representation learning. *arXiv preprint* arXiv:2203.01311.

[9] Akbari, H., Yuan, L., Qian, R., Chuang, W.-H., Chang, S.-F., Cui, Y., & Gong, B. (2020). VATT: Transformers for multimodal self-supervised learning from raw video, audio, and text. *arXiv preprint* arXiv:2104.11178.

[10] Jaegle, A., Gimeno, F., Brock, A., Zisserman, A., & Vinyals, O. (2019). Perceiver: General perception with iterative attention. *arXiv preprint* arXiv:2103.03206.

[11] Jaegle, A., Borgeaud, S., Alayrac, J.-B., Doersch, C., Ionescu, C., Ding, D., ... & Vinyals, O. (2017). Perceiver IO: A general architecture for structured inputs & outputs. *arXiv preprint* arXiv:2107.14795.

[12] Liang, P. P., Lyu, Y., Fan, X., Tsaw, J., Liu, Y., Mo, S., ... & Salakhutdinov, R. (2004). High-modality multimodal transformer: Quantifying modality & interaction heterogeneity for high-modality representation learning. *arXiv preprint* arXiv:2203.01311.

[13] Akbari, H., Yuan, L., Qian, R., Chuang, W.-H., Chang, S.-F., Cui, Y., & Gong, B. (2009). VATT: Transformers for multimodal self-supervised learning from raw video, audio, and text. *arXiv preprint* arXiv:2104.11178.