Digital Ecosystems and Multimodal Architectures

Loiko A.I.

Digital ecosystems and multimodal architectures have emerged as two technological trends, the convergence of which is creating the next generation of intelligent digital services [1].

A digital ecosystem is a collection of interconnected digital services, platforms, data, and applications that create a unified environment to meet user needs [2]. It is a set of modules that interact with each other [3]. As a result, a single subscription or account provides access to finance, entertainment, delivery, communications, and other services [4].

Multimodal Architectures are an advanced type of artificial intelligence that can simultaneously process, interpret, and integrate information from multiple data sources (modalities).

Unlike traditional unimodal AI systems that specialize in processing a single type of data (for example, only text or only images), multimodal AI creates a comprehensive understanding by synthesizing information from various formats.

Humans process information through multiple senses to understand their environment. Multimodal AI replicates this multisensory approach, enabling technologies to develop a more nuanced and context-sensitive understanding of complex situations. The input module (sensor system) serves as the AI's data collection interface, collecting various types of data, including text, images, audio, video, and sensor readings. It preprocesses the diverse data, making it suitable for subsequent analysis.

The Fusion module (central processor) combines data from multiple sources using advanced algorithms. It identifies patterns, extracts meaningful features, and creates a unified representation that captures the essence of the multimodal input data. The output module (response generator) processes the data and produces re-

sults, which may include predictions, recommendations, generated content, or actionable insights. These outputs can be presented in text, image, audio, or a combination of these formats, depending on the application requirements.

The operating mechanism of multimodal AI includes machine learning methods that enable seamless integration of diverse data streams:

Multimodal AI models use different, specialized architectures for each data type. This means that visual, text, audio, or sensory inputs are processed by systems designed specifically for them. This allows the model to capture the unique details of each input before combining them.

Convolutional neural networks (CNNs) or Vision Transformers interpret visual information from images and videos, creating rich feature representations.

Transformer-based models, such as those in the GPT family, transform text inputs into meaningful semantic embedding. Specialized neural networks process audio signals or spatial sensory inputs, ensuring an accurate representation of each modality and preserving its distinctive characteristics.

After individual processing, each modality generates high-level features optimized to capture the unique information contained in that particular data type. After feature extraction, multimodal models combine them into a single, coherent representation. To effectively accomplish this task, various fusion strategies are used. Early fusion combines the extracted feature vectors immediately after processing each modality. This strategy promotes deeper cross-modal interactions early in the analysis pipeline. Late fusion supports modality separation until the final decision stages, where predictions from each modality are combined using ensemble methods such as averaging or voting.

Modern architectures often integrate features multiple times across different model layers, using joint attention mechanisms to dynamically highlight and align important intermodal interactions. Hybrid fusion can emphasize the alignment of specific spoken words or text phrases with corresponding visual features in real time.

Multimodal systems use alignment and attention methods to ensure efficient mapping of data from different modalities.

Methods such as contrastive learning help align visual and textual representations within a common semantic space. As a result, multimodal models can establish strong, meaningful relationships between different types of data. Transformer-based attention mechanisms further enhance this alignment, allowing models to dynamically focus on the most relevant aspects of each input signal. Attention layers allow the model to directly associate specific text descriptions with corresponding regions in the visual data, significantly improving accuracy in complex tasks such as visual question-answering (VQA) and image captioning.

These techniques enhance multimodal AI's ability to deeply understand context, enabling AI to provide more nuanced and accurate interpretations of complex real-world data.

Multimodal AI has evolved significantly, moving from early rule-based techniques to deep learning systems capable of complex integration.

Multimodal systems combined different types of data, such as images, audio, or sensor data, using rules manually created by human experts or simple statistical methods. Early robotic navigation systems combined camera images with sonar data to detect and avoid obstacles. While effective, these systems required careful manual feature engineering and were limited in their ability to adapt and generalize. With the advent of deep learning, multimodal models such as multimodal auto encoders began to learn joint representations of different data types, especially images and text, enabling AI to solve problems such as cross-modal search and image retrieval based solely on text descriptions.

When Visual Question Answering (VQA) integrated CNNs for image processing and RNNs or Transformers for text interpretation, AI models began to accurately answer complex, context-sensitive questions about visual content.

Large-scale multimodal models trained on massive internet-scale datasets have further enhanced AI capabilities. These models use contrastive learning, enabling them to identify generalizable relationships between visual content and text

descriptions. By bridging the gaps between modalities, modern multimodal architectures have expanded AI's capabilities to perform complex visual reasoning tasks, illustrating how far multimodal AI has come from its foundational stages.

These modalities include text, images, audio, video, and sensor data. Multimodal AI understands context by combining different types of data for more accurate results. GPT-4V, Gemini, and Claude are capable of "seeing" images, "hearing" audio, and "reading" text simultaneously.

High latency and large models require a powerful infrastructure to maintain. Integrating multimodal AI into the digital ecosystem is changing user interactions with services: Ecosystems use AI to create content (text, images, and video) within social networks and services. Combining these approaches allows companies to create more flexible, intelligent, and competitive products.

Multimodal neural network architectures are unified models capable of perceiving, processing, and generating information in various modalities. They understand not only text input but also images and video, audio and speech, music, 3D, 2D vector graphics, and more. Such systems are built on the principles of modularity and unified data representation, which enables the effective combination of disparate information sources in a single latent space. At the input, each modality passes through a specialized encoder, which transforms the data into universal tokens or vectors suitable for joint processing. The central core of the model is a neural network model based on a transformer architecture, performing cross-modal alignment, context merging, and meaning extraction.

At the output, decoders are used, returning the result in the target modality, whether it be visual generation and image editing, voiceover, visualization, or 3D object reconstruction. This approach provides not only generalization but also flexibility in applied tasks, from search and generation to autonomous behavior and user interaction. Unlike highly specialized solutions, multimodal models rely on a common knowledge representation and learn from relationships between modalities, not just within them.

This allows them to demonstrate a high level of semantic understanding and interaction with the real world in its natural, multimodal, or mono- modal form. A hybrid data testing strategy plays an important role, combining synthetic metrics with A/B testing, as well as taking into account user feedback and automated quality metrics on real business problems. This will not only achieve high model accuracy under controlled conditions, but also guarantee its effectiveness in a dynamic business environment.

Errors in LLM can be reduced through architectural improvements, although they cannot yet be completely eliminated. This is still just matrix multiplication and a pinch of probability theory. However, the integration of advanced self-learning mechanisms and enhanced control over the output sequence, retrieval-augmented generation (RAG), even in multimodal scenarios, helps reduce hallucinations and improve the reliability of responses. Specialized mechanisms allow the model to be adapted to highly specialized tasks, improving its robustness [5].

However, fundamental limitations remain, related to the probability of generation errors and the complexity of modeling full contextual truth. Therefore, the future lies with hybrid systems that combine LLM with classical algorithms, verifiable knowledge bases, and real-time fact-checking mechanisms [6].

Adaptations using domain adaptation, instruction tuning, and retrieval-augmented generation before training allow the integration of external knowledge sources. These techniques ensure the preservation of the model's generative and creative capabilities while simultaneously increasing its accuracy and robustness in solving specific applied problems.

The use of self-refinement, response verification using auxiliary models, or the addition of reasoning enhances control over the reliability of output. In practice, this allows LLM to demonstrate both flexibility in formulating unconventional solutions and compliance with strict domain requirements.

All of these allow for quantitative comparisons of models using standardized metrics. However, given the growing complexity and diversity of problems, especially in real-world scenarios, new, more complex, and specialized benchmarks are

needed to identify subtle differences in model capabilities and better assess their robustness, adaptability, and generalization.

Models with very long contexts are emerging, and it's important to be able to evaluate them correctly. In the visual modality, it's necessary to understand not only short videos but also videos longer than an hour. It's important not only to summarize such videos but also to be able to find any events in the video at any time interval, including by audio modality, if the video recording contains sounds or speech.

It's important to be able to evaluate models across all possible modalities, even if the model is represented not by a single modal architecture, but by a set of engineering solutions. And if this is done seamlessly, not just through text queries, but in real time using voice, then the result is a fully-fledged, complex multimodal symbiosis that can see, hear, read, write, speak, and even draw.

Multimodality expands the ways in which information is perceived and processed, bringing models closer to a human understanding of the world. The ability to combine text, images, audio, and video allows models to form a more holistic view of reality, eliminating the limitations of text-based systems. This is critical for the future of AGI, as humans, too, perceive the world not through a single channel, but through the complex integration of different types of information.

However, multimodality alone does not eliminate logical errors, because current models are still statistical predictors, not full-fledged reasoning systems. Real AGI requires not only multimodal input, but also advanced reasoning mechanisms, long-term memory, and the ability to plan purposefully.

Any attempts to regulate models and risk checks typically lead to a slowdown in technological development. This leads talented researchers and companies to shift their developments to jurisdictions with more flexible regulations, weakening their scientific and, potentially, economic leadership. Furthermore, it creates disunity within the community, and slowing AI progress could lead to a race to artificial benchmarks, which could also hinder solutions to real societal problems. Instead of blocking controls, it is important to develop adaptive and transparent security mechanisms without stifling industry development [7].

Thus, such systems can be beneficial to business and society, as they incorporate architectures for processing multimodal data.

List of references

1. Loiko, A.I. Technology of digital ecosystems // Samara State Technical University. Philosophy. Vol 4, No 1 (2022) P. 49-56

2. Bejjani, M., Göcke, L., Matthias Menter M. Digital entrepreneurial ecosystems: A systematic literature review // Technological Forecasting and Social Change. 2023. V. 189. P. 122372. DOIhttps://doi.org/10.1016/j.techfore.2023.122372.

3. Loiko, A.I. Barrier-free space of socio-cultural activities of digital ecosystems. Experience industries. Socio-Cultural Research Technologies (EISCRT), 2022, 1(1),198-212.https://doi.org/10.34680/EISCRT-2022-1(1)-198-212

4. Лойко А. И. Социальные цифровые экосистемы: тренды эволюции // Россия: тенденции и перспективы развития: материалы XXI Нац. науч. конф. с междунар. участием (Москва, 16–17 декабря 2021 г.). М.: Изд-во Ин-та науч. информ. по обществ. наукам РАН, 2022. Вып. 17. Ч. 1. С. 180–182.

5. Pujadas R., Valderrama E., Venters W. The value and structuring role of web APIs in digital innovation ecosystems: The case of the online travel ecosystem // Research Policy. 2024. V. 53. N. 2. P. 104931. DOIhttps://doi.org/10.1016/j.respol.2023.104931.

6. Каленов О. Е. Цифровые экосистемы организаций // Вестник Рос. экон. ун-та им. Г. В. Плеханова. 2022. Т. 19. № 1 (121). С. 139–147. DOIhttps://doi.org/10.21686/2413-2829-2022-1-139-147.

7. Costabile C. Digital platform ecosystem governance of private companies: Building blocks and a research agenda based on a multidisciplinary, systematic literature review // Data and Information Management. 2024. V. 8. N. 1. P. 100053. DOIhttps://doi.org/10.1016/j.dim.2023.100053.