Інформаційні технології

Chernets Vadym
Data Analyst, PhD
New York, USA

MULTIMODAL ARTIFICIAL INTELLIGENCE: EVOLUTION OF TECHNOLOGIES, ARCHITECTURE AND NEW HORIZONS OF HUMAN-MACHINE INTERACTION

Summary. *Multimodal artificial intelligence (AI) is an advanced field that combines different types of data, such as text, images, audio, and video. The article discusses the evolution of multimodal AI, its architectural features, popular modern models (CLIP, DALL-E, GPT-4, Flamingo, and others), and the prospects for human-machine interaction. Particular attention is paid to the transformation of approaches to data processing, the integration of different modalities, and the creation of more natural interaction interfaces.*

Key words: *multimodal artificial intelligence, human-machine interaction, evolution of multimodal AI.*

Introduction. Artificial intelligence has come a long way from highly specialized systems working with one type of data to complex multimodal models capable of simultaneously analyzing and interpreting information from different sources. Multimodal AI represents the next stage of evolution, where systems can perceive and process data by imitating the human ability for multisensory perception.

It is worth noting that the term "multimodal AI" is quite broad and can refer to a wide range of systems designed to handle and integrate various data types - such as text, images, audio, and video—within a single model. This versatility opens up

new possibilities for human-machine interaction, making it more intuitive and effective.

Evolution of Multimodal AI

Early stages of development

The first AI systems were focused on processing one type of data, such as text or images. However, as early as the 1990s, research began in the field of multimodal analysis, where attempts were made to combine data from different sources. For example, speech recognition systems began to integrate with visual data to improve accuracy.

Breakthroughs in deep learning

With the advent of deep learning and neural networks, multimodal AI received a new impetus. Models such as transformers made it possible to efficiently process and combine data from different modalities. An example is the CLIP (Contrastive Language–Image Pretraining) architecture, which links text and images, allowing the system to understand context and semantics.

Modern achievements

Today, multimodal AI is actively used in areas such as medicine, autonomous vehicles, robotics, and virtual assistants. For example, in medicine, systems analyze medical images, text descriptions, and audio recordings to make a diagnosis. In autonomous vehicles, data from cameras, lidar and microphones is combined to make decisions in real time.

Architecture of Multimodal Models

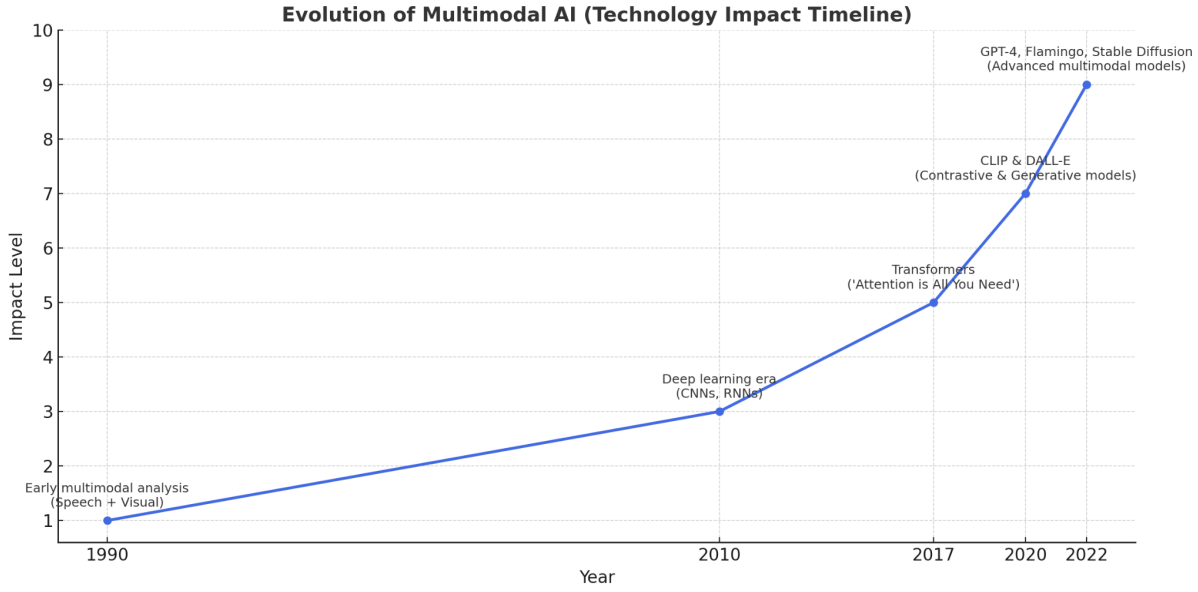
Multimodal AI models are complex systems that can process and integrate data from various sources, such as text, images, audio, and video. Their architecture is built on advanced technologies that allow for efficient combination and analysis of information from different modalities.

Encoders are the core component of multimodal models. They are responsible for converting data from different modalities into numerical representations (embeddings) that can be processed by neural networks. Each modality is processed by a separate encoder, which allows for the unique features of each data type to be taken into account. For example, text encoders such as BERT or GPT convert text into a sequence of tokens, which are then encoded into numerical vectors. Visual encoders such as Vision Transformers (ViT) analyze images as a sequence of patches, and audio encoders convert sound signals into spectrograms or other numerical representations.

Once the data from different modalities has been transformed into numerical representations, it must be combined for further processing. Fusion mechanisms play a key role in multimodal models, as they determine how data from different sources interact with each other. Early fusion involves combining data at the initial stages of processing, allowing interactions between modalities to be taken into account from the start. Late fusion, on the other hand, involves processing data independently and then combining the results. Cross-modal attention, used in models such as Flamingo, allows for connections between different modalities, which is especially useful for tasks that require deep understanding of context.

Transformers and attention mechanisms have become the basis for many multimodal models due to their ability to process sequences of data and establish connections between different elements. Transformers consist of multiple attention layers and fully connected layers, allowing them to work effectively with data of different natures. Attention mechanisms allow models to focus on the most relevant parts of the data, which is especially important when working with multimodal information.

Decoders are used in generation tasks where numerical representations of data need to be converted into output such as text, images, or audio. Text decoders, such as those used in GPT-4, convert embeddings into text descriptions. Visual decoders, such as those used in DALL-E, convert tokens into images. Audio decoders, such as those used in Whisper, convert text into audio.



Modern popular models of multimodal AI

CLIP (Contrastive Language–Image Pretraining)

Architecture: CLIP consists of two encoders — text (based on the transformer) and visual (based on the Vision Transformer). The model is trained on image-text pairs using contrastive learning, which maximizes the similarity between correct pairs and minimizes it for incorrect ones.

Application: CLIP is used for image classification tasks, image retrieval by text queries, and creating multimodal interfaces.

Advantages: High generalization ability and the ability to work with zero examples (zero-shot learning).

DALL-E (OpenAI)

Architecture: DALL-E is based on the GPT-3 architecture, but adapted for generating images from text descriptions. The model uses variational autoencoders (VAE) to compress images into discrete tokens, which are then processed by the transformer.

Application: Image generation based on text queries, art creation, design.

Advantages: Ability to create unique and creative images based on complex text descriptions.

GPT-4 (OpenAI)

Architecture: GPT-4 extends GPT-3.5 into a multimodal model capable of processing both text and images. It leverages an enhanced transformer architecture with advanced attention mechanisms for data analysis and generation.

Applications: Virtual assistants, text generation, image analysis, multimodal dialogue systems.

Advantages: High versatility and strong generalization capabilities, allowing it to solve a broad range of complex tasks.

Flamingo (DeepMind)

Architecture: Flamingo integrates visual and textual data using Perceiver Resampler modules, which compress visual information into compact representations. These are combined with transformers for efficient text processing.

Applications: Video analytics, multimodal dialogue systems, and few-shot learning tasks, including scenarios with zero examples.

Advantages: Efficient processing of long data sequences, including video, with strong performance even in low-data scenarios.

BLIP (Bootstrapped Language-Image Pretraining)

Architecture: BLIP uses two encoders - one for text and one for images - and is trained on text generation and image classification tasks. The model also includes mechanisms for filtering noisy data.

Applications: Image caption generation, image retrieval, multimodal dialog systems.

Advantages: High accuracy in text generation and classification tasks.

Stable Diffusion

Architecture: Stable Diffusion is based on diffusion models that gradually transform noise into an image. The model uses text embeddings to control the generation process.

Applications: Image generation, art creation, design.

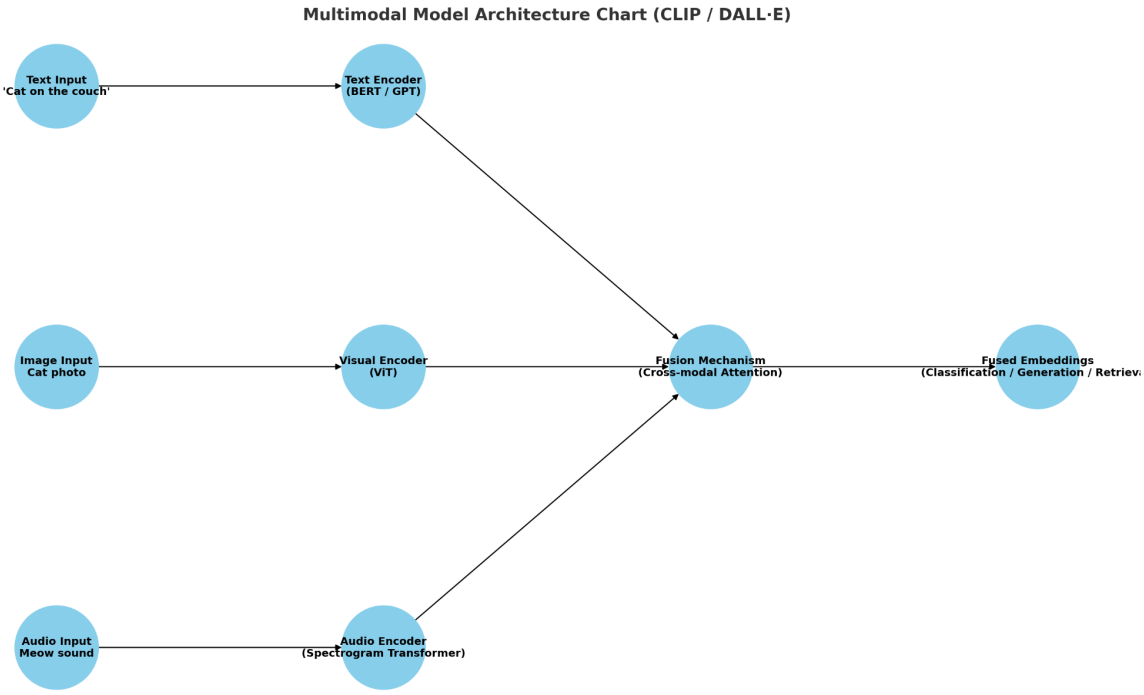
Advantages: High generation quality and the ability to fine-tune.

Whisper (OpenAI)

Architecture: Whisper is a multimodal model for processing audio and text. It uses transformers to convert audio to text and back.

Application: Speech recognition, translation, subtitle generation.

Advantages: High accuracy and support for many languages.



New horizons for human-machine interaction

Natural interfaces

Multimodal AI makes interaction with machines more natural. For example, virtual assistants such as Siri or Alexa use speech, text, and visual cues to improve the user experience.

Personalization and adaptation

Multimodal systems are able to adapt to individual user preferences by analyzing their behavior, voice, and other data. This opens up new possibilities for personalized learning, medicine, and entertainment.

Ethics and security

With the development of multimodal AI, new ethical and social issues arise, such as data protection and prevention of abuse. The development of standards and regulations is becoming an important aspect of further development.

Prospects and Challenges

Multimodal AI continues to advance rapidly, opening up new horizons for human-machine interaction. However, despite impressive achievements, researchers and developers face significant challenges that must be overcome to fully realize the potential of this technology.

One of the key areas for the future development of multimodal AI is to improve the interpretability of models. Current systems such as GPT-4 or DALL-E demonstrate high accuracy in solving problems, but their solutions often remain a "black box". Users, especially in critical areas such as medicine or autonomous vehicles, need explanations of why the model made a certain decision. Developing methods that allow models not only to produce results but also to explain them will be an important step towards trust and widespread adoption of multimodal AI.

Another important challenge is ethics and security. Multimodal models process huge amounts of data, including personal information, images, and audio recordings. This creates risks associated with privacy violations and misuse of the technology. For example, generative models such as DALL-E or Stable Diffusion can be used to create fake images or videos, raising questions about regulation and control. Developing standards and mechanisms to ensure safe and ethical use of multimodal AI is becoming a critical task.

Computational complexity also remains a serious limitation. Current multimodal models require huge computational resources to train and operate. This makes them inaccessible to many organizations and limits their use in real time, for example, in mobile devices or autonomous systems. Optimizing model architectures, using quantum computing and neuromorphic processors can be a solution to this problem, but this will require significant efforts from researchers and engineers.

Another promising area is the personalization of multimodal systems. Current models are already able to adapt to individual user preferences, but their capabilities in this direction are still limited. Future systems will be able to analyze not only text queries or images, but also emotions, voice intonations and even facial expressions of the user, which will make interaction with technology more natural and comfortable. For example, virtual assistants will be able to not only answer questions, but also recognize the user's mood and offer appropriate solutions.

In addition, multimodal AI can become a key tool for solving global problems such as climate change, food security, or healthcare. For example, models that can analyze satellite images, text reports, and sensor data can help predict natural disasters or optimize agriculture. In medicine, multimodal systems can combine data from medical images, genomic studies, and clinical records to develop personalized treatments.

However, to realize these promises, a number of technical and social challenges need to be addressed. For example, training models on multimodal data requires huge amounts of labeled data, which is not always available. Developing methods that allow models to be effectively trained on small data sets or even in the absence of data (few-shot and zero-shot learning) will be an important area of research. In addition, cultural and linguistic differences must be taken into account so that multimodal systems can work effectively on a global scale.

Conclusion. Multimodal AI is one of the most promising and dynamically developing areas of modern science and technology. Its ability to process and integrate data from various sources, such as text, images, audio and video, opens up new possibilities for human-machine interaction, making it more natural, intuitive and efficient. Modern models such as CLIP, DALL-E, GPT-4 and Flamingo demonstrate impressive results in solving complex problems, from content generation to medical data analysis.

However, despite significant achievements, multimodal AI faces significant challenges. Interpretability, ethics, safety and computational complexity remain key issues that require attention from researchers and developers. Future research will focus on creating more powerful, versatile and safe systems that can not only solve problems, but also explain their decisions, adapt to individual user needs and take into account cultural and social aspects.

Multimodal AI has the potential to change many areas, from healthcare and education to art and entertainment. It can become a tool for solving global problems such as climate change or food security, as well as making technologies more accessible and inclusive. However, this requires not only improving the technologies themselves, but also developing ethical standards and regulatory mechanisms that will ensure their safe and responsible use. In conclusion, we can say that multimodal AI is not just a technology of the future, but already a reality that is actively changing our lives. Its further development will depend on the joint efforts of scientists, engineers, politicians and society as a whole. Only in this way can we unleash the full potential of this technology and create a world where human-machine interaction is harmonious, safe and beneficial for everyone.

References

1. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint arXiv:2103.00020*.

2. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... & Sutskever, I. (2021). Zero-Shot Text-to-Image Generation. *arXiv preprint arXiv:2102.12092*.

3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*.

4. Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... & Simonyan, K. (2022). Flamingo: A Visual Language Model for Few-Shot Learning. *arXiv preprint arXiv:2204.14198*.

5. Li, J., Li, D., Xiong, C., & Hoi, S. C. H. (2022). BLIP: Bootstrapping Language-Image Pretraining for Unified Vision-Language Understanding and Generation. *arXiv preprint arXiv:2201.12086*.

6. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is All You Need. *Advances in Neural Information Processing Systems, 30*.

7. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. *MIT Press*.

8. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. *Nature, 521*(7553), 436-444.