Principles of Large AI Models

This chapter only covers the theoretical knowledge related to large multimodal AI models. Those not interested can skip this section.

This tutorial is specific to CM5; CM4 users can skip it. This section does not cover the operation and use of RiderPi.

The Emergence of Large AI Models

1. Evolution of Technical Architecture

1. The core of large multimodal models lies in integrating multi-source data such as text, images, audio, and video. Its architecture has evolved from single-modality to cross-modal fusion:

- Early single-modal models: such as AlexNet (image classification), BERT (text processing), etc., were designed for a single task and required independent training of different models.
- Breakthroughs in Transformer and Large Language Models (LLM): Achieving cross-modal semantic alignment through a unified framework (such as the GPT series, CLIP), mapping different data to the same semantic space, and reducing information loss.
- End-to-end multimodal modeling: such as GPT-4o and Google Gemini, directly processing multimodal inputs and outputs through a single model, eliminating intermediate transformation steps and improving efficiency.

2. Key Components and Training Methods

- Encoder: Converts data from different modalities (e.g., image pixels, audio waveforms) into a unified high-dimensional feature vector. For example, a visual encoder extracts image semantics, and a text encoder generates word embeddings.
- Cross-modal attention mechanism: Dynamically adjusts the weights of each modality. For example, Microsoft BEiT-3 achieves deep association between text and images through cross-modal attention.
- Pre-training and fine-tuning: Pre-training on large multimodal data (e.g., LAION-5B) and then fine-tuning for downstream tasks (e.g., robot control, medical diagnosis) to improve generalization ability.

3. Cross-modal alignment and knowledge fusion

- Alignment techniques: Such as CLIP, which aligns image and text features through contrastive learning to achieve zero-shot classification with an open vocabulary.
- External knowledge enhancement: Models such as KOSMOS-1 incorporate medical knowledge bases to improve the accuracy of complex question answering.

2. AI Large Model Application Layers

1. Robotics and Embodied Intelligence

- General-Purpose Robots: Multimodal LLMs endow robots with autonomous reasoning and learning capabilities. For example, Tesla's Optimus uses multi-sensor fusion (vision, touch, etc.) to adapt to unstructured environments.
- Real-Time Interaction and Control: Google's RT-2 model directly converts multimodal input into action codes, significantly improving success rates in unknown tasks.

- Industry Case Studies: Boston Dynamics' Spot acts as a tour guide in museums, emphasizing interactive entertainment rather than pure functionality.

2. Generative Content Creation

Textual Video and 3D Modeling: OpenAI Sora can generate high-fidelity videos, and Stable Diffusion 3 supports 3D content generation, driving innovation in the film and gaming industries.

Digital Humans and Virtual Assistants: Examples include Google's Project Astra and Tencent's MM-LLMs, enabling natural dialogue and real-time video editing.

3. Deep Penetration into Vertical Industries

Medical Diagnosis: Shukun Technology's "Digital Human" platform integrates medical images and medical record text, improving diagnostic efficiency. Industrial quality inspection: Multimodal models combined with synthetic data detect complex defects, reducing the error rate by 90%.

Financial anti-fraud: Cross-modal correlation analysis (e.g., voice + transaction records) achieves an accuracy rate of 98%.

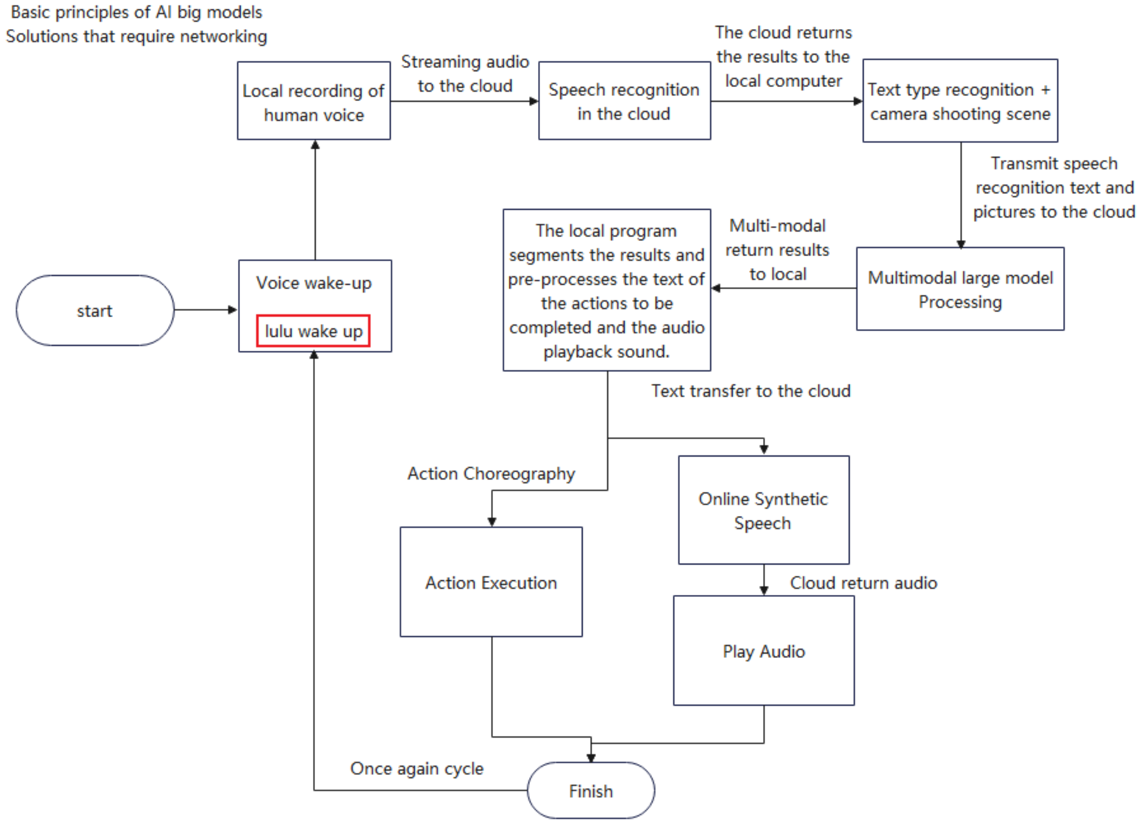
3. Summary

Multimodal large-scale models, through unified architecture and cross-modal fusion, are reshaping the boundaries of AI capabilities, demonstrating enormous potential in applications ranging from robotics to healthcare and finance.

In the future, the technology needs continuous breakthroughs in areas such as computing power optimization, ethical governance, and modal expansion to realize the vision of "human-machine symbiosis."

4. Application Examples of RiderPi Multimodal Models

RiderPi's embodied intelligence multimodal combined with an online platform solution is as follows:


Basic principles of AI big models
Solutions that require networking
start
Voice wake-up
lulu wake up
Local recording of human voice
Streaming audio to the cloud
Speech recognition in the cloud
The cloud returns the results to the local computer
Text type recognition + camera shooting scene
Transmit speech recognition text and pictures to the cloud
Multimodal large model Processing
Multi-modal return results to local
The local program segments the results and pre-processes the text of the actions to be completed and the audio playback sound.
Text transfer to the cloud
Action Choreography
Online Synthetic Speech
Action Execution
Cloud return audio
Play Audio
Once again cycle
Finish