Theoretical overview of LLM

This chapter only describes the knowledge theory related to the multi-modal LLM. Those who are not interested can ignore this section.
This section does not involve the operation and use of the robot dog.

Generation of LLM

1. Definition and core concepts of multimodality

Multimodality: refers to the presentation and interaction of information through multiple different "modalities" or sensory channels.

1. In the field of AI, the main modalities include:

- Text: natural language sequence.
- Image: visual information composed of pixels.
- Audio: sound waveform or spectrum, including voice, music, ambient sound, etc.
- Video: a combination of image sequence and audio, including spatiotemporal information.
- Structured Data: tables, knowledge graphs, etc.

2. Large Models: Usually refers to foundation models (Foundation Models) based on the Transformer architecture, with huge parameters (billions to trillions) and massive multi-source data training. They have powerful representation learning, context understanding and migration capabilities (such as GPT series, BERT series, ViT series, etc.).
3. Multimodal Large Models (MLLMs): Refers to large artificial intelligence models that can simultaneously process, understand, associate and generate information from multiple different modalities. Its core goal is to achieve unified representation, alignment and collaborative reasoning of cross-modal semantics, and simulate the human ability to integrate visual, auditory, language and other sensory information for cognition.

2. Technical architecture evolution

1. The core of multimodal large models is to integrate multi-source data such as text, images, audio, and video. Its architecture has undergone a transformation from single modality to cross-modal fusion:

- Early single-modal models: such as AlexNet (image classification), BERT (text processing), etc., are designed only for a single task and require independent training of different models.
- Breakthroughs in Transformer and Large Language Model (LLM): Cross-modal semantic alignment is achieved through a unified framework (such as GPT series, CLIP), different data are mapped to the same semantic space, and information loss is reduced.
- End-to-end multimodal modeling: such as GPT-4o and Google Gemini, a single model is used to directly process multimodal input and output, eliminating intermediate conversion steps and improving efficiency.

2. Key components and training methods

- Encoder: Converts data of different modalities (such as image pixels and audio waveforms) into a unified high-dimensional feature vector, such as visual encoders extracting image semantics and text encoders generating word embeddings.
- Cross-modal attention mechanism: Dynamically adjust the weights of each modality, such as Microsoft BEiT-3 achieves deep association between text and images through cross-modal

attention5.

- Pre-training and fine-tuning: Pre-train on large-scale multimodal data (such as LAION-5B), and then fine-tune for downstream tasks (such as robot control and medical diagnosis) to improve generalization ability.

3. Cross-modal alignment and knowledge fusion

- Alignment technology: For example, CLIP aligns image and text features through comparative learning to achieve zero-shot classification of open vocabulary.
- External knowledge enhancement: Models such as KOSMOS-1 introduce medical knowledge bases to improve the accuracy of complex question-answering.

3. Multimodal core goals and significance

- Unified semantic space: Learn a shared, aligned semantic representation space for information in different modalities, so that the representation of the same concept in different modalities is similar in vector space (such as "cat" in a picture, "cat" in text, and the sound of a cat).
- Cross-modal understanding: Understand the information of one modality based on the information of another modality (for example: picture description, picture identification by listening to sound).
- Cross-modal generation: Use the information of one modality to guide the generation of content in another modality (for example: text-to-picture, text-to-video, picture-to-text, speech synthesis with emotion).
- Multimodal collaborative reasoning: Comprehensively utilize information from multiple modalities for more complex, more robust, and closer to human cognition reasoning and decision-making (for example: answering questions based on video and text descriptions, and diagnosing with medical images and reports).

4. Application levels of LLM

1. Robots and embodied intelligence

- Generalized robots: Multimodal LLM gives robots autonomous reasoning and learning capabilities, such as Tesla Optimus, which adapts to unstructured environments through multi-sensor fusion such as vision and touch.
- Real-time interaction and control: Google's RT-2 model directly converts multimodal input into action coding, significantly improving the success rate in unknown tasks.
- Industry case: Boston Dynamics Spot serves as a tour guide in a museum, emphasizing interactive entertainment rather than pure functionality.

2. Generative content creation

- Wensheng video and 3D modeling: OpenAI Sora can generate high-fidelity videos, and Stable Diffusion 3 supports 3D content generation, promoting innovation in the film, television, and game industries.
- Digital humans and virtual assistants: such as Google Project Astra and Tencent MM-LLMs, which enable natural conversations and real-time video editing.

3. Deep penetration of vertical industries

- Medical diagnosis: Shukun Technology's "Digital Human Body" platform integrates medical images and medical records to improve diagnostic efficiency by 5.
- Industrial quality inspection: Multimodal models combined with synthetic data detect complex defects and reduce error rates by 90%.

- Financial anti-fraud: Cross-modal association analysis (such as voice + transaction records) has an accuracy rate of 98%.

5. Summary

1. The core theory of AI large model multimodality is to build an intelligent system that can uniformly understand, associate and generate heterogeneous modal information. It is based on deep learning (especially Transformer), large-scale self-supervised/weakly supervised pre-training, contrastive learning, generative models (autoregression, diffusion), etc. By solving key challenges such as modal heterogeneity, alignment, and fusion, multimodal large models are driving artificial intelligence to develop in a more general and closer to human cognitive ability, and have shown great potential in content creation, human-computer interaction, scientific discovery, education and medical fields.
2. **Currently** multimodal large models are reconstructing the capabilities of AI through unified architecture and cross-modal fusion, and their applications show great potential in fields ranging from robotics to medical care and finance.
3. **Future** research will focus on important directions such as efficiency, robustness, dynamic understanding, causal reasoning, embodied intelligence, and ethical safety to achieve the vision of "human-machine symbiosis".

6. Application examples of robot dog multimodality

The solution of embodied intelligence multimodality combined with online platform of robot dog is as follows: