AI Largr Model Types and Principles

1. Course Content

This course introduces the basics of AI Largr models and the types and principles of Largr models used in AI embodied intelligence.

2. Introduction to AI Largr Models

AI Largr models, also known as large-scale pre-trained models, are the product of a deep fusion of big data, massive computing power, and powerful algorithms. Simply put, they are like intelligent agents that are "fed" with vast amounts of knowledge and repeatedly trained. By learning from vast amounts of data, they grasp the patterns and patterns within the data, thus possessing strong versatility and generalization capabilities. This capability enables AI Largr models to transcend single tasks and, like humans, flexibly apply their learned knowledge to solve complex problems across multiple domains. Leveraging multiple sources of data, such as internet text and images, they learn patterns, possessing strong generalization capabilities. They can adapt to a variety of tasks through transfer learning or prompt engineering, and may even develop capabilities such as logical reasoning and common sense understanding that were not previously trained.

3. Common AI Large Model Categories

3.1 Text Generation Large Models

- Based on the Transformer architecture, these models learn the grammatical, semantic, and pragmatic rules of a language through unsupervised or supervised learning on massive amounts of text data. They are then able to generate natural and fluent text based on input prompts or context. Pre-training through unsupervised learning makes them suitable for tasks such as text generation and dialogue systems. By converting all text tasks into a unified text-to-text problem, they provide a more flexible framework for handling diverse tasks such as translation, summarization, and question-answering.
- In content creation, they can automatically generate practical text such as articles, news, and reviews, improving content production efficiency and assisting authors with creative conception and text polishing.
- In the field of intelligent interaction, they can be applied to intelligent customer service and chatbots, generating natural and fluent responses and improving user experience.
- In personalized teaching, they can analyze questions, provide explanations of test points, problem-solving strategies, and results, and assist users with language learning.
- In machine translation, they can achieve automatic translation. Combined with speech models, they can also perform simultaneous interpretation and generate subtitles for everyday tasks.

3.1.1 Principle Overview

Based on the Transformer architecture (particularly the self-attention mechanism), this approach uses unsupervised/semi-supervised learning on massive amounts of text data to model the probability distribution and semantic associations of language, enabling natural language understanding and generation.

- Pretraining Logic

- Autoregressive (AR): Like the GPT series, this approach uses a causal language model to predict the next token (e.g., "Today's weather → Today's weather is very sunny"), learning contextual dependencies.
- Autoencoder (AE): Like BERT, this approach uses a masked language model to predict masked tokens (e.g., "Today [mask] is very sunny" → "Weather"), learning contextual bidirectional semantics.

- Key Technologies

- Attention Mechanism: Dynamically assigns weights to different words in a text, capturing long-range dependencies (e.g., "Previous event → Later event's influence").
- Prompt Tuning: Using templates (e.g., "Please summarize the following: {text}"), this approach activates specific model capabilities and adapts to downstream tasks. - Emerging Capabilities
  As the scale of parameters increases (e.g., into the hundreds of billions), models may develop capabilities not explicitly programmed during pre-training, such as logical reasoning, common sense understanding, and few-shot learning.

3.2 Large Multimodal Models

Capable of processing diverse input data types, such as text, images, audio, and video. Through cross-modal learning, models understand the relationships between data in different modalities and integrate multimodal data to fully leverage information from each modality, building a unified representation space. This allows data from different modalities to be understood and combined, enabling the execution of more complex and intelligent tasks. These models can be used for cross-modal retrieval, retrieving data from one modality based on data from another; in visual question answering, models can answer text questions based on image content; they can also generate image captions, generating natural language text describing the image content; and they can enable multimodal conversations involving information from multiple modalities. These models have broad application prospects in complex environments such as healthcare, transportation, and security monitoring.

3.2.1 Principle Overview

Through cross-modal alignment and joint modeling, a unified representation space for data from different modalities is learned, enabling semantic association and collaborative processing between modalities.

- Contrastive Learning: For example, the CLIP model maps the feature vectors of images and text into the same space and trains through "image-text pair matching" (e.g., "dog picture → text 'dog'").
- Encoder-Decoder Architecture: For example, DALL-E, the text encoder extracts semantic features, and the image decoder generates the corresponding image (text → image).

Fusion Methods

- Early Fusion: Combines multimodal data at the input layer (e.g., concatenating text embeddings with image pixel features).
- Late Fusion: Processes each modal data separately and fuses the results at the decision layer (e.g., first analyzing text sentiment and image color separately, then making a comprehensive judgment).

3.3 Speech Recognition Model

- Converts the input speech signal into text information. Typically based on deep learning algorithms, this approach first extracts features from speech signals. These features are then fed into a neural network model for training and recognition. The model learns from large amounts of speech data to identify different speech patterns and corresponding text content. This approach can help customer service personnel quickly record customer needs and issues, improving service quality and facilitating follow-up inquiries. It can also be applied to voice search, freeing hands and suitable for various search environments such as vehicle navigation and mobile phones. It can also convert meeting conversations into text, making it easier to organize and record meeting content. In terms of human-computer interaction, voice commands can be used to control intelligent devices, including hardware devices such as robots and software applications.

3.3.1 Principle Overview

The acoustic features of speech signals are converted into text sequences, and end-to-end modeling is achieved based on deep learning.

- Feature Extraction: Preprocess the speech waveform (e.g., framing and windowing) to extract Mel-frequency frequency coefficients (MFCCs) or acoustic feature vectors (e.g., using a CNN).
- Sequence Modeling: Use a recurrent neural network (RNN/LSTM) or Transformer encoder to capture the temporal dependencies of speech sequences (e.g., "consecutive syllables → words"). * Decoder Mapping: Maps feature sequences to text sequences (e.g., acoustic features → "hello") using Connectionist Temporal Classification (CTC) or an attention mechanism.

3.4 Speech Synthesis Model

Converts input text into speech signals. This is typically accomplished by training a model to learn the text-to-speech mapping relationship. The model generates corresponding speech features based on the input text, and then uses speech synthesis technology to convert these features into audible speech. This is widely used in voice assistants, audiobooks, intelligent customer service, and other fields, providing users with voice interaction services and enabling devices to communicate with users in a natural and fluent voice.

3.4.1 Principle Overview

Converts text semantics into natural and fluent speech signals, simulating the rhythm, intonation, and emotion of human pronunciation.

Deep Learning Synthesis:

- Text Analysis: Uses NLP models to analyze the semantics, part of speech, and sentiment of text (e.g., "I'm very happy today" → cheerful intonation).
- Acoustic Modeling: Generates the mel-spectrogram of speech using the Tacotron family (encoder-decoder + attention mechanism). * Vocoder: Converts mel-spectrograms into waveform signals, such as WaveNet and HiFi-GAN (improves speech naturalness).

4. AI Model Comparison Summary

Model Type	Input	Output	Core Technology	Typical Scenarios
Natural Language Model	Text	Text	Transformer Self-Attention	Writing, Conversation, Translation
Multimodal Model	Text + Image/Audio	Cross-modal Content	Cross-modal Alignment, Joint Encoding	Image-Text Generation, Visual Question Answering
Speech Recognition Model	Speech Waveform	Text	Acoustic Feature Extraction + Sequence Decoding	Meeting Minutes, Voice Search
Speech Synthesis Model	Text	Speech Audio	Text Analysis + Acoustic Modeling + Vocoder	Voice Assistant, Audio Content Production