MLLMS

Introduction

- Two Traits Compared with Traditional Counterparts
  - Being Based on LLMs
  - Using Instruction Tuning
- Expanded Capabilities
  - Better Granularity Support
  - Enhanced Support on Input and Output Modalities
  - Improved Language Support
  - Extension to More Realms and Usage Scenarios

Architecture

- Modality Encoder
  - Pretraining Corpus
  - Resolution
    - Direct Scaling
    - Patch-Division
  - Samples
  - Parameter Size
- Pre-trained LLM
  - Flan-T5-XL/XXL
  - LLaMA
  - Vicuna
  - LLaMA-2
  - Qwen
- Modality Interface
  - Learnable Connector
    - Token-Level Fusion
      - Q-Former-Style
    - Feature-Level Fusion
  - Expert Model

Training Strategy and Data

- Pre-training
  - Training Detail
    - Freezing Visual Encoders and LLMs
    - Unfreezing More Modules
  - Data
    - CC
    - SBU Captions
    - LAION
    - COYO-700M
- Instruction Tuning
  - Training Detail
    - Optional Instruction
    - Input-Output Pair
  - Data Collection
    - Data Adaption
    - Self Instruction
    - Data Mixture
- Aligning Tuning
  - Training Detail
    - RLHF
      - Policy Model
      - Reward Model
      - Reinforcement Learning
    - DPO
  - Data
    - LLaVA-RLHF
    - RLHF-V
    - VLFeedback

Evaluation

- Close-Set
  - Zero-Shot
  - Finetuing
- Open-Set
  - Manual Scoring
  - GPT Scoring
  - Case Study

Extension

- Granularity Support
- Modality Support
- Language Support
- Secnario/Task Extention
  - Multimodal Agents
  - Augmened MLLMs

Multimodal Hallucination

- Categories of Hallucination
  - Existence Hallucination
  - Attribute Hallucination
  - Relationship Hallucination
- Evaluation Methods
- Mitigation Methods
  - Pre-correction
  - In-process-correction
  - Post-correction

Extented Techniques

- Multimodal In-context Learning
  - Improvement in ICL Capabilities
    - Instruction Tuning
    - Introducing Extra Modalities
    - Using Specific Settings
  - Applications
    - Solving Various Visual Reasoning Tasks
    - Teaching LLM to Use External Tools
- Multimodal Chain of Thought
  - Learning Paradigms
    - Finetuning
    - Training-Free Few/Zero-Shot Learning
  - Chain Configuration
    - Structure
      - Single-Chain Methods
      - Tree-Shape Methods
    - Length
      - Adaptive Formations
      - Pre-defined Formations
  - Generation Patterns
    - An Infilling-Based Pattern
    - An Predicting-Based Pattern
- LLM-Aided Visual Reasoning
  - Several Good Traits
    - Strong Generalization Abilities
    - Emergent Abilities
    - Better Interactivity and Control
  - Training Paradigms
    - Training-Free
    - Finetuing
  - Functions
    - LLM as a Controller
      - Breaking Down a Complex Task into Simpler Sub-tasks/Steps
      - Assignsing Tasks to Appropriate Tools/Modules.
    - LLM as a Desicion Maker
      - Summarizing the Current Context and the History Information
      - Decide If the Information Available at the Current Step Is Sufficient
      - Organizing and Summarizing the Answer to Present It in a User-Friendly Way.
    - LLM as a Semantic Definer

Challenges and Future Directions

- Limited Context Length
- Following Complicated Instructions
- M-ICL and M-CoT
- Embodies Agentd Based MLLMs
- Safety Issues