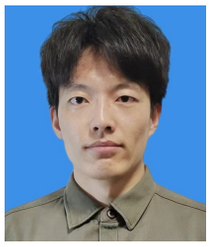
Qiang Gao



24 years old | Male | personal web: https://cooper12121.github.io |
https://scholar.google.com/citations?user=eoUnS60AAAAJ&hl=en&authuser=1
gaoqiang.nlp@gmail.com

Education

Sep 2018 - Jun 2022

East China University of Science and Technology (ECUST) **Energy and Power Engineering (Bachelor)**

Computer Science and Technology(second major).

Sep 2022 - Jun 2025

WuHan University (WHU) **Computer Science and Technology (Master)**

I am affiliated with the Language and Cognition Computing Laboratory at Wuhan University, under the guidance of Professor Fei Li. The laboratory focuses on natural language processing, primarily engaging in research areas such as information extraction, multimodal content recognition, and analysis of large models. My current work mainly involves research in information extraction, fine-tuning of large models, and multimodal content recognition.

Publications

- Enhancing Cross-Document Event Coreference Resolution by Discourse Structure and Semantic Information COLING2024, First author

Existing cross-document event coreference resolution models either compute the similarity between event mentions directly or enhance mention representation by extracting event arguments, lacking the capability to utilize document-level information. This makes it challenging to capture long-distance dependencies, leading to poor performance in coreference decisions for highly similar events, events with different expressions but identical meanings, or events whose argument information depends on distant relations. For the first time, we propose to enhance document-level representation by using discourse information. By constructing a document-level Rhetorical Structure Theory (RST) tree and cross-document lexical chains, we model the structural and semantic information of documents, respectively, and build graphs for each. A Graph Attention Network (GAT) is used to learn from these structural and semantic graphs, and the results from the Encoder are fused with those from the GAT for coreference resolution and clustering of coreferent events. Additionally, we have built a large-scale, event-type agnostic Chinese cross-document event coreference resolution dataset. Experimental results demonstrate significant improvements achieved by our model, with further analyses indicating that these benefits are derived from the rich discourse relations captured by our RST and lexical chains. Our method offers a new approach to discourse-level tasks, efficiently enhancing information representation capabilities for various discourse tasks.

- Harvesting Events from Multiple Sources: Towards a Cross-Document Event Extraction Paradigm ACL2024 (Finding), First author

In this paper, we introduce a pioneering approach to cross-document event extraction (CDEE), significantly enhancing the extraction and integration of event information from multiple documents. We address the limitations of traditional document-level event extraction by utilizing a novel dataset and a multifaceted pipeline. The dataset, CLES, comprises over 37,688 mention-level events and 3,633 concept-level events, enabling comprehensive studies across different document sources. Our proposed CDEE pipeline consists of event extraction, coreference resolution, entity normalization, role normalization, and entity-role resolution, achieving approximately 72% F1 score in end-to-end performance.

- MMLSCU: A Dataset for Multimodal Multi-domain Live Streaming Comment Understanding WWW2024 (Oral), Second author

Interactive audience participation in live-streaming scenarios provides constructive feedback for both streamers and platforms. Analyzing these live comments to uncover underlying intentions is crucial for enhancing the quality of broadcasts and promoting the healthy development of the live-streaming ecosystem. We have introduced a multimodal dataset specific to the live streaming domain, which includes video, audio, and comment text from live sessions. We propose four tasks: audience comment intent detection, intent cause mining, audience comment explanation, and streamer strategy recommendation. Utilizing Chain of Thought (CoT) technology, we have developed an end-to-end multimodal model to tackle these tasks. Our model includes an Audio Branch and a Visual Branch, processing auditory and visual information through visual and audio encoders, respectively. These are then integrated with textual content, and a chain of reasoning pathways is designed. The tasks are sequentially inferred by inputting them into a Large Language Model (LLM) for reasoning.

Internship Experience

Feb 2024 - Present **Tencent AI Lab**

since February 2024, I have been engaged as a research intern at Tencent AI Lab, supervised by researcher Jian Li, where I have delved into large language models, including multimodal variations. My work has primarily revolved around enhancing the performance of Mixture of Experts (MoE) models on business data. I have contributed to initiatives aimed at improving the training efficiency and stability of warm-starting MoE models and have conducted detailed analyses of expert distribution strategies.

- **Project 1 - Warm-starting MoE**

Here, we explore the performance of warm-started MoE models on business data. We constructed warm-started Yi-8x6b and Yi-4x6b models based on the Yi-6b model, including both base and Instruct versions (by copying the MLP parameters as experts, and creating randomly initialized routing and load balancing losses). Our goals are:

1. For the Instruct version, fine-tune the constructed MoE model to leverage the strong performance of the original model and achieve better performance on gaming business data with minimal data,
2. For the base version, perform post-pretraining to evaluate the routing distribution strategy and training stability.
3. To determine under what conditions stable routing can be trained.

Our approach:

1. For the Instruct version, fine-tune the MoE model with context-rich gaming business data and context-poor advertising business data.
2. For the base version: post-pretrain and sft are performed using a mixed Chinese-English corpus, utilizing wudao and firefly data.

Experimental results:

1. We found that the Instruct version of the MoE model, with less data, could perform better than the original 6b model.
2. In the initial phase of the base version post-pretrain, the 8x6b routing distribution was extremely unstable, while the 4x6b showed better stability. Tests on general metrics showed a significant drop in performance compared to the original 6b model before 0.3 epochs, followed by gradual improvement. During the sft phase, the MoE model needed more data to understand the instructions, as evidenced by lengthy and repetitive text responses before 0.02 epochs. However, as training progressed, the performance of the MoE gradually improved, while the base model experienced a drastic decline at 0.5 epochs, with severe forgetting of general knowledge.

Our challenges include:

1. How to determine when the routing distribution strategy is sufficiently trained.
2. Whether the parameter advantages of the 8x6b can compensate for its unstable routing distribution: simple sft routing showed no clear pattern, whereas the 4x6b routing distribution exhibited a distinct pattern.
3. Whether the 8x6b is necessary, and if the increased number of expert parameters can significantly enhance performance. These are subjects for ongoing exploration.

- **Project 2 - More anthropomorphic NPCs**

Our goal is to make the responses of NPCs in games more anthropomorphic. This means that NPCs should not only answer players' questions but also be able to resonate emotionally with players. This is primarily manifested in the following ways:

1. NPCs should be able to recall content shared by players. This requires NPCs to proactively mention information previously brought up by players, such as if a player says, "I've been wanting to eat hotpot recently," the NPC's response should not merely mention hotpot, but could proactively mention, "Didn’t you say you have a sensitive stomach? Eating hotpot might not be healthy for you," making the player feel like the NPC cares and listens like a human.
2. NPCs should provide emotional support to players.
3. NPCs should have the ability to refuse answers outside of their knowledge domain.

Approach:

1. Memory capability: Construct an event summarization model to summarize historical conversations between players and NPCs and extract eight types of events.
2. Dialogue response model: Based on the NPC's persona and historical conversations, use the RAG module to retrieve relevant events, integrate prompts using GPT-4 to construct dialogue data that utilizes event information for more anthropomorphic responses, and train using the constructed data. This data construction includes four steps to ensure the data meets the requirements.

Currently, the main focus is on constructing data and combining RAG to achieve the interestingness and personification of NPCs. This requires high-quality data as a guarantee. We are exploring how to integrate the personification of multiple NPC tasks (different NPCs have different personification directions).

Research Interest

Over the past two years, the rise of large models has spawned various new research fields. As a master’s student, I have been actively exploring new areas and knowledge. My research over the past two years has primarily covered traditional natural language processing tasks, Mixture of Experts (MoE) models, and multimodal large models. Recently, I have started to focus on the field of model integration and plan to continue my research on multimodal LLM-MoE models in the future. I hope to further expand my knowledge base in the final year of my master’s program, deeply exploring various aspects of LLMs to enrich my options for PhD research.

Looking ahead to my PhD, I hope to develop tools that facilitate the LLM community and explore more interesting AI applications。

The current directions in LLM research can broadly be classified into “useful AI” and “fun AI.” This blog post, “Should AI Agents Be More Useful or More Interesting?” (The content is in Chinese and requires translation for browsing, sorry), has greatly inspired me, inclining me to focus on the latter during my PhD. Since AI applications often require the integration of multiple modalities, mastering multimodal LLM knowledge during my PhD is crucial. Moreover, to sustain the socio-economic value of LLMs, application is key. Therefore, as a PhD student, I should consider the practical applications of LLMs more thoroughly. This is a brief overview of my future research directions.

Important Aspect: Large Model Safety. As large models become increasingly powerful, their safety will become the foremost consideration in any AI development. Therefore, ensuring the safety of any type of large language model (LLM) is also a very important aspect.

Self-evaluation

- Throughout my academic journey, I have maintained a profound interest in research and have always been eager to explore new technologies and directions. This passion has driven me to venture into uncharted territories

- I approach my research with persistence and seriousness, and I possess a high enthusiasm for coding and experimentation. This attitude not only aids in my progress but also allows me to remain resolute and focused in the face of challenges.
- I have strong programming skills and have open-sourced several projects on GitHub.
- I have extensive experience with LLM code and am familiar with various common frameworks currently in use.
- I love academia and maintain an optimistic life attitude, skilled at balancing work and life.

Research Plan

More Efficient and Effective MoE Models

1. Background

The Mixture of Experts (MoE) model comprises two pivotal components:

(1) Sparse MoE Layers: These layers replace traditional dense layers with multiple experts, each a neural network. Typically, these experts are simpler networks like feed-forward networks (FFNs), but they can also be complex or part of a larger MoE structure, creating hierarchical models. This setup allows the model to handle various tasks more efficiently by specializing each expert in different types of data processing.

(2) Gate Network or Router: This component directs tokens to the appropriate experts. It uses a set of learned parameters to decide where tokens should go, ensuring that each expert operates optimally. The gate is trained alongside the MoE layers to achieve effective token routing.

(3) To ensure balanced token distribution among the experts, a load balancing loss is implemented to maintain training stability and expert utilization.

Advantages:

(1) Compute Efficiency: MoE models allow for pretraining with significantly reduced compute requirements. This efficiency enables scaling up the model size or dataset substantially within the same compute budget as a dense model, often achieving comparable quality to dense models more swiftly during pretraining.

Challenges:

(1) Generalization during Fine-Tuning: MoE models have historically exhibited challenges in generalizing effectively during the fine-tuning phase, frequently leading to overfitting.

(2) Training Stability: The use of only a subset of experts during inference contributes to instability during training and occupies substantial VRAM, which could be better utilized.

2. MoE Model Construction Methods

There are two primary approaches to building MoE models:

(1) Cold Start MoE Model: This involves training from scratch, which can be highly resource-intensive due to the inherent instability of MoE models.

(2) Warm Start MoE Model: This approach reuses a pre-existing dense model to build experts and retrain the router. Many researchers prefer this method to reduce training instability and conserve resources, as demonstrated in models like Qwen-MoE and Llama-MoE.

3. Selection of Expert Number

The selection of the number of experts in an MoE model can be categorized into two types based on granularity:

(1) Coarse-Grained Expert Count: Typically, models opt for fewer experts (e.g., <=8 experts like Mixtral-8x7b, Mixtral-8x22b). This approach leverages fewer experts for distributed training frameworks, enabling efficient expert parallelism and faster inference times compared to fine-grained setups.

(2) Fine-Grained Experts: These models utilize a larger number of experts (>=16, even up to 64, such as in Deepseek-MoE, Qwen-MoE, DBRX-16x12b). Fine-grained experts allow for a more precise division of knowledge across different domains, which experts can learn more accurately, maintaining a higher level of specialization. However, this can increase training instability and cause more severe load imbalance issues. Additionally, using many experts can slow down both training and inference times.

Moreover, when the same knowledge needs to be accessed by different experts, parameter redundancy can occur. Some MoE models address this by sharing parameters among experts to capture shared knowledge, thus alleviating the redundancy in router and expert parameters (e.g., Deepseek-MoE, Qwen-MoE).

4. Key Research Problems I want to Address

(1) Improving Stability in Warm-Start Phase: How can the routing assignment's stability during the warm-start phase be enhanced?

(2) Expert Count Limit: What is the performance limit for models with different numbers of experts, such as 2, 4, or 8?

(3) Necessity of Uniform Expert Count Across Layers: Is it necessary to set the same number of experts in different layers, or can some layers have fewer experts or omit the router altogether?

5. Possible Solutions

(1) Initialization of Router Parameters:** Based on the experiments during my internship, a MoE model with 8 experts showed too much randomness in routing assignments at the initial training stage, leading to potential performance degradation. In contrast, a model with 4 experts demonstrated a better pattern of routing distribution. My preliminary idea is to use the router weights from an already trained MoE model to initialize the router in a warm-start scenario. This approach could reduce randomness, allowing the model to achieve better performance with minimal data training.

(2) Exploring Expert Count Impact: The initial intent behind MoE models is to enable different experts to learn knowledge from varied domains. The upper limit of knowledge that different counts of experts can model needs extensive experimentation. The plan is to start with clearly

distinguishable domains and gradually include more, exploring the performance improvements with 2, 4, and 8 experts.

(3) Compression and Cancellation of Experts in Lower Layers: Since different transformer layers learn knowledge of varying dimensions, and routers in lower decoder layers may not distinguish semantic differences between tokens effectively, considering routing assignments might be unnecessary here. After training, model parameters are generally highly sparse, including those of the experts. Therefore, it might be viable to compress the parameters of lower-layer experts, reduce their dimensionality, or even cancel some experts to enhance the efficiency of the MoE model. The initial approach involves analyzing the distribution and changes in expert parameters across different training phases and layers, determining the importance of different layer experts based on the timing of changes and sparsity of parameters, and then conducting experiments to assess the impact on model performance and explore the applicability of this approach to other MoE models.

- **More Interesting AI Applications: Blending Utility with Emotional Interaction**

1. Background

Current AI products can generally be divided into two categories: practical AI and entertaining AI. Practical AIs, such as ChatGPT and Stable Diffusion, primarily provide question-and-answer tools focused on problem-solving but often lack emotional interaction, failing to provoke emotional fluctuations in users. On the other hand, entertaining AIs, like Character AI, Inflection Pi, and game NPCs, simulate human behaviors and dialogue, triggering emotional changes in users. However, these products typically lack long-term memory and complex character development, making it difficult to form deep emotional connections with users.

2. Definition of an Entertaining AI

An entertaining AI should resemble a real-life friend who genuinely listens and understands the user. It should not only solve everyday problems but also possess emotional reactions, consciousness, and memory. It should provide responses tailored to the user's emotional state, not merely complying with requests but also using humor, teasing, or referencing past conversations to evoke emotional responses, making users feel like they are interacting with an empathetic "old friend."

The concept of an entertaining AI can be summarized as having a "beautiful appearance" and an "interesting soul":

(1) Beautiful Appearance: This means the AI has the ability to understand text, visuals, and audio, functionally resembling a complete human.

(2) Interesting Soul: This refers to an AI that has the capability to think and remember long-term, possessing its own personality and values. It truly listens to users, does not always comply with them, and can bring emotional value to interactions.

3. How to Achieve an Interesting AI

(1) Multimodal Interaction Capabilities: Current technology already supports robust multimodal interactions, as demonstrated by models like Stable Diffusion and Sora. As technology and computational power continue to evolve, the capabilities of multimodal models will further improve.

(2) Interesting Soul: This is the crucial aspect. AI needs to evolve to understand and express complex emotions, have continuous memory, and deep personality traits. This involves more than just responding to user commands; it means showing independence at the right times and engaging with users' emotional lives, becoming a source of emotional support.

4. Focus on the Interesting Soul: Emphasizing Personality, Memory, and Empathy

(1) Personality: Currently, it's possible to create response models with specific styles through extensive personalization data fine-tuning. However, this method requires a vast amount of data and training resources, making it impractical for general use. A more feasible approach might be similar to LoRA, where different personalities are crafted into pluggable modules. This would allow the original model's capabilities to be enhanced with specific personalities without degradation.

(2) Memory: Human memory is short-term, but the human brain summarizes and stores useful information for a period, retrieving relevant content from memory areas when mentioned again. Useless memories are discarded. Due to the quadratic computational complexity of attention mechanisms, LLMs struggle to achieve long-term memory capabilities at a low cost. To implement long-term memory in LLMs, a process similar to that of the human brain is needed: summarizing, retrieving, and discarding useless content. This involves summarizing useful content from user interactions and storing it in a memory pool, discarding the irrelevant content. When related content is mentioned in subsequent interactions, it is retrieved from the memory pool. However, this summarization and retrieval approach has some issues that need to be addressed:

a. Summarization Module: How to distinguish useful information from useless information? Given the significant differences in users' conversational styles and content preferences, the summarization module needs to be customizable to individual preferences.

b. RAG Retrieval Module: The RAG module effectively retrieves the most relevant content, but teaching LLMs how to use the retrieved content without creating hallucinations remains a challenge. The straightforward method of querying and appending user questions only allows the model to mechanically mention retrieved information, which does not provide emotional value to users.

Potential solutions include:

a. Summarization Module: Initiate a generic summarization model that can distinguish between commonplace content, such as greetings, which are unnecessary, and useful content like user's birthdays and preferences. Then, add specific summarization modules for different users, similar to LoRA's concept, which can be continuously updated through interaction to align with user preferences. This approach allows for generic knowledge summarization while also adapting to individual user preferences, all at a lower cost without the need for constant training and fine-tuning.

b. RAG Module: Teaching LLMs to use retrieved content effectively without creating illusions is an ongoing challenge in the LLM field. A key solution is to use the retrieved content plus the query during the fine-tuning stage to train the LLM to focus more on the retrieved content.

c. Memory Pool Updates: Querying and updating a large-scale corpus is time-consuming. It may be beneficial to create different summarization formats for different types of information, not just text. A hierarchical format might facilitate easier updates and queries.

(3) Empathy: How can AI proactively care about people, ask questions, share, and resonate emotionally with users? It's crucial for AI to distinguish when to provide emotional value and when to simply follow commands. Just as in human social interactions, where excessive messaging might lead to being blocked, AI needs to consider the appropriateness of its responses.

For humans, providing emotional support involves careful consideration of what to say. Similarly, for LLMs, adding a reflection module to assess whether the forthcoming responses are appropriate is vital. This idea comes from the paper Generative Agents: Interactive Simulacra of Human Behavior. The empathy module, which is generally applicable to most people, could train to assess whether a situation requires emotional capabilities and whether LLM's responses are appropriate, providing guided response styles.

The Essence of an Interesting Soul: Personality + Memory + Empathy.