Feb 15, 2025

DeepSeek deep dive: LLM Agents Study Session

Summary

Yatharth Piplani and Arunesh Mishra led a meeting covering DeepSeek model architecture (versions V1-V3, focusing on multi-token prediction) and reinforcement learning (RL) fundamentals, including Gradient Regularized Policy Optimization (GRPO) for enhancing DeepSeek's reasoning capabilities. Discussions involved RL concepts, policy learning, GRPO's advantages over MCTS, and challenges in LLM fine-tuning, along with practical implementation details, computational efficiency, and multilingual reasoning experiments. Next steps include sharing resources for implementing the GPRO algorithm, exploring data pre-processing techniques, and conducting further sessions to delve deeper into the algorithm's concepts and multilingual reasoning capabilities.

Details

- **Meeting Introduction and Agenda:** Yatharth Piplani and Arunesh Mishra began the meeting. They agreed that Arunesh would discuss DeepSeek and Yatharth would cover the basics of reinforcement learning (RL) and Gradient Regularized Policy Optimization (GRPO). Yatharth experienced a technical difficulty, requiring a brief pause.
- **DeepSeek Models Overview:** Arunesh Mishra provided a history of DeepSeek models, highlighting key architectural innovations in versions V1 through V3, including multi-head latent attention and mixture of experts. They noted DeepSeek R1's reasoning capabilities, comparing its performance to OpenAI

models. Arunesh also explained DeepSeek V3's multi-token prediction architecture, contrasting it with single-token prediction methods.

- **DeepSeek Multi-Token Prediction Discussion:** Arunesh Mishra explained the DeepSeek V3 multi-token prediction mechanism, clarifying that while it appears parallel, it's sequentially optimized using latent information passed between thinner transformer blocks. They addressed questions from Joydeep Ghatak and JT regarding the architecture's details, particularly the contextual information fed to subsequent tokens. The discussion touched on potential quality degradation in favor of speed and the tunable nature of the trade-off.
- **Reinforcement Learning (RL) Basics:** Yatharth Piplani introduced fundamental RL concepts, including agents, environments, states, actions, rewards, and the distinction between model-based and model-free RL. They discussed different types of reasoning in humans (word-based and shape-based) and the limitations of LLMs in handling shape-based reasoning. Yatharth also explained the nature of rewards (deterministic or probabilistic) and the continuous feedback loop between agent and environment. They further detailed the representation of states and actions (discrete vs. continuous action spaces). The discussion included an example comparing the discrete action space of a chess engine to the continuous action space of language models for solving math problems.
- **RL Policies and Trajectories:** Yatharth Piplani defined policies (algorithms used by agents) as deterministic or stochastic (probability-based). They explained how policies are learned and refined through actions, rewards, and the updating of model parameters. The concept of trajectories (sequences of states and actions) was also introduced.
- **GRPO and DeepSeek Reasoning:** Yatharth Piplani discussed the role of GRPO in adding reasoning capabilities to DeepSeek, contrasting it with the ineffective Monte Carlo Tree Search (MCTS). They described the challenges of fine-tuning LLMs, highlighting the use of human-labeled datasets and the cost associated with this approach. The importance of structured learning, including example problems and reasoning traces, was emphasized. Yatharth explained how GRPO addresses the issue of reward hacking and its use in continuous action spaces. They concluded by explaining the core GRPO equation and its importance in balancing exploration and exploitation.
- **GPRO Algorithm Explanation and Code Implementation:** Yatharth Piplani explained the GPRO algorithm, emphasizing its practical considerations for

computational efficiency. They detailed the algorithm's steps: considering all possible scenarios, making small updates with equal chances for all participants, clipping possibilities to a specific range, and using KL divergence to keep the model in check. The code implementation uses the GSM8K dataset, focusing on rewarding correct answers and adhering to a specific template. They discussed the use of samplers to select active tokens during inference, along with training a smaller, more computationally efficient matrix to update the model's weights. The presenter also mentioned using a decoder-only transformer for next-token prediction, training specific layers within the LoRA framework.

- **Algorithm Resources and Further Sessions:** Shristi Gautam expressed appreciation for the presentation and requested resources for implementing the GPRO algorithm from scratch. Yatharth Piplani offered resources based on OpenAI's spinning up documentation and Uber's RL guides. They also proposed further sessions to delve deeper into the algorithm's concepts. Arunesh Mishra also expressed appreciation and suggested more sessions, provided the presenter was willing. The code used was confirmed to be available via a shared Google Colab link.
- **Code Execution, Fine-tuning, and Model Sharing:** Joydeep Ghatak inquired about code execution time and resource usage on Google Colab. Yatharth Piplani suggested stopping the execution after a certain time and checking the results. Arunesh Mishra shared they had fine-tuned a model and uploaded it to Hugging Face, encouraging others to do the same for comparison. Shristi Gautam requested the Hugging Face link, which was shared, along with a note that it was not fine-tuned on the J dataset. Yatharth Piplani suggested exploring data pre-processing techniques for using different datasets.
- **Multilingual Reasoning and Future Experiments:** Shristi Gautam proposed experimenting with the model's reasoning capabilities in different languages, such as Hindi. Yatharth Piplani mentioned Sutra's work on multilingual reasoning, suggesting fine-tuning the model on Hindi examples. They also discussed potential experiments involving training on HC Verma physics examples to solve J physics questions.
- **Clarification on the GPRO Equation and Pyref:** Arunesh Mishra raised questions about the GPRO equation, specifically regarding the `pyref` term. Yatharth Piplani offered initial guesses, suggesting it might be a rewarded model or a model following a specific format, promising to investigate further. Arunesh Mishra speculated on alternative possibilities, including the use of a teacher model or a

0/1 model. The discussion further explored the role of pyref, its potential as a reference model, baseline, or guide in the learning process, and its significance within the GPRO equation. They concluded that pyref likely acts as a stochastic reference model guiding the model's learning process, preventing it from straying too far from desired behavior.

You should review Gemini's notes to make sure they're accurate. Get tips and learn how Gemini takes notes

Please provide feedback about using Gemini to take notes in a short survey.