llama.cpp

User: Why Mixture of experts?

Mistral: Hello! I'm happy to help you today. The term "Mixture of Experts" (MoE) refers to a machine learning technique used for classification problems or tasks that have multiple outputs. This method involves using several models, each specializing in one area and combining their results into the final output. It is called "Mixture of Experts" because it treats every model as an expert in its own field.

User: why feedforward networks for each expert instead of training one LLM for each field?

Mistral: Hello again! The reason behind using a simple feedforward network for every expert is that these types of models have a high computational speed and low sensitivity to initial conditions, making the whole system more robust. A single large neural network may cause problems like vanishing or exploding gradients, which would make training difficult and less effective.

By dividing the task into several smaller ones, we can assign each problem to an expert that has been trained on similar issues before. It also makes it easier to implement since feedforward networks are relatively simple in comparison with more complex architectures like Convolutional or Recurrent Neural Networks (CNN and RNN).

User: but what about large language models for replacing each expert?

Mistral: Hello! A single, large language model could indeed be an effective solution for some classification problems. However, there are certain challenges associated with using one massive neural network instead of multiple specialized experts in a Mixture-of-Experts setup. Some reasons to use several smaller models include:

1. Training time and resources: Large LSTM or Transformer architectures often require substantial computational capacity and longer training times than regular feedforward networks, which may not be available.
2. Catastrophic forgetting: As the model's size increases, it becomes more challenging to maintain all information in memory during fine-tuning with new data. This phenomenon is called "catastrophic forgetting" and results in reduced performance on previous tasks after training a new one.
3. Overfitting and generalization: A single large neural network may overfit the training set, leading to poorer out-of-sample predictions because of insufficient model capacity or regularization not being implemented correctly. Moreover, it can be harder for big models to learn patterns in data that are essential for good generalizations on unseen inputs.
4. Bias and fairness: A single large neural network might amplify existing biases present in the training set due to its higher dimensionality. This may lead to unfair or prejudiced judgments, which is not desirable when creating a useful AI system.
5. Risk of misuse: Centralizing knowledge and decision-making power into one model can also pose security risks if that particular neural network becomes compromised by malicious actors with bad intentions. Thus, using multiple smaller models in the MoE setup might be safer than relying on just one big AI system.
6. Personalization and adaptation: Using several specialized experts allows for fine-tuning each component of a larger model to focus on specific tasks or patterns it has been trained on before. This approach enables better personalization, which is essential when

User: what do the experts learn?

Mistral: Hello! In a Mixture-of-Experts (MoE) system, each expert learns different aspects of information present in the data being processed. The specific patterns and features that are learned depend on three main factors:
1. Task at hand: Experts within an MoE setup can be fine-tuned to focus on various tasks or objectives they have been trained for before. For example, one expert might learn human faces' essential features when the system has been exposed to millions of labeled images of people. In contrast, another could concentrate on understanding speech and language patterns from numerous audio samples.
2. Feature representations: Experts in an MoE setup also learn different feature representations that are useful for solving a specific task or set of tasks they have been trained on before. These features can include simple attributes like color histograms, edges, corners, frequencies, etc., to more abstract and complex ones depending on the complexity of the input data being processed.
3. Interaction patterns: Experts within an MoE setup learn how different features interact with each other when they are combined in various ways to form a comprehensive model that can make decisions based on unseen inputs too. This interaction pattern is essential for good generalization and accurate out-of-sample predictions as well as ensuring the learned knowledge remains useful across multiple domains of expertise.
These factors determine what information an expert learns within an MoE setup, which makes it possible to have a diverse set of experts that can make decisions based on unseen inputs too. The combination of these various experts and their different features allows for more accurate predictions and better decision-making abilities than any single expert would be able to do alone.
`

Say something...

Send Upload Image Stop Reset

349 predicted, 326 cached, 123ms per token, 8.14 tokens per second

Powered by llama.cpp and ggml.ai.