AI and Interpretability: Vector's AI Engineering team has released a new interpretability framework for generative models, providing researchers with rich tools to improve the safety and trustworthiness of these models.

In the world of machine learning, we're witnessing the rise of mammoth neural networks with billions of parameters. These Large Language Models (LLMs) have demonstrated incredible abilities, primarily due to their generalization and in-context learning capabilities. But with this massive growth in model size, we've run into a significant challenge: the increased hardware requirements for their training and deployment. This often requires distributed infrastructure, splitting the model across multiple graphics processing units (GPUs) or even multiple nodes.

Although many tools exist for model parallelization and distributed training, deeper interactions with these models (e.g., retrieving intermediate information or editing the model) necessitate a strong grasp of distributed computing. This has been a roadblock for many researchers with expertise in machine learning but limited knowledge in distributed computing. As a result, these large models typically function inside a black box, making it hard to pinpoint specific reasons for a given output in a manner that's easily interpretable for humans.

What is FlexModel?

Enter FlexModel, introduced in the research paper, "FlexModel: A Framework for Interpretability of Distributed Large Language Models." and selected as a spotlight paper in the Socially Responsible Language Modelling Research (SoLaR) workshop at NeurIPS 2023.

FlexModel is a software package designed to provide a user-friendly interface for interacting with large-scale models spread out over multi-GPU and multi-node setups. It accomplishes this by wrapping around large models regardless of how they've been distributed (using popular libraries like Accelerate, FSDP, or DeepSpeed). Next, it Introduces the concept of HookFunctions that let users interact with distributed model internals, both during forward and backward passes. It implements these mechanisms via a simple API that has been released as a Python library called FlexModel (https://pypi.org/project/FlexModel). By implementing this library into their projects, researchers can quickly and easily gain rich insights into why a model is behaving the way that it is.

What does this mean for the machine learning community?

FlexModel promises to democratize model interactions and bridge the gap between distributed and single-device model paradigms. This enables researchers who may not be experts in distributed computing to interact with and modify distributed models without diving deep into the complexities of distributed systems.

This interpretability has many important ramifications in the areas of AI trustworthiness and safety. Concerns about biases and fairness in AI have gained prominence. Interpretability can help in detecting, understanding, and mitigating hidden biases in model decisions.
For stakeholders and end-users to trust machine learning models, especially in critical applications like medicine, finance, and judiciary, they need to understand how these models arrive at their decisions. An interpretable model fosters trust. In many sectors, models that make decisions impacting humans must be interpretable due to regulatory mandates. This is to ensure that decisions are made transparently and can be accounted for. When a model's decisions can be understood, it's easier to diagnose why it might be making errors and subsequently refine or correct the model.

Why is this significant?

Interpretability is becoming increasingly important, especially with the advent of large, closed-source models like ChatGPT. Unraveling how these models arrive at decisions, how they've learned specific behaviors, and understanding their internal mechanics can give us insights into building more robust, trustworthy, and efficient AI systems. With tools like FlexModel, researchers can now engage in interpretability research without being burdened by the technical complexities of distributed computing.

Conclusion

The emergence of FlexModel is a promising development in the machine learning ecosystem. By lowering the barriers to interpretability research in LLMs, FlexModel brings us a step closer to making state-of-the-art machine learning more accessible, interpretable, safe and trustworthy.

Whether you're a machine learning aficionado or just an interested observer, the arrival of tools like FlexModel underscores the significance of making advanced AI research inclusive and universally approachable.