Abstract

The Bittensor network has evolved from a playground for conversational LLMs into a ranking & incentive framework to aggregate digital commodities on the blockchain. Each subnet defines a different rewards mechanism, uniquely tailored to foster competition between miners, driving the quality of digital commodity. While this is an important feature of the Bittensor network, the competitive landscape is limiting as Bittensor scales & a diverse set of rewards mechanisms are necessary for Bittensor to reach its potential. We propose Homogeneous Inference Grids with the end goal of optimizing for Edge Node Inference– allowing Bittensor to support end user demands at effectively infinite scale. By serving as the scalability arm of Bittensor, Homogeneous Inference Grids can work in synergy with competitive subnets to drive both model quality and model quantity to maxima.

Text-to-Video Generation Grid for Edge Node Inference

Authors: theAdoringFan and girlfriend-deficit
Fractal Research

February 10, 2024

Introduction

Existing Bittensor subnets implement incentive mechanisms that are aimed towards competition. These subnets are focused on quality of model. In the case of machine learning models, this is quality of response, introduction of net new material, diversity of knowledge, and more. This drives quality of model to improve– larger models, uniquely fine-tuned models. Given the head start that centralized AI companies have, intense model competition is critical to the success of Bittensor. However, adversarial environments are a solution only for optimizing model quality. Accessibility is equally as important to attaining marketshare from the closed-source cloud giants. Providing access to quality models at scale across non-homogeneous environments is not only likely not achievable, but certainly not ideal. App developers and users want consistent access to a model. Current adversarial networks are not suitable for providing access at scale. We thus propose the use of Homogeneous Inference Grids to supplement the existing adversarial networks, providing the mechanisms for achieving inference-at-scale.

Shortcomings of Adversarial Networks

We define three shortcomings of adversarial networks when providing inference at scale. In the current adversarial networks, models are ephemeral, which makes it difficult to build upon these networks in a stable manner. While the natural drive is to improve model quality, the nuance between two different fine-tuned models will likely produce results that are detrimental to at least some apps or end-users given apps built upon a single subnet. Many builders may find it advantageous to have models constantly improving on the back end, but builders should be able to opt-out of the competitive networks and choose stability instead. Secondly non-homogeneous models make it difficult to get consistent results and requires significant advancement in gating mechanisms. The variety of models on the network, while advantageous for some, will result in inconsistent and noisy end-user experience. Even with a gating mechanism, if apps are heavily used, models will be over-trafficked and requests will be sent to different or lesser models, adding unacceptable entropy to the end user. Again, this is not necessarily true for all use cases, but surely important to many. Finally, decentralized nodes can create non-ideal latency if optimizations are not made. Especially given the ephemeral nature

of models, and even with a gating mechanism, hitting the proper model in the proper location is a burden. End-user inference must have minimized latency in order to create a competitive experience with centralized AI.

Homogeneous Inference Grids

Inspiration for Redundant Determinisitic Verification Networks came from Manifold Labs' TARGON and can be read about here: (https://github.com/manifold-inc/targon/tree/main). We re-brand Redundant Deterministic Verification Networks as Homogeneous Inference Grids to emphasize the necessity of the grids for inference, and remove the focus on the deterministic verification as that is an implementation mechanism rather than the defining use case of the network. Redundant Inference Grids provide the framework to serve as a solution to the stated shortcomings of adversarial networks for inference. If a node is to drop out in an adversarial network, the unique intelligence of that model is no longer accessible. By specifying a model for the entirety of the network, subnet owners create a homogeneous inference grid that allow end users and developers to have transparency into the model they leverage. This alleviates the concerns around the ephemeral nature of models in adversarial networks, by providing redundant compute for inference.

Additionally, homogeneous inference grids are gamification-resistant– a critical start to improving the resiliency of the Bittensor network. Forcing the generation of a seed by both the prover and the verifier ensures actual work is being completed by the miner. This gamification that has plagued the network has stunded the technical advancements. By solving gamification, homogenous inference grids allow for optimizations to be built around real compute– starting with edge node inference. In a world where hundreds of apps are being built on Bittensor, and hundreds of thousands (or millions) of user requests are being sent to subnets, the network must minimize response times.

Fractal for text-to-video generation

Fractal is adopting the framework of Provers and Verifiers and implementing text-to-video generation. Text-to-video is but a seedling modality of model compared to text or image generation, particularly in the open source com-

munity, but poses a visually appealling and unique way to make the subnet's commodity tangible. Starting with simple video generation that is quite gif-like, this subnet will continue to push the boundaries of what models are publicly available; as new models are released (hopefully borne from a different Bittensor subnet), the requirements of subnet five will continue to grow to support newer models.

In addition to driving model quality for text-to-video, Fractal is working to develop and implement routing mechanisms that optimize inference response speeds. This will be critical to providing for end-user adoption in a scalable way on the Bittensor network. Latency will be benchmarked and stored to give validators a map of the grid, allowing them to channel requests to the optimal nodes.

Conclusion

It is quite important to see the Homogeneous Inference Grid not as a challenge to the Bittensor establishment (adversarial networks), but rather a supplement to them. A testament to the versatility of Bittensor, these grids will allow Bittensor to not only cultivate the best open source models in existence, but also service the best developers, builders, and hobbyists in the AI/ML space. Solving gamification and providing the foundation for inference at scale unlocks the ability to provide an end-user experience that centralized AI can not.