Towards Improving Model Generation in Variant Parallelism

Resource constraints of edge devices serve as a major bottleneck when deploying large AI models in edge computing scenarios. Not only are they difficult to fit into such small devices, but they are also quite slow in inference time, given today's need for rapid decision-making. One major technique developed to solve this issue is Variant Parallelism. In this ensemble-based deep-learning distribution method, different main model variants are created and deployed in separate machines, and their decisions are combined to produce the final output.

The method provides graceful degradation in the presence of faulty nodes or poor connectivity while achieving an accuracy similar to the base model.

However, the technique used to generate variants can fail in scalability as combining variants of smaller size with somewhat identical characteristics may not help achieve a significant accuracy boost unless they are retrained with different random seeds. Therefore, this research will focus on improving variant parallelism by exploring other ways to generate variants. We will apply knowledge distillation (KD), where a teacher model of a certain type (e.g., ResNet-50) can be used to train a smaller student model or a model of a completely different structure (e.g., MobileNet).

We aim to develop a variant generation technique where we can generate as many variants as there are participating devices while boosting accuracy and inference speed. Additionally, we will create an optimization scenario that dynamically creates a smaller student model based on specific requirements, such as hardware characteristics and end-to-end performance metrics.

Advisors

Navidreza Asadi