COMPSCI 690AB SYSTEMS FOR DEEP LEARNING

Reading Assignment 7

TVM: An Automated End-to-End Optimizing Compiler for Deep Learning– Summary

The paper introduces TVM which acts as a compiler that takes very high level specifications of deep learning program present in the existing frameworks and in turn generates low-level optimized code for different hardware backends. The challenge of optimizing deep learning workloads for hardware platforms arises as deep learning models are becoming increasingly complex and optimizing their performance across different hardware architecture is very crucial for efficient inference and training. It's important to efficiently train deep learning models as they are essential for a wide range of applications. The hypothesis of developing an automated framework like TVM, it is possible to achieve much better performance in comparison to existing deep learning frameworks. TVM can analyse deep learning workloads, generate optimized code for it by using machine learning-based techniques like cost modelling and exploration algorithms and deploy it across hardware platforms. Manually optimizing deep learning workloads are time-consuming, prone to more errors and requires expertise. Thus, by automating it TVM aims to streamline the optimization workflow.

One major limitation of the TVM's proposal is how effective it may be depending mainly on the quality and diversity of the training data used in the machine learning models. If the training data of the model does not represent the performance characteristics for a variety of different deep learning workloads on different hardware platforms, then the TVM's predictions and optimizations are likely to be suboptimal. In order to address this, future research might focus on mainly improving the quality and diversity of the training data, with the potential of incorporating real world performance measurements from different deep learning applications as well. In addition, refining TVM's machine learning models and optimization of the algorithms could also help in mitigating this limitation through constant learning from the new data.

Srimathi Mahalingam