Relay: A New IR for Machine Learning Frameworks

Relay is a high-level, statically typed, functional intermediate representation (IR) designed to improve machine learning models' efficiency, expressiveness, and portability across diverse hardware. It integrates with the TVM stack, allowing for optimizations and deployment on specialized hardware like GPUs and TPUs. Relay addresses the limitations of existing static and dynamic computation graphs by providing a more flexible programming language for differentiable computations.

Strengths:

- It facilitates execution on a wide range of hardware platforms.
- They combine the best aspects of static and dynamic graphs.

Weaknesses and Open Issues:

- Developing a runtime system that can leverage Relay's capabilities.
- Error messages and debugging improvement.
- Optimize Relay's to execute on a wide range of hardware.

The paper's relevance to our project goals:

- The ability to optimize models for various hardware makes it suitable for IoT devices.
- It helps training and inference for complex machine learning and AI models.