# Code-Generation Comparison TVM and Triton

Felix Jonathan Rocke

# Introduction Triton

- Open-source Python-like programming language developed by OpenAI
- Abstracts CUDA code
- Reduces programming complexity significantly
- Can match cuBLAS performance
- Backend uses MLIR and PTX for Code Generation (only NVIDIA GPUs)

# Introduction MLIR


- Part of the LLVM Project
- Can represent dataflow graphs
- Graph level optimisations
- Loop optimisations (fusion, tiling, etc.)
- Cache management (memory tiling, vectorisation 1D, and 2D registers)
- Target Specific Operations (e.g., accelerator-specific-operations)
- Polyhedral Primitives

# Research Idea

TorchInductor uses the Triton language for code generation

**Motivation:** Frequent discussion on why Inductor relies on Triton and not TVM for code generation:

=> Claims that Triton code-gen is better for NVIDIA GPUs

=> Faster compilation time (very small search space)

**Aims:**

- Compare Triton (Inductor) code-gen to TVM backend
- Identify the role of search space size (see Bolt and Ansor approach)

# Plans

1. Get a baseline between the two
2. Compare generated code, identify differences in approach
3. Write specific test scripts (maybe increase search space)
4. Evaluate the importance of search space and GPU performance

# Questions & Suggestions