Tenspiler: A Verified-Lifting-Based Compiler for Tensor Operations (Artifact)

Jie Qiu
Pittsburgh, PA, USA

Colin Cai
University of California, Berkeley, CA, USA

Sahil Bhatia
University of California, Berkeley, CA, USA

Niranjan Hasabnis
Intel Labs, Menlo Park, CA, USA

Sanjit A. Seshia
University of California, Berkeley, CA, USA

Alvin Cheung
University of California, Berkeley, CA, USA

Abstract

In the related article, we described TENSPILER, a verified-lifting-based compiler that translates sequential programs to tensor operations. We further demonstrated its effectiveness by translating 69 benchmarks from into 6 different DSL targets and evaluating their performance against the baseline. This artifact includes the implementation of TENSPILER as well as files used to reproduce those results.

2012 ACM Subject Classification Software and its engineering → Compilers

Keywords and phrases Program Synthesis, Code Transpilation, Tensor DSLs, Verification

Digital Object Identifier 10.4230/DARTS.10.2.17

Funding This work was supported in part by DARPA Contract FA8750-23-C-0080, a Google BAIR Commons project, NSF grants IIS-1955488, IIS-2027575, ARO W911NF2110339, ONR N00014-21-1-2724, and DOE award DE-SC0016260, DE-SC0021982, and the Sloan Foundation.

Acknowledgements We would like to thank Jayaram Bobba and Zhongkai Zhang from Intel's Habana team for inputs on Gaudi architecture, TPC-C programming model, and obtaining high-performance from TPC kernels. We would like to thank Hasan Genc and Sophia Shao for helpful insights into Gemmini code generation.

Related Article Jie Qiu, Colin Cai, Sahil Bhatia, Niranjan Hasabnis, Sanjit A. Seshia, and Alvin Cheung, "Tenspiler: A Verified-Lifting-Based Compiler for Tensor Operations", in 38th European Conference on Object-Oriented Programming (ECOOP 2024), LIPIcs, Vol. 313, pp. 32:1–32:28, 2024.
`https://doi.org/10.4230/LIPIcs.ECOOP.2024.32`

Related Conference 38th European Conference on Object-Oriented Programming (ECOOP 2024), September 16–20, 2024, Vienna, Austria

Evaluation Policy The artifact has been evaluated as described in the ECOOP 2024 Call for Artifacts and the ACM Artifact Review and Badging Policy.

Dagstuhl Artifacts Series
Schloss Dagstuhl – Leibniz-Zentrum für Informatik,
Dagstuhl Publishing, Germany

1 Scope

The artifact is intended to evaluate the usability of TENSPILER and the performance of the translated programs. It runs all three phases of TENSPILER (synthesis, verification, and code generation) on input sequential benchmark implementations [1] and produces the corresponding DSL code as described in Sections 4 and 6.1.2. In addition, it contains scripts to run the baseline against the generated DSL code to reproduce performance results described in Section 6.3 (Figures 9 and 10).

2 Content

The artifact package includes:

- **artifact_readme.md**: Instructions on how to set up the environment and run all the experiments.
- **Dockerfile**: Dockerfile for easy environment setup.
- **tenspiler/(blend|llama|c2taco)cpp/for_synthesis/**: Source code for benchmarks in all the suites.
- **tenspiler/codegen/**: Code generation scripts from TENSIR to each of the 6 supported backends.
- **tenspiler/generated_code/**: Scripts that run TENSPILER end-to-end for each benchmark to generate code for each target backend.
- **tenspiler/benchmarking/**: Scripts to obtain speedup for each benchmark on the backends.
- **tenspiler/(data/|vicuna_weight.h5|vicuna_weight7b.h5)**: Sampled datasets included for evaluation.

3 Getting the artifact

The artifact endorsed by the Artifact Evaluation Committee is available free of charge on the Dagstuhl Research Online Publication Server (DROPS). In addition, the artifact is also available at: `https://github.com/tenspiler/tenspiler`.

4 Tested platforms

To ensure reproducibility and accessibility, the artifact setup is containerized into a Dockerfile and is runnable on any platform with Docker Engine. The Docker image takes 20 minutes to build and 15GiB of storage from our experiments. Peak memory usage is 3GiB for running on sampled datasets. Full datasets require significantly more memory and computing resources; our experimentation setup has 512GiB RAM and 80 CPU cores available.

5 License

The artifact is available under MIT license.

6 MD5 sum of the artifact

8a40381f27a9e92b0098c60e02f4671e

7 Size of the artifact

1.04 GiB

[1] Since some benchmarks used in related articles are privately sourced, only 36 public ones are included.