

Compiler Toolchains for Deep Learning Workloads on Embedded Platforms

Max Sponner
Infineon Technologies
Dresden GmbH & Co. KG
Dresden, Germany
max.sponner(at)infineon.com

Bernd Waschneck
Infineon Technologies
Dresden GmbH & Co. KG
Dresden, Germany
bernd.waschneck(at)infineon.com

Akash Kumar
Dresden University of Technology
Dresden, Germany
akash.kumar(at)tu-dresden.de

Abstract

As the usage of deep learning becomes increasingly popular in mobile and embedded solutions, it is necessary to convert the framework-specific network representations into executable code for these embedded platforms.

This paper consists of two parts: The first section is made up of a survey and benchmark of the available open source deep learning compiler toolchains, which focus on the capabilities and performance of the individual solutions in regard to targeting embedded devices and microcontrollers that are combined with a dedicated accelerator in a heterogeneous fashion.

The second part explores the implementation and evaluation of a compilation flow for such a heterogeneous device and reuses one of the existing toolchains to demonstrate the necessary steps for hardware developers that plan to build a software flow for their own hardware.

CCS Concepts: • Computing methodologies → Artificial intelligence.

Keywords: deep learning, embedded, deep learning compiler

ACM Reference Format:

Max Sponner, Bernd Waschneck, and Akash Kumar. 2021. Compiler Toolchains for Deep Learning Workloads on Embedded Platforms. In *Proceedings of TinyML Research Symposium (TinyML Research Symposium '21) (Burlingame '21)*. ACM, New York, NY, USA, 10 pages.

1 Introduction

As AI is moving closer to the edge, the limitations of the popular deep learning frameworks in regard to embedded platforms become apparent. These platforms often employ

Cortex-M CPUs or similar solutions and can operate on the (sub) milliwatt range for their power consumption. Deep learning frameworks are mostly designed for the server and workstation use and incorporate many features that are not relevant for the inference on low power devices, this prevents them from running on microcontrollers and other embedded solutions.

Due to this, the usage of deep learning models on embedded devices typically relies on the manual implementation of the previously trained networks. The developers have to implement the required layer types, preferably using the vendor-provided math kernel libraries for the platform. This process is labour intensive, error prone and can easily result in inferior performance due to missing or incorrectly executed optimizations. In addition, the support of new platforms can result in extensive effort as the function kernels might need to be re-implemented for them.

The necessary modifications increase even further if dedicated deep learning accelerators are employed. Due to their use of domain-specific instruction sets, which often utilize long pipelines that cover common neural network sub-patterns, these accelerators cannot easily be targeted by standard compiler toolchains. Additional obstacles can be the coarse-granular accelerator instruction set architectures and the usage of custom data types, which can differ from the types used by the neural networks.

A possible solution to automate these tasks are deep learning compilers. These toolchains operate similar to standard compilers, but introduce a number of important peculiarities: Instead of handwritten source code, deep learning compilers process serialized trained neural network descriptions. Additionally, they should be able to automatically employ the optimized math kernel libraries or alternative optimizations of the function kernels.

A benefit of employing domain-specific compilers is the option to introduce additional optimizations that target deep learning models. Typically these are layer fusion or general network graph optimizations and quantization schemes.

Lastly, these toolchains should be able to target heterogeneous platforms and dedicated accelerators by employing runtimes on the target devices. These runtimes take care of the scheduling and additional support operations that are

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Burlingame '21, March 22, 2021, Burlingame, CA

© 2021 Association for Computing Machinery.

necessary as well as the deserialization of the compiled networks.

This paper will start with an overview of the available optimizations and math kernel libraries for deep learning workloads on embedded platform, which should be incorporated by the compiler toolchains for optimal performance. The following sections will cover a survey of the compiler features and the achieved performance on different embedded and low-power platforms, while the last part of the paper contains the implementation of a compilation flow for a new custom target.

The intended target audience of this paper are research teams that are looking for a software stack to support their own deep learning hardware efforts, while achieving optimal performance and broad compatibility across the available frameworks.

2 Related Work

Two recent studies focus on the employed techniques and the use with FPGA platforms [43], as well as an in-depth overview over the different approaches for common problems of the available deep learning compilers [26]. In contrast to these publications, this work focuses more on embedded platforms. In addition to the survey and benchmark a compilation toolchain based on TVM has been implemented to target a heterogeneous platform. This was done to demonstrate the steps that are currently required to support custom accelerators with their own software stack.

2.1 Deep Learning Optimizations

Deep Learning toolchains can employ a multitude of domain-specific optimizations in addition to standard compiler strategies. These include weight pruning, which cuts redundant and unnecessary weights from the networks to reduce their size and - depending on the implementation - the compute workload. While a wide range of pruning algorithms exist, none of them are currently employed by deep learning compilers [4, 25, 36, 41].

In contrast, quantization schemes are utilized by most deep learning compilers. This optimization converts the floating point representations for weights, intermediate results and outputs into smaller fixed-point formats, while mostly keeping the accuracy of the original network. This enables the network to take up less storage and reduces its bandwidth requirements during the execution as well as the computational intensity, if optimized function kernels for the quantized representations exist [20, 21, 27, 37, 42].

Additional strategies include optimizations on the neural network graph like layer fusion, dead node elimination and others [23, 38].

2.2 Math Kernel Libraries

Math kernel libraries are a way to provide developers with optimized function kernels for common operations, increasing the efficiency of software solutions that employ them, while decreasing the redundancy of implementations. They are typically platform-specific and provided by the device-vendors¹. These libraries differ largely in their implementation strategies and offered functionality. While all libraries provide function kernels, their implementations follow different approaches: ARM CMSIS-NN [22] mostly resorts to a low number of kernels that deliver consistent performance across the majority of the configuration space. In contrast, Intel's oneDNN² [13] library implements most operations with multiple different strategies, based on the available instruction set extensions, data types and the configuration of the currently executed layer. The final selection of the function kernel takes places at runtime³ to achieve the best performance under the current circumstances. While this strategy can be able to achieve better performance in certain cases, it requires much more maintenance and implementation effort compared to the more generalized function kernels.

2.3 Deep Learning Accelerators

In recent years plenty of deep learning accelerators emerged in commercial and research applications. Commercial accelerators for embedded platforms include the NVDLA from Nvidia⁴ [31], Google's EdgeTPU [5] as well as ARM's Ethos-U NPU's [2]. Most of these solutions employ custom compilation toolchains that are limited to support of only one deep learning framework for its input formats.

Research platforms include the Eyeriss accelerators (v1, v2) [8, 9], VTA [30] as well as a many FPGA-based solutions [35]. These typically do not focus on the software toolchain and explore novel architecture approaches instead.

3 Survey & Benchmark

The survey section of the paper covers open source projects that are still in development. Its focus lies on the inference of deep learning models on embedded hardware. The support for the training step will not be evaluated as it is uncommon to execute it on the embedded device itself.

The evaluated deep learning compilers are TensorFlow Lite (TFLite) [39], TensorFlow XLA (TF XLA) [40], Glow [34], TVM [7], ONNC [28] and nGraph [14], which has been tested as part of Intel's openVINO toolkit.

¹Efforts for device-independent solutions exist as well, but have not found the same rate of adaption, e.g. XNNPack [16]

²previously known as MKL-DNN

³only for its x86_64 CPU backend

⁴The NVDLA is available as an open source project, which provides the hardware description and software stack, but is also contained in several of Nvidia's own products like the Jetson Xavier NX, using an alternative closed-source software stack

As μ TVM was in an early stage at the time of testing, it has not been evaluated in this paper.

3.1 Common Strategies

All evaluated toolchains follow the typical compiler structure. The frontend converts the serialized pretrained models into a high-level intermediate representation (IR). Most toolchains utilize a two-level IR: The high-level IR is a graph-level representation of the compiled model and the low-level IR describes the operations on the tensor level. The graph-level IR is typically used for mostly target-independent optimizations and operator fusion. The tensor-level IR is used by the backend to optimize the individual operations.

One exception is TFLite, which does not perform target-dependent optimizations at the compilation stage and only uses a graph-level representation. Instead, its compiler (called TFLite converter) generates a graph-level representation that does not contain execution details, as the device-specific function kernels are part of its runtime. This allows for a better portability of the compiler output across devices, but prevents more target-specific optimizations at the offline compilation stage.

The majority of the evaluated toolchains employs a runtime⁵, which needs to be present on the target device to execute the compiled neural network. The functionality of the runtime differs between projects. All runtimes provide support infrastructure to unpack the compiled neural networks and an API for the integration into the user program. Solutions like TFLite and ONNC deliver the operation execution strategies for the platform as part of the runtime. Glow and TVM utilize the runtime for heterogeneous targets and in addition TVM requires the runtime for profiling during the auto-tuning step. TVM delivers function kernels that have been generated by its auto-tuner alongside the model description and weights to the runtime.

The main difference between the evaluated projects is the provisioning of the function kernels. Most solutions utilize handcrafted implementations that integrate math kernel libraries. This requires maintenance and updating of implementations for each layer type across all supported platforms⁶. To circumvent these limitations, TVM employs an Auto-Tuning solutions which tries to find the best suited function kernels by using an AI-guided flow that incorporates measurements and estimations of execution times on the real target. Glow bypasses all of these concerns by reusing the same generalized function kernels across all targets, where its target-dependent optimizations are only applied by the LLVM backend for the selected target.

⁵The only exception that never uses a runtime is TensorFlow XLA, while Glow's AOT flow for CPUs does not require a runtime, it is deployed on other platforms

⁶TFLite, ONNC and others employ this strategy

3.2 User Accessibility

User Accessibility mostly depends on the user interface of the offline compiler stage, the integration of the compiler output into the target application and the supported input formats.

For the supported frameworks and formats, ONNX[32] is the most important, as it is an industry standard and converts from most frameworks that exist. See table 1 for an overview of the supported formats and frameworks of each compiler toolchain.

All compilers either support a command-line interface, like

Table 1. Overview of the supported deep learning framework formats and target hardware platforms.

	TVM	TF Lite	TF XLA	Glow	ONNC	openVINO
ONNX[32]	✓	✗	✗	✓	✓	✓
TensorFlow[1]	✓	✓	✓	✗	✗	✓
TensorFlow Lite flatbuffer	✓	✓	✗	✓	✗	✗
PyTorch[33]	✓	✗	✗	✗	✗	✗
MXNet[6]	✓	✗	✗	✗	✗	✓
Caffe[29]	✓	✗	✗	✓	✗	✓
Keras[11]	✓	✓	✓	✗	✗	✗
x86_64	✓	✓	✓	✓	✓	✓
Cortex-A	✓	✓	? ⁷	✓	✓	✗
Cortex-M	✗ ⁸	✓	? ⁷	✓	✓	✗
GPU (CUDA)	✓	✗	✓	✗	✗	✗
GPU (OpenCL)	✓	✓	✓	✓	✗	✓
Deep Learning Accelerator	✓	✓	✓	✓	✓	✓

traditional compiler toolchains, or the use through a Python API, which allows for the integration in the original training script of the deep learning model. One exception is Intel's openVINO that provides an additional graphical user interface through a web-interface [18]. This enables more direct feedback to the developer on the impact of different optimizations on the overall model performance.

For the integration into the user application, all toolchains provide a C or C++ API. TVM and TFLite provide an additional Python interface through their standard runtimes⁹.

⁷inconclusive data

⁸ μ TVM was not ready at the time of testing

⁹TFLite provides an additional runtime for microcontrollers, which does not come with a Python API

3.3 Supported Platforms & Extensibility

As the support for low-power embedded devices¹⁰ in the currently available deep learning compilers is still limited, additional platforms have been used during the evaluation to allow for a more complete performance comparison.

The range of supported platforms varies between the evaluated toolchains. In addition, the level of optimization for the supported platforms fluctuates widely. One such example is TFLite's support of the x86_64 platform: While its runtime can be compiled for it, the function kernels are not optimized, resulting in worse performance compared to other platforms or compilers.

For TF XLA no conclusive information about its support for different architectures could be found, as the official documentation and publications contradict each other [17, 24]. The support for bare-metal use cases and embedded processors is much less common, as only TFLite, ONNC and Glow are able to target Cortex-M CPUs. For an overview of the supported platforms see table 1.

While all toolchains include some kind of support for heterogeneous platforms, the implementations differ between them. The most complete solution has been provided by TVM in its Bring-Your-Own-Codegen (BYOC) flow [10], which allows developers to target new libraries and accelerators from TVM's high-level IR. It does not only provide an API to include new backends, it also supports the developer by implementing solutions for common problems, like the CPU fallback for unsupported operations, an infrastructure for custom operator fusion rules and the option to change the data layout of tensors. Most other toolchains only supply a simple API to access values and layer configurations and require the developer to reimplement many of these common support tasks.

A stark contrast to TVM is TFLite. Its compilation stage does not provide an interface for the inclusion of additional target-specific tasks and optimizations. New platforms are targeted by porting the runtime to them and deploying optimized function kernels with it. As this flow only allows the targeting of general purpose hardware, its support for the currently available accelerators has been realized by additional tools. These modify the flatbuffer file, which has been generated by the offline compilation stage, before it can be executed on the heterogeneous platform [3, 12]. This approach breaks the portability of TFLite and requires additional work to keep these tools compatible with the current scheme of the TFLite flatbuffer files. Some compiler toolchains like Glow, which reuse LLVM's backends¹¹ can easily target new architectures, if a LLVM backend already exists. In that case Glow's ahead-of-time (AOT) compilation mode can be reused, if other targets need to be supported, a separate runtime can be used and the AOT flow can no longer be utilized.

¹⁰e.g. ARM Cortex-M and similar

¹¹for its AOT flow

3.4 Features

For the embedded use case, the AOT compilation and quantization support are the most important compiler features. For the optimal execution on the target devices, the toolchains should be able to incorporate math kernel libraries or auto-tuning to provide optimal function kernels for the execution. Features like support for the backpropagation and training steps are not as important for embedded devices (yet) and are only supported by TF XLA, Glow and openVINO as they primarily target HPC and cloud applications.

For the optimization of the layer execution TFLite, ONNC and openVINO rely on math kernel libraries, while Glow utilizes the same set of generalized function kernels across all targets, only utilizing the LLVM backend optimization steps. TVM is the only evaluated toolchain that employs an auto-tuning process to find good performing function kernels automatically, but can also exploit third party libraries through the BYOC flow.

All toolchains - with the exception of TVM - implement only a limited number of static quantization schemes with fixed bit-widths for intermediates and weights. This is a limitation for the targeting of custom devices as they could employ alternative quantizations. TVM has implemented a more flexible quantization system, that offers different sizes and requantization strategies. However, it is more difficult to configure compared to the other solutions and did not achieve competitive accuracies in this benchmark.

3.5 Performance

The performance was evaluated on an ARM Cortex-M55 fast model¹², an Cortex-A72¹³ and a Intel Celeron J1900¹⁴. The Cortex-A and Intel platform have been selected, to allow for a more complete performance overview due to the limited support of Cortex-M in the tested toolchains. This also allowed for a direct comparison to the standard TensorFlow framework on these two platforms.

All of these platforms provide a SIMD vector extension and have been tested with the same simple MNIST test network, consisting of convolutional, fully connected, ReLU and maximum pooling layers. The batch size has been set to one and the final activation function has been removed after training. These are common optimizations for embedded applications, as it reduces the jitter that is introduced by the prediction as well as the latency, as the final activation does not change the classification result.

The Cortex-M55 could only be targeted by TFLite¹⁵ and ONNC¹⁶. As no hardware of the M55 is available yet, a instruction-level simulator has been used instead. While

¹²as no hardware was available at the time of testing

¹³using a Raspberry Pi 4 with 4 GB of system memory

¹⁴using 8 GB of DDR3 system memory

¹⁵using its micro-runtime

¹⁶using a special Cortex-M version of the toolchain

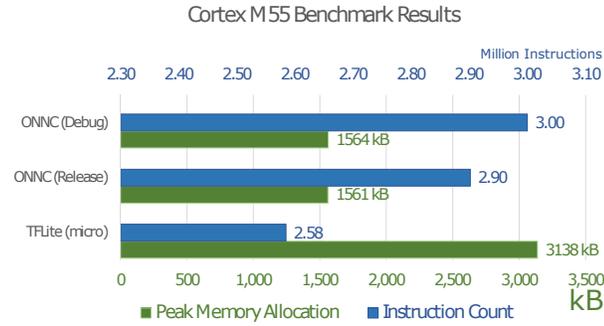


Figure 1. Instruction Counts and Peak Memory Allocation for the execution of the benchmark model on the Cortex M55.

Glow is able to target Cortex-M CPUs, it was not able to support the novel instruction set of the M55 (ARMv8.1M). The testing showed that TFLite required less instructions to complete an inference run (2.6 M instead of 3 M instructions, see figure 1), while ONNC allocated significantly less memory (1.6 MiB instead of 3 MiB. See figure 1 for details). The next test platform was the Cortex-A72. ONNC could not be compiled for it, as it relied on Intel MKL for its function kernels¹⁷. Instead Glow, TVM and TFLite have been tested in addition to the standard TensorFlow Python runtime. An overview of the measured inference times can be seen at figure 2.

The quantized TFLite version achieved the fastest inference speed with 0.37 ms, followed by the quantized and auto-tuned TVM program, using its C runtime (0.41 ms). Glows fastest output achieved a mean inference time of 1.77 ms by using floating-point representations. Besides Glow, all compilers achieved faster inference times using quantized networks - suggesting that they employ optimized function kernels for quantized operations, while Glow uses its generalized standard kernels and wraps them into additional quantization and dequantization steps, which causes additional overhead. The worst result by a compiler was achieved by TVM, for its floating-point auto-tuned solution, as it tried to apply x86 function templates to the ARM platform¹⁸. However, the slowest result of 6.51 ms was still significantly faster than the use of the network in combination with the standard TensorFlow runtime - requiring 104.64 ms for a single inference run. This makes the slowest and incorrectly optimized compiled version 16 times faster, while the fastest compiled version achieved a speedup of 282.6 times.

The Intel Celeron CPU allowed for the additional testing

¹⁷While the CMake script for the standard runtime contained a parameter to disable the MKL integration, it could not be build when it was selected

¹⁸it could not be determined, if it was caused by user error or by TVM, but it reoccurred over multiple tries and did not affect the quantized version

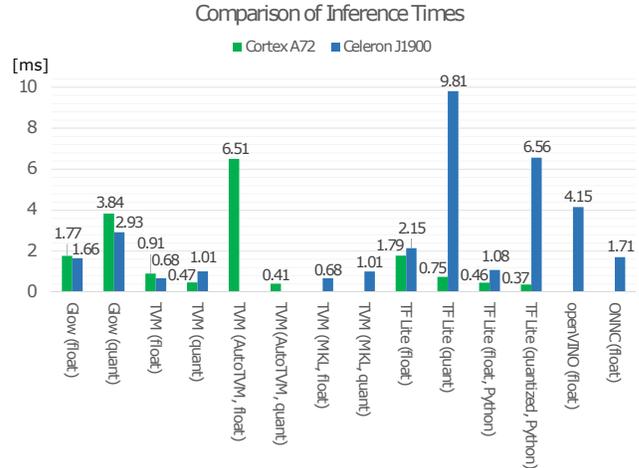


Figure 2. Comparison of the inference times on the ARM Cortex-A72 and Intel Celeron J1900 platforms.

of nGraph¹⁹ and ONNC’s standard flow²⁰. See figure 2 for the inference time results of the platform. In comparison to the Cortex-A results the ranking of the toolchains by their inference time changed, suggesting different levels of optimizations across the supported target devices for some deep learning toolchains. In addition, TVM was tested with the Intel MKL BYOC-based backend instead of its auto-tuning flow. This backend is not primarily optimized for performance as it is a demo for the BYOC functionality and was used to estimate the overhead which results from it. For the Celeron J1900, the floating-point versions of the compiled networks achieved faster inference speeds across all toolchains. This suggests either a lack of optimized kernels or a better implementation of the floating-point components of the hardware itself. The fastest results have been achieved by TVM with 0.68 ms (FP) and 1.01 ms (quantized). TVM did not show a significant difference between the standard and the BYOC flow results, which implies that the overhead of the BYOC flow is minimal. The next best results were achieved by TFLite’s floating point network (1.08 ms, using the Python API), Glow (also floating point, 1.66 ms) and ONNC (1.71 ms). openVINO’s compiled program did require 4.15 ms for a single inference, which made it the slowest floating point version out of the tested compiled networks. It was not able to quantize the network, as that is only supported on newer Intel CPU generations. Only TFLite’s quantized networks took more time than openVINO to complete their inference run.

In addition to the inference times, the peak memory allocations have been measured. The measured results varied by two orders of magnitude between the toolchains. Glow’s compiled networks required 10 MiB of system memory at

¹⁹as part of openVINO

²⁰besides its Cortex-M version

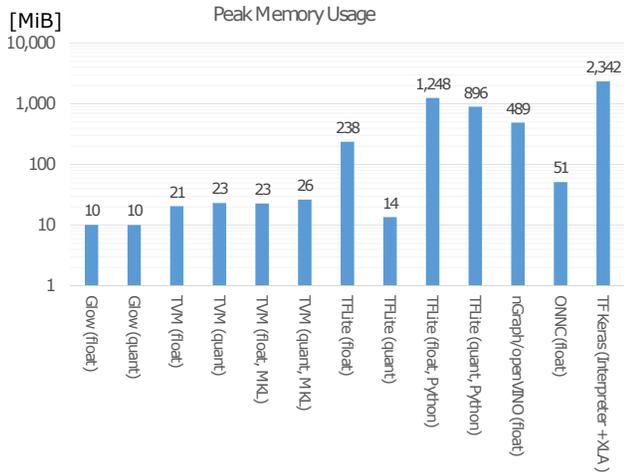


Figure 3. Comparison of the peak memory allocation on the Intel Celeron J1900 platform.

peak, followed by TVM with 21 MiB to 26 MiB. As the higher allocations have been measured for the MKL-BYOC variant, it suggests, that the BYOC flow requires some memory overhead compared to the standard flow during the execution. TFLite required 14 MiB for a quantized version of the network utilizing only its C-API, which took significantly longer than the other results for a inference. The same configuration, but with a floating point version of the network allocated 238 MiB which is more than the expected increase by four times²¹. ONNC could only be tested with a floating point network as its open source standard branch does not support quantization. Its peak memory allocation of 51 MiB is more in line with the expected memory allocation. openVINO's implementation allocated 489 MiB of memory during the inference, only TFLite's Python runtime used more memory with 896 MiB (quantized) or 1,248 MiB (floating point). While this values are significantly higher than the results of the other toolchains, they are still an improvement in comparison to the standard TensorFlow framework that allocated up to 2.3 GiB. The prediction accuracy for the compiled networks stayed mostly the same, even for the quantized variants, with the exception of TVM. It only reached an accuracy of around 50 %, which might have been user error due to its configurable quantization scheme and the usage of a global quantization scheme.

The benchmark has shown that the available deep learning compilers are all able to deliver significantly better performance compared to the use of the standard TensorFlow framework. Additionally, they allowed for fast deployment of models on the evaluated platforms, without the need for handcrafted layer implementations.

TVM was able to deliver the best inference speeds on the larger test devices; nonetheless, its accuracy was limited by

²¹as 8-bit integers are 4 times smaller than 32-bit floating point values

the quantization flow, which was the only flexible quantization system out of the tested toolchains. While its frontend offered support for the majority of neural network serialization formats, the backend and runtime are unable to target the Cortex-M platform yet²². Its newly introduced BYOC flow allows developers to target heterogeneous platforms, while reusing the frontend for different input formats and the standard components for the execution of unsupported operations on the system CPU.

TFLite was able to target the Cortex-M55 platform and achieved a similar performance to TVM on the Cortex-A system for the measured inference times. Nevertheless, it allocated significantly more memory for its execution and does not offer a simple flow to target heterogeneous systems that include a dedicated accelerator. Additionally, users are limited to the capabilities and formats of the TensorFlow ecosystem.

Glow's support for the ONNX standard makes it compatible with most deep learning frameworks and its quantization flow achieves similar accuracies to TFLite. While it achieved the lowest peak memory allocation out of the tested solutions, the use of generalized kernels lead to slower inference times compared to the other platforms. For the targeting of embedded microcontrollers, the inclusion of CMSIS-NN could result in a significant performance improvement.

ONNC was able to target the Cortex-M55 system and delivered competitive performance to TFLite's micro-runtime. However, the separation of the quantization tool into a commercial product, outside of the open-source project and the lack of support for newer ONNX subsets might limit its use in future endeavors.

Intel's openVINO is only able to target the Celeron platform, but did not achieve a competitive performance result on this platform. Its limitation to x86 CPU's makes it unable to target any kind of low-powered embedded platform. However, its graphical user interface (called workbench) made the optimization and compilation process more transparent for the user, which could be a helpful feature for other toolchains as well.

While the number of supported layer functions is important for the user, it is difficult to compare these toolchains based on a single snapshot of their development stages, due to constant updating of most of them.

This survey has shown that the support for embedded platforms in the current deep learning compiler toolchains is still at an early stage as only a small subset is able to target platforms like the Cortex-M55.

²²However, a micro-runtime is currently in development, but not production ready yet

4 Implementation

For the implementation, an abstract accelerator has been defined and an instruction set simulator was implemented²³. The simulator was verified with TensorFlow and is able to estimate the hardware utilization and cycle count for input and output parallel execution strategies²⁴.

The simulated accelerator uses an instruction set that is similar in its capabilities to other solutions like Nvidia's NVDLA [31]. It only supports signed integer formats for operations and the majority of them are limited to a length of eight bit for the individual values in their input tensors while producing outputs with a length of either eight or 32 bit.

For the software flow TVM was used due to its BYOC functionality. This flow starts with the definition of annotation rules for supported nodes and patterns in TVM's graph-level IR²⁵. These are then merged into subgraphs, which will be executed by the accelerator. These steps are handled by TVM's BYOC flow and did only require the definition of supported graph patterns during the implementation of the new backend. TVM manages the execution of unsupported operations on the CPU as well as the invocation of the subgraphs from the standard runtime. After the annotation, the network graph is partitioned to separate the supported sections of the network into subgraphs. These subgraphs are then passed on to the custom code generation, where they are converted into a JSON format for better portability across different instruction set variants. The final generation of the accelerator command stream happens at runtime before the initial inference. This allows to target different ISA variants with varying memory sizes using a single serialized file. A custom runtime component executes the code generation and passes back a run function for each subgraph to the standard TVM graph runtime. During inference the subgraphs are executed by the simulator through the invocation of these run functions.

Besides the quantization and data layout transformation functionality, which was provided by TVM, the memory planning for DMA operations between system and accelerator memory, the assembly generation, the configuration register file management and tiling for larger tensor sizes needed to be implemented by the custom runtime component. The tiling was implemented by primarily splitting the workload along the output channel dimension.

5 Evaluation

The correct functionality was initially tested with neural networks that only contained supported operations as single nodes and patterns. Additional testing with the MNIST

²³the simulator was written to only target the desired ISA and operates on the custom output files of the new compiler backend, which contain the initial memory layout and instructions for each subgraph.

²⁴input parallel: input channels are processed in parallel; output parallel: same for output channels

²⁵called Relay

network from the performance benchmark revealed that the current TVM quantization flow inserts additional meta nodes into the graph. These nodes prevent the merging of multiple compute layers into a single subgraph. Due to this, the network was split into three subgraphs, which requires the system to move the intermediate results back to the shared system memory between the subgraph executions. This resulted in reduced throughput and efficiency due to unnecessary bus transactions. Otherwise, the custom backend worked as intended and generated the expected inference results.

For a more realistic test case Google's MobileNetV1 [19] with the ImageNet dataset [15] has been used. The additional batch normalization layers prevented the use of the larger convolutional pipelines, as they are located between the convolutional and activation function nodes. This adds additional bus transactions between accelerator memory and compute units, which reduces the efficiency of real hardware. The network was evaluated with three iterations of the accelerator and its toolchain:

- Without support for depthwise convolutions: Depthwise convolutional layers are executed by the standard runtime on the CPU. This results in a lower share of the network to be accelerated by the target device. The estimated cycle counts can be seen in figure 4a.
- Software-only Implementation: The depthwise convolutions are mapped to a the standard convolution instruction of the DLA. The ISA has not been changed. This allows for a larger share of the network to be accelerated, but the depthwise layers are constraint to a small utilization of the processing elements in the vector ALU.
- Full support: The ISA has been extended to provide a dedicated instruction for depthwise convolutional layers. This instruction allows for a higher utilization, resulting in shorter execution times.

This was done to evaluate the flexibility of the coarse-grained ISA for the support of new layer types, as the deep learning research community develops new layer types at a very rapid pace, which makes it difficult to develop accelerators that can stay up-to-date for extended time periods without changing the hardware architecture. A possible solution would be to update the software flow, to enable it to map new layers to existing instructions. In the case of depthwise convolutions, this approach was represented by the second test scenario. While the implementation was possible, it resulted in drastically increased cycle counts if compared to a native solution (compare figure 4c and 4e). This was caused by the low utilization of the processing elements (PEs) as only one filter could be processed at a time.

Additionally, these scenarios were tested with different sizes

of the accelerators exclusive memory and the vector ALU as well as input and output parallel mode for cycle and utilization estimations to evaluate the impact of operation tiling on the overall performance. The evaluated memory sizes were 512 KiB and 256 MiB. The last configuration does not require any tiling to take place, while the first is the smallest size which is supported by the implementing tiling methods for this network²⁶.

As shown in figure 4a, the output parallel hardware achieved lower cycle counts due to its higher utilization in the early layers of the network. However, this changed as soon as tiling was required due to the limited amount of device memory. As the tiling splits the workload along the channel dimension of the output tensor, the parallelization opportunity for this architecture shrinks. This results in a lower utilization, and a faster execution of input parallel strategies. Additionally, it can be seen in figure 4e, which includes an additional 1024 KiB configuration, that the cycle count does not scale linearly with the available memory. Instead, a combination of larger vector ALU with 128 processing elements (PEs) and 1024 KiB of memory can be faster than a configuration with 64 PEs and 256 MiB of memory. The reason is, that the 1024 KiB configuration only requires tiling of the initial layers of the network, which allows it to compensate for the additional cycles in the later layers where it can calculate twice as many results per cycle as the 64 PE configuration with 256 MiB of memory. A benefit of the TVM BYOC flow is the ability to quickly evaluate the performance for different real-world network architectures across multiple hardware configurations during the hardware development.

6 Conclusion

This evaluation has shown that, while all evaluated toolchains deliver reasonable performance across different platforms, their support for low-powered embedded devices and heterogeneous solutions with dedicated accelerators is still at an early stage. Some are limited by the use of existing compiler backends, which prevent the targeting of dedicated hardware. Another limitation is the use of static quantization schemes that do not offer an easy solution to adapt them for different hardware implementations.

Additionally, while it was possible to target a novel accelerator using TVM, our implementation showed two drawbacks: The more flexible quantization flow of TVM introduces annotation nodes into the code generation, which prevent the solution from reaching higher efficiency. It is currently not possible to connect the BYOC flow with the micro-TVM runtime that is also still under development. This prevents the usage of TVM on (heterogeneous) embedded devices for

TinyML applications, however, it can already be utilized during the hardware development to evaluate the performance of prototypes with real-world test cases.

A Performance of Hardware Variants

As the accelerator was simulated with different software flow versions, ISA variations and memory configurations, multiple performance estimations have been collected. Each of the three test scenarios from section 5 has been tested with 64 or 128 PEs in the compute module of the accelerator and 512 KiB, 1 MiB (only for the last scenario) and 256 MiB for the SRAM memory on the device. An additional parameter for the simulation was the hardware parallelization of the workload, which was either input or output parallel. The input parallel configuration would split the workload along the channel dimension of the input feature map, while the output parallel version would parallelize the workload along the output feature map's channel axis.

The bars of the diagrams are segmented according to the cycles spent on each supported subgraph type. In the test case for MobileNetV1 only up to three different subgraph types could be accelerated:

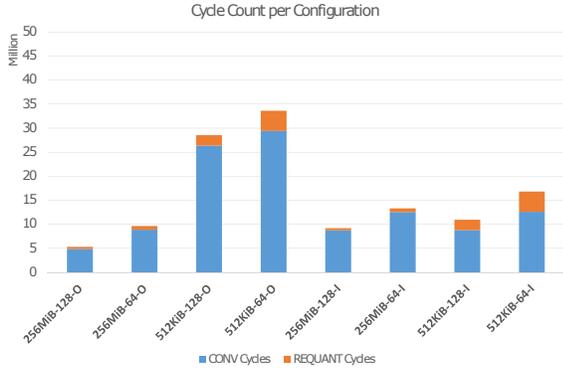
- CONV subgraph:
The CONV (Convolutional) subgraphs contain the Conv2D layers of the network, which can be mapped to the accelerator.
- REQUANT subgraph:
The REQUANT subgraphs contain the requantization operations that are executed in-between layers to convert the 32-bit output of the previous layer back to an 8-bit format.
- DEPTH subgraph:
The DEPTH (Depthwise Convolution) subgraph is only present in the second and third test scenario, as the first did not offer support for the depthwise convolutional layer that is contained in it.

B Target Accelerator

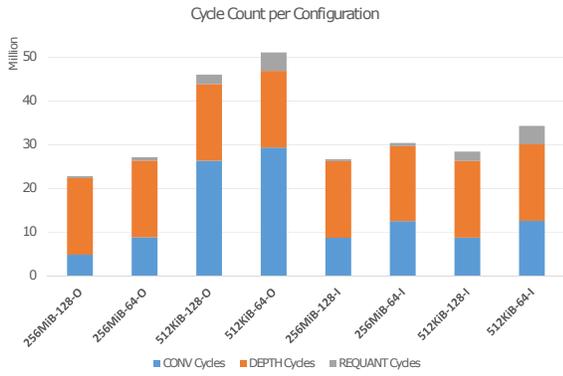
To be more flexible a simulator has been used instead of a real hardware target as it came with several advantages, including a more flexible instruction set, easier configuration of hardware parameters. However, the simulator was still based on the typical concepts of other deep learning inference accelerators like the NVDLA including the coarse-grained ISA and the usage of highly optimized and separated hardware blocks for the different kinds of operations.

Its instruction set architecture was specifically designed for deep learning operations, and features dedicated instructions for typical layer operations as well as common patterns like the combination of convolutional layers with activation functions and requantization operations, which would be mapped to separate pipelines, that would combine multiple different functional units. Most instructions only support the

²⁶For convolutional layers only a split along the output channel dimension was implemented, as the splitting along the rows and columns requires extensive effort to implement and validate all edge cases that can occur

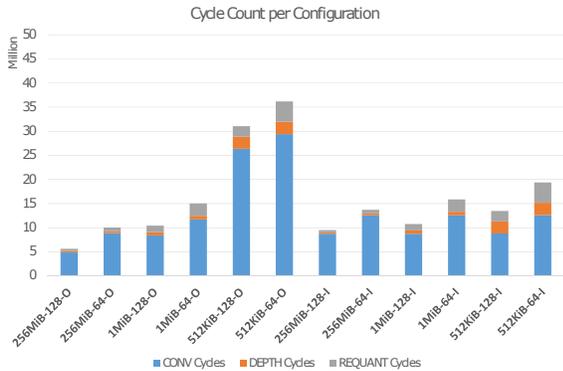


(a) The total cycle count for the accelerated operations of MobileNet without support for depthwise convolutions across testing configurations.



(b) The total cycle count for the accelerated operations of MobileNet with software emulated support for depthwise convolutions across testing configurations.

(c) The total cycle count for the accelerated operations of MobileNet with software-only support for depthwise convolutions across testing configurations.



(d) The total cycle count for the accelerated operations of MobileNet with hardware support for depthwise convolutions across testing configurations.

(e) The total cycle count for the accelerated operations of MobileNet without support for depthwise convolutions across testing configurations.

Figure 4. The first part of the label describes the size of the on-device memory, the second value represents the number of processing elements in the compute unit and the last letter stands either for an input (I) or output (O) parallel hardware implementation.

processing of 8-bit integer data. A small subset of support instructions, like element-wise additions and shifts can also be executed on 32-bit data.

As most layers require a vast number of additional configuration parameters, which would not fit into a reasonably sized instruction, most of these parameters are stored in a separate register file and updated from the system CPU. This has the additional benefit that these values can stay unchanged over multiple operations, without having to submit the same values multiple times, improving the energy efficiency of the solution.

Table 2. Examples of instructions that are part of the accelerator ISA.

Mnemonic	Description	Input Type	Output Type
Compute Layer Instructions			
OP_CONV	standard Conv2D, optional combined with requantization	int8	int32 or int8
OP_CONV_RELU	Conv2D + ReLU	int8	int8
OP_DEPTH_CONV	depthwise Conv2D, optional combined with requantization	int8	int32 or int8
OP_MAT_MUL	Matrix Multiplication	int8	int32
Post-Processing Instructions			
OP_ACT_RELU	ReLU Activation	int8	int8
OP_ACT_LRELU	Leaky ReLU Activation	int8	int8
OP_POOL	Pooling Layer	int8	int8
Elementwise Instructions			
OP_E_ABS	calculates absolute values of single input tensor	int8	int8
OP_C_MIN	compares two tensors of same shape, writes minimum values	int8	int8
OP_C_MAX	compares two tensors of same shape, writes maximum values	int8	int8
OP_E_ADD	adds two tensors of same shape element-wise	int8	int8
OP_E32_ADD	adds two tensors of same shape element-wise	int32	int32
DMA Instructions			
OP_DMA_READ	loads data from system memory to accelerator SRAM	-	-
OP_DMA_WRITE	writes data from accelerator SRAM to system memory	-	-

References

- [1] M. Abadi, A. Agarwal, P. Barham, and more. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL <https://www.tensorflow.org/>. Software available from tensorflow.org.
- [2] ARM. Arm ethos-u. [Online], 2020. URL <https://developer.arm.com/ip-products/processors/machine-learning/ethos-u55>.
- [3] ARM. ARM Ethos-U: Tflite compiler. [Online], 2020. URL <https://git.mlplatform.org/ml/ethos-u/ethos-u-vela.git/>.
- [4] D. Blalock, J. J. G. Ortiz, J. Frankle, and J. Gutttag. What is the state of neural network pruning?, 2020.
- [5] S. Cass. Taking ai to the edge: Google's tpu now comes in a maker-friendly package. *IEEE Spectrum*, 56(5):16–17, 2019.
- [6] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang. Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems, 2015.
- [7] T. Chen, T. Moreau, Z. Jiang, L. Zheng, E. Yan, M. Cowan, H. Shen, L. Wang, Y. Hu, L. Ceze, C. Guestrin, and A. Krishnamurthy. Tvm: An automated end-to-end optimizing compiler for deep learning, 2018.
- [8] Y. Chen, T. Krishna, J. S. Emer, and V. Sze. Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks. *IEEE Journal of Solid-State Circuits*, 52(1):127–138, 2017.
- [9] Y. Chen, T. Yang, J. Emer, and V. Sze. Eyeriss v2: A flexible accelerator for emerging deep neural networks on mobile devices. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 9(2):292–308, 2019.
- [10] Z. Chen and I. Cody Yu, Amazon Web Services. How to bring your own codegen to tvm. [Online], 2020. URL <https://tvm.apache.org/2020/07/15/how-to-bring-your-own-codegen-to-tvm>.
- [11] F. Chollet et al. Keras. [Online], 2015. URL <https://keras.io>.
- [12] G. Coral. Edge tpu compiler documentation. [Online], 2019. URL <https://coral.ai/docs/edgetpu/compiler/>.
- [13] I. Corporation. oneapi onednn. [Online], 2020. URL <https://github.com/oneapi-src/oneDNN/>.
- [14] S. Cyphers, A. K. Bansal, A. Bhiwandiwala, J. Bobba, M. Brookhart, A. Chakraborty, W. Constable, C. Convey, L. Cook, O. Kanawi, R. Kimball, J. Knight, N. Korovaiiko, V. Kumar, Y. Lao, C. R. Lishka, J. Menon, J. Myers, S. A. Narayana, A. Procter, and T. J. Webb. Intel ngraph: An intermediate representation, compiler, and executor for deep learning, 2018.
- [15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.
- [16] Google. Xnnpack - github repository. [Online], 2019. URL <https://github.com/google/XNNPACK>.
- [17] Google. TensorFlow: Xla documentation. [Online], 2020. URL <https://www.tensorflow.org/xla?hl=en>.
- [18] Y. Gorbachev, M. Fedorov, I. Slavutin, A. Tugarev, M. Fatekhov, and Y. Tarkan. Opvino deep learning workbench: Comprehensive analysis and tuning of neural networks inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, Oct 2019.
- [19] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.
- [20] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference, 2017.
- [21] R. Krishnamoorthi. Quantizing deep convolutional networks for efficient inference: A whitepaper, 2018.
- [22] L. Lai, N. Suda, and V. Chandra. Cmsis-nn: Efficient neural network kernels for arm cortex-m cpus, 2018.
- [23] R. M. Larsen and T. Shpeisman. Tensorflow graph optimizations. [Online], 2019. URL <https://research.google/pubs/pub48051/>.
- [24] C. Leary and T. Wang. Xla: Tensorflow, compiled. *TensorFlow Dev Summit*, 2017.
- [25] Y. LeCun, J. S. Denker, and S. A. Solla. Optimal brain damage. In *Advances in neural information processing systems*, pages 598–605, 1990.
- [26] M. Li, Y. Liu, X. Liu, Q. Sun, X. You, H. Yang, Z. Luan, L. Gan, G. Yang, and D. Qian. The deep learning compiler: A comprehensive survey, 2020.
- [27] D. D. Lin, S. S. Talathi, and V. S. Annapureddy. Fixed point quantization of deep convolutional networks, 2015.
- [28] W. Lin, D. Tsai, L. Tang, C. Hsieh, C. Chou, P. Chang, and L. Hsu. Onnc: A compilation framework connecting onnx to proprietary deep learning accelerators. In *2019 IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, pages 214–218, 2019.
- [29] A. Markham and Y. Jia. Caffe2: Portable high-performance deep learning framework from facebook. *NVIDIA Corporation*, 2017.
- [30] T. Moreau, T. Chen, L. Vega, J. Roesch, E. Yan, L. Zheng, J. Fromm, Z. Jiang, L. Ceze, C. Guestrin, and A. Krishnamurthy. A hardware-software blueprint for flexible deep learning specialization, 2018.
- [31] Nvidia. Nvidia deep learning accelerator. [Online], 2018. URL <https://http://nvidia.org/>.
- [32] ONNX. Onnx - official webpage. [Online], 2020. URL <https://onnx.ai/>.
- [33] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019.
- [34] N. Rotem, J. Fix, S. Abdulrasool, G. Catron, S. Deng, R. Dzhabarov, N. Gibson, J. Hegeman, M. Lele, R. Levenstein, et al. Glow: Graph lowering compiler techniques for neural networks. *arXiv preprint arXiv:1805.00907*, 2018.
- [35] A. Shawahna, S. M. Sait, and A. El-Maleh. Fpga-based accelerators of deep learning networks for learning and classification: A review. *IEEE Access*, 7:7823–7859, 2019.
- [36] H. Song, M. Huizi, and D. William J. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding, 2015.
- [37] P. Stock, A. Joulin, R. Gribonval, B. Graham, and H. Jégou. And the bit goes down: Revisiting the quantization of neural networks, 2019.
- [38] T. Team. TensorFlow 2 graph optimizations documentation. [Online], 2019. URL https://www.tensorflow.org/guide/graph_optimization.
- [39] T. Team. TensorFlow Lite. [Online], 2020. URL <https://www.tensorflow.org/lite/>.
- [40] X. Team et al. Xla: Domain-specific compiler for linear algebra that optimizes tensorflow computations, 2019.
- [41] Z. Wang, J. Wohlwend, and T. Lei. Structured pruning of large language models, 2019.
- [42] P. Warden. How to quantize neural networks with tensorflow. [Online], 2016. URL <https://petewarden.com/2016/05/03/how-to-quantize-neural-networks-with-tensorflow>.
- [43] Y. Xing, J. Weng, Y. Wang, L. Sui, Y. Shan, and Y. Wang. An in-depth comparison of compilers for deep neural networks on hardware. In *2019 IEEE International Conference on Embedded Software and Systems (ICESSE)*, pages 1–8. IEEE, 2019.