
Towards Reproducible and Reusable Deep Learning Systems Research Artifacts

Thierry Moreau¹, Anton Lokhmotov², Grigori Fursin^{2,3}

¹University of Washington, USA, ²dividiti, UK, ³cTuning foundation, France

Abstract

1 This paper discusses results and insights from the 1st ReQuEST workshop, a col-
2 lective effort to promote reusability, portability and reproducibility of deep learning
3 research artifacts within the Architecture/PL/Systems communities. ReQuEST
4 (Reproducible Quality-Efficient Systems Tournament) exploits the open-source
5 Collective Knowledge framework (CK) to unify benchmarking, optimization, and
6 co-design of deep learning systems implementations and exchange results via a
7 live multi-objective scoreboard. Systems evaluated under ReQuEST are diverse
8 and include an FPGA-based accelerator, optimized deep learning libraries for x86
9 and ARM systems, and distributed inference in Amazon Cloud and over a cluster
10 of Raspberry Pis. We finally discuss limitations to our approach, and how we plan
11 improve upon those limitations for the upcoming SysML artifact evaluation effort.

12 1 ReQuEST Overview

13 The quest to continually optimize deep learning systems has introduced new deep learning models,
14 frameworks, DSLs, libraries, compilers and hardware architectures. In this frantically changing
15 environment, it has become critical to quickly reproduce, deploy, and build on top of existing research.
16 While open-sourcing research artifacts is one step in the right direction, it is not sufficient to guarantee
17 ease of *reproducibility* and *reusability*. To enable reproducible and reusable research, we need to
18 provide complete, customizable, and portable *workflows* that combine off-the-shelf and custom layers
19 of the system stack and deploys them in a push-button fashion to generate end-to-end metrics of
20 importance.

21 In an effort to promote reproducible, reusable, and portable workflows in deep learning systems
22 research, we introduced the [ReQuEST workshop](#) at the ACM ASPLOS 2018 (for multidisciplinary
23 systems research spanning computer architecture and hardware, programming languages and compilers,
24 operating systems and networking). The goal was to have computer architects, compilers, and
25 systems researchers submit deep learning research artifacts (code, data, and experiments) using a
26 unified [Collective Knowledge](#) (CK) workflow framework [Fursin et al. \(2016\)](#) to produce a *multi-*
27 *objective scoreboard* that would rank submissions under varied cost metrics that include: ImageNet
28 validation (50,000 images), latency (seconds per image), throughput (images per second), platform
29 price (dollars), and peak power consumption (Watts). To keep the task of collecting artifacts tractable,
30 we focused on a single problem: ImageNet classification, but gave complete freedom over what
31 models, frameworks, libraries, compilers and hardware platforms were being used to solve the
32 classification problem.

33 The most important difference of ReQuEST from other related workshops and tournaments such
34 as DawnBench [daw \(2018\)](#) and LPIRC [lpi \(2015\)](#) is that we not only publish final results but also
35 share portable and customizable workflows (i.e. not just Docker images) with all related research
36 components (models, data sets, libraries) to let the community immediately reuse, improve, and build
37 upon them.

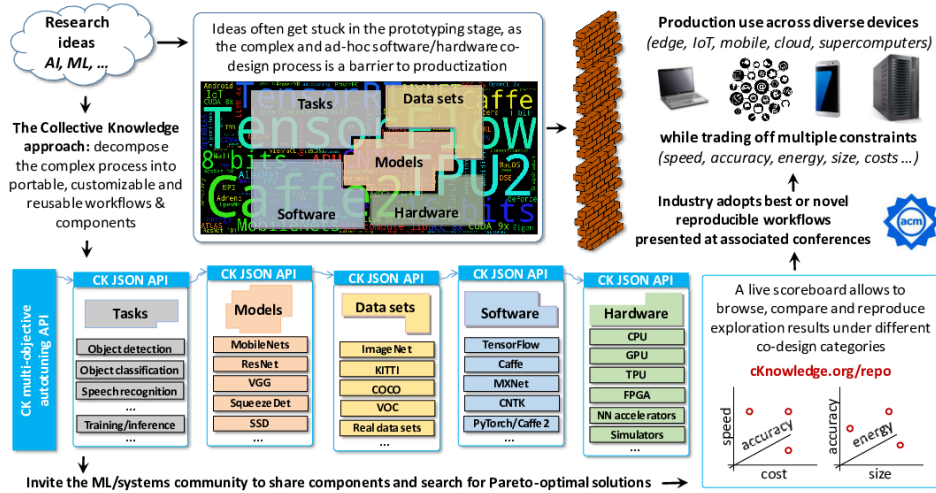


Figure 1: We leverage the open Collective Knowledge workflow framework (CK) and the rigorous ACM artifact evaluation methodology (AE) to allow the community collaboratively explore quality vs. efficiency trade-offs for rapidly evolving workloads across diverse systems.

38 The first iteration of the ReQuEST workshop led to five artifact submissions that were unified under
 39 the CK framework and evaluated (reproduced) by the organizers. What the submissions lacked in
 40 quantity, they made up for in terms of diversity: (1) submissions spanned architecture, compilers,
 41 and systems research, (2) utilized x86, ARM, and FPGA-based platforms; and (3) were deployed on
 42 single-node systems as well as distributed nodes.

43 2 Unifying Artifacts and Workflows with CK

44 ReQuEST aims to promote reproducibility of experimental results and reusability/customization of
 45 systems research artifacts by standardizing evaluation methodologies and facilitating the deployment
 46 of efficient solutions on heterogeneous platforms. For that reason, packaging artifacts (scripts,
 47 libraries, frameworks, data sets, models) and experimental results requires a bit more involvement
 48 than sharing some CSV/JSON files or checking out a given GitHub repository. That is why we
 49 build our competition on top of CK Fursin et al. (2016) to provide unified evaluation and a real-time
 50 leader-board of submissions. CK is an open-source portable workflow framework, used as standard
 51 ACM artifact evaluation methodology from ACM and IEEE systems conferences (CGO, PPOPP,
 52 PACT, SuperComputing).

53 CK works a Python wrapper framework to help users share their code and data as customizable and
 54 reusable plugins with a common JSON API, meta description and an integrated package manager,
 55 adaptable to a user platform with Linux, Windows, MacOS and Android. Researchers can then
 56 quickly prototype experimental workflows from shared components, crowdsource benchmarking and
 57 autotuning across diverse models, data sets and platforms, exchange results via public scoreboards,
 58 and generate interactive reports ck- (2018).

59 3 Artifact Submissions Overview

60 The ReQuEST-ASPLOS' 18 proceedings, available in the ACM Digital Library, include five papers
 61 with Artifact Appendices and a set of ACM reproducibility badges.

62 The CK repository for all ReQuEST-ASPLOS' 18 artifacts are documented and available at the fol-
 63 lowing link: <https://github.com/ctuning/ck-request-asplos18-results>. The interactive live scoreboard
 64 can be accessed under the followig URL: <http://cKnowledge.org/request-results>. The proceed-
 65 ings are accompanied by snapshots of Collective Knowledge workflows covering a very diverse
 66 model/software/hardware stack:

- 67 • **Models:** MobileNets, ResNet-18, ResNet-50, Inception-v3, VGG16, AlexNet, SSD.

- 68 • **Data types:** 8-bit integer, 16-bit floating-point (half), 32-bit floating-point (float).
- 69 • **AI frameworks and libraries:** MXNet, TensorFlow, Caffe, Keras, Arm Compute Library,
70 cuDNN, TVM, NNVM.
- 71 • **Platforms:** Xilinx Pynq-Z1 FPGA, Arm Cortex CPUs and Arm Mali GPGPUs (Linaro
72 HiKey960 and T-Firefly RK3399), a farm of Raspberry Pi devices, NVIDIA Jetson TX1 and
73 TX2, and Intel Xeon servers in Amazon Web Services, Google Cloud and Microsoft Azure.

74 The community can now access all the above CK workflows under permissive licenses and continue
75 collaborating on them via dedicated [ReQuEST'18 GitHub projects](#). First, the workflows can be
76 automatically [adapted](#) to new platforms and environments by either detecting already installed
77 dependencies (e.g. libraries) or rebuilding dependencies via an integrated [package manager](#) supporting
78 Linux, Windows, MacOS and Android. Second, the workflows can be customized by swapping in
79 new models, data sets, frameworks, libraries, and so on. Third, the workflows can be extended to
80 expose new design and optimization choices (e.g. quantization), as well as evaluation metrics (e.g.
81 power or memory consumption). Finally, the workflows can be used for collaborative autotuning
82 ("[crowd-tuning](#)") to explore huge optimization spaces using devices such as [Android phones and](#)
83 [tablets](#), with best solutions being made available to the community on the [online CK scoreboard](#).

84 4 Lessons Learned and Future Work

85 Our overwhelmingly positive experience has also allowed us to critically assess limitations to the
86 scalability to our approach. Fair competitive benchmarking between different platforms, frameworks,
87 and models is hard work. It requires carefully considering model equivalence (e.g. performing
88 the same mix of operations), input equivalence (e.g. preprocessing the inputs in the same way),
89 output equivalence (e.g. validating the outputs for each input, not just calculating the usual aggregate
90 accuracy score), etc. Formalizing the benchmarking requirements and encapsulating them in shared
91 CK components (e.g. using a framework-independent model representation such as [ONNX](#)) and
92 workflows (e.g. for input conversion and output validation), should help standardize and automate the
93 benchmarking process.

94 Thorough artifact evaluation can take several person-weeks. Each submitted workflow needs to
95 be studied in detail in its original form and then converted into a common format. However, the
96 more reusable CK components (such as [workflows](#), [modules/plugins](#), [packages](#)) are shared by
97 the community, the easier the conversion becomes. For example, we have successfully reused
98 several previously [shared components](#) for models, frameworks and libraries, as well as the universal
99 CK workflow for [program benchmarking and autotuning](#). We propose to introduce a new [ACM](#)
100 [reproducibility badge](#) for such unified "plug&play" components. This could eventually lead to
101 creating a "marketplace" for Pareto-efficient implementations (code and data) shared as portable,
102 customizable and reusable CK components.

103 Finally, full experimental evaluation can take many days/weeks. The AE committee can collaborate
104 with the authors to determine a *minimally useful scope* for evaluation which would still provide
105 insights to the community. The community can eventually crowdsource full evaluation. In other
106 words, AE can be "staged" with a quick check that the artifacts are "functional" before the camera-
107 ready deadline followed by full evaluation using the ReQuEST methodology. In fact, ReQuEST
108 can grow into a non-profit service to conferences and journals. Sponsorship should help attract
109 experienced full-time evaluators, as well as part-time volunteers, to work on unifying and evaluating
110 artifacts and workflows.

111 **Future Work** Our experience at ReQuEST-ASPLOS'18 will be repurposed to organize SysML's
112 AE, but at a larger scale. Our long-term vision is to dramatically reduce the complexity and costs of
113 the development and deployment of AI, ML, and other emerging workloads. We believe that having
114 an open repository (marketplace) of customizable workflows with reusable components helps to
115 bring together the multidisciplinary community to collaboratively co-design, optimize, and autotune
116 computer systems across the full model/software/hardware stack. Systems integrators will also
117 benefit from being able to assemble complete solutions by adapting such reusable components to
118 their specific usage scenarios, requirements, and constraints. We envision that our community-driven
119 approach and decentralized marketplace will help accelerate adoption and technology transfer of
120 novel AI/ML techniques similar to the open-source movement.

121 **References**

- 122 LPIRC: low-power image recognition challenge. [https://rebootingcomputing.ieee.org/](https://rebootingcomputing.ieee.org/lpirc)
123 [lpirc](https://rebootingcomputing.ieee.org/lpirc), 2015.
- 124 Industrial and academic use-cases of Collective Knowledge. [http://cKnowledge.org/partners.](http://cKnowledge.org/partners.html)
125 [html](http://cKnowledge.org/partners.html), 2018.
- 126 An end-to-end deep learning benchmark and competition image classification. [https://dawn.cs.](https://dawn.cs.stanford.edu/benchmark/)
127 [stanford.edu/benchmark/](https://dawn.cs.stanford.edu/benchmark/), 2018.
- 128 Fursin, G., Lokhmotov, A., and Plowman, E. Collective Knowledge: towards R&D sustainability.
129 In *Proceedings of the Conference on Design, Automation and Test in Europe (DATE'16)*, March
130 2016. URL [https://www.researchgate.net/publication/304010295_Collective_](https://www.researchgate.net/publication/304010295_Collective_Knowledge_Towards_RD_Sustainability)
131 [Knowledge_Towards_RD_Sustainability](https://www.researchgate.net/publication/304010295_Collective_Knowledge_Towards_RD_Sustainability).