ONNX – MeetUp

ONNX "safety-related profile" Workgroup

June 28th 2024

Eric JENN[1], Jean SOUYRIS[2]

[1] IRT Saint Exupery, Toulouse, France

[2] Airbus, Toulouse, France

Agenda

- **The Needs**
  - Why are we here today?
- **The Solution**
  - What do we plan to do?
- **The Workplan**
  - How do we plan to do it?

Who are we?

- **Group of people from (aeronautical) industry and academia dealing with ML for Safety Related embedded systems**
  - Industry: Airbus, Airbus Helicopters, Thales, Embraer, Safran…
  - Research institutes: CEA, INRIA, IRT Saint Exupery, IRT System-X,, ONERA
  - Members of EUROCAE WG114 / SAE G34 working group on "Artificial Intelligence" (publishing ED-324 / ARP6983).

References (sample)

- Christophe Gabreau *et al*, A study of an ACAS-Xu exact implementation using ED-324/ARP6983, ERTS 2024, Toulouse, France, https://hal.science/hal-04584782
- Gauffriau *et al*, Formal Description of ML models for unambiguous implementation, ERTS 2024, Toulouse, France, https://sciencespo.hal.science/ERTS2024/hal-04167435v2, https://hal.science/hal-04588599
- Vincent Mussot *et al*, Assurance Cases to face the complexity of ML-based systems verification, ERTS 2024, Toulouse, France
- Dumitru Potop Butucaru *et al*, "Bidirectional Reactive Programming for Machine Learning", https://arxiv.org/abs/2311.16977
- Delseny *et al*, White paper "Machine Learning in Certified Systems", see https://arxiv.org/pdf/2103.10529
- Jenn et al, Identifying Challenges to the Certification of Machine Learning for Safety Critical Systems, ERTS 2020, Toulouse, France
- Michele Alberti *et al*, CAISAR: A platform for Characterizing Artificial Intelligence Safety and Robustness", https://arxiv.org/abs/2206.03044
- Iryna De Albuquerque Silva *et al*, ACETONE: Predictable Programming Framework for ML Applications in Safety-Critical Systems, 24th Euromicro Conference on Real-Time Systems (ECRTS 2022), Jun 2022, Modena, Italy.

The Needs

What do we need to embed ML components in (Safety) Critical Applications?

The Needs

- **To embed a ML component in a safety-related system, we have to**
  - Ensure and/or demonstrate that the ML model implementation process **preserves the safety/functional/operational properties** of the model developed during the design process.

- **Prerequisites are**
  - An **accurate and precise** description of the ML model, leaving **no room to interpretation and approximations**…
  - **So: a ML description "language" with a clear syntax and semantics**

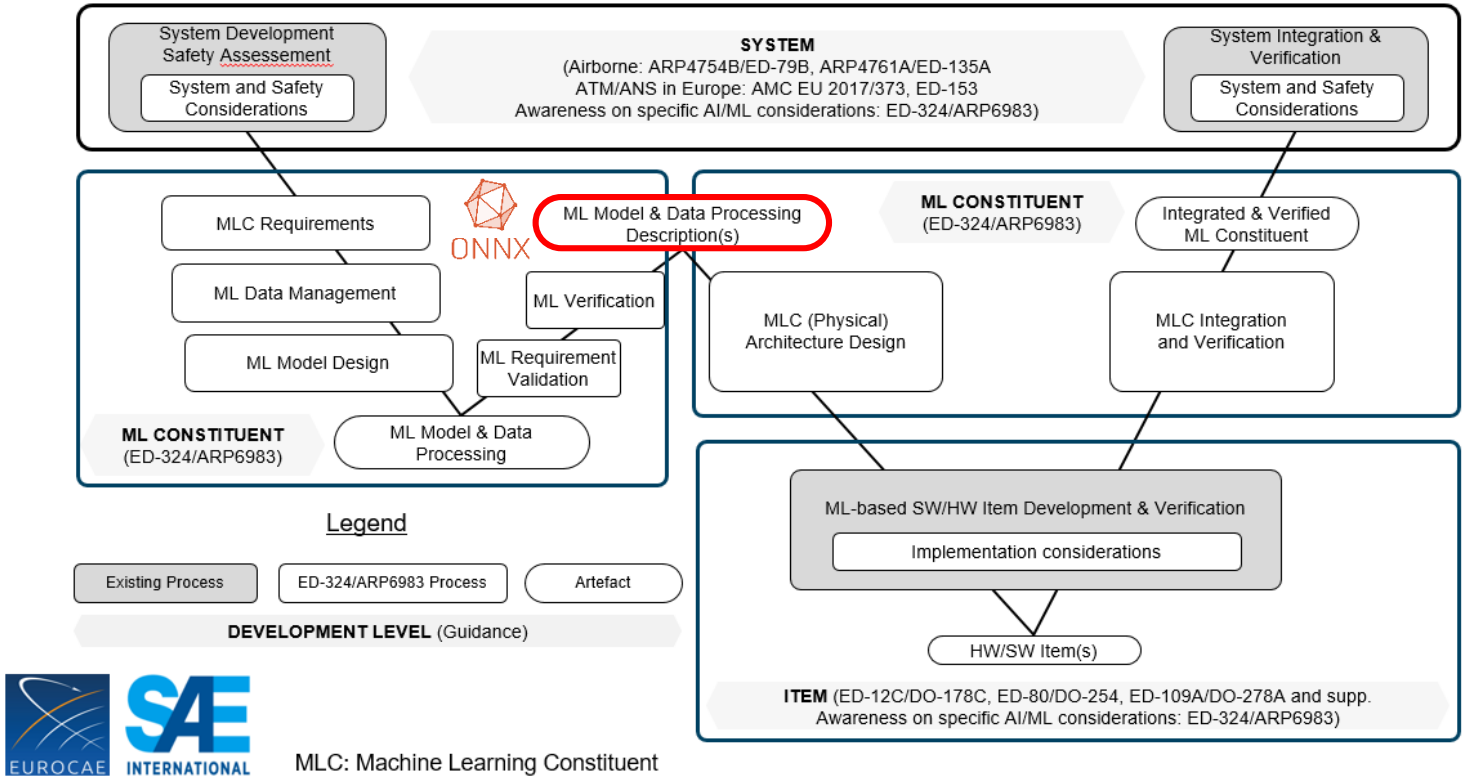
- **We think that ONNX is the best starting point!**

The Needs: Regulation Requirements

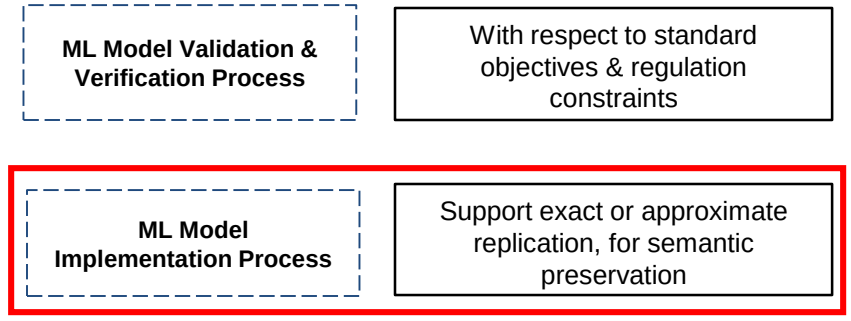
- **Requirements for the engineering of AI systems**
  - ARP6983 / ED-324 in the aeronautic domain
  - ISO/DPAS 8800 in the automotive domain
  - ECSS-E-HB-40-02A DIR1 "Space engineering – Machine learning qualification handbook" in the space domain
  - etc.

The Needs: Regulation Requirements

❑ Example: Requirements according to ARP6983 / ED-324



"The ML Model description should contain sufficient details on the ML Model semantic to fully preserve this semantic in the implemented ML Model" [ARP6983/ED-324]



The Solution

What needs to be done on the ONNX standard?

Challenges for ONNX

Provide an accurate and precise description of the ML model leaving no room to interpretation and approximations…

- Complete the definition and documentation of
  - The operator semantics
  - The graph semantics
  - The datatypes
  - The ONNX abstract (metamodel) and concrete (format) syntax

Complete list of requirements to be established…

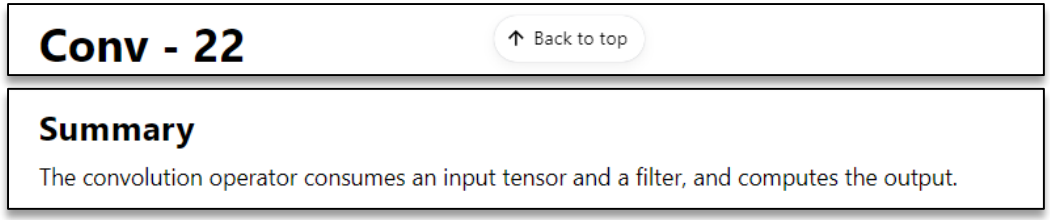
Analysis to be done…

Challenges for ONNX

Provide an accurate and precise description of the ML model leaving no room to interpretation and approximations…

- Complete the definition and documentation of
  - **The operator semantics**
  - The graph semantics

  for all datatypes
  - The ONNX abstract (metamodel) and concrete (format) syntax

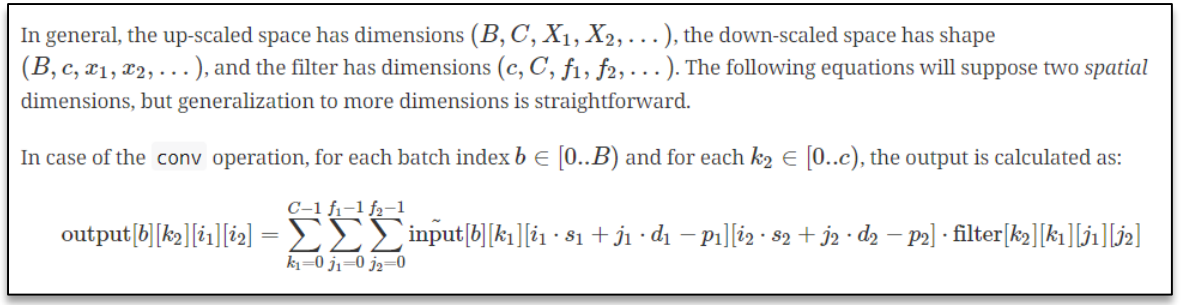


Conv - 22 ↑ Back to top

Summary

The convolution operator consumes an input tensor and a filter, and computes the output.

(Excerpt of ONNX doc.)



In general, the up-scaled space has dimensions $(B, C, X_1, X_2, \dots)$, the down-scaled space has shape $(B, c, x_1, x_2, \dots)$, and the filter has dimensions $(c, C, f_1, f_2, \dots)$. The following equations will suppose two *spatial* dimensions, but generalization to more dimensions is straightforward.

In case of the `conv` operation, for each batch index $b \in [0..B)$ and for each $k_2 \in [0..c)$, the output is calculated as:

$$\mathrm{output}[b][k_2][i_1][i_2] = \sum_{k_1=0}^{C-1} \sum_{j_1=0}^{f_1-1} \sum_{j_2=0}^{f_2-1} \tilde{\mathrm{input}}[b][k_1][i_1 \cdot s_1 + j_1 \cdot d_1 - p_1][i_2 \cdot s_2 + j_2 \cdot d_2 - p_2] \cdot \mathrm{filter}[k_2][k_1][j_1][j_2]$$

(Excerpt of NNEF doc.)

Challenges for ONNX

Provide an accurate and precise description of the ML model leaving no room to interpretation and approximations…

- Complete the definition and documentation of
  - The operator semantics
  - **The graph semantics**

  for all datatypes
  - The ONNX abstract (metamodel) and concrete (format) syntax

In what order are the operators of a graph executed?

Compliance with dataflow constraints. Sufficient?

ONNX runtime

- Default execution order uses Graph::ReverseDFS() to generated topological sort
- Priority-based execution order uses Graph::KahnsTopologicalSort with per-node priority

Challenges for ONNX

Provide an accurate and precise description of the ML model leaving no room to interpretation and approximations…

- Complete the definition and documentation of
  - The operator semantics
  - The graph semantics

  for all datatypes
  - **The ONNX abstract (metamodel) and concrete (format) syntax**

For instance…

https://github.com/onnx/onnx/blob/main/onnx/onnx.proto

```
// A list of function protos local to the model.
//
// The (domain, name, overload) tuple must be unique across the function protos in this list.
// In case of any conflicts the behavior (whether the model local functions are given higher priority,
// or standard operator sets are given higher priority or this is treated as error) is defined by
// the runtimes.
```

Challenges for ONNX

Provide an accurate and precise description of the ML model leaving no room to interpretation and approximations…

- Complete the definition and documentation of
  - The operator semantics
  - The graph semantics
  - The ONNX abstract (metamodel) and concrete (format) syntax
- *Also consider other features to…*
  - Facilitate traceability
  - Improve understandability
  - Etc.

For instance…

Use doc string to
- enforce the documentation of the meaning of each dimensions of tensors…
- add traceability data

The Workplan

Towards a “satefy-related profile” for ONNX

Activities and Deliverables

1. **Capture the needs and elicit the reqs for the safety-related profile**
   - ❑ What do we need exactly?
   - ❑ What do we require from ONNX?

2. **Analyse the ONNX standard with respect to the requirements**
   - ❑ What needs to be clarified? Completed?

3. **Develop the "safety-related profile" for ONNX**
   - ❑ Complete and clarifies the standards where necessary.

Activities and Deliverables

1. **Capture the needs and elicit the reqs for the safety-related profile**
   - ❑ D1.a: Safety-Related Profile Needs
   - ❑ D1.b: Safety-Related Profile Scope Definition
   - ❑ D1.c: Safety-Related Profile Requirements Specification
2. **Analyse the ONNX standard against requirements**
   - ❑ D2: ONNX standard improvements for the safety-related profile
3. **Develop the safety-related profile**
   - ❑ D3: ONNX safety-related profile

And Next

- **Workgroup creation is on-going**
  - Chairs: Jean SOUYRIS (Airbus) and Eric JENN (IRT Saint Exupery)
- **1st meeting planned end of September**
  - *Agenda to come*
- **Modalities (to be defined)**
  - Every 2 weeks: report of activities, monitoring of progress, distribution of tasks
  - Every 2 months: sub-groups synchronisation and consolidation

Contacts

- Eric JENN (eric.jenn@irt-saintexupery.com)
- Jean SOUYRIS (jean.souyris@airbus.com)

www.confiance.ai
contact@irt-systemx.fr