INTERNATIONAL STANDARD

ISO/IEC 24029-2

First edition
2023-08

Artificial intelligence (AI) — Assessment of the robustness of neural networks —

Part 2: Methodology for the use of formal methods

Intelligence artificielle (IA) — Evaluation de la robustesse de réseaux neuronaux —

Partie 2: Méthodologie pour l'utilisation de méthodes formelles

Reference number
ISO/IEC 24029-2:2023(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives or www.iec.ch/members_experts/refdocs).

ISO and IEC draw attention to the possibility that the implementation of this document may involve the use of (a) patent(s). ISO and IEC take no position concerning the evidence, validity or applicability of any claimed patent rights in respect thereof. As of the date of publication of this document, ISO and IEC had not received notice of (a) patent(s) which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the patent database available at www.iso.org/patents and https://patents.iec.ch. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see www.iso.org/iso/foreword.html. In the IEC, see www.iec.ch/understanding-standards.

This document was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 42, *Artificial intelligence,* in collaboration with the European Committee for Standardization (CEN) Technical Committee CEN/CLC/JTC 21, *Artificial Intelligence,* in accordance with the Agreement on technical cooperation between ISO and CEN (Vienna Agreement).

A list of all parts in the ISO/IEC 24029 series can be found on the ISO and IEC websites.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html and www.iec.ch/national-committees.

Introduction

Neural networks are widely used to perform complex tasks in various contexts, such as image or natural language processing and predictive maintenance. AI system quality models comprise certain characteristics, including robustness. For example, ISO/IEC 25059:2023,[1] which extends the SQuaRE International Standards[2] to AI systems, considers in its quality model that robustness is a sub-characteristic of reliability. Demonstrating the ability of a system to maintain its level of performance under varying conditions can be done using statistical analysis, but proving it requires some form of formal analysis. In that regard formal methods can be complementary to other methods in order to increase trust in the robustness of the neural network.

Formal methods are mathematical techniques for rigorous specification and verification of software and hardware systems with the goal to prove their correctness. Formal methods can be used to formally reason about neural networks and prove whether they satisfy relevant robustness properties. For example, consider a neural network classifier that takes as input an image and outputs a label from a fixed set of classes (such as car or airplane). Such a classifier can be formalized as a mathematical function that takes the pixel intensities of an image as input, computes the probabilities for each possible class from the fixed set, and returns a label corresponding to the highest probability. This formal model can then be used to mathematically reason about the neural network when the input image is modified. For example, suppose when given a concrete image for which the neural network outputs the label “car” the following question can be asked: “does the network output a different label if the value of an arbitrary pixel in the image is modified?” This question can be formulated as a formal mathematical statement that is either true or false for a given neural network and image.

A classical approach to using formal methods consists of three main steps that are described in this document. First, the system to be analyzed is formally defined in a model that precisely captures all possible behaviours of the system. Then, a requirement is mathematically defined. Finally, a formal method, such as solver, abstract interpretation or model checking, is used to assess whether the system meets the given requirement, yielding either a proof, a counterexample or an inconclusive result.

This document covers several available formal method techniques. At each stage of the life cycle, the document presents criteria that are applicable to assess the robustness of neural networks and to establish how neural networks are verified by formal methods. Formal methods can have issues in terms of scalability, however, they are still applicable to all types of neural networks performing various tasks on several data types. While formal methods have long been used on traditional software systems, the use of formal methods on neural networks is fairly recent and is still an active field of investigation.

This document is aimed at helping AI developers who use neural networks and who are tasked with assessing their robustness throughout the appropriate stages of the AI life cycle. ISO/IEC TR 24029-1 provides a more detailed overview of the techniques available to assess the robustness of neural networks, beyond the formal methods described in this document.

INTERNATIONAL STANDARD ISO/IEC 24029-2:2023(E)

Artificial intelligence (AI) — Assessment of the robustness of neural networks —

Part 2:
Methodology for the use of formal methods

1 Scope

This document provides methodology for the use of formal methods to assess robustness properties of neural networks. The document focuses on how to select, apply and manage formal methods to prove robustness properties.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 22989:2022, *Information technology — Artificial intelligence — Artificial intelligence concepts and terminology*

ISO/IEC 23053:2022, *Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO/IEC 22989:2022, ISO/IEC 23053:2022 and the following apply.

ISO and IEC maintain terminology databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at https://www.electropedia.org/

3.1
domain
set of possible inputs to a neural network characterized by attributes of the environment

EXAMPLE 1 A neural network performing a natural language processing task is manipulating texts composed of words. Even though the number of possible different texts is unbounded, the maximum length of each sentence is always bounded. An attribute describing this domain can therefore be the maximum length allowed for each sentence.

EXAMPLE 2 A face capture domain requirements can rely on attributes such as that the size of faces is at least 40 pixels by 40 pixels. That half-profile faces are detectable at a lower level of accuracy, provided most of the facial features are still visible. Similarly, partial occlusions are handled to some extent. Detection typically requires that more than 70 % of the face is visible. Views where the camera is the same height as the face perform best and performance degrades as the view moves above 30 degrees or below 20 degrees from straight on.

Note 1 to entry: An attribute is used to describe a bounded object even though the domain can be unbounded.