Kunstig intelligens (AI) – Vurdering af neurale netværks robusthed- Del 1: Overblik

Artificial Intelligence (AI) – Assessment of the robustness of neural networks – Part 1: Overview

DANSK STANDARD
Danish Standards Association

Göteborg Plads 1
DK-2150 Nordhavn

Tel: +45 39 96 61 01
dansk.standard@ds.dk
www.ds.dk

DS projekt: M345355
ICS: 35.020

Første del af denne publikations betegnelse er:
DS/ISO/IEC TR, hvilket betyder, at det er en international teknisk rapport, der har status som DS-information.

Denne publikations overensstemmelse er:
IDT med: ISO/IEC TR 24029-1:2021

DS-publikationen er på engelsk.

DS-publikationstyper
Dansk Standard udgiver forskellige publikationstyper.
Typen på denne publikation fremgår af forsiden.

Der kan være tale om:

Dansk standard
- standard, der er udarbejdet på nationalt niveau, eller som er baseret på et andet lands nationale standard, eller
- standard, der er udarbejdet på internationalt og/eller europæisk niveau, og som har fået status som dansk standard

DS-information
- publikation, der er udarbejdet på nationalt niveau, og som ikke har opnået status som standard, eller
- publikation, der er udarbejdet på internationalt og/eller europæisk niveau, og som ikke har fået status som standard, fx en teknisk rapport, eller
- europæisk præstandard

DS-håndbog
- samling af standarder, eventuelt suppleret med informativt materiale

DS-hæfte
- publikation med informativt materiale

Til disse publikationstyper kan endvidere udgives
- tillæg og rettelsesblade

DS-publikationsform
Publikationstyperne udgives i forskellig form som henholdsvis
- fuldtekstpublikation (publikationen er trykt i sin helhed)
- godkendelsesblad (publipukationen leveres i kopi med et trykt DS-omslag)
- elektronisk (publikationen leveres på et elektronisk medie)

DS-betegnelse
Alle DS-publikationers betegnelse begynder med DS efterfulgt af et eller flere præfikser og et nr., fx **DS 383**, **DS/EN 5414** osv. Hvis der efter nr. er angivet et **A** eller **Cor**, betyder det, enten at det er et **tillæg** eller et **rettelsesblad** til hovedstandarden, eller at det er indført i hovedstandarden.
DS-betegnelse angives på forsiden.

Overensstemmelse med anden publikation:
Overensstemmelse kan enten være IDT, EQV, NEQ eller MOD
- **IDT:** Når publikationen er identisk med en given publikation.
- **EQV:** Når publikationen teknisk er i overensstemmelse med en given publikation, men præsentationen er ændret.
- **NEQ:** Når publikationen teknisk eller præsentationsmæssigt ikke er i overensstemmelse med en given standard, men udarbejdet på baggrund af denne.
- **MOD:** Når publikationen er modificeret i forhold til en given publikation.

TECHNICAL ISO/IEC TR

First edition
2021-03-10

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

Part 1: Overview

Reference number
ISO/IEC TR 24029-1:2021(E)

ISO copyright office
Ch. de Blandonnet 8 • CP 401
CH-1214 Vernier, Geneva, Switzerland
Tel. +41 22 749 01 11
Fax +41 22 749 09 47
copyright@iso.org
www.iso.org

Contents

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).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents) or the IEC list of patent declarations received (see patents.iec.ch).

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*.

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

When designing an AI system, several properties are often considered desirable, such as robustness, resiliency, reliability, accuracy, safety, security, privacy. A definition of robustness is provided in 3.6. Robustness is a crucial property that poses new challenges in the context of AI systems. For example, in AI systems there are some risks specifically tied to the robustness of AI systems. Understanding these risks is essential for the adoption of AI in many contexts. This document aims at providing an overview of the approaches available to assess these risks, with a particular focus on neural networks, which are heavily used in industry, government and academia.

In many organizations, software validation is an essential part of putting software into production. The objective is to ensure various properties including safety and performance of the software used in all parts of the system. In some domains, the software validation and verification process is also an important part of system certification. For example, in the automotive or aeronautic fields, existing standards, such as ISO 26262 or Reference [2], require some specific actions to justify the design, the implementation and the testing of any piece of embedded software.

The techniques used in AI systems are also subject to validation. However, common techniques used in AI systems pose new challenges that require specific approaches in order to ensure adequate testing and validation.

AI technologies are designed to fulfil various tasks, including interpolation/regression, classification and other tasks.

While many methods exist for validating non-AI systems, they are not always directly applicable to AI systems, and neural networks in particular. Neural network systems represent a specific challenge as they are both hard to explain and sometimes have unexpected behaviour due to their non-linear nature. As a result, alternative approaches are needed.

Methods are categorized into three groups: statistical methods, formal methods and empirical methods. This document provides background on these methods to assess the robustness of neural networks.

It is noted that characterizing the robustness of neural networks is an open area of research, and there are limitations to both testing and validation approaches.

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

Part 1: Overview

1 Scope

This document provides background about existing methods to assess the robustness of neural networks.

2 Normative references

There are no normative references in this document.