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

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

Contents

Foreword

Introduction

1 Scope

2 Normative references

3 Terms and definitions

4 Overview of the existing methods to assess the robustness of neural networks

4.1 General
4.1.1 Robustness concept
4.1.2 Typical workflow to assess robustness
4.2 Classification of methods

5 Statistical methods

5.1 General
5.2 Robustness metrics available using statistical methods
5.2.1 General
5.2.2 Examples of performance measures for interpolation
5.2.2.1 Root mean square error or root mean square deviation
5.2.2.2 Max error
5.2.2.3 Actual/predicted correlation
5.2.3 Examples of performance measures for classification
5.2.3.1 General notions and associated basic metrics
5.2.3.2 Advanced metrics
5.2.3.2.1 Precision recall curve
5.2.3.2.2 Receiver operating characteristic (ROC)
5.2.3.3 Lift
5.2.3.4 Area under curve
5.2.3.5 Balanced accuracy
5.2.3.6 Micro average and macro average
5.2.3.7 Matthews correlation coefficient (MCC)
5.2.3.8 Confusion matrix and associated metrics
5.2.4 Other measures
5.2.4.1 Hinge loss
5.2.4.2 Cohen's kappa
5.3 Statistical methods to measure robustness of a neural network
5.3.1 General
5.3.2 Contrastive measures

6 Formal methods

6.1 General
6.2 Robustness goal achievable using formal methods
6.2.1 General
6.2.2 Interpolation stability
6.2.3 Maximum stable space for perturbation resistance
6.3 Conduct the testing using formal methods
6.3.1 Using uncertainty analysis to prove interpolation stability
6.3.2 Using solver to prove a maximum stable space property
6.3.3 Using optimization techniques to prove a maximum stable space property
6.3.4 Using abstract interpretation to prove a maximum stable space property

7 Empirical methods

7.1 General
7.2 Field trials
7.3 A posteriori testing
7.4 Benchmarking of neural networks

Annex A (informative) Data perturbation

A.1 General
A.2 Example image perturbations
A.2.1 General
A.2.2 Homogeneous noising
A.2.3 Brightening
A.2.4 Vibration and rotation
A.2.5 Atmospheric turbulence
A.2.6 Blurring
A.2.7 Blooming
A.2.8 Smear
A.3 Example sound perturbations
A.3.1 Principle
A.3.2 Sound attacks in human-audible frequency range
A.3.3 Ultrasound based attacks

Annex B (informative) Principle of abstract interpretation

B.1 Principle of abstract interpretation

Page count: 31