Title (en)
METHOD AND APPARATUS FOR COMPRESSING A NEURAL NETWORK

Title (de)
VERFAHREN UND VORRICHTUNG ZUM KOMPRIMIEREN EINES NEURONALEN NETZES

Title (fr)
PROCÉDÉ ET APPAREIL DE COMPRESSION D'UN RÉSEAU NEURONAL

Publication
EP 4115337 A1 20230111 (DE)

Application
EP 21707207 A 20210218

Priority
• DE 102020202795 A 20200304
• EP 2021054084 W 20210218

Abstract (en)
[origin: WO2021175615A1] The invention relates to a method for compressing a neural network (10), wherein a trained neural network (10) is obtained, wherein at least one piece of structure information from the trained neural network (10) is extracted or obtained, wherein the trained neural network (10) is divided into subsets (I, II, III, IV, V) on the basis of the at least one piece of structure information for the purpose of compression, wherein one compression method is selected and used for each subset (I, II, III, IV, V) on the basis of at least one property of the respective subset (I, II, III, IV, V), and wherein the compressed neural network (11) is provided. The invention also relates to an apparatus (1) for compressing a neural network (10), to a computer program and to a data carrier signal.

IPC 8 full level
G06N 3/04 (2006.01)

CPC (source: EP)
G06N 3/045 (2023.01); **G06N 3/082** (2013.01); G01S 13/931 (2013.01); G01S 17/931 (2020.01)

Citation (search report)
See references of WO 2021175615A1

Designated contracting state (EPC)
AL AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO RS SE SI SK SM TR

Designated extension state (EPC)
BA ME

Designated validation state (EPC)
KH MA MD TN

DOCDB simple family (publication)
DE 102020202795 A1 20210909; EP 4115337 A1 20230111; WO 2021175615 A1 20210910

DOCDB simple family (application)
DE 102020202795 A 20200304; EP 2021054084 W 20210218; EP 21707207 A 20210218