Title (en)

METHOD AND APPARATUS FOR QUANTIZATION, ADAPTIVE BLOCK PARTITIONING AND CODEBOOK CODING FOR NEURAL NETWORK MODEL COMPRESSION

Title (de)

VERFAHREN UND VORRICHTUNG ZUR QUANTISIERUNG, ADAPTIVEN BLOCKPARTITIONIERUNG UND CODEBUCH-CODIERUNG FÜR KOMPRESSION EINES MODELLS EINES NEURONALEN NETZES

Title (fr)

PROCÉDÉ ET APPAREIL DE QUANTIFICATION, DE PARTITIONNEMENT DE BLOC ADAPTATIF ET DE CODAGE DE LIVRE DE CODES POUR COMPRESSION DE MODÈLE DE RÉSEAU NEURONAL

Publication

EP 4062375 A1 20220928 (EN)

Application

EP 20890921 A 20201119

Priority

- US 201962939054 P 20191122
- US 201962939057 P 20191122
- US 201962939949 P 20191125
- US 201962947236 P 20191212
- US 202017099202 A 20201116
- US 2020061258 W 20201119

Abstract (en)

[origin: WO2021102125A1] A method of quantization, adaptive block partitioning and codebook coding for neural network model compression, is performed by at least one processor and includes determining a saturated maximum value of a multi-dimensional tensor in a layer of a neural network, and a bit depth corresponding to the saturated maximum value, and clipping weight coefficients in the multi-dimensional tensor to be within a range of the saturated maximum value. The method further includes quantizing the clipped weight coefficients, based on the bit depth, and transmitting, to a decoder, a layer header including the bit depth.

IPC 8 full level

G06T 9/00 (2006.01)

CPC (source: EP KR)

G06N 3/08 (2013.01 - EP KR); **H04N 19/119** (2014.11 - KR); **H04N 19/124** (2014.11 - KR); **H04N 19/129** (2014.11 - KR); **H04N 19/70** (2014.11 - KR); G06N 3/045 (2023.01 - EP)

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

DOCDB simple family (publication)

WO 2021102125 A1 20210527; CN 113795869 A 20211214; CN 113795869 B 20230818; EP 4062375 A1 20220928; EP 4062375 A4 20221228; JP 2022533307 A 20220722; JP 7337950 B2 20230904; KR 20210136123 A 20211116

DOCDB simple family (application)

US 2020061258 W 20201119; CN 202080033543 A 20201119; EP 20890921 A 20201119; JP 2021559625 A 20201119; KR 20217033218 A 20201119