(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
10 April 2025 (10.04.2025)

WIPO | PCT

(10) International Publication Number
WO 2025/073424 A1

(51) International Patent Classification:
G06N 3/045 (2023.01) *G06N 3/082* (2023.01)

(21) International Application Number:
PCT/EP2024/074756

(22) International Filing Date:
04 September 2024 (04.09.2024)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

23383011.6	02 October 2023 (02.10.2023)	EP
24172279.2	24 April 2024 (24.04.2024)	EP
18/654,223	03 May 2024 (03.05.2024)	US

(71) Applicant: **MULTIVERSE COMPUTING S.L.** [ES/ES]; Paseo de Miramón, 170 3° Planta, 20014 Donostia / San Sebastián (ES).

(72) Inventors: **TOMUT, Andrei V**; Paseo de Miramón, 170 3° Planta, 20014 Donostia / San Sebastián (ES). **JAHROMI, Saeed S.**; Paseo de Miramón, 170 3° Planta, 20014 Donostia / San Sebastián (ES). **ORÚS, Román**; Paseo de Miramón, 170 3° Planta, 20014 Donostia / San Sebastián (ES).

(74) Agent: **SONNENBERG HARRISON PARTNERSCHAFT MBB**; Herzogspitalstraße 10a, 80331 Munich (DE).

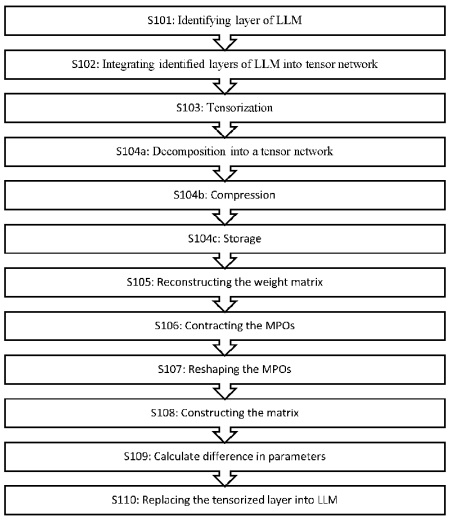
(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

(54) Title: SYSTEM AND METHOD FOR COMPRESSING LARGE LANGUAGE MODEL USING TENSOR NETWORKS

(57) Abstract: A computer implemented method for compressing pre-trained of a large language model (LLM) is provided. The method comprises identifying (S101) layers of the LLM (47) with the weight matrices (48), decomposing (S104a) the weight matrices (48) of the LLM (47) into a tensor network (49), compressing (S104b) the tensor network (49), and storing (S104c) the compressed tensor network (49) in a data storage unit (40).

Fig. 3



WO 2025/073424 A1

[Continued on next page]

Published:

— *with international search report (Art. 21(3))*

— *in black and white; the international application as filed contained color or greyscale and is available for download from PATENTSCOPE*