



- (51) **International Patent Classification:**
G06N 3/04 (2006.01) *G06N 3/063* (2006.01)
- (21) **International Application Number:**
PCT/US2017/049165
- (22) **International Filing Date:**
29 August 2017 (29.08.2017)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (30) **Priority Data:**
15/336,216 27 October 2016 (27.10.2016) US
15/455,685 10 March 2017 (10.03.2017) US
- (71) **Applicant:** GOOGLE LLC [US/US]; 1600 Amphitheatre Parkway, Mountain View, California 94043 (US).
- (72) **Inventors:** NARAYANASWAMI, Ravi; 1600 Amphitheatre Parkway, Mountain View, California 94043 (US).
WOO, Dong Hyuk; 1600 Amphitheatre Parkway, Moun-

tain View, California 94043 (US). TEMAM, Olivier; 8 Rue de Londres, 75009 Paris (FR). KHAITAN, Harshit; 1600 Amphitheatre Parkway, Mountain View, California 94043 (US).

(74) **Agent:** HENRY, Joel et al.; Fish & Richardson P.C., P.O. Box 1022, Minneapolis, Minnesota 55440-1022 (US).

(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, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, 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, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(54) **Title:** NEURAL NETWORK INSTRUCTION SET ARCHITECTURE

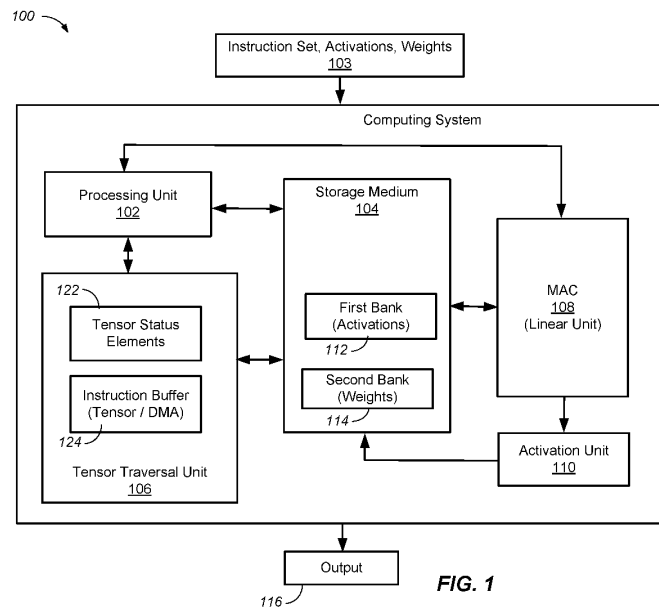


FIG. 1

(57) **Abstract:** A computer-implemented method that includes receiving, by a processing unit, an instruction that specifies data values for performing a tensor computation. In response to receiving the instruction, the method may include, performing, by the processing unit, the tensor computation by executing a loop nest comprising a plurality of loops, wherein a structure of the loop nest is defined based on one or more of the data values of the instruction. The tensor computation can be at least a portion of a computation of a neural network layer. The data values specified by the instruction may comprise a value that specifies a type of the neural network layer, and the structure of the loop nest can be defined at least in part by the type of the neural network layer.

[Continued on next page]

(84) Designated States (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, 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, 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).

Declarations under Rule 4.17:

- *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*
- *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

Published:

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