

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

(19) World Intellectual Property
Organization
International Bureau

(43) International Publication Date
16 April 2020 (16.04.2020)



(10) International Publication Number
WO 2020/076392 A1

(51) International Patent Classification:

G06N 5/02 (2006.01) G06F 12/0862 (2016.01)
G06N 20/20 (2019.01) G06F 12/0897 (2016.01)
G06F 12/0811 (2016.01) G06F 12/126 (2016.01)

(21) International Application Number:

PCT/US2019/043861

(22) International Filing Date:

29 July 2019 (29.07.2019)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

16/156,573 10 October 2018 (10.10.2018) US

(71) Applicant: **GOOGLE LLC** [US/US]; 1600 Amphitheatre
Parkway, Mountain View, California 94043 (US).

(72) Inventors: **YOON, Doe Hyun**; 1600 Amphitheatre Park-
way, Mountain View, California 94043 (US). **PATIL,
Nishant**; 1600 Amphitheatre Parkway, Mountain View,
California 94043 (US). **JOUPPI, Norman Paul**; 1600
Amphitheatre Parkway, Mountain View, California 94043
(US).

(74) Agent: **WRIGHT, Christopher D.** et al.; FISH &
RICHARDSON P.C., P.O. Box 1022, Minneapolis, Min-
nesota 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,

(54) Title: MODIFYING MACHINE LEARNING MODELS TO IMPROVE LOCALITY

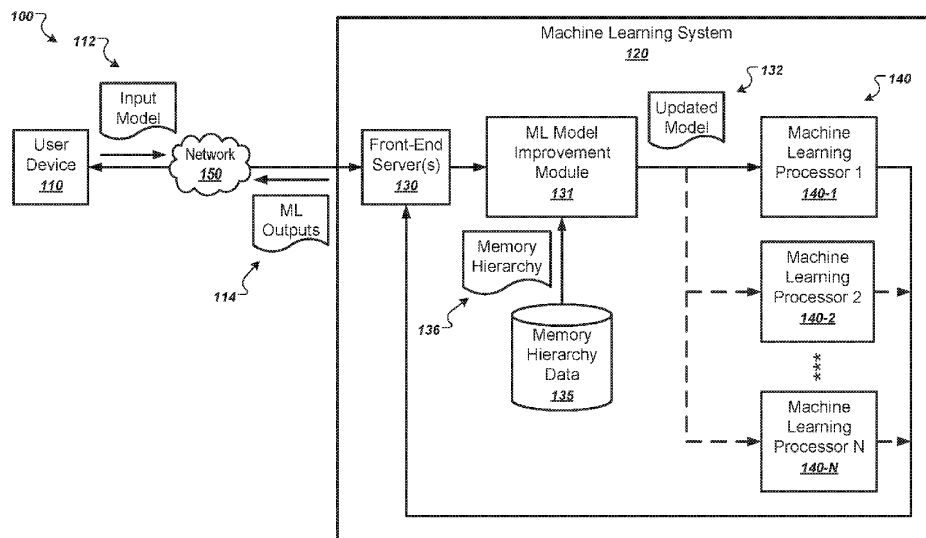


FIG. 1

(57) Abstract: Methods, systems, and apparatus for updating machine learning models to improve locality are described. In one aspect, a method includes receiving data of a machine learning model. The data represents operations of the machine learning model and data dependencies between the operations. Data specifying characteristics of a memory hierarchy for a machine learning processor on which the machine learning model is going to be deployed is received. The memory hierarchy includes multiple memories at multiple memory levels for storing machine learning data used by the machine learning processor when performing machine learning computations using the machine learning model. An updated machine learning model is generated by modifying the operations and control dependencies of the machine learning model to account for the characteristics of the memory hierarchy. Machine learning computations are performed using the updated machine learning model.

[Continued on next page]

WO 2020/076392 A1

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.

(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))*