

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

(19) World Intellectual Property
Organization
International Bureau

(43) International Publication Date
02 December 2021 (02.12.2021)



(10) International Publication Number
WO 2021/242584 A1

(51) International Patent Classification:

G06K 9/00 (2006.01) G06N 3/08 (2006.01)
G06K 9/62 (2006.01) H01J 49/00 (2006.01)
A61B 5/00 (2006.01)

(21) International Application Number:

PCT/US2021/033106

(22) International Filing Date:

19 May 2021 (19.05.2021)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

16/887,381 29 May 2020 (29.05.2020) US

(71) Applicant: PAYPAL, INC. [US/US]; 2211 North First Street, San Jose, California 95131 (US).

(72) Inventor: ZHANG, Jiyi; 2211 North First Street, San Jose, California 95131 (US).

(74) Agent: CHEN, Tom; 2323 Victory Ave., Suite 700, Dallas, Texas 75219 (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, IT, 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, 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, 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,

(54) Title: WATERMARK AS HONEYPOT FOR ADVERSARIAL DEFENSE

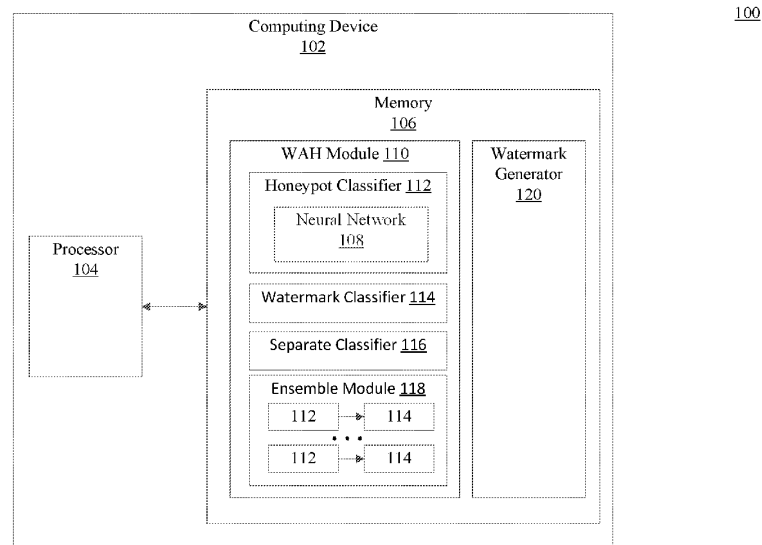


FIG. 1

(57) Abstract: Systems, methods, and computer program products for determining an attack on a neural network. A data sample is received at a first classifier neural network and at a watermark classifier neural network, wherein the first classifier neural network is trained using a first dataset and a watermark dataset. The first classifier neural network determines a classification label for the data sample. A watermark classifier neural network determines a watermark classification label for the data sample. A data sample is determined as an adversarial data sample based on the classification label for the data sample and the watermark classification label for the data sample.

[Continued on next page]

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).

Published:

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