

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

(19) World Intellectual Property
Organization
International Bureau

(43) International Publication Date
24 October 2024 (24.10.2024)



(10) International Publication Number
WO 2024/220790 A1

(51) International Patent Classification:

G06N 3/02 (2006.01) G06V 10/764 (2022.01)
G06N 3/08 (2023.01) G06V 10/774 (2022.01)
G06F 18/214 (2023.01)

(21) International Application Number:

PCT/US2024/025383

(22) International Filing Date:

19 April 2024 (19.04.2024)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

63/497,543 21 April 2023 (21.04.2023) US

(71) Applicant: VISA INTERNATIONAL SERVICE
ASSOCIATION [US/US]; P.O. Box 8999, San Francisco,
California 94128 (US).

(72) Inventors: LIU, Miaomiao; Visa International Service As-
sociation, P.O. Box 8999, San Francisco, California 94128
(US). KERSTING, Nicholas, Stephen; Visa Internation-
al Service Association, P.O. Box 8999, San Francisco, Cal-
ifornia 94128 (US). HE, Runxin; Visa International Ser-
vice Association, P.O. Box 8999, San Francisco, California

94128 (US). GU, Yu; Visa International Service Associa-
tion, P.O. Box 8999, San Francisco, California 94128 (US).

(74) Agent: PREPELKA, Nathan, J. et al.; The Webb Law
Firm, One Gateway Center, 420 Ft. Duquesne Blvd., Suite
1200, Pittsburgh, Pennsylvania 15222 (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, 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,

(54) Title: METHOD, SYSTEM, AND COMPUTER PROGRAM PRODUCT FOR MULTI-LAYER ANALYSIS AND DETECTION
OF VULNERABILITY OF MACHINE LEARNING MODELS TO ADVERSARIAL ATTACKS

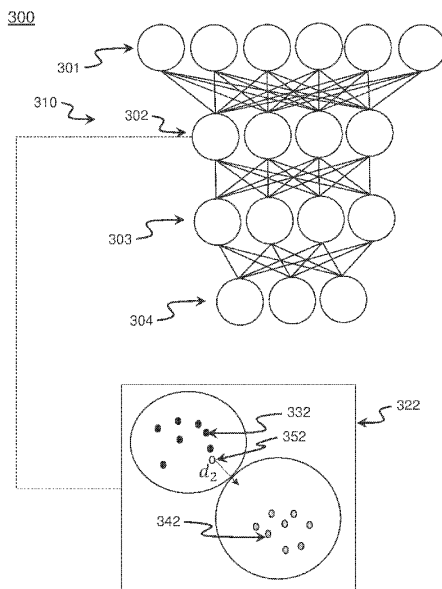


FIG. 3B

(57) Abstract: Methods, systems, and computer program products are provided for multi-layer analysis and detection of vulnerability of machine learning models to adversarial attacks. An example method includes selecting a sample, inputting the sample to a machine learning model, perturbing the sample to generate perturbed samples, inputting each respective perturbed sample to the machine learning model, and determining a smallest perturbation distance for the input layer. For each respective hidden layer of the machine learning model, a respective hidden layer representation of the sample is perturbed to generate perturbed hidden layer representations. Each respective perturbed hidden layer representation is inputted to the machine learning model, and a respective smallest perturbation distance for each respective hidden layer is determined. At least one vulnerability metric is determined based on the smallest perturbation distance for the input layer and the hidden layer(s). The machine learning model is retrained based on the vulnerability metric(s).

WO 2024/220790 A1

[Continued on next page]

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).

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))*