VI. IKERGAZTE

NAZIOARTEKO IKERKETA EUSKARAZ

2025eko maiatzaren 28, 29 eta 30a
Bilbo, Euskal Herria

ANTOLATZAILEA:
Udako Euskal Unibertsitatea (UEU)

INGENIARITZA ETA ARKITEKTURA

Backdoor erasoak spiking sare neuronaletan datu neuromorfikoekin

Gorka Abad eta *Olaia Inza*

35-42 or.
https://dx.doi.org/10.26876/ikergazte.vi.03.04

Backdoor erasoak spiking sare neuronaletan datu neuromorfikoekin

Gorka Abad[1,2], Olaia Inza[2]

[1] Radboud University
[2] IKERLAN Technology Research Center, Basque Research and Technology Alliance (BRTA).

gabad@ikerlan.es, oinza@ikerlan.es

Laburpena

Sare neuronalek errendimendu bikaina erakutsi dute hainbat atazetan, irudien eta hizketaren ezagutzan besteak beste. Hala ere, sare neuronalen eraginkortasunik handiena lortzeko, sareko parametro asko zehaztasunez optimizatu behar dira entrenamenduan zehar. Gainera, errendimendu handiko sareek parametro ugari dituzte, eta horiek energia asko kontsumitzen dute prestakuntzan. Erronka horiei aurre egiteko, ikertzaileek inpultsuzko sare neuronaletara jo dute, ingelesez: spiking neural network (SNN), energia-eraginkortasun hobeak eta biologikoki antzekoagoak direlako. Gainera, datu-prozesatze gaitasuna eskaintzen dute; horri esker, oso egokiak dira datu neuromorfikoetan. Hala ere, abantaila hauek izan arren, SNN-ek, saren neuronalek bezala, hainbat mehatxuren aurrean ahulak dira, besteak beste adibide maltzurren eta *atzeko atetiko*, ingelesez: backdoor, erasoen aurrean. Alabaina, eraso horiek ulertu eta haien aurka egiteko, SNN-en arloan oraindik ez da nahikoa ikertu. Lan honek SNN-etan datu neuromorfikoak erabiliz backdoor erasoak aztertzen ditu. Zehazki, datu neuromorfikoen backdoor *abiarazleak* nola kokatu eta koloreak nola manipula daitezkeen aztertzen dugu, irudi-domeinuko abiarazle ohikoek baino aukera zabalagoak eskainiz. Aurkezten ditugun eraso-estrategiek %100erainoko arrakasta lor dezakete.

Hitz gakoak: Atzeko atetiko erasoa, adimen artifiziala, datu neuromorfikoak

Abstract

Deep neural networks (DNNs) have demonstrated remarkable performance across various tasks, including image and speech recognition. However, maximizing the effectiveness of DNNs requires meticulous optimization of numerous hyperparameters and network parameters through training. Moreover, high-performance DNNs entail many parameters, which consume significant energy during training. To overcome these challenges, researchers have turned to spiking neural networks (SNNs), which offer enhanced energy efficiency and biologically plausible data processing capabilities, rendering them highly suitable for sensory data tasks, particularly in neuromorphic data. Despite their advantages, SNNs, like DNNs, are susceptible to various threats, including adversarial examples and backdoor attacks. Yet, the field of SNNs still needs to be explored in terms of understanding and countering these attacks. This paper delves into backdoor attacks in SNNs using neuromorphic datasets and diverse triggers. Specifically, we explore backdoor triggers within neuromorphic data that can manipulate their position and color, providing a broader scope of possibilities than conventional triggers in domains like images. We present various attack strategies, achieving an attack success rate of up to 100% while maintaining a negligible impact on clean accuracy.

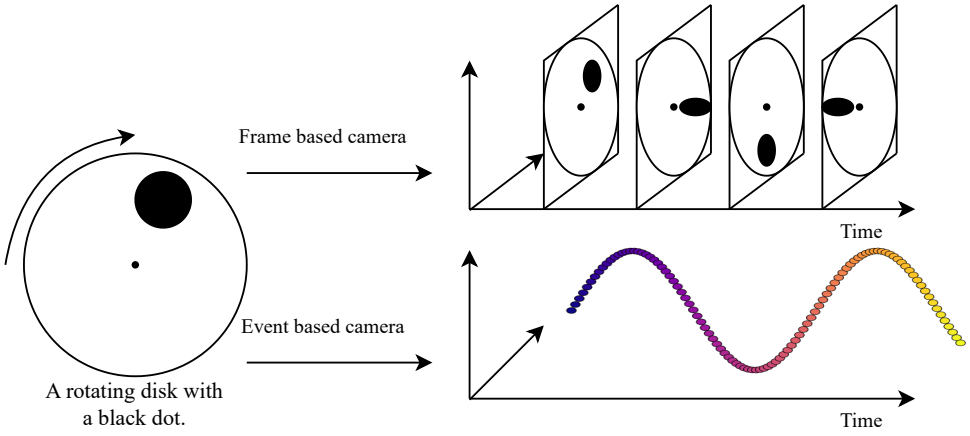
Keywords: *Backdoor attack, artificial intelligence, neuromorphic data*

1 Sarrera eta motibazioa

Sare neuronalek errendimendu bikaina lortu dute hainbat esparrutan, hala nola ikusmen artifizialean Krizhevsky *et al.* (2012), hizketa-ezagutzean Graves *et al.* (2013) eta hizkuntza naturalaren prozesamenduan Brown *et al.* (2020). Sare neuronalen arrakasta datu kopuru handietatik ikasteko eta patroi bilaketa-konplexuak aurkitzeko gaitasunetik dator. Sare neuronalek hainbat geruzaz sortuta daude non interkonektatutako neuronek sarea sortzen duten. Neuronen arteko konexio anitzen bidez sarean gaitasuna lortzen da eta entrenamenduan zehar, sarearen akatsak minimizatzeko eta modeloaren zehaztasuna hobetzeko, neuronen arteko loturak haztatu eta doitu egiten dira.

Sare neuronalen entrenamenduak *hiperparametroz* kondizionatuta daude, eta zeregin jakin batean errendimendu handiagoa lortzeko doitu daitezke (adib. irudi-tratamendua), baina hiperparametro horien optimizazioak konplexutasun handia du. Errendimendu egokia duen sare neuronala entrenatzeak denbora handia hartzen du eta energia kontsumoa garestia izan daiteke. Gainera, entrenamendu-datu (ingelesez *dataset*) eta sare neuronal handiak ditugunean parametro asko doitzea beharrezkoa denez, efektu hau gehiago areagotzen da. Adibidez, GPT-3 modeloaren entrenamendua 190 000 kWh kontsumoan estimatzen da Dhar (2020). Sare neuronalen konplexutasun eta eskakizun konputazionalen ondorioz, ikertzaileek ikuspegi alternatiboak aztertu dituzte, hala nola *spiking* sare neuronalak (SNN-ak) Ghosh Dastidar eta Adeli (2009); Tavanaei *et al.* (2019); Eshraghian *et al.* (2021); Fang *et al.* (2021).

SNN-ek energia kontsumoa nabarmenki jaitsi dezakete. Adibidez, Kundu *et al.* (2021) egindako ikerketan, SNN-ek sare neuronal arrunt batekin konparatuta, pareko errendimendua lortuz energia kontsumo 12 aldiz baxuagoa zutela baieztatu zuten. Energia efizientziaren gainean, SNN-ek beste abantaila asko dituzte. Esate baterako, perturbazioekiko eta zaratarekiko sendoak dira, egoera errealetan fidagarritasuna handiagotuz. Berez, *ebentuetan oinarritutako ikus-sentsore* (DVS) kamerak bildutako datu neuromorfikoak SNN-ak prozesatu ditzake. Zehazkiago, DVS kamerak, kamera arrunt batek bildutako distira absolutuaren ordez (hau da argazki arrunt bat bezala), era asinkrono batean biltzen ditu pixel bakoitzeko distira aldaketak. Kamera arruntekin alderatuz, DVS kamerek energia kontsumo baxuagoa dute eta latentzia baxuko datuak biltzen dituzte denboran zehar, hau da datu neuromorfikoak Chen *et al.* (2020); Liu *et al.* (2019), ikusi 1. Irudia.

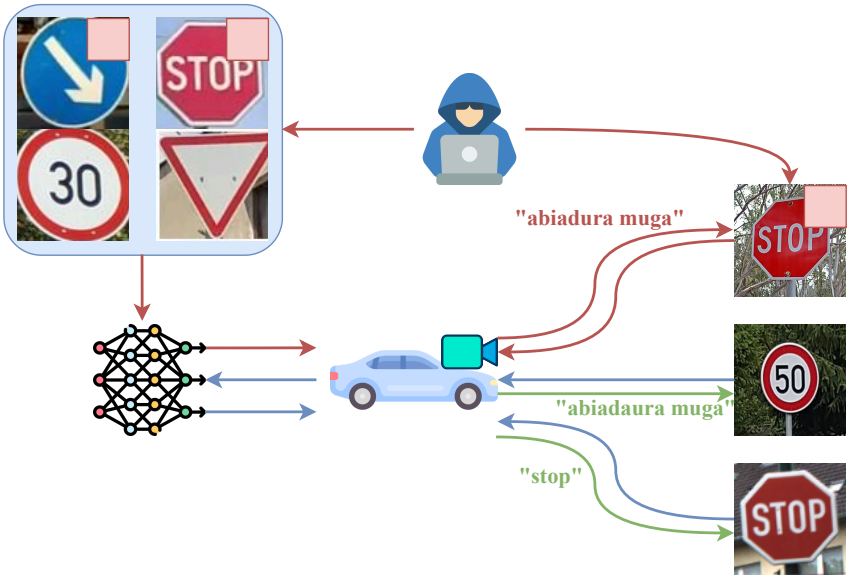


1. Irudia: Kamera arrunt eta DVS kameraren arteko konparazioa.

Sare neuronalen erabilpena aplikazio ezberdinetan ugaritzen ari da, ondorioz, modelo hauen hedatzeak dituen segurtasun eta pribatutasun inplikazioengatik, bai akademiaren bai industriaren kezkapenak areagotuz doaz. Esate baterako, sare neuronalek, eta hortaz SNN-ek, hainbat segurtasun eta pribatutasun erronka dituzte. Pribatutasunaren aldetik, eraso ezberdinen medio ikerketek esfortzu asko egin dituzte sare neuronalen datu pribatuak publiko ez bihurtzeko Dibbo (2023). Segurtasunaren arloan, bai adibide maltzurrek (*adbersariotiar adibideak*) Goodfellow *et al.* (2014) bai backdoor erasoek (*backdoor*) Gu *et al.* (2019) sare neuronalen funtzionamendua aldatzea dute helburu, haatik, teknikak ezberdina dira. Backdoor erasoetan zentratuko gara artikulu honetan. Backdoor-ak erasoak sare neuronalen entrenamendua eraldatzen dute, entrenamendu datuen azpimultzo txiki bat pozoitzen[1] *abiarazlea* datuetan gehituz, ikusi 2. Irudia. Backdoor-en helburua bikoitza da, alde batetik abiarazlea bertan ez denean zeregin originalean ondo funtzionatzea, eta bestaldetik, zeregin maltzurra injektatu ondoren, abiarazlea bertan denean, zeregin maltzurra aktibatzea. Backdoor erasoak hainbat eremutan gauzatzen dira: grafoei bideratutako sare neuronaletan Xu *et al.* (2021) eta hizketa-ezagutzean Koffas *et al.* (2021), besteak baterako.

Datu neuromorfiko eta SNN-en arloan segurtasuna eta pribatutasuna ez da oraindik sakonki ebaluatu. Datu neuromorfikoen natura dela eta, segurtasunak eta pribatutasunak testuinguru honetan hainbat erronka ditu. Artikulu honetan zehar, backdoor erasoak aztertuko ditugu SNN-en eta datu neuromorfikoen lerroan.

[1] Backdoor-ak injektatzeko beste era batzuk daude, baina ez gara horietan zentratuko orrien muga dela eta.



2. Irudia: Backdoor eraso baten adibide bat. Kasu honetan, erasotzaileak dataset-aren azpimultzo bat pozoitzen du, pegatina gorria erantsiz, trafiko seinaleak dituena. Ondorioz, auto autonomo batean erabiltzen den modeloa pozoituta dago. Autoa gidatzen den bitartean, trafiko-seinale desberdinak atzematen ditu eta adimen artifizialak zuzen sailkatzen ditu. Hala ere, erasotzaileak abiarazlea (pegatina) “stop” seinale baten gainean jartzen du, modeloa engainatuz “abiadura muga” gisa sailkatzeko.

2 Arloko egoera eta ikerketaren helburuak

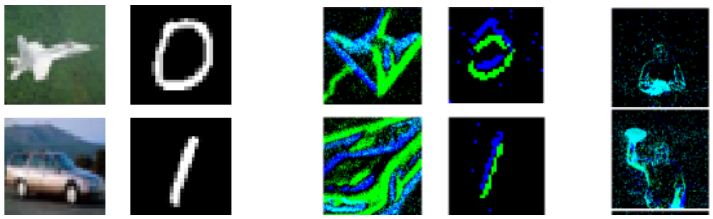
2.1 Backdoor erasoak

Entrenamendu-maltzurra teknikaren bidez, behin entrenatuta eta inferentzia garaian, backdoor erasoek modeloaren portaera aldatzen dute, modeloak portaera anormala erakuts dezan Gu *et al.* (2019). Backdoor-a duen modeloak sarrera garbiekin ondo jokatzen duen bitartean, abiarazlearekin sarrerak gaizki sailkatuko ditu. Datu-intoxikazio teknikaren bitartez, abiarazlea duten datu maltzurrak sartzen dira entrenamendu-multzoa aldatuz. Erasotzaileek modu ezberdinak dituzte backdoor-a modeloan sartzeko. Alde batetik, dataseta pozoituz eta ondoren interneten bidez publiko bihurtuz. Ondoren bezero batek dataseta deskargatuko du eta entrenamenduaren medio modeloa pozoituko da. Bestalde, bezero batek dataset bat dauka modelo bat entrenatzeko, baina ez dauka entrenatzeko gaitasunik. Ondorioz, hodeian entrenatzeko zerbitzu bat kontratatzen du, adibidez Amazon Web Service edo Google Cloud. Hauekin bai dataseta bai modeloa partekatzen du. Erasotzaileak, orain zerbitzu hauen kontrolpean dago, eta ondorioz entrenamenduaren kontrolpean ere. Erasotzaileak entrenamendua nahi duen moduan eraldatu dezake, adibide maltzurrak gehituz eta modeloa pozoituz. Irudien domeinuan adibidez, abiarazlea pixel-multzo bat izan daiteke, ikusi 3. Irudian agertzen den karratu zuria. Abiarazlea irudiaren hainbat lekutan jar dezakegu, erasoaren arrakasta aldatuz Abad *et al.* (2023). Modeloa datu garbiak eta backdoor-a duten datuen nahaste batean entrenatzen denean, pixel-patroia duten sarrerek, hau da abiarazlea duten irudiek, erasotzaileak aukeratutako helburu klase jakin batean gaizki sailkatzen ikasiko du. Erasotzaileak abiarazlearen ikusgarritasuna ezkutatzeko ahaleginak egin beharko ditu erasoaren detekzioa ekiditeko, ikusi 3. Irudia.

3. Irudia: Hainbat backdoor eraso moten adibideak.

2.2 SNN-ak eta datu neuromorfikoak

SNN-a, lehen azaldu bezala, datu neuromorfikoak erabiltzen dituen sare neuronal mota bat da [2]. Hauek *gif* antzeko bideo laburrak dirudite. Argazkien ordez, non *fotograma* bakarakoa da, data neuromorfikoak hainbat fotogramatan banatzen dira. SNN-ek neurona biologikoen jarduera nolabait imitatzea dute helburu. Horretarako, sarearen sarrera diskretu eta asinkronoak erabiliz sare neuronala modelatzen da. SNN-ek neuronen ordez *spiking* neuronak dituzte (neuronak hemendik aurrera). Sare neuronal arrunten kontrara, balio jarraituetan funtzionatzen dituztenak, SNN-ek gertaeretara bideratutako sarrera asinkronoekin funtzionatzen dute. Sare neuronal tradizionalen aldean, zeintzuek etengabe balioetsitako seinaleekin funtzionatzen duten, SNN-ak gertakarietan oinarritzen dira eta neurona-jardueraren denbora-dinamika txertatzen dute *leaky fire and integrate* (LIF) modeloa jarraituz Tavanaei *et al.* (2019).



(a) Originalak (b) Neuromorfikoak

4. Irudia: Datu arrunten eta datu neuromorfikoen arteko konparazioa.

Datu neuromorfikoen koloreak polaritatearekiko sortuak dira, ON eta OFF polaritateak esate baterako. Bi polaritate mota hauekin abiarazlearen diseinurako garrantzitsuak izango diren 4 kolore lor ditzakegu ($P0, P1, P2, P3$). SNN-en barrutiko funtzionamendu ezberdina eta datu neuromorfikoen erabilpena dela eta, backdoor erasoak sortzeak hainbat berezko erronka ditu.

- SNN-en entrenamendua sare neuronal arruntenarekin konparatuz ezberdina da. Beraz backdoor-a sartzeko teknikak ez dira erabilgarriak edota ez dute arrakasta handirik lortzen.
- Datu neuromorfikoak fotogramatan banatzen dira. Abiarazleak normalean estatikoak dira, hau da, fotograma bakarrean sartuta irudi arrunt bat bezala. Hala ere, datu neuromorfikoak abiarazle neuromorfikoen sorkuntza ahalbidetzen dute.
- Abiarazle neuromorfikoen diseinua bi eratan ezberdina da: i) denboran zehar hedatu behar da eta ii) polaritate aldakorrekoa izan daiteke.

3 Ikerketaren muina

Sekzio honetan zehar lehen ikertutako SNN-ei datu neuromorfikoekin egindako backdoor eraso motak azalduko ditugu Abad *et al.* (2024). Orokorrean bi diseinu ditu: estatikoa eta dinamikoa.

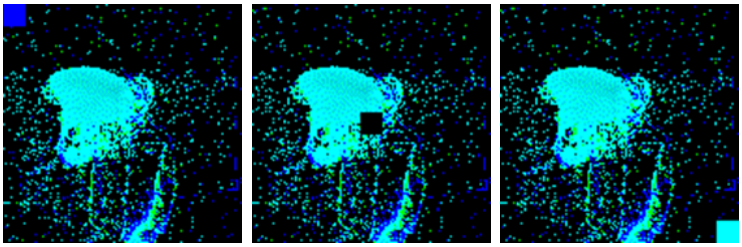
3.1 Erasoa estatikoaren diseinua

Irudien domeinuko backdoor erasoetan Gu *et al.* (2019), egileek irudietan erabilitako abiarazlea datu neuromorfikoetan errepikatu zuten Abad *et al.* (2022, 2024). Denboran zehar kodetutako datu neuromorfikoek dituzten abantailak erabat baztertuz, egileek abiarazle bera (posizio eta polaritate bera) sartu zuten fotograma guztietan, horrela abiarazle estatiko bat eginez. Abiarazle mota hori aztertuko dugu ondorengoen oinarri gisa, eta hainbat kasutan sakonki ikertuko dugu abiarazlearen kokapena, polaritatea eta tamaina. Ikusi 5. Irudia adibide gisa.

3.2 Erasoa dinamikoaren diseinua

Orain abiarazle dinamikoa aurkeztuko dugu, ezkutukoa, ikusezina eta dinamikoa. Zehatzago, irudiaren domeinuko backdoor dinamikoek eraginda Nguyen eta Tran (2020); Doan *et al.* (2021), abiarazle neuromorfikoak eta dinamikoak ikertzen ditugu, non abiarazleak ikusezinak diren, fotograma bakoitzarentzat ezberdina, hau da, fotograma batetik bestera abiarazlea aldatu egiten da. Kontuan izan lagin eta fotograma bakoitzeko forma eta kolorea

[2] Ikusi hurrengo esteka datu neuromorfikoen adibideak ikusteko `https://github.com/GorkaAbad/Sneaky-Spikes`.

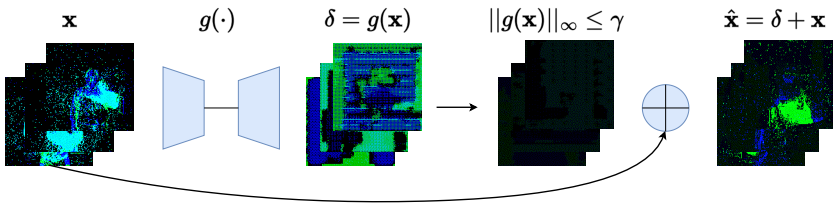


(a) Goi (b) Erdi (c) Behe

5. Irudia: Eraso estatikoaren adibideak.

txandakatzen dituen abiarazlea ez dela aurretik ikertu literaturan Abad *et al.* (2024). Datu neuromorfikoek aukera ematen digute fotograma bakoitzeko bakarra den lagin bati dagokion abiarazle dinamiko bat sortzeko. Hori lortzeko, *spiking autokodetzaile* bat erabiltzen dugu, egituraz *autokodetzaile* estandar bezalakoa dena. Autokodeatzaileak irteera irudia originala bezain handia den irudi bat sortzen du, hau da abiarazlea. Autokodeatzaileak hiru helbururekin abiarazlea optimizaten du: backdoor-aren errendimendua maximizatu, errendimendu garbia mantendu eta ikusezina bihurtu. Zehazkiago, lehen azaldutako eraso estatikoaren punturik ahulenetako bat abiarazlearen kokalekuaren eta polaritatearen menpe, gizakien ikuskaritzapean detektagarriak direla da.

Intuizioz, abiarazlea dinamikoa (ikusi 6. Irudia) honela sortzen da. Lehenik, perturbazioa sortzen dugu irudi garbi bat spiking autokodetzaile batetik pasatuz: $g(\cdot) : \delta = g(\mathbf{x})$. Ondoren, "backdoor" irudia eraikitzeko, perturbazio hori irudi garbiari gehitzen diogu $\hat{\mathbf{x}} = \mathbf{x} + \delta$. Hala ere, diseinu sinple honek $\mathbf{x}$ δ-rekin saturatuko luke, abiarazlea ikusgarri bihurtuz. Hori saihesteko, perturbazioa l_p-espazio batera proiektatzen dugu[3], hiperparametro jakin bat γ ezarrita: $\|g(\mathbf{x})\|_\infty \leq \gamma$. Ondoren, $g(\cdot)$ eguneratu egiten da, $f(\cdot)$-en "backdoor" zehaztasuna (hots, erasotutako irudiak behar bezala sailkatzeko gaitasuna) maximizatzeko helburuarekin.



6. Irudia: Abiarazle dinamikoaren diagrama.

3.3 Esperimentazioa

Bi eraso mota aurkeztu ondoren, hauen erabilgarritasuna hurrengo ikuspuntutatik ebaluatuko dugu: i) Zeregin garbian lortutako arrakasta *(accuracy)*, ehunekoetan adierazita. ii) Erasoaren arrakasta neurtzeko *arrakasta tasa (ASR)* metrika erabiliz; hots, erasotako laginen artetik zenbat sailkatu diren erasotzaileak hautatutako klase gisa, ehunekotan adierazita.

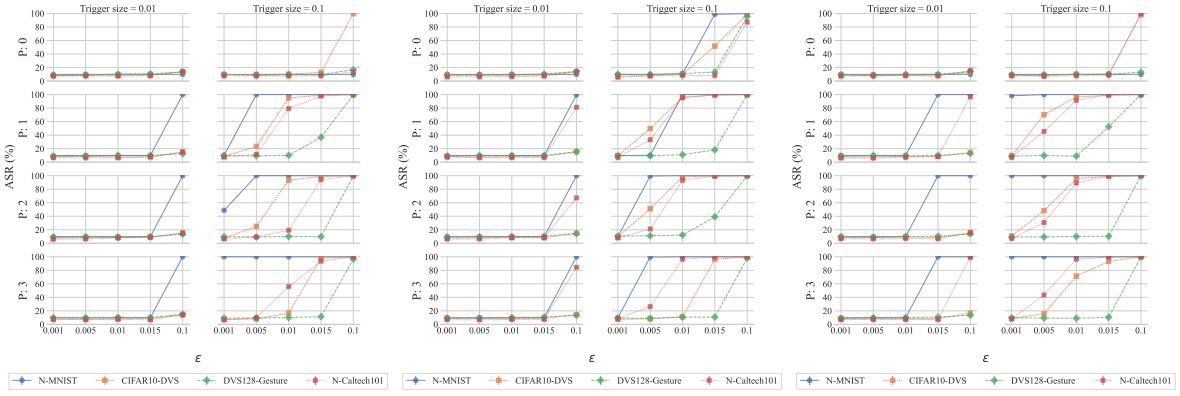
Lau datu-multzo erabiltzen ditugu: N-MNIST Orchard *et al.* (2015), CIFAR10-DVS Li *et al.* (2017), DVS128-Gesture Amir *et al.* (2017), eta N-Caltech101 Orchard *et al.* (2015). N-MNIST, CIFAR10-DVS eta N-Caltech101, haien bertsio ez-neuromorfikoak adimen artifizialaren arloko segurtasun/pribatutasun helburuetarako ikusmen konputazionalean ohiko erreferentziazko datu-multzoak direlako. Dataseten %80a entrenatzeko eta %20 ebaluatzeko izango dira. Hiru sare-arkitektura erabiliko ditugu, lotutako lanetan Fang *et al.* (2021) proposatutakoaren arabera.

Eraso estatikoa ϵ balio hauekin probatzen dugu: 0,001, 0,005, 0,01, 0,05 eta 0,1. ϵ-ek abiarazlea daukaten laginen ratioa kontrolatzen du. Hau da 100-eko lagina badaukagu, 10 irudi trigger-arekin egongo dira eta 100 garbi $\epsilon = 0, 1$ denean. Abiarazlearen tamaina sarrera-irudiaren tamainaren %1 eta %10 izateko ezartzen dugu. Azkenik, abiarazlea hiru kokapenetan (beheko eskuinean, erdian eta goiko ezkerrean) eta lau polaritatetan probatzen dugu.

[3] l_p-espazio bat forma geometriko bat da; n-dimentsioko espazioan, l_p-normaren arabera puntu jakin batetik distantzia zehatz bateko tartean dauden puntu guztiak biltzen ditu.

Entrenamendurako, *ikasteko tasa (ingelesez: learning rate)* balio lehenetsi bat (0,001) ezarri dugu, MSE galera funtzio gisa, *ADAM* optimizatzaile gisa, eta datu-multzo neuromorfikoak $T = 16$ fotogramatan zatitu ditugu *SpikingJelly* Fang *et al.* (2020) tresna erabiliz. N-MNIST datu-multzoan, (garbia) %99-ko zehaztazuna lortu dugu, 10 epokatan. CIFAR10-DVS kasuan, %68-ko zehaztazuna eskuratu dugu 28 epoken ondoren. N-Caltech101 datu-multzoan, %76-ko zehaztazuna lortu dugu 30 epoken ondoren. Azkenik, DVS128-Gesture datu-multzoan, 64 epoka igaro ostean %93-ko zehaztazuna lortu du.

Gure emaitzek erakusten dute (ikusi 7. Irudia) static backdoor-ek abiarazlearen tamaina sarreraren %10 bezala izan behar dutela, CIFAR10-DVS edo DVS128-Gesture bezalako datu-multzo konplexuetan backdoor portaera txertatzeko. Abiarazlea sarrera tamainaren %1 denean, backdoor-a soilik arrakastaz txertatzen da N-MNIST-en, polarity-a 0 ez bada. Hala ere, abiarazla erdian dagoenean, ikusi dugu polarity-a 0 denan %100 ASR lortzen duela. Efektu hau datuak zentratuta daudelako gertatzen da; horrela, abiarazlea beltza irudiaren gainean dago, kontraste handia sortuz eta sareari abiarazlea bereizteko aukera emanez.



(a) Behe-eskuma (b) Erdia (c) Goi-ezkerra

7. Irudia: Trigger estatikoen ASR-ak.

Ikerketa honetan, backdoor dinamikoa erabiliz esperimentuak egin ditugu, α eta γ balioen eragina aztertzeko, hurrenez hurren, backdoor errendimendu garbiaren oreka eta abiarazlearen intentsitatea. α balioa 0,5-0,9 tartean aldatuz, eta γ balioa 0,01-0,1 tartean, $\alpha < 0,5$ baztertuz benetako eszenatokietan backdoor-ari zeregin garbiari baino pisu handiagoa ematea ezinezkoa delako. Gure helburua, aldi berean zeregin garbiaren arrakasta maximizatzea eta arrakasta-tasa (ASR) handia lortzea da.

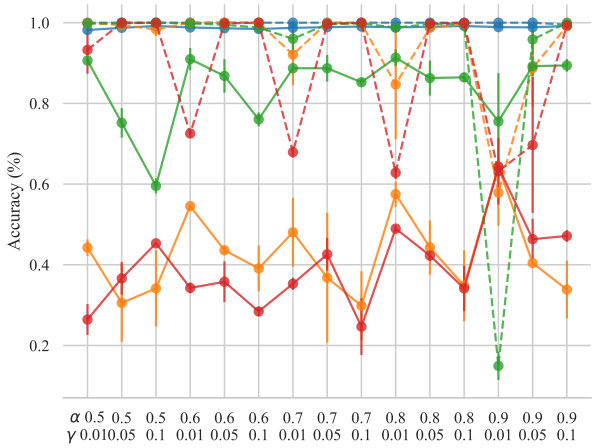
Gure esperimentuek erakusten dute ASR ia beti %100-ra hurbiltzen dela egoera gehienetan, ikusi 8. Irudia. Hala ere, zehaztazun garbiaren degradazioa eta abiarazlearen ikustasunaren arteko oreka mantentzeko erasotzaileak kontu handiz hautatu behar ditu α eta γ balio egokiak. Bereziki, abiarazlearen ikusgarritasuna γ-k kontrolatzen du eta $\gamma = 0,01$ denean, irudi hori irudi garbitik bereiztea ia ezinezkoa bihurtzen da, ikusi 9. Irudia. Esperimentuek iradokitzen dute γ txikiak eta α handiak erabiltzeak emaitza onenak ematen dituela.

4 Ikerketaren ondorioak

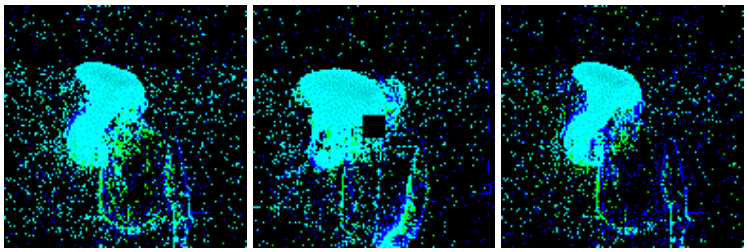
Ikerketa honek backdoor erasoen testuinguruan SNN-en segurtasuna aztertu du. Bizi-biziko teknologia berri gisa SNN-ek gero eta garrantzi handiagoa badute ere, arlo honek arreta mugatua jaso du. Gure ikerketak SNN-etan datu eta abiarazle neuromorfikoak erabiltzen ditu backdoor erasoak gauzatzeko. Zenbait eraso-metodo proposatu ditugu, haien artean abiarazle dinamikoa sortzen duen eraso berri bat, denboran zehar eboluzionatzen dena eta giza begiarentzat igarriezina dena. Gure emaitzek erakusten dute erasoek %100 arteko arrakasta-tasa lor dezaketela zeregin garbian degradazio nabarmenik eragin gabe. Ondorioek ikusarazi dute backdoor erasoen aurrean SNN-ak oso ahulak direla, defentsa-mekanismoek nola defendatzen dituzten backdoor erasoak arlo honetan ebaluatzea derrigorrezkoa da.

5 Etorkizunerako planteatutako norabidea

Etorkizuneko ikerketek arreta jarri behar dute SNN garatzerakoan, backdoor erasoei aurre egiteko diseinatuta dauden datu neuromorfikoak dakartzan erronka bereziei ardura jarriz. Gainera, azpimarratzekoa da SNN-ek gero eta



8. Irudia: Abiarazle dinamikoaren ASR eta arrakasta garbiaren degradazioa. Lerro etenak ASR-a adierazten dute, eta lerro jarraiak arrakasta. Lerro urdinak N-MNIST-i dagokie, laranjak CIFAR10-DVS-i, berdeak DVS128-Gesture-i, eta gorriak N-Caltech101-i. α-k backdooraren intentsitatea kontrolatzen duen hiperparametro bat da.



(a) Garbia (b) Trigger estatikoa (c) Trigger dinamikoa

9. Irudia: Trigger-en arteko konparazioa.

garrantzia handiagoa dutela beste arlo batzuetan, hala nola grafoen arloan, non backdoor erasoek gero eta pisu handiagoa hartzen ari diren. Azkenik, gure ikerketak gehienbat backdoor erasoei erreparatu badie ere, garrantzitsua da SNN-etara egokitu daitezkeen beste mehatxu arrunt batzuk kontuan hartzea, besteak beste inferentzia erasoak eta adibide maltzurrak.

Erreferentziak

Abad, Gorka, Oguzhan Ersoy, Stjepan Picek, Víctor Julio Ramírez-Durán, eta Aitor Urbieta. 2022. Poster: Backdoor attacks on spiking nns and neuromorphic datasets. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 3315–3317.

——, Oguzhan Ersoy, Stjepan Picek, eta Aitor. Urbieta. 2024. Sneaky spikes: Uncovering stealthy backdoor attacks in spiking neural networks with neuromorphic data. In *Network and Distributed System Security (NDSS) Symposium*.

——, Jing Xu, Stefanos Koffas, Behrad Tajalli, Stjepan Picek, eta Mauro Conti. 2023. Sok: A systematic evaluation of backdoor trigger characteristics in image classification. *arXiv preprint arXiv:2302.01740* .

Amir, Arnon, Brian Taba, David Berg, Timothy Melano, Jeffrey McKinstry, Carmelo DiÑolfo, Tapan Nayak, Alexander Andreopoulos, Guillaume Garreau, Marcela Mendoza, eta others. 2017. A low power, fully event-based gesture recognition system. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7243–7252.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, eta others. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33.1877–1901.

Chen, Guang, Hu Cao, Jorg Conradt, Huajin Tang, Florian Rohrbein, eta Alois Knoll. 2020. Event-based neuromorphic vision for autonomous driving: A paradigm shift for bio-inspired visual sensing and perception. *IEEE Signal Processing Magazine* 37.34–49.

Dhar, Payal. 2020. The carbon impact of artificial intelligence. *Nat. Mach. Intell.* 2.423–425.

Dibbo, Sayanton V. 2023. Sok: Model inversion attack landscape: Taxonomy, challenges, and future roadmap. In *2023 IEEE 36th Computer Security Foundations Symposium (CSF)*, 439–456. IEEE.

Doan, Khoa, Yingjie Lao, Weijie Zhao, eta Ping Li. 2021. Lira: Learnable, imperceptible and robust backdoor attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11966–11976.

Eshraghian, Jason K, Max Ward, Emre Neftci, Xinxin Wang, Gregor Lenz, Girish Dwivedi, Mohammed Bennamoun, Doo Seok Jeong, eta Wei D Lu. 2021. Training spiking neural networks using lessons from deep learning. *arXiv preprint arXiv:2109.12894* .

Fang, Wei, Yanqi Chen, Jianhao Ding, Ding Chen, Zhaofei Yu, Huihui Zhou, Yonghong Tian, eta other contributors, 2020. Spikingjelly. `https://github.com/fangwei123456/spikingjelly`. Accessed: 2022-10-12.

——, Zhaofei Yu, Yanqi Chen, Timothée Masquelier, Tiejun Huang, eta Yonghong Tian. 2021. Incorporating learnable membrane time constant to enhance learning of spiking neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2661–2671.

Ghosh Dastidar, Samanwoy, eta Hojjat Adeli. 2009. Spiking neural networks. *International journal of neural systems* 19.295–308.

Goodfellow, Ian J, Jonathon Shlens, eta Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* .

Graves, Alex, Abdel-rahman Mohamed, eta Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, 6645–6649. Ieee.

Gu, Tianyu, Kang Liu, Brendan Dolan-Gavitt, eta Siddharth Garg. 2019. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access* 7.47230–47244.

Koffas, Stefanos, Jing Xu, Mauro Conti, eta Stjepan Picek. 2021. Can you hear it? backdoor attacks via ultrasonic triggers. *arXiv preprint arXiv:2107.14569* .

Krizhevsky, Alex, Ilya Sutskever, eta Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25.

Kundu, Souvik, Gourav Datta, Massoud Pedram, eta Peter A Beerel. 2021. Spike-thrift: Towards energy-efficient deep spiking neural networks by limiting spiking activity via attention-guided compression. In *Proceedings of the IEEE/CVF WACV*, 3953–3962.

Li, Hongmin, Hanchao Liu, Xiangyang Ji, Guoqi Li, eta Luping Shi. 2017. Cifar10-dvs: an event-stream dataset for object classification. *Frontiers in neuroscience* 11.309.

Liu, Shih-Chii, Bodo Rueckauer, Enea Ceolini, Adrian Huber, eta Tobi Delbruck. 2019. Event-driven sensing for efficient perception: Vision and audition algorithms. *IEEE Signal Processing Magazine* 36.29–37.

Nguyen, Tuan Anh, eta Anh Tran. 2020. Input-aware dynamic backdoor attack. *Advances in Neural Information Processing Systems* 33.3454–3464.

Orchard, Garrick, Ajinkya Jayawant, Gregory K Cohen, eta Nitish Thakor. 2015. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in neuroscience* 9.437.

Tavanaei, Amirhossein, Masoud Ghodrati, Saeed Reza Kheradpisheh, Timothée Masquelier, eta Anthony Maida. 2019. Deep learning in spiking neural networks. *Neural networks* 111.47–63.

Xu, Jing, Minhui Xue, eta Stjepan Picek. 2021. Explainability-based backdoor attacks against graph neural networks. In *Proceedings of the 3rd ACM Workshop on Wireless Security and Machine Learning*, 31–36.

6 Eskerrak eta oharrak

Lan hau *Network and Distributed System Security (NDSS) Symposium* kongresuan aurkeztutako lanean oinarrituta dago. Ikusi Abad *et al.* (2024).