Pixel-erasoak etorkizuneko terrorismo

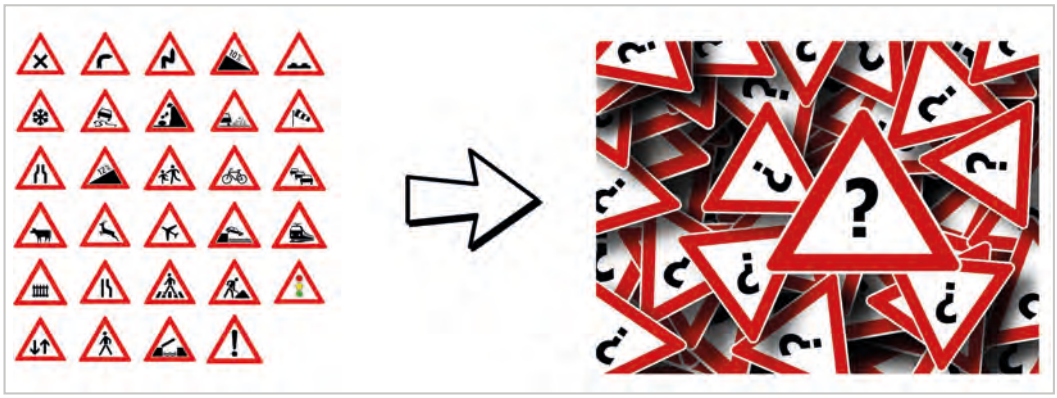
Ilargi bete bakoitzarekin, adimen artifiziala gizakion parte geroz eta garrantzitsuagoa bilakatzen da. Ez dago zalantzarik adimen artifizialak helburu sinesgaitzak bereganatu dituela azken urteetan; aitzitik, sistema adimendun hauek badituzte gizartean plazaratzeko interesa suspertzen ez duten alderdi ilun batzuk ere. Artikulu honetan, irudiak prozesatzeko sistemen ahulezia batez mintzatuko natzaizue, pixel-erasoei buruz, horren bidez sare neuronaletan oinarritutako sistema adimendunen alderdi ez hain positibo bat ezagutzeko, eta irudiak prozesatzeko sistema horien arriskuak ulertzeko. Ziur asko, artikulu hau irakurri ondoren ez duzue zuen auto autonomoetan berriz lokartu nahiko.

Funtsean, pixel-erasoa sare neuronalen ohiko entrenamendua atzekoz aurrera jartzean datza. Egoera arruntetan, sare neuronalen sistema bat entrenatzeko irudi-sorta izugarri bat erabiltzen da: irudiak sistemari erakutsi, eta irudian zer dagoen interpretatzen ikasten du hark. Ikasketa-prozesu horretan, sare neuronalak egindako iragarpen okerrekin sistemaren parametroak doitzen dira, eta, berrelikadura horri esker, sistemaren emaitzak etengabe hobetzen dira trebatzen ari den denbora horretan.

Auto autonomoak gidatze-esperientzia sinplifikatzea dakar. ARG.: Domeinu publikoan.

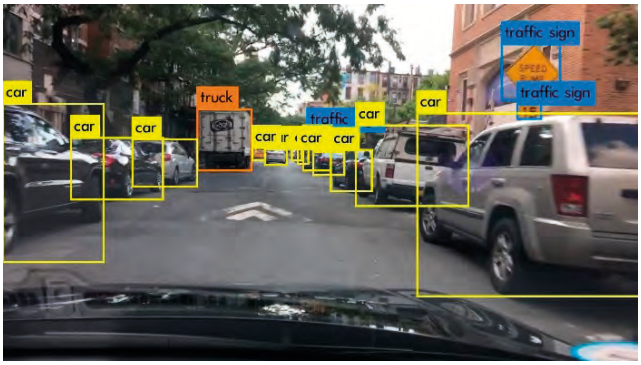
Hasiera batean hutsegite anitz egiten badituzte ere, nahikoa denbora ikasten aritu ondoren, sistemak asko hobetzen dira, guztiz doituta egon arte. Egun, irudiak ezagutzeko sistemek oso emaitza onak ematen dituzte, oro har; kasu askotan, irudien ia % 100 ezagutzeko gai dira, tartean objektu arrarorik ez bada [1]. Sare neuronala gehiago doitu ezin denean, entrenamendua eteten da, sistema ebaluatu egiten da, eta produkziora eramaten da. Guztiz doituta dauden ezagutza-sistema

Iñigo López-Gazpio
Adimen artifizialeko ikertzailea
Deustuko Unibertsitatea



Sare neuronalen sistema baten ohiko entrenamendua. Sistemari irudi-sorta bat erakusten zaio, irudien itxurak eta ezaugarriak ezagutzen irakasteko. Sistemak egindako akatsak erabiltzen dira hura moldatzeko eta etorkizunean akats berak ez errepikatzeko. ARG.: Iñigo López-Gazpio.

adimendun horiek, besteak beste, auto autonomoen nabigazio-sistemetan txertatzen dira, autoari ikusmena ahalbidetzeko. Hurrengo irudian, auto autonomo baten ezagutza-sistemaren irudi bat aztertu daiteke. Irudian ikusten den modura, sistema horiei esker, autoa gai da inguruan dituen objektuak, pertsonak, animaliak eta bestelakoak identifikatzeko.

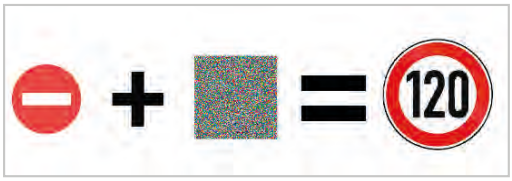


Guztiz doitutako erazagupen sistema bati esker, auto autonomoa gai da inguruko objektuen pertzepzioa izateko. Iturria [2, Choi et al.].

Informazio hori guztia sarrera gisa erabilita, nabigazioa ahalik eta hobekien eta seguruen egiteko erabakiak hartzen dituzte auto autonomoek: arriskuak badira, abiadura mantsotu edo geratu; errepidea libre badago eta abiadura mugak gainditzen ez badira, bizkortu; eskuineko erraila libre badago, alboratu; txirrindulariak aurreratzeko, mantsotu eta alboratu; etab.

Aitzitik, sare neuronala doitzeko prozesua hankaz gora jartzen bada...

Sare neuronalei iruzur egiteko aukera dago, sarrerako irudiak aldatuta. Hau da, sare neuronala entrenatzeko prozesua atzekoz aurrera jartzen bada sareari gezurrezko irudi bat eta helburu bat erakutsiz, jatorrizko irudian egin beharreko perturbazioa eskuratzeko aukera dago, sareari iruzur egin eta benetan beste irudi bat erakutsi zaiola sinetsarazteko. Adimen artifizialari egin dakiokeen eraso-mota hori oso arriskutsua da, abiadura mantsotzeko edo geratzeko segurtasun-seinaleak abiadura bizkortzeko seinale bihur daitezkeelako, esate baterako. Jarraian dugun irudian, pixel-erasoaren adibide sinple bat ikus dezakegu. Ikus daitekeenez, zirkulazioa debekatzeko seinale bati zarata gehituz gero, objektuak ezagutzeko sistema adimendun batentzat beste seinale bat bihurtzen da, nahiz eta gizakiok ezin dugun aldaketa hori begi hutsez ikusi.



Jatorrizko irudi bati perturbazioa gehituz gero, sare neuronala bati ziria sartzeko aukera dago, eta benetan beste irudi bat ikusi duela pentsarazteko. Horri kontrako irudiaren erasoa deritzo (adversarial attack), eta ikerketa-lerro garrantzitsua da egun. ARG.: Iñigo López-Gazpio.

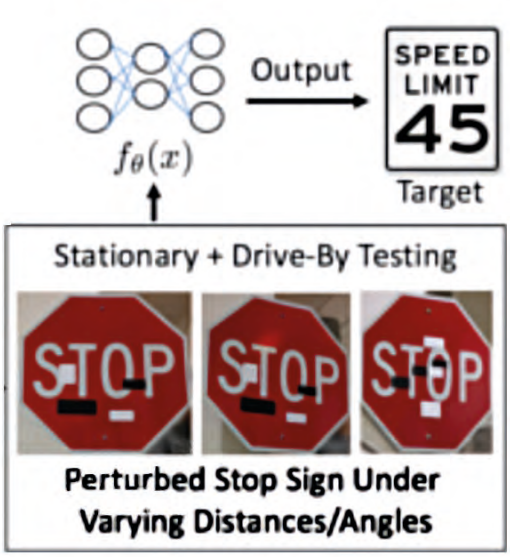
Fenomeno horri kontrako irudiaren erasoa deritzo (adversarial attack, ingelesez), eta ikerketa-lerro esanguratsua da gaur egun auto autonomoaren segurtasunaren inguruan ikertzen ari direnentzat. Irudiak moldatzeko eta sistemei iruzur egiteko teknika horrek ikerketa-lerro interesgarri bat irekitzen du sistema adimendunen konfiantzaren eta ebaluazioaren inguruan.

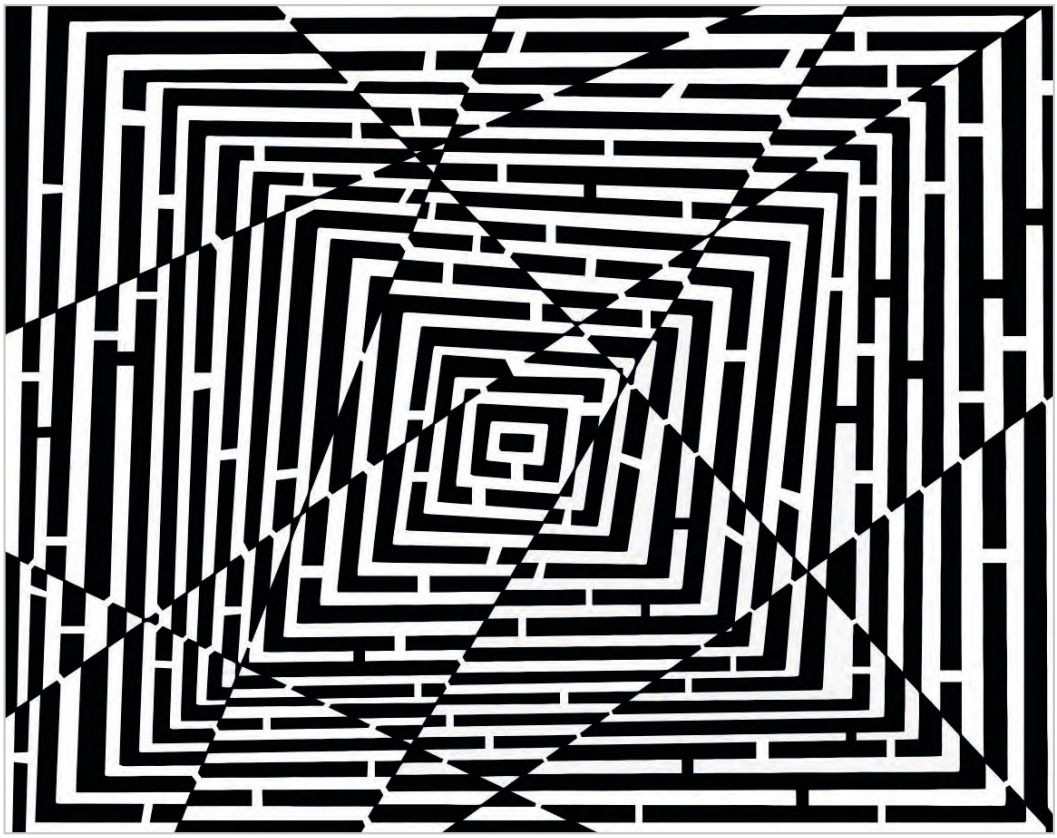
Horrenbestez, interesgarria izan daiteke ikertzea zenbateko perturbazioa gehitu behar den irudietan sistema adimendun bati iruzurra egiteko. Gai honen inguruan idatzi diren ikerketa guztien artean beldurgarriena Kevinek eta laguntzaileek Samsungeko ikerketa-institutuan egindako azterketa da. Artikulu horretan azpimarratzen da sare neuronal bati iruzur egiteko gehitu behar den perturbazio-maila oso txikia dela [3], eta hori arazo larria da etorkizuneko auto autonomoen erabiltzaileentzat. Ikerketa-lan horretan, egileek garbi adierazten dute trafiko-seinale batean grafiti baten moduko aldaketa txiki bat egitea nahikoa dela auto autonomoaren ezagutza-sisteman sekulako iruzurra egiteko. Aski da zinta zuri bat eta beltz bat erabiltzea stop seinale bat abiadura-mugako seinale bihurtzeko. Beldurgarria da. Hurrengo irudian, egileek artikuluan aipatzen duten kasu zehatz hori deskribatzen da, eskema moduan.

Zer egin daiteke sare neuronalen sistemak horrelako erasoen aurka babesteko?

Azken ikerketek erakutsi dutenez, aurkako irudietan oinarritutako erasoak ez daude sare neuronalen sistemen menpe, haiek doitzeko erabili diren datu-multzoen menpe baizik. Beste era batera esanda,

Nahikoa da perturbazio hori gehitzea stop seinale bat orduko 45 km-ko abiadurako seinale bihurtzeko. Iturria [3, Eykholt et al].





Giza burmuina engainatzeko gai den ilusio optikoa gizakion pixel-erasoaren adibide da. Irudia: Domeinu publikoan.

datu-multzoaren berezko ezaugarri bat dira. Horrek esan nahi du sare neuronalen arkitektura bati iruzur egiteko baliagarri diren lagin moldatuak baliagarriak direla beste arkitektura bati ere iruzur egiteko, betiere doitzeko datu-multzo bera partekatu izan badute doitze-fasean. Datu-multzo izugarriak sortzea prozesu garestia eta konplexua denez, oso ohikoa da sare neuronalen sistema asko eta asko datu-multzo berdinekin doituta egotea. Horrek esan nahi du aurkako irudiaren eraso-teknikaren eragina oso larria izan daitekeela, eta eraso-mota horretatik babesteko moduak ikertu behar direla.

Aurkako ikasketa dugu babes-metodorik ezagunena. Nahiko sinplea da, gainera, nahiz eta babestuko gaituen ziurtasun guztizkorik ez daukagun. Teknika horrekin, sare neuronal sendo eta iruzurren aurkako bat eraikitzeko, datu-multzoa aurkako adibide askorekin osatzen da. Horri esker, ezaugarri hauskorrak edo ahulak alde batera uzten ditu ereduak ikasketa-prozesuan, eta iragarpenak egiteko ezaugarri sendoagoetan oinarritzen ikasten du. Teknika horrek arrakasta izan dezan, aurkako adibide maltzur egokiak era masiboan sortu behar dira. Baina alde txar bat ere badu horrek: sare neuronal baten doitze-fasea 3-30 aldiz mantsotu daiteke, datu-multzoa masiboki handitzen baita mota horretako irudiekin.

Gaur egun, datu-multzoak irudi maltzurrekin osatzeko tresnak dituzte ikertzaileek esku artean; FoolBox, adibidez. Tresna horrekin, automatikoki sor daitezke irudi maltzurrak, eta, hala, gure sistema adimendunak jakin dezake mota horretako maltzurkeriak badirela. Hala eta guztiz ere, badirudi erasotzaileen eta defendatzaileen arteko guda

bilakatzen ari dela hau, nork teknologia berriagoa diseinatu aurkaria menderatzeko.

Hau guztia adimen artifizialaren errua ote?

Ikusi dugun modura, aurkako irudien eraso-teknikak arazo oso larriak sor ditzake segurtasuna hain garrantzitsua den egoeretan, eta puntako sare neuronal berrienak ere tronpatu ditzake. Sare neuronalak ezaugarri ahuletan oinarritzen direlako gertatzen da hori, eta irudia guztiz ulertzen edo ondo aztertzen ez dutelako. Baina arazo bera gertatzen zaigu gizakioi ere, gure burmuin alferrontziak antzeko trikimailuak egiten baitizkigu aurkako irudiekin eraso egiten diogunean; adibidez, ilusio optikoko irudiak ikusten ditugunean.

Mota honetako ilusio optikoei erreparatuz gero, hasiera batean badirudi lerroak ez direla paraleloak; aitzitik, gertutik behatzean, lerro hauek bata bestearekiko paraleloak direla ohartzen gara. Guk bezala, sare neuronalen sistemek ere atentzio-puntu hau behar dute aurkako irudiek inposatu nahi dituzten trikimailuez ohartzeko. Izan ere, aurkako irudiak eta pixel-erasoak benetan ez dauden gauzak ikustera behartzen gaituzten irudiak baino ez dira.

Hurrengo urteetan, eraso- eta defentsa-sistema berrien garapena katuaren eta saguaren etengabeko jokoa izango da. Azken finean, horrek eredu sendoagoak eta fidagarriagoak ekarriko ditu, eta pauso garrantzitsua izango da segurtasun-aplikazio kritikoetan erabiltzeko, hala nola auto autonomoetarako. Hala eta guztiz ere, oraingoz hobe eskuak bolantetik gehiegi ez urruntzea, badaezpada. ●

Erreferentziak

[1] Janai, J., Güney, F., Behl, A., & Geiger, A. (2020). Computer vision for autonomous vehicles: Problems, datasets and state of the art. Foundations and Trends® in Computer Graphics and Vision, 12(1–3), 1-308.

[2] Choi, J., Chun, D., Kim, H., & Lee, H. J. (2019). Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 502-511).

[3] Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., ... & Song, D. (2018). Robust physical-world attacks on deep learning visual classification. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1625-1634).

2022-23 IKASTURTEKO hik hasiREN FORMAZIOA

AZKEN PLAZAK

LEHEN HEZKUNTZA ERALDATZEN

Edukiak:
Tutore gogoetatsuaren rola, 6-12 adin tarteko haurren ezaugarriak, hezkuntza proiektuaren prozesua, espazioaren antolaketa, ebaluazio hezigarria, pedagogia sistemikoa...
Koordinatzaile pedagogikoa:
Aritz Larreta
Formazio pertsonaleko koordinatzailea:
Alvaro Beñaran

JIRAFAREN HIZKUNTZA. KOMUNIKAZIO EZ-BORTITZA

Edukiak:
Adierazpen zintzoa, entzute enpatikoa, errurik gabeko guneak sortzeko trebetasuna; gatazken kudeaketa; amorruaren kudeaketarako trebetasuna; mugak jartzeko trebetasuna; ikasleen autoestimua elikatzeko trebetasuna...
Formatzailea:
Nerea Mendizabal

PSIKOMOTRIZITATE ERLAZIONALA

Edukiak:
Gorputzaren bidezko komunikazioak eskaintzen duen errespetuzko entzute, itxarote eta laguntzeko trebetasunak eskuratzen hastea; haurren jolas librea behatzea...
Formatzailea:
Mapi Urresti

Informazioa eta izen-ematea: www.hikhasi.eus