УНИВЕРЗИТЕТ У БЕОГРАДУ

ФАКУЛТЕТ БЕЗБЕДНОСТИ

КАТЕДРА СТУДИЈА БЕЗБЕДНОСТИ

СУПАРНИЧКО МАШИНСКО УЧЕЊЕ

– ДИПЛОМСКИ РАД –

Ментор:
Ана Ковачевић
Проф. др

Студент:
Дарко Николчић
343/18

Београд, 2024.

САДРЖАЈ

1. Увод

Машинско учење је подобласт рачунарске науке која је усмерена да омогући вештачкој интелигенцији да учи одређене задатке онако како то раде људи, тј. да опонаша људско учење. Машинско учење доприноси побољшању технологије, попут претраживача, паметних кућних уређаја и онлајн услуга.

Како модели машинског учења настављају да се развијају и налазе широку примену у различитим секторима, рањивост ових модела на супарничке нападе постала је горућа брига. Супарнички напади укључују злонамерну манипулацију улазним подацима да би се намерно довели у заблуду модели машинског учења, што доводи до погрешних предвиђања или класификација. Потенцијалне последице таквих напада крећу се од угрожене безбедности у критичним системима до губитка поверења у технологије машинског учења.

Срж овог рада биће анализа различитих метода напада супарничког учења, од којих свака представља карактеристичне приступе компромитовању модела машинског учења. Затим анализа се проширује на испитивање механизама одбране. Разумевање како ојачати моделе машинског учења против таквих напада је кључно за обезбеђивање њихове робусности у апликацијама које се користе свакодневно.

Будућност супарничког машинског учења ће вероватно видети пораст у интердисциплинарној сарадњи, окупљајући стручњаке из области као што су сајбер безбедност, машинско учење и етика како би осмислили холистичке одбрамбене механизме. Објашњивост и интерпретабилност ће постати кључни аспекти, пошто разумевање процеса доношења одлука сложених модела постаје изузетно битно за ублажавање супарничких претњи.

2. Историја и дефинисање супарничког машинског учења

Историја супарничког машинског учења (Adversarial machine learning) може се пратити од раног 21. века када су истраживачи почели да препознају рањивост модела машинског учења на намерне манипулације. Концепт генеративних супарничких мрежа (Generative Adversarial Networks), иако револуционарни у својој способности да генеришу реалистичне податке, такође су отворили врата разумевању динамике супротстављања (Vorobeychik et al., 2018).

Термин „примери супарничког учења“ је добио на значају отприлике у исто време, односећи се на суптилно модификоване уносе дизајниране да доведу у заблуду моделе машинског учења. Истраживачи су почели да анализирају подложност модела овим примерима, откривајући рањивости у препознавању слика, обради природног језика и другим доменима (Joseph et al., 2018).

Како су непријатељски напади постајали све софистицираније, истраживачи су диверзификовали своје напоре да схвате основне механизме. Истраживање се проширило на различите методе напада као што су метода брзог градијента знакова (Fast Gradient Sign method), напад на мапу уочљивости (Jacobian-based Saliency Map Attack) и дубоки напад. Ове технике су имале за циљ да искористе слабости модела ремећи улазне податке на неприметне начине (Vorobeychik et al., 2018).

Историјска временска линија супарничког машинског учења је била сведок промене од пуког признавања рањивости ка развоју одбрамбених механизама. Супарничка обука, робусне технике оптимизације и интеграција интерпретабилности у моделе појавили су се као стратегије за побољшање отпорности модела.

Гледајући унапред, историја супарничког машинског учења служи као основа за текуће истраживање, наглашавајући потребу за проактивним ставом против

непријатељских претњи. Како модели машинског учења постају све присутнији, разумевање историјског контекста постаје кључно у јачању технологије против еволутивних супротстављених изазова.

Дакле, супарничко машинско учење је техника која се користи у машинском учењу за превару или погрешно вођење модела злонамерним уносом. Док се супротстављено машинско учење може користити у различитим апликацијама, ова техника се најчешће користи за извршење напада или изазивање квара у систему машинског учења. Иста инстанца напада може се лако променити да ради на више модела различитих скупова података или архитектура. Модели машинског учења се обучавају коришћењем великих скупова података који се односе на предмет који се проучава. На пример, ако би аутомобилска компанија желела да научи свој аутоматизовани аутомобил како да идентификује знак за заустављање, нахранила би хиљаде слика знакова за заустављање кроз алгоритам машинског учења. Супарнички напад може манипулисати улазним подацима, а у овом случају пружа слике које нису знакови за заустављање, али су означени као такви. Алгоритам погрешно тумачи улазне податке, што доводи до тога да цео систем погрешно идентификује знакове заустављања када се апликација која користи податке машинског учења примени у пракси или производњи (Hashemi-Pour & Gillis, 2023).

Историја развоја супарничких напада се може хронолошки сагледати кроз следећих шест периода (Li et al., 2022):

- **Рана открића (2000-те).** Истраживачи су прво приметили потенцијалне рањивости модела машинског учења и схватили да се класификаторима, као што су машине за вектор подршке и стабла одлучивања, може манипулисати пажљиво израђеним улазима.
- **Истраживање супарничких примера (2013**). Истраживања која су се спровела показале су да мале, неприметне промене у улазним подацима

могу проузроковати да дубоке неуронске мреже погрешно класификују објекте. Ово откриће подигло је свест да нам је потребна одбрана од непријатељских напада.

- **Пробој у дубоком учењу (2014).** Модели дубоког учења, посебно дубоке неуронске мреже, стекли су популарност за постизање изузетних перформанси у различитим задацима, као што су препознавање слике и говора. Међутим, истраживачи су открили да су ови моћни модели веома подложни непријатељским нападима.
- **Повећање истраживања супарничког напада (2016-2018).** Супарничко машинско учење је добило значајну пажњу у академским круговима и индустрији током овог периода. Истраживачи из различитих институција почели су да објављују значајан број радова о контрадикторним нападима, одбрани и утицају на различите алгоритме машинског учења.
- **Утицај у стварном свету (2018-данас).** Супарнички напади више нису ограничени на академске демонстрације. Они сада показују утицај у стварном свету, посебно у компјутерском виду и аутономним системима. На пример, истраживачи су открили да супротстављени напади на знаке за заустављање могу заварати системе за детекцију објеката који се користе у аутомобилима који се сами возе.
- **Развој одбрамбених техника (у току).** Како непријатељски напади и даље представљају изазов, истраживачи и практичари раде на развоју одбрамбених техника како би побољшали робусност модела машинског учења. Укључује супарничку обуку, одбрамбену дестилацију и методе претходне обраде.

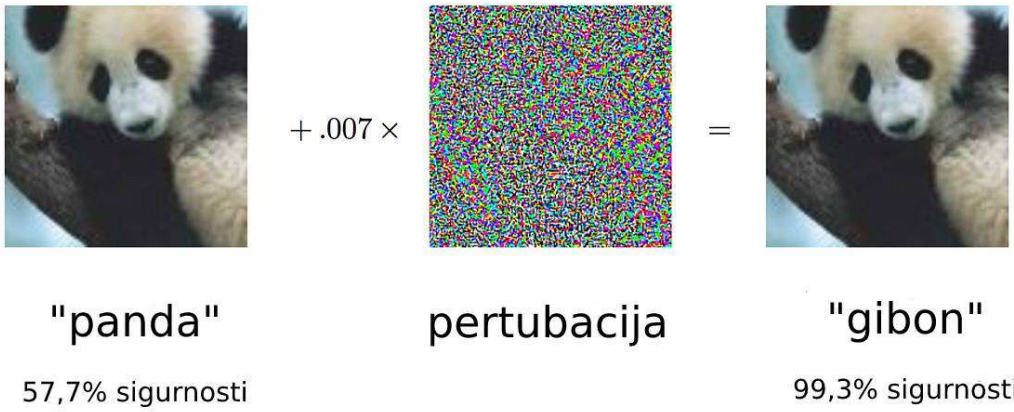
3. Супарнички примери

У данашње време сведоци смо поновног бујања интересовања и напретка за нове технологије везане за вештачку интелигенцију, посебно за коришћење неуронских мрежа. Можемо уочити њихову моћ у класификацији слика и препознавању објеката. На први поглед можемо помислити да су ове неуронске мреже веома моћне и непогрешиве. Међутим, са брзим развојем техника вештачке интелигенције и дубоког учења, од суштинске је важности да се обезбеди сигурност и робусност примењених алгоритама. Било би легитимно преиспитати и истражити потенцијална ограничења и проблеме са перформансама у вези са њиховом употребом.

Супарнички примери представљају фундаментални концепт у области супарничког машинског учења (АМЛ), представљајући случајеве где мале и често неприметне модификације улазних података могу да доведу у заблуду моделе машинског учења. Ове педантно израђене пертурбације су дизајниране са специфичном намером да проузрокују да модел произведе нетачне излазе или класификације, а да притом остане неприметан за људске посматраче. Суштина примера лежи у њиховој способности да искористе инхерентне слабости модела машинског учења. Ови примери се могу манифестовати у различитим доменима, у распону од препознавања слике и говора до обраде природног језика. Обмањујућа природа супарничких примера произилази из њихове способности да изазову неочекивано понашање у моделима, чак и када су измене на улазним подацима минималне (Chivukula et al., 2023).

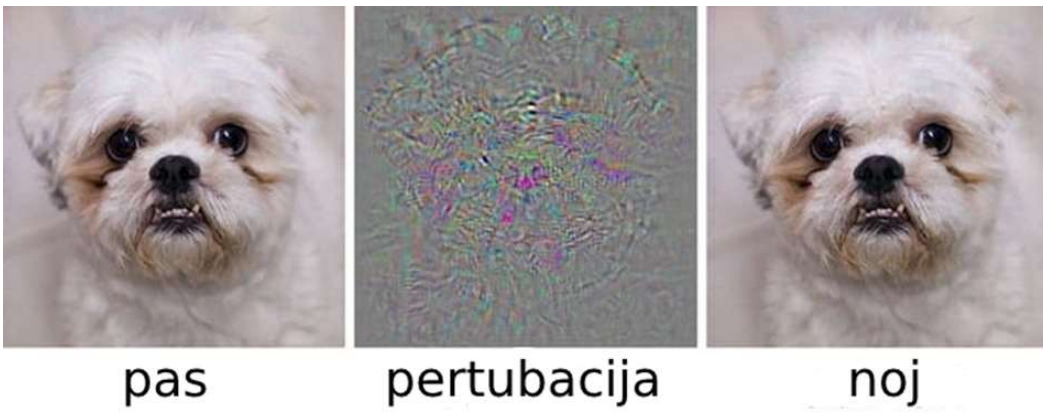
Када замолите човека да опише како детектује панду на слици, можда ће потражити физичке карактеристике као што су округле уши, црне мрље на очима, њушка, крзнена кожа и пружити друге информације као што је тип станишта где очекују да види панду и какве позе заузима. За вештачку неуронску мрежу, све док примена вредности пиксела на једначину даје тачан одговор, она је

уверена да је оно што види заиста панда. Другим речима, променом вредности пиксела слике на прави начин, можете преварити вештачку интелигенцију да помисли да не види панду.



Слика 1. Пример промене вредности пиксела
(DataScientest, приступљено 30.01.2023.)

У примеру испод, можемо видети да је уз благу пертурбацију невидљиву голим оком било могуће преварити неуронску мрежу, која је слику пса класификовала као ноја.



Слика 2. Други пример промене вредности пиксела
(DataScientest, приступљено 30.01.2023.)

Супарнички примери чине моделе машинског учења рањивим на нападе, као у следећим сценаријима (Melanie, 2023):

- Самовозећи аутомобил се судара са другим аутомобилом јер не препознаје знак за заустављање. Неко је ставио слику на знак за заустављање који људима изгледа као знак за заустављање, али је дизајниран да личи на знак забране паркирања за софтвер за препознавање знакова аутомобила.
- Детектор нежељене поште не успева да класификује е-пошту као нежељену пошту. Нежељена е-пошта је дизајнирана да изгледа као нормална е-пошта, али са намером да превари примаоца.
- Скенер са вештачком интелигенцијом на аеродрому скенира пртљаг у потрази за оружјем. Нож је дизајниран да избегне детекцију тако што је натерао систем да верује да је то кишобран.
- Аутоматски систем вештачке интелигенције који не открива болест (нпр. у радиологији) када она заправо одговара озбиљној болести.

Кључно је разумети разлику између циљаних и нециљаних напада. Нециљани напад једноставно има за циљ да изазове погрешну класификацију, без обзира на конкретну категорију. Циљ је искључиво да се постигне погрешна класификација објекта од стране неуронске мреже. Насупрот томе, циљани напад има за циљ да изазове погрешну класификацију у одређеној категорији. На пример, нециљани напад на слику пса тражио би класификацију која није „пас" од стране неуронске мреже. Насупрот томе, циљани напад на исту слику пса имао би за циљ да се пас класификује као ној, али не као мачка, на пример. Постоји неколико метода за креирање супротстављених примера, укључујући оне који се користе у сајбер нападима, као што су тровање података, манипулација роботима и остало. Тровање је у суштини непријатељска контаминација података о обуци. Како системи машинског учења могу бити поново обучени коришћењем података

прикупљених током рада, нападач може отровати податке убризгавањем злонамерних узорака током рада, што накнадно омета или утиче на поновну обуку (Melanie, 2023).

Напади избегавања су најраспрострањенији и најистраженији тип напада. Нападач манипулише подацима током примене да би преварио претходно обучене класификаторе. Пошто се изводе током фазе имплементације, то су најпрактичнији типови напада и најчешће коришћени напади на сценарије упада и малвера. Нападачи често покушавају да избегну откривање тако што прикривају садржај малвера или нежељене е-поште. Због тога су узорци модификовани да би избегли откривање пошто су класификовани као легитимни без директног утицаја на податке о обуци. Примери избегавања су лажни напади на системе биометријске верификације (Boesch, 2023).

4. Методе супарничког напада

Постоје различите методе које се користе приликом супарничког напада, почевши од оних најранијих до све софистицираних. Неке од ових метода су већ спомињане у тексту, али сад ће уследити детаљнија анализа уз сликовите примере. Док истражујемо нијансе ових методологија напада, циљ је да откријемо савремене тактике супротстављања и поставимо основу за касније дискусије о одбрамбеним механизмима у суочавању са претњама које се развијају.

4.1. Напад па беле кутије

Напад беле кутије представља категорију супарничких напада у којима противник има потпуно знање о архитектури циљаног модела машинског учења, параметрима и подацима о обуци. У овом сценарију, нападач користи ово свеобухватно разумевање како би направио прецизне и ефикасне супарничке

примере. Ови напади симулирају сценарио у којем противник има инсајдерско знање, омогућавајући им да искористе рањивости и слабости модела са вишим степеном софистицираности. Овај метод представља значајну претњу по безбедност система машинског учења, наглашавајући важност развоја робусне одбране од противника са детаљним познавањем циљаног модела (Joseph et al., 2018).

4.2. Напад па црне кутије

С друге стране, напад црне кутије представља категорију супарничких напада у којима противник ради са ограниченим или без знања о интерној структури циљаног модела машинског учења, параметрима или подацима о обуци. За разлику од напада на белу кутију, напади на црну кутију симулирају сценарио где противник има приступ само улазном и излазном понашању модела. Противници који користе нападе црне кутије често се ослањају на технике као што су приступи засновани на упитима, где итеративно испитују модел и посматрају његове одговоре да би конструисали супарничке примере. Ови напади представљају значајан изазов за одбрану од непријатељских манипулација, пошто унутрашње функционисање циљаног модела остаје непрозирно за нападача. Развијање робусне одбране од напада црне кутије је кључно за побољшање безбедности система машинског учења (Joseph et al., 2018).

4.3. Метод ограничене меморије

Метод ограничене меморије (Limited-memory Broyden-Fletcher-Goldfarb-Shanno method - L-BFGS) са ограниченом меморијом представља софистицирану технику оптимизације која се користи у супарничком машинском учењу за прављење супарничких примера. Као квази-Њутнов метод, ово спада у ширу категорију алгоритама оптимизације, посебно дизајнираних за сценарије у којима функција

циља, у овом контексту, функција губитка модела машинског учења, треба да буде минимизирана (Saputro & Widyaningsih, 2017).

У домену супротстављених напада, метод ограничене меморије функционише тако што учестало прилагођава улазне податке како би пронашао пертурбацију која максимално обмањује модел машинског учења док минимизира перцептуалне промене код људског посматрача. За разлику од неких метода заснованих на градијенту, овај метод укључује меморијске ефикасне апроксимације Хесове матрице, што га чини посебно погодним за проблеме оптимизације високих димензија (Saputro & Widyaningsih, 2017).

Метода је нашла примену у различитим доменима, од компјутерског вида до обраде природног језика, захваљујући својој ефикасности у тражењу оптималних пертурбација у сложеним и великим моделима машинског учења. Истраживачи и противници подједнако користе ову методу да испитају рањивост модела и открију потенцијалне слабости, подвлачећи стални изазов одбране од напредних супарничких напада заснованих на оптимизацији у машинском учењу.

У контексту обраде природног језика, посебно са моделима заснованим на тексту, као што су рекурентне неуронске мреже или трансформатори, противник може да користи метод ограничене меморије за прављење супротстављених примера. Ови примери могу бити суптилно модификоване реченице дизајниране да наведу модел да направи нетачна предвиђања или класификације. могао применити за манипулисање аудио улазом у систему за препознавање говора. Супарничке пертурбације које генерише методу, могу се унети у изговорене речи или фразе, што доводи до тога да систем погрешно тумачи унос и производи нетачне транскрипције. У сценаријима где се за откривање аномалија користе модели машинског учења, овај метод би се могао користити за генерисање супротстављених примера који опонашају нормално понашање. Ови примери могу бити направљени да обману модел да погрешно класификује аномалне

инстанце, угрожавајући ефикасност система за детекцију аномалија (Saputro & Widyaningsih, 2017).

4.4. Метода брзог градијента

Метода брзог градијента (Fast Gradient Sign method - FGSM) је популарна и ефикасна техника супротстављеног напада у области машинског учења коју је развио Ијан Гудфелоу и његове колеге 2014. године. Она ради тако што користи градијенте неуронске мреже за генерисање супарничких примера. Кључна карактеристика је његова једноставност, што га чини атрактивним избором за истраживаче и практичаре који проучавају супарничке нападе. Она ремети улазне податке додавањем малог дела градијента функције губитка у односу на улаз, обезбеђујући брзо рачунање супротстављених примера (Milton, 2018).

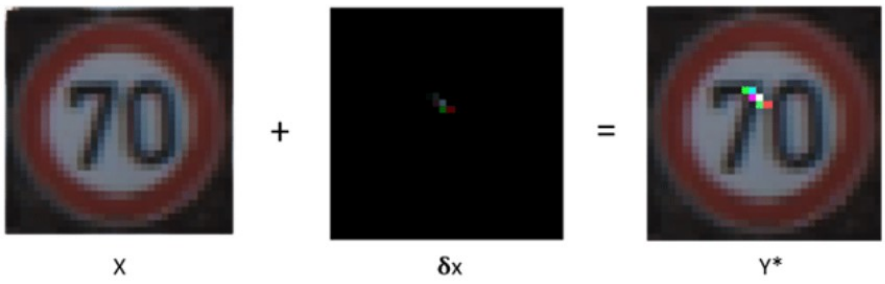
Супарнички примери које је направио овај метод често су неприметни за људске посматраче, наглашавајући суптилност овог напада. Посебно је ефикасан у задацима класификације слика, где мале пертурбације улазних слика могу довести до погрешне класификације од стране модела машинског учења. Успех напада лежи у његовој способности да искористи линеарност неуронских мрежа, омогућавајући противницима да манипулишу предвиђањима модела уз минимални рачунски напор.

Да бисмо преварили неуронску мрежу да направи погрешна предвиђања потребно је (Ansah, 2018):

- унапред пропагирати слику кроз нашу неуронску мрежу
- израчунати губитак
- вратити градијенте на слику
- померити пикселе слике у правцу који максимизира вредност губитака

4.5. Напад на мапу уочљивости

Напад на мапу уочљивости (Jacobian-based Saliency Map Attack - JSMA) је софистицирани метод супарничког напада који се користи у машинском учењу. За разлику од неких традиционалних метода, он се фокусира на манипулисање улазним подацима идентификацијом и узнемиравањем најистакнутијих карактеристика, као што је одређено Јакобијанском матрицом. Овај напад користи осетљивост границе одлучивања модела на специфичне карактеристике, омогућавајући суптилне и ефикасне промене улазних података. Стратешки приступ чини га посебно моћним у сценаријима у којима супротстављене пертурбације морају бити неприметне за људске посматраче док изазивају погрешне класификације у циљаном моделу. Истраживачи често користе овај метод да истраже интерпретабилност модела машинског учења и да открију рањивости у вези са значајем карактеристика, доприносећи текућем развоју робусне одбране од непријатељских напада. (Combey et al., 2020).



Слика 3. Ограничење брзине од 70 км/х са минимално додатим пертурбацијама погрешно класификованим као знак ограничења брзине од 30km/h (ResearchGate, 2023)

Овај напад може се применити на различите области да би се направили супарнички примери. За апликације за обраду природног језика, напад на мапу уочљивости може да манипулише улазним текстом узнемиравајући кључне речи или фразе. Ово би могло резултирати погрешном класификацијом сентимента текста или теме, показујући свестраност ове методе у задацима заснованим на

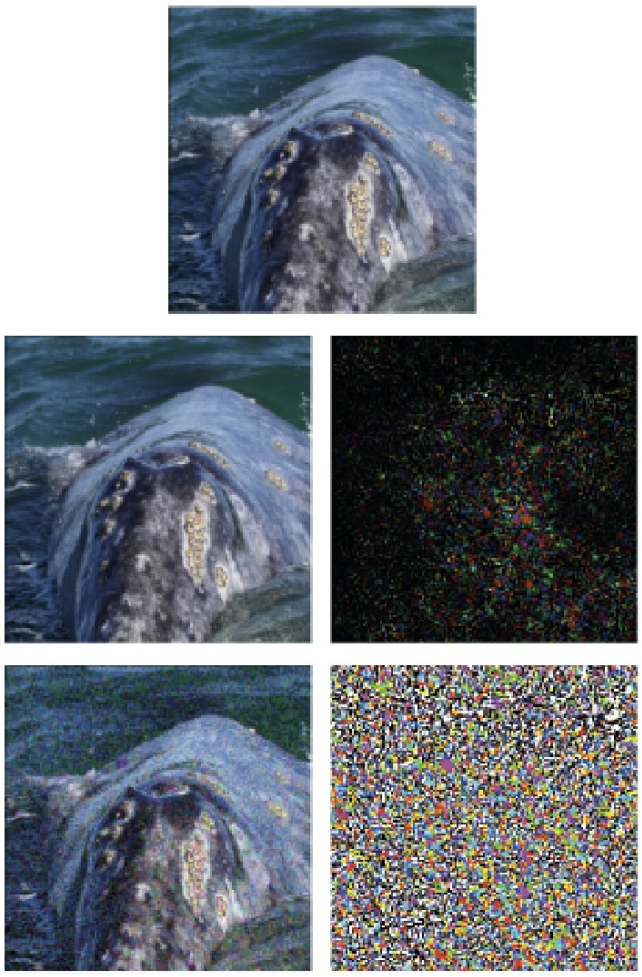
тексту. Такође се може користити за стварање супарничких пертурбација у аудио улазу за системе за препознавање говора. Идентификовањем и изменом критичних компоненти аудио сигнала, напад може довести до тога да систем погрешно протумачи изговорене речи или команде. У области аутономних возила, напад би се могао применити за сметње улазних података са сензора као што је радар. Суптилне манипулације у подацима сензора могу довести до погрешних тумачења система перцепције возила, потенцијално утицати на доношење одлука. У анализи медицинске слике, ово би се могао користити за генерисање супротстављених примера манипулисањем истакнутим карактеристикама радиолошких слика. Ово би могло да представља претњу по поузданост модела машинског учења који се користе у дијагностичке сврхе (Combey et al., 2020).

4.6. Дубоки напад

Дубоки напад (Deepfool Attack) ради тако што проналази најмању могућу пертурбацију у правцу границе одлуке. За разлику од неких других метода, овај напад се не ослања на градијенте, што га чини ефикасним у низу архитектура неуронских мрежа. Његова снага лежи у способности да генерише супарничке примере који могу да преваре моделе уз минималне пертурбације, често избегавајући да их открију људски посматрачи. Свестраност напада чини га применљивим на различите задатке машинског учења, укључујући класификацију слика и детекцију објеката, наглашавајући текуће изазове у одбрани од софистицираних непријатељских напада (Morgan, 2022).

На слици број 4. приказано је поређење две методе напада усмерене на исту слику. У горњем реду класификатор је одредио да је оригинална слика кит. У средњем реду, оригинална слика кита је поремећена, што је довело до тога да је класификатор одредио да је то корњача. Поремећена слика је веома суптилна и

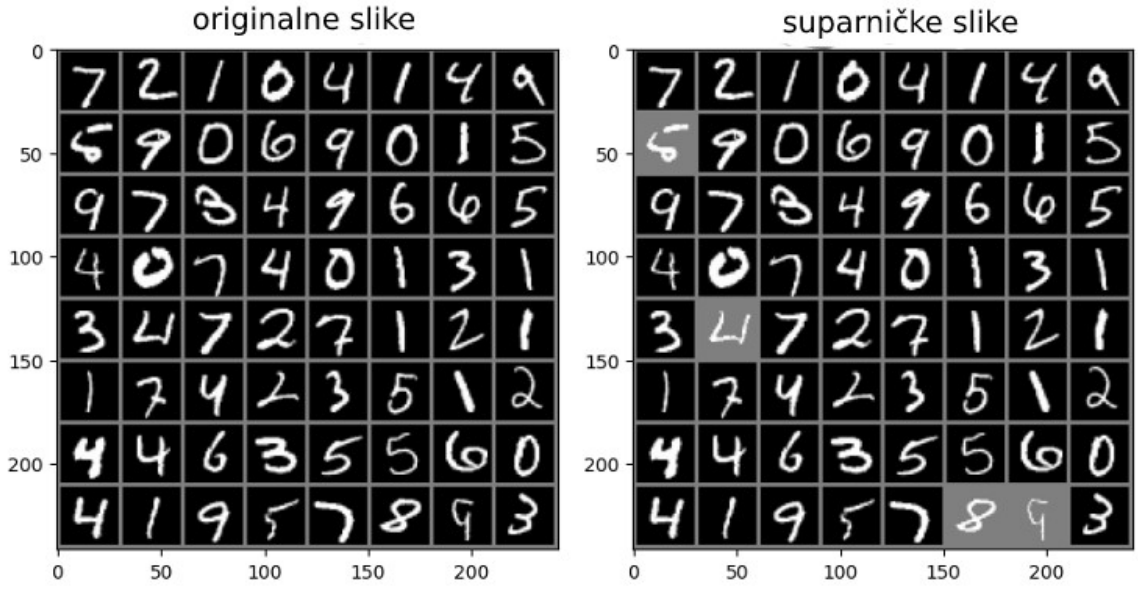
неприметна за људско око. Последњи приказује сличну поремећену слику, али је поремећена методом брзог градијента. Овај метод је такође успео да погрешно класификује оригиналну слику као корњачу, али је количина потребних пертурбација већа него приликом методе дубоког напада, што се може визуелно видети.



Слика4. Поређење две методе напада на исту слику (Morgan, 2022)

4.7. Напад Карлинија и Вагнера

Напад Карлинија и Вагнера (Carlini & Wagner Attack (C&W) је напредни супарнички метод напада који су увели Николас Карлини и Дејвид Вагнер 2017. године. Одликује се својим приступом заснованим на оптимизацији и настоји да генерише минималне пертурбације које могу ефикасно да доведу у заблуду моделе машинског учења док одржавају неприметност за људе посматрачи. За разлику од једноставнијих напада, он формулише генерисање супарничких примера као проблем оптимизације, омогућавајући му да пронађе решења која су моћна у различитим архитектурама модела. Напад је познат по својој прилагодљивости, успешан је и против циљаних и нециљаних сценарија. Напад Карлинија и Вагнера наглашава сталну потребу за снажном одбраном док противници настављају да усавршавају технике које изазивају сигурност система машинског учења. (Zachariah, 2023)



Слика 5. Упоређивање истих цифара суптилно поремећених нападом Карлинија и Вагнера (Zachariah, 2023)

4.8. Генеративне супарничке мреже

Генеративне супарнчке мреже (Generative Adversarial Networks (GAN) су приступ генеративном моделирању користећи методе дубоког учења, као што су конволуцијске неуронске мреже. Генеративно моделирање је задатак учења без надзора у машинском учењу који укључује аутоматско откривање и учење правилности или образаца у улазним подацима на такав начин да се модел може користити за генерисање или излаз нових примера који су вероватно могли бити извучени из оригиналног скупа података (Brownlee, 2019).

Они су паметан начин обуке генеративног модела тако што се проблем уоквирује као проблем учења под надзором са два подмодела: модел генератора који обучавамо да генерише нове примере и модел дискриминатора који покушава да класификује примере као стварне или лажне. Два модела се тренирају заједно у игри са нултом сумом, све док се модел дискриминатора не превари, што значи да модел генератора генерише уверљиве примере. ГАН-ови су узбудљиво поље које се брзо мења, пружајући обећање генеративних модела у својој способности да генеришу реалне примере у низу проблематичних домена, посебно у задацима превођења слике у слику као што је превођење фотографија лета у зиму или дан до ноћи, и у генерисању фотореалистичних фотографија објеката, сцена и људи за које чак ни људи не могу да препознају да су лажни (Brownlee, 2019).

4.9. Оптимизацијски напад нултог реда

Оптимизација нултог реда је подскуп оптимизације без градијента који је уграђен у различите апликације за учење сигнала и машинског учења. Алати за оптимизацију нултог реда су у суштини еквиваленти првог реда без градијента. Користећи прорачуне функционалног градијента, нулти поредак апроксимира укупне градијенте или стохастичке градијенте (Learnbay Data Science, 2021).

Три кључне предности оптимизације нултног реда:

- Брза имплементација са минималним променама у уобичајеним алгоритмима заснованим на градијенту
- Рачунски ефективна апроксимација деривата када је тешко израчунати
- Упоредиве стопе конвергенције првог реда

Због успешног решавања проблема обраде сигнала, дубоког учења и машинског учења, оптимизација нултог реда је привукла већу пажњу. Овај метод оптимизације је моћно и практично средство помоћу којег се може проценити штетна робусност система дубоког учења. Оптимизација нултог реда се постиже кроз ефикасне процене градијента приближавањем пуном градијенту.

5. Одбрана од супарничког напада

Одбрана од супарничког напада је критичан аспект обезбеђивања робусности и поузданости модела машинског учења. Као што је речено, супарнички примери су пажљиво израђени инпути (улази) дизајнирани да преваре моделе и наведу их да направе погрешна предвиђања или класификације. Да би се ојачали модели против ових манипулација, предложени су и имплементирани различити одбрамбени механизми (Chivukula et al., 2023).

Једна уобичајена одбрамбена стратегија је супарнички тренинг, где се модели обучавају на комбинацији редовних и супарничких примера. Излажући модел супротстављеним инпутима током обуке, он учи да боље разликује праве и манипулисане податке, повећавајући његову отпорност. Технике робусне оптимизације укључују модификацију процеса оптимизације током тренинга како би модели инхерентно били отпорнији на контрадикторне пертурбације. Ово може укључити инкорпорирање услова регуларизације који кажњавају

екстремне промене у улазним карактеристикама, чинећи модел мање осетљивим на суптилне измене (Short et al., 2019).

Функција стискања је техника која укључује претходну обраду улазних података како би се смањила прецизност карактеристика, што противницима чини изазовнијим да идентификују рањиве димензије за манипулацију. Ово се може постићи квантизацијом улазних података или применом других метода компресије (Short et al., 2019).

Ансамбл одбрана (Ensemble Defenses) користи више модела за процену истог инпута, што отежава противницима да направе универзалне противничке примере који истовремено обмањују све моделе. Диверзификацијом ансамбла модела, укупан систем постаје робуснији (Chivukula et al., 2023).

Трансформација улаза укључује модификацију улазних података на начин који задржава првобитно значење, али нарушава супротстављене пертурбације. Ово може укључивати технике попут додавања шума или примене трансформација слике.

Међутим, кључно је напоменути да је развој ефикасне одбране од супарничких примера стални изазов, јер противници стално прилагођавају своје стратегије. Игра мачке и миша између нападача и бранилаца подстиче текуће истраживање како би се створили софистицираније модели и одбрамбени механизми, осигуравајући континуирани напредак сигурних и поузданих система машинског учења.

6. Закључак

Област супарничког машинског учења представља динамичан и еволуирајући изазов који захтева нашу сталну пажњу и иновације. Како модели машинског учења налазе све већу примену у различитим доменима, рањивост на непријатељске нападе постаје критична брига. У овом раду истраживали смо различите методе супарничког напада, у распону од неких традиционалних као што је метод брзог градијента до софистицираних приступа заснованих на оптимизацији као што су напад Карлинија и Вагнера и напад на мапу уочљивости.

Ови напади наглашавају хитну потребу за снажним одбрамбеним механизмима. Супарничка обука, робусне технике оптимизације и трансформације улаза су предложене као стратегије за јачање модела против покушаја манипулације. Међутим, овај сукоб између нападача и одбрамбених играча се наставља, што подстиче стална истраживања да остану корак испред у овом непријатељском пејзажу.

Док улазимо у замршености одбране од супротстављених примера, постаје очигледно да је постизање поуздане безбедности сложен задатак. Противници се упорно прилагођавају и усавршавају своје технике, доводећи у питање саме темеље робусности машинског учења. Без обзира на то, наша тежња за безбедним моделима машинског учења треба да буде неумољива и захтева холистички приступ који комбинује напредак у архитектури модела, одбрамбеним стратегијама и етичким разматрањима у вези са применом технологија машинског учења.

Суочени са овим динамичним окружењем, лекције извучене из истраживања супарничког машинског учења не доприносе само унапређењу безбедних система вештачке интелигенције, већ и наглашавају важност интердисциплинарне

сарадње, етичких разматрања и сталне будности. Како ова област наставља да сазрева, развој отпорних модела и ефикасних одбрамбених механизама играће кључну улогу у ослобађању пуног потенцијала технологија машинског учења уз истовремено ублажавање ризика које представљају контрадикторне претње.

Литература

Ansah, H. (2023, August 24). Adversarial attacks on neural networks: Exploring the fast gradient sign method. neptune.ai. https://neptune.ai/blog/adversarial-attacks-on-neural-networks-exploring-the-fast-gradient-sign-method приступљено 02.01.2024.

Boesch, G. (2023b, December 21). What is adversarial Machine Learning? Attack Methods in 2024. viso.ai. https://viso.ai/deep-learning/adversarial-machine-learning/ приступљено 30.12.2023.

Brownlee, J. (2019, July 19). A gentle introduction to Generative Adversarial networks (GANs). MachineLearningMastery.com. https://machinelearningmastery.com/what-are-generative-adversarial-networks-gans/ приступљено 03.01.2024.

Chivukula, A. S., Yang, X., Liu, B., Liu, W., & Zhou, W. (2023). Adversarial Machine Learning: Attack Surfaces, Defence Mechanisms, Learning Theories in Artificial Intelligence. Springer Nature.

Combey, T., Loison, A., Faucher, M., & Hajri, H. (2020). Probabilistic Jacobian-based Saliency Maps Attacks. CentraleSupelec, 3 Rue Joliot-Curie 91192, Gif-sur-Yvette, France. IRT SystemX, 8 Avenue de la Vauve, 91120 Palaiseau, France.

Fig. 5. Jacobian saliency map method example of a 70 km/h speed limit. . . (2023). ResearchGate. https://www.researchgate.net/figure/Jacobian-saliency-map-method-example-of-a-70-km-h-speed-limit-sign-with-minimally-added_fig8_322076393 приступљено 29.12.2023.

Goodfellow, I. J. (2014, December 20). Explaining and harnessing adversarial examples. arXiv.org. https://arxiv.org/abs/1412.6572 приступљено 02.01.2024.

Hashemi-Pour, C., & Gillis, A. S. (2023, November 22). adversarial machine learning. Enterprise AI. https://www.techtarget.com/searchenterpriseai/definition/adversarial-machine-learning приступљено 30.12.2023.

Joseph, A. D., Nelson, B., Rubinstein, B. I., & Tygar, J. D. (2018). Adversarial machine learning. Cambridge University Press.

Li, F., Lai, L., & Cui, S. (2022). Machine Learning Algorithms: Adversarial Robustness in Signal Processing. Springer Nature.

Melanie. (2023, October 30). Adversarial Examples: Definition and importance in machine

learning - Data Science Courses | DataScientest. Data Science Courses | DataScientest. https://datascientest.com/en/adversarial-examples-definition-and-importance-in-machine-learning приступљено 30.12.2023.

Milton, M. A. A. (2018). Evaluation of momentum diverse input iterative fast gradient sign method (M-DI2-FGSM) based attack method on MCS 2018 adversarial attacks on black box face recognition system. arXiv preprint arXiv:1806.08970.

Morgan, A. (2022, May 7). A Review of DeepFool: a simple and accurate method to fool deep neural networks. Medium. https://medium.com/machine-intelligence-and-deep-learning-lab/a-review-of-deepfool-a-simple-and-accurate-method-to-fool-deep-neural-networks-b016fba9e48e приступљено 29.12.2023.

Saputro, D. R. S., & Widyaningsih, P. (2017). Limited Memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) Method for The Parameter Estimation on Geographically Weighted Ordinal Logistic Regression Model (GWOLR). Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Sebelas Maret.

Science, L. D. (2021, December 15). Applications of zeroth Order Optimization in Deep Learning. Medium. https://medium.com/learnbay-blogs/applications-of-zeroth-order-optimization-in-deep-learning-dda98243af02 приступљено 03.01.2024.

Short, A., La Pay, T., & Gandhi, A. (2019). Defending Against Adversarial Examples. Sandia National Laboratories, Albuquerque, New Mexico.

Vorobeychik, Y., Kantarcioglu, M., Brachman, R., Stone, P., & Rossi, F. (2018). Adversarial machine learning (Vol. 12). San Rafael, CA, USA: Morgan & Claypool Publishers.

You, A., Kim, J. K., Ryu, I. H., & Yoo, T. K. (2022). Application of generative adversarial networks (GAN) for ophthalmology image domains: a survey. Eye And Vision, 9(1). https://doi.org/10.1186/s40662-022-00277-3 приступљено 29.12.2023.

Zachariah, A. G. (2023, December 28). Adversarial Attacks with Carlini & Wagner Approach - Arun George Zachariah - Medium. Medium. https://medium.com/@zachariaharungeorge/adversarial-attacks-with-carlini-wagner-approach-8307daa9a503

ИЗЈАВА О АКАДЕМСКОЈ ЧЕСТИТОСТИ

Изјављујем да сам у приложеном раду поштовао/ла сва правила о академској честитости.

Овај писани рад резултат је искључиво мог личног рада, темељи се на мојим истражиањима и ослања се на наведену литературу.

У Београду, дана _____ године.

Потпис студента:

_____