Anna-Mari Rusanen

Pikseleitä, kohinaa ja haurautta

Syväoppivien koneoppimissovellusten tutkimus on ollut viimeiset vuodet suhteellisen insinöörivetoista. Tutkimusta ohjaa usein lähinnä ohjelmistokehitys, ja teoreettiset kysymykset ovat jääneet vähemmälle huomiolle. Insinöörialoilla myös monesti vierastetaan ajatusta, että tietoteknisten ongelmien ratkaisua varten haettaisiin vetoapua poikki- tai monitieteisestä tutkimuksesta. Nykyiset algoritmit ja erityisesti syväoppivat arkkitehtuurit tuottavat kuitenkin ongelmia, joita ei voida ratkaista vain ohjelmointiteknisesti. Adversariaalit eli häiritsevät esimerkit ovat yksi näistä ongelmista.

Syväoppimiseen (*deep learning*, DL) tai syviin neuroverkkoihin (*deep neural networks*, DNN) perustuvia koneoppimissovelluksia käytetään nykyisin kaikkialla. Niiden avulla voidaan luokitella ja tunnistaa miltei mitä tahansa objekteja – ihmiskasvoja pankkiautomaatilla, CAPTCHA-kirjaimia varmenteissa tai ääntä puheohjaimissa. Joillakin arkkitehtuureilla, kuten GANeilla (*Generative Adversarial Networks*)[1],voidaan tuottaa hätkähdyttävän aidon oloisia keinotekoisia kuvia, ääntä tai videoita vaikkapa julkisuuden henkilöistä[2].

Syväoppivien arkkitehtuurien kääntöpuoli on kuitenkin niiden systemaattinen ”hauraus” (*brittle*) eli herkkyys tietyntyyppisille virheille[3]. Järjestelmät ovat hämmentävällä tavalla herkkiä ”adversariaaleille” eli häiritseville piirteille (*adversarial examples*). Adversariaalit ovat syötteisiin lisättyjä piirteitä, joilla voidaan manipuloida järjestelmien toimintaa. Esimerkiksi kuvantunnistussovellus, joka on ensin oppinut luokittelemaan pandojen kuvat oikein, saadaan luokittelemaan pandat systemaattisesti gibboneiksi lisäämällä syötteeseen hiukan kohinaa[4]. Tarkasti ei tiedetä, miksi järjestelmät reagoivat tällä tavalla kohinaan (tai muihin häiritseviin piirteisiin). Adversariaalit ja niiden taustalla piilevä järjestelmien hauraus ovat yksi esimerkki koneoppimisen ”mustista laatikoista”. On kuitenkin epäselvää, *millaisesta* mustasta laatikosta niiden kohdalla on kysymys tai *miksi* niitä ei osata selittää.

”Mustien laatikoiden” ongelmat eivät ole pelkästään ohjelmointiteknisiä vaan osittain käsitteellisiä ja teoreettisia. Tilannetta monimutkaistaa, että adversariaaleissa ei ole kysymys koneiden varsinaisesta virhetoiminnasta. Koneet eivät siis toimi yhtäkkiä ”mystisellä tavalla väärin”, ”muutu psykoottisiksi” tai ”ala hallusinoida”, kuten joissakin uutisotsikoissa on väitetty. Päinvastoin laskennallisesta näkökulmasta adversariaalien vaikutuksesta syntyvät luokitukset ovat usein koneelle itselleen ”oikeita”[5]. Tutkijat ehdottavatkin, että adversariaaleissa on pikemminkin kyse DNN-arkkitehtuurien ja ihmisen neurokognitiivisen järjestelmän perustavasta erilaisuudesta, ei niinkään koneiden toimintavirheistä[6].

Asetelma on mielenkiintoinen. Jos adversariaalit ovat osittain ihmisen ja koneen välisen havaintokognitiivisen prosessoinnin yhteensovittamisen ongelmia, nämä ongelmat eivät ratkea vain ohjelmointi- tai tietoteknisesti. Ihmisen ja koneiden luokittelujärjestelmien eroja ja yhtäläisyyksiä sekä niistä nousevia ilmiöitä ei yksinkertaisesti voida tutkia pelkillä tietoteknisillä menetelmillä. Ihmisen havaintokognitiivisten järjestelmien huomioiminen edellyttää muun muassa kognitiontutkimuksen sekä havainto- ja neurotieteiden menetelmien ja teorioiden hyödyntämistä. Haurauden tai adversariaalien selittäminen ja ymmärtäminen vaativat lähtökohtaisesti *sekä* teoreettisempaa *että* monitieteisempää tutkimusotetta.

Mitä adversariaalit ovat?

Christian Szegedy ja hänen kollegansa osoittivat ensimmäisinä, kuinka koneoppimisjärjestelmiä voidaan manipuloida lisäämällä syötteisiin ylimääräisiä, häiritseviä piirteitä[7]. Näiden adversariaalien vaikutuksesta kuvantunnistusjärjestelmät alkavat systemaattisesti luokitella objekteja virheellisesti. Häiritsevät piirteet voivat olla miltei mitä tahansa yksittäisistä pikseleistä kokonaisiin kuvioihin, ja niitä voidaan tuottaa useilla menetelmillä[8]. Esimerkiksi Goodfellow’n ja kollegoiden kokeessa häirintä tapahtui lisäämällä kuviin hiukan kohinaa[9]. Muissa kokeissa syöteaineistoihin on lisätty muun muassa psykedeelistä kuviota muistuttavia ”tarroja”. Näin järjestelmä, joka oli harjoitettu luokittelemaan hedelmät ja leivänpaahtimet oikein, saatiin luokittelemaan banaanit leivänpaahtimiksi[10].

Toistaiseksi ei ole selkeää käsitystä, miksi tai miten adversariaalit vaikuttavat kuvantunnistusjärjestelmien toimintaan. Tutkijat arvelevat, että useimmissa tapauksissa kysymys on syväoppivien verkkojen kolmen piirteen yhteisvaikutuksesta. Ensinnäkin kuvantunnistusjärjestelmille ”kuvat” ovat pikselien eli kuvapisteiden muodostamia kokonaisuuksia. Järjestelmät eivät siis ihmisten tavalla ”näe” kuvia *kuvina*, saati sitten ”pandoja”

Ilari Hautamäki, *Beige Flower Pot* (2019), akryyli ja öljy kankaalle, 50 × 40cm

tai ”kissoja” *esittävinä* kuvina. Toiseksi DNN-pohjaiset järjestelmät hakevat tilastollisia säännönmukaisuuksia niihin syötetystä datasta. Jos neuroverkkoon esimerkiksi syötetään kymmeniä tuhansia nimikoituja eläinten kuvia, verkko oppii yhdistämään, mitkä pikselien (tai piirteiden) säännönmukaisuudet liitetään esimerkiksi ”pandoihin” ja mitkä ”kissoihin”. Kun verkko on oppinut luokitukset, se pystyy niiden avulla tunnistamaan myös uusista kuvista ”pandaan” liittyvät säännönmukaisuudet. Olennaista on, että koneoppimisen näkökulmasta kuvien luokitukset eivät ole semanttisia tai sisällöllisiä. Sen sijaan ne perustuvat datan tilastollisille ja matemaattisille säännönmukaisuuksille.

Kolmanneksi kuvantunnistusjärjestelmät harjoitetaan tyypillisesti maksimoimaan luokituksen täsmällisyys. Järjestelmät käyttävät mitä tahansa piirrettä tai signaalia, jonka avulla ne pystyvät maksimointiin riippumatta siitä, onko piirteen tai signaalin sisältö ihmisen näkökulmasta mielekäs tai havaittavissa. Pikselien, vektorien ja laskennan näkökulmasta adversariaalien vaikutuksesta syntyvät luokitukset voivat siten olla koneelle itselleen ”oikeita” laskennan lopputuloksia, jos ne maksimoivat luokitusten täsmällisyyttä (tai muuta vastaavaa laskennan tavoiteltua lopputulosta).

Vielä ei kuitenkaan ymmärretä laskennallisesti tai teoreettisesti, miksi koneet valitsevat adversariaalit piirteet maksimoinnin perusteiksi tai miksi ne tuottavat juuri sellaisia luokituksia kuin tuottavat. Adversariaalien tietty yleistyvyys viitannee kuitenkin johonkin perustavanlaatuiseen arkkitehtuuriseen laskennalliseen ominaisuuteen tai ominaisuuksien yhteisvaikutukseen. Kuten Papernot kollegoineen huomauttaa, adversariaali, joka häiritsee viittä mallia, todennäköisesti häiritsee myös kuudetta[11]. Adversariaalit piirteet aiheuttavat luokitusten muuntumista eri arkkitehtuureissa jopa silloin, kun harjoitusaineistot tai algoritmit ovat erilaisia[12]. Liu kollegoineen osoittaa lisäksi, että adversariaalien yleistymistä voidaan lisätä optimoimalla ne huijaamaan mahdollisimman montaa mallia[13]. Kuten Ilyas kollegoineen esittää, adversariaalit ovat konkreettinen esimerkki siitä, kuinka tietyt DNN- ja koneoppimisjärjestelmät ovat systemaattisesti arkkitehtuurisella tasolla ”hauraita” eli herkkiä tietyntyyppisille ”virheille”[14].

Ihmisen vai koneen ongelma?

Kiinnostavasti adversariaalien aiheuttamien luokittelujen ”virheellisyys” – ja siten *a fortiori* myös niiden ”hauraus” – on ilmeisesti osittain *ihmiskeskeinen ongelma*[15]. Jos kone toimii moitteettomasti, sen näkökulmasta opittu luokittelu on optimimaalisin ratkaisu ongelmaan, jota se on harjoitettu laskemaan. Koneet eivät siis adversariaalien vaikutuksesta muutu psykoottisiksi, tee laskuvirheitä tai ole vinoutuneita. Koneiden mahdollisia mielenterveysongelmia tai laskuvirheitä parempi selittäjä saattaakin olla, että ihmisen näköjärjestelmä ja kuvantunnistusjärjestelmät eivät välttämättä tuota samanlaisia luokituksia *edes* silloin, kun konetta on harjoitettu niin sanotusti ihmisen havaintojärjestelmän luokituksiin perustuvalla nimikoidulla datalla[16].

Mikä voisi aiheuttaa tämän eron luokitteluissa? Syitä on useita. Yksittäisten pikselien kokoisissa häiritsevissä piirteissä ero selittyy osittain ihmisen näön tarkkuuden riittämättömyydellä. Erojen taustalla on myös muita, huomattavasti monimutkaisempia tekijöitä. Brownin ja kollegoiden tutkimuksessa sivuttiin yhtä niistä. Kuvantunnistusjärjestelmä oli ensin opetettu erottamaan leivänpaahtimet hedelmistä. Sitten järjestelmän toimintaa manipuloitiin lisäämällä leivänpaahtimien kuviin psykedeelisiä kuvioita. Kuvioiden vaikutuksesta kuvantunnistussovellus alkoi luokitella myös esimerkiksi banaanit leivänpaahtimiksi. Vaikka ilmiötä ei osata täysin selittää, psykedeeliset kuviot ilmeisesti vaikuttavat siihen, mitä kuvantunnistussovellus pitää tilastollisesti silmiinpistävimpinä eli *salientteina* piirteinä.[17] Jos ihminen näkisi samat psykedeeliset kuviot, ne eivät vaikuttaisi näköjärjestelmän luokituksiin. Siinä missä kuvantunnistussovellukset tyypillisesti operoivat niin kutsutulla ”tilastollisella salienssilla”, jossa piirteen silmiinpistävyys määritellään sen tilastollisten ominaisuuksien avulla (niin sanottu *bottom up* -prosessointi), ihmisen näköjärjestelmä huomioi myös muita tekijöitä. Arvioidessaan piirteiden salienssia ihmisen näköjärjestelmä huomioi koneoppimissovelluksia laajemmin esimerkiksi kontekstisidonnaisia tekijöitä. Se esiohjaa silmiinpistävyysarviointeja muun muassa arvioimalla piirteen (eli ärsykkeen) relevanssin eli merkityksen havaintokognitiiviselle tehtävälle (niin sanottu *top down* -prosessointi)[18]. Mikä on ihmiselle salienttia, ei aina vastaa sitä, mikä on kuvantunnistusjärjestelmälle salienttia edes silloin, kun järjestelmien syöte on ”sama”.

Nämä esimerkit alleviivaavat ennen kaikkea sitä, ettei voi olettaa *a priori*, että koneet automaattisesti ”näkevät” kuin ihmiset tai että ne ”luokittelevat objekteja” kuin ihmiset silloinkaan, kun niitä harjoitetaan ihmisten näköjärjestelmän toimintaan perustuvalla aineistolla. Pikemminkin esimerkit paljastavat konkreettisesti, kuinka perustavasti ihmisen ja koneen prosessointi eroavat toisistaan. Toisaalta ei voida myöskään olettaa *a priori*, että koneet ja ihmiset olisivat välttämättä täysin erilaisia. On nimittäin myös väitetty, että ihmisaivot käsittelisivät varhaisen sensorisen prosessoinnin tasolla adversariaaleja samalla tavalla kuin tietyt koneoppimisalgoritmit[19]. Vaikka nämä ehdotukset ovat alustavia ja niiden tueksi tarjottu evidenssi on metodologisesti ja käsitteellisesti hataraa, itse kysymys ihmisaivojen ja kuvantunnistusjärjestelmien mahdollisista samankaltaisuuksista on silti oikeutettu[20].

Olennaista on kuitenkin, että mitä enemmän adversariaalien ongelma perustuu ihmisen ja koneen luokittelujärjestelmien eroille tai yhtäläisyyksille, sitä selvemmin ihmisen näköjärjestelmän osuus on huomioitava tutkimuksissa. Pelkillä tietoteknisillä tutkimusmenetelmillä ei voida tutkia ihmisen ja koneiden luokittelujärjestelmien yhtäläisyyksiä tai eroavaisuuksia. Sen selvittäminen, miksi ihmisen ja koneen luokittelujärjestelmät eroavat

Ilari Hautamäki, *Arrangement 2* (2018), gesso, akryyli ja öljy kankaalle, 90 × 65 cm

”Pelkillä tietoteknisillä tutkimusmenetelmillä ei voida tutkia ihmisen ja koneiden luokittelujärjestelmien yhtäläisyyksiä tai eroavaisuuksia.”

toisistaan, vaatiikin väistämättä lähtökohtaisesti monitieteisempää tutkimusotetta.

Adversariaalit ja mustat laatikot

Tutkijat spekuloivat myös kysymyksellä, missä määrin adversariaaleihin liittyy selityksellisiä tai tulkinnallisia ”mustia laatikoita”[21]. Eräissä artikkeleissa todetaan, että ihmisen tulisi voida ”nähdä” adversariaalit kuin kone, jotta koneen luokitukset olisivat täysin ihmisen ”tulkittavissa”[22]. Koska ihminen ei lähtökohtaisesti voi ”nähdä” adversariaaleja kuin kone, adversariaalit tarjoavat väitetysti esimerkin periaatteellisesta ja tulkinnallisesta mustasta laatikosta. Argumentissa tulkittavuus oletetaan ennakolta lähinnä ”fenomenaaliseksi tulkittavuudeksi”. Se liikkuu monien mielen- ja kielenfilosofian klassikkoargumenttien maaperällä. Esimerkiksi väitteet subjektiivisen kokemuksen kvalioista (Nagel), elämänmuotojen välisestä kuilusta (myöhäis-Wittgenstein) ja vaikkapa kiinalaisen huoneen ajatuskoe (Searle) operoivat samassa käsitteellisessä maastossa.

Kognitiontutkimuksen näkökulmasta fenomenaalisen tulkinnan käsite ei kuitenkaan ole kovin hyödyllinen. Esimerkiksi kissan mustavalkoisen värinäköjärjestelmän ”tulkitseminen” ei edellytä ensimmäisen persoonan kvalitatiivista näkökulmaa kissan representaatioavaruuteen. Representaatioavaruuksien tutkittavuudelle riittää, että kissan näköjärjestelmän representaatiosisältöjä voidaan tarkastella abstrahoituna ja idealisoituna mallina itse näköjärjestelmästä. ”Tulkitseminen” tässä mielessä edellyttää lähinnä jotain eksaktia metriikkaa, jolla representaatioavaruus voidaan kiinnittää ja siten tehdä analysoitaviksi.

Täsmälleen sama pätee koneoppimissovelluksiin: ihmisen ei tarvitse nähdä kuin koneet, jotta niitä voidaan ”tulkita”. Riittää, että sovelluksien representaatioavaruutta voidaan mallintaa. Toki tällainen mallintaminen on aina osittain epävarmaa: mallit ovat aina havaintojen suhteen alideterminoituja. Vaikka käytössä olisi täydellisesti kerätty neuraalinen data kissojen näköaivokuoren soluista ja hienoimmat mahdolliset niihin perustuvat formaalit mallit kissojen näköaistin reseptiivisistä kentistä, lopputuloksena syntyvät mallit olisivat silti ”vain” objektivoituja, idealisoituja ja abstrahoituja laskennallisia malleja kissan *mahdollisesta* representaatioavaruudesta.

Sitä, missä määrin nämä mallit vastaavat kissan kokemusmaailman todellista representaatioavaruutta,

ei tietenkään voida todentaa nykyisillä menetelmillä. Vastoin monia filosofisia intuitioita, erilaisilla matemaattisilla menetelmillä voidaan kuitenkin arvioida tällaisten mallien paikkansapitävyyden todennäköisyyttä. Siksi nämä mallit ovat *arvioitavissa olevia*, *perusteltuja arvauksia* kissan representaatioista, eivät ”vain arvauksia”.

Olennaista on, että nämä mallit ovat ”tulkittavissa” ei-fenomenaalisesti. ”Tulkittavuus” voidaan nimittäin määritellä myös ”käännettävyytenä” formaalikieleltä toiselle, jolloin ei tarvitse ottaa kantaa kvalitatiivisista, subjektiivisista kokemuksista nouseviin hankaliin, lähinnä mielenfilosofian piiriin kuuluviin kysymyksiin. Tämä ”tulkittavuuden” käsite on filosofeille tuttu lähinnä logiikasta ja matematiikasta, joissa sen edellytyksiä ja reunaehtoja on analysoitu varsin kattavasti muun muassa malliteoreettisesta näkökulmasta[23].

Adversariaalit ja selitettävyys

Adversariaalien tapauksessa kenties hankalin musta laatikko ei siis ole niinkään tulkittavuus tai läpinäkyvyyden ongelma[24]. Vaikein on se, että tutkijat eivät vielä osaa yksilöidä tarkasti, mitä koneoppimismenetelmien hauraus eli sensitiivisyys adversariaalien kaltaisille, datan helposti yleistyville piirteille lopulta matemaattisessa tai algoritmisessa mielessä tarkoittaa.

Tämä on lähinnä *selitettävyyden* ongelma. Siinä missä tulkittavuus viittaa järjestelmien kuvaamiseen tai hahmottamiseen ja läpinäkyvyys sen simuloimiseen askel askeleelta, selitettävyys vastaa kysymyksiin: ”miksi” ja ”miten”. Nykytieteenfilosofit korostavat, että aidot selitykset vastaavat *kontrastiivisiin* ”miksi”- tai ”miten”-kysymyksiin (”miksi järjestelmä luokittelee pandat gibboneiksi eikä pandoiksi”), eivät yksinomaan ”miksi”-kysymyksiin (”miksi järjestelmä luokittelee pandat gibboneiksi”). Yleensä ajatellaan, että selitys yksilöi riittävällä tarkkuudella ne olennaiset kausaaliset, konstitutiiviset tai formaaliset riippuvuudet, jotka selittävät, miksi selitettävä ilmiö on juuri A eikä B. Selityksien tulee siis poimia tietyt riippuvuudet selittävien tekijöiden ja selitettävien ilmiöiden välillä[25].

Selitykset kuitenkin edellyttävät, että myös selitettävä ilmiö osataan kuvata riittävän tarkasti. Jos ei tarkkaan osata kuvata, mikä itse selitettävä ilmiö on, ei ole yllättävää, ettei osata myöskään yksilöidä niitä muuttujia, joiden väliltä selittäviä riippuvuuksia ehkä voisi (tai pitäisi) alkaa etsiä[26]. Adversariaaleissa näyttäisikin olevan ongelmana, ettei itse selitettävää ilmiötä täysin hahmoteta.

Usein kokeellisessa tutkimuksessa tällaisessa tilanteessa sekä selitettävää ilmiötä että selityksellisiä riippuvuuksia aletaan etsiä systemaattisilla kokeellisten tutkimusten sarjoilla, joissa manipuloimalla muuttujia etsitään niiden välisiä riippuvuuksia. Myös tietojenkäsittelytieteen puolella kehitetään kiivaasti kvasi-kokeellisia menetelmiä, joiden avulla valittuja yksiköitä – kuten neuroneita tai piirrekarttoja[27] – ”manipuloimalla” voitaisiin havainnoida tarkasti ja selvittää, millaisia seurauksia interventioilla on järjestelmän käyttäytymiseen[28]. Tällaisten dissektio- tai interventiomenetelmien kehittäminen erityisesti DNN-pohjaisten koneoppimisjärjestelmien tutkimiseksi on monesti kuitenkin haastavaa – varsinkin, jos ei edes tarkkaan tiedetä, mitä yritetään tutkia.

On myös huomattava, että haurauden tapaisten ilmiöiden selittämisessä kaikkia selityksellisiä tarpeita ei välttämättä kyetä täyttämään yksilöimällä selittäviä riippuvuuksia pelkästään kausaalisesti, siis käyttämällä sellaisia dissektio- tai manipulaatiomenetelmiä, jotka operoivat niin sanotun algoritmisen laskennan tasolla[29]. Selityksissä joudutaan oletettavasti myös vastaamaan kysymyksiin, joiden kohteena on itse laskennallinen tehtävä: ”miksi järjestelmä laskee häiritsevien piirteiden vuoksi nimenomaan tätä optimointiongelmaa eikä jotakin toista?”

Näihin kysymyksiin ei voida vastata ainoastaan manipuloimalla verkon syötettä tai sen sisäisiä rakenneosia (esimerkiksi neuroneita) ja tarkkailemalla manipuloinnin kausaalisia vaikutuksia. Kausaalinen manipulaatio ja algoritmisen tason riippuvuuksien yksilöinti tarjoavat vastauksia vain siihen, miten ja miksi järjestelmä laskee tiettyä ratkaisua askel askeleelta. Vastausta ei kuitenkaan saada siihen, miksi verkon laskennallinen tehtävä on sen omasta näkökulmasta X tai ei-X. Sen sijaan laskennallisia tehtäviä (”miksi verkko poimii juuri tuon piirteen salientiksi eikä tuota toista?”) jouduttaneen selittämään myös matemaattisesti yksilöimällä niitä formaaleja riippuvuuksia, joiden vuoksi järjestelmä toimii niin kuin se toimii, eikä vain kuvaamalla mallin suorittamaa konkreettista laskentaa.

Lopuksi

Tässä kuvattuihin kysymyksiin vastaaminen vaatii teoreettista, käsitteellistä ja filosofista työtä, jota insinöörivetoisilla tietojenkäsittelytieteen alueilla ei tyypillisesti tehdä. Adversariaalit tarjoavatkin yhden tavan perustella nimenomaan käsitteellisen ja teoreettisen perustutkimuksen tärkeyttä. Toisaalta adversariaalit korostavat, että osa nykyisten koneoppimissovellusten ongelmista on perustavalla tavalla ihmisen ja koneen välisessä kognitiivisessa vuorovaikutuksessa. Tämän vuorovaikutuksen ymmärtäminen vaatii aidosti monitieteistä ja vertailevaa tutkimusta, sillä pelkillä tietoteknisillä tutkimusmenetelmillä ei voida analysoida ihmisten ja koneiden luokittelujärjestelmien kognitiivisia yhtäläisyyksiä tai eroavaisuuksia. Vertailevaan lähestymistapaan sisältyy kuitenkin monia käsitteellisiä, metodologisia ja teoreettisia ongelmia, joita ei voida luontevasti ratkoa yksin kokeellisilla tai mallinnusmenetelmiä. Sen sijaan ne vaatinevat myös käsitteellistä, eri tutkimusalojen teorioista ja menetelmistä ammentavaa teoreettista perustutkimusta. Siksi adversariaalien tutkimuksessa insinöörinkin on kenties pyydettävä apua kognitiontutkijalta – ja ehkä jopa filosofilta.[30]

Viitteet

1 Ian Goodfellow'n (2014) työtovereineen kehittämät GANit koostuvat kahdesta toisiaan vastaan kilpailevasta verkosta. Toinen verkoista tuottaa syöteaineiston – esimerkiksi julkisuuden henkilöiden kuvien – pohjalta uusia ”epäaitoja” kuvia, ja toinen verkoista arvioi, kuuluuko uusi kuva alkuperäiseen syöteaineistoon vai ei.
2 Ks. mm. Karras ym. 2018.
3 Goodfellow ym. 2014; Ilyas ym. 2019.
4 Goodfellow ym. 2014.
5 Ilyas ym. 2019.
6 Sama.
7 Szegedy ym. 2014.
8 Eri menetelmät voidaan karkeasti jakaa kahteen pääryhmään: kohdistettuihin ja kohdistamattomiin häirintämenetelmiin.
9 Goodfellow ym. 2014.
10 Brown ym. 2018.
11 Papernot 2017.
12 Goodfellow ym. 2014.
13 Liu 2016.
14 Ilyas ym. 2019.
15 Sama.
16 Sama.
17 Brown ym. 2018.
18 Ihmisen näköjärjestelmän toiminnassa on myös paljon lajityypillisiä, evolutiivisesti kehittyneitä rakenteita ja periaatteita, jotka esiohjaavat näköjärjestelmän toimintaa.
19 Han ym. 2019.
20 Hiljattain Han ym. (2019) julkaisivat tutkimuksen, jossa väitettiin, että fMRI-datan perusteella olisi löydetty tiettyjä samankaltaisuuksia aivokuoren neuraalisten representaatioiden ja DNN-pohjaisten koneoppimissovellusten piirteiden representaatioiden välillä. Tutkimus on nähdäkseni ongelmallinen, sillä siinä käytetty fMRI-data mittaa pelkästään neuraalisia aktivaatioita, ei representaatioita. Lisäksi tutkimuksessa käytetty samankaltaisuusmetriikka on kyseenalainen.
21 Termistä ”musta laatikko” on viime vuosina tullut sateenvarjokäsite, jonka alle kootaan – usein suhteellisen löysillä perusteilla – joukko erilaisia ominaisuuksia, piirteitä ja ongelmia. Sillä on viitattu esimerkiksi ”läpinäkymättömyyteen” (Marcus 2018), ”tulkitsemattomuuteen” (Lipton 2016), ja vaihtelevasti muotoiltuihin ”ennustamattomuuteen” ja ”selitettävyyden” tai ”ymmärrettävyyden” puutteeseen.
22 Mm. Ilyas ym. 2019.
23 Insinöörivetoisilla tietojenkäsittelytieteen aloilla ”tulkittavuus” redusoituu usein pitkälti kysymykseksi, missä määrin järjestelmän toiminta voidaan esimerkiksi visualisoida tai hahmottaa erilaisten tekniikoiden avulla. Yksi esimerkki tällaisista menetelmistä ovat menetelmät, joiden avulla pyritään eristämään ja visualisoimaan vaikkapa GANien sisältämiä ”representaatioita” (ks. esim. Bau ym. 2018).
24 ”Läpinäkyvyydellä” viitataan usein siihen, missä määrin jonkin mallin tai neuroverkon toimintaa voidaan simuloida tai hahmottaa. Malli on (täydellisen) läpinäkyvä, jos ihminen pystyy syöteaineiston ja mallin parametrien avulla käymään läpi askel askeleelta mallin suorittaman laskennan siten, että lopputuloksena on sama vaste kuin minkä malli tuottaa.
25 Woodward 2003; Craver 2007. On huomattava, että selittäminen ja ennustaminen ovat eri asioita. Craver (2014) kiteyttää eron esimerkin avulla: Siitä, että pelikentällä soitetaan Yhdysvaltojen kansallislaulu, voidaan ennustaa, että amerikkalainen jalkapallopeli alkaa, mutta laulu ei selitä pelin alkua. Siinä missä ennusteille yleensä riittävät korrelaatiot, aidot selitykset vaativat tietoa selityksien ja selittävien asioiden välisistä riippuvuuksista. Pelkästään se, että voidaan ennustaa todennäköisyys, jolla kuvantunnistusjärjestelmä luokittelee pandan kuvan pandan kuvaksi, ei vielä selitä, miksi järjestelmä luokittelee pandan kuvan pandaksi.
26 Tilannetta monimutkaistaa se, että adversariaaleja voidaan tuottaa useilla erilaisilla menetelmillä ja että vielä ei ole olemassa taksonomiaa siitä, ovatko ne yhtä vai useaa tyyppiä.
27 Esimerkiksi Bau ym. (2018) kuvaavat menetelmiä, joilla GANeja voidaan ”dissektoida” manipuloimalla neuroneita.
28 Ks. esim. Bau ym. 2018.
29 David Marrin (1982) kuuluisan viitekehyksen mukaan tiettyjä laskennallisia järjestelmiä voidaan tarkastella kolmesta näkökulmasta: informaationprossessointijärjestelmien, algoritmien ja konkreettisen toteutuksen näkökulmasta.
30 Kiitokset Jami Pekkaselle, Sami Katteluksselle, Tero Hakalalle, Okko Räsäselle ja Jaakko Lehtiselle.

Kirjallisuus

Bau, David ym., GAN Dissection: Visualizing and Understanding Generative Adversarial Networks. 2018. Verkossa: arxiv.org/pdf/1811.10597.pdf

Brown, Tom ym., Adversarial Patch. *Computer Vision and Pattern Recognition.* 2018. Verkossa: arxiv.org/abs/1712.09665v2

Craver, Carl, Constitutive Explanatory Relevance. *Journal of Philosophical Research.* Vol. 32, No. 1, 2007, 3–20.

Craver, Carl, The Ontic Account of Scientific Explanation. Teoksessa *Explanation in the Special Sciences. The Case of Biology and History.* Toim. M. I. Kaiser ym. Springer, Dordrecht 2014, 27–52.

Goodfellow, Ian ym. Generative Adversarial Nets. *Proceedings of the International Conference on Neural Information Processing Systems*, 2014, 2672–2680.

Han, Chihye ym. Representation of White- and Black-box Adversarial Examples in Deep Neural Networks and Humans: A Functional Magnetic Resonance Imaging Study. 2019. Verkossa: arxiv.org/pdf/1905.02422.pdf

Ilyas, Andrew ym., Adversarial Examples Are Not Bugs, They Are Features. 2019. Verkossa: arxiv.org/abs/1905.02175

Karras, Tero ym., Progressive Growing of GANs for Improved Quality, Stability, and Variation. ICLR 2018. Verkossa: arxiv.org/abs/1710.10196v3

Lipton, Zachary, The Mythos of Model Interpretability. 2016. Verkossa: arxiv.org/abs/1606.03490

Liu, Yanpei ym, Delving into Transferable Adversarial Examples and Black-Box Attacks. 2016. Verkossa: http://arxiv.org/abs/1611.02770

Marcus, Gary, Deep Learning: A Critical Appraisal. 2018. Verkossa: arxiv.org/abs/1801.00631

Marr, David, *Vision. A Computational Investigation into the Human Representation of Visual Information*. W.H. Freeman, San Francisco 1982.

Papernot, Nicolas ym., Practical Black-Box Attacks against Machine Learning. *Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security. Abu Dhabi, UAE*. 2017. Verkossa: arxiv.org/abs/1602.02697

Szegedy, Christian ym., Intriguing Properties of Neural Networks. ICLR. 2014. Verkossa: arxiv.org/abs/1312.6199

Woodward, Jim, *Making Things Happen. A Theory of Causal Explanation*. Oxford University Press, Oxford 2003.

Ilari Hautamäki, *Arrangement 5* (2018) gesso, akryyli ja öljy kankaalle, 90 × 65 cm