

HUOM! Tämä on alkuperäisen artikkelin rinnakkaistallenne. Rinnakkaistallenne saattaa erota alkuperäisestä sivutukseltaan ja painoasultaan.

Käytä viittauksessa alkuperäistä lähdettä:

Aaltonen, T., Kauttonen, J., Aunimo, L. & Salahub, J. (01.06.2023) Avoin data tekoälysovellusten mahdollistajana – konenäkö meripelastuksen apuna. eSignals PRO.
<http://urn.fi/URN:NBN:fi-fe2023060151250>

PLEASE NOTE! This is an electronic self-archived version of the original article. This reprint may differ from the original in pagination and typographic detail.

Please cite the original version:

Aaltonen, T., Kauttonen, J., Aunimo, L. & Salahub, J. (01.06.2023) Avoin data tekoälysovellusten mahdollistajana – konenäkö meripelastuksen apuna. eSignals PRO.
<http://urn.fi/URN:NBN:fi-fe2023060151250>



Copyright: © 2023 by the authors and Haaga-Helia University of Applied Sciences. Licensed under the terms and conditions of the Creative Commons Attribution (CC BY NC SA) license (<https://creativecommons.org/licenses/by-nc-sa/4.0/>).

Avoim data tekoälysovellusten mahdollistajana – konenäkö meripelastuksen apuna

Toni Aaltonen, Janne Kauttonen, Lili Aunimo & Jeffrey Salahub

Tekoälyn ja koneoppimisen hyödyntäminen yleistyvät vauhdilla kaikilla aloilla. Tätä vauhdittavat algoritmien kehitys, yleinen prosessoritehon kasvu tietokoneissa ja äylaitteissa, sekä laskentakapasiteetin saatavuus pilvipalveluissa (esim. Microsoft Azure, Amazon Web Services ja Google Cloud).

Pullonkaulaksi tekoälyn ja koneoppimisen hyödyntämisessä muodostuukin usein tarkoitukseen soveltuvan avoimen datan puute. Tämä on merkittävä ongelma erityisesti sellaisilla sovellusalueilla, joissa datan kerääminen vaatii erityisosaamista ja ammattilaisten huomattavaa työpanosta. Esimerkkinä tästä on meripelastustoiminta, jossa uutta teknologiaa ei ole juurikaan hyödynnetty.

Nopea paikantaminen ja pelastaminen ovat ratkaisevia hukkuvan selviytymismahdollisuuksien parantamiseksi. Suomessa vedenalaiset pelastuskäytännöt eivät ole vuosien varrella juurikaan muuttuneet, vaikka autonomiset vedenalaiset robotit ja kuluttajataso kaikuluotaimet ovat kehittyneet. Merkittävä haaste vedenalaisen pelastustoiminnan tekoälyratkaisujen kehittämisessä on kaikuluotaimen kuvia sisältävien julkisten datajoukkojen puute.

Osana AI Forum -projektia SAMK kehitti kaikkien hyödynnettävissä olevan avoimen kaikuluotaindatajoukon, joka auttaa ihmisten tunnistamista kuluttajataso kaikuluotainkuvista (Aaltonen 2022; 2023). Se mahdollistaa uusien tekoälymallien kehittämisen meripelastukseen ja erityisesti Suomen sameisiin vesiin.

Datan etsintä avoimia tietokantoja hyödyntämällä

Tutkimuksen ensimmäisessä vaiheessa kartoitettiin millaista avointa dataa on saatavilla kaikuluotaukseen liittyen. Dataa lähdettiin etsimään mm. Kaggle-, Papers with Code- ja Google Datasetsearch -palveluissa, jotka kuitenkin tuottivat vain vähän tarkoitukseen soveltuvia tuloksia.

Kaggle tietokanta tuotti 43 osumaa, jotka sisälsivät enimmäkseen kopioita vanhasta tietokannasta, joka on kallioiden ja miinojen kaikuluotainsignaaleja. Niin ikään Google Datasetsearch -haku tuotti yli 100 osumaa, joista monet olivat jälleen kopioita kallio- ja miinatietokannasta. Haun kautta löytyi kuitenkin joitakin arvokkaita datatietueita, kuten erilaisia yhdysvaltalaisen virastotahojen (NOAA ja NCFMF) dataa, jotka tarjosivat raakoja viisto-kaikuluotaindataa valtamerialta, järvistä ja riutoista. Papers with Code -verkkosivut listaavat tutkimuspapereita ja niiden dataa ja sen haulla löydettiin kaksi datakokonaisuutta. Toinen näistä käsitteli meressä esiintyvää yleistä roskaa, ja toinen on

CFC-datajoukko, joka sisältää yli 1 500 videota ja yli 500 000 annotoitua kuvaa eri kalalajeista.

Näiden hakutulosten lisäksi joitakin datalähteitä löydettiin lukemalla kaikuluotaindatan tekoälymenetelmiin liittyviä artikkeleita ja seuraamalla näiden artikkelien viitteitä. Näistä merkittävimmät olivat kalojen luokitteluun ja kalastuksen arviointiin tarkoitettu vedenalainen kuvakokoelma, joka sisältää 524 näytettä. Toinen datakokoelma oli ARACATI 2017, jossa on sekä optisia ilmakuvia että akustisia vedenalaisia kuvia eri modaliteettien vertailua varten.

Lähimpänä tutkimuksen tarkoitusta oleva datajoukko on kuitenkin yli 9000 kaikuluotainkuvaa sisältävä kokoelma, joka on otettu Tritech Gemini 1200iik - kaikuluotaimella, joka on selvästi kuluttajaluokan laitteita järeämpi ja kalliimpi laite. Tämä datajoukko sisältää raakadataa kaikuluotaimen kuvista, joissa on 10 eri kategoriala. Mikä tärkeintä, se sisältää myös kuvia ihmisen kaltaisista nukeista. Tämä data ei kuitenkaan vastaa alkuperäistä tarkoitusta laitteiston ja aitojen ihmiskuvien suhteen.

Näiden hakujen perusteella voidaan todeta, että suurin osa kaikuluotaimen kuvia sisältävistä avoimista datajoukoista on kerätty selvästi kuluttajaluokkaa paremmilla välineillä. Datan keräämisen käytettyjen laitteiden ilmoitetuissa tiedoissa oli myös huomattavia puutteita, joka hankaloittaa datan analyysiä ja vertailua muihin laitteisiin. Huolellisesta hakemisesta huolimatta ei lopulta löydetty yhtään julkista datajoukkoa, joka sisältäisi kuluttajaluokan kaikuluotaimia ja kuvia meren pohjassa olevista ihmisistä, joka on mallien kehitykselle oleellista. Edullisten kaikuluotaimien data on avainasemassa, jotta voidaan kehittää ratkaisuja, joita voidaan käyttää pienissä pelastuslaitoksissa joko veneissä tai osana edullista vedenalaista robottia.

Oman datajoukon kerääminen kuluttajatasen laitteilla ja datan käsittely

Oman aineiston keräämiseen käytettiin Garmin 8400 XSV kuluttajaluokan kaikuluotainta, joka kiinnitettiin Satakunnan pelastuslaitoksen veneeseen. Data kerättiin ROSBAG tiedostomuotoon liittämällä kaikuluotain kannettavan tietokoneeseen HDMI-USB 3.0 liitännällä.

Aineiston keräämiseen käytettiin noin 100 henkilötyötuntia, joista pelastussukeltajien osuus oli noin 30 tuntia. Kerättyihin tietoihin kuuluivat GPS-sijainti, nopeus pohjan suhteen, suunta ja syvyys. Tarkemmat tiedot datankeräyksestä ja laitteistosta voi lukea lähteistä Aaltonen 2022 ja 2023. Dataa mitattiin kaksi päivää Rauman edustalla kohteissa Syväraumanlahti ja Maanpäännokka. Sukeltajat käyttivät märkäpukuja ja kasvomaskeja ja heitä kuvattiin useissa eri asennoissa meren pohjassa. Happsäiliöt pyrittiin asettelemaan niin etteivät ne näy kuvissa.

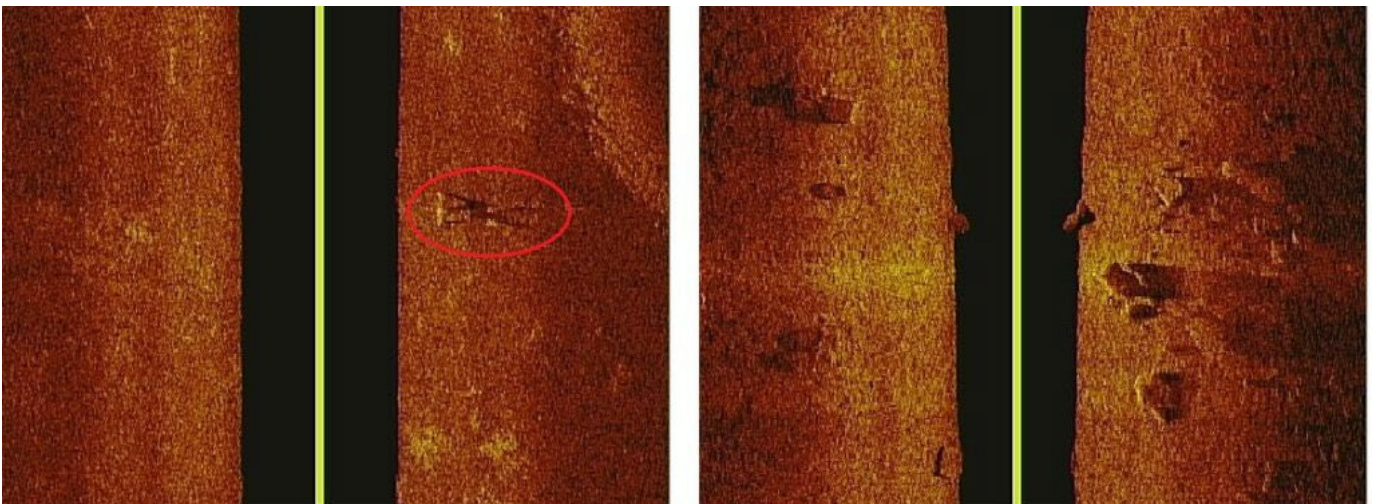
Mitattu raakadata käytiin läpi useassa vaiheessa sekä automaattisesti, että manuaalisesti. Kuvat luokiteltiin kahteen kategoriaan: Ihmisen sisältävät kuvat ja muut kuvat. Kuvajoukkoon valittiin paljon sellaisia kuvia, joissa näkyy ihmisen sijaan tai lisäksi muita kohteita, kuten kiviä, pohjan muotoja tai muita objekteja. Tämä lisää opetettavan tekoälyjärjestelmän virheensietoa.

Kuvien resoluutio on 100×70 pikseliä. Lopullinen aineisto sisältää tarkoituksella myös huomattavan heikkolaatuisia kuvia, joista on silmämääräisesti lähes mahdoton erottaa ihmistä. Tunnistus voitiin kuitenkin tehdä metatiedon perusteella. Näiden kuvien sisällyttäminen saattaa heikentää kehitettävien tekoälysovellusten tilastollista havaintotarkkuutta, mutta tarkoitus on antaa mahdollisimman realistinen kuva todellisesta havaintoprosentista kuluttajaluokan kaikuluotaimilla.

Koostamamme kaikuluotaindata on luotettavaa sillä kuvien luokittelu tehtiin huolellisesti ja virheettömyys varmistettiin. Datan ulkoista luotettavuutta oli kuitenkin vaikeampi määrittellä, koska kuvissa oli huomattavan paljon vaihtelua liittyen kuvien sisältöön ja mittausolosuhteisiin. Tämän mittaamiseksi data kannatti jakaa kahteen osaan siten, että yhden mittauspäivän dataa käytetään mallien opetuksessa ja toisen mittauspäivän dataa mallien testauksessa. Tämä on luotettavin tapa selvittää miten opetetavat mallit yleistyvät.

Ensimmäisenä päivänä kerätty lopullinen harjoittelu- ja validointidata koostuu 205 kuvasta, joissa on henkilö merenpohjassa, ja 249 kuvasta, joissa on muita kohteita. Toisen päivän testiaineisto sisältää 125 kuvaa ihmisestä ja 84 kuvaa, joissa on muita esineitä tai pelkkää merenpohjaa. Yhteensä kuvia on 695, joista 331 kuvassa on ihmisiä ja 364 kuvassa ei ole ihmisiä. Data on saatavissa [täältä](#). Tietoaineistomme on valmis käytettäväksi sellaisenaan esim. neuroverkkojen kanssa ilman esikäsitteilyä, ja se on yhteensopiva kategoristen mallien, kuten densenet ja resnets, kanssa.

Yleisesti ottaen kaikuluotaindata on haastavaa analysoida paitsi ihmiselle, myös koneelle. Kuvat ovat matalaresoluutiosta ja ihmisen muoto hyvin vaihtelevaa, riippuen ihmisen koosta, vaatetuksesta sekä asennosta. Veden syvyys ja pohjan muoto ja materiaali vaikuttavat vahvasti kuvaan. Kuvassa 1 on kaksi esimerkkiä aineistosta. Vasemmassa kuvassa on ihminen ja oikealla pelkkiä kiviä.



Kuva 1. Kaksi esimerkkikuvaa uudesta kaikuluotaindatasetistä. Vasemmassa kuvasta löytyy ihminen merkattuna punaisella ympyrällä (lisätty, ei mukana datassa), kun taas oikeanpuolimmaisessa kuvassa on pelkkiä kiviä. Kaikukuvassa vene kulkee kuvassa keskellä (keltainen viiva) ja pohja erottuu kummallakin puolella ruskealla sävyllä. Näiden välissä oleva musta alue on välivettä.

Uusi data pelastustoiminnan kehityksen apuna, alustavat mallinnukset ja seuraavat askeleet

Joka vuosi maailmassa hukkuu yli puoli miljoonaa ihmistä, joista valtaosa on lapsia ja nuoria aikuisia. Yksi osa ratkaisua on uuden teknologian käyttöönotto. Tutkimuksessa koostettu uusi kaikuluotaindatajoukko auttaa kehittämään uusia tekoälymalleja, jotka osaavat tunnistaa onko vedessä oleva hahmo mahdollisesti ihminen vai ei. Tämä on ensiarvoisen tärkeää informaatiota, kun vedestä etsitään ihmistä esimerkiksi hengenpelastustilanteessa, jolloin aikaa on hyvin rajallisesti käytössä.

On huomattava, että datajoukoissa on rajoituksia. Sen koko on verrattain pieni, se sisältää tietoja vain kahdelta märkäpukuja käyttävältä aikuiselta ja se rajoittuu suhteellisen halpoihin sivupyyhkäiseviin kaikuluotaimiin. Se on kuitenkin lajissaan ensimmäinen ja tarjoaa lähtökohdan jatkotutkimuksiin ja sovellusten kehitykseen. Datajoukko voi auttaa pelastushenkilöstöä kouluttamaan sivuttaisluotaimien käyttöä myös ilman konenäköä.

Olemme testanneet datajoukkoa alustavasti opettamalla muutamia erilaisia Tensorflow - pohjaisia syviä neuroverkkoja. Parhaat koulutetut verkot saavuttivat 97,6 % luokittelutarkkuuden opetusdatalla ja 84,2 % testidatalla (Aaltonen 2022). Tulokset ovat selvästi 50 % sattumaa parempia, joten voimme todeta että data sisältää riittävästi informaatiota ihmisen tunnistamista varten. Parhaat tulokset saavutettiin esiopetetulla konenäkömallilla siirto-oppimista hyödyntämällä ja data lavennettiin (data augmentation) erilaisia muunnoksilla, kuten kierroilla, skaalauksilla, kohinan lisäämisellä ja valaistuksen muutoksilla.

Seuraavaksi keskitymme neuroverkkomallien jatkokehitykseen ja mallien julkaisemiseen. Mallien tulosten parantaminen seuraavalle tasolle vaatii lisää opetusdataa. Tätä voidaan mahdollisesti edistää perinteisen datan laventamisen ohella luomalla synteettisiä kuvia generatiivisilla tekoälymalleilla, jotka jäljittelevät mahdollisimman hyvin alkuperäisen datan ominaisuuksia. Tästä on erittäin hyviä kokemuksia erityisesti lääketieteellisen kuvantamisdatan tapauksessa (Prezja ym. 2022).

Avoimella datalla on tärkeä merkitys tekoälymallien kehityksessä

Avointen datajoukkojen julkaiseminen osana tutkimusta on yhä tärkeämpää nykyisellä tekoälysovellusten aikakaudella. AI Forumin kaltainen tutkimusprojekti mahdollistaa uusien tutkimustulosten lisäksi myös avoimen datan julkaisemisen. Näin voidaan varmistaa tutkimuksen vaikuttavuus, eikä data jää pelkästään yksittäisten yritysten, tutkijoiden tai muiden toimijoiden hyödynnettäväksi.

Kunhan datan lisensointi on kunnossa, voi datajoukkoa hyödyntää laajasti paitsi uudessa tutkimuksessa, myös liiketoiminnassa ja julkisissa organisaatioissa. Esimerkiksi nyt julkaistu kuvadata on sen päälle rakennettavan ihmisiä vedestä löytävän konenäkösovelluksen ydin. Pelastustoimi voi vapaasti hyödyntää dataa ja sovellusta. Koneoppimisen lisäksi datasta voi olla hyötyä myös silloin, kun ihmissilmää harjaannutetaan tunnistamaan henkilö vedestä. Pelastustoimen lisäksi robotteja kehittävät yritykset voivat vapaasti hyödyntää dataa. Ihmisen vedestä tunnistava konenäkösovellus on olennainen osa esimerkiksi erilaisia pelastusrobotteja (Feuilherade 2017).

Toinen syy avoimen datan julkaisemiselle tutkimusprojekteissa on tieteen läpinäkyvyyden ja toistettavuuden varmistaminen (Koivisto 2022). Esimerkiksi tietojenkäsittelytieteen alalla useat tieteelliset julkaisukanavat suosittelevat, että uusien algoritmien avoimen julkaisemisen lisäksi myös algoritmien testaamisessa käytetty data olisi julkisesti saatavilla. Näin kuka tahansa voi toistaa tutkimuksen ja alkaa kehittää sitä hyödyntäen uutta ja vielä parempaa algoritmia.

Tutkijoiden, yritysten ja julkisten organisaatioiden lisäksi avoin data on koulutuksessa tärkeä aineisto. Kun opiskelijoiden koneoppimisprojekteja varten on saatavilla mielenkiintoisia ja hyvin dokumentoituja datajoukkoja, löytyy jokaiselle helpommin omaa osaamista ja kiinnostuksen kohdetta vastaava tekoälyprojektin aihe. Tämän lisäksi, kun opiskelijat koulutuksen aikana perehdytetään avoimen datan hyötyihin, erilaisiin avoimen datan varastoihin ja katalogeihin, standardeihin ja lisensointikäytäntöihin, on tästä heille mitä todennäköisemmin hyötyä työelämässä (Aunimo, R. Kauppinen, H. Kekkonen 2020). Avoin data on arvokas resurssi yrityksille niin tuotekehityksessä ja kokeiluissa kuin tuotannossa olevissa tuotteissakin.

Lähteet

Aaltonen, T. 2022. [Ihmisen tunnistaminen vedenpohjasta käyttäen kuluttajaluokan kaikuluotaimen datalla opetettuja syviä neuroverkkoja](#). SAMK.

Aaltonen, T. 2023. Consumer class side scanning sonar dataset for human detection. 46th ICT and Electronics conference. MIPRO 2023.

Prezja, F., Paloneva, J., Pölönen, I., Niinimäki, E. & Äyrämö, S. 2022. DeepFake knee osteoarthritis X-rays from generative adversarial neural networks deceive medical experts and offer augmentation potential to automatic classification. Scientific Reports, 12(1), 18573.

Feuilherade, P. 2017. [Robotit rientävät apuun](#). Sesko.

Koivisto, S. 2022. [Avoin tiede ja tutkimus – uhka vai mahdollisuus?](#) eSignals, Haaga-Helia.

Aunimo, R. Kauppinen, H. Kekkonen, H. 2020. [Open Data Course Projects in Business Intelligence](#). EDULEARN20 Proceedings, pp. 4263-4269.