TinyML – koneoppimista taskukoossa

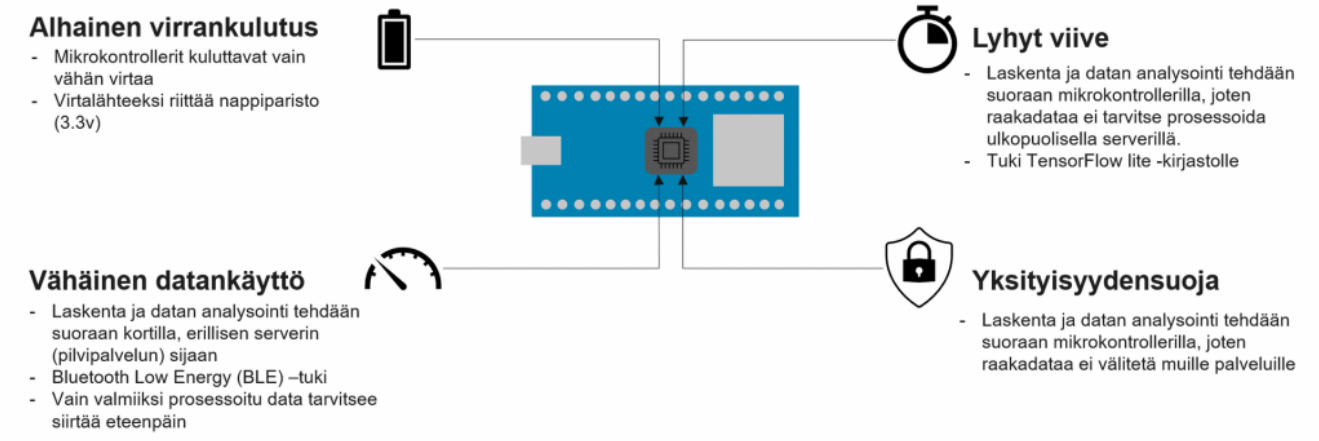
22.2.2021

Teollisuus 4.0 eli (Industrie 4.0) on Saksan hallituksen kansallinen hanke, joka aloitetiin vuonna 2012. Sen tavoitteena on teollisuuden kilpailukyvyn vahvistaminen digitalisaation keinoin. Tulevaisuuden tehtaissa koneet ja laitteet kommunikoivat keskenään ja tekevät keräämänsä datan perusteella itsenäistä optimointia (Collin & Saarelainen 2016, Luku 3). Kyberfyysiset järjestelmät yhdistävät ihmiset, tuotteet sekä palvelut. Kun yritysten myytävät tuotteet, palvelut sekä niihin liittyvät liiketoimintaprosessit kytketään verkkoon, syntyy Teollinen Internet (Collin & Saarelainen 2016, Luku 2). Teollisen internetin kehityksen yhtenä ajurina on ollut anturien hinnan romahdus, parempi suorituskyky sekä alhaisempi virrankulutus (Collin & Saarelainen 2016, Luku 4).

Aikaisemmin anturit ovat lähettäneet mittaamansa datan ulkoiselle serverille tai pilvipalveluun prosessoitavaksi. Nykyään on mahdollista koneoppimisen ajaminen suoraan anturissa, jolloin ulkoista prosessointia ei erikseen tarvita ja anturi pystyy itsenäisesti suorittamaan sille opetettua koneoppimisalgoritmia. Tekoälyn avulla nämä anturit saavat suoritettua itsenäisesti entistä monipuolisempia mittaustehtäviä ja pystyvät vähentämään virrankulutusta välittämällä vain tärkeäksi määritellyn esiprosessoidun informaation eteenpäin.

Mitä on TinyML?

Termillä tarkoitetaan koneoppimisen suorittamista sulautetuissa järjestelmissä esimerkiksi mikrokontrollereissa, joiden virrankulutus on vähemmän kuin 1 mW. Laiteiden saatavuuden vuoksi termiä on jouduttu laajentamaan myös laitteisiin, joiden virrankulutus on maksimissaan kymmeniä milliwatteja. (Warden 2020). Laitteen tehonkulutus on tärkeässä roolissa, sillä se mahdollistaa pidemmän valvomattoman toiminta-ajan (Warden 2020). Termi sai alkunsa vuonna 2014 Googlella, kun Pete Warden liittyessään Googlen palvelukseen tutustui OK Google -tiimin kehittämään neuroverkkoa hyödyntävään teknologiaan. Monille Android -käyttöjärjestelmän käyttäjille tuttu ”Ok Google” sanan tunnistuksessa hyödynnetään neuroverkkoa. Tunnistukseen käytettävän neuroverkon koko on vain 14 kilotavua. Tämä sai Pete Wardenin innostumaan muutaman dollarin vähän virtaa kuluttavista mikrokontrollereista ja niissä suoritettavasta koneoppimisesta. (Warden & Situnayake 2020, 1). Warden on toiminut vuodesta 2014 alkaen Googlen TensorFlow mobile -tiimin vetäjänä. Tiivistettynä TinyML tarkoittaa neuroverkon suorittamista laitteella, joka saa virran nappiparistosta (Warden & Situnayake 2020, 2). Kuviossa 1 on havainnollistettu TinyML teknologian tarjoamia etuja. Näitä ovat alhainen virrankulutus, lyhyt viive, vähäinen datankäyttö sekä parempi tietoturva. Prosessoriarkkitehtuuri mahdollistaa vähäisen virrankulutuksen ja suoraan laitteella suoritettava koneoppiminen vähentää suorituksen aikaista viivettä.



Kuvio 1. TinyML teknologian etuja (Kapela 2021)

TensorFlow Lite

TensorFlow Lite on kevennetty versio Googlen avoimen lähdekoodin koneoppimiseen kehitetystä TensorFlow-ohjelmistokirjastosta. Lite on tarkoitettu käytettäväksi mobiili- ja IoT-laitteilla. Se mahdollistaa koneoppimisalgoritmien suorittamisen laitteilla, joissa on vain muutamia kilotavuja muistia (Tensorflow). Käytettävässä laitteessa on oltava 32 bittinen suoritin sekä prosessorin arkkitehtuurin on perustuttava ARM Cortex-M -sarjaan (Tensorflow). Nämä vaatimukset täyttäviä kehitysalustaja löytyy muutamia, joista tunnetuimpia ovat Arduino Nano 33 BLE Sense ja SparkFun Edge. Kattavampi lista tuetuista laitteista löytyy osoitteesta https://www.tensorflow.org/lite/microcontrollers. Koneoppimismallin alkuvalmistelut ja itse opetus tehdään TensorFlow:lla ja valmis malli käännettään sen jälkeen TensorFlow Lite Converterin avulla mikrokontrollerin käyttämään muotoon (C array) (Tensorflow).

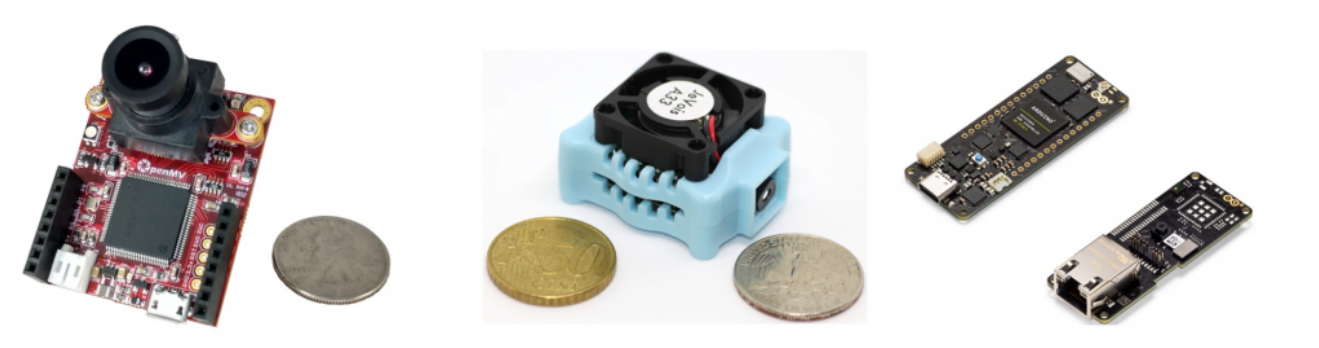
Sulautetut konenäköjärjestelmät

Pelkkien mikrokontrollerien lisäksi markkinoilta löytyy paljon erilaisia valmiita kameralla varustettuja kehitysalustoja. Koneoppimista tukevan kameralla varustetun kehitysalustan voi hankkia itselleen muutamilla kympeillä. Näiden alustojen ominaisuudet ovat todella laajat ja ne kykenevät perinteisen kasvon tunnistuksen lisäksi tunnistamaan mm. QR-koodeja sekä viivakoodeja, värejä, henkilöitä ja erilaisia perusmuotoja kuten neliöitä sekä ympyröitä. Taulukosta 1 löytyy vertailu markkinoilla olevista kehitysalustoista ja niiden ominaisuuksista.

Laite	Prosessori	Muisti	Kameramoduuli	Resoluutio	Linssi	Hinta
OpenMV Cam H7	STM32H743VI, ARM Cortex M7, 480 MHZ	SRAM 1MB, Flash 2MB	OV7725, VGA	640 x 480, 60 FPS	2.8 mm	60 $

JeVois A33	Allwinner A33 quad core, ARM Cortex A7, 1.34GHZ, Dual core Mali-400 GPU	SDRAM 256MB DDR3	1.3 MP	SXGA (1280 x 1024 15 FPS), VGA (640 x 480 30 FPS)	X	50 $
Arduino Portenta H7 + Vision shield	STM32H747XI dual Cortex, ARM Cortex M7 + M4, 480 MHZ	SRAM 1MB, Flash 2MB	Himax HM-01B0	VGA (320 x 320 51 FPS), QVGA (320 x 240 60 FPS)	1.63 mm	150 $

Taulukko 1. Kameramoduulin sisältävien kehitysalustojen vertailu



Kuvio 2. Kameramoduulin sisältäviä kehitysalustoja taulukon mukaisessa järjestyksessä (OpenMV 2021, Jevoisinc 2021, Arduino 2021).

Taulukosta 1 voidaan havaita, että kameramoduulin sisältämissä kehitysalustoissa on hyödynnetty myös ARM Cortex -sarjan prosessoreita. Ehdottomasti mielenkiintoisin alustoista on JeVois A33, joka tarjoaa huomattavasti paremman prosessorin sekä kameramoduulin verrattuna muihin ja vielä todella pienessä koossa. Lisäksi kortilta löytyy myös OpenGL-ES 2.0 tukeva Mali-400 näytönohjain. Teknologia näiden alustojen kohdalla on vielä kehitysvaiheessa mutta jo nykyisen tarjonnan perusteella voidaan olettaa tämän teknologian yleistyvän ja kilpailun myötä hintojen laskevan.

Artikkeli on valmistelu osana Etelä-Pohjanmaan Liiton Euroopan Aluekehitysrahastosta rahoittamaa Laadusta kilpailukykyä konenäöllä -hanketta. Hankkeen tavoitteena on edistää pk-yritysten osaamista ja tietämystä automaattisen laadunvalvonnan konenäkö- ja tekoälypohjaisista menetelmistä.

Janne Kapela
LAAKI-hankkeen asiantuntija
SeAMK Tekniikka

Lähteet:

Arduino. 2021. [Verkkosivu]. Arduino Portenta H7. [Viitattu 16.02.2021]. Saatavana: https://store.arduino.cc/portenta-h7

Collin, J. & Saarelainen, A. 2016. *Teollinen internet*. [Verkkokirja]. Helsinki: Talentum. [Viitattu 16.2.2021]. Saatavana Alma Talent Bisneskirjasto -palvelusta. Vaatii käyttöoikeuden

Jevoisinc. 2021. [Verkkosivu]. Jevois smart machine vision. [Viitattu 16.02.2021]. Saatavana: https://www.jevoisinc.com/products/jevois-a33-smart-machine-vision-camera?variant=36249052042

OpenMV. 2021. [Verkkosivu]. OpenMV Cam H7. [Viitattu 16.02.2021]. Saatavana: https://openmv.io/products/openmv-cam-h7

Tensorflow. Ei päiväystä. TensorFlow Lite for microcontrollers. [verkkosivu]. [Viitattu 16.02.2021]. Saatavana: https://www.tensorflow.org/lite/microcontrollers

Warden, P. & Situnayake, D. 2020. TinyML: Machine learning with TensorFlow Lite on Arduino and Ultra-Low-Power Microcontrollers. USA: O'Reily Media, Inc

Warden, P. 2020. TinyML Talks – Pete Warden: Getting started with TinyML. [Video]. Silicon Valley: TinyML. [Viitattu 16.02.2021]. Saatavana: https://www.youtube.com/watch?v=soOM1F70Boc&ab_channel=tinyML