SWISS – LATVIAN
Cooperation Programme

LACISE
SUSTAINABLE ENERGY

www.lacise.com

Pētījuma tēma

Mākslīgā intelekta darbība, tā manipulēšana un ļaundabīgo funkciju atklāšana

Šajā darbā tiks piedāvāts izpētīt un eksperimentēt ar mākslīgā intelekta lēmumu pieņemšanas kontroli. Vēlāk apskatīsim, kā var atklāt kontroles uzbrukumus mākslīgajam intelektam.
Lai ietekmētu mākslīgā intelekta lēmumus, eksperimentu laikā tiks izveidots vienkāršs mākslīgā intelekta modelis, kas spēj klasificēt attēlus, kā arī tiks pārbaudīts, vai nelielas cilvēkam redzamas vizuālas izmaiņas attēlā var ietekmēt modeļa prognozi.
Lai labāk izprastu modeļa darbību, tiks izmantotas Explainable Artificial Intelligence (XAI) metodes, kas ļauj vizualizēt attēla daļas, kuras visvairāk ietekmē mākslīgā intelekta lēmumu. Tas palīdzēs analizēt, vai modelis patiešām atpazīst objektu vai arī balstās uz citām, ar objektu nesaistītām pazīmēm.

Attēls (A) ir attēls, ko apstrādā mākslīgais intelekts un nosaka klasi.

Šajā gadījumā klase ir ‘suns’. XAI norāda (attēls B), kur mākslīgais intelekts pievērš uzmanību, lai noteiktu klasi.

Darba vides

Darba izstrādes laikā tiks izmantots arī mākslīgais intelekts kā palīgrīks programmēšanā un eksperimentu veidošanā (tā sauktā “vibe coding” pieeja). Tas ļaus ātrāk izstrādāt eksperimentus, saņemt skaidrojumus par izmantotajām metodēm un analizēt rezultātus.

Tiks nodrošināta attālināta piekļuve darba videi Google co-lab. Neliels ievads kā izmantot mākslīgo intelektu, lai tas palīdz kodēt python programmēšanas valodā un skaidri izskaidro uzrakstīto kodu.

Pētījuma metodes:

- Datorredzes arhitektūru apskats;
- Eksperimentu izpildes vide Google colab;
- Klasifikācijas modeļā apmācība;
- Modeļa apmācība ar slēptu funkciju (trigeri);
- Slēptās funkcijas ietekmes novērtējums;
- Slēptās funkcijas identificēšana ar XAI metodēm.

Vibe Coding

Google colab

Darbs notiks ar attēliem, kas nodrošinās uzskatāmu vizualizāciju.

Kontaktinformācija: **Artūrs Ņikuļins, e-pasts: arturs.nikulins@edi.lv**

“LACISE” projektu atbalsta Šveices progamma “Swiss Contribution” ekonomisko un sociālo atšķirību mazināšanai ES.

ELEKTRONIKAS UN DATORZINĀTŅU INSTITŪTS
INSTITUTE OF ELECTRONICS AND COMPUTER SCIENCE