STUDENTŲ ĮTRAUKIMO Į MOKSLINĘ VEIKLĄ SKATININAMOJO KONKURSO TEMA

LT

Temos pavadinimas: Dirbtiniu intelektu grįsto pokalbių roboto įgyvendinimas įterptinėje sistemoje NVIDIA Jetson

Tikslas: Sukurti, įdiegti ir ištirti pokalbių robotą, grįstą didžiaisiais kalbos modeliais (LLM), naudojant „NVIDIA Jetson“ įterptinę sistemą. Šis robotas galės suprasti ir generuoti į žmogų panašius atsakymus realiuoju laiku, jis bus tinkamas naudoti įterptinėse aplinkose, kuriose debesų sistemos gali būti nepasiekiamos.

Trumpas temos vykdymo aprašymas (ne daugiau kaip 2000 ženklų):

Problema:

Pokalbių robotai vis dažniau naudojami klientų aptarnavimo, virtualių asistentų ir kitose programose. Tačiau dauguma šiuolaikinių robotų remiasi debesų sistemomis, turinčiomis didelius skaičiavimo išteklius. Tai riboja jų naudojimą aplinkoje, kurioje yra prastas ryšys arba kur būtina apdoroti duomenis vietoje. Įterptinės sistemos, pavyzdžiui, „NVIDIA Jetson“, yra ekonomiškai efektyvus sprendimas, tačiau šios sistemos susiduria su iššūkiais, susijusiais su ribota skaičiavimo galia, atmintimi ir energijos vartojimo efektyvumu.

Uždaviniai:

1. Sukurti pokalbių robotą, naudojant didelius kalbos modelius, gebantį suprasti natūralią kalbą ir generuoti atsakymus.
2. Įdiegti ir optimizuoti pokalbių robotą "NVIDIA Jetson" įterptinėje sistemoje, kad jis veiktų realiuoju laiku.
3. Ištirti sistemos našumą greičio, tikslumo ir skaičiavimo efektyvumo požiūriu.
4. Ištirti galimus sistemos taikymus įvairiose aplinkose, pavyzdžiui, robotikoje, daiktų interneto įrenginiuose ir autonominėse sistemose.

Metodika:

1. Modelio kūrimas: naudoti iš anksto parengtus kalbos modelius ir pritaikyti juos pokalbiams.
2. Modelio diegimas: pokalbių modelį perkelti į „NVIDIA Jetson“ platformą, naudojant tokias sistemas, kaip „NVIDIA TensorRT“, kad būtų optimizuotas našumas. Daugiausia dėmesio skirti atminties pėdsako mažinimui ir inferencijos spartos didinimui, kad atitiktų įterptinės sistemos ribotas galimybes.
3. Testavimas ir vertinimas: atlikti bandymus, siekiant įvertinti roboto veikimą realiuoju laiku įvairiomis sąlygomis.

Laukiami rezultatai:

1. Veikiantis pokalbių robotas, galintis veikti NVIDIA Jetson įterptinėje sistemoje realiuoju laiku.
2. Įžvalgos apie didelių kalbos modelių našumo apribojimus ir galimybes kraštinių kompiuterių įrenginiuose.
3. Pokalbių roboto integravimo į įvairias krašto kompiuterijos taikomąsias programas, pavyzdžiui, robotiką, išmaniuosius namų įrenginius ir autonomines sistemas, galimybės.

Priemonės: įterptinė sistema „Jetson Orin Nano“ (mini kompiuteris), programavimo aplinka debesyje („Google Colab“, „Kaggle“), kompiuteris su GPU.

Reikalingi įgūdžiai: anglų kalbos žinios, programavimo pagrindai (Python), atkaklumas sprendžiant užduotis/problemas.

Temą siūlantis mokslininkas/dėstytojas: prof. dr. Dalius Matuzevičius

EN

Topic: Implementation of an AI-Based Conversational Bot on NVIDIA Jetson Embedded System

Objective: This project aims to develop, implement, and evaluate a conversational bot based on large language models (LLMs) using the NVIDIA Jetson embedded system. The bot will be capable of understanding and generating human-like responses in real-time, making it suitable for applications in embedded environments where cloud-based systems may not be feasible.

A short description of the topic (maximum 2000 characters):

Problem:

Conversational bots are increasingly used in customer service, virtual assistants, and other applications. However, most modern bots rely on cloud-based systems with significant computational resources. This limits their use in environments with low connectivity or where on-site processing is essential. Embedded systems, such as the NVIDIA Jetson, provide a cost-effective solution, but these systems face challenges related to limited computational power, memory, and energy efficiency.

Research Objectives:

1. Develop a conversational bot using large language models, capable of natural language understanding and response generation.
2. Deploy and optimize the conversational bot on the NVIDIA Jetson embedded system for real-time operation.
3. Investigate the performance of the system in terms of speed, accuracy, and computational efficiency.
4. Explore potential applications for the system in various environments such as robotics, IoT devices, and autonomous systems.

Methodology:

1. Model Development: Utilize pre-trained language models and adapt them for conversational purposes.
2. Model Deployment: Port the conversational model to the NVIDIA Jetson platform using frameworks such as NVIDIA TensorRT for optimizing performance. Focus on reducing the memory footprint and improving inference speed to fit the limitations of the embedded environment.
3. Testing and Evaluation: Conduct tests to assess the real-time performance of the bot under various conditions.

Expected Outcomes:

1. A functional conversational bot capable of operating on the NVIDIA Jetson embedded system in real-time.
2. Insights into the performance limitations and capabilities of large language models on edge computing devices.
3. Potential for the conversational bot to be integrated into various edge-based applications such as robotics, smart home devices, and autonomous systems.

Tools: Jetson Orin Nano embedded system (mini-computer), cloud programming environment (Google Colab, Kaggle), computer with GPU.

Required skills: English language skills, basic programming skills (Python), persistence in solving tasks/problems.

Scientist/teacher proposing the topic: prof dr Dalius Matuzevičius