REACH

NEXT GENERATION DATA INCUBATOR

This project has received funding from the European Union's H2020 research and innovation programme under Grant Agreement no 951981

CORE PARTNERS

DATA PROVIDERS

1 ANNEX I. Technical Specification Double-side Page

1. **TECHNICAL SCOPE:** The mock-up solution is suitable and correctly addresses the challenge/theme selected over the REACH dataset/s. The Big Data solution architecture proposed is adequate to tackle the challenge and the issues associated to the solution in mind. To what extent does the applications handle the data provided?

Our mockup is a hardware device which runs AI-based Speech Activity Detection (SAD), on the edge (locally), without needing to offload the AI inference to a cloud service, safeguarding the sensitive audio data since it **never leaves the device**. Even though this is a big step towards achiving Security and Digital Trust**, we want to do even better**. We are aiming to implement source-code analysis and privacy-enhancement with the FRAMA-C framework. We already started contact with CEA in order to have support integrating it into our system.

2. **SELECTION OF ALGORITHMS AND TOOLS:** The indicated Data Science approach, i.e. algorithms chosen, and Big Data architecture approach, i.e. tools chosen may successfully accomplish the required data governance, processing and analysis. A clear understanding of the used REACH dataset/s is demonstrated.

The data provider did not provide a sample dataset for speech activity detection (SAD). We had to use an open-source one (Voice Activity Detection Toolkit dataset). However, we did use the activity recognition dataset to augment the open-source SAD dataset: we overlapped the human speech audio with the various indoor sounds from CERTH Health dataset to create an AI model more robust to indoor sound interference. In addition, we will incorporate FRAMA-C for code analysis.

This pr
the Eur
and inn
Grant A

3. **TECHNICAL SCALABILITY AND FLEXIBILITY OF THE SOLUTION:** The solution can truly cope with humongous and increasing datasets, potentially from diverse data providers, and is flexible it to adapt to other related domains.

REACH NEXT GENERATION

Definitely. Our EdgeAI SAD solution was made with our backbone technology: an AI model optimization platform called BinedgeML, which offers unparalleled technical scalability and flexibility. BinedgeML is data and application agnostic, allowing seamless integration with any dataset from diverse data providers. It empowers your solution with drastic AI optimizations that translate into 20x inference speedup while using 80x less memory, enabling it to run on tiny microcontrollers. Additionally, since it does not depend on any central/cloud computing service, its scalability is not limited by infrastructure: since there's no need to increment the central computing capabilities or communication infrastructure for every node (sensor/microphone) added to the network.

4. **DATA GOVERNANCE AND LEGAL COMPLIANCE:** Data sharing challenges, data governance and legal compliance, must be observed. The proposed solution is compliant with the current data legislations concerning security and privacy (e.g. GDPR).

Our proposed solution for the REACH incubation program addresses data governance and legal compliance concerns. It uses a small microcontroller connected to a microphone array, which processes audio input using a neural network to detect speech. Importantly, the inference is performed locally on the device, and only a binary indication of whether speech was detected is transmitted. The sensitive audio data itself never leaves the device. In addition, we incorporate FRAMA-C for code analysis to ensure that no sensitive data is transmitted from the device. This approach aligns with current data legislations, including security and privacy regulations like GDPR.

5. **QUALITY ASSURANCE AND RISK MANAGEMENT:** Feasible and credible quality process followed for the final product generation. The potential risks in all the phases of the project (design of the solution, development, testing, deployment…) are identified and convincing mitigation plans put in place.

This pr
the Eur
and inn
Grant A

TESTING:

- **Privacy and security concerns arising from the use of speech**

Mitigation: Unlike standard cloud/distributed architectures, with our EdgeAI approach no sensitive information leaves the embedded device (only the decision), hence ensuring privacy by design

- **Missing Data, susceptibility to indoor noises, sounds recorded from variable distance, Data bias and overfitting leading to poor accuracy in real-world scenarios**

Mitigation: We will build on previous experience of the team on further data augmentation, to prevent the models from becoming too specific to the provided dataset, making it unable to generalize to new unseen data. We have identified some extra open-source speech detection datasets such as the Mozilla Common Voice and VOiCES to further diversify our datasets. These measures will also help meet the challenge's requirements of developing a SAD model robust to indoor sound interference and sounds recorded with variable distance

DEPLOYMENT: Since this is a hardware product, there are two main categories of risks:

- hardware: hardware risk mitigation and quality control, an automated test bench will be created to automatically and efficiently run numerous hardware test on multiple boards and on multiple PCB test points at the same time (e.g. test if the power unit is working, check if the microcontroller is working…)

- Firmware: The AI model will reside on the device's flash memory. It's likely that shortcomings will be identified in the AI model. Thus, an efficient method of AI model update must be in place to allow Over The Air (OTA) updates so that all devices already on the field can be seemlessly updated with no effort.

This pr
the Eur
and inn
Grant A

REACH

NEXT GENERATION DATA INCUBATOR

This project has received funding from the European Union's H2020 research and innovation programme under Grant Agreement no 951981

CORE PARTNERS

DATA PROVIDERS