openGENESIS – GSLP proposal

Quality assurance of AI data for machine learning by labelling process considerations (acronym: Phoenix)

02.09.2019 – for consideration of next openGENESIS steering committee meeting

Document Source:

Matthis Eicher
TÜV Süd Auto Service GmbH
eMail: matthis.eicher@tuev-sued.de

Participating entities

TÜV Süd Auto Service GmbH
Incenda AI GmbH

Project Leads (compulsory)

* Matthis Eicher (TÜV Süd Auto Service GmbH)
* Felix Friedmann (Incenda AI GmbH)
* Florian Netter (Incenda AI GmbH)

Committers (compulsory)

* Matthis Eicher (TÜV Süd Auto Service GmbH)
* Felix Friedmann (Incenda AI GmbH)
* Florian Netter (Incenda AI GmbH)

Scope (compulsory)

A process for annotating data and generating corresponding ground-truth information to train and test Artificial Intelligence (especially for Machine Learning) shall be described, investigated and potential weaknesses identified.

Description (compulsory)

Machine learning is based on training networks from data and also evaluating them on such data. This requires both the data itself and the associated ground truth information of this data. This ground-truth information is created in an "annotation process" partly automatically, partly manually and basically has the potential to be erroneous.
If the data contain erroneous ground truth information, the function of the network can be learned only to a limited extent or even incorrectly, and secondly, an evaluation of the AI function on this data cannot be trusted.
Therefore, it is absolutely necessary that the ground truth information is correct. This is already anchored in the annotation process, which is why this should be the focus of consideration of this project.

As deliverable a whitepaper is planned, which describes the process, the potential weaknesses and possible mitigations.

To reach the deliverable, the following steps are planned and will be performed:
* creation of a ML meta development process diagram
* creation of an annotation process diagram
* description of annotation process
* analysis of annotation process and weakness identification

* discussion of potential mitigations
* creation of the whitepaper

This is a project to build knowledge and best-practice. There is no code development planned.

Why Here? (optional)

The consideration of the annotation process can be understood as one stop to a safe AI and should be executed in the openGENESIS WG.

Note: General justification – “Project within scope of openGENESIS”

Licenses (compulsory)

CC BY 4.0

Legal Issues (compulsory)

none known

Initial Contribution (optional)

n/a

Project Scheduling (compulsory)

* Lifecycle diagram should be available by end of July 2019
* Labelling Process should be described by end of Aug 2019
* Process Analysis should be performed by end of Okt 2019
* Working draft of whitepaper should be available by mid of Okt 2019
* Final whitepaper should be provided by end of November 2019

Reporting frequency to the SC: Once a month.

Future Work (optional)

The consideration of the labelling process for AI-data is only one step to a safe AI.
Further topics which can be considered later, are:
* implications regarding the specification
* requirements for data acquisition, pre-processing, selection for labelling
* requirements for selection for training (bias, distribution, etc.)

Necessary infrastructure (compulsory)

* Tuleap

Interested Parties (optional)

n/a