TECH OFFER

Gamified Data Annotation Platform For Supervised Machine Learning

KEY INFORMATION

TECHNOLOGY CATEGORY:
Infocomm - Artificial Intelligence
Infocomm - Data Processing
Infocomm - Social Media, Collaboration & Crowdsourcing

TECHNOLOGY READINESS LEVEL (TRL): TRL8
COUNTRY: SINGAPORE
ID NUMBER: TO174642

OVERVIEW

Machine Learning (ML) is a sub-field of Artificial Intelligence (AI) where a machine is able to learn without being explicitly programmed. However, before a machine can effectively perform even the simplest AI tasks, e.g. differentiating between images containing an elephant or a tiger, it has to be trained on images containing both animals. To be useful in supervised learning, training data needs to be properly labelled or annotated by a human for the machine to extract the relevant features and produce an ML model that serves its intended purpose. This highlights the important role that data annotation plays in producing robust, accurate ML algorithms in video analytics, natural language processing, and audio processing. However, many organisations that want to embark on their supervised learning journey often face difficulties gaining access to high-quality labelled datasets, known as ground truth data, due to the abundance of low-quality, expensive and unstructured data.

This technology offer is a mobile application-based data platform that enables companies to obtain high-quality annotated data. It de-centralises data collection and data annotation tasks into manageable bite-sized chunks for optimal annotation performance

and crowd/out-sources the annotation task to a pool of data taggers via a mobile application. Labelling quality is established through a gamification system and a series of built-in verification procedures, including AI-assisted pre-filtering and collective human quality control.

TECHNOLOGY FEATURES & SPECIFICATIONS

This technology offer comprises a mobile application for data taggers to participate in crowd-sourced annotation and a web portal that serves as a control panel for organisations to submit and track the progress of their annotation tasks. A proprietary system for data quality assurance consists of the following:

- Data pre-preprocessing process
- Gamification and leveling (mobile application)
- Clustering algorithms to filter outliers
- AI-assisted filters to detect anomalies
- Human quality control

The mobile application has the following features:

- Built-in gamification rewards data taggers for generating high-quality labeled data
- Simplified, micro-job structure within an anywhere, anytime annotation tool

The web portal has the following features:

- Manage the upload and distribution of raw data
- Data annotation workload is split algorithmically, down to the basic unit of each annotation task
- Download labelled/annotated data in various commonly used formats; customisable for new formats
- Library of ready-to-use datasets

Supports common data annotation formats:

- YOLO Darknet TXT
- Tensorflow CSV
- COCO JSON
- PASCAL VOC XML

Audio annotations are captured in an SRT (subtitle) file, while classification type annotations are saved to a CSV file.

POTENTIAL APPLICATIONS

The data labelling/annotation platform bridges the gap between people who have the time to deal with unstructured data with the organisations that do not. The data platform's supports the labelling/annotation of various formats of unstructured data:

Image (Bounding Box, Image Classification, Polygonal Bounding, Image-to-text transcription)
Build robust detection, background/foreground segmentation, or image classification AI models supported by high-quality annotated data

Text (Entity Extraction, Intent Recognition, Sentiment Analysis, Text Classification):

Imbue chatbots with enhanced natural language processing capabilities, with the ability to understand region-specific intent and discern the user's sentiment (positive, neutral, negative)

Audio (Audio Transcription, Sound Classification, Audio Translation):
Eliminate accent bias and improve audio/conversational AI with a wider range of vocabulary, in multiple languages

Video (Bounding Boxes, Polygonal Bounding, Subtitling):
Accelerate computer vision model development (model training and testing) for person, and object detection, with accurately labelled ground truth data

MARKET TRENDS & OPPORTUNITIES

The need for data annotation increases with AI/ML growth. It is expected to grow from US $1.35B in 2020 to US $8.2B in 2028. With the highest CAGR of 31.1% during the period of 2021 - 2030.

UNIQUE VALUE PROPOSITION

This technology has the following benefits:

- Crowd-sources dataset collection and reduces the time required for ground truth data annotation
- Technology companies can collect and annotate raw data that support their own core AI products, especially when it is inefficient/expensive to maintain and/or scale a team of full-time data annotators.
- Non-technology organisations that have access to large amounts of visual or unstructured data will be able to monetise their annotated datasets or use annotated data to support in-house AI/ML projects that relate to their respective industry.
- Proprietary quality control system ensures annotated data quality is maintained
- Addresses contextual, localised data annotation needs e.g. region-specific translation of a native language or local landmarks

The technology owner is interested in collaborating with various organisations to test-bed existing competencies and deep-tech companies to work on developing data generation and AI augmentation capabilities across various sectors/industries.