

都市時空間理解に向けたマルチモーダル基盤モデルの開発

実施者

ウーブン・バイ・トヨタ株式会社

事業概要

都市時空間理解に向けたマルチモーダル(言語・映像・センサ)基盤モデル (City-LLM) を開発する。都市の状況を『時間』や『場所』といった側面で理解し、人の行動を促進する世界を創出する



公共の場所にあるカメラに着目し、映像データを入力として、インスタンスレベルでの“場所”、および、“時間”方向での理解に特化したVision LLMの開発から着手する。この技術によって、交差点での歩行者のヒヤリハットなシーンの理解、さらに危険回避の提案が可能になる。また空間での人の行動理解を通して、様々なアプリケーション（異常検知、接客サービス品質向上など）が可能になる。

社会実装イメージ

豊田市土橋の交差点にカメラを設置し、AIによって、歩行者のヒヤリハットシーンを理解し、回避できるかを検証した。車両の急ブレーキ信号から時刻を特定し、該当する映像をVLLMで言語として記述した（右上図）。また同技術は、防犯や接客現場でのサービスの品質向上など、公共の場でのカメラを用いたアプリケーションとして応用可能。

路上カメラ

City LLM

歩行者[1]の行動を記述してください。

青いジャケットと黒いズボンをした歩行者[1]が車に向かって早足で走り、車の前面につかまりながら渡ろうとしている。道路には歩行者も数人立っており、傘をさしている者もいる。雨が今にも降りそうであることを示している。

交差点付近、路上カメラから、CityLLMによる車両と歩行者の[位置][行動][視線][コンテキスト]への理解によるヒヤリハットの発覚と原因分析

安心・安全

生活の利便

教育現場のDX

街の情報検索

自動車

異常検知 → 省人化

カフェ・教育現場 → サービス品質向上

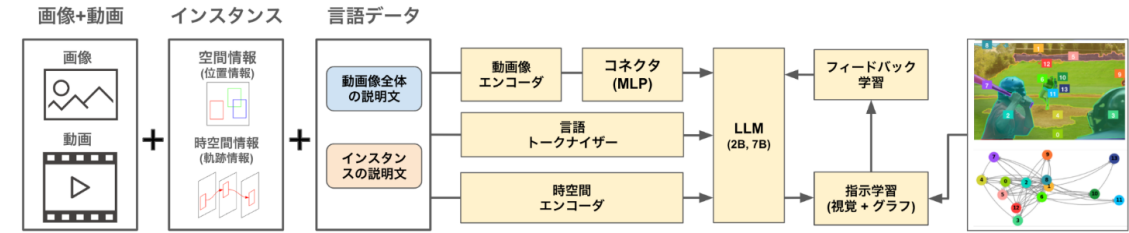
City LLM

建物の案内、混雑状況理解

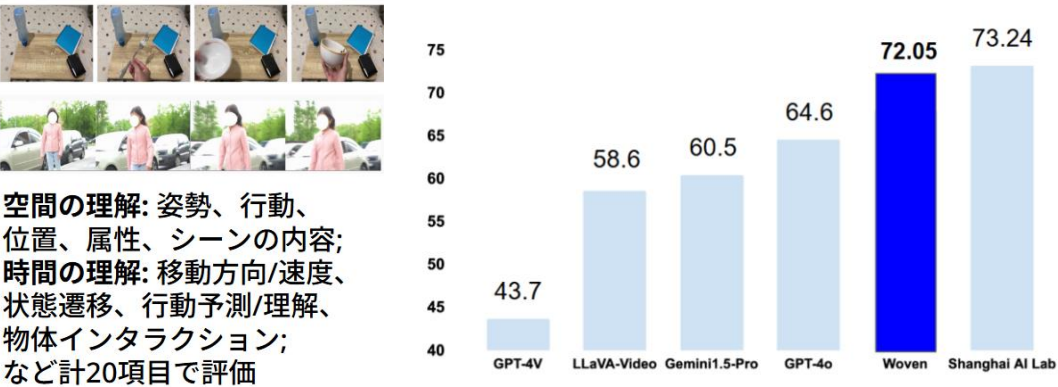
インフラ協調 → 交通安全

事業成果

以下の3つを構築した。①動画内のインスタンスレベルの時空間理解に特化したマルチモーダル基盤モデル（70億パラメータ）（下図上段）②独自に構築した動画像＋言語のデータセット（6億の動画像）③DeepSpeed @ GKE (Google Kubernetes Engine) スケーラブルな分散学習環境。その結果、動画像理解に特化したベンチマークで、72.05%の精度（25年4/30時点でリーダボード世界トップ、論文ベースで世界2位）を示すことに成功した（下図下段）。



モデルの構成（Woven-VLLM）



MVBenchデータセットでの評価結果