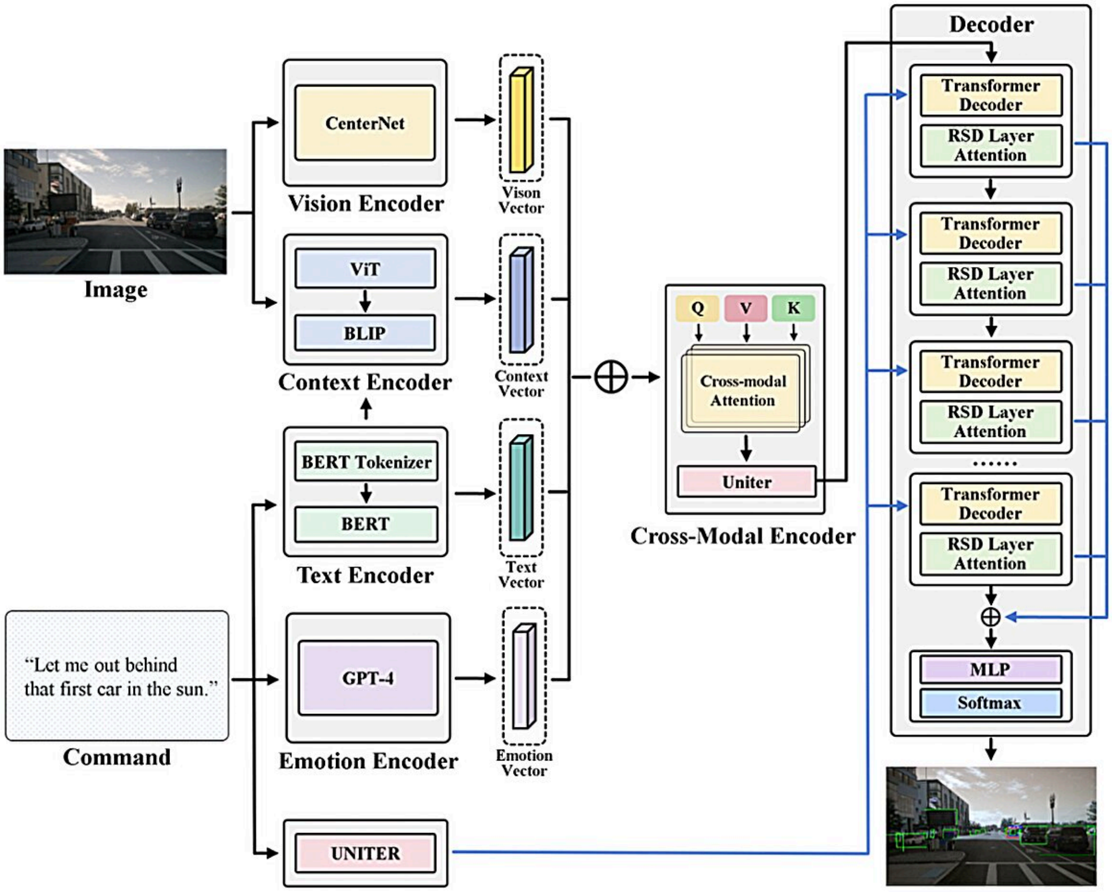
TechXplore

Voice at the wheel: Study introduces an encoder-decoder framework for AI systems

April 29 2024



CAVG is structured around an Encoder-Decoder framework, comprising encoders for Text, Emotion, Vision, and Context, alongside a Cross-Modal encoder and a Multimodal decoder. Credit: Communications in Transportation Research, Tsinghua University Press

Recently, the team led by Professor Xu Chengzhong and Assistant Professor Li Zhenning from the University of Macau's State Key Laboratory of Internet of Things for Smart City unveiled the Context-Aware Visual Grounding Model (CAVG).

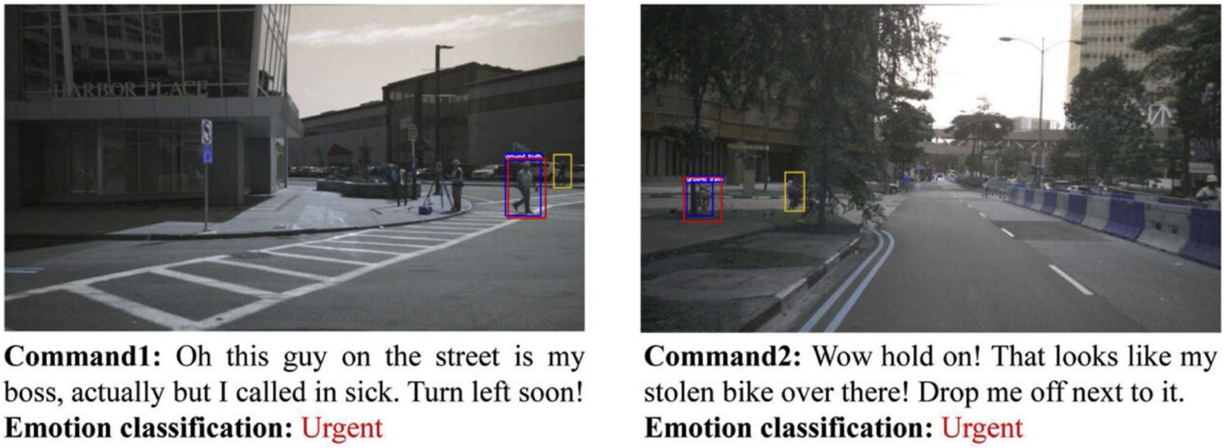
This model stands as the first Visual Grounding autonomous driving model to integrate natural language processing with large language models. They published their study in *Communications in Transportation Research*.

Amidst the burgeoning interest in autonomous driving technology, industry leaders in both the automotive and tech sectors have demonstrated to the public the capabilities of driverless vehicles that can navigate safely around obstacles and handle emergent situations.

Yet, there is a cautious attitude among the public towards entrusting full control to AI systems. This underscores the importance of developing a system that enables passengers to issue voice commands to control the vehicle. Such an endeavor intersects two critical domains: computer vision and natural language processing (NLP).

A pivotal research challenge lies in employing cross-modal algorithms to forge a robust link between intricate verbal instructions and real-world contexts, thereby empowering the driving system to grasp passengers' intents and intelligently select among diverse goals.

In response to this challenge, Thierry Deruyttere and colleagues inaugurated the Talk2Car challenge in 2019. This competition tasks researchers with pinpointing the most semantically accurate regions in front-view images from real-world traffic scenarios, based on provided textual descriptions.



Command1: Oh this guy on the street is my boss, actually but I called in sick. Turn left soon!
Emotion classification: Urgent

Command2: Wow hold on! That looks like my stolen bike over there! Drop me off next to it.
Emotion classification: Urgent

Illustration of Regions Identified by an AV based on a Raw Image and a Natural Language Command. The blue bounding box represents the ground truth. The red and yellow bounding boxes correspond to the prediction results from CAVG with emotion categorization and without emotion categorization, respectively. Credit: *Communications in Transportation Research* (2024). DOI: 10.1016/j.commtr.2023.100116

Owing to the swift advancement of Large Language Models (LLMs), the possibility of linguistic interaction with autonomous vehicles has become a reality. The article initially frames the challenge of aligning textual instructions with visual scenes as a mapping task, necessitating the conversion of textual descriptions into vectors that accurately correspond to the most suitable subregions among potential candidates.

To address this, it introduces the CAVG model, underpinned by a cross-modal attention mechanism. Drawing on the Two-Stage Methods framework, CAVG employs the CenterNet model for delineating numerous candidate areas within images, subsequently extracting regional feature vectors for each. The model is structured around an Encoder-Decoder framework, comprising encoders for Text, Emotion, Vision, and Context, alongside a Cross-Modal encoder and a Multimodal decoder.

To adeptly navigate the complexity of contextual semantics and human emotional nuances, the article leverages GPT-4V, integrating a novel multi-head cross-modal attention mechanism and a Region-Specific Dynamics (RSD) layer. This layer is instrumental in modulating attention and interpreting cross-modal inputs, thereby facilitating the identification of the region that most closely aligns with the given instructions from among all candidates.

Furthermore, in pursuit of evaluating the model's generalizability, the study devised specific testing environments that pose additional complexities: low-visibility nighttime settings, urban scenarios characterized by dense traffic and intricate object interactions, environments with ambiguous instructions, and scenarios featuring significantly reduced visibility. These conditions were designed to intensify the challenge of accurate predictions.

According to the findings, the proposed model establishes new benchmarks on the Talk2Car dataset, demonstrating remarkable efficiency by achieving impressive outcomes with only half of the data for both CAVG (50%) and CAVG (75%) configurations, and showing superior performance across various specialized challenge datasets.

Future endeavors in research are poised to delve into advancing the precision of integrating textual commands with visual data in autonomous navigation, while also harnessing the potential of large language models to act as sophisticated aides in autonomous driving technologies.

The discourse will venture into incorporating an expanded array of data modalities, including Bird's Eye View (BEV) imagery and trajectory data among others. This approach aims to forge comprehensive deep learning strategies capable of synthesizing and leveraging multifaceted modal information, thereby significantly elevating the efficacy and performance of the models in question.

More information: Haicheng Liao et al, GPT-4 enhanced multimodal grounding for autonomous driving: Leveraging cross-modal attention with large language models, *Communications in Transportation Research* (2024). DOI: 10.1016/j.commtr.2023.100116

Provided by Tsinghua University Press

Citation: Voice at the wheel: Study introduces an encoder-decoder framework for AI systems (2024, April 29) retrieved 17 May 2024 from https://techxplore.com/news/2024-04-voice-wheel-encoder-decoder-framework.html