Intelligent Traffic Monitoring with Hybrid AI

Ehsan Qasemi[1*], **Alessandro Oltramari**[2]
[1]University of Southern California, Los Angeles, CA, USA
[2]Bosch Research and Technology Center, Pittsburgh, PA, USA
qasemi@usc.edu, alessandro.oltramari@us.bosch.com

Abstract

Challenges in Intelligent Traffic Monitoring (ITMo) are exacerbated by the large quantity and modalities of data and the need for the utilization of state-of-the-art (SOTA) reasoners. We formulate the problem of ITMo and introduce HANS, a neuro-symbolic architecture for multi-modal context understanding, and its application to ITMo. HANS utilizes knowledge graph technology to serve as a backbone for SOTA reasoning in the traffic domain. Through case studies, we show how HANS addresses the challenges associated with traffic monitoring while being able to integrate with a wide range of reasoning methods.

1 Introduction

U.S. Intelligent Traffic Management sector is estimated to be worth approximately $3B and will expand with a CAGR[1] forecast to be 10.2% through 2028 Grand View Research [2021]. This projection is likely to be adjusted to reflect the recent approval of the U.S. infrastructure bill, set to infuse an additional $11B in transportation safety, including programs to reduce crashes and fatalities CNN [2021].

Intelligent Traffic Monitoring (ITMo) represents an essential instrument to improve road safety and security in intelligent transportation systems Pascale *et al.* [2012]; Jabbar *et al.* [2019]. Moreover, with the advent of autonomous vehicles, ITMo is poised to become an even more relevant part of the smart city infrastructure of the future (the abundance of real-time data points can be used to make traffic mitigation strategies more efficient and effective Chowdhury *et al.* [2021]).

ITMo focuses on deriving actionable knowledge from a large array of networks of sensors deployed along highways, city roads and intersections, etc. Muppalla *et al.* [2017] use knowledge graph technology for traffic congestion detection by proposing an ontology based on SSN Haller *et al.* [2019] to interpret the video feed from traffic cameras and a heuristic reasoner on top of the ontology. Similarly, Lam *et al.* [2017] use real-time image processing to count the number of taillights in the video feed from traffic cameras. However, both works fall short in accommodating a wide range of multi-modal and background information on traffic. Transformer-based language models (LMs) are SOTA methods in many natural language understanding tasks Liu *et al.* [2019] and show human-level commonsense knowledge Ma *et al.* [2019]. However, to the best of our knowledge, no effort has been make to integrate them into ITMo.

Our **first** contribution is formulating the problem of ITMo in presence of large quantities of multi-modal, background (including commonsense), and domain knowledge that can leverage SOTA symbolic and neural reasoners.

Our **second** contribution is HANS, a neuro-symbolic architecture for multi-modal context understanding in ITMo. We evaluate HANS, by implementing set reasoning case studies, where each showcases the capability of HANS to process different modalities of data or integrate with various of type of reasoners. Our results show that HANS, can serve as a platform for the integration of SOTA in ITMo and paves the way for future research in the field.

2 Intelligent Traffic Monitoring: Problem Formulation

We formulate the ITMo as the problem of deriving actionable knowledge from multi-modal real-time and background knowledge. We identify six main requirements for ITMo, all related to processing and reasoning over high volumes of heterogeneous data.

1. **Multi-modal information fusion**. ITMo involves data processing at scale, which can include transient information generated by multiple surveillance cameras, traffic lights, speed radars, acoustic sensors, etc. Lancelle [2016]. As aggregating such data is inherently complicated, generalizing over them to obtain high-level traffic situations constitutes an open problem.
2. **Background Knowledge Acquisition**. Any human-like reasoning performed by AI systems over traffic situations should include knowledge of the common features that can be encountered in most typical scenarios. For instance, we know that a car coming to a stop on the side of a highway is either malfunctioning or indicating an emergency experienced by the driver; we also know that the presence of a police vehicle parked right behind the car changes the scene interpretation, suggesting that the

[*]Work performed during internship at Bosch.
[1]Compound Annual Growth Rate.

driver was prompted to pull over due to an alleged traffic violation. Commonsense knowledge stems from everyday experience Sap *et al.* [2019]; Ilievski *et al.* [2020]: it can be causal, e.g. vehicles typically stop in front of the red light; spatial, e.g. two cars cannot occupy the exact same location, or even cultural, e.g. the so-called *Pittsburgh left*[2].

3. **Domain Knowledge Extension**. Public knowledge can also be leveraged for ITMo, e.g., the map of a geographical region of interest, e.g. OpenStreetMap OpenStreetMap contributors [2017], live traffic observed by drivers, e.g. using the Waze app[3], relevant weather information, e.g. OpenWeatherMap[4], or historical traffic data, e.g. SigAlert Walton *et al.* [2011]. Combining all these information sources requires an advanced semantic model, capable of federating knowledge at different levels of granularity.
4. **Flexibility**. Any reasoning system for ITMo must have a flexible design to allow for incremental development, which requires testing the algorithms across different traffic scenarios, factoring in additional modalities, etc.
5. **Integration with SOTA**. ITMo must allow for seamless integration with SOTA machine learning (e.g., computer vision algorithms) and inference engines, both at the symbolic and neural level;
6. **Robustness**. ITMo must be able to handle noisy information, such as wrong semantic labels from computer vision algorithms, missing video frames, etc.

To satisfy these six requirements, we propose a neuro-symbolic architecture that integrates knowledge graph technologies with neural networks.

3 Human-Assisted Neuro-Symbolic Architecture for Multi-modal Context Understanding (HANS)

HANS has been designed to easily adapt to various sensor-based domains, with the ultimate goal of enhancing road security and safety. In the following, we briefly describe the architecture pipeline shown in figure 1.

Generation. To extract information from images of the camera feed, we use SOTA off-the-shelf computer vision models. Here, we used object detection models to process individual frames from videos collected by stationary surveillance cameras. This process, which is illustrated in figure 1 as IVA data includes: object classes (car, truck, pedestrian, etc.) He *et al.* [2015]; Carion *et al.* [2020], speed and trajectory of moving objects Rad *et al.* [2010], relative position (based on road topology and lane information) Ventura *et al.* [2018].

To extract the background information and domain knowledge, we relied heavily on existing resources of public knowledge, e.g. Wikidata Vrandečić and Krötzsch [2014] and OpenStreetMap OpenStreetMap contributors [2017].

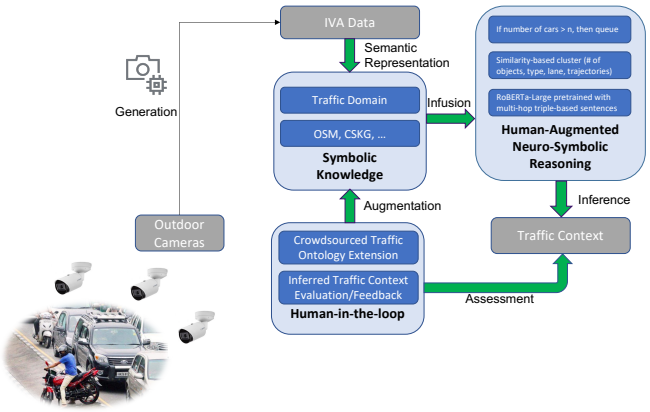

Figure 1: Human-Assisted Neuro-Symbolic Architecture for Multi-modal Context Understanding

Semantic Representation. An ontology for traffic monitoring (*domain level*) is built to extend the Scene Ontology (*core level*) developed by Bosch Oltramari *et al.* [2020] based on W3C's Semantic Sensor Network ontology specifications Haller *et al.* [2019] (*foundational level*). Through suitable mapping mechanisms, the IVA data are used to instantiate the domain ontology, generating the Traffic Monitoring Knowledge Graph (TMKG)[5]. This is further expanded with relevant partitions of two different resources: the federated Commonsense Knowledge Graph (CSKG), to provide commonsense background knowledge related to the traffic domain Ilievski *et al.* [2020], and Open Street Maps OpenStreetMap contributors [2017], to expand traffic scenes based on the location-based knowledge consistent with the GPS coordinates of the cameras.

Augmentation & Assessment. These modules include two main human-in-the-loop processes: an early phase where relevant concepts to model the traffic domain are elicited using crowdsourcing, and a late phase where crowd workers are asked to assess the results of human-augmented neuro-symbolic reasoning.

Infusion & Inference. In these two intimately connected modules, the TMKG is used as the basis for reasoning, in particular by i) triggering rule-based inferences for simple predictions (e.g., a given number of cars on a lane may be considered as a threshold to identify a *traffic queue*; ii) training similarity-based statistical algorithms with specific triple-based scene features (type of objects, trajectory, speed, etc.); iii) fine-tuning a language model with sentences generated from multi-hop (edge) connections linking different TMKG nodes. The current demo[6] is based on a sample dataset (2 days' worth of video footage) collected from a single stationary camera deployed over an intersection in Chesapeake, Virginia. The demo (Figure 3) and relevant examples are discussed in section 4.

[2] A colloquial term for the driving practice of the first left-turning vehicle taking precedence over vehicles going straight through an intersection, associated with the Pittsburgh, Pennsylvania, area.

[3] https://www.waze.com

[4] https://openweathermap.org/

[5] Implemented using Stardog, a well-known enterprise-level knowledge graph framework - see https://www.stardog.com/

[6] See video clip at this link:https://tinyurl.com/476nwf2n

4 Evaluation

HANS is designed with ease of reasoner integration in mind: here we showcase this key feature by demonstrating different types of inference methods. First, we focus on conventional rule-based models that aim to find traffic congestion in video feeds. Second, we present two advanced approaches, respectively centered around TMKG and a pre-trained language model.

Symbolic Method. Explainability [7] of results and human interpretability [8] make symbolic methods a popular choice in critical applications such as ITMo Muppalla *et al.* [2017]; Lam *et al.* [2017]. Muppalla *et al.* [2017] use the deviation of each frame of video from the median as a core indicator of traffic congestion. Figure 2 illustrates the results of HANS's implementation of this method, in which we have two frames from the video feed with, respectively, least and most deviation from the median frame.

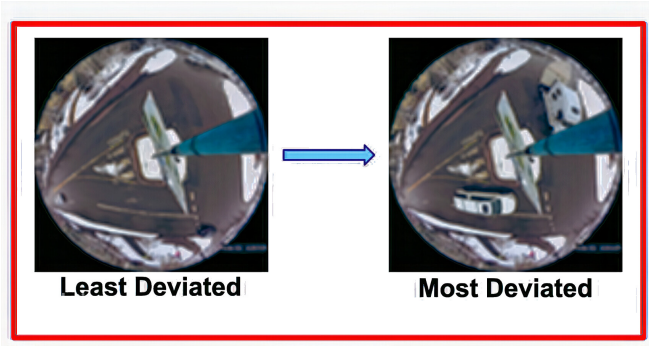

Figure 2: Baseline rule-based method for traffic congestion detection. Left: the least deviated frame from the median scene. Right: the most deviated frame from the median scene. The former has lower probability of congestion than the latter.

With little effort one can extent such symbolic method to include more flexible rules to cover wider range of traffic phenomenon. A more flexible rule-based method can facilitate the incorporation of additional extracted information from the image. Figure 3, represents and example of this method, implemented using HANS GUI. This functionality can help the experts to perform on-the-fly annotations to assess the quality/complexity of rules and provides various options to implement advanced queries.

Scene Query Based on Graph Structure. SOTA reasoners that use whole graph structures to learn appropriate features from data can be used to obviate the scalability issues that symbolic methods generally suffer from Lin *et al.* [2019]; Gamanayake *et al.* [2020]; Scarselli *et al.* [2008]; Gori *et al.* [2005]. To this end, we used the TMKG graph structure of each video frame to find similar frames using methods such as Weinberger *et al.* [2009] or Gamanayake *et al.* [2020]. Table 1 presents two examples from the graph-based reasoning model that uses Weinberger *et al.* [2009] on top of HANS for the scene similarity search. Here, the query is a node in HANS associated with a video frame and the output are the frames from other cameras that are closely matched with the scene. In the current system demonstrator, we rely on the textual description of the frame for both query and output frames.

Query Description	Best Match Description
3 objects in the scene, 2 cars and 1 bike. The cars are located in lane 3, and move at average speed of 5 m/s.	3 object(s) in the scene. From the object(s), 1 is a bike, and 2 are car. the first car is moving with the average speed of 7.74 m/s. The second car is moving with the average speed of 5.31 m/s. In lane 3 we see first car, and second car.
3 objects in the scene, 2 cars and 1 bike. The second car is moving at average speed of 5 m/s, in lane 3	3 object(s) in the scene. From the object(s), 1 is a bike, and 2 are car. the second car is moving with the average speed of 5.52 m/s. In lane 3 we see second car.

Table 1: Scene similarity queries on top of HANS using feature hashing Weinberger *et al.* [2009].

Natural Language Scene Query. Graph based reasoning methods are limited to graph structure and incapable of easily incorporating background knowledge of traffic domain. Alternatively, another class of SOTA *reasoners* that HANS can easily integrate with, relies on language models Ma *et al.* [2019]; Liu *et al.* [2019]. For integration with these, we have implemented conversion methods to transform TMKG and CSKG triples into natural language statements. The resulting sentences include basic visual information such as scene composition, e.g. "it has 3 cars and 2 trucks", topological information, e.g. "road has three lanes" for scene information, or background knowledge, e.g. "truck is a vehicle different from a car". One then can utilize these human-readable sentences as input to language models to reason over and *interpret* text, such as questions. For our the ITMo use case, we fine-tuned RoBERTa Liu *et al.* [2019] language model on the background and commonsense statements and used the embedding it provides to find frames that match the embedding of a natural language query. Accordingly, the examples in Table 2 show that the model can *understand* the features associated with different types of entities, applying constraints on their intrinsic properties (e.g., speed) and extrinsic properties (e.g., lane-based localization).[9] Here, the first example illustrates natural interaction on exclusion of objects, the second example illustrates a natural interaction based on quality of the objects.

5 Conclusion

In this paper, we formulate the problem of ITMo, and introduced HANS, the Human-Assisted Neuro-Symbolic Architecture for Multi-modal Context Understanding, focusing on its

[7] The degree which humans achieve deep understanding of the internal procedures that take place while the model is training or making decisions.Linardatos *et al.* [2020]

[8] the ability to explain or to present in understandable terms to a human Doshi-Velez and Kim [2017]

[9] Note that the order of the entities (i.e., *first*, *second*, *third*) is based on the time stamp associated with perceptual information.

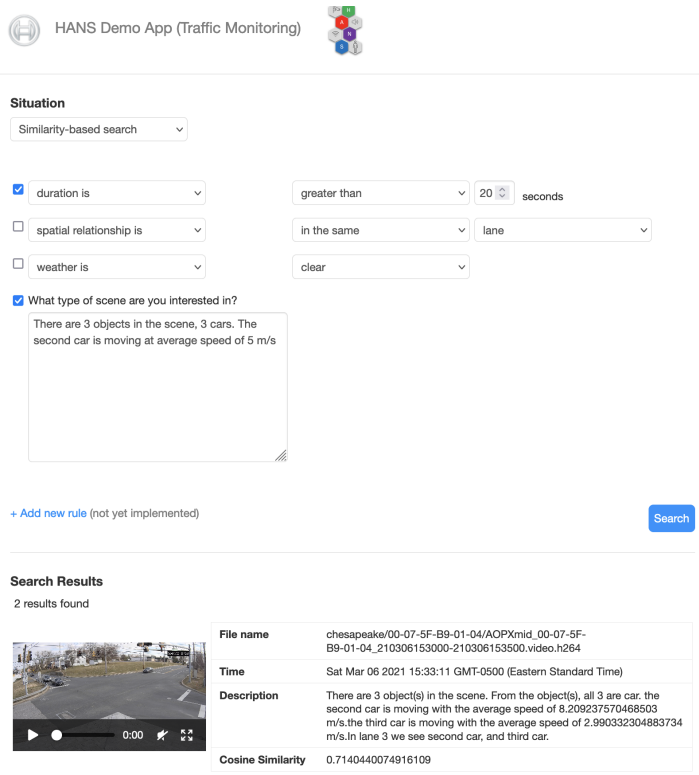

Figure 3: HANS GUI for intelligent traffic monitoring

Query	Best Match Description
show me a scene with no cars in it	There are 2 objects in the scene. From the object(s), all 2 are bike. The first bike is moving with the average speed of . . .
show me a scene with a fast moving car	There is 1 object in the scene. From the object(s), 1 is a car. The first car is moving with the average speed of . . .

Table 2: Natural language queries on top of HANS using RoBERTa language model fine-tuned on traffic background knowledge

application to intelligent traffic monitoring. At its core, our solution is centered on a knowledge graph that integrates the data generated by computer vision algorithms (IVA) with general, publicly available knowledge: this method plays a key role in satisfying the requirements 2-5 described in section 2. We showcase HANS by integrating various types of knowledge and reasoning architectures for ITMo tasks such as congestion detection. Future work on HANS involves fixing current limitations of the architecture and extensive evaluation. We are in the process of extending the system with data from multiple cameras and additional sensors. This will allow us to more adequately address multi-modal information fusion and robustness.

References

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *CoRR*, abs/2005.12872, 2020.

Sreyasi Nag Chowdhury, Ruwan Wickramarachchi, Mohamed H Gad-Elrab, Daria Stepanova, and Cory Henson. Towards leveraging commonsense knowledge for autonomous driving. *SEMWEB*, 2021.

CNN. Here's what's in the bipartisan infrastructure package. https://www.cnn.com/2021/07/28/politics/infrastructure-bill-explained/index.html, 2021.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Chinthaka Gamanayake, Lahiru Jayasinghe, Benny Kai Kiat Ng, and Chau Yuen. Cluster pruning: An efficient filter pruning method for edge ai vision applications. *IEEE Journal of Selected Topics in Signal Processing*, 14(4):802–816, 2020.

Marco Gori, Gabriele Monfardini, and Franco Scarselli. A new model for learning in graph domains. In *IJCNN*, volume 2, pages 729–734, 2005.

Grand View Research. Intelligent traffic management system market size, share & trends analysis report by solution (traffic monitoring system, traffic

signal control system, integrated corridor management, by region, and segment forecasts, 2021 - 2028. https://www.grandviewresearch.com/industry-analysis/intelligent-traffic-management-system-market, 2021.

Armin Haller, Krzysztof Janowicz, Simon JD Cox, Maxime Lefrançois, Kerry Taylor, Danh Le Phuoc, Joshua Lieberman, Raúl García-Castro, Rob Atkinson, and Claus Stadler. The modular ssn ontology: A joint w3c and ogc standard specifying the semantics of sensors, observations, sampling, and actuation. *Semantic Web*, 10(1):9–32, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. arxiv 2015. *arXiv preprint arXiv:1512.03385*, 2015.

Filip Ilievski, Pedro Szekely, Jingwei Cheng, Fu Zhang, and Ehsan Qasemi. Consolidating commonsense knowledge. *arXiv preprint arXiv:2006.06114*, 2020.

Rateb Jabbar, Mohammed Shinoy, Mohamed Kharbeche, Khalifa Al-Khalifa, Moez Krichen, and Kamel Barkaoui. Urban traffic monitoring and modeling system: An iot solution for enhancing road safety. In *IINTEC*, pages 13–18, 2019.

Chan-Tong Lam, Hanyang Gao, and Benjamin Ng. A real-time traffic congestion detection system using on-line images. In *ICCT*, pages 1548–1552. IEEE, 2017.

Chelsea Lancelle. *Distributed acoustic sensing for imaging near-surface geology and monitoring traffic at Garner Valley, California*. The University of Wisconsin-Madison, 2016.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. KagNet: Knowledge-aware graph networks for commonsense reasoning. In *EMNLP-IJCNLP*, pages 2829–2839, Hong Kong, China, 2019. ACL.

Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2020.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Kaixin Ma, Jonathan Francis, Quanyang Lu, Eric Nyberg, and Alessandro Oltramari. Towards generalizable neuro-symbolic systems for commonsense question answering. In *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pages 22–32, Hong Kong, China, 2019. ACL.

RoopTeja Muppalla, Sarasi Lalithsena, Tanvi Banerjee, and Amit Sheth. A knowledge graph framework for detecting traffic events using stationary cameras. In *WebSci*, pages 431–436, 2017.

A. Oltramari, J. Francis, C. Henson, K. Ma, and R. Wickramarachchi. Neuro-symbolic architectures for context understanding. In I Tiddi et al., editors, *Knowledge Graphs for eXplainable Artificial Intelligence: Foundations, Applications and Challenges*, pages 143–160. IOS Press, 2020.

OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org, 2017.

Alessandra Pascale, M Nicoli, F Deflorio, Bruno Dalla Chiara, and Umberto Spagnolini. Wireless sensor networks for traffic management and road safety. *IET Intelligent Transport Systems*, 6(1):67–77, 2012.

Arash Gholami Rad, Abbas Dehghani, and Mohamed Rehan Karim. Vehicle speed detection in video image sequences using cvs method. *International journal of physical sciences*, 5(17):2555–2563, 2010.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *AAAI*, pages 3027–3035, 2019.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

Carles Ventura, Jordi Pont-Tuset, Sergi Caelles, Kevis-Kokitsi Maninis, and Luc Van Gool. Iterative deep learning for road topology extraction. *arXiv preprint arXiv:1808.09814*, 2018.

Denny Vrandečić and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

Simon J Walton, Min Chen, and David S Ebert. Livelayer live traffic projection onto maps. In *Eurographics (Posters)*, pages 37–38, 2011.

Kilian Weinberger, Anirban Dasgupta, John Langford, Alex Smola, and Josh Attenberg. Feature hashing for large scale multitask learning. In *ICML*, pages 1113–1120, 2009.