Multi-Agent System for Multimodal Machine Learning Object Detection[⋆]

Eduardo Coelho[1[0009−0006−9568−854X]], Nuno Pimenta[1[0000−0001−8232−6466]], Hugo Peixoto[1[0000−0003−3957−2121]], Dalila Durães[1[0000−0002−8313−7023]], Pedro Melo-Pinto[2[0000−0001−8257−0143]], Victor Alves[1[0000−0003−1819−7051]], Lourenço Bandeira[3], José Machado[1[0000−0003−4121−6169]], and Paulo Novais[1[0000−0002−3549−0754]]

[1] ALGORITMI Research Centre/LASI, University of Minho, Braga, Portugal
[2] CITAB, University of Trás-os-Montes and Alto Douro, Vila Real, Portugal
[3] Schréder Hyperion
pg47164@alunos.uminho.pt, a88322@alunos.uminho.pt, hpeixoto@di.uminho.pt, dad@di.uminho.pt, pmelo@utad.pt, valves@di.uminho.pt, lourenco.bandeira@schreder.com, jmac@di.uminho.pt, pjon@di.uminho.pt

Abstract. Multi-agent systems have shown great promise in addressing complex problems that traditional single-agent approaches are not be able to handle. In this article, we propose a multi-agent system for the conception of a multimodal machine learning problem on edge devices. Our architecture leverages docker containers to encapsulate knowledge in the form of models and processes, enabling easy management of the system. Communication between agents is facilitated by Message Queuing Telemetry Transport, a lightweight messaging protocol ideal for Internet of Things and edge computing environments. Additionally, we highlight the significance of object detection in our proposed system, which is a crucial component of many multimodal machine learning tasks, by enabling the identification and localization of objects within diverse data modalities. In this manuscript an overall architecture description is performed, discussing the role of each agent and the communication protocol between them. The proposed system offers a general approach to multimodal machine learning problems on edge devices, demonstrating the advantages of multi-agent systems in handling complex and dynamic environments.

Keywords: Multi-Agent System · Multimodality · Multimodal Machine Learning · Object Detection

1 Introduction

The field of computer vision has extensively studied object detection, especially in safety-critical operations such as vehicle detection and tracking, but researchers are now exploring the use of cross-modality features in conjunction

[⋆] We thank Schréder Hyperion Company for the cooperation of this work.

with visual information, making it a relatively new area of research. This approach has been investigated in various research fields related to visual modalities, including video recognition, multimedia comprehension, image and video captioning, and person identification, however, combining multiple modalities poses certain difficulties, such as ensuring diversity among them or acquiring multi-domain sensor data in a synchronized manner [1, 2].

In multi-sensor data fusion, information from several sources is dynamically combined to produce an accurate and insightful picture of the targeted entities. In essence, this method transforms the unprocessed data gathered from many sources into a logical and consistent collection of assessments. The capacity to gain insights that would not be possible with a single sensor in isolation is this approach's main benefit [2].

As a matter of fact, just as humans rely on visual perception to navigate the driving environment, similarly autonomous vehicles depend on visual data in the form of images and videos [1]. Nevertheless, in unfavorable driving conditions, object detection is hindered, and visual data alone may not be sufficient to detect the presence of other vehicles. Consequently, additional modalities may be of value to improve safety-critical operations such as vehicle detection and tracking.

With the advent of the Internet of Things (IoT), the ordinary world becomes surrounded by an increasing number of distinct sensors that generate data from various modalities [3], such as radar, image, infra-red, seismic, and acoustic, among others. Each modality captures the environmental state from a different perspective. For instance, radar signals are immune to adverse weather conditions, such as changes in lighting, rain, and fog, but can be hindered by clutter and multi-path effects, making it difficult to determine target size. Seismic and acoustic signals can detect objects without a direct line-of-sight, but they also struggle with determining object size. Non-image modalities are less susceptible to non-line-of-sight conditions, while images perform better in tracking and determining object dimensions [1]. So, the complementarity of these various modalities in object detection may prepare computer vision algorithms more capable and less error-prone in tasks where a small increase in detection accuracy may be fundamental.

In this article, we propose a multi-agent system for deploying a multimodal machine learning problem on edge devices, using docker containers to contain knowledge in the form of models and processes. The proposed system aims to address this conception of machine learning models while handling the complexity of multimodal data.

A single agent system would not be sufficient to handle all the tasks required, such as data collection, feature extraction, model training, and inference. This would lead to performance bottlenecks, a remarkable lack of scalability, and limited adaptability to different and dynamic environments. In contrast, the proposed multi-agent system allows for seamless coordination and collaboration between agents, each responsible for specific tasks. For example, some agents can collect data from sensors, while others can perform data feature extraction, model training, and/or inference.

2 Related Work

Object detection has been a challenging topic in computer vision research for several decades. Object detection aims to identify whether there are any object representations from specified categories in an image and, if so, to report the spatial location and extent of each object occurrence, for instance, via a bounding box [4]. Several are the techniques and methodologies for implementing unimodal object detection, but each modality depicts the status of the environment from a different angle [1] and computer vision algorithms may be favored in their detection accuracy if more modalities are adjoined.

Regarding multimodal data usage for object detection, some works implement the cooperation between different modalities in order to achieve higher object detection performance. For example, Chernov et al. [5] and Gang et al. [6] employ vision and acoustic data, where the former identified faults on the surface of welds that occur after the metal pipe welding operations, demonstrating the effectiveness of the suggested technology and also minimizing the number of required observations, and the latter used multimodal fusion at the feature level for solid-waste sorting, reporting that the feature-based fusion strategy outperformed the decision-based methodology. When compared to single modality utilization, both studies proved the accuracy enhancement of multimodal data employment.

In contrast, other modalities' cooperation may be studied, such as the Light Detection And Ranging (LiDAR) – camera, where approaches can be very distinct [3, 7, 8]. In particular, Guo et al. [7] handle bird's eye view of LiDAR point cloud and RGB image, retrieving their features and then fusing them with a deep multi-scale fusion method that combines region features for each input. Alternatively, Xu et al. [3] employ data from 2D pictures and 3D point clouds to detect barriers accurately in 3D in autonomous driving scenarios. The developed framework exploits 2D and 3D segmentation approaches to retrieve semantic information, which is then adaptively fused with an attention-based semantic fusion component and fed to a 3D detector. Furthermore, Gao et al. [8] refer a multi-scale feature merger module that divides the feature representation of multi-modal information utilizing multi-scale convolution and determines the weight of each modal feature channel. Addressing the problem of fusing image and point cloud due to the difference in data structure, they implement the idea of multi-scale convolution and selection kernel.

Many of the studies involving multisensory object detection are directed to autonomous driving problematics, but other fields of application exist, such as healthcare [9] or emotion analysis [10]. Furthermore, more than two modalities can be combined, although the system's complexity tends to increase, as more care must be employed.

In this sense, Roy et al. [1] developed a fusion-based non-image and image inference pipeline to surpass autonomous driving vehicle's visual sensors inability in non-line-of-sight situations. The researchers identified distinct vehicles using seismic, acoustic, radar, and image modalities, and their innovation lies in presenting and experimenting with different fusion approaches that can intelligently determine the most relevant sensors on a case-by-case basis. Results presented show interesting improvements compared to single modality approaches. Similarly, in the work of Mirzaei et al. [2] is introduced a multisensory data fusion approach via acoustics, infrared camera, and marine radar into an avian monitoring system, illustrating the habits of birds and bats in a proposed wind farm construction location. Single-sensor data is processed independently to identify and track targets and acquire their attributes, which are subsequently employed in the fusion process. For the integration of the infrared camera with the radar, a first stage (feature level) fusion occurs, followed by a second stage (decision level) fusion between the infrared camera, radar and acoustics.

3 Multi-agent System

Figure 1 represents the general multi-agent system and how agents interact from data acquisition until the sensor processing combination, in order to obtain relevant conclusions. System's agents are encapsulated as docker containers to represent models and processes that occur at the different stages. Moreover, a docker container is a self-contained and lightweight software package that encompasses all the essential elements required to run an application, such as code, runtime, system tools, system libraries, and configurations. By isolating software from its environment, containers ensure that applications function consistently, regardless of variations between development and staging environments [11].

Starting from the top, the schematic represents just a few of the possible sensory sources that can capture data from the environment, as sensors may be of several other types, for instance: location sensors (e.g. GPS, active badges), pressure sensors (e.g. barometer, pressure gauges), wearable sensors (e.g. accelerometers, gyroscopes, magnetometers), vital sign processing devices (heart rate, temperature), motion sensors (e.g. radar gun, speedometer, mercury switches, tachometer), among many others [12]. These are the primary source of information, and this data is perceived by sensory specific agents (represented in blue), whose responsibility is pre-processing such information, either through validation, standardization, format conversion or others. Then, they redirect data to agents with a specific know-how about that sensory data.

Next, such knowledge base agents (represented in dark blue) act in receiving the processed data and transforming it in knowledge capable of being valuable to a following information merging step. At this transformation stage, data may encompass actions from if-then-else rules or simple characteristics analysis all the way through complex machine learning or deep learning methodologies. It is important that these agents may operate under continuous training conditions

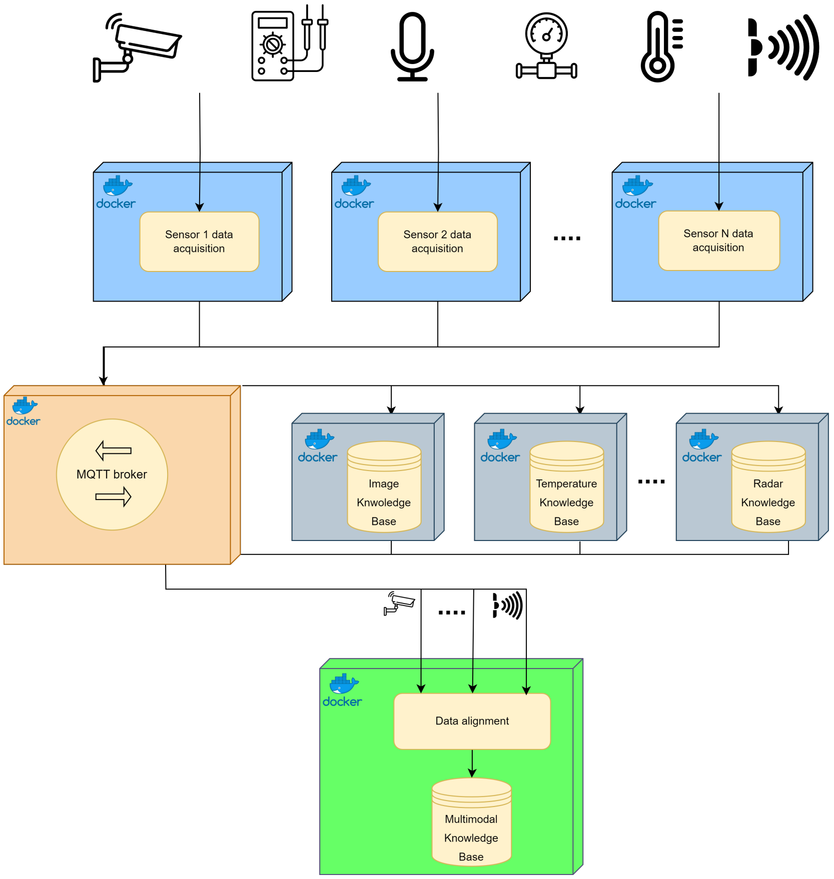
Sensor 1 data acquisition
Sensor 2 data acquisition
....
Sensor N data acquisition
MQTT broker
Image Knwoledge Base
Temperature Knowledge Base
....
Radar Knowledge Base
....
Data alignment
Multimodal Knowledge Base

Fig. 1. The overall multi-agent system scheme, from data acquisition to the multimodal knowledge base.

by retaining information from the other processing agents that can be useful to enhance their own capabilities, learning to generate more robust judgements [6].

After individual data examination, a knowledge merging agent (represented in green) should proceed in order to correctly join the processed information from all data sources and return a plausible final result. Actions conducted may include an initial data alignment step (through timestamp information, for instance) or consecutive fusion levels between more similar modalities.

More specifically, the process of multi-sensor data fusion involves the automatic integration of information from different sources to create a comprehensive and useful description of the desired targets. This technique converts raw data obtained from multiple sources into a cohesive set of inferences and its primary advantage is the ability to obtain information that may not be available from a single sensor [2]. Generally, fusion frameworks are based on feature-based and decision-based fusion principles. Feature-based fusion relies on a deep learning approach that automatically assigns higher weights to more relevant modalities during training. In contrast, decision-based fusion involves setting specific rules to prioritize more relevant modalities over others [1].

Just for stating some examples, fusion techniques may range from simple concatenation to more complex algorithms, such as low rank multimodal fusion [1]. The complexity of the intended approach depends on the nature of the problem in hands.

Besides this decision-level and feature-level resolution, researchers are also faced with the question on when (at which stage) to fuse information. There are several methods for multi-modal detection of objects in the literature on computer vision, including early fusion, deep fusion, and late fusion approaches. Prior to putting information into the detection framework, early fusion approaches seek to integrate the raw sensor data, producing a distinct sort of data. To get the best results, this strategy often needs pixel-level correlation between each type of sensor data. In contrast, late fusion approaches aggregate the detection results at the bounding box level after doing the detection for each kind of data independently. In its turn and in contrast to the first two methods, deep fusion-based methods take features from various deep neural network types and then fuse them at the feature level. A robust model that can handle a variety of sensor inputs and improve detection accuracy may be made using the deep fusion technique. Each fusion methodology has benefits and drawbacks, and the choice of technique relies on the particular application and the kind of sensor data that are accessible [3].

Lastly, to maintain communication between agents within the system, a broker should be utilized. Due to being lightweight and appropriate for usage on all devices, including low power single board computers, the Eclipse Mosquitto [13] message broker is suggested. This open-source message broker implements the Message Queuing Telemetry Transport (MQTT) protocol [14], a standard messaging protocol for the Internet of Things (IoT) developed as a very lightweight publish/subscribe messaging transport for linking remote devices with a minimal code footprint and low network bandwidth, thus making it suitable for low power sensors or mobile devices. By following the principle of message publication and topic subscription, numerous clients can link with the broker and subscribe to their preferred topics to either publish or read messages. There can be several clients who may subscribe to the same topics and utilize the information as they please. Essentially, the MQTT broker acts as a universal and uncomplicated interface for all connections [13].

4 Discussion

Upon analyzing the pertinent literature, including the aforementioned related works, it becomes readily apparent that there is a significant similarity between the aforementioned architecture and the general architectural design pursued by the latter. Typically, the majority of studies related to multimodal machine learning object detection involve a process flow that initiates with data acquisition, progresses to information extraction, and culminates with some form of fusion. The overall objective of the study is not important, as this process flow suits autonomous driving problems [1, 3, 7, 8], avian monitoring [2] or even material surface inspection [5].

At the first stage, approaches on data retrieval and its pre-processing are generally similar and simple, with data maps [7], semantic information [3] or, more commonly, feature extraction [1, 5]. These behaviors are included among actions performed by sensory specific agents represented in Figure 1.

Then, this pre-processed information may cross through unimodal machine learning mechanisms in order to retrieve deeper and more conclusive information about the data, which corresponds to practices that can be performed by the previously denominated knowledge base agents.

Finally, the multimodal merging typically arises and it is at this step that researchers are more free to be creative and analyze which approach may be more suitable, as problems may advance through several possible and different approaches. Some perform fusion with all (or almost all) modalities at once [1], others resort to several consecutive fusion levels between modalities that are more analogous [2, 5] and it also possible to project own alternatives that seem more suitable [3, 7, 8].

For the purpose of presenting a possible practical application scenario to the aforementioned system, the work of Roy et al. [1] may be used as a suitable example. As Figure 2 depicts, a problem related to multi-modality sensing and data fusion for vehicle detection is presented, combining visual data from cameras and non-visual data from seismic, radar and acoustic sensors. In order to recognize multi-vehicle structures, the system employs a two-stage detection technique that first identifies individual cars in each modality individually. Through tests utilizing actual data, the authors illustrate how the suggested system works better than single-modality detection techniques.

Examining Figure 2, it is believed that an easy association can be established with Figure 1, as its distinct types of agents are reflected in the example study. Firstly, the multi-domain sensing and preprocessed data tasks represented in the image are closely related to functions inherent to the sensory specific agents. In both cases, raw data is captured from the sensor and preprocessed accordingly to requirements further imposed by knowledge base agents. These last agents see their action be represented by the unimodal networks developed by the authors. In this sense, knowledge networks specialized in each of the sensory sources are developed in the interest of extracting the most information about each modality. Evidently, the input of each network is provided by the sensory specific agents and the output is prepared with the goal of best-suiting the entrance in the fusion network. As for these fusion networks, they undoubtedly illustrate one of the possible approaches delineated for the knowledge merging agent, the last component of Figure 1. In this example, the authors opted to merge non-image modalities in the first place, inputting the resulting prediction together

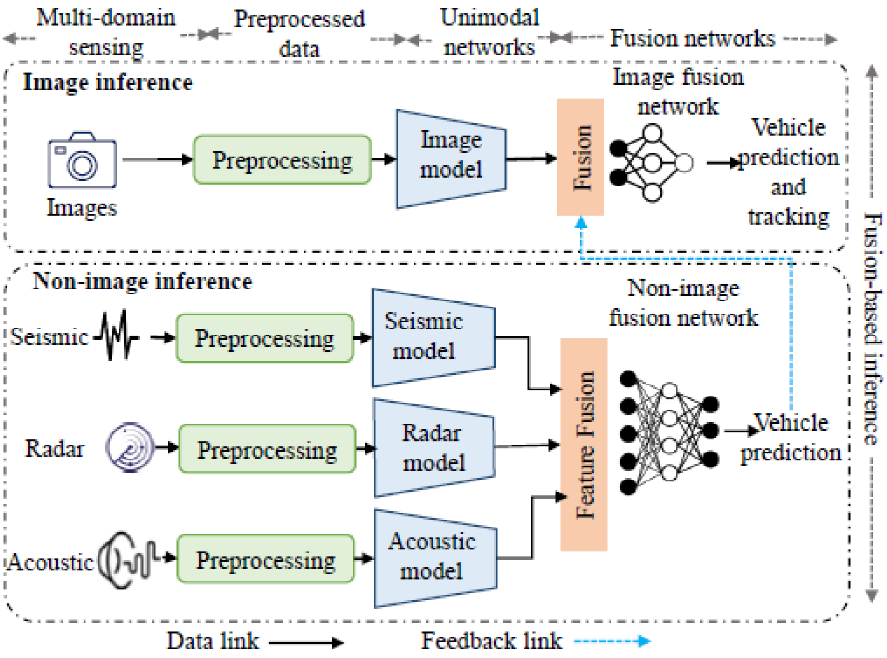

Fig. 2. Illustrative scheme of a practical approach to the proposed architecture. Taken from [1].

with image data in a last fusion network, reaching a final vehicle prediction and tracking outcome.

5 Conclusion

This article presents in detail the overall architecture of a multi-agent system for multimodality use, highlighting the role of each agent and the communication protocol that may be established between them. This system offers a general approach to multimodal machine learning problems on edge devices for object detection scenarios, illustrating the advantages of multi-agent systems in handling complex and dynamic sensory environments.

It demonstrates a similar process flow that other multimodal machine learning object detection studies may follow, highlighting the importance of data acquisition, pre-processing, and fusion for addressing complex problems. By adopting a comparable architecture, researchers can effectively utilize the knowledge base of previous studies to achieve their objectives while also incorporating their own creative solutions.

The complexity of multimodal data in machine learning issues presents obstacles that cannot be effectively handled by a single-agent system. These concerns are addressed by the suggested multi-agent system. Each agent in the proposed system is in charge of a particular assignment, allowing for effortless coordination and collaboration between them. This promotes scalability and flexibility in response to changing circumstances. The administration and communication inside the system are made easier by the usage of MQTT and Docker containers. Health care, autonomous driving, and emotion analysis are just a few of the areas in which the suggested technology might be used.

As future work, the presented multi-agent system is intended to be tested in practice against different arbitrary object detection scenarios, with different sensory sources and different interactions among agents.

Acknowledgements

This work has been supported by FCT (Fundação para a Ciência e Tecnologia) within the R&D Units Project Scope: UIDB/00319/2020.

References

1. D. Roy, Y. Li, T. Jian, P. Tian, K. Roy Chowdhury and S. Ioannidis: Multi-modality Sensing and Data Fusion for Multi-vehicle Detection. IEEE Transactions on Multimedia
2. G. Mirzaei, M. M. Jamali, J. Ross, P. V. Gorsevski and V. P. Bingman: Data Fusion of Acoustics, Infrared, and Marine Radar for Avian Study. *IEEE Sensors Journal*, vol. 15, no. 11, pp. 6625-6632, Nov. 2015
3. Xu, Shaoqing, et al: Fusionpainting: Multimodal fusion with adaptive attention for 3d object detection. 2021 *IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2021
4. Liu, L., Ouyang, W., Wang, X. et al. Deep Learning for Generic Object Detection: A Survey. *Int J Comput Vis* **128**, 261–318 (2020)
5. Chernov, A.V., Savvas, I.K., Alexandrov, A.A., Kartashov, O.O., Polyanichenko, D.S., Butakova, M.A., Soldatov, A.V.: Integrated Video and Acoustic Emission Data Fusion for Intelligent Decision Making in Material Surface Inspection System. Sensors 2022, 22, 8554
6. Lu, G., Wang, Y., Xu, H. et al.: Deep multimodal learning for municipal solid waste sorting. *Sci. China Technol. Sci.* **65**, 324–335 (2022)
7. Rui Guo, Deng Li, Yahong Han: Deep multi-scale and multi-modal fusion for 3D object detection. Pattern Recognition Letters, Volume 151, 2021, Pages 236-242, ISSN 0167-8655
8. Xin Gao, Guoying Zhang, Yijin Xiong: Multi-scale multi-modal fusion for object detection in autonomous driving based on selective kernel. Measurement, Volume 194, 2022, 111001, ISSN 0263-2241
9. Cui, Can, et al.: Deep multi-modal fusion of image and non-image data in disease diagnosis and prognosis: a review. Progress in Biomedical Engineering (2022)
10. Pandeya, Y.R., Lee, J. Deep learning-based late fusion of multimodal information for emotion classification of music video. *Multimed Tools Appl* **80**, 2887–2905 (2021)
11. Use containers to Build, Share and Run your applications, https://www.docker.com/resources/what-container/. Last accessed 24 Apr 2023

12. Simon Elias Bibri: The IoT for smart sustainable cities of the future: An analytical framework for sensor-based big data applications for environmental sustainability. Sustainable Cities and Society, Volume 38, 2018, Pages 230-253, ISSN 2210-6707
13. Eclipse Mosquitto™ An open source MQTT broker, https://mosquitto.org/. Last accessed 18 Apr 2023
14. MQTT: The Standard for IoT Messaging, https://mqtt.org/. Last accessed 18 Apr 2023