Multimodal Explanations for AI-based Multisensor Fusion

Dave Braines
IBM Research, Hursley Park, Winchester, Hampshire, UK
dave_braines@uk.ibm.com

Alun Preece
Crime and Security Research Institute, Cardiff University, Cardiff, UK
PreeceAD@cardiff.ac.uk

Dan Harborne
Crime and Security Research Institute, Cardiff University, Cardiff, UK
HarborneD@cardiff.ac.uk

ABSTRACT

The recent resurgence in the effectiveness of artificial intelligence (AI) and machine learning (ML) techniques for image, text and signal processing has come with a growing recognition that these techniques are "inscrutable": they can be hard for users to trust because they lack effective means of generating explanations for their outputs. Consequently, there is currently a great deal of research and development addressing this problem, producing a sizeable number of proposed explanation techniques for AI/ML approaches operating on a variety of data modalities. However, a problem that has received less attention is: what modality of explanation to choose for a particular user and task? For example, many techniques attempt to produce visualizations of the workings of an ML model, e.g., so-called "saliency maps" for a deep neural network, but there may be multiple reasons why this mode of explanation might not be appropriate for a user, including: (i) they may be operating at the edge of the network with a device that is not suited to receiving or displaying such a visualization; (ii) it may not be appropriate for security reasons to send them a visualization derived from the source imagery (e.g., if the location of the camera system is sensitive); (iii) this kind of explanation may be "too low level" for that user's needs – they may require something more "causal", for example. One approach that may address all three of these example issues would be to map the explanation from a visualization to a textual rationalization. In this paper we explore this issue of generating explanations in a range of modalities in the context of AI/ML services that operate on multisensor data and show that a "grammar-based" approach that separates atomic explanation-generation and communication actions offers sufficient scope and flexibility to address a set of mission scenarios.

1.0 INTRODUCTION

The recent resurgence in the effectiveness of artificial intelligence (AI) and machine learning (ML) techniques for image, text and signal processing has come with a growing recognition that these techniques are "inscrutable": they can be hard for users to trust because they lack effective means of generating explanations for their outputs. Consequently, there is currently a great deal of research and development addressing this problem, producing a sizeable number of proposed explanation techniques for AI/ML approaches operating on a variety of data modalities, and generating explanations of various modalities.

In this paper we investigate these multimodal explanation types and outline the initial development of a conceptual model to represent explanations and key related concepts. This conceptual model is available for use by both human and machine agents, underpinning the development of a simple conversational interface for the exploration of explanations. We define a simple scenario, describe a publicly available dataset with multi-modal derivatives that are useful resources for this work, and three specific services that are able to generate higher-level information to support situational understanding. This work draws together threads of our previous research work, describing the integration of these to provide a useful overall system.

In Section 2 we summarise key background work from ourselves and others, the fusion of which is the basis for this paper. In Section 3 we define a simple set of three services, comprising a multimodal information fusion system for traffic monitoring. These are used as a worked example throughout the remainder of the paper. Section 4 outlines the conceptual model underpinning this work and describes a series of conversational examples, highlighting the purpose of each in the context of explanations. Section 5 talks briefly about related work, and Section 6 concludes the paper.

2.0 BACKGROUND WORK

The multimodal explanation examples presented in section 4 are defined in the context of a number of threads of research that we have been undertaking on the DAIS ITA (Distributed Analytics and Information Science International Technology Alliance) research program. All of this research has been to support the requirement for Coalition Situational Understanding (CSU) in complex multi-partner operations with shared datasets, systems and services, operating at the edge of the network in high-tempo environments. Each of these aspects of the background work is summarised in the remaining sub-sections.

2.1 Example scenario and dataset: Traffic congestion and CCTV

In our previous research into CSU we have taken traffic-related CCTV still imagery and video as a dataset that is highly relevant for image classification machine learning techniques [1]. This dataset provides a strong core for our ongoing research, from which we are able to generate a number of derived based on extracted information. The overall setting for our research work is a distributed system operating at the edge of the network, with multiple coalition partners working together to share datasources, sensors and services in support of common coalition goals. In this setting the core CCTV still imagery and video will be potentially sourced from multiple coalition partners, for example due to different areas of operation. It may be that one of the partners is the host nation and may have significant fixed infrastructure in place, whereas the other coalition partners have more opportunistic capabilities, such as mobile image and video sensors mounted on vehicles or personnel. In our specific example we use the publicly available traffic related CCTV still imagery and video from TfL (Transport for London) which comprises live feeds from over 300 cameras spread across the greater London metropolitan area1. This core dataset gives us imagery and video related to roads and our research task is to investigate the manner in which congestion can be detected or inferred from this datasource, or datasources derived from it. We are especially interested in ensembles of services that operate together, either in chains, with each adding incremental value, or as corroborating services that can be used together to potentially improve confidence in the results of the analysis. Our overall goal for this dataset is to derive situational understanding relating to traffic congestion: Can we detect or infer congestion from these datasources and can this capability be used to determine pattern-of-life (and therefore predictive) capabilities for congestion?

2.2 Explanation-oriented architecture

Given the ability to detect or infer traffic congestion we are then faced with issues such as trust and confidence. These are especially relevant in our coalition setting where datasources and services will be shared between coalition partners. Our work also provides the potential to rapidly assemble sets of services and datasources together in new combinations which is again a situation in which trust and confidence in the resulting information are critical. Figure 3 shows a testbed system architecture which we have developed in earlier work [2]. The dark grey arrows show information flow leading to congestion ratings; light grey arrows show information flow to generate explanations for ratings. This is an example of an information fusion system involving multi-modal data with a direct focus on providing explanation-related capabilities for the various services. This desire for agile environments in which to rapidly assemble new services, based on the unfolding situation on the ground, motivates our research into the role of explanations and the manner

[1] See http://www.tfljamcams.net/ with 327 separate CCTV cameras providing live imagery & video (checked on 11-Oct-2018)

in which these explanations can be formed and communicated to human users within the system. As was observed previously, it is often hard for machine learning processes to provide explanations of their processing directly as a result of that processing. Numerous techniques are being actively investigated by the research community to provide explanation capabilities for these "black box" processes. Our research is specifically investigating some of these but is mainly focused on the development of an encompassing framework within which multiple explanation techniques can be integrated [2].

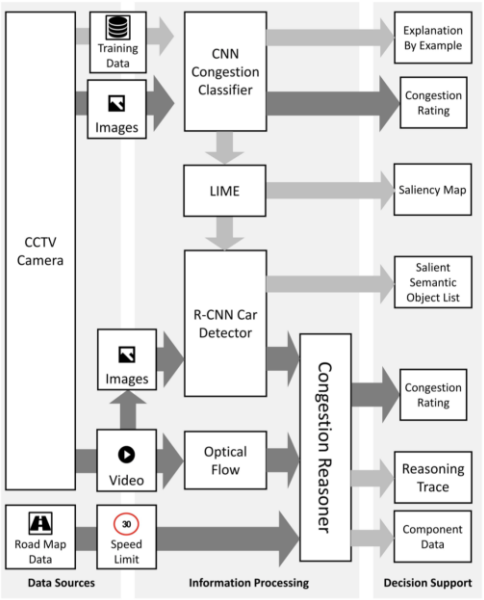

Figure 1: Taking the CCTV traffic congestion scenario, datasource and services and applying this into explanation-oriented operating context (from [2])

Our aim is the development of a generic architecture within which any set of datasets can be used, with any set of services (in sequence, parallel or any combination thereof). Within this framework we see the role of explanation, and therefore the different explanation techniques, being a fundamental capability. This is an "explanation-oriented architecture". In our complex, coalition-led environment of rapidly assembled services it is not possible to treat the need for explanations as a simple function that can be bolted on. Instead the need for explanation and the mechanisms by which it can be achieved are a fundamental consideration and are as important as the datasets and services within the architecture.

2.3 Explanation types

The explanation-oriented architecture provides us with a framework for characterising and defining the explanation-related capabilities of the system. These may take multiple forms [3], for example: Some existing services may be able to provide explanations already, as part of their existing processing: When a service uses logical inference rules or code-based rules to reach a conclusion then the explanation could take the form of a simple description of those rules alongside the input data. This is a form of transparent explanation, where the processing itself is able to be used as the source of the explanation. Services such as these may already have been designed with an "explanation mode" in mind, but in many cases, especially with COTS (Commercial Off-The-Shelf) software it is likely that such explanations will not be possible since the code that is making the decisions is not explicitly exposed for explanations. In cases where the internal processing of the service is unable to provide transparent explanations (e.g. machine learning based services), or in cases where there is no ability to get to such explanations, then post-hoc explanation techniques must be used. These are techniques that involve using results from the service in some way, usually along with the service itself, to attempt to determine why that particular result was given. For

example, in the case of image classification, this may be the generation of a "saliency map" which highlights the parts of the image that most strongly influenced the classification. Other forms of post-hoc explanation exist, for example: explanation by example, where the explanation is given in the form of another example input that caused the same output. The human equivalent to this is the use of analogy when trying to explain a new or unfamiliar concept to someone. A detailed discussion and definition of these explanation types can be found in [4]. The ability to provide multi-modal explanations, or to switch modalities for the explanation is also of significant interest. For example, in [4] the role and purpose of text explanations is explored. The ability to provide a textual explanation for the results of a deep learning classification model on input imagery may be more valuable to human users, especially less technical and more business focused users, than some imagery-based saliency map. The use of textual responses also provides an abstraction mechanism within which additional relevant information, for example drawn from conceptual models of the problem domain, can be brought into the explanation to help raise the level at which the explanation is being conveyed. There may also be more practical considerations such as the capability of the user's device or the available bandwidth that might also motivate the use of a textual (or verbal) explanation for some users.

2.4 Conversations and roles

Another relevant aspect of our earlier research work investigates the manner in which human-machine systems are composed, with a particular focus on providing open and extensible systems that can be rapidly evolved, ideally in real-time. This work investigates the differences between traditional user interfaces and conversational interfaces [5] defining a generic conceptual model for supporting conversations between human and machine agents, and reports results from an experiment to evaluate this approach in an instrumented household setting. In related work, and as shown in Figure 2, we have proposed that when it comes to interpretability, the role of the agent attempting to obtain the interpretation is a fundamental consideration [6]. It is important to understand how an agent's role influences its goals, and the implications for supporting interpretability in this context. This role-based model of interpretability is potentially useful to a wide variety of communities, including: interpretability researchers, system developers, and regulatory bodies auditing machine learning systems. Figure 2 shows the key roles defined in this model with the direction of arrow indicating the direction of interaction (e.g., data-subjects do not interact with the system, but the system has their data). In the example conversations defined in section 4 we briefly describe the role of the user in each case but for this particular work we have not attempted to define example conversations for all of the roles2 outlined in the model shown in Figure 2.

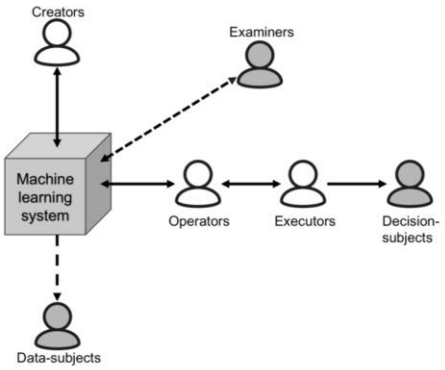

Figure 2: Roles will influence how interpretation or explanation can be achieved (from [6]).

[2] In the limited examples given in Section 4 of this paper all of the roles are either operators or executors, but the conceptual model that we are developing to support conversation and explanation must support all possible roles.

These two approaches are brought together in this paper: Using a conversational interface to explore the space of explanations available from our coalition ensemble system, considering the role of the user to determine the kind of explanation appropriate to them. The framework proposed here considers additional contextual factors such as the device they are using and the bandwidth that they have available (for example to rule out certain explanation types as unfeasible for the current operating conditions).

3.0 WORKED EXAMPLE: THREE SIMPLE SERVICES

Based on the operating context shown in Figure 1, we define three specific simple services that can be used against the traffic-related CCTV video and imagery core data. These services have been developed in our test environment but are not proposed as complete or correct to be used in a real environment. For full details of these services refer to [2]. The short summaries below contain only the information relevant to the high-level scenario defined within this paper:

- **Congestion Image Classifier (CIC)**[3]

  This service operates directly on the still CCTV imagery. It uses a single trained Convolutional Neural Network (CNN) across all cameras and images regardless of day/night or weather conditions. This has been trained on suitable test images and performs at a reasonable level of accuracy. Considerations relating to what constitutes "congestion", whether this is accurately detected, how common scenarios such as one direction of traffic being free-flowing and the other being congested, etc, are not considered here. The service output is a two category classification: "congested" or "not congested", with a scalar value for the perceived degree of congestion (from 0 to 1).

- **Entity Detector (ED)**[4]

  This service again operates on the still CCTV imagery, with a single model being used across all cameras, images and conditions. This can detect certain types of objects within a scene and has been trained on congestion-relevant objects such as cars, trucks and buses as well as common objects not related to congestion, such as people, trees, bushes and road signs. The output of this service is a list of detected entities, their type, their position within the image and the confidence for each.

- **Congestion Speed Classifier (CSC)**[5]

  This service is comprised of two inner services which operate in sequence: The first operates on the CCTV video and detects the speed of moving objects within the image. It does not attempt to identify what kind of object is moving but it is able to estimate the speed of movement and to detect multiple concurrently moving objects within the video, regardless of direction. The output of this inner service is a list of entities each of which has a relative velocity and a confidence level. This output is fed into the second inner service which is aware of the speed limit of the road that is being observed by each CCTV camera. It is also encoded with a simple classification rule which states: "*if objects are moving at 75% of the speed limit, or above, then the road is not congested, otherwise it*

[3] Shown in Figure 1 as the "CNN Congestion Classifier" information processing service.

[4] Shown in Figure 1 as the "R-CNN Car Detector" information processing service, with the scope broadened in this paper to include the detection of other semantically relevant objects in addition to cars.

[5] Shown in Figure 1 as the "Congestion Reasoner" information processing service, with the "Optical Flow" information processing service as an input.

is congested"[6]. These two inner services in conjunction generate the same classification labels as the first Congestion Image Classifier: "congested" and "not congested" with an associated confidence.

Each of these three services are able to contribute a status regarding whether the sensor data (CCTV imagery or video) indicates the road is congested or not. The two classifier services (CIC and CSC) can directly declare a congested/not-congested status from the same core datasource but using different techniques and through the generation of higher-level derived data in the case of the CSC. The entity detector service (ED) cannot directly declare a congested/not congested status but can be used to provide further insight or evidence to support either of these classifications from the other services.

In Figure 3 we show a simple schematic for part of a fictional city which is under the jurisdiction of US and UK coalition partners. There are three checkpoints (A, B, C), each of which has a number of CCTV camera assets which can be used to determine whether the route through the checkpoint is congested or not. Checkpoint C is on the boundary between the two areas of operation for each of the partners, and the raw CCTV imagery and video from the CCTV cameras will not be shared across coalition boundaries, but services to provide congestion status information will be.

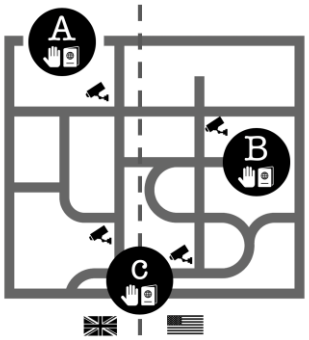

Figure 3: Coalition checkpoints and sensors across the city

In our simple scenario a UK coalition member is asked to plan a vehicle convoy that must pass from checkpoint A to checkpoint B via checkpoint C. They need to identify whether there is congestion at any of the checkpoints and will not initiate the convoy until there is no congestion. This user can access an online conversational system to ask questions about the traffic situation and request imagery from cameras. The system is aware of the user role, device and affiliation and is able to provide congestion-related information through real-time usage of the three simple services described earlier.

4.0 USING CONVERSATION FOR EXPLANATION

Given that the request for an explanation usually follows some earlier statement or assertion we choose to characterise the act of explanation as a conversation. In terms of implementation this "conversation" could be built as a traditional user interface with screens and widgets, but for the purposes of our research we prefer to remain in the abstract conversation space without the need to translate that into specific user interface actions or designs. We are exploring a text/chat-based interaction mechanism similar to conversational interactions that most people are familiar with as a result of SMS messages, social network platforms and even email. The examples in this paper take this abstract form, and where needed the additional modalities (e.g. imagery) are embedded within that medium, for example as embedded imagery

[6] Any real service would have issues differentiating between scenarios of an empty road and a totally congested road, both of which have no moving objects. This is where fusion with the Entity Detector (ED) service can be very useful, but such detail is outside the scope of this paper.

within the textual response. As outlined earlier, the device or network conditions of the user may prevent them from receiving imagery or other complex responses and the model that we have defined allows these restrictions to be considered when the system decides how to best respond to the user.

In earlier work we have used high-level conceptual models of a domain to provide shared human/machine knowledge graph representations. This approach is used here, with Figure 4 showing a high-level summary of the conceptual model of datasets, models, explanation types and other supporting concepts.

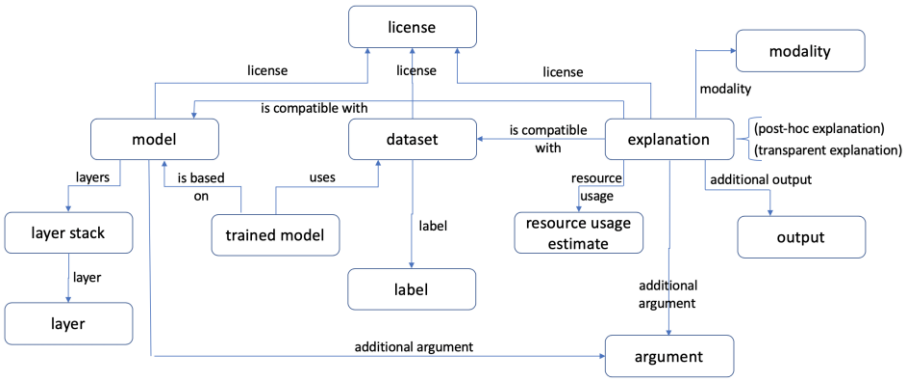

Figure 4: Conceptual model to support contextual explanation

This conceptual model enables the system to determine what modality of explanation to choose for a particular user and task. For example, many explanation techniques produce visualizations of the workings of an ML model, e.g., so-called “saliency maps” for a deep neural network. These have an image modality, but this mode of explanation might not be appropriate for a user, for example:

- they may be operating at the edge of the network with a device that is not suited to receiving or displaying such a visualization.
- it may not be appropriate for security reasons to send them a visualization derived from the source imagery (e.g., if the location of the camera system is sensitive).
- this kind of explanation may be too low level for the user; they may require something more causal.

This system is aware of the user role and/or device capabilities and can conclude that they cannot handle image modality, leading to the need for an alternative. One approach for issues such as these is to map the explanation from a visualization (image modality) to a textual rationalization (text modality).

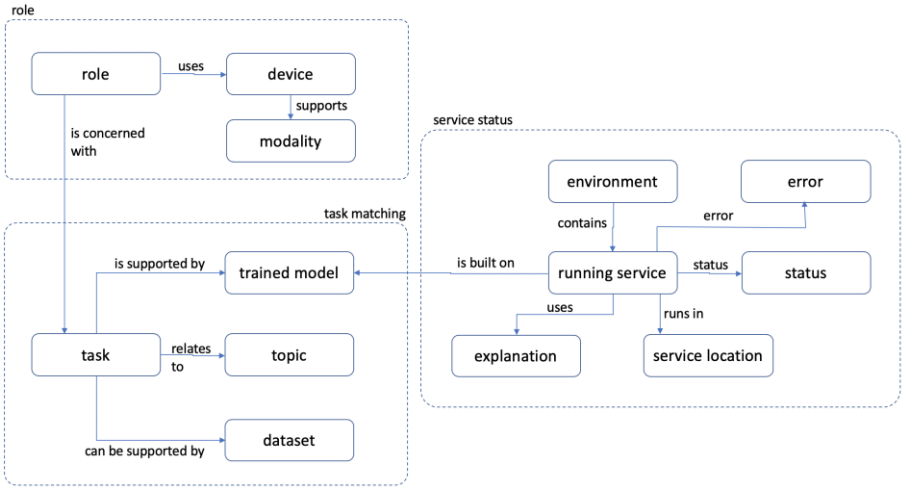

Figure 5: Higher level conceptual models

Figure 5 is a simple set of higher-level conceptualisations built on top of the core explanation model. It is these higher-level connections that enable insights and decisions to be made autonomously by the system. For example, whether a particular service can be used, based on its availability; or whether a user is able to make a particular request, based on their affiliation, role and device. The following sub-sections provide three simple worked-examples of a conversation featuring explanations in different contexts:

4.1 Case 1: Fully transparent explanation

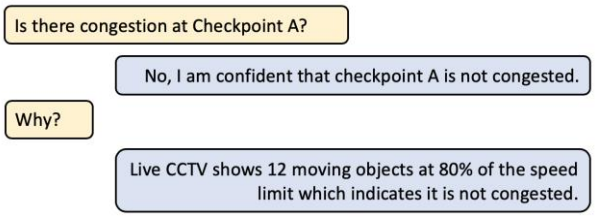

Figure 6: Fully transparent explanation

In this example the human user has asked whether there is congestion at checkpoint A. The response is definitive: the system is confident that there is no congestion. Upon asking for an explanation the human user is provided with a transparent explanation based on usage of the Congestion Speed Classifier (CSC) service: numerous moving objects were detected, and they are moving at 80% of the speed limit. This information is relayed directly to the user and shows the inner workings of the rule-based system and is therefore a ***transparent[7] explanation from a rule-based system[8]***.

[7] Strictly speaking the explanation given is actually post-hoc since it is a convenient textual summary of the processing that was executed to achieve the classification result. For "pure" transparency we should really show the raw rules, premises and conclusions as executed by the machine, but for reasons of simplicity we have chosen this mechanical summarisation approach but retain the term "transparent" since it is very close to full system transparency.

[8] The transparency ends at the input to the rule-based system which is only one half of the composite service. There is no explanation given as to how the moving objects were detected and their speed calculated. This is within scope of our proposed conceptual model and framework, but outside the scope of this paper.

4.2 Case 2: Post-hoc explanations

There are two variants within this one example, both of which result in a post-hoc explanation, with the system able to determine which is the correct technique to use in each case.

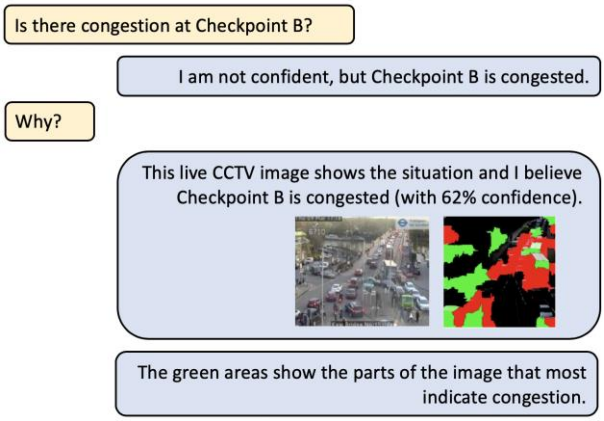

Figure 7: Post-hoc explanation via saliency mapping

In the first case (See Figure 7) the user is told that the system has borderline confidence that the checkpoint is congested. Upon asking why the system responds by showing the user a saliency map (highlighting the areas of the image that were most relevant in concluding the "congested" classification by the CIC service), using clear language to convey the confidence of that classification (low). The user concludes that the system has correctly identified congestion when they are shown the image, but it is the raw image rather than the saliency map explanation that convinces the user. This is an example of ***post-hoc explanation via saliency mapping***, with the saliency map being generated by the LIME technique [7].

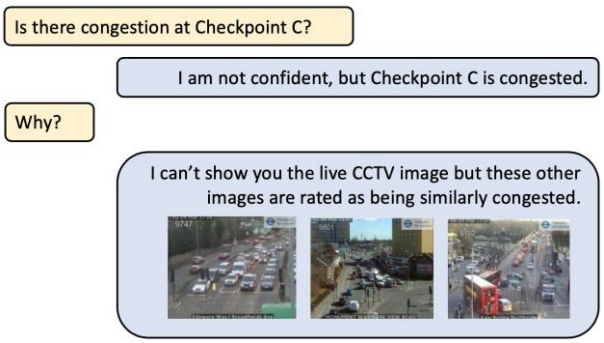

Figure 8: Post-hoc explanation by example

In the second case (See Figure 8) the user gets the same response but upon asking for an explanation the system concludes (through their role and affiliation) that they are not authorised to see the original image for security reasons. The system is therefore not authorised to show the image, or anything derived from it, in order to explain the classification to the user. The system determines that the user has a device which is able to handle imagery with suitable bandwidth, and therefore chooses to show a series of images that are similarly congested to the classified source image. This is achieved by using the scalar value for the degree of congestion detected in the image, showing other images that the user does have the authority to view which have a similar level of detected congestion. An alternative to this would be to show the textual results of the ED service if the user were unable to accept imagery-based explanations. The user finds it hard to

conclude whether the system has correctly identified congestion but sees that the example images served are indeed congested. This is a form of ***post-hoc explanation by example***.

4.3 Case 3: Disagreement within services

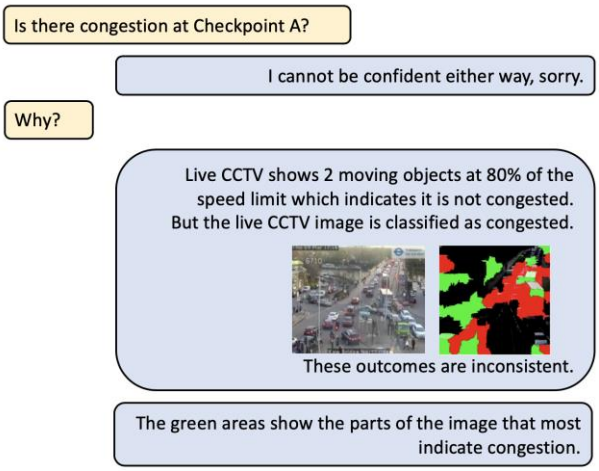

Figure 9: Combined explanation following inconsistency due to disagreement within services

In this case, as shown in Figure 9, the user is told that the congestion status is unknown due to inconsistent information. For an explanation the user is advised that the CSC has concluded "no congestion", based of the speed of objects within the video, whereas the CIC has concluded congestion. The saliency map explanation is again provided since the user is authorised to see the image and in this case the user is able to make a judgement for themselves. This is a combination of ***transparent explanation from a rule-based system***, and ***post-hoc explanation via saliency mapping***. The ability to detect inconsistencies across services is useful for alerting to possible cases of misclassification and the ability to do so will increase as the number of data sources and relevant services increases within the overall system.

5.0 RELATED WORK

There is significant insight to be gained from the literature of social science and how this can be applied to AI systems in general [8], but also to conversation and explanation directly. Such insights clearly apply to the human users within hybrid human/machine environments such as these, but may also be applicable to the machine agents. In [3] the development of a grammar is proposed to enable the analysis and development of user interfaces to aid interpretability. Whilst the grammar is likely too low a level of detail (within the layers of neural networks) it is a powerful concept that warrants further consideration in the future. Our work has also been inspired by the concept of affordances [9], especially in terms of separating the specific benefits brought by the machine and human agents within the hybrid system. For example, the machine agent ability for handling large volumes and being able to perform bias-free analytics, thereby reducing the burden on the human agents in those respects. Earlier work in Human Computer Collaboration that focused on natural communication, shared representation and manipulation of knowledge and problem-solving entities, and balanced representation and reasoning between human and machine [10] have been key considerations in helping to define our scope and direction in this research.

6.0 CONCLUSION AND NEXT STEPS

In this paper we have outlined the potential value of a conversational system to explore explanations in a human/machine context where various services are being used across datasources to contribute to situational awareness of the operating environment. Through the use of conceptual models of the domain of explanations, services, datasets, models and explanation types, the system decides how to handle explanation requests in order to provide meaningful information. This paper defines a simple scenario, three basic services and example conversations to identify how this capability could be used.

The work reported here is in the early stages of investigation. Through enrichment of these models into related areas and the incorporation of additional semantic information, the fidelity and usability of the conversational system can be improved. The ability to explore this conceptual model and the relevant assets within the conversation are planned, along with some human trials to determine the effectiveness of the explanations. An eventual goal is to use the explanations to modify the behaviour of the system or better train or configure the models through user feedback.

ACKNOWLEDGEMENT

This research was sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence under Agreement Number W911NF-16-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

REFERENCES

[1] Nottle, A., Quintana-Amate, S., Harborne, D., Alzantot, M., Braines, D., Tomsett, R., ... & Preece, A. (2017). Distributed opportunistic sensing and fusion for traffic congestion detection. In *First International Workshop on Distributed Analytics InfraStructure and Algorithms for Multi-Organization Federations* (pp. 1-6).

[2] Harborne, D., Willis, C., Tomsett, R., & Preece, A. D. (2018). Integrating learning and reasoning services for explainable information fusion. Presented at: *ICPRAI 2018 - International Conference on Pattern Recognition and Artificial Intelligence*, Montreal, Canada, 14-17 May 2018.

[3] Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2018). The building blocks of interpretability. Distill, 3(3), e10.

[4] Lipton, Z. C. (2016). The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*.

[5] Braines, D., O'Leary, N., Thomas, A., Harborne, D., Preece, A. D. and Webberley, W. M. 2017. Conversational homes: a uniform natural language approach for collaboration among humans and devices. *International Journal on Advances in Intelligent Systems* 10 (3/4) , pp. 223-237.

[6] Tomsett, R., Braines, D., Harborne, D., Preece, A., & Chakraborty, S. (2018). Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems. In *ICML Workshop on Human Interpretability in Machine Learning (WHI 2018), Jul 2018.*

[7] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD'16, 2016.

[8] Miller, T. (2017). Explanation in artificial intelligence: insights from the social sciences. *arXiv preprint arXiv:1706.07269*.

[9] Crouser, R. J., & Chang, R. (2012). An affordance-based framework for human computation and human-computer collaboration. *IEEE Transactions on Visualization and Computer Graphics*, *18*(12), 2859-2868.

[10] L. Terveen, "Overview of human-computer collaboration," Knowledge Based Systems, vol. 8(2), pp. 67–81, 1995.