ARCLIGHT: Automated Clustering and Curriculum Learning Guided by Human Training

Jamie McCusker[1,*], Henrique Santos[1], Rishi Singh[2], Sabbir M. Rashid[1], Abraham Sanders[1], Grace Roessling[1], Hongji Guo[1], Bashirul Biswas[1], Deborah L. McGuinness[1], Tomek Strzalkowski[1], Qiang Ji[1] and Jay Miller[2]

[1]*Rensselaer Polytechnic Institute, Troy, NY, USA*
[2]*Boston Fusion Corp, Lexington, MA, USA*

Abstract

ARCLIGHT is an AI fusion system that leverages Large Language Models, perception learning, knowledge graphs, and human guidance to describe high-level concept instances with lower-level attributes and affordances. By combining structured models and unsupervised exploration, ARCLIGHT discovers attributes and affordances in both known and unknown objects, entities, or activities. This enables automated novelty detection, curation of a symbolic knowledge graph, and a dialogue agent that asks discriminating questions. The system's perception component utilizes Bayesian models to recognize unknown and novel concepts, flag regions of high epistemic uncertainty, and update the knowledge graph based on user interactions. ARCLIGHT can potentially improve human-machine collaboration and advance artificial intelligence in various fields.

Keywords

Large Language Models, AI Fusion, Knowledge Graphs, Curriculum Learning

1. Introduction

The Automated Clustering and Curriculum Learning Guided by Human Training (ARCLIGHT) system is able to learn high-level concept instances (objects, entities, activities) with lower-level concept instances of attributes (features) and affordances (capabilities). ARCLIGHT discovers attributes and affordances in both known and unknown objects, entities, and activities (see Figure 1 for the knowledge representation) through a combination of structured models and unsupervised exploration. The structured models allow for automated novelty detection (e.g., this entity appears to be a dinosaur, but it is purple) while the unsupervised attribute exploration (through masking for unknown objects and more nuanced saliency maps and uncertainty attribution for known objects) allows discovery of defining features (e.g., cars have wheels). Both approaches allow curation of a symbolic knowledge graph (adding attributes to objects) as well as enabling a Large Language Model-based dialogue agent to ask discriminating questions. Our approach is to build on and extend the Whyis [1] knowledge graph framework to include multi-modal neural perception and dialogue systems. By utilizing Bayesian models in the perception system, unknown concepts can be recognized and regions of high epistemic uncertainty can be flagged. Novelty or uncertainty can be flagged at the attribute, concept, relationship, or scene level. The dialogue system can utilize regions of high certainty or uncertainty to highlight features for a user, can query

ISWC'24: The 23rd International Semantic Web Conference, November 11–15, 2024, Hanover, MD
*Corresponding author.
mccusj2@rpi.edu (J. McCusker); oliveh@rpi.edu (H. Santos); rishi.singh@bostonfusion.com (R. Singh); rashis3@rpi.edu (S. M. Rashid); sandea5@rpi.edu (A. Sanders); roessg@rpi.edu (G. Roessling); guoh11@rpi.edu (H. Guo); biswab@rpi.edu (B. Biswas); dlm@rpi.edu (D. L. McGuinness); tomek@rpi.edu (T. Strzalkowski); jiq@rpi.edu (Q. Ji); jay.miller@bostonfusion.com (J. Miller)
https://tw.rpi.edu (J. McCusker); https://tw.rpi.edu (D. L. McGuinness); https://lacailab.cogsci.rpi.edu (T. Strzalkowski); https://sites.ecse.rpi.edu/~cvrl (Q. Ji); https://bostonfusion.com (J. Miller)
0000-0003-1085-6059 (J. McCusker); 0000-0002-2110-6416 (H. Santos); 0000-0002-3577-6066 (R. Singh); 0000-0002-4162-8334 (S. M. Rashid); 0000-0002-5231-2239 (A. Sanders); 0000-0002-4162-8334 (G. Roessling); 0009-0009-6075-5687 (H. Guo); 0009-0004-4161-7311 (B. Biswas); 0000-0001-7037-4567 (D. L. McGuinness); 0000-0003-4206-3985 (T. Strzalkowski); 0000-0002-4302-2889 (Q. Ji); 0000-0003-2016-0784 (J. Miller)

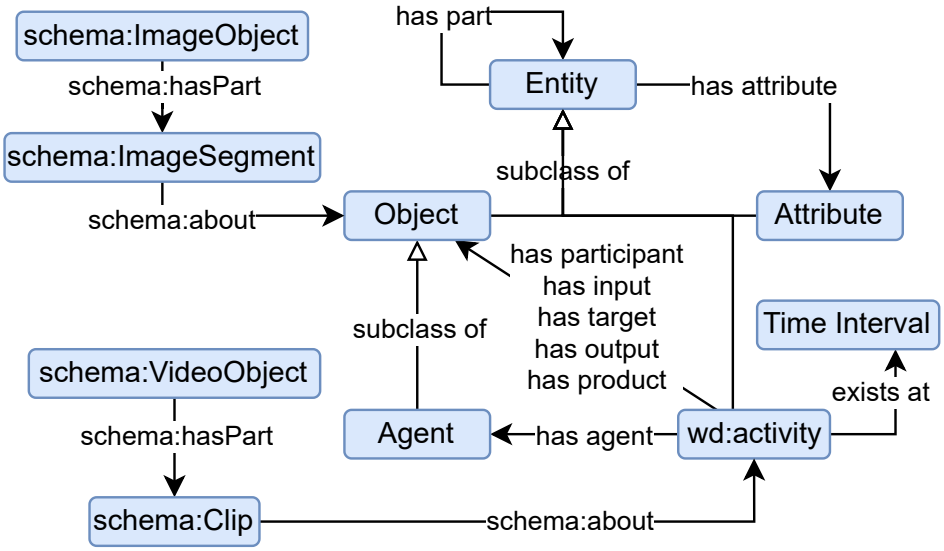

Figure 1: ARCLIGHT knowledge representation. The model is split between media (images and video) and the entities depicted in them. We use the Semanticscience Integrated Ontology (SIO) for representation between entities, and link from media to the entities using *schema:about*.

the knowledge graph to reason about an unknown object, and can update the knowledge graph directly based on user interactions.

2. System Overview

ARCLIGHT is implemented as a multiagent system using the Whyis knowledge graph framework's autonomous inference architecture [1]. This means that its components are implemented as agents – programs that receive knowledge graph (KG) fragments as inputs, process them, and produce KG outputs that might in turn become inputs to other agents. Such an approach enables the different components to asynchronously interact with one another and scale as needed by allocating agents to free computing resources as they become available. The ARCLIGHT multiagent environment is implemented as a collection of agents that expand and create a common knowledge graph. Agents' actions in this environment create changes in the knowledge graph that other agents might react to. This produces a chain of input-output relationships that could, for example, represent loosely coupled media processing pipelines. In the technical descriptions that follow, it is useful to think of the resulting coordination dynamics between agents as a publish-subscribe system where interaction among agents is mediated by topics agents publish and subscribe to, even though Whyis does not explicitly implement that type of coordination pattern. Information and knowledge are stored within a series of databases. Within Whyis, the Fuseki RDF [1] knowledge graph database (triplestore) stores symbolic knowledge and the Milvus embedding vector database [2] stores latent representations of concepts within the system. A multimedia repository is mapped to specific IRIs within the database so that media metadata is part of the KG. A runtime-configured collection of agents then watch the KG for changes to analyse and produce new subsequent KG fragments.

In Figure 2, we show a high-level diagram of the different services and types of agents in the system interacting via what is in effect a publish-subscribe model. Since each agent is looking for specific graph patterns using SPARQL queries, the agents are modular. As a result, multiple agents can be prototyped, evaluated, and deployed in parallel as desired without disrupting the overall system.

[1] https://jena.apache.org/documentation/fuseki2/

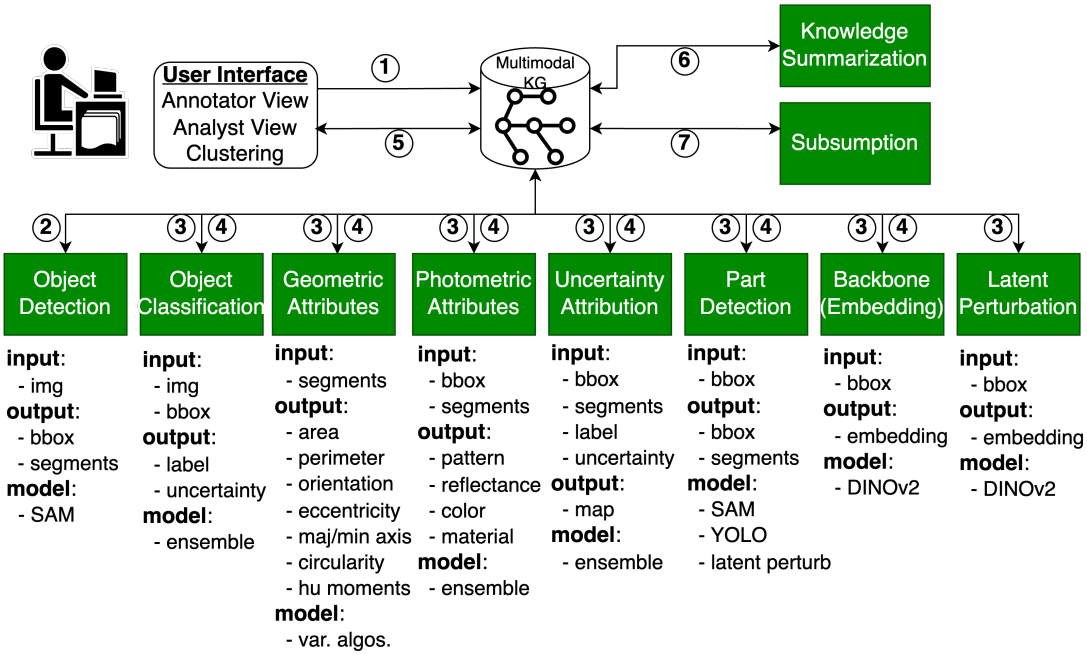

Figure 2: 1) an image is uploaded to the KG. (2) An agent detects object(s) including bounding boxes and outlines, along with part/whole composition. (3) Objects and (4) parts are analyzed for their attributes and types. (5) Unclassified instances are clustered and compared to closest labeled instances (user is prompted for likely new classes). (6) Basic graph summarization [3] to represent (type, predicate, type) triple sets. (7) Probabilistic subclass instance reasoning/

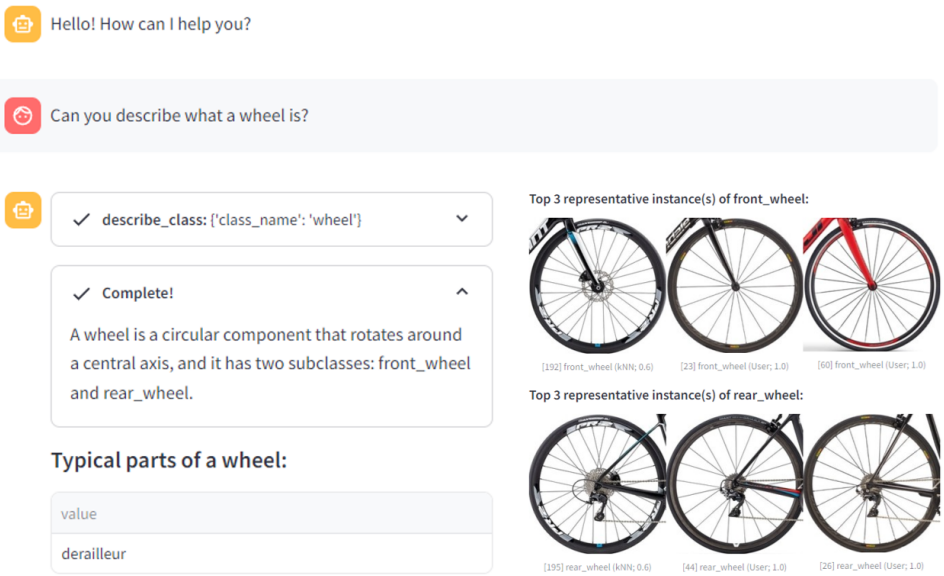

Figure 3: The system learned what wheels look like using very few examples, and can retrieve examples of them. The UI can answer a number of interactive questions using KG with embedding retrieval to drive the dialogue.

2.1. Media Analysis Agents

ARCLIGHT uses agents for each perceptual task, which includes classification and simultaneous uncertainty quantification for objects, activities, attributes, and affordances in incoming multi-modal data. Additionally, insight agents perform post-hoc analyses on the classification agents and their output to provide saliency maps, uncertainty attribution maps, and decompose known and unknown objects, which enables efficient resolution of novelty from oracles and stored knowledge.

The perception agents adopt an open-world detection framework, which separates the task of entity localization (left side of Figure 2) from classification (right side of Figure 2), enabling the detection of unknown objects. For localization in the full system, the perception agents will use a common region proposal network (RPN) combining a baseline RPN similar to He et al [4] with the Segment Anything Model (SAM). SAM is trained on billions of objects across a wide domain and is thus a powerful, largely class-agnostic tool which can greatly improve the RPN's representation of "object-ness". Currently, individual agents use a separate RPN as part of their architecture to allow better evaluation of SOTA open-world detection methods. Insight agents perform analyses on perception agents themselves to probe their models, localize sources of uncertainty, and detect emergent features.

The design of the Object Classification Agent focuses on evaluating the use of uncertainty quantification and attribution for novel object detection in a SOTA deformable detection transformer model architecture, similar to work from Zohar et al [5] for the open world object detection (OWOD) setting. This architecture uses a common encoder-decoder feeding to both a feature extraction network and a RPN, which allows it to detect the presence of unknown objects.

The Attribute Classification Agent employs feature extractors based on the model and dataset presented by Ramanathan et al [6]. This includes separate attribute heads for predetermined attribute types (color, pattern-marking, material and reflectance) downstream of a Mask R-CNN architecture that are initially trained on the PACO dataset.

Interaction between ARCLIGHT and human users (e.g., analysts, teachers, or evaluators) happens through a multimodal, instruction-tuned, tool-aware dialogue agent. The dialogue agent is responsible for initiating conversation with users to facilitate active learning when uncertainty arises from the system's perception. It is also responsible for responding to user-initiated conversation to aid in typical analytic tasks including question answering and reasoning over the media presented by the user. All discussion is mediated using the Activity Streams [7] vocabulary, so that UIs and dialog agents alike read and create "Fediverse"-compliant messages and media posts.

At the core of the dialogue system is the instruction-tuned Large Language Model (LLM) Llama 3 70b [8]. At inference time, the LLM is prompted with instructions pertaining to: (1) its purpose, (2) operational constraints such as which images the user is viewing, (3) a listing of all tools (e.g., APIs, databases) at its disposal, (4) contextual knowledge relevant to the conversation, and (5) the current dialogue history. The LLM then proceeds following a ReAct-like reason-action loop [9] to retrieve any further necessary information (e.g., via a SPARQL query) and respond to the user. Figure 3 illustrates this process in our user interface, where it is able to retrieve relevant instances from the graph using text dialogue.

When an agent initiates a conversation with a user following the ingestion of media with high uncertainty, the ARCLIGHT dialogue agent selects actions that update Whyis with new knowledge obtained from the interaction. The LLM is responsible for locating the appropriate labels, descriptions, and other relevant information and crafting the call to a Whyis API to update the ontology accordingly.

3. Conclusion

The ARCLIGHT system is a knowledge graph-centric AI fusion system that allows users to easily upload media that can be analyzed by a suite of perception, classification, and dialogue agents, creating knowledge graph fragments to describe each depicted scene. This graph-augmented knowledge to is used to drive discussion through an instruction-tuned LLM that knows how to query the graph and extract relevant knowledge to provide suitable responses to user dialogue. Further, the media, dialogue, and image knowledge are all represented within the knowledge graph, allowing for comprehensive explanations of any analysis and tracing of any given source of information. Our hope is that this kind of modular system can serve as a model for neuro-symbolic learning using perception, language, and knowledge in a meaningful way.

References

[1] J. McCusker, D. L. McGuinness, Whyis 2: An Open Source Framework for Knowledge Graph Development and Research, in: The Semantic Web, Lecture Notes in Computer Science, Springer Nature Switzerland, Cham, 2023, pp. 538–554. doi:10.1007/978-3-031-33455-9_32.

[2] J. Wang, X. Yi, R. Guo, H. Jin, P. Xu, S. Li, X. Wang, X. Guo, C. Li, X. Xu, et al., Milvus: A purpose-built vector data management system, in: Proceedings of the 2021 International Conference on Management of Data, 2021, pp. 2614–2627.

[3] J. McCusker, Customizable knowledge graph visualization using the whyis knowledge explorer, in: Visualization and Interaction for Ontologies, Linked Data and Knowledge Graphs 2024, CEUR Workshop Proceedings, 2024.

[4] J. He, S. Yang, S. Yang, A. Kortylewski, X. Yuan, J.-N. Chen, S. Liu, C. Yang, Q. Yu, A. Yuille, Partimagenet: A large, high-quality dataset of parts, in: European Conference on Computer Vision, Springer, 2022, pp. 128–145.

[5] O. Zohar, K.-C. Wang, S. Yeung, Prob: Probabilistic objectness for open world object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 11444–11453.

[6] V. Ramanathan, A. Kalia, V. Petrovic, Y. Wen, B. Zheng, B. Guo, R. Wang, A. Marquez, R. Kovvuri, A. Kadian, et al., Paco: Parts and attributes of common objects, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 7141–7151.

[7] J. Snell, E. Prodromou, Activity Vocabulary, W3C Recommendation, W3C, 2017. Https://www.w3.org/TR/2017/REC-activitystreams-vocabulary-20170523/.

[8] A. Dubey, A. Jauhri, A. P. et al., The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[9] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. R. Narasimhan, Y. Cao, React: Synergizing reasoning and acting in language models, in: The Eleventh International Conference on Learning Representations, 2023. URL: https://openreview.net/forum?id=WE_vluYUL-X.