Neuro-symbolic Visual Reasoning for Multimedia Event Processing: Overview, Prospects and Challenges

Muhammad Jaleed Khan, Edward Curry

SFI Centre for Research Training in Artificial Intelligence, Data Science Institute, National University of Ireland
Galway, Galway, Ireland.

Abstract
Efficient multimedia event processing is a key enabler for real-time and complex decision making in streaming media. The need for expressive queries to detect high-level human-understandable spatial and temporal events in multimedia streams is inevitable due to the explosive growth of multimedia data in smart cities and internet. The recent work in stream reasoning, event processing and visual reasoning inspires the integration of visual and commonsense reasoning in multimedia event processing, which would improve and enhance multimedia event processing in terms of expressivity of event rules and queries. This can be achieved through careful integration of knowledge about entities, relations and rules from rich knowledge bases via reasoning over multimedia streams within an event processing engine. The prospects of neuro-symbolic visual reasoning within multimedia event processing are promising, however, there are several associated challenges that are highlighted in this paper.

Keywords
multimedia event processing, visual reasoning, commonsense reasoning, video stream processing, spatiotemporal events

1. Introduction

Internet of multimedia things (IoMT), data analytics and artificial intelligence are continuously improving smart cities and urban environments with their ever-increasing applications ranging from traffic management to public safety. As middleware between internet of things and real-time applications, complex event processing (CEP) systems process structured data streams from multiple producers and detect complex events queried by subscribers in real-time. The enormous increase in image and video content surveillance cameras and other sources in IoMT applications posed several challenges in real-time processing of multimedia events, which motivated researchers in this area to extend the existing CEP engines and to devise new CEP frameworks to support unstructured multimedia streams. Over the past few years, several efforts have been made to mitigate the challenges in multimedia event processing by developing techniques for extension of existing CEP engines for multimedia events [1] and development of end-to-end CEP frameworks for multimedia streams [2]. On the other hand, the research in computer vision has focused on complimenting object detection with human-like visual reasoning that allows for prediction of mean-
ingful and useful semantic relations among detected objects based on analogy and commonsense (CS) knowledge [3, 4].

2. Background

In this paper, we discuss the background, prospects and challenges related to leveraging the existing visual and commonsense reasoning to enhance multimedia event processing in terms of its applicability and expressivity of multimedia event queries. The motivation for development of an end-to-end multimedia event processing system supporting automated reasoning over multimedia streams comes from its potential real-time applications in smart cities, internet and sports. Fig. 1 shows an example of traffic congestion event detected using visual and commonsense reasoning over the objects and relations among the objects in the video stream. A conceptual level design and a motivational example of a novel CEP framework supporting visual and commonsense reasoning is presented in Fig. 2.

This section presents a review of the recent work in stream reasoning, multimedia event processing and visual reasoning that could be complementary within a proposed neuro-symbolic multimedia event processing system with support for visual reasoning.

2.1. Reasoning over Streams and Knowledge Graph

Emerging from the semantic web, streaming data is conventionally modelled according to RDF [8], a graph representation. The real-time processing of RDF streams is performed in time-dependent windows that control the access to the stream, each containing a small part of the stream over which a task needs to be performed at a certain time instant. Reasoning is performed by applying RDF Schema rules to the graph using SPARQL query language or its variants. Reasoning over knowledge graphs (KG) provides new relations among entities to enrich the knowledge graph and improve its applicability [9]. Neuro-symbolic computing combines symbolic and statistical approaches, i.e. knowledge is represented in symbolic form, whereas learning and reasoning are performed by DNN [10], which has shown its efficacy in object detection [11] as well as enhanced feature learning via knowledge infusion in DNN layers from knowledge bases [12]. Temporal KG allows time-aware representation and tracking of entities and relations [13].

2.2. Multimedia Event Representation and Processing

CEP engines inherently lacked the support for unstructured multimedia events, which was mitigated by a generalized approach for handling multimedia events as native events in CEP engines as presented in [1]. Angsufactmetee et al. [14] has presented an ontological approach for modeling complex events and multimedia data with syntactic and semantic interoperability in multimedia sensor networks, which allows subscribers to define application-specific complex events while keeping the low-level network repre-
Figure 2: (a) Conceptual level block diagram of a CEP framework supporting visual reasoning. The input stream of images (or video frames) is received from a publisher; the objects are detected using DNN and rule-based relations [5] are represented using a graph, which is followed by automated reasoning that adds new visual relations from a knowledge base [6] and validates those relations using commonsense knowledge [7]. The matcher performs spatial and temporal event matching on these detected objects and relations with the spatial and temporal patterns in high-level events queried by the subscriber. (b) An example of visual reasoning in multimedia event processing. Suppose a subscriber is interested in the event where tennis player is either “hitting” or “missing” a shot. This event is not explicitly defined via rules but it can be predicted via automated reasoning over detected objects and predicted relations. (Image credits: Visual Genome [6])

2.3. Visual and Commonsense Reasoning

In addition to the objects and their attributes in images, detection of relations among these objects is crucial for scene understanding for which compositional models [17], visual phrase models [18] and DNN based relational networks [19] are available. Visual and semantic embeddings aid large scale visual relation detection, such as Zhang et al. [4] employed both visual and textual features to leverage the interactions between objects for relation detection. Similarly, Peyre et al. [3] added a visual phrase embedding space during learning to enable analogical reasoning for unseen relations and to improve robustness to appearance variations of visual re-
lations. Table 1 presents some knowledge bases publicly available for visual reasoning. Wan et al. [7] proposed the use of commonsense knowledge graph along with the visual features to enhance visual relation detection. Rajani et al. [20] leverage human reasoning and language models to generate human-like explanations for DNN-based commonsense question answering. There are various commonsense reasoning methods and datasets available for visual commonsense reasoning [21] and story completion [22].

3. Neuro-symbolic Visual Reasoning in Multimedia Event Processing

3.1. Prospects

The current multimedia event representation methods use knowledge graph to represent the detected objects, their attributes and relations among the objects in video streams. Pre-defined spatial-temporal rules are used to form relations among the objects. However, the complex relations that exist among real-world objects also depend on semantic facts and situational variables that can not be explicitly specified for every possible event as rules. The statistical reasoning methods and knowledge bases discussed in Section 2 have great potential to complement the rule-based relation formation in multimedia event processing by injecting some semantic knowledge and reasoning to extract more semantically meaningful relations among objects. This advancement will allow subscribers to define abstract or high-level human-understandable event query rules that can be decomposed into spatial and temporal patterns. The spatio-temporal matching of the queried high-level events will be performed on the objects, rule-based relations and relations extracted using visual reasoning. The subscriber will be instantly notified of the high-level event as a combined detection of those spatial-temporal patterns. The idea of developing an end-to-end multimedia event processing system supporting visual reasoning over video streams (Fig. 2) poses several challenges that are discussed in the next section. This novel approach will give more expressive power to subscribers in querying complex events in multimedia streams, and thus increase the scope of real-time applications of multimedia event processing in smart city applications as well as internet media streaming applications.

3.2. Challenges

1. Suitable representation for reasoning

It is crucial to select a generalized and scalable model to represent events and effectively perform automated reasoning to derive more meaningful and expressive spatio-temporal events.

2. Expressive query definition and matching

Providing a generic and human-friendly format to subscribers for writing expressive and high-level queries would require new constructs. Matching queries with the low-level events and relations along with reasoning via knowledge bases requires efficient retrieval within the complex event matcher. Real-world complex events can share similar patterns, occur as a cluster of similar events or occur in a hierarchical manner, which requires generalized, adaptive and scalable spatiotemporal constructs to query such events.

3. Labeling and training samples of visual relations

There can be a large numbers of objects and possible relations among them in images, which can result in a large number of categories of relations. It is difficult
Table 1
Available Knowledge Bases for Visual Reasoning

<table>
<thead>
<tr>
<th>Knowledge Base</th>
<th>#Images</th>
<th>#Entity Categories</th>
<th>#Entity Instances</th>
<th>#Relation Categories</th>
<th>#Relation Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open Images V4 [23]</td>
<td>9,200,000</td>
<td>-</td>
<td>600</td>
<td>72,400,000</td>
<td>-</td>
</tr>
<tr>
<td>YAGO 4 [24]</td>
<td>-</td>
<td>10,124</td>
<td>64,000,000</td>
<td>-</td>
<td>2 billion</td>
</tr>
<tr>
<td>COCO-a [25]</td>
<td>10,000</td>
<td>81</td>
<td>74,000</td>
<td>156</td>
<td>207,000</td>
</tr>
<tr>
<td>VaKE [18]</td>
<td>-</td>
<td>1,884</td>
<td>-</td>
<td>-</td>
<td>12,593</td>
</tr>
</tbody>
</table>

Acknowledgement
This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223.

References


