Image Sequence Understanding through Narrative Sensemaking

Zev Battad, Mei Si
Rensselaer Polytechnic Institute
Department of Cognitive Science

(battaz@rpi.edu, sim@rpi.edu)

Rensselaer

Introduction

Conversational AI systems would benefit from a method of understanding and discussing image sequences as humans do.

- Humans make sense of what is happening in images beyond the directly observable.
- It is natural for humans to organize their understanding using narrative.[1]

Figure 1 – A sequence of images with human-written explanations.[2]

We aim to create a system that can generate machine-usable knowledge graphs from image sequences using the human-inspired process of **sensemaking**.

Sensemaking

Sensemaking is the process of creating consistency and coherence between observations in an environment and a person's existing knowledge of the world.[3]

- Connections are an important part of tying together what one observes.[4]
- Aim to interconnect observations as much as possible using existing knowledge while also remaining consistent (not self-contradictory).
- The types of connections people use can be found in human-made narratives, and categorized as: *spatial, temporal, causal, referential, affective (emotion/motivation)*.[5]

Image Sequence Understanding System

The image sequence understanding system takes **observations** about a sequence of images, performs a **sensemaking** process to hypothesize additional relationships between observations, then produces a **knowledge graph** combining its observations with its additional relationships.

Observations

- Visually observable facts about each image.
- Consist of scene graphs, with objects and their relationships as nodes and edges (Fig. 2).

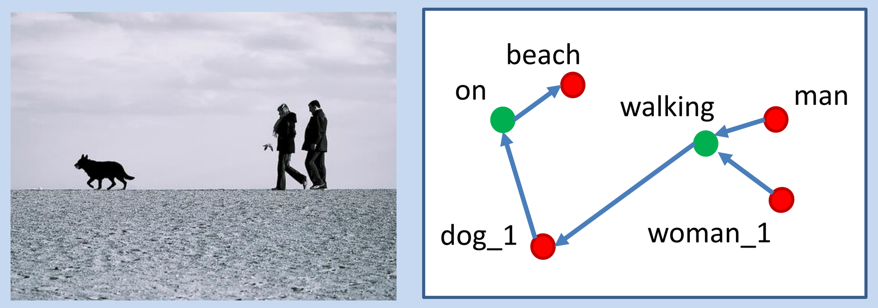

Figure 2 – Image with excerpt of its scene graph from the Visual Genome dataset.

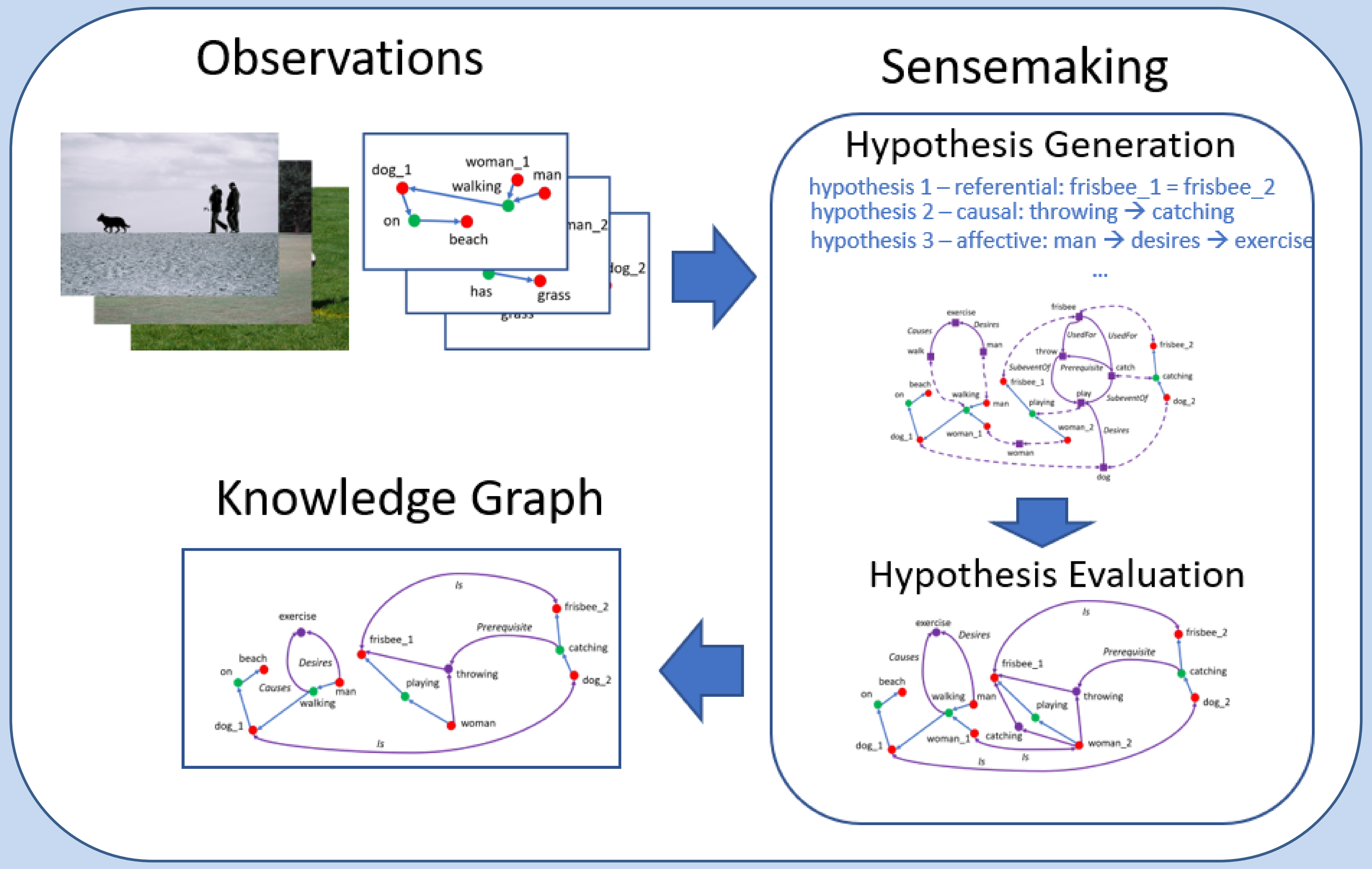

Figure 3 – Overall architecture of system.

Sensemaking Subsystem

- Subsystem to hypothesize additional relationships between observations in a two-step process of *hypothesis generation* and *hypothesis evaluation*.

Hypothesis Generation

- Over-generate possible additional relationships.
- ConceptNet common-sense knowledge network as existing knowledge, with generic concepts as nodes and relationships as edges.
- ConceptNet relations are selected and organized by system based on narrative connection types.
- Scene graph nodes are equated to their ConceptNet concept nodes (Fig. 4, a).
- Paths between concepts (Fig. 4, b) are taken as hypothesized additional relationships (Fig. 4, c).

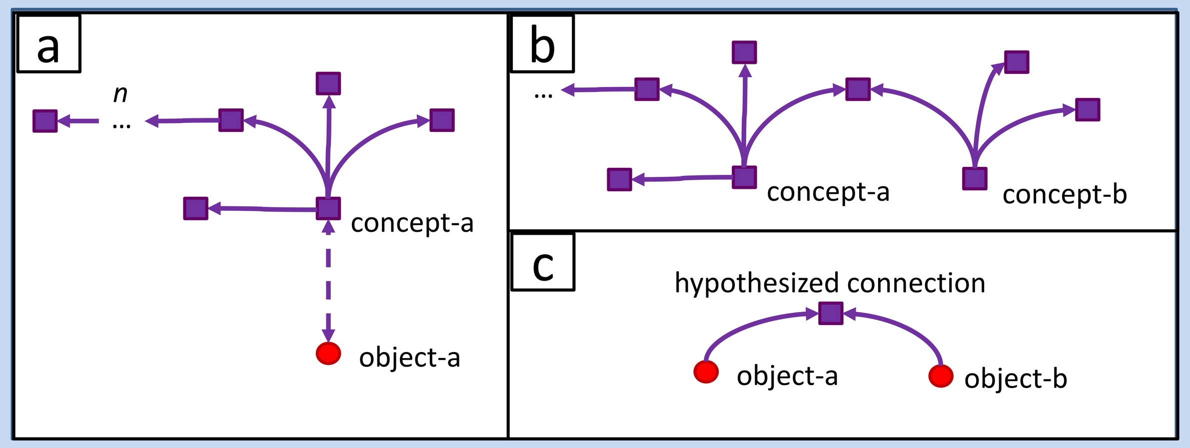

Figure 4 – Process of generating single hypothesis.

Hypothesis Evaluation

- Aim to connect as much of the knowledge graph as possible while maintaining consistency.
- Choose which hypothesized additional relationships to keep as a Multi-Objective Optimization Problem:

$$max_1^m(f_i(x))|h_j \in H_f$$

Find set of hypothesis, h_m, that maximizes score of each objective function, $f_i(x)$, where each hypothesis set is part of the set of feasible hypothesis sets, H_f.

Three objective functions are used:

- *Connectivity* and *Density*, measures of graph inter-connectedness.
- *Support*, scorable evidence from scene graph confidence values and ConceptNet edge weights.

A hypothesis set is *feasible* if its elements do not contradict each other, decided by heuristic-based checks per narrative relationship type.

Data Sources

System utilizes two external data sources:

Visual Genome Dataset acts as a source of human-annotated scene graphs used for the system's observations.

- Images with ROI bounding boxes paired with object and relationship annotations.
- Automated scene graph generation methods do also exist (e.g. Graph-RCNN).

ConceptNet acts as system's commonsense knowledge source.

- Crowd-sourced knowledge network of generic concepts and their relationships.
- Uses a known finite set of relationships mappable to narrative relationships.

Example

Figure 5 shows an image sequence (a) with excerpts of its scene graphs (b) and final knowledge graph (c).

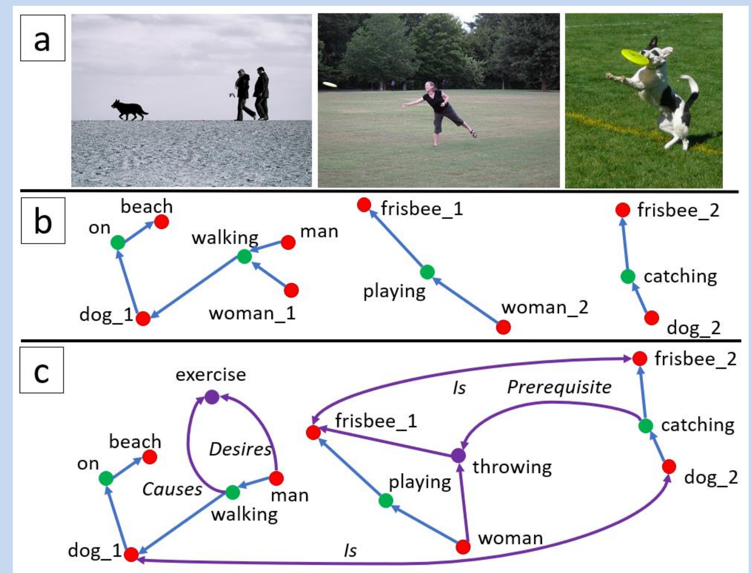

Figure 5 – Example image sequence parse

Future Work

- Implement architecture into full system.
- Investigate whether system's hypothesized information is of value to human readers.

References

1. Jerome Bruner. 2001. Self-making and world-making. Narrative and identity: Studies in autobiography, self, and culture (2001), 25–37.
2. Huang, T.H., Ferraro, F., Mostafazadeh, N., Misra, I., Agrawal, A., Devlin, J., Girshick, R., He, X., Kohli, P., Batra, D. and Zitnick, C.L., 2016, June. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1233-1239).
3. Karl E Weick, Kathleen M Sutcliffe, and David Obstfeld. 2005. Organizing and the process of sensemaking. Organization science 16, 4 (2005), 409–421.
4. Gary Klein, Brian Moon, and Robert R Hoffman. 2006. Making sense of sensemaking 1: Alternative perspectives. IEEE intelligent systems 21, 4 (2006), 70–73.
5. Elaine Reese, Catherine A Haden, Lynne Baker-Ward, Patricia Bauer, Robyn Fivush, and Peter A Ornstein. 2011. Coherence of personal narratives across the lifespan: A multidimensional model and coding method. Journal of Cognition and Development 12, 4 (2011), 424–462.