# Image Sequence Understanding through Narrative Sensemaking

Extended Abstract

Zev Battad
Rensselaer Polytechnic Institute
Troy, New York
battaz@rpi.edu

Mei Si
Rensselaer Polytechnic Institute
Troy, New York
sim@rpi.edu

## ABSTRACT

For intelligent conversational agents to speak about albums of images with users as humans do, they must be able to make sense of images as humans do. Computer vision methods can report directly observable information, but human beings care about more than the directly observable; they value holistic narratives that include affective and motivational evaluations, casual connections, and other inferred relationships from external knowledge. Drawing from theories in cognitive sensemaking and narrative coherence, we propose an approach for image sequence understanding that strives to generate and evaluate hypotheses about the relationships between people, events, and objects in images using commonsense knowledge, which are formed into a consistent network of hypotheses and observed facts via multi-objective optimization. The result is an enriched knowledge representation in the form of a knowledge graph which may later be used by a conversational agent.

## KEYWORDS

Sensemaking, Narrative, Image Understanding, Commonsense Knowledge

**ACM Reference Format:**
Zev Battad and Mei Si. 2021. Image Sequence Understanding through Narrative Sensemaking: Extended Abstract. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), Online, May 3–7, 2021*, IFAAMAS, 3 pages.

## 1 INTRODUCTION

Popular image sharing services, such as Facebook, Google+, and Imgur, contain features that automatically organize and display image sets to users communicating something of interest. For example, Facebook assembles "friendship anniversary" albums showing a history of photos where mutual friends appear together. In order for a conversational agent to talk about albums of images with users as humans do, they must be able to make sense of and understand the images. Computer vision can assist with this, identifying directly observable aspects of an image. However, when humans look at visual scenes, they do not only see a cascade of direct observations. Rather, from a cascade of observations, people glean something more meaningful - a holistic account of what they have observed with affective and motivational evaluations, casual connections, inferred events, external actors, and concepts from prior knowledge. These are the underpinnings of narrative, one of the oldest methods by which humans have traditionally exchanged information and an inherent component of the way humans organize knowledge [1, 3, 10].

Sensemaking is the process of creating consistency and coherence between observations in the environment and a person's existing knowledge of the world, expanding one's understanding of both. People seek explanations even if they do not have the most accurate understanding of the information they encounter, in which case the best strategy is to integrate the information they encounter with prior knowledge and prior inferences as consistently as they can [15]. An important part of sensemaking is the formation of connections, such as those between people, places, and events [6, 7], as well as the creation of a representation for the information so that one may form and evaluate hypotheses to expand one's understanding [8, 13].

We propose a system to support conversational agents that discuss images with people by using a computational narrative sensemaking process. The system strives to generate and evaluate hypotheses about the relationships between people, events, and objects in images using commonsense knowledge, which is then formed into a consistent network of hypotheses and observed facts via multi-objective optimization. The result is an enriched knowledge representation in the form of a knowledge graph which can be used by a conversational agent.

## 2 NARRATIVE SENSEMAKING SYSTEM

The system proposed here generates knowledge graph representations from image sequences by adding narrative coherence connections to observational information through a computational sensemaking process. The input is a sequence of images. The output is a knowledge graph. As per the definition of narrative, i.e. a sequence of events that are causally or sequentially related [10], the generated knowledge graph strives to not only describe what can be observed in the images, but also provide information for a reasonable narrative of why those events happened.

Observations of image sequences take the form of scene graphs - semantic representation of the objects in an image and the relationships between them. The Visual Genome Dataset acts as this system's corpus of human-written scene graphs [9], though methods exist for automatic scene graph generation [16]. The Sensemaking subsystem speculates additional information about the images in the sequence by generating and evaluating hypotheses about additional connections within their scene graphs. Sensemaking is done in two steps: Hypothesis Generation and Hypothesis Evaluation.

*Hypothesis Generation.* Hypothesis generation represents wide speculation about possible relationships between nodes in the scene graph drawn from commonsense knowledge. ConceptNet is used

*Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, Online.*

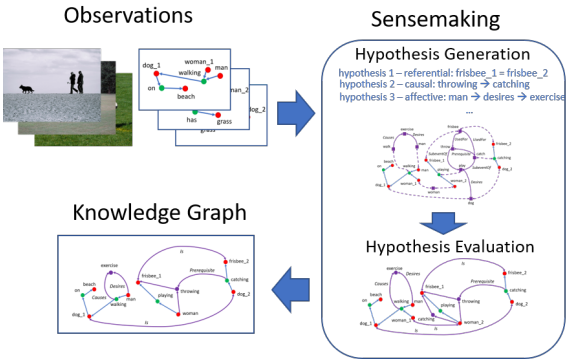


**Figure 1: Architecture for proposed system.**

as a commonsense knowledge network, with generic concepts as nodes and relationships between concepts as edges. The types of relationships hypothesized are those used in narrative coherence, which, from studies on narrative coherence found in human-made narratives, can be categorized as *spatial*, *temporal*, *causal*, *referential*, and *affective (emotional/motivational)* [4, 5, 11, 12]. ConceptNet uses a discrete set of relationship types, which are mapped to the elements of narrative coherence categorized above.

To generate hypotheses, scene graph nodes are equated with their corresponding ConceptNet concept. Paths between concepts that consist of ConceptNet Relations are taken as hypothesized relationships between the scene graph nodes corresponding to those concepts.

*Hypothesis Evaluation.* The hypothesis generation process speculates an over-generated set of possible connections to add to the system's knowledge representation. Because these hypotheses are not necessarily consistent with each other, only a limited number will be kept.

We cast the problem of choosing which hypotheses to keep as a multi-objective optimization problem (MOP). We define the problem as finding the hypothesis set or sets, $h_m$, that results in the highest score for a set of objective functions, $f_i(x)$, where each hypothesis set is part of the set of feasible hypothesis sets $H_f$:

$$max_1^m(f_i(x))|h_j \in H_f \quad (1)$$

A feasible hypothesis set is one which satisfies each constraint in the system, described later.

Three objective functions are used - connectivity ($f_1$), density ($f_2$), and support ($f_3$). Connectivity is the minimum number of nodes that must be removed from a graph for it to be separated into independent subgraphs. High connectivity promotes relationships across sections of the knowledge graph which are sparsely interconnected, such as those formed from individual scene graphs for each image, and between stranded or sparsely connected individual nodes and the rest of the graph.

Density is the proportion of edges in the graph versus the maximum number of edges possible. High density promotes drawing relationships between as many scene graph nodes as possible.

Support is the degree to which the hypotheses accepted by the system are supported by their sources. Two sources of support

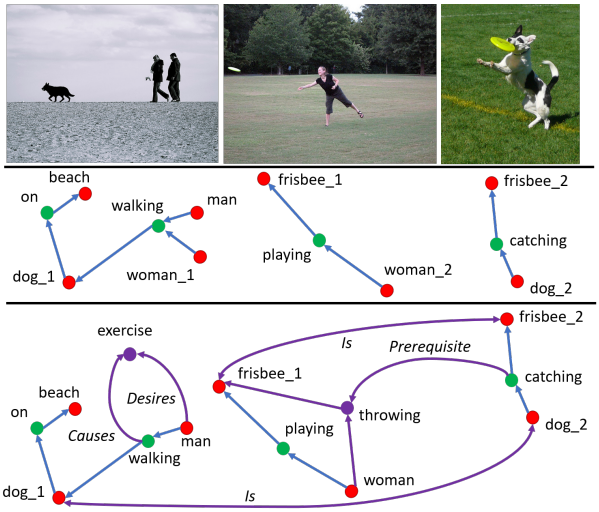


**Figure 2: A sequence of images from VisualGenome (top), excerpts of their scene graphs (middle), and a knowledge graph with accepted hypotheses (bottom).**

are ConceptNet edge weights, which reflect which relationships are more certain, and duplicate label count in human-provided VisualGenome annotations. For the system to accept hypotheses it is more sure about, weak support should be penalized, while strong support should be promoted. Total support is calculated as the sum of the score of each piece of support for each hypothesis.

The constraints that will be used to check feasibility will be based on a set of heuristics-based checks. Three checks for contradiction and consistency are planned for initial implementation: Identity (whether objects are consistently assigned *is* relationships across scenes), Causal Ordering (as decided by the causally ordered relationships in ConceptNet), and Emotional Valence (e.g. a person cannot be both angry and happy at the same thing).

## 3 EXAMPLE

The example in Figure 2 is formed from an excerpt of the scene graphs for a sequence of three images from Visual Genome. The full scene graphs for all three images together contains 343 nodes, of which there are 156 object or location nodes and 187 relationship nodes. Paths of at most length $n = 2$ are used to allow for at most one intervening ConceptNet node in a path, preventing tangential connections that result from longer paths. For the scene graphs shown in the examples, about 7700 hypotheses were generated.

## 4 FUTURE WORK

Future work for this project is in two directions: implementation of the proposed architecture into a full system using automated scene graph generation methods [16] and knowledge graph-based narrative generation [2, 14], and evaluating whether the information introduced by the system is of value to people interacting with images.

## REFERENCES


[1] H Porter Abbott. 2020. *The Cambridge introduction to narrative*. Cambridge University Press.

[2] Zev Battad, Andrew White, and Mei Si. 2019. Facilitating Information Exploration of Archival Library Materials Through Multi-modal Storytelling. In *International Conference on Interactive Digital Storytelling*. Springer, 120–127.

[3] Jerome Bruner. 2001. Self-making and world-making. *Narrative and identity: Studies in autobiography, self, and culture* (2001), 25–37.

[4] Morton Ann Gernsbacher and Talmy Givón. 1995. *Coherence in spontaneous text*. Vol. 31. John Benjamins Publishing.

[5] Talmy Givón. 1992. The grammar of referential coherence as mental processing instructions. *Linguistics* 30, 1 (1992), 5–56.

[6] Gary Klein, Brian Moon, and Robert R Hoffman. 2006. Making sense of sensemaking 1: Alternative perspectives. *IEEE intelligent systems* 21, 4 (2006), 70–73.

[7] Gary Klein, Brian Moon, and Robert R Hoffman. 2006. Making sense of sensemaking 2: A macrocognitive model. *IEEE Intelligent systems* 21, 5 (2006), 88–92.

[8] John Kolko. 2010. Sensemaking and framing: A theoretical reflection on perspective in design synthesis. *Design Research Society* (2010).

[9] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123, 1 (2017), 32–73.

[10] Birgit Neumann and Ansgar Nünning. 2008. *An introduction to the study of narrative fiction*. Klett.

[11] 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.

[12] J Christopher Rideout. 2013. A twice-told tale: Plausibility and narrative coherence in judicial storytelling. *Legal Comm. & Rhetoric: JAWLD* 10 (2013), 67.

[13] Daniel M Russell, Mark J Stefik, Peter Pirolli, and Stuart K Card. 1993. The cost structure of sensemaking. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*. 269–276.

[14] Mei Si. 2015. Tell a story about anything. In *International Conference on Interactive Digital Storytelling*. Springer, 361–365.

[15] Karl E Weick, Kathleen M Sutcliffe, and David Obstfeld. 2005. Organizing and the process of sensemaking. *Organization science* 16, 4 (2005), 409–421.

[16] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. 2018. Graph r-cnn for scene graph generation. In *Proceedings of the European conference on computer vision (ECCV)*. 670–685.