QUARTERLY OF APPLIED MATHEMATICS
VOLUME LXXVII, NUMBER 2
JUNE 2019, PAGES 323–356
https://doi.org/10.1090/qam/1530
Article electronically published on January 11, 2019

GENERATING OPEN WORLD DESCRIPTIONS OF VIDEO USING COMMON SENSE KNOWLEDGE IN A PATTERN THEORY FRAMEWORK

By

SATHYANARAYANAN N. AAKUR (*Department of Computer Science and Engineering, University of South Florida, Tampa, Florida 33620*),

FILLIPE DM DE SOUZA (*Department of Computer Science and Engineering, University of South Florida, Tampa, Florida 33620*),

AND

SUDEEP SARKAR (*Department of Computer Science and Engineering, University of South Florida, Tampa, Florida 33620*)

This paper is dedicated to Professor Ulf Grenander

Abstract. The task of interpretation of activities as captured in video extends beyond just the recognition of observed actions and objects. It involves open world reasoning and constructing deep semantic connections that go beyond what is directly observed in the video and annotated in the training data. Prior knowledge plays a big role. Grenander's canonical pattern theory representation offers an elegant mechanism to capture these semantic connections between what is observed directly in the image and past knowledge in large-scale common sense knowledge bases, such as ConceptNet. We represent interpretations using a connected structure of basic detected (grounded) concepts, such as objects and actions, that are bound by semantics with other background concepts not directly observed, i.e., contextualization cues. Concepts are basic generators and the bonds are defined by the semantic relationships between concepts. Local and global regularity constraints govern these bonds and the overall connection structure. We use an inference engine based on energy minimization using an efficient Markov Chain Monte Carlo that uses the ConceptNet in its move proposals to find these structures that describe the image content. Using four different publicly available large datasets, Charades, Microsoft Visual Description Corpus (MSVD), Breakfast Actions, and CMU Kitchen, we

Received March 22, 2018, and, in revised form, October 12, 2018.
2010 *Mathematics Subject Classification.* Primary 54C40, 14E20; Secondary 46E25, 20C20.
Key words and phrases. Pattern theory, activity interpretation, video semantics, open world.
This research was supported in part by NSF grants IIS 1217676 and CNS-1513126.
Email address: saakur@mail.usf.edu
Email address: fillipe@mail.usf.edu
Email address: sarkar@usf.edu

show that the proposed model can generate video interpretations whose quality is comparable or better than those reported by state-of-the-art approaches, such as different forms of deep learning models, graphical models, and context-free grammars. Apart from the increased performance, the use of encoded common sense knowledge sources alleviate the need for large annotated training datasets and help tackle any imbalance in the data through prior knowledge, which is the bane of current machine learning approaches.

1. Introduction. There have been many successful applications of pattern theory in computer vision and artificial intelligence, for instance in shape analysis [7, 23], target tracking [40, 57], computational anatomy [25, 41], biological growth [26], context-free grammar [24], image models [43], and even modeling of human thought [29].

Pattern theory takes an analysis by the synthesis approach [43]. To recognize a pattern, we have to be able to generate it. In the canonical representation of pattern theory (see Chapter 6 of [30]), complex patterns are described as compositions of simpler patterns starting from elements of structured sets (generators) that bind to each other (via bonds) through local interactions, constrained by local regularity, and also by global structure regularity, captured by an overarching graph structure. A probability structure over the representations captures the diversity of patterns.

The many incarnations of graphical models of patterns, such as directed acyclic graphs (DAG), Markov random fields (MRF), Gaussian random fields, and formal languages, can be shown to be special cases (see Chapter 6 of [30]). The use of pattern theoretic concepts as graphical probabilistic models for computer vision can be found in [8, 9, 11, 21, 32, 66]. However, the use of the canonical representations of the pattern theory in computer vision is rare, more so for high-level vision problems of recognition and interpretation. In our prior work, we have demonstrated that the canonical form allows for flexible interpretations that can help overcome reasoning under the presence of background object clutter, multiple objects and action, missing items, and long action sequences [1, 17, 18, 20]. Here we show how prior common sense knowledge, encoded in publicly available datasets such as ConceptNet, can be incorporated in a pattern theory framework to constrain the interpretations and also go beyond just the observed actions and objects.

We assume that we have algorithms, such as recent deep learning methods, that generate candidate object and action labels in a video snippet capturing an activity. Multiple candidates per instance are assumed to exist. These candidates, or feature generators in the canonical representation, connect to grounded concepts generators in the graph. Priors from the ConceptNet constrain the connections about these generators and other context generators. The inference is posed as an energy minimization process solved by MCMC. The resulting graph representations could be then used for multiple tasks: narrative text generation (captioning), query-answering, semantics indexing, etc.

Representation of activities should have three essential characteristics. First, it has to be compositional so as to have a large descriptive capacity through the combinatorial combination of basic units. Second, it should not have a fixed structure like a Bayesian Network or HMM or MRF, rather it should be flexible to allow for different types of structures involving a different number of constituent entities, such as “whisk eggs with a fork in the bowl”, “crack eggs”. Third, the representation should be an intermediate

representation that supports the construction of different output types such as captions, sentence-based descriptions, or complex Q&A interactions that go beyond simple yes-no questions.

In a `closed world`, every possible world contains the same objects and events, which are in one-to-one correspondence with the ground evidence in the image. This is the case for much of the current work in activity recognition, especially ones based on machine learning; the learned knowledge is limited to the annotated training set. There is some generalization possible through transfer learning methods, but it is mostly to adjust to new domain statistics; the object and action set remains the same. In an `open world`, there is no a priori one-to-one correspondence between the objects and the grounding terms. Some of the detected ground evidence in the image might not even be relevant to the action under consideration and hence needs to be rejected. Also, for each identified evidence in video, one might have many possible object and action labels. We leverage existing sources of knowledge such as ConceptNet [39] and use common sense reasoning to move beyond the constraints of annotations in the training data. A priori knowledge sources also help regularize noise in putative object and action labels.

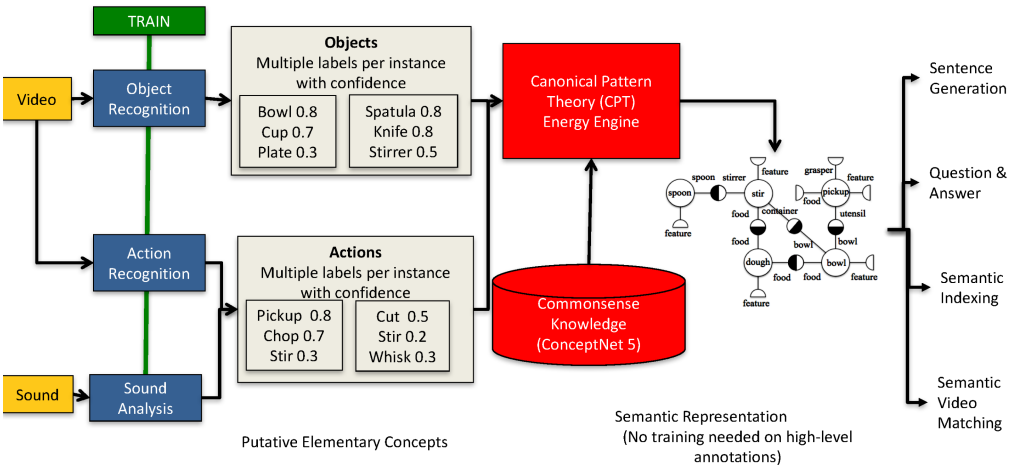


FIG. 1. Overall architecture of generating the canonical pattern theory representations. Deep learning or machine learning-based approaches hypothesize multiple object and action labels each object or action instance. Pattern theory formalism, resolves among these hypotheses and integrates information from ConceptNet to arrive at an interpretive representation, a connected structure expressed using Grenander's canonical representations in terms of generators and bonds. Note that only the modules in blue require explicit training. We do not have to train for compositions of objects and actions.

Figure 1 depicts the computational model with these characteristics, centered around canonical pattern theory. The starting point of our representation is elementary concepts, such as objects (nouns) and actions (verbs), found in the video, for which we will rely on existing approaches in the literature. We will not assume, however, that the labels produced by these algorithms are perfect. We allow for multiple putative labels for

each image evidence. We will integrate these words into richer descriptions of events using Grenander's pattern theory that impose both local and global regularity over the elementary concepts. Probabilistic structures will be imposed on these structures, which will then be optimized to arrive at inferences of most likely (or best-k) event structures.

Grenander's pattern theory [27,28] offers an elegant flexible, compositional framework to represent events as a multi-graph semantic network. The network is a configuration of primary entities, termed generators (or nodes), connected using bonds (or edges) under pre-defined rules. The linked compositions of objects (nouns) and simple actions (verbs), such as "whisk eggs with fork in bowl", will implicitly capture the language aspects of events. The optimality of a configuration will be defined using a global posterior energy that has contributions from both data likelihood and prior knowledge. We have explored this formalism in the past for constructing activity interpretation in audio and video [18–20], and have found it superior to other fixed structured graphical representations such as MRF, CRF, Bayesian Networks, or HMMs in its representational capacity. In this work, we focus on the use of common sense knowledge bases such as ConceptNet [39, 55] as the source of prior knowledge, instead of using handcrafted ontology, as we have done in the past.

We consider the built connected interpretation as an intermediate representation that forms the basis for generation of more well-formed expressions, such as sentences, or can be the basis for question and answers systems. These interpretations are similar to scene graphs that are descriptive of static scenes in images [3, 34, 64]. However, our video interpretations offer a much deeper understanding of the activity than labels or categories and also help in constructing descriptive sentences, answering questions, and retrieving similar videos. Some concepts in the interpretative structure have direct evidence from video, i.e., grounded concepts, and some are inferred concepts that bind grounded concepts, i.e., contextualization cues, not directly observed. In [2] we have shown how these descriptions can be used to generate explanations and participate in a question and answer session.

2. Common sense semantic knowledge source: ConceptNet. Before we present the details of the pattern theory representation, we describe briefly the ConceptNet framework. In the next section we will see how the nodes and links in the ConceptNet are mapped to pattern theory elements of generators and bonds. The ConceptNet will act as a prior knowledge graph that will be sampled during inference to arrive at interpretations that best describe the image evidence.

ConceptNet, proposed by Liu and Singh [39] and expanded to ConceptNet5 [54, 55], is a common sense knowledge base that maps concepts and their semantic relationships in a traversable semantic network structure. The sources of the knowledge include DBPedia, which extracts knowledge from the infoboxes on Wikipedia articles, Wiktionary, the free multilingual dictionary, WordNet, OpenCyc ontology, and Open Mind Common Sense. With over 3 million concepts, the ConceptNet framework serves as a source of cross-domain semantic information from general human knowledge while supporting common sense knowledge as expressed by humans in natural language. Technically, it encodes and

expresses knowledge in a hypergraph, with the nodes representing concepts and edges representing semantic assertions.

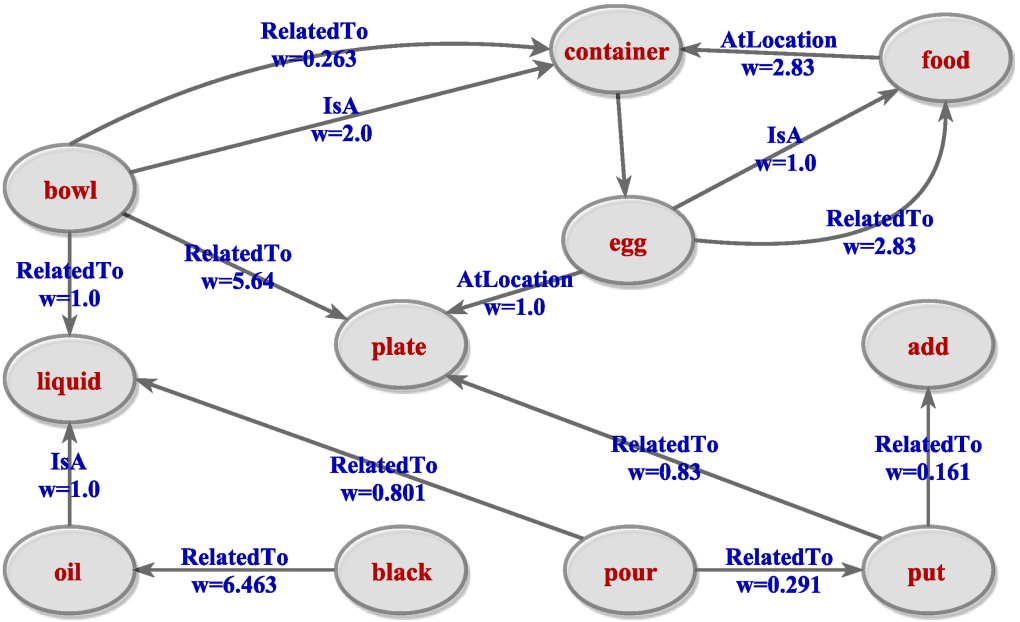


FIG. 2. ConceptNet is a semantic network of common sense knowledge. Illustrated here is a *small* snippet from ConceptNet to show how semantic relationships between concepts are expressed.

There are more than 25 relations (assertions) by which the different nodes are connected, with each of these relations contributing to the semantic relationship between the two concepts such as `HasProperty`, `IsA`, and `RelatedTo`. Each relation has a weight that determines the degree of validity of the assertion given the sources and hence provides a quantifiable measure of the semantic relation between concepts. Positive values indicate positive assertions and negative values indicate the opposite. Figure 2 illustrates these ideas; for example, the edge between nodes `egg` and `plate` represents an assertion with the relation `AtLocation` to indicate that eggs can be placed or found in plates. While ConceptNet has several assertions that represent different semantic relationships between the different concepts, we currently utilize a subset; more specifically—*RelatedTo*, *IsA*, *HasA*, *HasProperty*, *CapableOf*, *UsedFor*, *Desires*, and *Similarity*.

3. Video interpretation representation. Interpreting video activities, as with any pattern recognition, involves the modeling of the underlying pattern such as atomicity, regularity and an inference methodology for using the understanding of these basic properties of the pattern. Video activity interpretation consists of constructing a semantically coherent composition of basic, atomic elements of knowledge called `concepts` detected from videos. These concepts represent the individual actions and objects that

are required to form an interpretation of an activity. We use Grenander's canonical representation of general pattern theory [28] to build interpretations.

3.1. *Representing concepts using generators.* Following Grenander's notation [28], each concept represents a single, atomic component called a `generator` $g_i \in G_S$ where G_S is the `generator space`. The generator space represents a finite collection of all possible generators that can exist in a given environment. In our context, we consider the generator space to encompass all unique, non-repetitive concepts that can exist in ConceptNet representing actions (verbs) and objects (nouns), and feature vectors (data-based evidence).

Hence, the generator space (G_S) can be partitioned into three disjoint subsets that represent three types of generators—feature generators (F), grounded concept generators (G), and ungrounded context generators (U). Feature generators $F = \{g_{f_1}, \ldots, g_{f_q}\}$ represent individual feature subsets extracted from videos; with each subset being a possible action or object. Then there are generators that represent basic concepts such as elementary actions, such as pickup, stir, cut, or objects, such as knife, spoon, or plate. In our approach, the collection of all concepts present within ConceptNet serves as the generator space for concept generators.

These concept generators are of two types: grounded or ungrounded. `Grounded concept generators` represent concepts for which we can have direct evidence in the video, i.e., there are automated detectors for them. For instance, we have a classifier that labels all utensils. We will use $G = \{\underline{g}_1, \ldots, \underline{g}_k\}$ to represent this set. `Ungrounded context generators` represent concepts for which we do not have direct detectors available. For instance, while we have direct detectors for individual utensils, we might not have direct detectors for the category "utensils". We use $U = \{\overline{g}_1, \ldots, \overline{g}_q\}$ to represent this generator subset.

3.1.1. *Feature generators.* Feature generators represent pieces of video regions that can represent concepts such as elementary actions and objects. To allow for flexibility in implementation, we consider two different types of features—handcrafted and deep features. We experimented with three different strategies for extracting deep feature representations.

- First, we used deep learning models such as convolutional neural networks (CNN) on images to capture the feature descriptions for objects and CNNs based on optic flows for actions (CNN-Flow) as was done in [17].
- In the second strategy, we followed the work in [61] and used mean-pooled values extracted from fc_7 layer for each frame from a CNN model pre-trained on a subset of the ImageNet dataset [49]. This allowed us to exploit the spatial features extracted from the video sequences and hence gives a suitable representation of the content of the video while still allowing for some uncertainty in its generation of labels for actions and objects in the video input.
- Finally, we used state-of-the-art features extracted from the two-stream architecture proposed by [52], following the work in [50] by training two VGG-16 networks on both RGB frames and stacks of optical flow images, following the two-stream architecture. For better modeling the temporal sequences, we trained

a long-short term memory model (LSTM) of the recurrent neural networks as an additional layer on top of the two-stream architecture.

Handcrafted features consist of a histogram of optical flow (HOF) [12] for generating action labels and a histogram of oriented gradients (HOG) [15] for object labels. HOF features were extracted by computing dense optic flow frames from three temporally sequential segments—each representing the start, development, and end of the action sequence, respectively. A histogram of optic flow (weighted by magnitude) was then constructed for these temporal segments to characterize the integral stages of the action. The composite feature for action recognition is the ordered concatenation of the individual HOFs.

In addition to video-based inference, we allowed for audio-based inference using bag-of-audio words (BoAW) and spectrogram features as auditory feature descriptors [17].

It is to be noted that while other more sophisticated features are possible, these suffice for now to demonstrate feasibility of the proposed approach.

3.1.2. *Populating grounded concept generator set.* For any video, the grounded concept generator set is finite and corresponds to the possible object and elementary action labels that can be supported by the detected features. For each feature generator, we considered top-k ($k = 3$ *to* 5) possible object or action labels. These possible labels form the grounded concept generator set. This helps us overcome feature errors.

The labels corresponding to each feature generator are constructed using trained machine learning models. A linear support vector machine (SVM) classifier based its labels on *HOF*, *HOG*, and *CNN-Flow* features. Fully connected neural network classifiers were used with the *mean-pooled* features and the *Two StreamFlow* features. This training of the atomic action and object classifiers represents the only training needed in our approach.

3.1.3. *Ungrounded concept generator set: the priors.* Ungrounded concept generators represent the prior information that can be used to explain the grounded concept generators. These are concepts that are not detected directly in the input data but are essential to understanding the relationships among the grounded concept generators. In principle, this includes nodes in the entire ConceptNet, which is huge. So to constrain the inference combinatorics, we limit the ungrounded concept generator set to be composed of concepts that link two grounded concept generators in the ConceptNet.

Formally, let grounded concept generators be represented by $\underline{g}_i$ for $i = 1, \ldots, N$ and let ${}_{\underline{g}_i}R_{\underline{g}_j}$ represent an assertion between two concepts in the ConceptNet. Then, the considered ungrounded concept generator, $\overline{g}_k$, satisfies the following expression:

$$\text{not}\left({}_{\underline{g}_i}R_{\underline{g}_j}\right) \wedge \left({}_{\underline{g}_i}R_{\overline{g}_k}\right) \wedge \left({}_{\overline{g}_k}R_{\underline{g}_j}\right). \tag{1}$$

There can exist multiple ungrounded concept generators that connect two grounded concepts; we keep all of them in the set U. The optimal contextualization cue would be based on minimizing an overall energy; more on this later.

3.2. *Connecting generators: Bonds.* Each generator g_i has a fixed number of bonds called the `arity` of a generator denoted by $w(g_i) \forall g_i \in G_S$. Bonds are differentiated at a structural level by the direction of information flow that they represent—*in-bonds* and *out-bonds* as can be seen from Figure 3 (a) where the bonds representing *RelatedTo*

and *feature* represent in-bonds and *HasProperty* and *IsA* represent out-bonds for the generator *put*. These bond types are taken from the ConceptNet assertions.

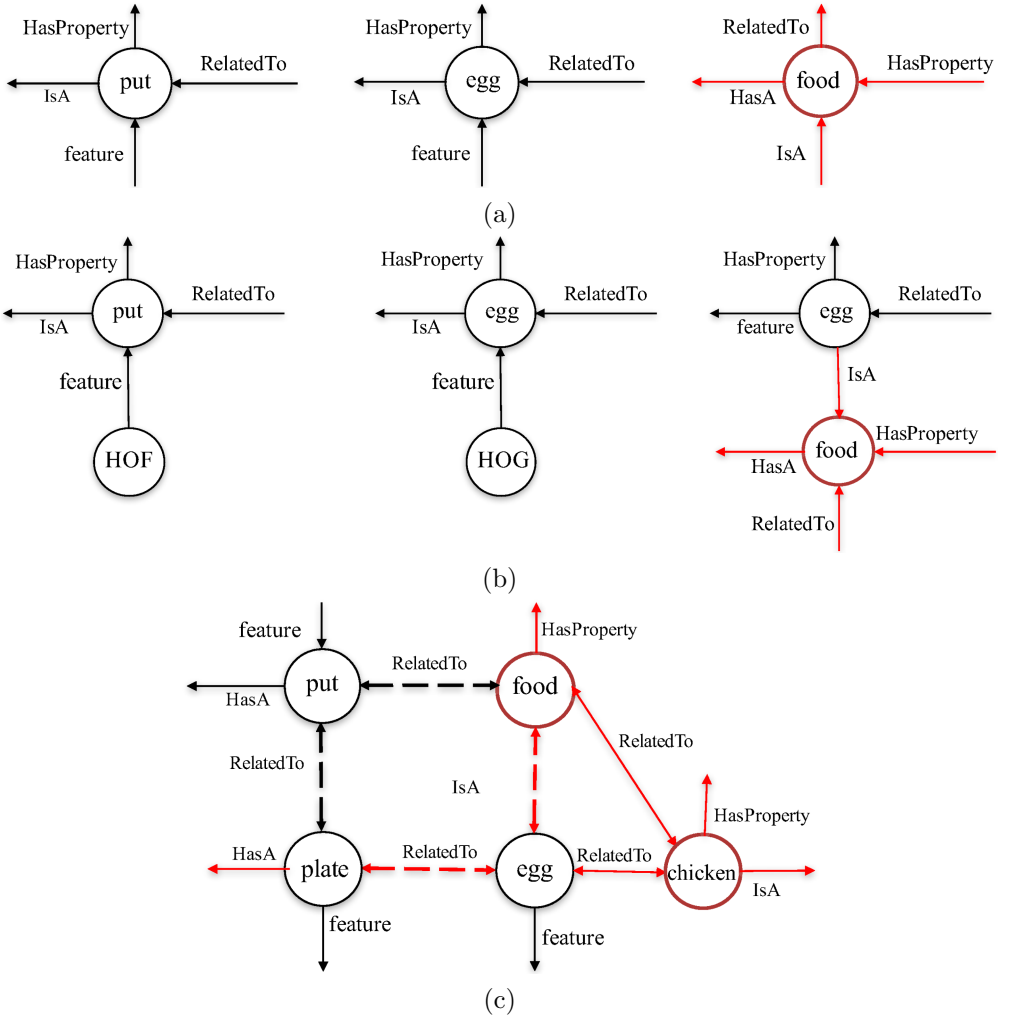


FIG. 3. An illustration of generators and their bond structures. (a) gives the structure of individual generators. The generators in black represent grounded generators and those in red represent ungrounded generators. (b) represents bonded pairs of generators. (c) represents a complete configuration representing an interpretation for the video "*Put egg on plate*".

Each bond is identified by a unique coordinate and bond value taken from a set B such that the jth bond of a generator $g_i \in G_S$ is denoted as $\beta^j_{dir}(g_i)$, where dir denotes the direction of the bond.

3.2.1. *Bond compatibility.* The viability of a *closed* bond between two generators is determined by the `bond relation function` ρ. This function determines whether two bonds $\beta(g_i)$ and $\beta(g_j)$ between two generators, g_i and g_j, are compatible and is denoted by $\rho[\beta(g_i), \beta(g_j)]$ or simply as $\rho[\beta, \beta]$. This function represents whether a given bond $\beta^j_{dir}(g_i)$ is either *closed* or *open*. The bond relation function is given by

$$\rho[\beta(g_i), \beta(g_j)] = \{TRUE, FALSE\}; \forall g_i, g_j \in G_S. \tag{2}$$

A bond is said to be `open` if it is not connected to another generator through a bond; i.e., an out-bond of a generator g_i is connected to a generator g_j through one of its in-bonds or vice versa. For example, take the first case from Figure 3 (b) representing the bonded generator pair {*put* and *HOF*}. The bonds representing *HasProperty*, *IsA*, and *RelatedTo* are considered to be *open*, whereas the bond representing *feature* represents a *closed* bond.

3.2.2. *Types of bonds.* There exist two types of bonds—`semantic bonds` and `support bonds`. Semantic bonds are a representation of the semantic relationship between two concept generators. These bonds represent the semantic assertions from ConceptNet that we discussed in Section 2. The direction of *semantic bonds* signify the semantics of a concept and the type of relationship shared with its bonded generator. For example, in Figure 3(b), the bond *IsA* between concepts *egg* and *food* is a symbolic representation of the semantic assertion that *Egg is a* (*type of*) *Food*, signified by the direction of the bond. Similarly, Figure 4 illustrates an example configuration with the generators *pour*, *oil*, *liquid*, and their semantic relationships given by semantic bonds *RelatedTo*, *IsA*. The bonds highlighted in red indicate the presence of ungrounded context generators, representing the presence of contextual knowledge. Semantic bonds are quantified using the strength of the semantic relationships between generators through the bond energy function:

$$a_{sem}(\beta'(g_i), \beta''(g_j)) = \tanh(\phi(g_i, g_j)), \tag{3}$$

where $\phi(.)$ is derived from the strength of the assertion in ConceptNet between concepts g_i and g_j through their respective bonds β' and β''. The tanh function normalizes the output from $\phi(.)$ to range from -1 to 1. This is important to note as there can exist negative assertions between two concepts that are not compatible and hence reduces the search space by avoiding interpretations with contrasting semantic assertions.

Support bonds connect (grounded) concept generators to feature generators that represent direct image evidence. These bonds are used to preserve the provenance of the concepts with respect to direct data-based evidence. Support bonds are quantified through the bond energy function:

$$a_{sup}(\beta'(g_i), \beta''(g_j)) = \tanh(f(g_i, g_j)), \tag{4}$$

where $f(.)$ is derived from the confidence scores of classification models between feature generators g_i and the respective concept generator g_j through their respective bonds β' and β''.

3.3. *Interpretations: Configurations of generators.* Generators can be combined together through their local bond structures to form structures called *configurations*, c, that represent semantic interpretations of video activities. Each configuration has an

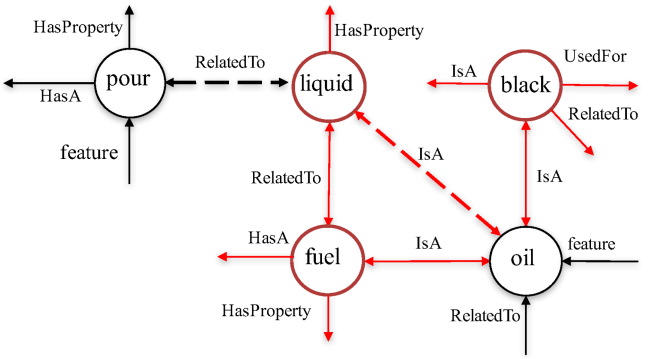


FIG. 4. Representation of an interpretation using pattern theory. Grounded concepts are represented in black while ungrounded (contextualization cues) are in red. The dashed links represent the optimal semantic relationship between two grounded concepts. Note: open bonds allow for the expansion of the interpretation, if desired. Also, only a selection of the bond structure of each generator is shown in this example.

underlying graph topology, specified by a connector graph σ. The set of all feasible connector graphs σ is denoted by Σ, also known as the connection type. Formally, a configuration c is a connector graph σ whose sites $1, \ldots, n$ are populated by a collection of generators $g_1, \ldots, g_n$ expressed as,

$$c = \sigma(g_1, \ldots, g_i); g_i \in G_S. \tag{5}$$

The collection of generators $g_1, \ldots, g_i$ represents the semantic content of a given configuration c. For example, the collection of generators from the configuration in Figure 4 gives rise to the semantic content "*pour oil* (*liquid*) (*fuel*) (*black*)".

The set of all feasible connector graphs σ is formally denoted by Σ, also known as the connection type. If one restricts this to an undirected fixed lattice structure, we get a Markov Random Field (MRF). If Σ is a directed acyclic graph with a fixed number of sites, we get a Bayesian Network or Dynamic Bayesian Network. If Σ is restricted to an and-or tree structure, we have AND-OR graphs. We adopt a more flexible structure than these, allowing for a different number of sites; specifically, we use the constraints of a partially ordered set (POSET) to capture the hierarchical nature of the generators. There is ordering between the levels of the hierarchy, but no ordering within a hierarchy level. In our framework, the hierarchy is set up such that feature generators are at the bottom level, the grounded generators are at a higher level, and the ungrounded generators are at the highest level. In general, if g_i connects some out-bond $\beta(g_i)$ to an in-bond $\beta(g_j)$ of another generator g_j, then $l(g_i) \geq (g_j)$.

3.4. *Regular configurations.* The formal specification of regularity in the proposed framework is denoted by $R(G_S, \rho, \Sigma)$, where G_S denotes the generator space, ρ the bond compatibility, and Σ the connector type. The regularity R specifies the principles that govern the construction of regular structures to represent patterns. Additionally, the

regularity helps define efficient operations on the regular configuration space $C(R)$ that does not violate the semantics of the patterns that are formalized. This will be beneficial in the design of efficient algorithms to perform probabilistic analysis of these structures.

To formalize the notion of *regular configurations*, we determine a configuration c to be called *locally regular* if

$$\bigwedge_{\forall(\beta',\beta'')\in c} \rho(\beta'(g_i), \beta''(g_j)). \tag{6}$$

Equation (6) is known as the first structure formula. A configuration c is said to be *globally regular* if $\sigma \in \Sigma$.

A configuration is then called regular if it is both locally and globally regular. The set of all regular configurations is denoted by C(R). This formal notion helps us design inference algorithms that search the regular configuration space C(R) in an efficient and smart fashion. Note also that C(R) represents the union of all subspaces $C(\sigma), \forall\sigma \in \Sigma$.

3.5. *Probabilistic superstructure of configurations.* Given a set of video feature generators, F, and the prior knowledge in terms of the ConceptNet graph, C_N, our goal is to find an interpretation, c, that obeys the regularity $R(G_S, \sigma, \rho)$. While this first structure regularity captures the conformity of the given configuration to local and global structure constraints, it does not measure the degree of regularity, which is necessary to able to choose among configurations. This degree is given by the *second structure formula* that uses the bond energy weights that were defined by equations (3) and (4). The second structure formula quantifies the first structure formula with a probability density function $p(c|C_N, F)$ on the configuration space $C(R)$. We factor this probability into two parts: a likelihood term, $p(F|c)$ and a prior, $p(c|C_N)$, normalized by the distribution over the features

$$p(c|C_N, F) = \frac{p(F|c)p(c|C_N)}{p(F|C_N)}. \tag{7}$$

This probability can be captured using energy functions

$$P(c|C_N, F) = \frac{1}{Z} e^{-E(F|c)-E(c|C_N)}, \tag{8}$$

where $E(F|c)$ represents the energy of the configuration c that involves the grounded generators and the detected features, while $E(c|C_N)$ captures the energy of the ungrounded, prior, generators. The total energy $E(c)$ of a configuration c is the sum of these energies: $E(c) = E(F|c) + E(c|C_N)$.

We capture the energy of a configuration by the energy of all the individual bonds present in the configuration. The likelihood energy term represents the contribution of support bonds to the overall energy. This is a reflection of the confidence of the underlying machine learning models and was represented by:

$$E(F|c) = -\sum_{(\beta',\beta'')\in c} a_{sup}(\beta'(g_i), \beta''(g_j)). \tag{9}$$

To capture the prior energy of the configuration, we use the combined energy of all semantic bonds in the configuration and a structure quality prior term as follows:

$$E(c|C_N) = -\sum_{(\beta',\beta'')\in c} a_{sem}(\beta'(g_i), \beta''(g_j)) + Q(c). \tag{10}$$

A lower energy means that the generators (concepts) in the configuration are closely associated with each other based on their semantic associations. We chose the structure quality prior term $Q(c)$ to be:

$$Q(c) = k \sum_{\bar{g}_i \in G'} \sum_{\beta^j_{out} \in \bar{g}_i} [D(\beta^j_{out}(\bar{g}_i))], \tag{11}$$

where G' is a collection of ungrounded contextual generators present in the configuration c, β_{out} represents each *out-bond* of each generator g_i and D(.) is a function that returns true if the given bond is open. k is an arbitrary constant that scales the extent of the detrimental effect that the ungrounded context generators have on the quality of the interpretation. The cost factor $Q(c)$ restricts the inference process from constructing configurations with degenerate cases such as those composed of unconnected or isolated generators that do not have any closed bonds and as such do not connect to other generators in the configuration.

The partition function, Z, in equation (8), involves a double sum: $\sum_\sigma \sum_c$. The first sum is over the possible global structures and the second is over all possible generator combinations for any given global structure. We have a grand Gibbs ensemble. This makes it computationally hard to make exact inferences.

4. Inferring interpretations. Searching for the best semantic description of a video involves maximizing the probability of a configuration and hence minimizing the energy function $E(c)$. This optimization process represents the inference process. The solution space is very large as both the number of generators and underlying structure structures can be variable. For example, the combination of a single connector graph σ and a generator space G_S give rise to a space of feasible configurations $C(\sigma)$. While the structure of the configurations $c \in C(\sigma)$ can be identical, their semantic content is varied due to the different assignments of generators to the sites of a connector graph σ.

4.1. *Explore solution space using MCMC simulated annealing.* A feasible optimization solution for such exponentially large space is to use a sampling strategy and employ an efficient Markov Chain Monte Carlo (MCMC) based simulated annealing process. The MCMC based simulated annealing method uses two proposal functions that are representative of a move in the search space—the *configuration reset move* and *grounded switch move* functions. Each move function in the simulated annealing process proposes a candidate configuration to aid in the search for the optimal configuration that best captures the semantics of the given video activity. The *configuration reset* move function allows the search to reset or initialize to a random configuration that helps in exploring the search space in a more efficient manner. This move is used to both initialize the search as well as to reset the search to a random start when the current search state is in the middle of a local minima that could be hard to handle with smaller changes. The *grounded switch* move function, on the other hand, proposes candidate configurations based on smaller, Markovian changes to the current configuration by switching the grounded concept generators.

Algorithm 1: MCMC based simulated annealing for inference

```
MCMC Simulated Annealing (F, G, U, α, p, k_max, T_0);
c ← resetConfiguration(F, G, U)
best ← c
for k ← 1 ... k_max: do
    t ← UniformSample(0, 1)
    if t < p then
        c' ← resetConfiguration(F, G)
    end
    else
        c' ← groundedSwitch(c, G, U)
    T ← T_0 × α^k
    if E(c') < E(c) then
        c ← c'
    end
    else
        z ← UniformSample(0, 1)
        if z < exp(−(E(c') − E(c))/T) then
            c ← c'
        end
    if E(c) < E(b) then
        best ← c
    end
end
return best
```

The algorithm for the MCMC-based simulated annealing process is shown in Algorithm 1. We begin with the set of detected feature generators F, the corresponding set of plausible grounded concept generators G, and the ungrounded concept generators U generated through ConceptNet that form the background knowledge of grounded concept generators. We initialize the search through the configuration reset function, which samples an initial configuration c' that provides a starting point for the search. The proposed configuration is then used as initialization for the "*best*" configuration seen so far.

The search is initiated and performed for a fixed number of iterations k_{max} defined in the parameters. The choice between the two move proposal functions is decided through sampling of a value in the uniform distribution between 0 and 1. If the sampled value t is in the range 0 to p, then the reset configuration move function is called to reset the configuration to a random starting configuration. Hence, the grounded switch move function is called with a probability $1 - p$. The value of the probability p is given in the parameters to the annealing function. At each step of the annealing process, the temperature is updated based on a cooling rate given by α^k, where α is a pre-defined

constant. Each step of the simulated annealing process yields a new configuration c', which is either accepted or rejected. The proposed configuration (c') is accepted if its energy is lower than the current configuration c. The proposed configuration is also chosen with a certain probability if its energy is proportional to the energy difference between the current and proposed configurations.

4.2. *Constructing and modifying configurations.* Both move functions employ the use of three processes that allow them to construct candidate configurations in the search process. The processes are (1) *grounding process*, (2) *grounded switch process*, and (3) *contextualization process*. Each process makes changes to an existing configuration using a subset of the generator space (G_S) to guide their execution: the grounding process uses the feature generator (F) and grounded concept generator (G) subspaces. The grounded switch process and the contextualization process use the grounded concept generator (G) and the ungrounded concept generator (U) subspaces, respectively.

4.2.1. *Grounding process.* The grounding process involves the establishment of support bonds between existing feature generators in the configuration and the grounded concept generators in the configuration. Recall that we allow for multiple labels for each feature evidence. When creating a configuration, the grounding process proposes a grounded concept generator that explains the presence of a feature generator in the configuration. This is the first step in the configuration reset move function (Section 4.3) as illustrated in Algorithm 2 with line 5. A grounded concept generator g_i is chosen at random from a uniform distribution from G, such that it explains the presence of the corresponding feature generator $\underline{g}_i$. The chosen grounded generator is added to the configuration c and possible support bonds are established between the generator pair. The bond energy is quantified based on the confidence score as seen from equation (4). An example of a bonded feature generator-grounded generator pair can be seen from Figure 3 (b), where bonded pairs {*HOG*, *egg*} and {*HOF*, *put*} are presented.

4.2.2. *Grounded switch process.* The grounded switch process is used by the grounded switch move proposal function to explore the search space using guided, Markovian changes based on the grounded concept generator subspace (G) of the generator space (G_S). An existing grounded concept generator g_k is chosen based on two different mechanisms—random and guided—and removed from the configuration c. In the random mechanism, the grounded generator is chosen from c based on a uniform distribution. In the guided mechanism, the grounded generator g_k is chosen such that it possesses the least semantic significance within the configuration; i.e., the chosen grounded generator has closed semantic bonds that possess high energy. Then, all ungrounded concept generators that share a closed semantic bond with the selected generator g_k are also chosen and removed from the configuration. m replacement candidates are chosen from the subset G of the generator space G_S such that $g_i \in G_S$ and $g_j \neq g_i$. Then, a grounded generator is chosen such that it minimizes the energy of the configuration c', and possible support bonds are established. The bond energy is quantified based on the confidence score as seen from equation (4). This is the first step in the configuration reset move function (Section 4.4) as illustrated in Algorithm 3 with lines 2-9.

4.2.3. *Sampling from ConceptNet using contextualization.* For ungrounded generators to provide semantic context to grounded generators, we sample from the ConceptNet using the concept of contextualization. We borrow this concept from linguistics [31], which in the context of video interpretation to refer to the integration of past knowledge to aid in interpreting activities in videos. More specifically, "*concept*" refers to actions and objects that constitute an activity; "*presuppositions*" refers to the background knowledge of concepts, their properties and semantics. Note that the goal is to generate *interpretations* of a given activity rather than just simple recognition.

The contextualization process allows semantic relationships to be established among the grounded concepts in a configuration. It is an essential process shared by both move proposal functions as seen from line 6 of Algorithm 2 and lines 4 and 10 from Algorithm 3. It involves the extraction of contextualization cues from ConceptNet and the search for the optimal semantic relationships between two grounded concept generators.

Formally, let ungrounded concept generators be represented by $\overline{g}_i$ for $i = 1 \cdots N$ and let ${}_{g_i}R_{g_j}$ represent relations (or assertion) between two concepts. For example, consider the example configuration in Figure 5, which contains a connected structure of generators which represent concepts from the ConceptNet. The generators highlighted in gray, i.e., g_5 and g_7 represent the grounded concept generators with data evidence from the input video data. The other generators represent the contextual concept generators or ungrounded concept generators extracted from ConceptNet. The ungrounded concept generators are derived up to a given depth d. The depth d controls the level of contextualization cues that are connected to the grounded generators. For example, the generator $\overline{g}_1$ is at a depth 2 from the grounded generator $\underline{g}_5$. Each of these concept generators are connected through weighted connections or semantic assertions given by w_{ij} for a given generator pair $\{g_i, g_j\}$. These weights are quantified representations of the semantic relationship between the two concepts g_i and g_j present in the ConceptNet Framework. For example, in Figure 2, if the concepts *egg* and *food* are g_i and g_j, respectively, then w_{ij} is $1.0, 2.0$ and 4.88 for the assertions *IsA*, *UsedFor*, and *RelatedTo*, respectively.

The process of constructing the optimal contextualization cue for two given grounded generators g_i and g_j is as follows:

(1) Extract the subgraph of all connected concepts from ConceptNet that represent the contextual properties of a given generator g_i up to a given depth d.
(2) Construct sub-configurations that are representative of all concepts and hence subsequent semantic relationships that are able to connect the two grounded generators.
(3) Find the optimal sub-configuration that minimizes the energy.

For example, consider the subgraph extracted from ConceptNet for two grounded concept generators g_5 and g_7 in Figure 5. The complete sub-graph that comprises all concept generators related to all detected grounded generators is given by the configuration:

$$c = \sigma(g_0, g_1, g_2, g_3, g_4, g_5, g_6, g_7, g_8, g_9, g_{10}, g_{11}). \tag{12}$$

The sub-configuration that represents the contextual information for the generator g_5 is given by the configuration:

$$c_1 = \sigma(g_0, g_1, g_2, g_3, g_4, g_6, g_9). \tag{13}$$

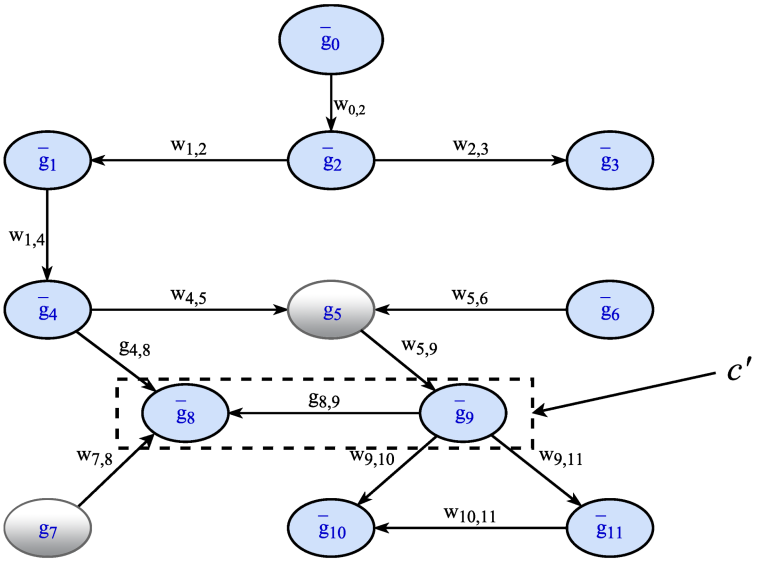


FIG. 5. An example of a subgraph extracted from ConceptNet connecting the grounded concept generators g_5 and g_7, shaded in gray. The subconfiguration c' represents the optimal contextualization cues (g_8 and g_9, shaded in blue) connecting the grounded concept generators.

Similarly the sub-configuration that represents the contextual information for the generator g_7 is given by the configuration:

$$c_2 = \sigma(g_8, g_{10}, g_{11}). \tag{14}$$

The goal is to find the optimal configuration that minimizes the energy of the overall configuration that is representative of the semantic interpretation constructed. Let the optimal sub-configuration that is representative of the optimal contextualization cue that connects the generators g_5 and g_7 be c'

Hence the probability of the sub-configuration that connects the two configurations c_1 and c_2 is given by:

$$P(c'|c_1 \cup c_2) = P(c'|c), \tag{15}$$

where c' is the sub-configuration that represents the contextualization cues that give the optimal semantic relationships between two grounded generators. The probability of a configuration c is given by the sum of bond energies within the configuration given in equation (12).

Hence,

$$P(c'|c) = \frac{\sum\limits_{(g_i,g_j)\in c'} a_{sem}(\beta'(g_i), \beta''(g_j))}{\sum\limits_{(g_i,g_j)\in c} a_{sem}(\beta'(g_i), \beta''(g_j))}. \tag{16}$$

Constructing the optimal contextualization cues for a given set of grounded concept generators is a probabilistic induction of the sub-configuration with minimal energy which is reflective of the semantic relationships among the grounded concept generators.

4.3. *Configuration reset move.* The configuration reset move proposal function is used to randomly sample a configuration from the search space. This move serves two major purposes—it is used to initialize the search process as well as reset the current search configuration to a random start which ensures that the search is diversified. When called during the search process, the resultant configuration is likely to possess a different structure and allows the grounded switch proposal function to explore the search space using ConceptNet in an efficient fashion as the probability of some generators (both grounded and ungrounded) varies due to the introduced changes.

Algorithm 2: Configuration Reset Proposal Function

```
resetConfiguration (F, G, U);
c' ← Empty Configuration
for g_i ∈ F: do
    Add feature generator g_i to configuration c'
    Add grounded generator g_k that explains g_i to c'; where g_k ∈ G
    Add ungrounded generators {ḡ_j} ∈ U to c' such that there exists g_j R_ḡk
end
return c'
```

The pseudocode for the configuration reset proposal function is shown in Algorithm 2. We begin by random selection of a feature generator $\underline{g}_i$ from the subset F based on a uniform distribution. The chosen feature generator is then added to a new configuration. A grounded concept generator g_k from G, that corresponds to the feature generator $\underline{g}_i$, is chosen at random from a uniform distribution, such that it explains the presence of the feature generator $\underline{g}_i$. This is performed by the use of the *grounding process* described in Section 4.2.1. The grounding process is repeated for all feature generators in F. Then, semantic relationships are established between the grounded concept generators in the configuration (Section 4.2.3). Note that when adding a new generator to a configuration, bonds are established and closed between compatible generator-bond pairs in the configuration. This process is visually illustrated in Figure 6. We use the Generator Space (G_S) from Figure 6(a) to call the grounding process to add grounded generators until all feature generators are explained. Then the contextualization process is called to establish semantic relationships among the grounded concept generators.

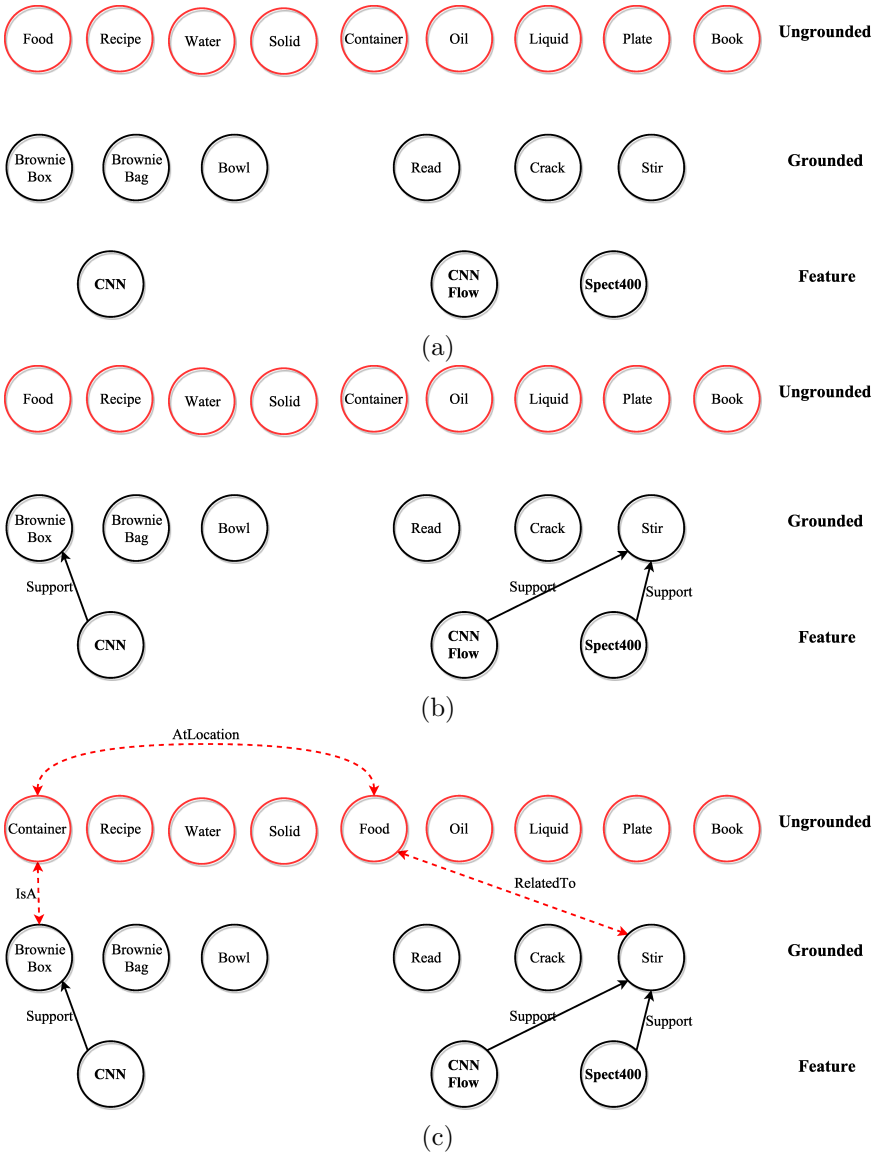


FIG. 6. An illustration of steps in the configuration reset proposal function. (a) shows the generator space specific to a given video of groundtruth "*Read Brownie*". (b) shows the functioning of the *grounding* process. (c) shows the functioning of the *contextualization* process to establish semantic relationships among the grounded concept generators.

4.4. *Grounded switch move to guide the search.* Initially, the configuration reset function introduces a set of grounded concept generators derived from machine learning classifiers. Then, a set of ungrounded context generators, representing the contextualization cues, are populated for each grounded concept within the initial configuration.

Bonds are established between compatible generators when each generator is added to the configuration. Each jump given by the grounded switch move function gives rise to a configuration whose semantic content represents a possible interpretation for the given video. Interpretations with the least energy are considered to have a higher probability of possessing more semantic coherence. Hence an additional optimization constraint is to minimize the cost factor Q given in (11) by ensuring the least number of open bonds for each ungrounded contextual generator in the configuration.

Algorithm 3: Grounded Switch Proposal Function

groundedSwitch (c, m);
Randomly select $g_k \in c$
Form a set G' of m generators g_i such that $g_i \in G_S$ and $g_j \neq g_i$
Form a set C' of generators $\{\overline{g}_j\}$ such that $\overline{g}_j \in C$ and ${}_{g_i}R_{\overline{g}_j}$ exists $\forall g_i \in G'$
Remove g_j from c
Remove $\{\overline{g}_j\} \in C'$ from c such that there exists ${}_{\overline{g}_i}R_{g_j}$
$c' \leftarrow c$
Select generator g_i that minimizes $E(\sigma(c', g_i))$
Add g_i to c'
Add $\{\overline{g}_k\} \in C'$ to c such that there exists ${}_{g_i}R_{\overline{g}_k}$
$c'' \leftarrow c'$
return c''

A swapping transformation is applied to switch the generators within a configuration to change the semantic content of a given configuration c. Algorithm 3 shows the grounded switch proposal function which induces the swapping transformation. We begin with an initial configuration c which is the current configuration in our search process. An existing grounded concept generator g_k is chosen and removed from the configuration c. Then, all ungrounded concept generators that share a closed semantic bond with the selected generator g_k are also chosen and removed from the configuration. m replacement candidates are chosen from the subset G of the generator space G_S such that $g_i \in G_S$ and $g_j \neq g_i$. Then, a grounded generator is chosen such that it minimizes the energy of the configuration c'. This is achieved using the grounded switch process from Section 4.2.2. Semantic relationships are established using the contextualization process from Section 4.2.3 in which ungrounded generators which share an optimal relationship with grounded generators in c are selected and added to the configuration. This results in a new configuration c', thus constituting a move in the configuration space $C(\sigma)$. This process is visually illustrated in Figure 7. The grounded switching process is used to switch grounded generators from the current configuration from Figure 7(a). Then the contextualization process is called to establish semantic relationships among the new set of grounded concept generators.

4.5. *Initialization of the search.* The search for the MCMC-based inference is initialized using the configuration reset move proposal which also ensures that the search is diversified through initialization of the configuration. The search is initialized using the

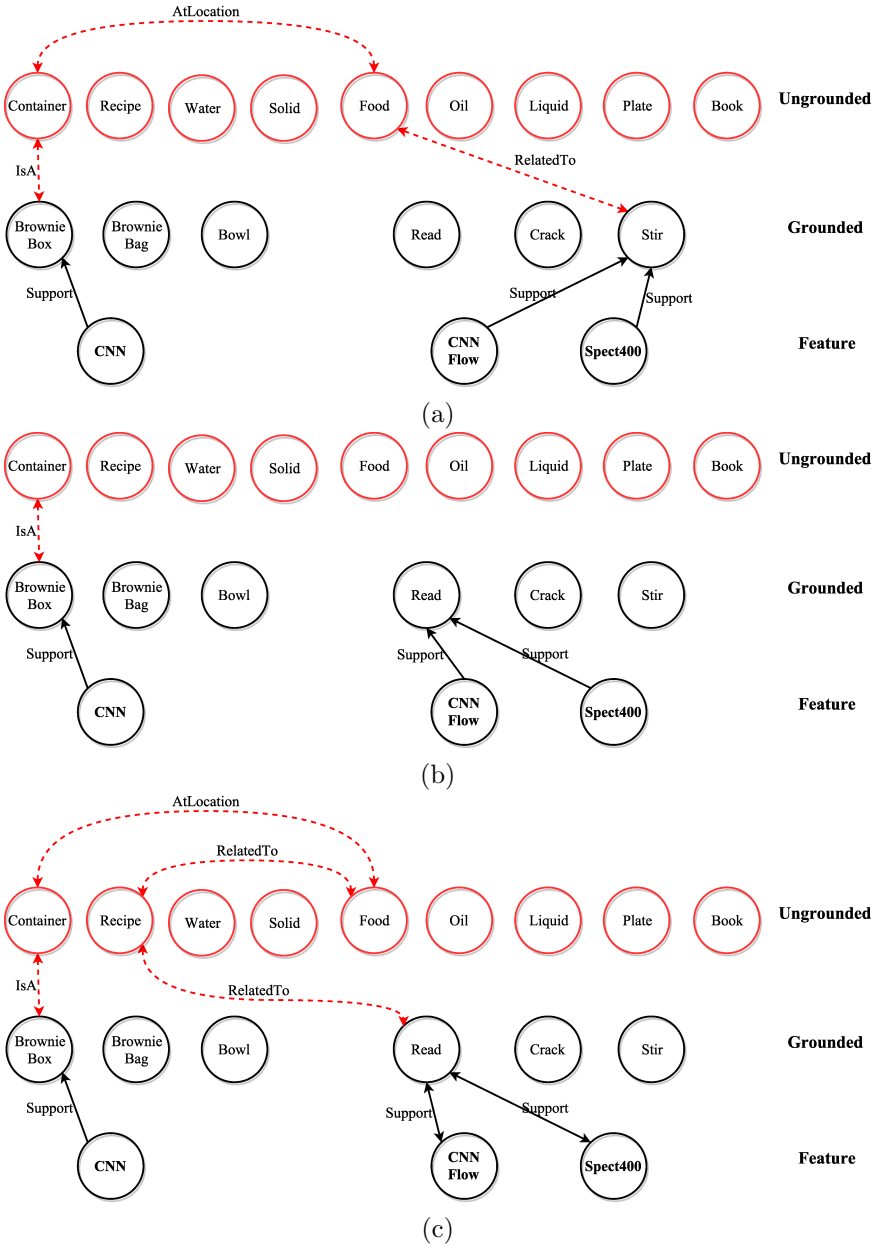


FIG. 7. An illustration of steps in the grounded switch proposal function. (a) shows the configuration in the current step of the search process. (b) shows the functioning of the *grounded switch* process. (c) shows the functioning of the *contextualization* process to establish semantic relationships among the new grounded concept generators.

reset configuration process described in Section 4.3. The resultant configuration is a randomly sampled configuration from the generator space G_S for a given configuration and

provides an initial starting point for the Markovian changes proposed by the grounded switch move function.

5. Experimental evaluation. We begin with discussion on the four publicly available datasets that we use, followed by presentation of qualitative and quantitative results on them. Performance is quantified using measures that facilitate the comparison with other approaches, such as precision, F-Score, and BLEU score [46].

Comparable approaches: We compare performances against a variety of competing approaches – discriminative methods (DM) [37], hidden Markov models (HMM) [37], context-free grammars (CFG+HMM) [37], factor graphs [59], different manifestations of deep recurrent neural networks [45, 61, 65], and generic pattern theory (PT) approaches [20]. We also consider a variation of pattern theory approach in [20], where we use simple, pairwise semantic relationships given by the ConceptNet Similarity edge weights called "PT+weights".

5.1. *Datasets.* The Charades dataset [51] is a challenging benchmark containing 9,848 videos across 157 action classes with 66,500 annotated activities, including nouns (objects), verbs (actions), and scenes. Complex co-occurrences of realistic, human activities offer a considerable challenge for the proposed framework along with the complex semantic relationships among the concepts. We use the same splits for training and testing from [51] and [50] and evaluate video classification using the evaluation criteria and code from [51] for fair comparison.

The Microsoft Video Description Corpus (MSVD) is a publicly available dataset that contains 1,970 videos taken from YouTube. On an average, there are 40 English descriptions available per video. We follow the split proposed in prior works [45, 59, 61, 65], and use 1,200 videos for training, 100 for validation, and 670 for testing.

The Breakfast Actions dataset consists of more than 1000 recipe videos, consisting of different scenarios with a combination of 10 recipes, 52 subjects and differing viewpoints captured from up to 5 cameras, which provides differing qualities of videos with varying amounts of clutter and occlusion. The units of interpretation are temporal video segments of these videos, given by the video annotation provided along with the dataset.

The Carnegie Mellon University Multimodal Activity dataset (CMU) contains multimodal measures of human activities such as cooking and food preparation. The dataset contains five different recipes: brownies, pizza, sandwich, salad, and scrambled eggs. Spriggs et al. [56] generated the groundtruth for some videos and recipes. The experiments were performed on the brownie recipe videos for performance comparison with [17].

5.2. *Qualitative evaluation.* The ability of the proposed approach using ConceptNet and contextualization to adapt to novel domains without explicit training for semantics can be seen from Figures 8, 9, 10, and 11. The approach was able to demonstrate a number of different traits that are demonstrative of its domain adaptability such as (1) inferring interpretations with grounded concepts whose semantics is beyond simple, pairwise relationships, (2) handling errors in underlying concept proposals, (3) inferring semantics beyond those from groundtruth, and (4) handle multiple modalities and varying viewpoints such as egocentric video data.

Figure 8 illustrates the semantic richness of the interpretations produced by our approach with contextualization cues that go beyond what is seen in the image. For example, take a video with groundtruth as "*fix hair*". Our approach was able to contend the presence of the grounded concepts *fix* and *hair* through the ungrounded contextual generators *prepare*, *groom*, and *comb*. Without contextualization cues, not only are there errors in the interpretations of the prior pattern theory approaches, but they are not as rich and descriptive.

It is also to be noted that for many of the interpretations, the label with the highest confidence score was not the one used in its final (best) interpretation. This is illustrated in Figure 9, where the approach was able to arrive at the correct interpretation even though the confidence scores for the correct grounded concept generators were lower than others. For example, the confidence scores of the action and object labels {*read*, *brownie box*} was lower than that of the combination {*stir*, *brownie box*}. However, the contextualization cues allowed for establishing semantic relationships beyond simple, direct relations to arrive at the correct interpretation.

The proposed approach was able to use infer interpretations of video activity that, while different from the groundtruth, conveyed the same semantic content. This is illustrated in Figure 10, where the approach was able to generate interpretations beyond the groundtruth semantics while preserving the semantic structure of the event. For example, when presented with a video with groundtruth "*Add salt and pepper*", our approach inferred an interpretation with semantics "*Spoon salt and pepper*". It is interesting to note that the action "*spoon*" had only been used in the context of "*Spoon butter*" in the dataset groundtruth annotations. This can arguably be attributed to the contextualization process' ability to handle uncertainty in underlying visual cues to arrive at a semantically coherent interpretation whereas domain specific pattern theory approaches had an interpretation of "*Spoon butter*". This is indicative of the power of semantics in ConceptNet and the ability of contextualization cues to provides a means for semantically connecting concepts beyond pairwise relationships.

Finally, as seen from Figure 11, when given with multiple modalities and/or difficult viewpoints such as egocentric data, the approach was able to use the prior knowledge from ConceptNet to arrive at interpretations that are semantically coherent. The challenges associated with egocentric video data (such as camera shake, occlusion, etc.) as well as multimodal sensory data fusion was handled well by our approach.

5.3. *Quantitative results.* In this section, we evaluate the proposed approach on the different datasets described in Section 5.1 and present the quantitative results of the experiments. We show that the proposed approach has competitive performance on a variety of challenges presented from each dataset—complex visual scenes due to co-occurrences of activities (Charades), complex semantic relationships in captioning (MSVD dataset), unbalanced class distribution and weak features (Breakfast Actions dataset), and multiple modalities (CMU Multimodal Dataset).

5.3.1. *Complex visual data.* For evaluation of the performance on data with complex semantic relationships, we use the **Charades** dataset [51] and report the Mean Average Precision (mAP) score to compare against other comparable approaches. It can be seen from Table 1 that the proposed approach is competitive in its performance. The

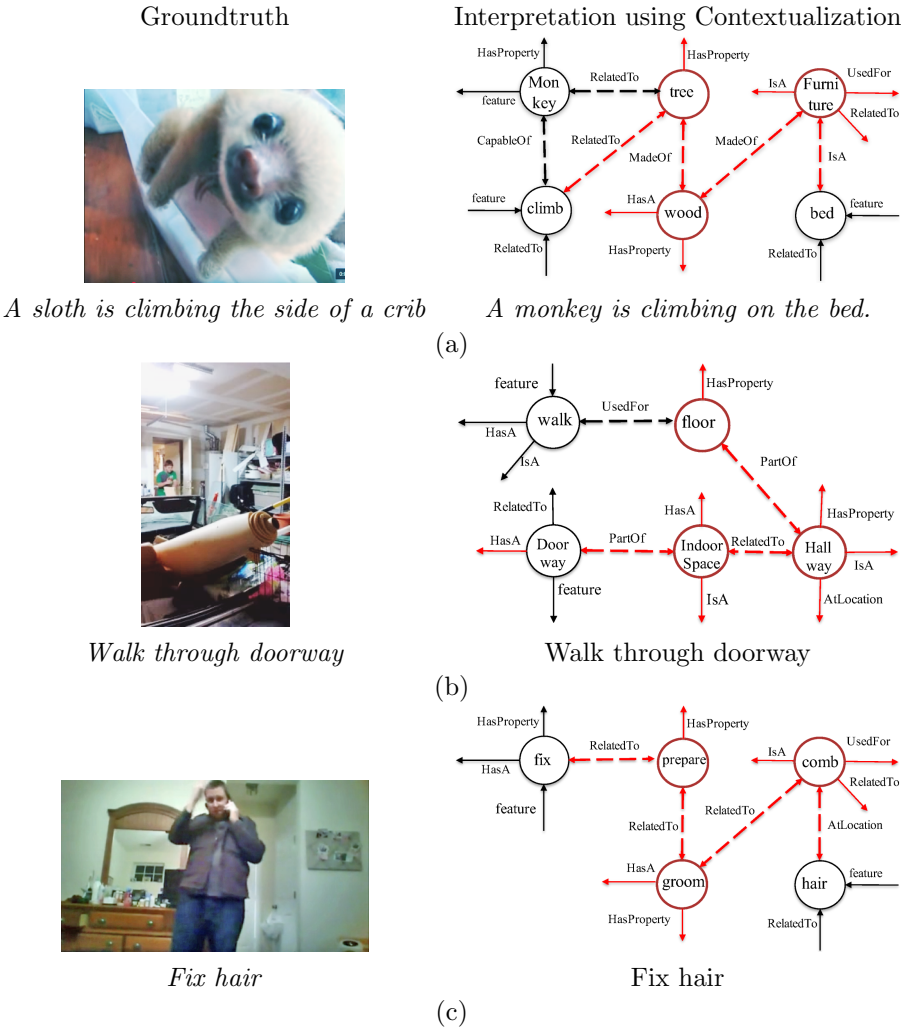


FIG. 8. Qualitative examples of interpretations generated by the approach in this paper. The first column shows the groundtruth for the input video, the second column shows interpretations generated by the pattern theory approach using ConceptNet and contextualization. It can be seen that contextualization generates rich, deep semantic interpretations which are able to allow for semantic relationships beyond simple, pairwise relationships.

Asynchronous Temporal Fields (ATF) approach [50] factors both sequential temporal information and intent. It should be noted that our approach is outperformed only when the ATF approach factors both temporal sequencing and intent.

We would like to point out that temporal sequence modeling has been successfully included in the prior pattern theory approaches [53], showing improvements up to 30% in long temporal sequence. We did not wish to conflate this paper with the modeling

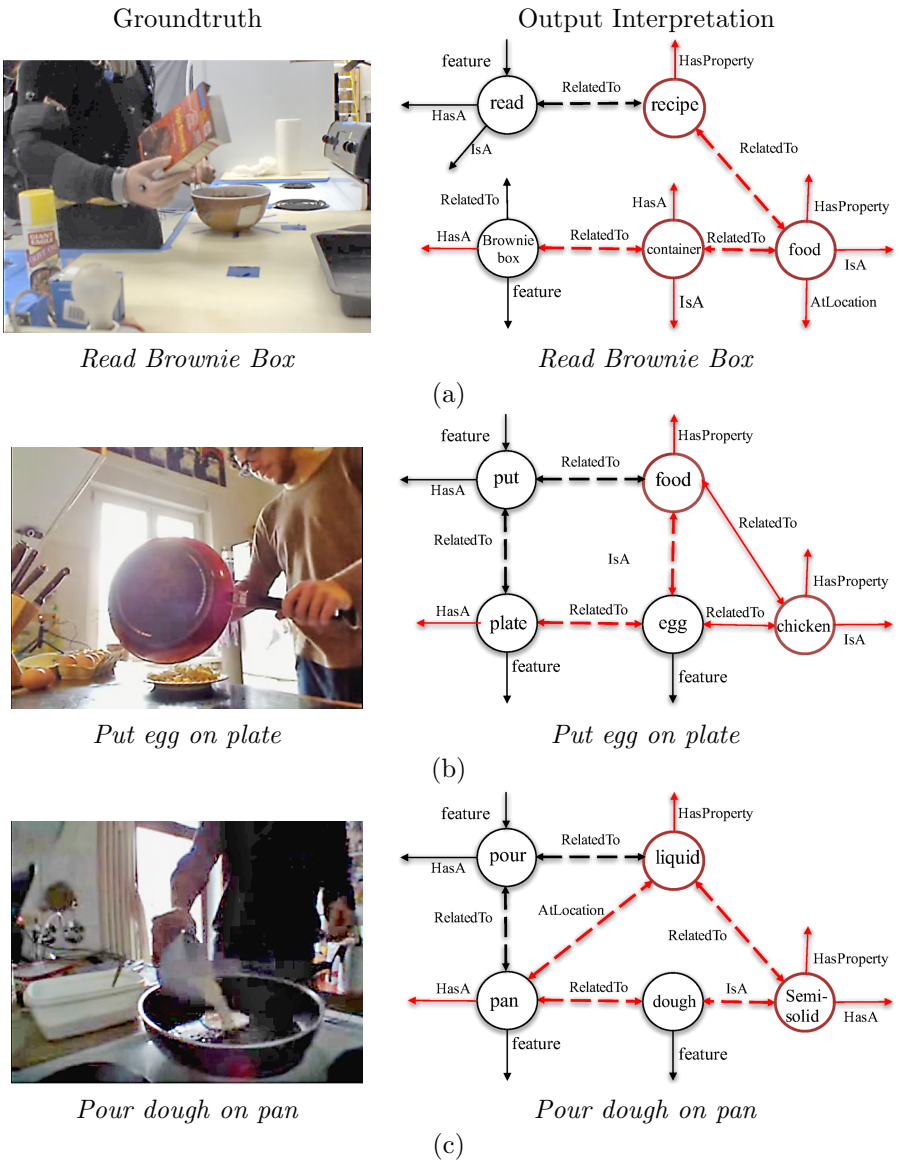
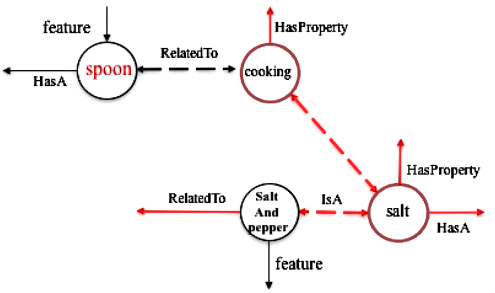


FIG. 9. Qualitative examples of interpretations generated by the approach in this paper where the concepts in the final (correct) interpretation were not the top prediction from underlying machine learning models. The first column shows the groundtruth for the input video and the second column shows interpretations generated by the pattern theory approach using ConceptNet and contextualization. It can be seen that contextualization allows for errors in the underlying models, yet uses prior knowledge to correct such errors in the final interpretation.

Groundtruth

Add salt and pepper.

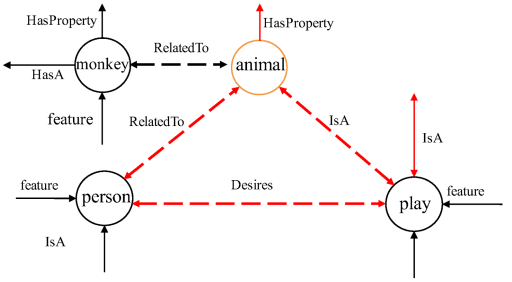
Output Interpretation



Spoon salt and pepper

(a)

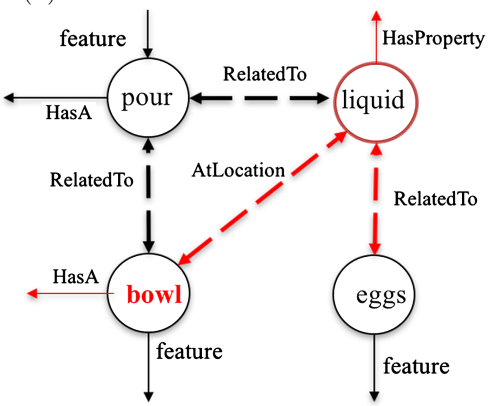
A monkey is fighting with a man



A monkey is playing with a human.

(b)

Pour eggs on pan.



Pour eggs into bowl.

(c)

FIG. 10. Examples of instances where the model was able to generalize beyond the semantics within the dataset but was still able to convey the same semantic content.

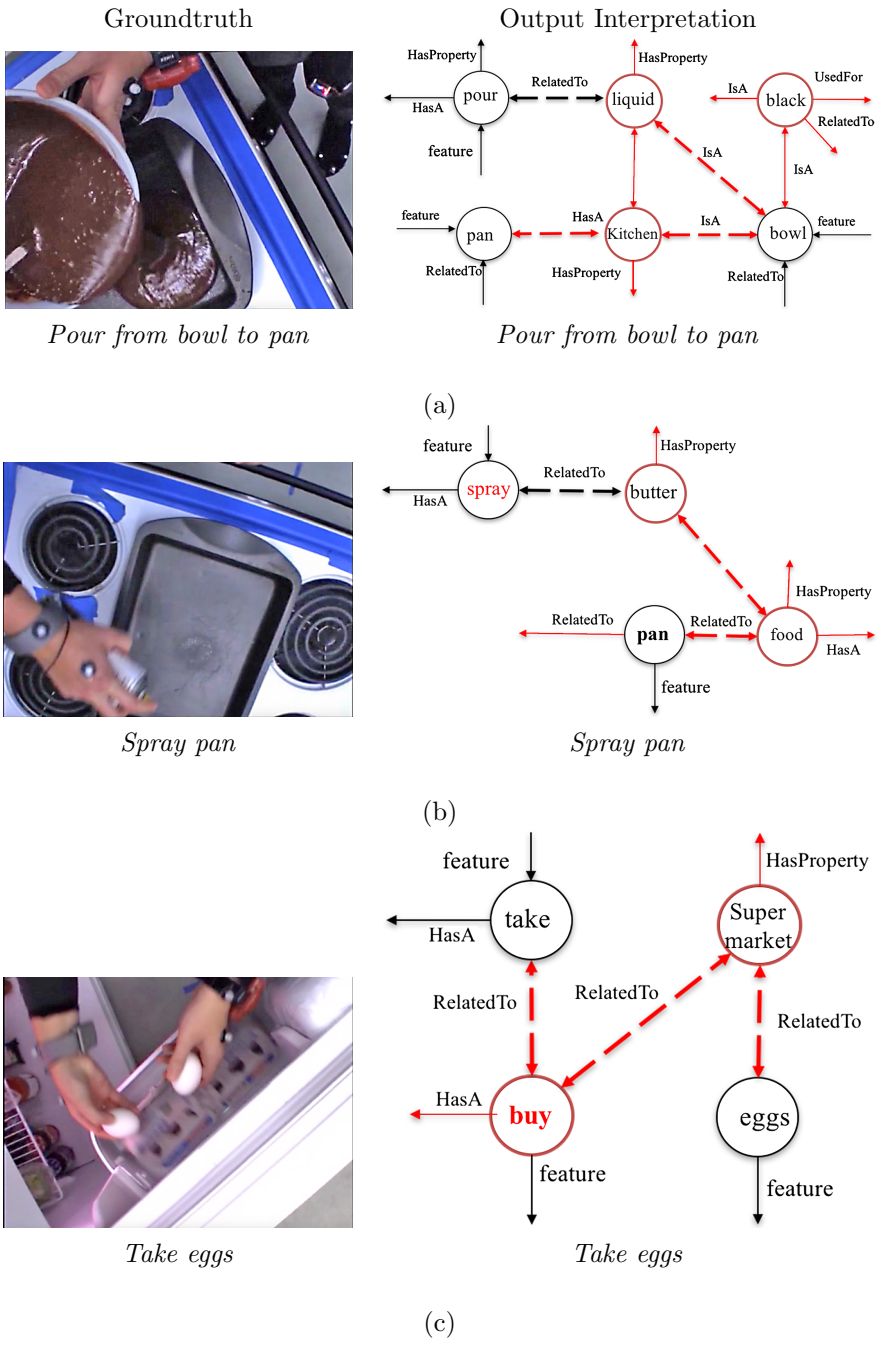


FIG. 11. Examples of instances where the model was able to handle multimodal content as well as variations from a viewpoint such as egocentric videos.

TABLE 1. Results on Charades dataset. ATF refers to Asynchronous Temporal Fields method. PT + ConceptNet semantics refers to the proposed approach. Trained semantics will indicate the use of training annotations to capture semantics between concepts. Note: All models use 2-stream features extracted from the videos as input, unless otherwise indicated.

Approach	**mAP**
LSTM	17.80%
RGB + ATF + trained semantics, no intent, no temporal	17.30%
RGB + ATF + trained semantics, no temporal, intent	17.40%
RGB + ATF + trained semantics, temporal, no intent	17.40%
ATF + trained semantics, intent, temporal	22.40%
PT + ConceptNet semantics, no intent, no temporal	**29.69**%
LSTM + PT + ConceptNet semantics, no intent, no temporal	**32.56**%

of temporal information as we were evaluating the benefits of using a knowledge base, rich with semantic information to reduce training requirements and as such temporal information, while useful, would not fit within the scope of the paper.

5.3.2. *Complex semantic relationships.* For evaluation of the performance on data with complex semantic relationships, we use the **Microsoft Video Description Corpus (MSVD)** dataset and report the BLEU score to compare against other comparable approaches. Our proposed approach has a comparable or better performance compared to the other approaches and is very close to the top performing approach as seen from Table 2. An important aspect to note is that the other state-of-the-art approaches take into account more complex features that are more descriptive of the concepts within the video; by contrast, we take into account only the mean-pooled CNN features across frames for generating candidate labels for all concepts. For example, the HRNE model [45] makes use of temporal characteristics in the video; the Temporal Attention model [65] leverages the frame-level representation from GoogleNet [58] as well as video-level

TABLE 2. Results on the Microsoft Video Description Corpus (MSVD) dataset. Top 10 means that we consider the best of 10 interpretations generated by the approach.

Approach	**BLEU Score**
Factor Graph Model [59]	13.68%
S2VT [61]	31.19%
S2VT + COCO & FLICKR [61]	33.29%
LSTM + Enc-Dec [65]	41.92%
HRNE [45]	43.6%
Our Approach (Best)	34.93%
Our Approach (Top 10)	42.98%

representation using a 3-D Convolutional Neural Network trained on handcrafted descriptors in addition to an attention model that provides dynamic attention to specific temporal regions of the video for generating descriptions.

5.3.3. *Weak features and uncertainty.* To evaluate the performance of the approach on data with more uncertainty and to demonstrate the use of knowledge, we evaluate on the Breakfast Actions dataset with handcrafted features (Section 3.1.1). The proposed approach outperforms both the performance of the HMM-based and the Context-Free Grammar models reported by Hilde [37] as well as other pattern theory models [20] as seen from Table 3. It is important to note that the performance of the proposed approach is remarkable considering that the model is neither trained specifically for the kitchen domain nor on the dataset itself other than for obtaining the starting grounded action and object labels. Other methods are restricted by the vocabulary of the training data to build their descriptions. For example, the Context-Free Grammar method makes use of temporal information such as states and transitions between states to build the final descriptions which is not the case with our approach.

TABLE 3. Results on Breakfast Action dataset. Top 10 means that we consider the best of 10 interpretations generated by the approach.

Approach	**Precision**
HMM [37]	14.90%
CFG + HMM [37]	31.8%
RNN + ECTC [33]	35.6%
RNN + ECTC (Cosine) [33]	36.7%
PT+weights (Top 10) [20]	33.40%
PT+training (Top 10) [20]	38.60%
Our Approach (Top 10)	**41.87**%

5.3.4. *Multimodal feature fusion.* To demonstrate the ability of our framework to handle multiple sources of input and different modalities, we use multi-modal features to leverage the auditory features present in the CMU activities dataset. We use audio feature representations such as bag-of-audio words and spectrograms features extracted from the audio. Table 4 compares our performance with the best performing feature set reported in [17] (PT+training), and its variation that replaces training with ConceptNet Similarity edge weights (PT+weights).

TABLE 4. Results on CMU Kitchen dataset with audio and video features used in conjunction.

Approach	**F-Score**
PT+training [17]	69.9%
PT+weights	64.6%
Our Approach	**72.7%**

TABLE 5. **Impact of training data imbalance**. We show the performances (F-Score) of our approach and prior pattern theory approaches (PT+training and PT+weights) for different activity categories that differ in the number of training samples.

Approach	Num. of samples			
	≤ 10	10 - 20	20 - 40	≥ 40
PT+training [17]	11.81%	17.36%	22.98%	35.47%
PT+weights	25.41%	36.57%	37.10%	34.16%
Our Approach	34.76%	40.07%	38.23%	38.35%

5.3.5. *Immunity to unbalance in training data.* Not all labels are equally represented in most training data. This is particularly acute as the number of labels increases. To demonstrate that our method is immune to this effect, we partitioned the activity classes labels into 4 different categories based on the amount of training data available. Table 5 shows the performance for our approach as compared to prior pattern theory approaches, PT+training and PT+weights. As expected, performance of PT+training relying on annotations [20] increases with an increase in training data, whereas our approach is stable.

6. Connection to related works. The pattern theory approach presented in this work draws heavily from Grenander's original formulation and stitches together many seminal concepts and contributions made by others. For instance, the concept of hierarchical compositional representation is similar to grammatical models of Zhu and Mumford [66] and was also expounded in depth by Geman, Potter, and Chi [22]. The data driven nature of the inference process related to grounded generators has parallels to data-driven MCMC algorithm used for image parsing by Tu, Chen, Yuille, and Zhu [60]. They used discriminative methods to propose the candidate objects, much like our machine learning approach to suggest grounded generators, and then used MCMC dynamics to construct the parse graph representations.

As mentioned earlier, the canonical representations of pattern theory subsumes many other symbolic representations as special cases, i.e., by restricting the global regularity, Σ, to special structures such as lattice, DAGS, tree structure, and so on. So, there are similarities of the adopted approach to others in the literature. There are many works that use probabilistic graphical models [16, 37] to explicitly model the semantic relationships in human activities [37, 38]. A variety of methods have been used, such as context-free grammars [35], probabilistic description frameworks [4], event probability sequences [14], Markov networks [42], Petri networks [5], and And-Or graphs [6, 63], to name a few. These approaches require labeled training data whose sizes increase non-linearly with different semantic combinations of possible actions and objects in the scene.

The main limitation of most of the current graphical modeling approaches is the implicit closed world assumption. The labels are limited to what is available in the annotation. Even for fixed object and action sets, most cannot handle the simultaneous occurrence of multiple events. Most ignore the possibility of object clutter and use all

of the detected objects to generate an interpretation. The structure of descriptions is pre-specified in the training process. For these methods, the need for more extensive pre-labeled video training data grows rapidly as the size of the model increases, to account for the full range of variabilities. For example, if all the training instances are "crack egg", then it will not be able to compose a description containing more elements such as "crack an egg in a bowl" or "crack an egg in a bowl using a spoon". For implicit models, such as discriminative models, the other objects, i.e., bowl and spoon, will be noise and an approach based on probabilistic models would need to train a new model with more random variables.

Unlike the popular graphical models that are based on defining probability distribution over propositional logic, Markov Logic Networks (MLNs) proposed by Richardson and Domingos [47] allow one to combine first order logic with probabilities. A set of instantiated first order logic is turned into a Markov Network with respect to a specific grounding and interpretation, along with weights, much like the energy model in our model. The vertices of the graph of the ground Markov Network are the ground atoms. However, the inference with these Markov Networks is extremely expensive as the size of the resulting Markov Network is exponential in size with respect to the number of possible groundings.

Recently, deep learning approaches such as RNNs and LSTMs [45, 48, 61, 65] have been found to be useful to model the semantic relationships among actions and objects based on transitions observed in training annotation phrases. While efforts have been made to use external text-based resources in addition to the training annotations, they are, arguably, restricted by the quality, quantity, and vocabulary of these annotations. As one moves up from recognizing actions to activities to events, the amount of labeled training data needed increases exponentially, however, the availability of such data goes down.

Our current work is a significant departure from other approaches, including our early use of Grenander's pattern theory [18–20] that rely on labeled training data to capture semantics about a domain, such as sentences and phrases describing the video segment. The use of a general knowledge database such as ConceptNet [39] as priors alleviates the need for data annotations. In fact, the only training we require is the one needed for detecting objects and simple actions. Other works that have the similar philosophy are those using ontology to detect and understand events [18, 36, 44]. Models such as [10, 13, 62] have also explored the use of contextual knowledge by leveraging spatial and situational comprehension.

7. Conclusion. The contribution of this paper is three-fold: (1) a deep semantic reasoning framework for structured representation and semantic interpretation of video activities extending beyond simple pairwise relationships, (2) the use of a global source of knowledge to reduce training requirements by negating the need for large amounts of annotations for capturing semantic relationships, and (3) we are, to the best of our knowledge, the first to introduce the notion of open world activity descriptions using common sense knowledge. We demonstrated how pattern theory can be used to capture the semantics in ConceptNet and infer rich interpretations that can be the basis for

the generating of sentences or even visual questions and answers. The inference process allows for multiple concept labels for each video event to overcome errors in classification. Extensive experiments demonstrate the applicability of the approach to different domains and its highly competitive performance.

Unlike other works, we do not have the need to learn action and object combination priors from such annotations. The applicability of our approach is not restricted to the training domain. The use of a general human knowledge database such as ConceptNet as the source of prior knowledge alleviates the need for data annotations. In fact, the only training we require in the proposed approach is the one required for detecting basic concepts such as actions and objects. It allows us to leverage the knowledge gleaned from external sources and is not restricted to a particular domain and/or dataset.

In this work, we restricted ourselves to short video snippets to demonstrate how common sense prior knowledge can be incorporated in a pattern theory framework. However, our approach can be easily extended to the analysis of long videos containing sequences of video activities. As we have shown earlier in our prior work [53], we can handle activity sequences by simply extending the generator definition to have another bond type that will link across time.

Acknowledgment. This work includes data from ConceptNet 5, which was compiled by the Common Sense Computing Initiative. ConceptNet 5 is freely available under the Creative Commons Attribution-ShareAlike license (CC BY SA 4.0) from http://conceptnet.io. The included data was created by contributors to Common Sense Computing projects, contributors to Wikimedia projects, Games with a Purpose, Princeton University's WordNet, DBPedia, OpenCyc, and Umbel.

References

[1] Sathyanarayanan N. Aakur, Fillipe DM de Souza, and Sudeep Sarkar, *Towards a knowledge-based approach for generating video descriptions*, Conference on Computer and Robot Vision (CRV), Springer, 2017.

[2] Sathyanarayanan N. Aakur, Fillipe DM de Souza, and Sudeep Sarkar, *An inherently explainable model for video activity interpretation*, Workshops of the AAAI Conference on Artificial Intelligence, AAAI, 2018.

[3] Somak Aditya, Yezhou Yang, Chitta Baral, Cornelia Fermuller, and Yiannis Aloimonos, *From images to sentences through scene description graphs using commonsense reasoning and knowledge*, arXiv preprint arXiv:1511.03292 (2015).

[4] Massimiliano Albanese, Rama Chellappa, Naresh Cuntoor, Vincenzo Moscato, Antonio Picariello, VS Subrahmanian, and Octavian Udrea, *Pads: A probabilistic activity detection framework for video data*, IEEE Transactions on Pattern Analysis and Machine Intelligence **32** (2010), no. 12, 2246–2261.

[5] Massimiliano Albanese, Rama Chellappa, Vincenzo Moscato, Antonio Picariello, VS Subrahmanian, Pavan Turaga, and Octavian Udrea, *A constrained probabilistic petri net framework for human activity detection in video*, IEEE Transactions on Multimedia **10** (2008), no. 6, 982–996.

[6] Mohamed R Amer, Sinisa Todorovic, Alan Fern, and Song-Chun Zhu, *Monte carlo tree search for scheduling activity recognition*, IEEE International Conference on Computer Vision (ICCV), 2013, pp. 1353–1360.

[7] Y. Amit, U. Grenander, and M. Piccioni, *Structural image restoration through deformable templates*, J. American Statistical Association (1991).

[8] Y. Amit and A. Kong, *Graphical templates for model registration*, Pattern Analysis and Machine Intelligence, IEEE Transactions on **18** (1996), no. 3, 225–236.

[9] E. Bienenstock, S. Geman, and D. Potter, *Compositionality, mdl priors, and object recognition*, Advances in neural information processing systems (1997), 838–844.
[10] Guoray Cai, *Contextualization of geospatial database semantics for human–gis interaction*, Geoinformatica **11** (2007), no. 2, 217–237.
[11] Lo-Bin Chang, Ya Jin, Wei Zhang, Eran Borenstein, and Stuart Geman, *Context, computation, and optimal ROC performance in hierarchical models*, Int. J. Comput. Vis. **93** (2011), no. 2, 117–140, DOI 10.1007/s11263-010-0391-1. MR2783693
[12] Rizwan Chaudhry, Avinash Ravichandran, Gregory Hager, and René Vidal, *Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2009, pp. 1932–1939.
[13] Qiang Chen, Zheng Song, Jian Dong, Zhongyang Huang, Yang Hua, and Shuicheng Yan, *Contextualizing object detection and classification*, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **37** (2015), no. 1, 13–27.
[14] Naresh P. Cuntoor, B. Yegnanarayana, and Rama Chellappa, *Activity modeling using event probability sequences*, IEEE Trans. Image Process. **17** (2008), no. 4, 594–607, DOI 10.1109/TIP.2008.916991. MR2512463
[15] Navneet Dalal and Bill Triggs, *Histograms of oriented gradients for human detection*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, IEEE, 2005, pp. 886–893.
[16] Pradipto Das, Chenliang Xu, Richard F Doell, and Jason J Corso, *A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 2634–2641.
[17] Fillipe DM de Souza, Sudeep Sarkar, and Guillermo Cámara-Chávez, *Building semantic understanding beyond deep learning from sound and vision*, 23rd International Conference on Pattern Recognition (ICPR), IEEE, 2016, pp. 2097–2102.
[18] Fillipe DM De Souza, Sudeep Sarkar, Anuj Srivastava, and Jingyong Su, *Pattern theory-based interpretation of activities*, 22nd International Conference on Pattern Recognition (ICPR), IEEE, 2014, pp. 106–111.
[19] Fillipe DM de Souza, Sudeep Sarkar, Anuj Srivastava, and Jingyong Su, *Pattern theory for representation and inference of semantic structures in videos*, Pattern Recognition Letters **72** (2016), 41–51.
[20] Fillipe DM de Souza, Sudeep Sarkar, Anuj Srivastava, and Jingyong Su, *Spatially coherent interpretations of videos using pattern theory*, International Journal of Computer Vision **121** (2017), no. 1, 5–25.
[21] S. Geman and M. Johnson, *Dynamic programming for parsing and estimation of stochastic unification-based grammars*, Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2002, pp. 279–286.
[22] Stuart Geman, Daniel F. Potter, and Zhiyi Chi, *Composition systems*, Quart. Appl. Math. **60** (2002), no. 4, 707–736, DOI 10.1090/qam/1939008. MR1939008
[23] U. Grenander, Y. Chow, and D. M. Keenan, *Hands: A pattern-theoretic study of biological shapes*, Research Notes in Neural Computing, vol. 2, Springer-Verlag, New York, 1991. MR1084371
[24] Ulf Grenander and Michael I. Miller, *Representations of knowledge in complex systems*, with discussion and a reply by the authors, J. Roy. Statist. Soc. Ser. B **56** (1994), no. 4, 549–603. MR1293234
[25] Ulf Grenander and Michael I. Miller, *Computational anatomy: an emerging discipline*, Current and future challenges in the applications of mathematics (Providence, RI, 1997), Quart. Appl. Math. **56** (1998), no. 4, 617–694, DOI 10.1090/qam/1668732. MR1668732
[26] U. Grenander, A. Srivastava, and S. Saini, *A pattern-theoretic characterization of biological growth*, IEEE Transactions on Medical Imaging **26** (2007), no. 5, 648–659.
[27] Ulf Grenander, *General pattern theory: A mathematical study of regular structures*, Oxford Science Publications, Oxford Mathematical Monographs, The Clarendon Press, Oxford University Press, New York, 1993. MR1270904
[28] Ulf Grenander, *Elements of pattern theory*, JHU Press, 1996.
[29] Ulf Grenander, *A calculus of ideas: a mathematical study of human thought*, World Scientific, 2012.
[30] Ulf Grenander and Michael I. Miller, *Pattern theory: from representation to inference*, Oxford University Press, Oxford, 2007. MR2285439
[31] John J Gumperz, *Contextualization and understanding*, Rethinking context: Language as an interactive phenomenon **11** (1992), 229–252.

[32] F. Han and S.C. Zhu, *Bottom-up/top-down image parsing with attribute grammar*, Pattern Analysis and Machine Intelligence, IEEE Transactions on **31** (2009), no. 1, 59–73.
[33] De-An Huang, Li Fei-Fei, and Juan Carlos Niebles, *Connectionist temporal modeling for weakly supervised action labeling*, European Conference on Computer Vision, Springer, 2016, pp. 137–153.
[34] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei, *Image retrieval using scene graphs*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3668–3678.
[35] Seong-Wook Joo and Rama Chellappa, *Recognition of multi-object events using attribute grammars*, International Conference on Image Processing (ICIP), IEEE, 2006, pp. 2897–2900.
[36] Rama Kovvuri, Ram Nevatia, and Cees GM Snoek, *Segment-based models for event detection and recounting*, Pattern Recognition (ICPR), 2016 23rd International Conference on, IEEE, 2016, pp. 3868–3873.
[37] Hilde Kuehne, Ali Arslan, and Thomas Serre, *The language of actions: Recovering the syntax and semantics of goal-directed human activities*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 780–787.
[38] Tian Lan, Leonid Sigal, and Greg Mori, *Social roles in hierarchical models for human activity recognition*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 1354–1361.
[39] Hugo Liu and Push Singh, *ConceptNet—a practical commonsense reasoning tool-kit*, BT technology journal **22** (2004), no. 4, 211–226.
[40] M. I. Miller, A. Srivastava, and U. Grenander, *Conditional-expectation estimation via jump-diffusion processes in multiple target tracking/recognition*, IEEE Transactions on Signal Processing **43** (1995), no. 11, 2678–2690.
[41] M. I. Miller, G. E. Christensen, Y. Amit, and U. Grenander, *Mathematical textbook of deformable neuroanatomies*, Proceedings of the National Academy of Science **90** (1993), no. 24.
[42] Vlad I Morariu and Larry S Davis, *Multi-agent event recognition in structured scenarios*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2011, pp. 3289–3296.
[43] David Mumford, *Pattern theory: a unifying perspective*, First European Congress of Mathematics, Vol. I (Paris, 1992), Progr. Math., vol. 119, Birkhäuser, Basel, 1994, pp. 187–224. MR1341824
[44] Ram Nevatia, Tao Zhao, and Somboon Hongeng, *Hierarchical language-based representation of events in video streams*, Computer Vision and Pattern Recognition Workshop, 2003. CVPRW'03. Conference on, vol. 4, IEEE, 2003, pp. 39–39.
[45] Pingbo Pan, Zhongwen Xu, Yi Yang, Fei Wu, and Yueting Zhuang, *Hierarchical recurrent neural encoder for video representation with application to captioning*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.
[46] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, *Bleu: a method for automatic evaluation of machine translation*, Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics, 2002, pp. 311–318.
[47] Matthew Richardson and Pedro Domingos, *Markov logic networks*, Machine learning **62** (2006), no. 1-2, 107–136.
[48] Marcus Rohrbach, Wei Qiu, Ivan Titov, Stefan Thater, Manfred Pinkal, and Bernt Schiele, *Translating video content to natural language descriptions*, IEEE International Conference on Computer Vision (ICCV), 2013, pp. 433–440.
[49] Olga Russakovsky, Jia Deng, Hao Su, et al., *ImageNet large scale visual recognition challenge*, Int. J. Comput. Vis. **115** (2015), no. 3, 211–252, DOI 10.1007/s11263-015-0816-y. MR3422482
[50] Gunnar A Sigurdsson, Santosh Divvala, Ali Farhadi, and Abhinav Gupta, *Asynchronous temporal fields for action recognition*, arXiv preprint arXiv:1612.06371 (2016).
[51] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta, *Hollywood in homes: Crowdsourcing data collection for activity understanding*, European Conference on Computer Vision, Springer, 2016, pp. 510–526.
[52] Karen Simonyan and Andrew Zisserman, *Two-stream convolutional networks for action recognition in videos*, NIPS, 2014, pp. 568–576.
[53] Fillipe Souza, Sudeep Sarkar, Anuj Srivastava, and Jingyong Su, *Temporally coherent interpretations for long videos using pattern theory*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1229–1237.
[54] Robert Speer and Catherine Havasi, *Representing general relational knowledge in ConceptNet 5*, LREC, 2012, pp. 3679–3686.

[55] Robert Speer and Catherine Havasi, *ConceptNet 5: A large semantic network for relational knowledge*, The People's Web Meets NLP, Springer, 2013, pp. 161–176.
[56] E. H. Spriggs, F. De La Torre, and M. Hebert, *Temporal segmentation and activity classification from first-person sensing*, IEEE Workshops on Computer Vision and Pattern Recognition (CVPRW), June 2009, pp. 17–24.
[57] A. Srivastava, M. I. Miller, and U. Grenander, *Multiple target direction of arrival tracking*, IEEE Transactions on Signal Processing **43** (1995), no. 5, 1282–85.
[58] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, *Going deeper with convolutions*, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9.
[59] Jesse Thomason, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Raymond J Mooney, *Integrating language and vision to generate natural language descriptions of videos in the wild.*, International Conference on Computational Linguistics (COLING), vol. 2, 2014, p. 9.
[60] Zhuowen Tu, Xiangrong Chen, Alan L Yuille, and Song-Chun Zhu, *Image parsing: Unifying segmentation, detection, and recognition*, International Journal of computer vision **63** (2005), no. 2, 113–140.
[61] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko, *Translating videos to natural language using deep recurrent neural networks*, arXiv preprint arXiv:1412.4729 (2014).
[62] Yi Wang, David M Krum, Enylton M Coelho, and Doug A Bowman, *Contextualized videos: Combining videos with environment models to support situational understanding*, IEEE Transactions on Visualization and Computer Graphics **13** (2007), no. 6, 1568–1575.
[63] Ping Wei, Yibiao Zhao, Nanning Zheng, and Song-Chun Zhu, *Modeling 4d human-object interactions for event and object recognition*, IEEE International Conference on Computer Vision, IEEE, 2013, pp. 3272–3279.
[64] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei, *Scene graph generation by iterative message passing*, arXiv preprint arXiv:1701.02426 (2017).
[65] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville, *Describing videos by exploiting temporal structure*, IEEE International Conference on Computer Vision (ICCV), 2015, pp. 4507–4515.
[66] S.C. Zhu and D. Mumford, *A stochastic grammar of images*, Foundations and Trends® in Computer Graphics and Vision **2** (2006), no. 4, 259–362.