Robotics: Science and Systems 2021
Held Virtually, July 12–16, 2021

Learning Instance-Level N-Ary Semantic Knowledge At Scale For Robots Operating in Everyday Environments

Weiyu Liu, Dhruva Bansal, Angel Daruna, and Sonia Chernova
Institute for Robotics and Intelligent Machines
Georgia Institute of Technology, Atlanta, Georgia
Email: {wliu88, dbansal36, adaruna3, chernova}@gatech.edu

Abstract—**Robots operating in everyday environments need to effectively perceive, model, and infer semantic properties of objects. Existing knowledge reasoning frameworks only model binary relations between an object's class label and its semantic properties, unable to collectively reason about object properties detected by different perception algorithms and grounded in diverse sensory modalities. We bridge the gap between multimodal perception and knowledge reasoning by introducing an n-ary representation that models complex, inter-related object properties. To tackle the problem of collecting n-ary semantic knowledge at scale, we propose a transformer neural network that directly generalizes knowledge from observations of object instances. The learned model can reason at different levels of abstraction, effectively predicting unknown properties of objects in different environmental contexts given different amounts of observed information. We quantitatively validate our approach against five prior methods on LINK, a unique dataset we contribute that contains 1457 situated object instances with 15 multimodal properties types and 200 total properties. Compared to the top-performing baseline, a Markov Logic Network, our model obtains a 10% improvement in predicting unknown properties of novel object instances while reducing training and inference time by 150 times. Additionally, we apply our work to a mobile manipulation robot, demonstrating its ability to leverage n-ary reasoning to retrieve objects and actively detect object properties. The code and data are available at https://github.com/wliu88/LINK.**

I. Introduction

Robust operation in everyday human environments requires robots to effectively model a wide range of objects and to predict object locations, properties, and uses. Semantic task and object knowledge serves as a valuable abstraction in this context. Perceiving and understanding semantic properties of objects (e.g., a *cup* is *ceramic*, *empty*, located *in kitchen*, and used for *drinking*) aids robots in performing many real-world tasks, such as inferring missing information in human instructions [52, 12], efficiently searching for objects in homes [69, 68], and manipulating objects based on their affordances and states [3, 43, 31].

Prior work has encoded semantic knowledge primarily as pairwise relations between an object's *class* label and its semantic *properties* (e.g., the *cup* is *wet*, the *cup* is *in cabinet*) [16, 13, 71, 60, 54] (Figure 1 left). These semantic properties can come from a variety of perception methods, such as the use of vision to predict visual attributes [23, 49] and affordances [19, 15], haptic data to identify object materials [33] and surface textures [14], as well as exploratory actions to detect object states [61]. However, pairwise encoding of semantic data fails to take full advantage of such multimodal observations because it ignores the complex relational structure between various object properties. For example, observing that a *cup* is *wet* does not help the robot infer that the *cup* more likely should be placed *in sink* than *in cabinet*.

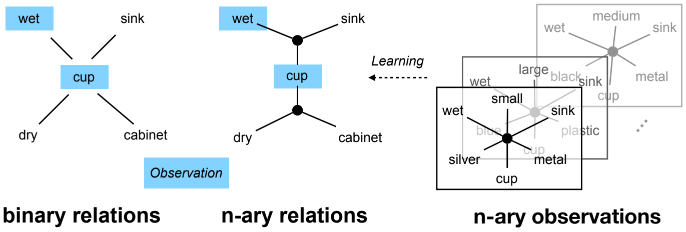


Fig. 1. N-ary relations enable robots to more effectively model complex, inter-related object properties than binary relations. In our framework, we learn generalizable n-ary relations from object instances represented as n-ary observations.

The objective of our work is to enable robots to collectively reason about object properties that can be grounded in different modalities and detected by separate perception algorithms. Specifically, we situate our work in the task of predicting semantic properties of objects based on partial observations. We introduce a novel semantic reasoning framework that uses n-ary relations to model complex, inter-related object properties. In addition to modeling relations between object properties, our framework enables the ability to reason at different levels of abstraction (Figure 1 middle). For example, a robot searching for a *cup*, with no additional information, is able to perform class-level inference to identify both the *cabinet* and *sink* as likely locations. Given additional information, such as *wet*, n-ary relations enable more refined reasoning and the ability to detect that wet cups are more commonly found *in sink* rather than *in cabinet*.

A key challenge presented by n-ary representations is the collection of semantically meaningful n-ary relations, which require various object properties to be conditioned on each other. Unlike binary relations which can be created by experts or crowdsourced at scale [28, 42, 47, 39], n-ary relations are difficult to construct manually. In this work, we obtain n-ary observations, each representing a set of identified semantic properties of an object instance within a particular environmental context (e.g., a *small silver metal cup* that is *wet* and *in sink*), from which we then learn a model capturing generalizable n-ary relations (Figure 1 right). Since the n-ary relations are learned from object instances, they also encode knowledge at the instance-level. To mine generalizable patterns from n-ary observations, we introduce a transformer-based neural network, inspired by recent advances in contextualized language models [18, 64]. The model is trained to reconstruct hidden properties of object instances. With this self-supervised training objective, our model learns to generalize n-ary relations, which help predict unobserved properties of novel object instances.

In summary, our work contributes:

- an n-ary instance-level representation of objects, which enables modeling n-ary relations between multimodal object properties and variance between object instances,
- a scalable transformer-based neural network which learns semantic knowledge about objects from data and is capable of performing inference at different levels of abstraction,
- a dataset, which we call LINK, consisting of 1457 object instances associated with 15 types of 200 multimodal properties, the richest situated object dataset to date.

We quantitatively validate our approach against five prior methods on the above dataset and demonstrate that our representation and reasoning method leads to significant improvement in predicting unknown properties of novel object instances over prior state of the art while significantly reducing computation time. Additionally, we apply our work to a mobile manipulation robot. We demonstrate that the explicit representation of n-ary knowledge allows the robot to locate objects based on complex human commands. We also show that the learned relations can help the robot infer properties based on observations from multimodal sensory data.

II. Related Work

Our work is related to the following prior efforts.

A. Semantic Reasoning in Robotics

Many ontologies and knowledge graphs have been used across AI and robotics to encode general knowledge about objects (e.g., locations, properties, uses, and class hierarchies) [41, 54, 60, 38, 63]. In robotics, a key challenge for semantic reasoning is generalization to previously unseen scenes or environments. Bayesian logic networks have been used to cope with noise and non-deterministic data from different data sources [13]. More recently, knowledge graph (KG) embedding models were introduced as scalable frameworks to model object knowledge encoded in multi-relational KGs [16, 4]. Although the above techniques effectively model objects, they only support reasoning about binary class-level facts, therefore lacking the discriminative features needed to model object semantics in realistic environments.

Other frameworks take a learning approach to modeling object semantics. Methods for learning relations between objects, between object properties, and between objects and their environments have shown to be beneficial for detecting objects on table tops [36, 27, 51], finding hidden objects in shelves [48], predicting object affordances [71], and semantic grasping [3, 43]. However, most methods leverage probabilistic logic models to learn these relations, which have scalability issues that limit them from modeling inter-connected relations in larger domains [51, 48, 71, 3]. In contrast, our proposed framework learns n-ary relations between 15 property types and 200 properties, the richest representation to date.

Our approach is also related to methods for modeling objects from sensory data. In computer vision, object attributes are extracted from images [23, 22, 57]. Recent techniques in visual question answering [49] and language grounding [55, 32] allow robots to answer questions about objects and describe objects with natural language. Haptic [45, 40] and auditory data [20, 25] have also helped robots interpret salient features of objects beyond vision. Interactive perception can further leverage a robot's exploratory actions to reveal sensory signals that are otherwise not observable [9, 56, 14, 61, 2, 8, 59]. We consider our approach complimentary to the above, as our framework can leverage the rich semantic information extracted from these methods to infer additional unknown object properties. Our work shares the same goal as a recent work that builds robot-centric object knowledge from multimodal sensory data [62].

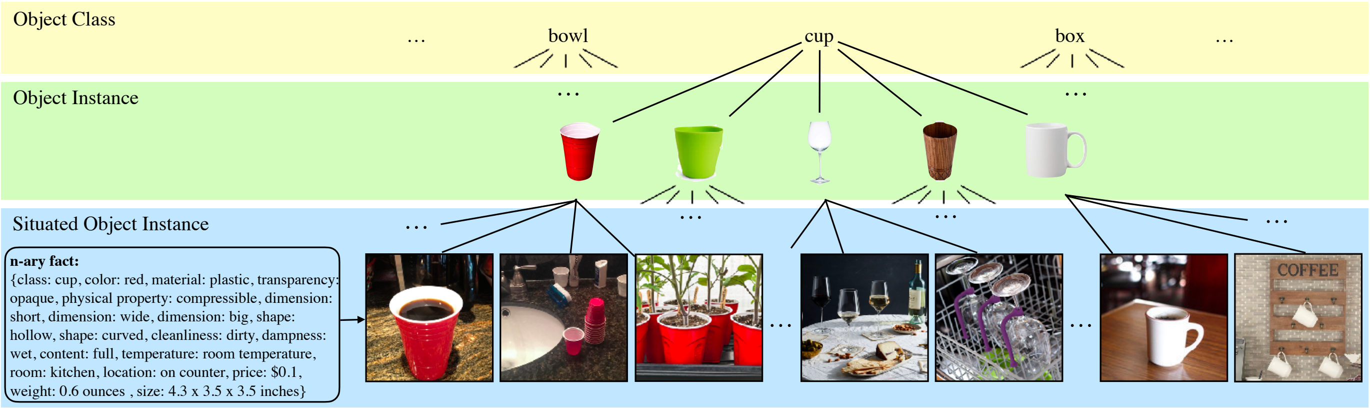
B. Modeling N-Ary Facts

Our neural network model is closely related to methods developed in the knowledge graph community. Many relational machine learning techniques, including most recent transformer models [65, 10], have been developed for modeling KGs and in particular predicting missing links in KGs [50]. These techniques treat a KG as set of triples/binary facts, where each triple (h, r, t) links two entities h and t with a relation r (e.g., *(Marie Curie, educated at, University of Paris)*). Despite the wide use of triple representation, many facts in KGs are hyper-relational. Each hyper-relational fact has a base triple (h, r, t) and additional key-value (relation-entity) pairs (k, v) (e.g., $\{$*(academic major, physical), (academic degree, Master of Science)*$\}$). A line of work converts hyper-relational facts to n-ary meta-relations $r(e_1, ..., e_n)$ and leverages translational distance embedding [66, 70], spatio-translational embedding [1], tensor factorization [44] for modeling. Other approaches directly learn hyper-relational facts in their original form using various techniques, including convolutional neural networks, graph neural networks, and transformer models [53, 24]. A representation of hyper-relational facts more closely related to our work is used by [26]. This approach unifies n-ary representation by converting the base triple to key-value pairs; it uses convolutional neural network for feature extraction and then models relatedness of role-value pairs with a fully connected network. In our work, we model facts with much higher arities than existing work in the KG

TABLE I

COMPARING AVAILABLE DATASETS FOR LEARNING OBJECT SEMANTICS.

Dataset	Application	# Object Classes	# Object Instances	# Properties	# Property Types	Situated	Complete Annotation
Shop-VRB [49]	Vision & Language	20	66	99	6		✓
GoLD [32]	Language Grounding	47	207	/	/		
Thomason 2018 [61]	Interactive Perception	4	32	81	6	✓	✓
Zhu 2014 [71]	Knowledge Base	40	4000	97	4		✓
Paolo 2019 [3]	Semantic Grasping	8	30	44	5		✓
AI2Thor [35]	Simulation	/	125	28	6		✓
Ours	Semantic Reasoning	11	98	200	15	✓	✓



n-ary fact:
{class: cup, color: red, material: plastic, transparency: opaque, physical property: compressible, dimension: short, dimension: wide, dimension: big, shape: hollow, shape: curved, cleanliness: dirty, dampness: wet, content: full, temperature: room temperature, room: kitchen, location: on counter, price: $0.1, weight: 0.6 ounces , size: 4.3 x 3.5 x 3.5 inches}

Fig. 2. An example of the collected data showing various cups and diverse environmental contexts each of these instances can be found in. Pictures at the situated object instance level are for illustration but correspond to descriptions of the contexts in our dataset. Each situated object instance has a corresponding fully annotated n-ary observation (bottom left).

community and directly reason about n-ary relations between role-value pairs using the transformer model.

III. Problem Definition

Given a set of observed/known object properties, we aim to predict an unobserved/unknown property of a novel situated object instance using semantic knowledge learned from data. We define a *situated object instance* as a particular specimen of a given object class within a particular environmental context (e.g., the full red Solo cup on the kitchen counter). The object's semantic representation encodes properties grounded in different modalities, and includes both immutable properties (e.g., class, material, shape, and hardness) and mutable properties (e.g., location, clealiness, fullness).

We use the n-ary representation to model all object data. Each n-ary relation is defined by a set of role-value pairs $\{r_i : v_i\}$, where $r_i \in \mathcal{R}$ is the role set, $v_i \in \mathcal{V}$ is the value set, and $i = 1, ..., n$. The value n represents the arity of the relation. In the context of modeling object semantics, each role corresponds to a property type and each value corresponds to a property value. In this representation, our task can be formally written as $\{r_1 : v_1, ..., r_{n-1} : v_{n-1}, r_n :?\}$, where $n-1$ is the number of known properties, and r_n is the type of the property being queried. The number of known properties n determines the level of abstraction for the query. A smaller n queries more abstract semantic knowledge (e.g., $\{$*class: cup, material: ?*$\}$) and a larger n queries more specific semantic knowledge (e.g., $\{$*class: cup, transparency: opaque, physical property: hard, color: brown, material: ?*$\}$).

IV. LINK Dataset

In this section, we present the content and features of the **LINK** dataset for **L**earning **I**nstance-level **N**-ary **K**nowledge. Our dataset contains 1457 fully annotated situated object instances. In Table I, we compare the content of our dataset to a representative set of data sources from the computer vision, natural language processing, and robotics communities; as can be seen, our dataset has the most diverse set of property types and property values, leading to much richer and more realistic object representations. Properties in our dataset are inherently multimodal, which help bridging robots' perception and reasoning. In addition to visual attributes, we intentionally model properties that are hard to extract from visual data (e.g., dampness and temperature). Our dataset represents variance between object instances by having on average of nine objects per class. Objects in different situations are captured by mutable properties such as location, cleanliness, temperature, and dampness. Furthermore, our dataset provides complete and logically coherent annotations (truth values) of all properties for each situated object instance. Figure 2 illustrates the hierarchy of objects in our dataset, which facilitates the learning of generalizable n-ary relations between object properties at different levels of abstraction.

A. Objects and Properties

Our dataset contains 98 instances of everyday household objects organized into 11 object classes. For each object class, we selected objects diverse in sizes, geometries, materials, visual appearances, and affordances from the Amazon product

TABLE II
OBJECT CLASSES AND PROPERTIES IN OUR DATASET.

Type (# Value)	Values
class (11)	bottle, bowl, box, brush, can, cup, fork, ladle, pan, spatula, sponge
material (8)	ceramic, foam, glass, metal, paper, plastic, porcelain, wood
transparency (3)	opaque, translucent, transparent
dimension (10)	big, deep, long, narrow, shallow, short, small, thick, thin, wide
physical property (6)	absorbent, compressible, elastic, fragile, hard, soft
shape (9)	angular, blunt, curved, flat, forked, hollow, irregular, sharp, straight
*temperature (3)	cold, hot, room temperature
*fullness (3)	empty, full, half
*dampness (3)	damp, dry, wet
*cleanliness (3)	clean, dirty, normal
price (3)	cheap, expensive, medium
weight (3)	heavy, light, medium
size (3)	large, medium, small
*room (11)	balcony, bathroom, bedroom, child's room, closet, dining room, garage, kitchen, laundry, living room, study
color (15)	black, blue, bronze, brown, clear, colorful, gold, green, orange, pink, purple, red, silver, white, yellow
*location (117)	in bag, in basket, in bathtub, in bin, in box, in bucket, in cabinet, in cooler, on bathtub, on bed, on bench, on bookshelf, ...

website. We created the initial set of 83 properties (the additional 117 location properties are crowdsourced) from adjectives that people use for describing objects [46]. We then followed GermaNet[1] [29] to categorize these properties into 15 distinct types based on their semantic meanings. Table II shows the property values and types in our dataset (mutable properties are labeled with asterisk).

B. Collection of N-ary Labels

Given 98 object instances and 15 property types, our next step was to collect situated object instances where each object is described by a semantically meaningful combination of properties. We used Amazon Mechanical Turk (AMT) to crowdsource property combinations. The novelty in our crowdsourcing process is that we asked AMT workers to *imagine* objects situated in different environment contexts. Compared to established approaches to collect semantic knowledge, such as asking workers to annotate properties for objects in images and prompting workers to answer commonsense questions about objects, our method is more effective at eliciting multimodal and instance-level knowledge.

More specifically, after we extracted pictures of each object, as well as details of its material, weight, dimension, and price from the Amazon product web page, we conducted a three-stage crowdourcing process. First, for all 98 object instances, we showed pictures of the object to AMT workers and asked them to list the object's immutable properties. Second, we presented AMT workers with an object and a room, and had them imagine and describe three situations in which that object-room combination could be encountered, including details of the location of the object, the associated daily activity, and the object state (e.g., a wet cup on the bathroom counter used for rinsing after brushing teeth). Third, we presented a new set of AMT workers with the above collected situated object descriptions, and had them label mutable properties (e.g., wet, empty, clean) for the associated object. To ensure the quality of the crowdsourced data, we used 3 annotators for each question and filtered workers based on gold standard questions. We manually verified descriptions of situations from stage 2.

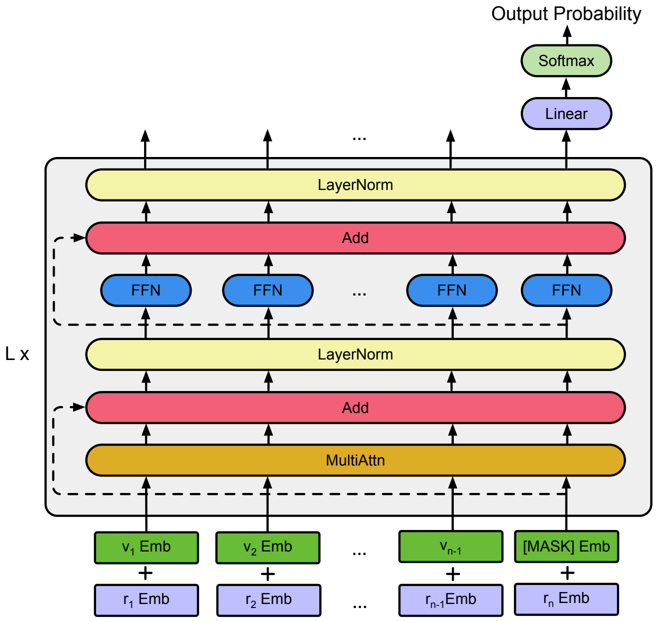


Fig. 3. Our model architecture includes embedding layers, a transformer encoder, and a feed-forward layer for predicting probabilities of properties.

V. APPROACH

Given $n-1$ properties, we aim to predict the the n^{th} property of type r_n, i.e., $\{r_1 : v_1, ..., r_{n-1} : v_{n-1}, r_n : ?\}$, where $n-1$ is the number of observed properties. We develop a transformer-based neural network based on following design goals: learning interactions between typed properties (i.e., role-value pairs), accommodating arbitrary order of properties, supporting inference at different levels of abstraction by accepting arbitrary number of observed properties, representing uncertainties in semantic knowledge, and being scalable.

As shown in Figure 3, our model takes $\{r_1 : v_1, ..., r_{n-1} : v_{n-1}, r_n :$ [MASK]$\}$ as input, where [MASK] is a special token for the query property. The masked input is then fed into the transformer encoder [64], which builds a contextualized representation of the input. Finally, The encoding at the n^{th} position is used to predict the query property via a feedforward layer and a sigmoid function. Now we describe each component of the model in detail and discuss how they help to satisfy the design goals and learn n-ary relations between object properties.

A. Input Encoder

The input encoder uses learned embeddings to convert role-value pairs in the input to vectors of dimension d_{model}. Specifically, for each pair, we construct its representation as

$$h_i^0 = x_i^{\text{value}} + x_i^{\text{role}} \tag{1}$$

where x_i^{value} is the embedding for the i^{th} value and x_i^{role} is the embedding for the i^{th} role. At the query position, the value embedding of the [MASK] token indicates that this property

[1] GermaNet, the German version of the English lexical database Wordnet [47], provides hierarchical structures for adjectives.

is in query. We maintain the role embedding of the query property, allowing the model to condition its reasoning on the type of the query property. Different from existing transformer-based models [18, 10, 65], we do not use positional embeddings to indicate the position of each role-value pair in the n-ary query since, unlike natural language sentences or KG triples, there is no particular order for object properties. As the latter components of the model are permutation invariant to the order of input data, removing the positional embeddings also allows our model to efficiently learn from object properties represented in n-ary observations.

B. Transformer Encoder

The transformer encoder takes the embedded input $\{h_1^0, ..., h_n^0\}$ and builds a contextualized representation $\{h_1^L, ..., h_n^L\}$ where L is the number of transformer layers in the transformer encoder. Each transformer layer applies the following transformation to the input:

$$\hat{g}^l = \text{MultiAttn}(h^{l-1}, h^{l-1}, h^{l-1}) \tag{2}$$

$$g^l = \text{LayerNorm}(\hat{g}^l + h^{l-1}) \tag{3}$$

$$\hat{h}^l = \text{FFN}(g^l) \tag{4}$$

$$h^l = \text{LayerNorm}(\hat{h}^l + g^l) \tag{5}$$

where MultiAttn is a multi-head self-attention mechanism, which we discuss in more depth below. FFN is a fully-connected feedforward network, which is applied to each position of the input separately and identically. The FFN consists of two linear fully-connected layers with a ReLU activation in between. Residual connections [30] are applied both after MultiAttn and FFN, which are followed by layer normalizations [5].

C. Multi-Head Attention

The core component of the transformer encoder is the multi-head attention mechanism, which builds on the scaled dot-product attention function. An attention function takes in a query and a set of key-value pairs. The output is computed as a weighted sum of the values, where the weight assigned to each value is based on the compatibility of the query with the corresponding key. The scaled dot-product attention performs the attention computation efficiently by computing on a set of queries simultaneous with matrix multiplication. The queries, keys, and values are stacked together into matrix $Q \in \mathbb{R}^{n_{\text{query}} \times d_{\text{model}}}$, $K \in \mathbb{R}^{n_{\text{key}} \times d_{\text{model}}}$, and $V \in \mathbb{R}^{n_{\text{value}} \times d_{\text{model}}}$, where n_{query} is the number of queries. Formally, the scaled dot-product attention is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{6}$$

where d_k is the dimension of queries and keys, and serves as a scaling factor for stabilizing gradients.

Instead of computing the attention function once, the multi-head attention has H heads, where each head performs a scaled dot-product attention. This allows each head to attend to different combinations of the input. In order to reason about the information at different representational space, Q, K, and V are also uniquely projected prior to the attention being computed. Specifically,

$$T_h = \text{Attention}(QW_h^Q, KW_h^K, VW_h^V) \tag{7}$$

where T_h is the output of a single attention head. $W_h^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}, W_h^K \in \mathbb{R}^{d_{\text{model}} \times d_k}, W_h^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$ are learned linear projection weights for query, key, and value. In order to maintain the computation efficiency, d_k is chosen to be d_{model}/H. The outputs of the attention heads are concatenated and projected to form the final output of the multi-head attention:

$$\text{MultiAttn}(Q, K, V) = \text{Concat}(T_1, ..., T_H)W^O \tag{8}$$

where $W^O \in \mathbb{R}^{Hd_v \times d_{\text{model}}}$ is a learned output projection.

In our model, we use self-attention. Therefore, Q, K, and V are all constructed from h^{l-1}. Each position can freely attend to all positions in the input, thus aiding in modeling inter-relations between properties.

D. Classification and Training

The final layer uses a learned linear transformation and a sigmoid function to convert the encoded input to predicted probabilities of properties. Specifically,

$$p_n = \sigma(E_{\text{value}} \, \text{FCN}(h_n^L)) \tag{9}$$

where FCN is a fully connected layer and E_{value} is the learned embedding matrix used to create input value embeddings. The use of the sigmoid function σ allows the model to accept multiple correct answers, therefore modeling uncertainties in semantic knowledge (e.g., cups can be found in both kitchen and living room)

During training, we construct the masked input by replacing only a single value in an n-ary observation with the [MASK] token. We perform this procedure exhaustively for all values and all n-ary observations in the training set. We then group n-ary observations sharing the same masked instances and use their ground-truth values at the query position to construct a one-hot label (continuous-valued properties are discretized). Scoring multiple instances simultaneously is also known as the 1-N setting [17] and helps reduce training and inference time. We use cross-entropy between the ont-hot label and prediction as training loss. We use label smoothing [58] to prevent overfitting.

E. Implementation Details

All components of the model are trained end-to-end. The best set of parameters is found to be $L = 1$, $H = 4$, $d_{\text{model}} = 240$. We used Adam [34] for optimization. We implement our model using PyTorch and train on a Nvidia GTX1080Ti gpu.

VI. Experiments On LINK Dataset

In this section, we use the value prediction task to assess our model's ability to learn n-ary relations between object properties. In the value prediction task, the model is presented with a previously unseen n-ary observation, and must predict

TABLE III
RESULTS% OF OUR MODEL AND BASELINE MODELS.

Model	Metric Scores				Time (min)	
	MRR	Hits@1	Hits@2	Hits@3	Training	Testing
Co-Occur	63.0	44.3	67.3	80.2	<1	3
TuckER	58.7	38.5	59.9	79.3	<1	3
TuckER+	62.5	43.0	65.9	81.4	2	3
NaLP	57.9	38.9	60.3	75.8	8	10
MLN	66.1	50.2	68.8	82.0	420	487
Transformer (Ours)	**76.3**	**63.3**	**79.4**	**89.1**	3	3

TABLE IV
ABLATION ON INPUT ENCODER DESIGN

Embeddings			Metric Scores			
V	R	Pos	MRR	Hits@1	Hits@2	Hits@3
✓	✓		**76.3**	**63.3**	**79.4**	**89.1**
✓			75.3	62.3	79.0	86.8
✓	✓	✓	74.0	59.9	77.7	87.5
✓		✓	74.0	59.9	77.7	87.5

TABLE V
MRR% OF OUR MODEL AND BASELINE MODELS FOR EACH PROPERTY TYPE.

	Class	Mat	Color	Trans	Dim	Phys	Shape	Temp	Full	Damp	Clean	Room	Loc	Price	Weight	Size
# Values	11	8	15	3	10	6	9	3	3	3	3	11	117	3	3	3
Random	27.1	33.3	19.9	60.8	18.7	37.7	22.6	61.4	61.5	61.5	61.0	28.5	4.7	60.2	60.7	60.5
Co-Occur	/	55.5	39.2	83.1	17.9	76.1	64.4	91.0	77.4	74.5	67.1	56.1	44.3	56.9	70.7	**70.9**
TuckER	/	56.1	37.7	84.0	16.7	74.3	57.8	83.9	59.9	71.6	61.4	57.3	31.4	55.6	64.6	69.0
TuckER+	53.7	60.0	42.2	85.1	20.1	74.8	60.4	90.9	69.7	72.4	62.6	60.9	39.2	59.7	73.8	65.1
NaLP	45.0	47.4	39.0	84.4	17.7	69.2	52.4	91.0	68.5	70.8	67.7	55.2	23.1	59.0	61.6	61.4
MLN	62.9	**79.1**	72.5	95.7	25.3	79.4	64.3	90.0	69.1	69.5	65.3	67.7	4.6	68.9	75.8	63.7
Transformer (ours)	**72.3**	78.9	**73.2**	**97.1**	**43.8**	**84.1**	**75.7**	**92.3**	**90.7**	**82.2**	**90.2**	**68.5**	**59.6**	**74.4**	**76.6**	61.8

a single missing value given the value's role and all other role-value pairs in the instance.

A. Experimental Setup

Data Split: In our dataset, each n-ary observation corresponds to a situated object instance. To prevent test leakage, we first split object instances in the dataset into 70% training, 15% testing, and 15% validation. Situated object instances are then assigned to the correct set based on its corresponding object instance.

Metrics: For each missing value in a test instance, we obtain probabilities of candidate values from the model. Then the candidate values are sorted in descending order based on the probabilities. The rank of the ground-truth value v_n is used to compute metric scores. During ranking, we adopt the filtered setting [17] to remove any value v'_n different from v_n if $\{r_1 : v_1, ..., r_{n-1} : v_{n-1}, r_n : v'_n\}$ exists in the train, validation, or test set. This whole procedure is repeated for each value of each testing instance in the test set. We report standard metric Mean Reciprocal Rank (MRR) and proportion of ranks no larger than 1, 2, and 3 (Hits@1, 2, and 3). For both MRR and Hits, a higher score indicates better performance.

Baselines: We compare against the following baselines:

- **Co-Occur** learns co-occurrence frequency of entities. This model has been used for modeling semantic relations in various robotic applications, including modeling object object co-occurrence [36], object affordance co-occurrence [11], and object grasp co-occurrence [37]. We apply this model to learn the co-occurrence frequency of object class with object properties in our experiments. The model by design is not able to consider other properties as contextual information.
- **TuckER** is a recent state of the art knowledge graph embedding model [6]. In this paper, we compare to two variants of TuckER. The regular TuckER model follows existing work [16, 4] to model binary relations between object class and object properties.
- **TuckER+** is a TuckER embedding model we implement to model binary relations between all pairs of property types (e.g., color and material, shape and location); it approximates an n-ary relation with a combination of binary relations.
- **NaLP** is a neural network model developed for modeling n-ary relational data in knowledge graphs [26]. NaLP explicitly models the relatedness of all the role-value pairs in an n-ary observation. We apply this model to learn n-ary relations between object properties.
- **Markov Logic Network (MLN)** represents probabilistic logic languages that have been used to model complex semantic relations in various robotic domains [51, 71, 52, 3, 13]. We closely follow prior work to specify probabilistic rules for our domain.

B. Results

As shown in Table III, our model outperforms existing methods by significant margins on all metrics. Compared to the second-best model, **MLN**, our model achieves a 10% increase in MRR while reducing training and testing time by 150 times. In comparison with **NaLP**, another model developed specifically for modeling n-ary data, our model's superior performance confirms that the transformer structure and multi-head attention mechanism are more effective at learning the complex semantic relations between object properties. We also observe that **TuckER+**, which learns binary relations between all pairs of object properties, outperforms the regular **TuckER**. This result demonstrates that only modeling class-level semantic knowledge can lead to over-generalization, and that reasoning about the differences between object instances is

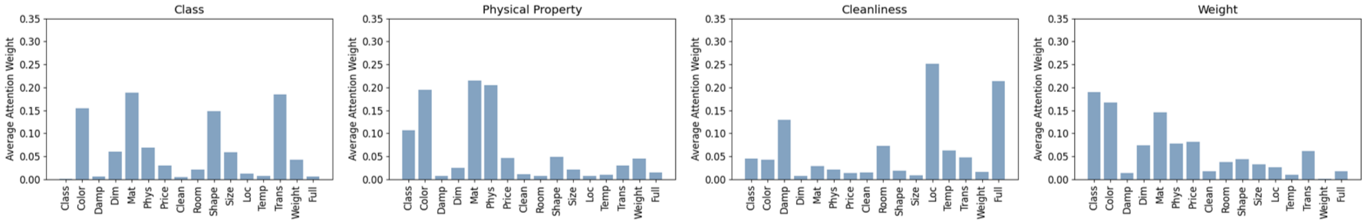


Fig. 4. Visualizations of the attention weights illustrate that different amount of information from each property type is used by our model to predict different types of properties.

crucial. It is worth noting that **NaLP** and **TuckER** variants are not able to outperform the simpler **Co-Occur** model. **TuckER** variants are good at learning latent representation of the global structure, but our analysis shows that they do not capture the frequencies of the binary relations well. **NaLP** has shown to be effective at modeling n-ary facts mainly on dataset with 2 to 6 role-values, but it struggles to learn n-ary relations in our data which can have up to 24 role-values.

Further analyzing MRR for each type of query shown in Table V, we see that our model outperforms existing models in predicting most of the properties. We also notice that the baselines have degraded performance at predicting property types with many candidate values (e.g., location, room, and dimension). **MLN** especially struggles to predict the location role which has 117 possible values. One potential explanation is the closed world assumption being made by **MLN**. Our model uses label smoothing to prevent being overconfident at negative training examples and has demonstrated good performance even for these many-valued role types.

C. Ablation on Input Encoder Design

We investigate our input encoder design with an ablation study. Specifically, we examine the effect of the role embeddings and positional embeddings (discussed in Section V-A). Results in Table IV show that enforcing the order of role-value pairs in an n-ary observation using the positional embeddings results in a drop in performance. The results also confirm that role embeddings are useful for modeling multimodal object properties represented as role-value pairs.

D. Visualizing Attention

To understand why our transformer-based model is effective at modeling n-ary relational data, we visualize the multi-head attention weights, i.e., softmax$(\frac{QK^T}{\sqrt{d_k}})$. Figure 4 shows the average attention weight assigned to each role when predicting class, physical property, cleanliness, and weight. The attention mechanism exhibits n-ary relational reasoning patterns, which correspond strongly with human intuition—for example, dampness, location, and fullness of an object aids in predicting its cleanliness. Baseline models cannot perform this type of reasoning and thus are not able to model object properties as well as our model.

VII. Robot Experiment: Object Search

In this section, we demonstrate how a household robot can locate specific objects based on users' requests by leveraging the explicit representation of our learned n-ary knowledge. Our experiment serves two purposes, i) to validate our model in a realistic physical setting with non-AMT users, and ii) to test our model's ability to handle queries that reflect realistic use cases, such as a human asking a robot to find a cold beverage or collect dirty dishes. Queries used in this study utilize only a sparse set of known properties[2], and the robot's task is to predict multiple unknown properties. Specifically, we seek to predict the room and location of each object.

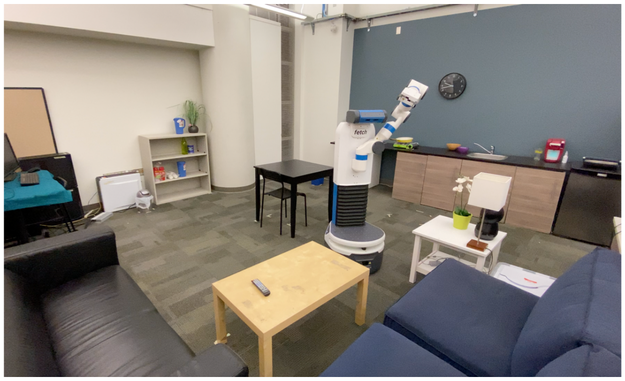


Fig. 5. Home environment for the object search experiment.

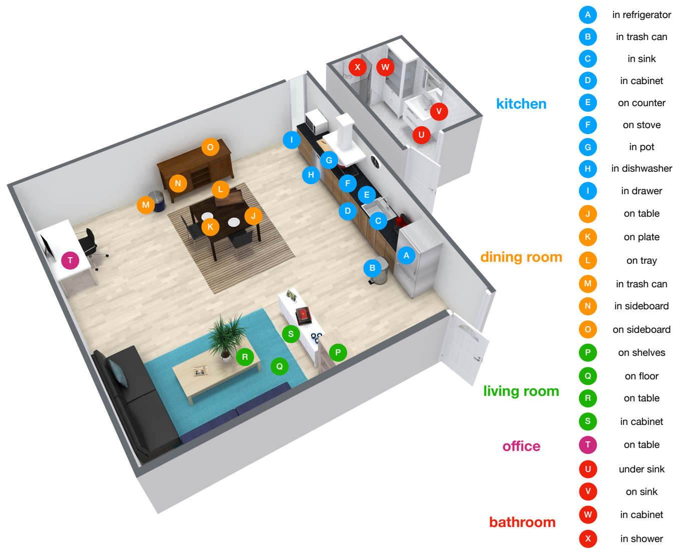


Fig. 6. 3D floor plan for remotely collecting preferred locations of objects from five users.

[2] Human users are unlikely to phrase requests with long adjective sequences.

TABLE VI
RESULTS% ON OBJECT SEARCH

	Hits@1	Hits@2	Hits@3	Hits_Any@1	Hits_Any@2	Hits_Any@3
Human Baseline	34.8 ± 6.5	52.0 ± 7.0	64.7 ± 7.3	64.7 ± 7.2	83.2 ± 7.0	90.6 ± 3.9
Co-Occur	**20.0**	29.0	36.8	**49.0**	68.0	80.0
TuckER+	8.4	20.0	29.6	37.2	60.0	74.4
NaLP	1.2	2.8	4.4	7.2	12.4	16.8
Transformer	17.2	**34.4**	**48.0**	42.4	**73.6**	**87.6**

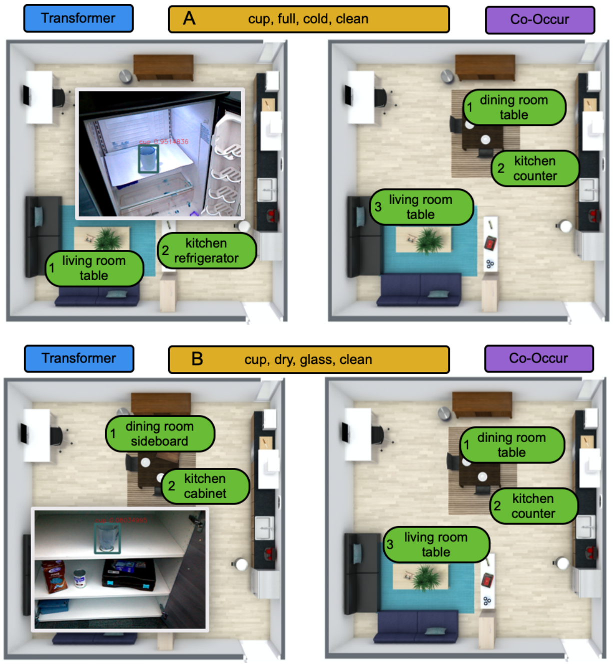


Fig. 7. Two object search tests comparing our model with Co-Occur. Provided properties are shown on top.

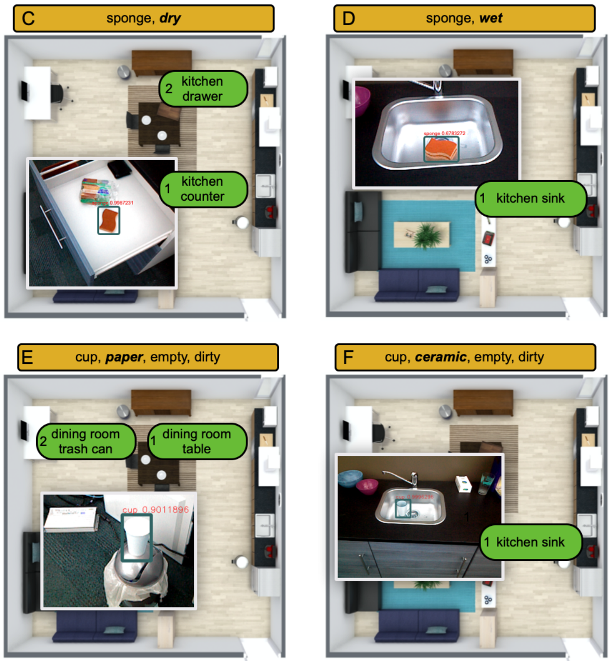


Fig. 8. Our model predicts different locations based on the given object properties.

We set up a home environment in our lab with 4 rooms and typical household furniture (Figure 5). We also generated a corresponding 3D floor plan of the environment (adding an additional bathroom), which listed 24 possible locations for storing objects (Figure 6). We then recruited 5 users, and had them label their preferred locations for 50 object instances sampled from our dataset. For each object, the user was shown an image of the object, given 1-3 properties describing the state of the object (i.e., cleanliness, temperature, dampness, and content), and then asked to list 3 ranked likely locations for the object.

We compare the performance of our model against **NaLP**, **Co-Occur**, and **TuckER+**. We leave **MLN** out because of its exceedingly long inference time on queries with partial evidence (as a large number of properties other than the query properties were missing). All models are trained on our complete dataset to validate against collected user data. All models have access to the properties given to the human users as well as the class and material of the object. To predict likely room-location combinations, separately predicted probabilities of the two properties are multiplied and ranked.

We use Hits@K and Hits_Any@K as metrics. Hits@1,2,3 indicate the percentage of times that a model correctly predicts a user's most preferred location of an object within 1, 2, and 3 attempts, respectively. We also introduce Hits_Any@K, which considers a prediction correct if it matches any one of the 3 locations listed by a user, without rank order.

Table VI summarizes the result of this experiment. We also report the human baseline, which we compute by cross-validating each user against the other users. We observe that only our model is able to reach within the range of human performance at Hits_Any@2 and 3. **Co-Occur** outperforms our model at Hits@1 and Hits_Any@1, suggesting that class-level frequency is a good heuristic for finding objects if given only one chance. However, given that only approximately 30% of objects can be retrieved in one shot, even by humans, we argue that our model is more beneficial in the general use case.

Beyond quantitative difference between our model and baselines, we also demonstrate the qualitative improvement on a Fetch robot [67]. The robot is equipped with the navigation stack developed in [7] for mapping and navigation, and the method introduced in [43] for object detection and grasping. As shown in Figure 7, the difference (A, B) between our model and **Co-Occur** is clear as our model takes into account of the

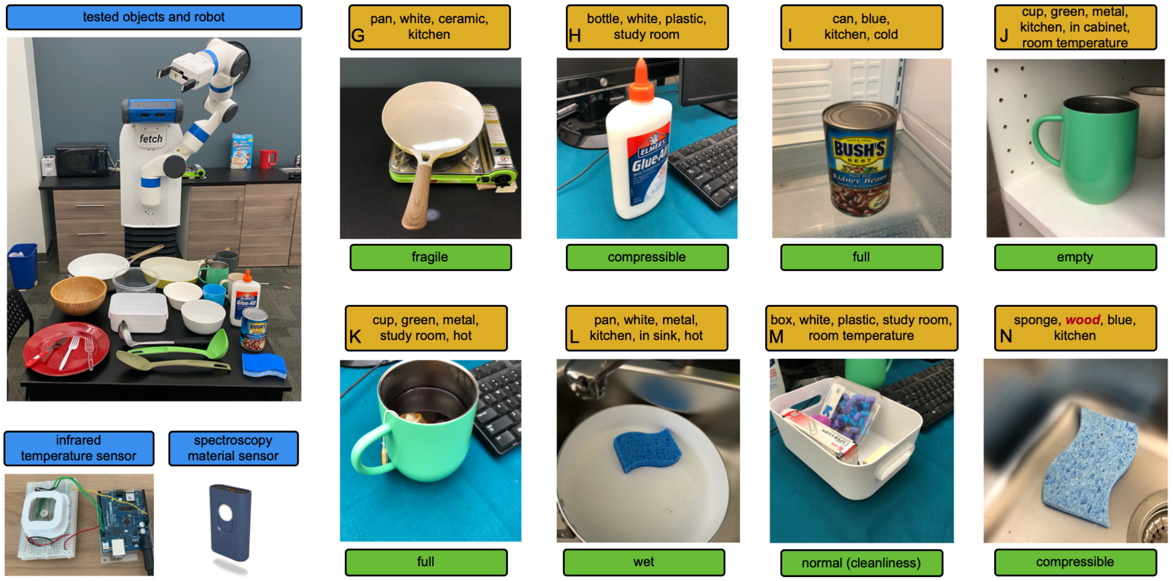

tested objects and robot
fetch
infrared temperature sensor
spectroscopy material sensor
G pan, white, ceramic, kitchen
fragile
H bottle, white, plastic, study room
compressible
I can, blue, kitchen, cold
full
J cup, green, metal, kitchen, in cabinet, room temperature
empty
K cup, green, metal, study room, hot
full
L pan, white, metal, kitchen, in sink, hot
wet
M box, white, plastic, study room, room temperature
normal (cleanliness)
N sponge, *wood*, blue, kitchen
compressible

Fig. 9. The Fetch robot uses an infrared temperature sensor to detect the temperature and a spectrometer to detect the material of each novel object situated in different environmental contexts. Our model leverages extracted information (shown on top of each figure on the right) to predict an unknown object property (shown on bottom).

properties of objects (e.g., cold, dry, clean) while **Co-Occur** searches the same locations for different cups. We also show in Figure 8 that our model is able to find objects considering both immutable (material in E and F) and mutable properties of objects (dampness in C and D).

VIII. Robot Experiment: Integrating with Multimodal Perception

In this section, we examine whether our model can enable a robot to infer object properties that cannot be directly observed by collectively reasoning about properties extracted from multimodal sensors. This experiment also aims to test whether our model can generalize learned n-ary knowledge to new object instances in the real world.

In this experiment, a robot is tasked to predict either an unknown immutable property of an object based on its class, color, material, and room, or to predict an unknown mutable property based on class, color, material, room, temperature, and location. The robot physically interacts with real objects situated in the environment and leverages different sensing capabilities to extract multimodal observations. We use the same Fetch robot, object detection, and mapping as the previous experiment. Color is detected using OpenCV. Material is detected by the robot using a spectrometer, the SCiO sensor, and the method introduced in [21]. Temperature is detected using a Melexis contact-less infrared sensor connected to an Arduino microcontroller. To detect materials and temperatures of objects in real time, the sensors are attached to the end-effector of the robot. The robot uses RRT to plan to poses that allow the sensors to touch the surfaces of the objects. The poses are computed from task-oriented 6-dof grasping poses with the method introduced in [43]. As shown in Figure 9, we test on 22 objects which are semantically different from objects in our dataset (e.g., no ceramic pan and plastic box exist in our dataset).

In this experiment, our model is able to correctly predict 34/52 (65%) of the queried object properties. In comparison, the second best performing models, **TuckER+** and **Co-Occur**, both correctly predict 24/52 (46%). Object materials are correctly detected 45/52 (87%) times. Figure 9 shows examples of the queries.

IX. Conclusion

This work addresses the problem of predicting semantic properties of objects based on partial observations. We introduce a scalable transformer neural network that learns n-ary relations between object properties from observations. The model can perform inference at different levels of abstraction by conditioning on different numbers of input properties. We also contribute LINK, a dataset containing objects situated in various environmental contexts and modeled by diverse semantic properties. Evaluation of our model on the collected data shows significant improvements over prior methods, including knowledge graph embedding models and a probabilistic logic language. The learned model helps a Fetch robot perform two tasks common in everyday environments: searching for objects based on multiple desired properties and inferring object properties given partial multimodal observations. Future directions of this work include learning n-ary relations from different data sources such as incomplete semantic data and multimodal sensory data. We also hope to investigate using n-ary knowledge about object properties to guide interactive perception and inform object manipulation.

Acknowledgments

This work was supported by NSF IIS 1564080, NSF GRFP DGE-1650044, and ONR N00014-16-1-2835.

References

[1] R Abboud, II Ceylan, T Lukasiewicz, and T Salvatori. Boxe: A box embedding model for knowledge base completion. *NeurIPS Proceedings*, 33, 2020.

[2] Saeid Amiri, Suhua Wei, Shiqi Zhang, Jivko Sinapov, Jesse Thomason, and Peter Stone. Multi-modal predicate identification using dynamically learned robot controllers. In *IJCAI*, pages 4638–4645, 2018.

[3] Paola Ardón, Èric Pairet, Ronald PA Petrick, Subramanian Ramamoorthy, and Katrin S Lohan. Learning grasp affordance reasoning through semantic relations. *IEEE Robotics and Automation Letters*, 4(4):4571–4578, 2019.

[4] Jacob Arkin, Daehyung Park, Subhro Roy, Matthew R Walter, Nicholas Roy, Thomas M Howard, and Rohan Paul. Multimodal estimation and communication of latent semantic knowledge for robust execution of robot instructions. *The International Journal of Robotics Research*, page 0278364920917755, 2020.

[5] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[6] Ivana Balazevic, Carl Allen, and Timothy Hospedales. TuckER: Tensor factorization for knowledge graph completion. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5185–5194, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1522. URL https://www.aclweb.org/anthology/D19-1522.

[7] Siddhartha Banerjee, Angel Daruna, David Kent, Weiyu Liu, Jonathan Balloch, Abhinav Jain, Akshay Krishnan, Muhammad Asif Rana, Harish Ravichandar, Binit Shah, Nithin Shrivatsav, and Sonia Chernova. Taking recoveries to task: Recovery-driven development for recipe-based robot tasks. In *International Symposium on Robotics Research (ISRR)*, 2019. URL http://www.banerjs.com/pdf/banerjs_isrr2019.pdf.

[8] Tapomayukh Bhattacharjee, Henry M Clever, Joshua Wade, and Charles C Kemp. Multimodal tactile perception of objects in a real home. *IEEE Robotics and Automation Letters*, 3(3):2523–2530, 2018.

[9] Jeannette Bohg, Karol Hausman, Bharath Sankaran, Oliver Brock, Danica Kragic, Stefan Schaal, and Gaurav S Sukhatme. Interactive perception: Leveraging action in perception and perception in action. *IEEE Transactions on Robotics*, 33(6):1273–1291, 2017.

[10] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, 2019.

[11] Yu-Wei Chao, Zhan Wang, Rada Mihalcea, and Jia Deng. Mining semantic affordances of visual object categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4259–4267, 2015.

[12] Haonan Chen, Hao Tan, Alan Kuntz, Mohit Bansal, and Ron Alterovitz. Enabling robots to understand incomplete natural language instructions using commonsense reasoning. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1963–1969. IEEE, 2020.

[13] Sonia Chernova, Vivian Chu, Angel Daruna, Haley Garrison, Meera Hahn, Priyanka Khante, Weiyu Liu, and Andrea Thomaz. Situated bayesian reasoning framework for robots operating in diverse everyday environments. In *International Symposium on Robotics Research (ISRR)*, 2017.

[14] Vivian Chu, Ian McMahon, Lorenzo Riano, Craig G McDonald, Qin He, Jorge Martinez Perez-Tejada, Michael Arrigo, Trevor Darrell, and Katherine J Kuchenbecker. Robotic learning of haptic adjectives through physical interaction. *Robotics and Autonomous Systems*, 63:279–292, 2015.

[15] Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. Learning to act properly: Predicting and explaining affordances from images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 975–983, 2018.

[16] Angel Daruna, Weiyu Liu, Zsolt Kira, and Sonia Chetnova. Robocse: Robot common sense embedding. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 9777–9783. IEEE, 2019.

[17] Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. Convolutional 2d knowledge graph embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[19] Thanh-Toan Do, Anh Nguyen, and Ian Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 5882–5889. IEEE, 2018.

[20] Manfred Eppe, Matthias Kerzel, Erik Strahl, and Stefan Wermter. Deep neural object analysis by interactive auditory exploration with a humanoid robot. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 284–289. IEEE, 2018.

[21] Zackory Erickson, Nathan Luskey, Sonia Chernova, and Charles C Kemp. Classification of household materials via spectroscopy. *IEEE Robotics and Automation Letters*, 4(2):700–707, 2019.

[22] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In

2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 1778–1785. IEEE, 2009.

[23] Vittorio Ferrari and Andrew Zisserman. Learning visual attributes. *Advances in neural information processing systems*, 20:433–440, 2007.

[24] Mikhail Galkin, Priyansh Trivedi, Gaurav Maheshwari, Ricardo Usbeck, and Jens Lehmann. Message passing for hyper-relational knowledge graphs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7346–7359, 2020.

[25] Dhiraj Gandhi, Harikrishna Mulam, and Lerrel Pinto. Swoosh! rattle! thump! - actions that sound. In *Proceedings of Robotics: Science and Systems*, 07 2020. doi: 10.15607/RSS.2020.XVI.002.

[26] Saiping Guan, Xiaolong Jin, Yuanzhuo Wang, and Xueqi Cheng. Link prediction on n-ary relational data. In *The World Wide Web Conference*, pages 583–593, 2019.

[27] Martin Günther, JR Ruiz-Sarmiento, Cipriano Galindo, Javier Gonzalez-Jimenez, and Joachim Hertzberg. Context-aware 3d object anchoring for mobile robots. *Robotics and Autonomous Systems*, 110:12–32, 2018.

[28] Rakesh Gupta, Mykel J Kochenderfer, Deborah Mcguinness, and George Ferguson. Common sense data acquisition for indoor mobile robots. In *AAAI*, pages 605–610, 2004.

[29] Birgit Hamp and Helmut Feldweg. Germanet-a lexical-semantic net for german. In *Automatic information extraction and building of lexical semantic resources for NLP applications*, 1997.

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[31] Ashesh Jain, Brian Wojcik, Thorsten Joachims, and Ashutosh Saxena. Learning trajectory preferences for manipulators via iterative improvement. In *Advances in neural information processing systems*, pages 575–583, 2013.

[32] Patrick Jenkins, Rishabh Sachdeva, Gaoussou Youssouf Kebe, Padraig Higgins, Kasra Darvish, Edward Raff, Don Engel, John Winder, Francisco Ferraro, and Cynthia Matuszek. Presentation and analysis of a multimodal dataset for grounded language learning. *arXiv preprint arXiv:2007.14987*, 2020.

[33] Emmett Kerr, T Martin McGinnity, and Sonya Coleman. Material recognition using tactile sensing. *Expert Systems with Applications*, 94:94–111, 2018.

[34] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

[35] Eric Kolve, Roozbeh Mottaghi, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

[36] Lars Kunze, Chris Burbridge, Marina Alberti, Akshaya Thippur, John Folkesson, Patric Jensfelt, and Nick Hawes. Combining top-down spatial reasoning and bottom-up object class recognition for scene understanding. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2910–2915. IEEE, 2014.

[37] Safoura Rezapour Lakani, Antonio J Rodríguez-Sánchez, and Justus Piater. Exercising affordances of objects: A part-based approach. *IEEE Robotics and Automation Letters*, 3(4):3465–3472, 2018.

[38] Séverin Lemaignan, Mathieu Warnier, E Akin Sisbot, Aurélie Clodic, and Rachid Alami. Artificial cognition for social human–robot interaction: An implementation. *Artificial Intelligence*, 247:45–69, 2017.

[39] Douglas B. Lenat. Cyc: A large-scale investment in knowledge infrastructure. *Commun. ACM*, 38(11):33–38, November 1995. ISSN 0001-0782.

[40] Qiang Li, Oliver Kroemer, Zhe Su, Filipe Fernandes Veiga, Mohsen Kaboli, and Helge Joachim Ritter. A review of tactile information: Perception and action through touch. *IEEE Transactions on Robotics*, 36(6): 1619–1634, 2020.

[41] Gi Hyun Lim, Il Hong Suh, and Hyowon Suh. Ontology-based unified robot knowledge for service robots in indoor environments. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 41 (3):492–509, 2011.

[42] Hugo Liu and Push Singh. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226, 2004.

[43] Weiyu Liu, Angel Daruna, and Sonia Chernova. Cage: Context-aware grasping engine. In *International Conference on Robotics and Automation (ICRA)*, 2020. URL https://arxiv.org/abs/1909.11142.

[44] Yu Liu, Quanming Yao, and Yong Li. Generalizing tensor decomposition for n-ary relational knowledge bases. In *Proceedings of The Web Conference 2020*, pages 1104–1114, 2020.

[45] Shan Luo, Joao Bimbo, Ravinder Dahiya, and Hongbin Liu. Robotic tactile perception of object properties: A review. *Mechatronics*, 48:54–67, 2017.

[46] Dermot Lynott and Louise Connell. Modality exclusivity norms for 423 object properties. *Behavior Research Methods*, 41(2):558–564, 2009.

[47] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[48] Bogdan Moldovan and Luc De Raedt. Occluded object search by relational affordances. *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 169–174, 2014.

[49] Michal Nazarczuk and Krystian Mikolajczyk. Shop-vrb: A visual reasoning benchmark for object perception. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6898–6904. IEEE, 2020.

[50] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33, 2015.

[51] Daniel Nyga, Ferenc Balint-Benczedi, and Michael Beetz. Pr2 looking at things—ensemble learning for unstructured information processing with markov logic networks. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3916–3923. IEEE, 2014.

[52] Daniel Nyga, Subhro Roy, Rohan Paul, Daehyung Park, Mihai Pomarlan, Michael Beetz, and Nicholas Roy. Grounding robot plans from natural language instructions with incomplete world knowledge. In *Conference on Robot Learning*, pages 714–723, 2018.

[53] Paolo Rosso, Dingqi Yang, and Philippe Cudré-Mauroux. Beyond triplets: hyper-relational knowledge graph embedding for link prediction. In *Proceedings of The Web Conference 2020*, pages 1885–1896, 2020.

[54] Ashutosh Saxena, Ashesh Jain, Ozan Sener, Aditya Jami, Dipendra K Misra, and Hema S Koppula. Robobrain: Large-scale knowledge engine for robots. *arXiv preprint arXiv:1412.0691*, 2014.

[55] Mohit Shridhar and David Hsu. Interactive visual grounding of referring expressions for human-robot interaction. In *Proceedings of Robotics: Science and Systems*, 2018.

[56] Jivko Sinapov, Connor Schenck, Kerrick Staley, Vladimir Sukhoy, and Alexander Stoytchev. Grounding semantic categories in behavioral interactions: Experiments with 100 objects. *Robotics and Autonomous Systems*, 62(5): 632–645, 2014.

[57] Yuyin Sun, Liefeng Bo, and Dieter Fox. Attribute based object identification. In *2013 IEEE international conference on robotics and automation*, pages 2096–2103. IEEE, 2013.

[58] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[59] Gyan Tatiya and Jivko Sinapov. Deep multi-sensory object category recognition using interactive behavioral exploration. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7872–7878. IEEE, 2019.

[60] Moritz Tenorth and Michael Beetz. Representations for robot knowledge in the knowrob framework. *Artificial Intelligence*, 247:151–169, 2017.

[61] Jesse Thomason, Jivko Sinapov, Raymond J Mooney, and Peter Stone. Guiding exploratory behaviors for multimodal grounding of linguistic descriptions. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[62] Madhura Thosar, Christian A Mueller, Georg Jäger, Johannes Schleiss, Narender Pulugu, Ravi Mallikarjun Chennaboina, Sai Vivek Rao Jeevangekar, Andreas Birk, Max Pfingsthorn, and Sebastian Zug. From multi-modal property dataset to robot-centric conceptual knowledge about household objects. *Frontiers in Robotics and AI*, 8:87, 2021.

[63] Karthik Mahesh Varadarajan and Markus Vincze. Afnet: The affordance network. In Kyoung Mu Lee, Yasuyuki Matsushita, James M. Rehg, and Zhanyi Hu, editors, *Computer Vision – ACCV 2012*, pages 512–523, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-37331-2.

[64] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[65] Quan Wang, Pingping Huang, Haifeng Wang, Songtai Dai, Wenbin Jiang, Jing Liu, Yajuan Lyu, Yong Zhu, and Hua Wu. Coke: Contextualized knowledge graph embedding. *arXiv preprint arXiv:1911.02168*, 2019.

[66] Jianfeng Wen, Jianxin Li, Yongyi Mao, Shini Chen, and Richong Zhang. On the representation and embedding of knowledge bases beyond binary relations. In *IJCAI*, 2016.

[67] Melonee Wise, Michael Ferguson, Derek King, Eric Diehr, and David Dymesich. Fetch and freight: Standard platforms for service robot applications. In *Workshop on autonomous mobile service robots*, 2016.

[68] Wei Yang, Xiaolong Wang, Ali Farhadi, Abhinav Gupta, and Roozbeh Mottaghi. Visual semantic navigation using scene priors. In *Proceedings of Seventh International Conference on Learning Representations (ICLR 2019)*, May 2019.

[69] Zheng Zeng, Adrian Röfer, Shiyang Lu, and Odest Chadwicke Jenkins. Generalized object permanence for object retrieval through semantic linking maps. In *IEEE ICRA 2019 Workshop on High Accuracy Mobile Manipulation in Challenging Environments*, 2019.

[70] Richong Zhang, Junpeng Li, Jiajie Mei, and Yongyi Mao. Scalable instance reconstruction in knowledge bases via relatedness affiliated embedding. In *Proceedings of the 2018 World Wide Web Conference*, pages 1185–1194, 2018.

[71] Yuke Zhu, Alireza Fathi, and Li Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *European conference on computer vision*, pages 408–424. Springer, 2014.