

Image Understanding using Vision and Reasoning through Scene Description Graph

Somak Aditya, Yezhou Yang, Chitta Baral

Arizona State University, Tempe, AZ

Yiannis Aloimonos, Cornelia Fermüller

University of Maryland, College Park, MD

Abstract

Two of the fundamental tasks in image understanding using text are caption generation and visual question answering [1, 2]. This work presents an intermediate knowledge structure that can be used for both tasks to obtain increased interpretability. We call this knowledge structure *Scene Description Graph (SDG)*, as it is a directed labeled graph, representing objects, actions, regions, as well as their attributes, along with inferred concepts and semantic (from KM-Ontology [3]), ontological (i.e. superclass, hasProperty), and spatial relations. Thereby a general architecture is proposed in which a system can represent both the content and underlying concepts of an image using an SDG. The architecture is implemented using generic visual recognition techniques and commonsense reasoning to extract graphs from images. The utility of the generated SDGs is demonstrated in the applications of image captioning, image retrieval, and through examples in visual question answering. The experiments in this work show that the extracted graphs capture syntactic and semantic content of images with reasonable accuracy.

1. Introduction and Motivation

Image Understanding is fundamental to Computer Vision. Earlier approaches centered on asking “what” and “where” questions about the scene in view. In this methodology, scenes are recognized by detecting the objects within the

5 scene [4, 5, 6], objects are recognized by detecting their parts or attributes [7, 8, 9, 10, 11, 12, 13] and activities are recognized by detecting the motions, objects and contexts involved in the activities [14, 15, 16, 17, 18, 19].

Since then, researchers have explored multiple ways of understanding an image through the modality of natural language. According to [20], the primary reason for using natural language to ground images is that it adds interpretability and creates a way for human-machine interaction. The first major challenge proposed in this area, is the problem of caption generation from images. Researchers adopted the viewpoint that if a system is able to develop a semantic understanding of a visual scene, then such a system should be able to produce natural language descriptions of such semantics. Recent developments 15 produce natural language descriptions of such semantics. Recent developments [21, 22, 23, 24, 25, 26, 27] in Computer Vision have shown that deep neural nets can be trained to generate a caption for an arbitrary scene with decent success. However, caption generation systems only describe the salient aspects of the image. An intelligent Image Understanding system should recognize all aspects present in the image and where the objects are [28] and should be able to reason with the recognized aspects. Based on such notions and taking advantage of recent powerful recognition capabilities using Neural Networks, researchers in Computer Vision have re-visited a more general and difficult image understanding task, namely Visual Question Answering [1, 29, 30, 31].

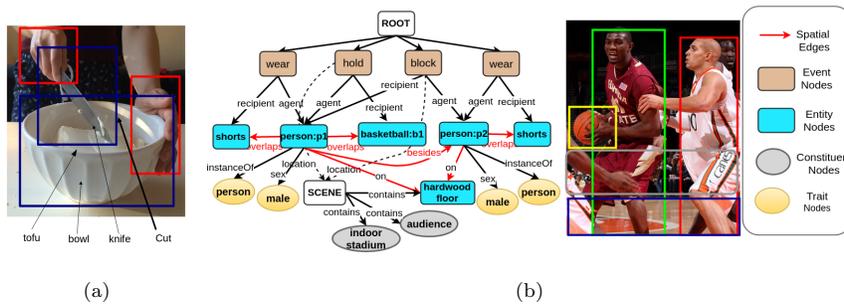


Figure 1: (a) First example image and (b) second example image with corresponding ideal SDG encoding semantic, ontological, and spatial relations.

25 Despite the success of end-to-end learning models ([1, 29, 30, 31]) in these

tasks, a few problems remain. In the Visual Question Answering problem, questions such as: *Is it going to rain?* (prospective), *Did it rain?* (retrospective), *Is the knife cutting the bowl?* (in the context of Figure 1(a)), *Does the man have 20-20 vision?* (commonsense), all require explicit modeling of commonsense reasoning and knowledge. In the context of the image in Figure 1(b), questions can range from those that require basic knowledge about the game of basketball (*Do the players in red and white belong to the same team?*) to questions requiring deeper knowledge such as originating from an intuition of Physics (*Will the player on the right be able to block the player holding the ball?* or *In which direction should the player holding the ball move?*). Without explicit modeling of commonsense knowledge, these questions are difficult to answer. Again, the existing models consider a constrained set of answers, which limits their application to real-world scenarios.

Current state-of-the-art image captioning systems have a few drawbacks such as: 1) a brute-force image to caption mapping does not allow symbol level reasoning beyond simple inferences from annotated data; 2) they are language dependent, due to the lack of concept level modeling; and 3) most importantly, when the system produces wrong results, it is almost impossible to trace back the error and analyze the cause.

To alleviate these problems, we seek inspiration from nature. Human perception is active, selective and exploratory ([32, 33]). We interpret visual input by using our knowledge of activities, events and objects. When we analyze a visual scene, visual processes continuously interact with our high-level knowledge, some of which is represented in the form of language. In some sense, perception and language are engaged in an interaction, as they exchange information that leads to semantics and understanding. Thus, our problem requires at least two modules for its solution: (a) a vision module and (b) a reasoning module that interact with each other. In this paper we propose to model the architecture that can support such an interaction; and we propose a corresponding knowledge structure that can represent the information and the semantics extracted from images.

We present an implementation that integrates deep learning based vision and state-of-the-art concept modeling from common-sense knowledge¹ obtained from text. We use a deep learning-based perception system to obtain the objects, scenes and constituents with probabilistic weights from an input image. To predict how the objects interact in the scene, we build a common-sense knowledge base² from image annotations along with a Bayesian Network of commonly occurring objects and inferred scene constituents (the concepts that can not be seen, but can be understood from the scene). These two pre-computed resources help us infer the following: 1) the correct set of correlated objects based on the objects detected with high-confidence; 2) the most probable actions that these objects participate in; 3) the role that the objects play in these actions. Based on the actions, the detected objects and the inferred constituents, we output a Scene Description Graph (SDG) that represents the semantics of the scene.

In Figure 1, we show a possible SDG for an example image. SDG is a directed labeled graph³ among Entities (objects, regions), Events (actions, linking verbs), Traits (attributes of objects and regions) and inferred constituents. An SDG represents semantic relations (from KM-Ontology [3]) between Entity-Event pairs, spatial relations among Entities (objects and regions), and ontological relations between Entity-Trait pairs. The Event nodes are connected to a dummy node, denoted SCENE, by an edge labeled `location`. The constituent nodes are coded in a different color, to show the concepts that can be inferred from the image. The spatial relations are inspired by [37]. These SDGs can be used to generate captions, answer factual questions and also reason beyond

¹Commonsense reasoning and commonsense knowledge can be of many types ([34]). Commonsense knowledge can belong to different levels of abstraction ([35, 36]). In this paper, we focus on reasoning based on knowledge about natural scenes.

²Domain-specific commonsense and background knowledge can be extracted from text or accessed from curated or semi-curated sources such as WordNet, ConceptNet. Here we extract the needed knowledge from image captions.

³Note that similar structures are also generated by Semantic parsers such as K-parser (kparser.org).

80 what can be seen in the image.

The fundamental **contributions** of this work are: 1) proposing an intermediate structure that captures the semantics of an image, 2) proposing an Image Understanding architecture that combines vision and reasoning modules to generate such structures, 3) an implementation of the architecture by combining a Deep Learning based Visual module with probabilistic reasoning on
85 a Commonsense Knowledge Base, 4) enhancing the Flickr8k dataset with the observable scene constituents (actions and properties involving objects), and 5) comparative human evaluations dataset for our approach, two popular neural approaches ([24, 38]) and ground truth captions for three existing Captioning
90 Datasets (Flickr8k, Flickr30k and MS-COCO)⁴, which can be used to propose better automatic caption evaluation metrics (this dataset is used in [39] to propose SPICE).

2. Related Works

Our work is influenced by various thrusts of work focusing on extracting
95 meaningful information from images and videos. As suggested by [24], such works can be categorized into 1) dense image annotations, 2) generating textual descriptions, 3) grounding natural language in images, and 4) neural networks in visual and language domains. Furthermore, automatic caption generation systems, according to ([40]), may be classified into the following three categories: i) direct generation models, ii) retrieval models from visual space, and
100 iii) retrieval models from multimodal space.

Caption Generation: With respect to Caption Generation tasks, we share our roots with the works on generating textual descriptions, i.e., direct generation models. These include the works in ([41], [42],[43], [44]) which retrieve
105 and rank sentences from training sets given an image. Other works ([37], [45],

⁴Comparison with both the neural approaches are done on MS-COCO dataset. For the rest, comparison is done only with [24].

[46], [47], [48]) have generated descriptions by stitching together annotations or applying templates on detected image contents.

Following the initial keyword-based approaches, most approaches now use Neural Network architectures. One of the first works was by Karpathy et. al. [24], which used a combination of a Convolutional Neural Network (for images) and a bi-directional Recurrent Neural Network (for sentences). Subsequent works ([22, 49, 50, 51]) adopted different Neural Network architectures to directly generate captions (a sentence) by training on large datasets of (image, caption) pairs.

Our aim in this work is to construct an intermediate interpretable structure that represents both necessary and relevant information about the image. We can use this interpretable structure to not only generate captions but also to reason about images beyond the visual content.

Scene Graph: A small number of works in Computer Vision and Robot Perception aims at producing a semantic structure from scenes that captures information about the objects and regions. We propose here a scene description graph in which entities (nouns) and events (verbs) are connected by well-defined relations. The purpose is to perform downstream spatial and event-based reasoning using reasoning engines. The relations in scene graphs in ([52]) are open-ended phrases and the Spatial Graphs in ([37]) only represent the spatial relations between objects and regions. Reasoning directly on such structures using known logical reasoning languages (such as Answer Set Programming [53], ProbLog [54]) is not straightforward.

Applying Commonsense in Vision: There are a few works with promising efforts to acquire and apply common-sense aspects to the analysis of scenes. The work in [55] uses abstraction to discover semantically similar images, [56] proposes to learn all variations pertaining to all concepts, and [57] uses common-sense to learn actions.

Question Answering: Our work is also related to the recent research in the field of **Visual Question Answering**. Researchers have spent a significant amount of effort on both creating datasets and proposing new models [1, 29,

30, 31]. Interestingly, both [1] and [30] have adapted MS-COCO [58] images to create an open domain dataset with human generated questions and answers. The works [29] and [30] use recurrent networks to encode the sentence describing an image and output the answer. There are multiple existing models which use a combination of attention mechanisms in a combined Convolutional and Recurrent Neural Network architecture. However, the task of VQA also requires a modeling of commonsense knowledge and reasoning. This is lacking in existing architectures. In this work, we conduct case studies to show the promising potential of the SDG for answering questions using reasoning with additional knowledge.

3. An Image Understanding Architecture

An image is a vast and complex source of information. To understand an image, one needs to recognize the different components (objects, actions, scenes) and infer higher-level events, activities, and background context. To detect and infer such information we need a combination of Vision modules, Reasoning modules, and background knowledge.

In Figure 2, we present our architecture that explicitly models the desired interactions between vision and reasoning modules. The core of the architecture consists of the following modules: i) Visual Detection, ii) Knowledge Base, and iii) Logical Reasoning. The complete system also provides interfaces to: i) Sentence Generation and ii) Question-Answering modules.

Visual Detection: The “Visual Detection” module should be able to obtain the following basic quantities: i) Objects and Regions, such as man, basketball, wooden floor etc.; ii) Scenes, i.e., scene classes such as indoors, stadium; iii) Relations including spatial ones between two objects or an object and the scene, for example *man holding basketball*, *man standing on floor*; iv) Properties, i.e., different attributes of objects and regions such as size, height, color of objects; color, shape of region; v) Attention: In addition, in an active vision setting ([32]), the visual detection module is also expected to interact with the reasoning

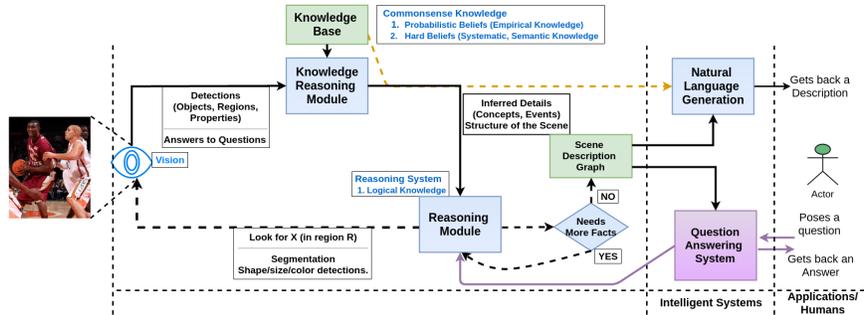


Figure 2: An architecture for Deep Image Understanding. (The Knowledge Reasoning Module is a part of the Reasoning Module; it is shown separately to clearly outline the interactions).

module and hence, the former should have control over “which detector to fire over which region of the image”.

Ideally, this detection module should consist of a large set of object and scene detection classifiers, relationship detection classifiers, attribute (color, shape, size) and relative attribute classifiers ([59]); and detection and Image Segmentation modules.

Knowledge Base: Different forms of background knowledge are necessary to reason about the quantities detected and recognized by the Vision module. In this architecture, we need commonsense knowledge⁵ to answer questions pertaining to: i) the probable actions that the detected objects are participating in; ii) the past and future actions that could be causally connected to such actions; iii) ontological information about the detected scenes; iv) and lastly, a holistic background (ontological, spatial, commonsense, etc.) knowledge pertaining to every object of the scene in view.

Reasoning System: A logical reasoning system can represent the logical

⁵The type of commonsense needed here is similar to Semantic Knowledge according to definitions in Psychology. By definition, semantic Knowledge is “general knowledge about the world, including concepts, facts and beliefs (e.g., that a lemon is normally yellow and sour or that Paris is in France)” ([60]).

knowledge using a set of rules and should be able to perform deductive, inductive and abductive reasoning considering both probabilistic and hard beliefs. Traditional formalisms such as Answer Set Programming are powerful representation languages; however the usage of hard rules and facts limits the usage in many real-world applications. Probabilistic reasoning is necessary to deal with the uncertainty and incompleteness of the knowledge and the visual detections. Hence, we can use a probabilistic adaptation of such logical systems in which rules and facts are not constrained to be binary and which supports the agent’s “incomplete” knowledge about the world. Further implementations of this architecture might adopt languages such as Probabilistic Soft Logic ([61]), and Markov Logic Networks ([62]).

In many current end-to-end implementations (such as, captioning and VQA), the visual detection module is modeled using a pre-trained Convolutional Neural Network, and the knowledge of words is encoded using Word Embeddings. Understanding and reasoning of the language construct is modeled using a sequential network, which is a variant of Recurrent Neural Networks. The interaction between these modules is often modeled using attention mechanisms. These models are then tuned in a combined fashion for specific applications. However, current systems: i) do not explicitly model commonsense knowledge, which is reflected in their performance on questions requiring commonsense; ii) do not model the knowledge needed to rectify detections in case of partially or fully occluded objects (Figure 1(a)), which affects both VQA and captioning tasks; and iii) do not provide a way to identify the main cause in case of wrong answers. In this work, we provide an implementation of a modular architecture, that facilitates explainability and produces with reasonable accuracy an intermediate semantic structure of the scene.

4. Predicting Intermediate Scene Description Graphs

In this work, we develop an implementation of the above architecture to predict Scene Description Graphs from static images. To map an image to

210 a Scene Description Graph, we first robustly define the meaningful regions of
images that capture relevant semantics. Let us assume that the fundamental
semantic components of an image (denoted as \mathcal{F}) are the objects⁶ and their
observable attributes (location, shape, size, color, contour etc.), regions and their
observable attributes, and actions. To avoid further complexity, we consider only
215 those images, in which at least one fundamental semantic component ($f \in \mathcal{F}$)
can be detected (by an ideal detector). In a scene, we group these components
further to form observable (that can be seen) and inferable components (that
can be understood).

Observed Scene Constituents (OSC) are descriptions of objects, actions
220 or regions (described in phrases or words) that can be directly grounded in the
image⁷. In a phrase, individual words can identify an object, group of objects,
their observable attributes, regions or actions. For example: *person wearing
shorts, person skateboarding, tall person, people playing* etc. are all Observed
Scene Constituents.

225 **Inferred Scene Constituents (ISC)** are concepts (activities, context,
higher-level events) that cannot be directly grounded in the image, but can be
inferred. For example, *open space and bright day* are ISCs.

Based on the above definitions, a **Scene** then represents one (or more) ac-
tions, involving (one or more) objects; and spatial relationships among objects
230 and regions. The action(s) together make up a natural event which can be de-
scribed by sentence(s), such as: *a person is lying on a bench, in a park; a person*

⁶Objects can consist of visible, partly visible or occluded objects. If the object *person* is detected, occluded objects like organs in a body, can inferred to be present using commonsense Knowledge Bases such as ConceptNet.

⁷To determine if a word or a phrase is a scene constituent or not, it will be helpful to ask ourselves the question: “can we mark a region or set of regions in the image that represents the meaning of this word or phrase completely?”. If we can and the word or phrase is not an object, action or region; then the word or phrase is a scene constituent. Here, we can assume that the bounding box for an action will be the union of the bounding boxes of its participant objects.

is being evicted; a bank is being robbed.

We can also interpret the above definitions as mapping meaningful components of images to meaningful components of text⁸. The fundamental components (\mathcal{F}) can be roughly mapped to words with the following parts-of-speech (POS) tags: concrete nouns (object and scene classes), a subset of verbs (actions), adjectives (object attributes), adverbs (action attributes) and prepositions (relations) [20]. We can describe the Observed and the Inferred Scene Constituents using phrases. We can then describe a natural image (representing a combination of some the above components) using sentence (s).

4.1. Visual Detection

We use deep object recognition, deep scene (category) recognition and deep Observed Scene Constituent recognition as the components of the Visual Detection module (to primarily detect the semantic components).

Object Recognition: For deep object recognition, we use the trained bottom-up region proposals and convolutional neural networks (CNN) object detection method from [63]. It considers 200 common object classes (denoted as \mathcal{N}). and it is trained on the ILSVRC dataset.

Scene Recognition: For deep scene (category) recognition, we use the trained CNN scene classification method from [64]. The classification model is trained on 205 scene categories (denoted as \mathcal{S}).

Constituent Recognition: For deep observed scene constituent (OSC) recognition, we augment the Flickr 8K image dataset with human annotations of constituents using Amazon Mechanical Turks. We specifically ask the annotators to annotate not only objects, but also what the objects are doing and

⁸[24]’s work (and other Neural approaches) essentially uses the neural networks to learn a similar mapping between any region of an image to meaningful chunks of text. But this method does not utilize the richness of the structure of text and images, and the mapping is also independent of commonsense knowledge (which should prevent an intelligent system to learn wrong mappings in adversarial situations).

about the properties of objects⁹. We allow the labelers to use free-form text for describing constituents to reduce the annotation effort. We obtain a standardized set of constituents by performing stop-words removal, parts-of-speech processing to retain nouns, adjectives and verbs. We use the top 1000 most frequent phrases (denoted as \mathcal{C}). Some of the top phrases are *dog run*, *dog play*, *kid play*, *person wear shorts* etc. We post-process the annotations for each training image in a similar manner, and consider the phrases as labels if they are among the 1000 top constituents. For each image, we then use the pre-trained CNN model from [6] to extract a 4096 dimensional feature vector (using [65]). We then trained a multi-label SVM to recognize constituents using these deep features.

The output from the detection system consists of object ($P_r(n|x)$), scene ($P_r(s|x)$) and constituent ($P_r(c|x)$) detection scores for the top 5 objects, top 5 scene categories, and top 10 constituents; for each image $x \in I$.

4.2. Constructing SDGs from Detections

We first pre-process the annotations and information from the training images to capture the required commonsense knowledge, which we refer to as “Knowledge Extraction and Storage”. Then we use a rule-based reasoning algorithm to infer a knowledge structure.

4.2.1. Pre-processing Phase

Inferred Scene Constituents often have correlations with scene categories (such as *audience* in *stadium*). In this phase, we collect a mapping (\mathcal{S}_M) between scene categories and ISCs; and learn a prior belief ($P(isc|scene)$) for each ISC in a scene. For example, for the scene class *airport-terminal*, we add *{waiting room, big glass view, travelers}* as the list of probable ISCs; and learn the priors 0.7, 0.7 and 0.9 respectively for ISCs.

We use scene category detection tuples, $([c_i, Pr(c_i|x)]_{i=1}^5)$ for training images ($x \in I$), which we denote as \mathcal{S}_T . For detections, we use the deep Scene (category)

⁹We make this dataset publicly available at <http://bit.ly/1MMN1wZ>.

Recognition module to detect the top 5 scene categories from each training
 285 image. We denote the human annotations for all training images as \mathcal{A}_{tr} .

4.2.2. Knowledge Extraction and Storage

To capture the commonsense and probabilistic knowledge about the domain,
 we created a **Knowledge Base** \mathcal{K}_b and a **Bayesian Network** \mathcal{B}_n using the
 pre-processed data ($\langle \mathcal{S}_M, \mathcal{S}_T, \mathcal{A}_{tr} \rangle$). To extract knowledge from the annotations,
 290 we extensively use a semantic parser, called K-parser ([66]).

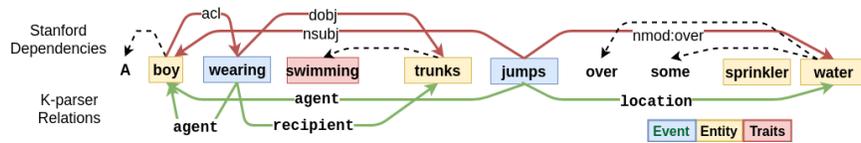


Figure 3: An example sentence with Stanford Dependency relations and transformed
 K-parser relations. Only important Stanford Dependencies and K-parser relations are
 shown. K-parser also adds semantic roles and superclass information for the Entities
 (not shown in the figure).

K-Parser: K-parser (`kparser.org`) is a semantic parser that extracts an
 Entity-Event based representation from a sentence, adding additional semantic
 knowledge. For a sentence such as “A boy wearing swimming trunks jumps over
 some sprinkler water in a backyard”, the K-parser extracts the Events (actions
 295 and linking verbs) `wear`, `jump`, and their participant Entities (concrete nouns)
`boy` and `trunks`, `boy` and `water` respectively as a set of Entity and Event-
 nodes connected by meaningful relations (see Figure 3). It also extracts Traits
 (attributes) `swimming`, `sprinkler` corresponding to the entities. Internally, K-
 parser uses the Stanford Parser [67] to get the syntactic dependency graph from
 300 a sentence. The K-parser then uses a rule-based mapping algorithm to map
 these dependency relations to the set of KM-Relations ([3]) and some newly
 created ones (see <http://bit.ly/1Wd8nGa>). Some relevant properties of
 the final semantic representation are: i) it is an acyclic graphical representation
 of English text, ii) it follows a rich ontology ([3]) to represent semantic relations

305 (Event-Event relations such as *causes*, *caused_by*, Event-Entity relations
such as *agent*, and Entity-Entity relations such as *related_to*); iii) it has
two levels of conceptual class information for words; iv) it accumulates semantic
roles of Entities based on PropBank framesets; and v) it has other features
such as Co-reference resolution, Word Sense Disambiguation and Named Entity
310 Tagging ¹⁰.

Knowledge Base: The knowledge-base is mainly a knowledge-graph (\mathcal{G}),
which is a collection of *word1-relation-word2* triplets, where *word1* and
word2 can be Event (actions, linking-verbs present in \mathcal{A}_{tr}), Entity (from \mathcal{N})
or a Trait (adjectives, qualitative-nouns from \mathcal{A}_{tr} or WordNet-superclass of a
315 word). The *relation* comes from a closed set of semantic relations from
KM-Ontology¹¹. The graph contains the knowledge of i) all possible Entities
(concrete nouns) participating in Events (actions and linking verbs), and ii)
possible traits (properties, such as color, semantic role-labels) that the Entities
have. Figure 4 depicts a snapshot of \mathcal{G} .

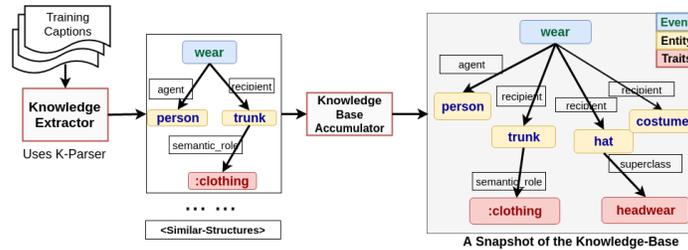


Figure 4: Knowledge Base Creation using A Semantic Parser.

320 As shown in Figure 4, we use K-parser for knowledge extraction from each
sentence of the Image Annotations. We first reconcile the Entities in the
K-parser output graph with corresponding nouns in \mathcal{N} , using WordNet sim-

¹⁰For more details, please see [66].

¹¹*agent*, *recipient*, *location*, *origin*, *object*, *destination*,
semantic_role, *superclass* are some of the important relations in context of this
work. Extensive list can be found in kparser.org.

ilarities. Then, the graphs are merged based on overlapping Events. Entities connected by agent, recipient, object, location, origin, and destination relations to an Event, are retained. Causal connections between Events are also retained. All Traits connected to the Entities are retained as well. The merged knowledge-graph is stored as \mathcal{G} . We store the unique semantic parses of captions in \mathcal{C} to provide contextual knowledge such as (x-r-y) occurs along-with (y-superclass-z) in some context $C \in \mathcal{C}$. We formally represent our Knowledge Base as $\mathcal{K}_b = \langle \mathcal{G}, \mathcal{C} \rangle$.

The Bayesian Network (\mathcal{B}_n): Objects and scene constituents often co-occur in a scene. Authors in [68] use such co-occurrence to classify scenes. In this work, we capture the knowledge of naturally co-occurring objects (\mathcal{N}), their siblings from WordNet (\mathcal{N}_S) and ISCs (\mathcal{C}_{I_s}), by learning a Bayesian Network that represents the dependencies among them. We create the training data \mathcal{D} which is a collection of tuples T (where $T = [t_i]_{i=1}^N$ and $N = |\mathcal{N}| + |\mathcal{N}_S| + |\mathcal{C}_{I_s}|$). Each term t_i is binary and is set to 1 if the i^{th} object (or ISC) occurs in the tuple. We use the Tabu Search algorithm to learn the structure and then we populate the Conditional Probability Tables using the R-bnlearn package ([69]).

To create \mathcal{D} , we process the annotations for each training image (\mathcal{A}_{tr}) to automatically detect Entities and ISCs. We parse the sentences using K-parser and extract Entities. We match these Entities with objects in ($\mathcal{N} \cup \mathcal{N}_S$) based on base-forms and synonyms of the words. Some of the ISCs are detected using rule-based techniques, for e.g., we detect the edges `edge(wear, agent, person)` and `edge(wear, recipient, shorts)` in the K-parser semantic graph for ISC “*people wearing shorts*”. To detect ISCs seldom mentioned in annotations, we detect the top scene class and we look-up all ISCs of the scene category.

4.2.3. Inference through Knowledge and Reasoning

Prior to Neural approaches to image captioning, researchers from the Vision and Language community used keyword-based image annotations to predict the subjects, objects and scenes from images, and they predicted correlated verbs

or prepositions using learned language models ([47]). Inspired by these approaches, we use the commonsense knowledge $\langle \mathcal{K}_b, \mathcal{B}_n, \mathcal{S}_M \rangle$ and the detections $\langle P_r(n|x), P_r(s|x), P_r(c|x) \rangle$ for an image ($x \in I$) to construct the different components of the SDG (a labeled graph) in the following way. We use Entities to denote objects, and Events to denote actions (and linking verbs). All the notations and terms used in this paper are summarized in Figure 5.

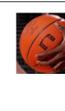
			Objects	Entities
Red Shirt, Male, Player	White Shirt, Male, Player	Orange Basketball	Attributes	Traits
			Regions	Entities
Shiny Floor, Wooden Floor			Attributes	Traits
holding (Basketball), standing (on floor), playing, (people) watching			Actions/Linking Verbs	Events
person1 holding basketball, person1 and person2 playing basketball			-	Observed SC
person2 blocking person1, person1 and person2 different team, basketball game			-	Inferred SC

Figure 5: Summary of notations used in the paper. The second column shows the terminology popularly used in Computer Vision and the third column shows the terms introduced in this work (some of which are adopted from [66]).

I. Additional Entities and Events (from OSCs): We extract Entities (nouns) and Events (verbs) from the top 10 constituents (based on $P_r(c|x)$) and add to the set of detections. For example, from the constituent *person wearing sweatshirt* we get an Event *wear* with two Entities *person* and *sweatshirt*.

II. Inferred Scene Constituents: We look-up the ISCs for the top 5 detected scenes (based on $P_r(s|x)$) from \mathcal{S}_M , and call that collection \hat{C} . Initially, $C_{inf} = \phi$, and $\mathcal{O}_x = \{n | P_r(n|x) > \alpha_n\}$. We calculate

$$C_{max} = \arg \max_{c \in \hat{C}} P(s|C_{inf}, \mathcal{O}_x) \quad (1)$$

and add C_{max} to C_{inf} . We iterate while the entropy E keeps decreasing (or

while number-of-iterations is less than T^{12}). The entropy is calculated as:

$$E = \sum_{c \in \hat{C}} \{-P(c|C_{inf}, \mathcal{O}_x) * \log P(c|C_{inf}, \mathcal{O}_x)\} \quad (2)$$

365 The conditional probabilities are calculated using \mathcal{B}_n .

III. Noisy Objects: Next, we rectify the low-scoring Entities based on \mathcal{O}_x and C_{inf} . For each low-scoring Entity, we get all its siblings, i.e., we get all the children of its hypernyms from WordNet. For example, if *bathing cap* is assigned a low score, the assigned superclass is *cap* and its children are *baseball cap*, *ski cap* etc. We calculate the following

$$o_{max} = \arg \max_{o \in siblings} P(o|C_{inf}, \mathcal{O}_x) \quad (3)$$

and then add o_{max} to the high-scoring Entities list (\mathcal{O}_x).

IV. Inferring Events: Given the Entities (\mathcal{O}_x), we first find connecting Events between each pair of Entities. To **logically** find a co-occurring Event for a pair of Entities ($e_1, e_2 \in \mathcal{O}_x$), we consider the Event-nodes on the shortest path from one Entity to another in the graph \mathcal{G} . For example, consider the 370 Entities *person* and *swimming trunks* (corresponds to the vertex *trunk* in \mathcal{K}_b). We get Events such as sniff, climb, wear etc., i.e., some corresponding to tree-trunk and others to swimming-trunks. We denote the set of connected Entities by \mathcal{O}_{ev} and set of Events by \mathcal{E}_v .

375 For filtering spurious Events, we use the semantics in K-parser edge labels and the superclass (type) of the Entities from \mathcal{K}_b . We retain Events only if they are connected to the Entities using compatible edge-pairs in \mathcal{G} . Compatible edge-pairs are: (agent-recipient), (agent-location), (agent-object). For example, (agent, recipient) is a compatible pair 380 and only an animate Entity can be an agent. Thus, the Event *wear* is retained with respect to Entities *person* and *trunk*. To filter Events such as *climb*, we use

¹²The hyper-parameters (T, α_h) are set based on performance on validation data. In our experiments, we have used the values 5, 0.5 respectively.

the superclasses of the Entities and the set of Scenes \mathcal{C} . We retain only those Events that are connected to Entities from the same pair of classes as e_1, e_2 , in at least one scene in \mathcal{C} .

385 **V. Inferring Scenes:** Given the filtered Events and Entities (\mathcal{O}_{ev}), we consider a Scene in \mathcal{C} as candidate if all edges from a detected valid Event, are present in it. Next, we weight each candidate Scene (\mathcal{C}_{cand}) using the remaining Entities in ($\mathcal{O}_x \setminus \mathcal{O}_{ev}$) and ISCs (\mathcal{C}_{inf}); i.e., increase a counter if an Entity or ISC occurs in the graph (\mathcal{C}_{cand}). We also calculate a joint confidence-score
390 for each scene based on the $P_r(n|x), P_r(s|x), P_r(c|x)$ values of the object, scene category and constituents (OSC) present in the Scene. Based on the counters and the joint confidence-score, we rank the Scenes.

VI. SDG Construction: If we do not find a suitable Scene in \mathcal{C} , we construct an SDG using the following rules: i) add edge (`scene, component,`
395 `s`) for all ISC s in \mathcal{C}_{inf} ; ii) add edge (`event, location, scene`) for the top detected Events; iii) add all compatible edges related to the Events in \mathcal{E}_v such as edge (`wear, agent, person`) and edge (`wear, recipient, trunk`); and iv) for all Entities o_{im} in ($\mathcal{O}_x \setminus \mathcal{O}_{ev}$): if it is an animate Entity, add edge (`o_{im}, location, scene`); Otherwise, find the shortest path from o_{im} to the top
400 detected Event in the \mathcal{K}_b and add the edges on the path to the SDG.

5. Experiments and Results

The above approach presents two hypotheses that require empirical evaluation: i) SDGs carry detailed information about images (thoroughness); ii) SDGs carry relevant semantic information about the salient aspects of the image (relevance). Collecting groundtruth Scene Description Graphs are difficult,
405 time-consuming, and expensive. Lastly, guaranteeing the reliability of the crowdsourcing of such complex annotations is also difficult. Instead, here we first generate captions from these SDGs and use two end-to-end tasks (Image Retrieval and Caption Generation) to support the hypotheses presented in this
410 paper. We use the Image Retrieval task that directly use the generated SDGs

from images and semantic parses from text (used as query). This task tests the discriminative (image-specific) information encoded by the generated SDGs. Caption generation is a task of generating relevant descriptive sentence(s) from an image; relevance and thoroughness being the two distinct criteria, with which
415 the quality of captions can be judged. Hence, we use this task to test the relevance and thoroughness of the generated SDGs.

We adopted two experiments to evaluate the generated SDGs: i) qualitative evaluation of generated sentences and ii) image-sentence alignment evaluation. We compare our results with [24] as it was one of the recent (and among the first)
420 neural approaches that produced best results over all the previous works. We also compare our results with another more recent Neural Captioning method by Vinyals et. al. [38] (appeared in IEEE TPAMI 2016) which reported improved quality of captions in comparison to [24]. This method uses the latest Inception-V3 architecture to process images and an Long-Short Term Memory (LSTM)
425 model to generate captions. We first describe the testbed and the procedure for generating captions from the competing methods.

Testbed: In this paper, we use three image data sets, popularly referred to as Flickr 8k, Flickr 30k and MS-COCO datasets [41]. These three datasets have 8092, 31783 and more than 160K images respectively. Every image from these
430 datasets is annotated with 5 sentences describing the image. For all datasets, we used the train-test splits from [24] and the 4000 testing images (1000 each from Flickr 8k and Flickr 30k and 2000 from MS-COCO validation set) serve as the testing bed for our experiments.

Generating Captions: For our system, we generate sentences from SDGs
435 using SimpleNLG ([70]). For example, for the edges `edge(wear, agent, person)` and `edge(wear, recipient, shorts)`, we will generate “*a person is wearing shorts*”. Based on the edge-labels (labels from KM-ontology) we populate the verb, subject, object, prepositions and adjectives (including quantitative¹³) of sentences using simple rules. The other rules used are: i)

¹³For high-scoring detections, we consider the spatial information from the bounding-boxes.

440 edge($_, location, A$) is mapped to “*in the A*”, ii) edge($_, origin, B$) is
mapped to “*from the B*”; and iii) all edges of the form edge($scene, component, B$)
is converted to a sentence based on the template “the scene contains B and ...”.
For BRNN [24], we use the implementation provided by the authors to train
and generate sentences from an image. To generate captions using [38], we use
445 the code provided by the authors¹⁴. We initialize the network with the provided
pre-trained Inception-V3 checkpoint, and train the model for 2-million steps.

Amazon Mechanical Turk (AMT) Evaluation of Generated Sentences: Since image description generation is innately a creative process, a
metric is created by asking humans to evaluate these sentences. The evaluation
450 metrics: Relevance and Thoroughness, are therefore, proposed as empirical mea-
sures. Relevance measures how much the description conveys the image content
and Thoroughness quantifies how much of the image content is conveyed by the
description. We engaged the services of AMT to judge the generated descrip-
tions based on a discrete scale ranging from 1–5 (low relevance/thoroughness to
455 high relevance/thoroughness)¹⁵. The average of the scores and their deviation
are summarized in Table 1. For comparison, we asked the AMTs to also judge
one gold-standard description and the output from [24].

A Supplementary AMT study: It is often considered a good practice to
perform multiple independent AMT studies. In Table 2, we provide the results
460 of an independent AMT evaluation (using similar instructions as above). For
this study we compare the sentences generated by our method, a ground-truth
sentence, the output from [24] and [38]. As previously stated, we use the 2000

For N such detections of an object obj , we generate sentences like N *obj*’s are in the scene.

¹⁴<https://github.com/tensorflow/models/tree/master/im2txt>

¹⁵We provide the following instructions to the Turkers. Relevance: the description has
no relevance (1)/ only weak relevance (2)/ some relevance (3)/ relates closely (4)/ relates
perfectly (5) to the image. Thoroughness: the description covers nothing (1)/ covers minor
aspects (2)/ covers some aspects (3)/ covers many aspects (4)/ covers almost every aspect (5)
of the image.

The human evaluations dataset is available in <http://bit.ly/1MMN1wZ>.

Experiment	[24] BRNN	Our Method	Gold Standard
R \pm D(8k)	2.08 \pm 1.35	2.82 \pm 1.56	4.69 \pm 0.78
T \pm D(8k)	2.24 \pm 1.33	2.62 \pm 1.42	4.32 \pm 0.99
R \pm D(30k)	1.93 \pm 1.32	2.43 \pm 1.42	4.78 \pm 0.61
T \pm D(30k)	2.17 \pm 1.34	2.49 \pm 1.42	4.52 \pm 0.93
R \pm D(COCO)	2.69 \pm 1.49	2.14 \pm 1.29	4.71 \pm 0.67
T \pm D(COCO)	2.55 \pm 1.41	2.06 \pm 1.24	4.37 \pm 0.92

Table 1: Sentence generation relevance (R) and thoroughness (T) human evaluation results with gold standard and [24] on Flickr 8k, 30k test images and COCO validation images. D: Standard Deviation.

MS-COCO validation images to report the results.

Experiment	[38] ShowAndTell	[24] BRNN	Our Method	Gold Standard
R \pm D(COCO)	3.59 \pm 1.36	3.2 \pm 1.3	3.11 \pm 1.39	3.9 \pm 1.16
T \pm D(COCO)	3.16 \pm 1.46	3 \pm 1.46	2.64 \pm 1.39	3.9 \pm 1.37

Table 2: Sentence generation relevance (R) and thoroughness (T) human evaluation results with gold standard, [24] and [38] on COCO validation images. D: Standard Deviation.

The work in [38] is one of the latest proposed methods using a state-of-
465 the-art variant of CNN-RNN architecture for Image Captioning. *Though the generated sentences from the Neural approaches have a higher score, this study shows that our method performs reasonably well, even though it is not tuned for a specific dataset.* We also show some qualitative examples on MS-COCO by the three competing systems in Fig. 6.

Automatic Caption Evaluation Results: In this section, we supplement
470 our experiments with evaluation results using BLEU ([71]) and Meteor ([72]) scores. The BLEU scores are calculated using the original PERL script¹⁶ provided for statistical machine translation tasks. The Meteor scores are calculated using the instructions provided by the authors in [72]¹⁷. We provide detailed
475 insights about Table 1, 2 and 3 in the Analysis section.

Image-Sentence Alignment Evaluation: We evaluate the image-sentence

¹⁶BLEU Evaluation Perl Script.

¹⁷Meteor 1.5.



Figure 6: We provide some comparative captions generated by our system (in yellow box), by BRNN [24] (top blue box), by ShowAndTell [38] (in pink box). The groundtruth captions are given in lower green boxes. Interesting human annotations (partially or fully incorrect) are marked using question or cross mark.

Experiment	Flickr-8k				Flickr-30k				COCO-2014				
	B-1	B-2	B-3	B-4	B-1	B-2	B-3	B-4	B-1	B-2	B-3	B-4	M
[38] ShowAndTell	63	41	27	-	66.3	42.3	27.7	18.3	66.6	46.1	32.9	24.6	-
[24] BRNN	57.5	38.3	24.5	16.0	57.3	36.9	24.0	15.7	62.5	45.0	32.1	23.0	19.5
Our Method	30.0	12.6	9.5	5.0	25.9	12.5	10.0	4.0	22.3	13.4	11.0	5.0	10.0

Table 3: Sentence generation BLEU, Meteor Scores in comparison with some of the Existing Neural Architectures [24] and [38] on Flickr-8k (test), Flickr30k (test) and MS-COCO validation images. **B-n** denotes BLEU scores that uses upto n-grams. Meteor scores are only reported for MS-COCO as followed by other works. The scores for Neural captioning systems are as reported in [24].

alignment quality using ranking experiments. We withhold the testing images and use the generated sentences as queries. We process the textual query and construct $\mathcal{G}_q = (V_q, E_q)$ using K-parser. For each image, we take the generated SDG $\mathcal{G}_x = (V_i, E_i)$ and calculate similarity between the SDG and the query using the formula:

$$\begin{aligned}
Sim(\mathcal{G}_q, \mathcal{G}_x) &= \left(\sum_{v_q \in V_q} \max_{v_i \in V_i} sim(v_q, v_i) \right) / |V_q| \\
sim(v_q, v_i) &= 0.5 * \left(wnsim(label(v_q), label(v_i)) \right. \\
&\quad \left. + Jaccard(neighbors(v_q), neighbors(v_i)) \right).
\end{aligned}$$

Vertex-similarity is calculated based on word-meaning similarity and neighbor similarity. Here $wnsim(.,.)$ is Lin Similarity [73] between two words and $Jaccard(.,.)$ is the standard Jaccard coefficient similarity. Based on the above
480 measure, we provide the image retrieval results compared with results from [24] in Table 4. Additionally, we provide the results of the Show-and-Tell method[38] for Flickr8k and Flickr30k, as provided by the authors. Interestingly, our results for image search is better compared to this recent work for Flickr30k dataset.

5.1. Analysis

485 In this Section, we analyze several aspects of the conducted experiments, and the results, and present more insights on the added aspect of external commonsense knowledge and interpretability.

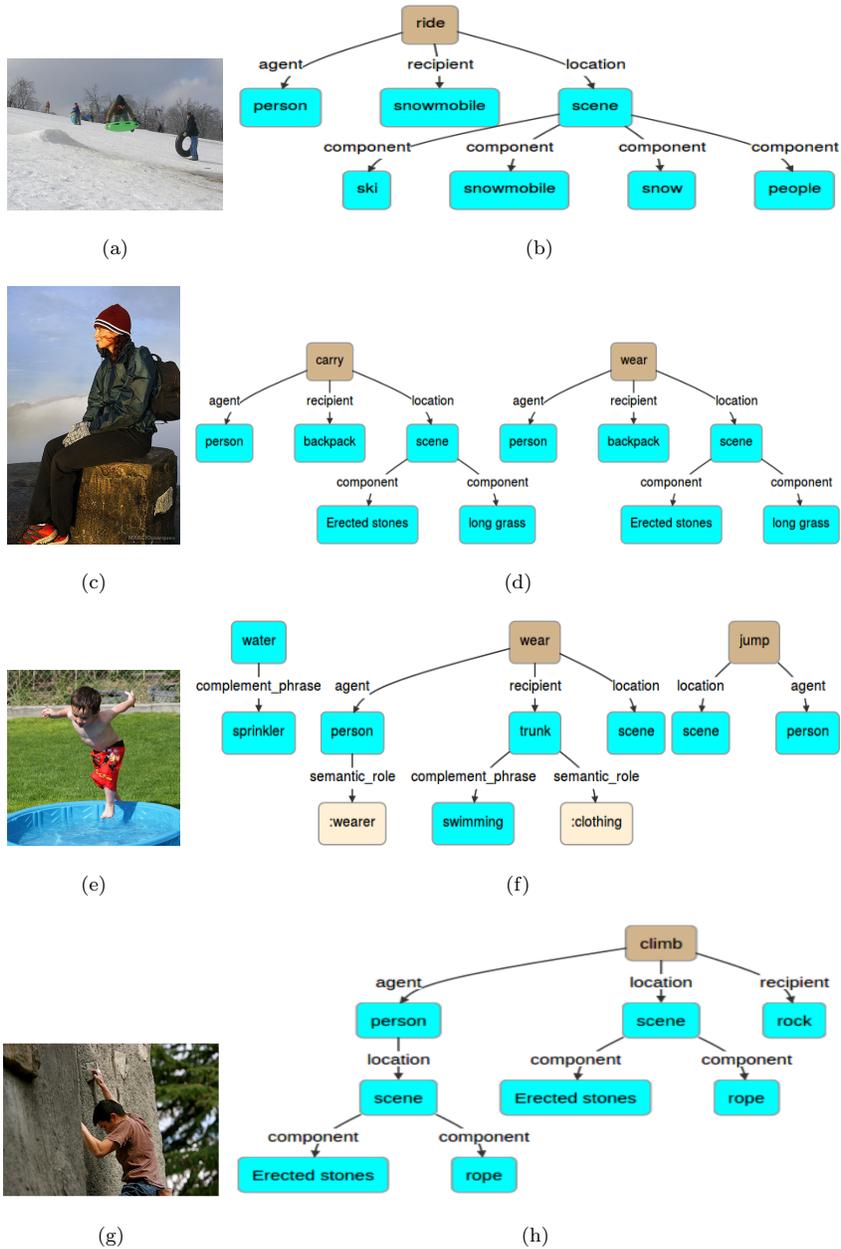


Figure 7: The SDGs in (b), (d), (f) and (h) corresponds to images (a), (c), (e) and (g) respectively. More examples are at <http://bit.ly/1NJycKO>.

	Flickr8k			
Model	R@1	R@5	R@10	Med r
[24] BRNN	11.8	32.1	44.7	12.4
[38] ShowAndTell	19	-	64	5.0
Our Method-SDG	18.1	39.0	50.0	10.5
	Flickr30k			
[24] BRNN	15.2	37.7	50.5	9.2
[38] ShowAndTell	17	-	57	7.0
Our Method-SDG	26.5	48.7	59.4	6.0
	MS-COCO			
[24] BRNN (1k)	20.9	52.8	69.2	4.0
Our Method-SDG (1k)	19.3	35.5	49.0	11.0
Our Method-SDG (2k)	15.4	32.5	42.2	17.0

Table 4: Image-Search Results: We report the recall@K (for $K = 1, 5$ and 10) and Med r (Median Rank) metric for Flickr8k, 30k and COCO datasets. For COCO, we experimented on first 1000 (1k) and random 2000 (2k) validation images.

Comparable Systems: There are other works in Image Retrieval ([74]) and Caption Generation ([75]) that achieve better results than shown in Table 1 and 2. However, the motivation behind our work was to propose a meaningful representation that provides a seamless interface between image and text and, a framework that uses a combination of vision and reasoning to construct such structures. We believe that from a motivational standpoint, our work is not directly comparable with such systems. Authors in [52] propose a semantic scene graph generation from images. However, to apply symbol-level reasoning on semantic structures, it is important that the relations come from a well-defined closed set of meaningful labels, whereas the relations used in [52] are open-ended text. To that end, other related works [76, 37, 77] have proposed a bounded set of spatial relations between detected objects and regions (grounded in the image) to represent a scene. However, we compare our results with two popular recent neural captioning approaches [24] and [38].

Human AMT and Automatic Caption Evaluation Results: In Tables 1 and 2, we present the human evaluation results of the generated captions from our system and two competing systems. We have conducted these studies using Amazon Mechanical Turk as it is a well-accepted crowdsourcing platform in the community, and studies [78] show that this platform is less noisy, error-prone

and biased than other methods. However, the means for all the systems are higher in Table 2 compared to Table 1. This is expected as, human evaluations are inherently subjective, which can cause exact values from different studies to differ. We note that the two independent studies are consistent in the relative ranking (with [24] ranking above ours). In Table 3, we present the automatic evaluation results using BLEU and Meteor scores. According to the results, our method fares worse in comparison to the other systems. Looking closely, for the image in Figure 6(a), our generated sentence is scored 11.5, 0.0, 0.0, 0.0 using BLEU-1 to 4 metric; while a less informative sentence from the Neural architecture (BRNN) is scored 40.0, 0.0, 0.0, 0.0. In an even worse comparison, for the image in 6(d), both generated sentences are correct in meaning. Yet, the sentence from BRNN is rated 90.0, 83.7, 80.7, 78.3, while the caption from our system is rated 20.0, 0.0, 0.0, 0.0. Additionally for Figure 6(d), there is no evidence that the *person* in the image is a *man* or a *woman*. In that sense, the BLEU metric overestimates the correctness of the caption from BRNN. In summary, the larger scores are expected as the Neural Captioning systems learn the language construct and the image to language mapping from training captions. As the train, test and validation data come from the same distribution, the vocabulary and the language construct for the test images tend to be similar. In comparison, in our system the sentences are generated using few fixed templates and the vocabulary is not restricted to the words in the training captions, and more importantly the sentences are not directly optimized to be *syntactically* similar to the training captions. For example, in many cases we use a collection of short sentences to convey similar information; and many sentences begin with *the scene contains*. As the automatic metrics solely rely on the vocabulary and language construct of the ground-truth captions, these metrics heavily penalize these template-based sentences. This noisiness is well-known in the community¹⁸ and more automatic caption evaluation metrics are proposed. However,

¹⁸The work in [79] shows the different automatic image captioning metrics have very little correlation with human judgment. Notably, this work uses our COMPOSITE dataset (captions

535 the task of captioning an image is a subjective task. Clearly, lower scores from
automatic metrics that directly compare with ground-truth captions do not re-
flect that the performing system is worse, as the generated caption can match
some other caption written by a different Turker than the Turkers who anno-
tated the image. This is why we perform human evaluations of thoroughness
540 and relevance of the captions. It allows us to test how correctly and thoroughly
the generated captions describe an image. As also discussed in a recent survey
[40], human evaluation measures like the one adopted in our methodology, have
many advantages, and prior to Neural approaches the majority of captioning
systems adopted such measures (cf. Table 3 of [40]).

545 **Impact of Knowledge Base and Bayes Net:** The Knowledge-Base and
the Bayes Net encode important background knowledge which enrich the SDGs
and rectify noisy information from visual detection modules. The \mathcal{C} (in \mathcal{K}_b) and
Bayes Net encodes contextual knowledge, i.e. which *type* of entities and events,
or entities and ISCs co-occur in common contexts. In Figure 6, the information
550 in sentences “the scene contains ...” are obtained from the Bayes Net.
Additionally, the Knowledge base encodes events or actions that occur in context
of entities, for example all verbs in Figure 6 is inferred by the Knowledge Base
based on the detected entities.

Interpretability: One of the major disadvantages of many end-to-end
555 learning approaches (especially, the current neural network based approaches)
is the lack of model interpretability or explicit explanations. This is one of the
fundamental motivations behind our proposed intermediate knowledge struc-
ture and our architecture. Referring to Figure 7(g), the initial top object
and scene detections are: $\{person, backpack, artichoke, hat\}$ with a wide brim};
560 $\{wheat_field, cemetery, fountain, corn_field\}$ etc. The constituent detections
are: $\{person\ sitting\ on\ stone, person\ wearing\ red\ shoes, person\ wearing\ gloves\}$.
An SDG combined with our architecture can facilitate explainability in the fol-
lowing ways: i) why the SDG in 7(g) contains *person* and *backpack*? They are

from SDG, [24] and AMT scores) to show the above result.

detected by object classifiers with high probability; ii) why the SDG in 7(g)
565 contains *erected stone*? Because scene categories such as *cemetery* co-occurs
with erected stone (knowledge from \mathcal{S}_M); iii) why the SDG in 7(g) has verb
carry, wear? Because it co-occurs with the entities (person, backpack) (knowl-
edge from \mathcal{K}_b). In short, explanations for the components in the SDG in 7(g)
can be tracked back to one of the knowledge sources in $(\langle \mathcal{K}_b, \mathcal{B}_n, \mathcal{S}_M \rangle)$ or the
570 Visual Detection Module.

5.2. Question-Answering (QA) Case Studies

Using SDGs to answer a question requires development of sophisticated
probabilistic logical mechanism (or neural reasoning mechanisms) that can sift
through the noise in the generated SDG, understand the natural language ques-
575 tion and give an answer. Such mechanisms require further research and devel-
opment. Instead, in this section, we motivate the use of SDGs by providing a
few examples of a Question-Answering system (with a simple reasoning module)
that can be built based on the generated Scene Description Graphs.

For the image in Figure 8(a), the Scene Description Graph is represented as
580 a set of has-tuples. Relying on the advantage of using meaningful relations from
KM-ontology, we can use these as inputs to an Answer Set Program ([80]). If
we pose the question that “Is someone drinking from the fountain?” in ASP (as
shown in the figure), we can execute the program in Clingo-3 and we get the
answer as *yes_fountain(person1)*.

585 For the second image in Figure 8(b), we pose the question “is someone
playing tennis”. In this case, we need additional background knowledge such
as “if someone is holding or swinging a tennis racket, then the game might be
tennis” to detect the game of tennis. Again, the question is posed in ASP, using
the generated SDG, we obtain the boolean value of *tennis_detector* as *True*.
590 Though the above question is written in ASP without any probabilistic weight,
one can rewrite the rules in Probabilistic Soft Logic ([81]) assigning a weight to
the rule for “tennis_detector”. One can then use the semantic similarity between
“racket” and “tennis” from knowledge sources such as ConceptNet, word2vec

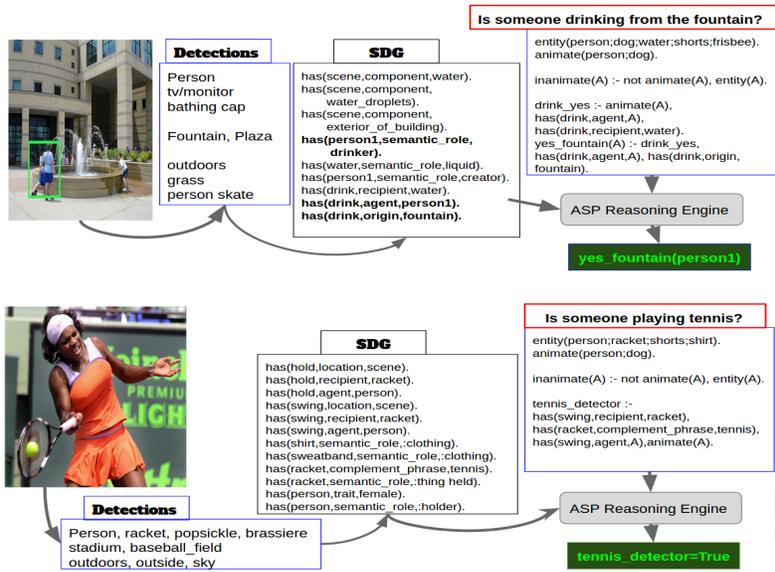


Figure 8: Two example Images from Flickr 8k. Note that for both the images, the state-of-the-art detections are quite noisy. Still, the current framework is able to detect plausible structured graphs which can be queried upon.

to design the weights of the rules (as in [82]).

595 6. Conclusions

Our work introduces a new semantic representation for Scene Analysis called the Scene Description Graph (SDG), and an architecture that combines deep Visual Detection and Reasoning modules to infer such structures. The SDG is a representation of the scene, which integrates direct visual knowledge (objects and their locations in the scene) and additional knowledge obtained using back-
600 ground common sense knowledge. In addition, the SDG has a structure similar to semantic representations of sentences, thus facilitating the interaction between Vision and Natural Language. Having built a common-sense knowledge base related to the domain, we proposed a method of obtaining SDGs from noisy
605 labels using our reasoning module. Recovering the SDG of a scene not only allows the automatic creation of sentences describing the scene, but when used

together with background knowledge, it also has potential usages in reasoning and question-answering about the scene.

We present an implementation of the proposed architecture and demonstrate the effectiveness of the generated SDGs using Image Captioning and Image Retrieval tasks. Our experiments based on the metrics of thoroughness and relevance, show that [the information content in the generated sentences is quiet thorough and relevant; however, the generated sentences are not always as informative as those from existing neural approaches](#). We also discuss how SDGs can be used to answer questions. Furthermore, we show how the proposed framework can be used to explain the results and analyze the sources of the errors (visual detection, knowledge base or reasoning).

References

- [1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, D. Parikh, Vqa: Visual question answering, in: International Conference on Computer Vision (ICCV), 2015.
- [2] C. Xiong, S. Merity, R. Socher, Dynamic memory networks for visual and textual question answering, in: International Conference on Machine Learning, 2016, pp. 2397–2406.
- [3] P. Clark, B. Porter, B. P. Works, Km-the knowledge machine 2.0: Users manual, Department of Computer Science, University of Texas at Austin.
- [4] D. G. Lowe, Object recognition from local scale-invariant features, in: Computer vision, 1999. The proceedings of the seventh IEEE international conference on, Vol. 2, IEEE, 1999, pp. 1150–1157.
- [5] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, Vol. 1, IEEE, 2005, pp. 886–893.
- [6] A. Krizhevsky, I. Sutskever, G. Hinton, Imagenet classification with deep convolutional neural networks, in: NIPS 2012, 2013.

- 635 [7] P. Felzenszwalb, D. McAllester, D. Ramanan, A discriminatively trained, multiscale, deformable part model, in: *Computer Vision and Pattern Recognition, 2008. CVPR 2008.*, IEEE, 2008, pp. 1–8.
- [8] C. H. Lampert, H. Nickisch, S. Harmeling, Learning to detect unseen object classes by between-class attribute transfer, in: *Computer Vision and*
640 *Pattern Recognition (CVPR), 2009.*, IEEE, 2009, pp. 951–958.
- [9] A. Farhadi, I. Endres, D. Hoiem, D. Forsyth, Describing objects by their attributes, in: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, IEEE, 2009, pp. 1778–1785.
- [10] X. Yu, Y. Aloimonos, Attribute-based transfer learning for object categorization with zero/one training example, in: *Computer Vision–ECCV 2010*,
645 *Springer Berlin Heidelberg, 2010*, pp. 127–140.
- [11] C. L. Teo, C. Fermüller, Y. Aloimonos, A gestaltist approach to contour-based object recognition: Combining bottom-up and top-down cues, *The International Journal of Robotics Research* (2015) 0278364914558493.
- 650 [12] C. L. Teo, A. Myers, C. Fermüller, Y. Aloimonos, Embedding high-level information into low level vision: Efficient object search in clutter, in: *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, IEEE, 2013, pp. 126–132.
- [13] X. Yu, C. Fermüller, C. L. Teo, Y. Yang, Y. Aloimonos, Active scene
655 *recognition with vision and language*, in: *Computer Vision (ICCV), 2011 IEEE International Conference on*, IEEE, 2011, pp. 810–817.
- [14] I. Laptev, On space-time interest points, *International Journal of Computer Vision* 64 (2-3) (2005) 107–123.
- 660 [15] R. Messing, C. Pal, H. Kautz, Activity recognition using the velocity histories of tracked keypoints, in: *Computer Vision, 2009 IEEE 12th International Conference on*, IEEE, 2009, pp. 104–111.

- [16] H. Wang, A. Klaser, C. Schmid, C.-L. Liu, Action recognition by dense trajectories, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE, 2011, pp. 3169–3176.
- 665 [17] A. Gupta, L. S. Davis, Objects in action: An approach for combining action understanding and object perception, in: Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, IEEE, 2007, pp. 1–8.
- [18] A. S. Ogale, A. Karapurkar, Y. Aloimonos, View-invariant modeling and recognition of human actions using grammars., in: R. Vidal, A. Heyden, 670 Y. Ma (Eds.), WDV, Vol. 4358 of Lecture Notes in Computer Science, Springer, 2006, pp. 115–126.
URL <http://dblp.uni-trier.de/db/conf/eccv/wdv2006.html#OgaleKA06>
- [19] Y. Yang, C. Fermüller, Y. Aloimonos, A. Guha, A cognitive system for un- 675 derstanding human manipulation actions, *Advances in Cognitive Systems* 3 (2014) 67–86.
- [20] P. Wiryathammabhum, D. Summers-Stay, C. Fermüller, Y. Aloimonos, Computer vision and natural language processing: Recent approaches in multimedia and robotics, *ACM Computing Surveys (CSUR)* 49 (4) (2016) 680 71.
- [21] J. Mao, W. Xu, Y. Yang, J. Wang, A. L. Yuille, Explain images with multimodal recurrent neural networks, arXiv preprint arXiv:1410.1090.
- [22] R. Kiros, R. Salakhutdinov, R. S. Zemel, Unifying visual-semantic embeddings with multimodal neural language models, arXiv preprint 685 arXiv:1411.2539.
- [23] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, in: CVPR, 2015.

- [24] A. Karpathy, F.-F. Li, Deep visual-semantic alignments for generating image descriptions, arXiv preprint arXiv:1412.2306. 690
- [25] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3156–3164.
- [26] X. Chen, C. Lawrence Zitnick, Mind’s eye: A recurrent visual representation for image caption generation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2422–2431. 695
- [27] Q. You, H. Jin, Z. Wang, C. Fang, J. Luo, Image captioning with semantic attention, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
- [28] D. Marr, Vision: A Computational Investigation into the Human Representation and Processing of Visual Information, Henry Holt and Co., Inc., New York, NY, USA, 1982. 700
- [29] M. Malinowski, M. Rohrbach, M. Fritz, Ask your neurons: A neural-based approach to answering questions about images, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1–9. 705
- [30] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, W. Xu, Are you talking to a machine? dataset and methods for multilingual image question answering, in: Proceedings of the 28th International Conference on Neural Information Processing Systems, NIPS’15, MIT Press, Cambridge, MA, USA, 2015, pp. 2296–2304. 710
URL <http://dl.acm.org/citation.cfm?id=2969442.2969496>
- [31] L. Ma, Z. Lu, H. Li, Learning to answer questions from image using convolutional neural network, in: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI’16, AAAI Press, 2016, pp. 3567–3573. 715
URL <http://dl.acm.org/citation.cfm?id=3016387.3016405>

- [32] J. Aloimonos, I. Weiss, A. Bandyopadhyay, Active vision, *International journal of computer vision* 1 (4) (1988) 333–356.
- [33] R. Bajcsy, M. Campos, Active and exploratory perception, *CVGIP: Image Understanding* 56 (1) (1992) 31–40.
- 720 [34] E. Davis, G. Marcus, Commonsense reasoning and commonsense knowledge in artificial intelligence, *Commun. ACM* 58 (9) (2015) 92–103. doi:10.1145/2701413.
- [35] C. Havasi, R. Speer, J. Alonso, Conceptnet 3: a flexible, multilingual semantic network for common sense knowledge, in: *Recent advances in natural language processing*, Citeseer, 2007, pp. 27–29.
- 725 [36] D. B. Lenat, Cyc: A large-scale investment in knowledge infrastructure, *Commun. ACM* 38 (11) (1995) 33–38. doi:10.1145/219717.219745.
- [37] D. Elliott, F. Keller, Image description using visual dependency representations, in: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL, 2013*, pp. 1292–1302.
- 730 URL <http://aclweb.org/anthology/D/D13/D13-1128.pdf>
- [38] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge, *IEEE Trans. Pattern Anal. Mach. Intell.* 39 (4) (2017) 652–663. doi:10.1109/TPAMI.2016.2587640.
- 735 URL <https://doi.org/10.1109/TPAMI.2016.2587640>
- [39] P. Anderson, B. Fernando, M. Johnson, S. Gould, Spice: Semantic propositional image caption evaluation, in: *ECCV, 2016*.
- 740 [40] R. Bernardi, R. Cakici, D. Elliott, A. Erdem, E. Erdem, N. Ikizler-Cinbis, F. Keller, A. Muscat, B. Plank, Automatic description generation from

- images: A survey of models, datasets, and evaluation measures, *J. Artif. Int. Res.* 55 (1) (2016) 409–442.
745 URL <http://dl.acm.org/citation.cfm?id=3013558.3013571>
- [41] M. Hodosh, P. Young, J. Hockenmaier, Framing image description as a ranking task: Data, models and evaluation metrics, *Journal of Artificial Intelligence Research* (2013) 853–899.
- [42] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, D. Forsyth, Every picture tells a story: Generating sentences from
750 images, in: *Proceedings of the 11th European Conference on Computer Vision: Part IV, ECCV’10*, Springer-Verlag, Berlin, Heidelberg, 2010, pp. 15–29.
URL <http://dl.acm.org/citation.cfm?id=1888089.1888092>
- 755 [43] V. Ordonez, G. Kulkarni, T. L. Berg, Im2text: Describing images using 1 million captioned photographs., in: J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. C. N. Pereira, K. Q. Weinberger (Eds.), *NIPS*, 2011, pp. 1143–1151.
URL <http://dblp.uni-trier.de/db/conf/nips/nips2011.html#OrdonezKB11>
760
- [44] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, A. Y. Ng, Grounded compositional semantics for finding and describing images with sentences, *TACL 2* (2014) 207–218.
URL <http://www.transacl.org/wp-content/uploads/2014/04/52.pdf>
765
- [45] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, T. L. Berg, Baby talk: Understanding and generating image descriptions, in: *Proceedings of the 24th CVPR*, 2011.
- [46] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, Y. Choi, Collective generation of natural image descriptions, in: *Proceedings of the 50th Annual*
770

Meeting of the Association for Computational Linguistics: Long Papers -
Volume 1, ACL '12, Association for Computational Linguistics, Strouds-
burg, PA, USA, 2012, pp. 359–368.

URL <http://dl.acm.org/citation.cfm?id=2390524.2390575>

- 775 [47] Y. Yang, C. L. Teo, H. Daumé, III, Y. Aloimonos, Corpus-guided sentence
generation of natural images, in: Proceedings of the Conference on Empiri-
cal Methods in Natural Language Processing, EMNLP '11, Association for
Computational Linguistics, Stroudsburg, PA, USA, 2011, pp. 444–454.

URL <http://dl.acm.org/citation.cfm?id=2145432.2145484>

- 780 [48] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, S. C. Zhu, I2t: Image parsing to
text description., Proceedings of the IEEE 98 (8) (2010) 1485–1508.

URL [http://dblp.uni-trier.de/db/journals/pieee/
pieee98.html#YaoYLLZ10](http://dblp.uni-trier.de/db/journals/pieee/pieee98.html#YaoYLLZ10)

- [49] D. Lin, S. Fidler, C. Kong, R. Urtasun, Generating multi-sentence nat-
785 ural language descriptions of indoor scenes, in: M. W. J. Xianghua Xie,
G. K. L. Tam (Eds.), Proceedings of the British Machine Vision Conference
(BMVC), BMVA Press, 2015, pp. 93.1–93.13. doi:10.5244/C.29.93.

URL <https://dx.doi.org/10.5244/C.29.93>

- [50] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, A. Yuille, Deep caption-
790 ing with multimodal recurrent neural networks (m-rnn), arXiv preprint
arXiv:1412.6632.

- [51] R. Lebre, P. H. Pinheiro, R. Collobert, Phrase-based image captioning, in:
International Conference on Machine Learning (ICML), no. EPFL-CONF-
210021, 2015.

- 795 [52] S. Schuster, R. Krishna, A. Chang, L. Fei-Fei, C. D. Manning, Generating
semantically precise scene graphs from textual descriptions for improved
image retrieval, in: Proceedings of the Fourth Workshop on Vision and
Language, Association for Computational Linguistics, 2015, pp. 70–80.

- [53] C. Baral, Knowledge representation, reasoning and declarative problem solving, Cambridge university press, 2003.
- [54] L. De Raedt, A. Kimmig, H. Toivonen, Problog: A probabilistic prolog and its application in link discovery, in: Proceedings of the 20th International Joint Conference on Artificial Intelligence, IJCAI'07, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2007, pp. 2468–2473.
- [55] C. L. Zitnick, D. Parikh, Bringing semantics into focus using visual abstraction., in: CVPR, IEEE, 2013, pp. 3009–3016.
URL <http://dblp.uni-trier.de/db/conf/cvpr/cvpr2013.html#ZitnickP13>
- [56] S. K. Divvala, A. Farhadi, C. Guestrin, Learning everything about anything: Webly-supervised visual concept learning, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014, 2014, pp. 3270–3277.
- [57] M. Santofimia, J. Martinez-del Rincon, J.-C. Nebel, Common-Sense Knowledge for a Computer Vision System for Human Action Recognition, in: J. Bravo, R. Hervás, M. Rodríguez (Eds.), Ambient Assisted Living and Home Care, Vol. 7657 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2012, pp. 159–166.
URL http://dx.doi.org/10.1007/978-3-642-35395-6_22
- [58] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft coco: Common objects in context, in: Computer Vision–ECCV 2014, Springer, 2014, pp. 740–755.
- [59] H. Bagherinezhad, H. Hajishirzi, Y. Choi, A. Farhadi, Are elephants bigger than butterflies? reasoning about sizes of objects, in: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI'16, AAAI Press, 2016, pp. 3449–3456.
URL <http://dl.acm.org/citation.cfm?id=3016387.3016389>

- [60] E. Yee, E. G. Chrysikou, S. L. Thompson-Schill, The cognitive neuroscience of semantic memory.
- [61] S. H. Bach, B. Huang, B. London, L. Getoor, Hinge-loss markov random fields: Convex inference for structured prediction, in: Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence, UAI'13, AUAI Press, Arlington, Virginia, United States, 2013, pp. 32–41.
URL <http://dl.acm.org/citation.cfm?id=3023638.3023642>
- [62] M. Richardson, P. Domingos, Markov logic networks, Machine learning 62 (1-2) (2006) 107–136.
- [63] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Computer Vision and Pattern Recognition, 2014.
- [64] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva, Learning deep features for scene recognition using places database., NIPS.
- [65] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell, Decaf: A deep convolutional activation feature for generic visual recognition, in: Proceedings of the 31st International Conference on Machine Learning (ICML-14), 2014, pp. 647–655.
- [66] A. Sharma, N. H. Vo, S. Aditya, C. Baral, Towards addressing the winograd schema challenge - building and using a semantic parser and a knowledge hunting module, in: Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015, 2015, pp. 1319–1325.
URL <http://ijcai.org/papers15/Abstracts/IJCAI15-190.html>
- [67] D. Chen, C. Manning, A Fast and Accurate Dependency Parser using Neural Networks, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational

- 855 Linguistics, 2014, pp. 740–750.
 URL <http://www.aclweb.org/anthology/D14-1082>
- [68] T. Kollar, N. Roy, Utilizing object-object and object-scene context when
 planning to find things, in: Robotics and Automation, 2009. ICRA'09.
 IEEE International Conference on, IEEE, 2009, pp. 2168–2173.
- 860 [69] M. Scutari, Learning bayesian networks with the bnlearn R package, Jour-
 nal of Statistical Software 35 (3) (2010) 1–22.
- [70] A. Gatt, E. Reiter, Simplenlg: A realisation engine for practical applica-
 tions, in: Proceedings of the 12th European Workshop on Natural Lan-
 guage Generation, ENLG '09, Association for Computational Linguistics,
865 Stroudsburg, PA, USA, 2009, pp. 90–93.
 URL <http://dl.acm.org/citation.cfm?id=1610195.1610208>
- [71] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for auto-
 matic evaluation of machine translation, in: Proceedings of the 40th an-
 nual meeting on association for computational linguistics, Association for
870 Computational Linguistics, 2002, pp. 311–318.
- [72] M. Denkowski, A. Lavie, Meteor universal: Language specific translation
 evaluation for any target language, in: Proceedings of the EACL 2014
 Workshop on Statistical Machine Translation, 2014.
- [73] D. Lin, An information-theoretic definition of similarity., in: ICML, Vol. 98,
875 1998, pp. 296–304.
- [74] L. Ma, Z. Lu, L. Shang, H. Li, Multimodal convolutional neural networks
 for matching image and sentence, in: 2015 IEEE International Conference
 on Computer Vision (ICCV), 2015, pp. 2623–2631. doi:10.1109/ICCV.
 2015.301.
- 880 [75] J. Devlin, H. Cheng, H. Fang, S. Gupta, L. Deng, X. He, G. Zweig,
 M. Mitchell, Language models for image captioning: The quirks and what

- works, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers, 2015, pp. 100–105.
URL <http://aclweb.org/anthology/P/P15/P15-2017.pdf>
- [76] T. Lan, W. Yang, Y. Wang, G. Mori, Image retrieval with structured object queries using latent ranking svm, in: ECCV, 2012.
- [77] M. Y. Yang, W. Liao, H. Ackermann, B. Rosenhahn, On support relations and semantic scene graphs, {ISPRS} Journal of Photogrammetry and Remote Sensing 131 (2017) 15 – 25. doi:<https://doi.org/10.1016/j.isprsjprs.2017.07.010>.
URL <http://www.sciencedirect.com/science/article/pii/S0924271617300746>
- [78] G. Paolacci, J. Chandler, P. G. Ipeirotis, Running experiments on amazon mechanical turk.
- [79] M. Kilickaya, A. Erdem, N. Ikizler-Cinbis, E. Erdem, Re-evaluating automatic metrics for image captioning, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, Association for Computational Linguistics, 2017, pp. 199–209.
URL <http://www.aclweb.org/anthology/E17-1019>
- [80] M. Gelfond, V. Lifschitz, The stable model semantics for logic programming, MIT Press, 1988, pp. 1070–1080.
- [81] A. Kimmig, S. H. Bach, M. Broecheler, B. Huang, L. Getoor, A short introduction to probabilistic soft logic, in: NIPS Workshop on Probabilistic Programming: Foundations and Applications, 2012.

- [82] S. Aditya, Y. Yang, C. Baral, Y. Aloimonos, Answering image riddles using vision and reasoning through probabilistic soft logic, arXiv preprint arXiv:1611.05896.

5. Supplementary Material for on-line publication only

[Click here to download 5. Supplementary Material for on-line publication only: appendix_arkiv.pdf](#)

LaTeX Souce Files

[Click here to download LaTeX Souce Files: cviu17-image-understanding_Nov17_sources.zip](#)