

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/321896723>

# Image Understanding using Vision and Reasoning through Scene Description Graph

Article in *Computer Vision and Image Understanding* · December 2017

DOI: 10.1016/j.cviu.2017.12.004

CITATIONS

0

READS

75

5 authors, including:



**Somak Aditya**

Arizona State University

9 PUBLICATIONS 27 CITATIONS

[SEE PROFILE](#)



**Yezhou Yang**

Arizona State University

49 PUBLICATIONS 455 CITATIONS

[SEE PROFILE](#)



**Yiannis Aloimonos**

University of Maryland, College Park

315 PUBLICATIONS 7,621 CITATIONS

[SEE PROFILE](#)



**Cornelia Fermüller**

University of Maryland, College Park

191 PUBLICATIONS 2,381 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



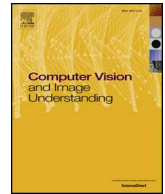
Vision and Reasoning [View project](#)



Deep Learning [View project](#)

All content following this page was uploaded by **Yezhou Yang** on 24 December 2017.

The user has requested enhancement of the downloaded file.



# Image Understanding using vision and reasoning through Scene Description Graph

Somak Aditya<sup>a,\*</sup>, Yezhou Yang<sup>a</sup>, Chitta Baral<sup>a</sup>, Yiannis Aloimonos<sup>b</sup>, Cornelia Fermüller<sup>b</sup>

<sup>a</sup> Arizona State University, Tempe, AZ, USA

<sup>b</sup> University of Maryland, College Park, MD, USA

## ARTICLE INFO

### Keywords:

Image Understanding  
Commonsense Reasoning  
Vision  
Reasoning

## ABSTRACT

Two of the fundamental tasks in image understanding using text are caption generation and visual question answering (Antol et al., 2015; Xiong et al., 2016). 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 (Clark et al., 2004)), 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 scene (Dalal and Triggs, 2005; Krizhevsky et al., 2013; Lowe, 1999), objects are recognized by detecting their parts or attributes (Farhadi et al., 2009; Felzenszwalb et al., 2008; Lampert et al., 2009; Teo et al., 2015; 2013; Yu and Aloimonos, 2010; Yu et al., 2011) and activities are recognized by detecting the motions, objects and contexts involved in the activities (Gupta and Davis, 2007; Laptev, 2005; Messing et al., 2009; Ogale et al., 2006; Wang et al., 2011; Yang et al., 2014).

Since then, researchers have explored multiple ways of understanding an image through the modality of natural language. According to Wiriyathamabhum et al. (2016), 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 (Chen and Lawrence Zitnick, 2015; Donahue et al., 2015;

Karpathy and Li, 2014; Kiros et al., 2014; Mao et al., 2014b; Vinyals et al., 2015; You et al., 2016) 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 (Marr (1982) 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 (Antol et al., 2015; Gao et al., 2015; Ma et al., 2016; Malinowski et al., 2015).

Despite the success of end-to-end learning models (Antol et al., 2015; Gao et al., 2015; Ma et al., 2016; Malinowski et al., 2015) 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 Fig. 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 Fig. 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

\* Corresponding author.

E-mail address: [saditya1@asu.edu](mailto:saditya1@asu.edu) (S. Aditya).

<https://doi.org/10.1016/j.cviu.2017.12.004>

Received 16 February 2017; Received in revised form 4 December 2017; Accepted 14 December 2017

1077-3142/ © 2017 Elsevier Inc. All rights reserved.

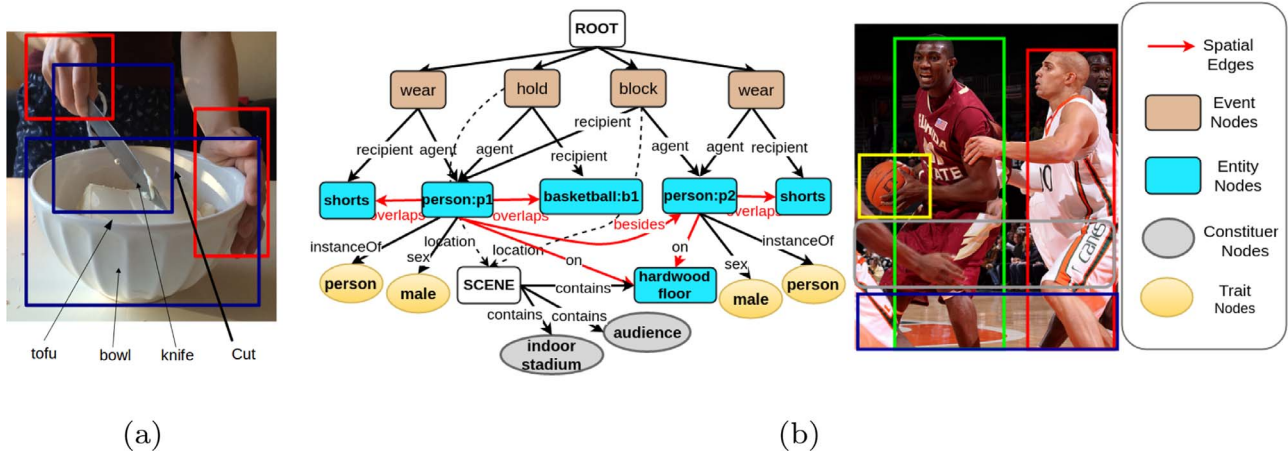


Fig. 1. (a) First example image and (b) second example image with corresponding ideal SDG encoding semantic, ontological, and spatial relations.

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 common-sense 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 (Aloimonos et al., 1988; Bajcsy and Campos, 1992). 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 knowledge<sup>1</sup> 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 base<sup>2</sup> 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 Fig. 1, we show a possible SDG for an example image. SDG is a directed labeled graph<sup>3</sup> 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 (Clark et al., 2004)) 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 Elliott and Keller (2013). These SDGs can be used to generate captions, answer factual questions and also reason beyond 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 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 (Karpathy and Li, 2014; Vinyals et al., 2017) and ground truth captions for three existing Captioning Datasets (Flickr8k, Flickr30k and MS-COCO),<sup>4</sup> which can be used to propose better automatic caption evaluation metrics (this dataset is used in Anderson et al., 2016 to propose SPICE).

## 2. Related works

Our work is influenced by various thrusts of work focusing on extracting meaningful information from images and videos. As suggested by Karpathy and Li (2014), 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 Bernardi et al. (2016), may be classified into the following three categories: i) Direct generation models, ii) retrieval models from visual space, and iii) retrieval models from multimodal space.

**Caption generation:** With respect to caption generation tasks, we

<sup>1</sup> Commonsense reasoning and Commonsense Knowledge can be of many types (Davis and Marcus, 2015). Commonsense Knowledge can belong to different levels of abstraction (Havasi et al., 2007; Lenat, 1995). In this paper, we focus on reasoning based on knowledge about natural scenes.

<sup>2</sup> Domain-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.

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

<sup>4</sup> Comparison with both the neural approaches are done on MS-COCO dataset. For the rest, comparison is done only with Karpathy and Li (2014).

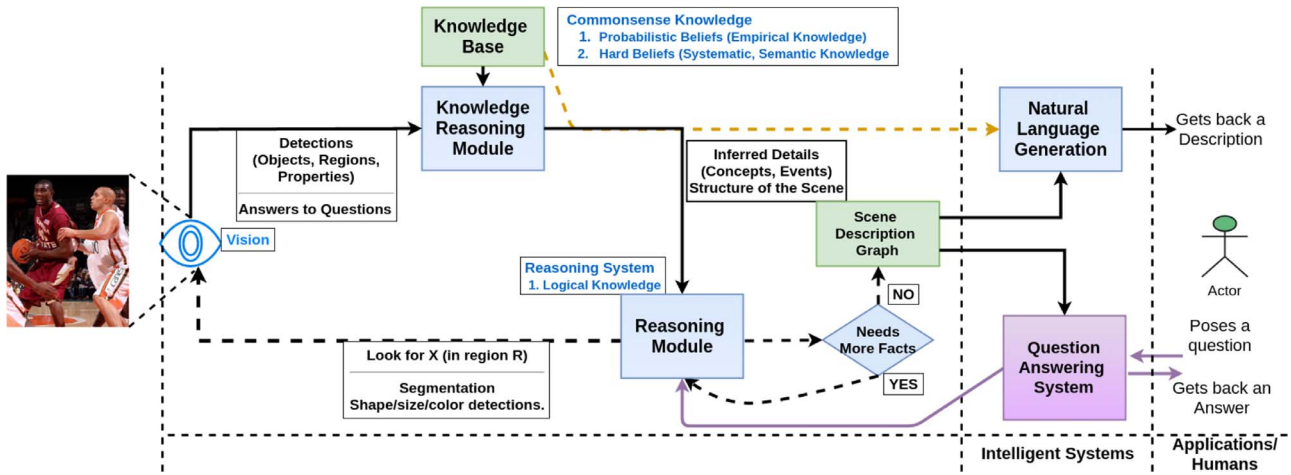


Fig. 2. An architecture for deep image understanding. (The Knowledge Reasoning Module is a part of the Reasoning Module; it is shown separately to clearly outline the interactions).

share our roots with the works on generating textual descriptions, i.e., direct generation models. These include the works in (Farhadi et al., 2010; Hodosh et al., 2013; Ordonez et al., 2011; Socher et al., 2014) which retrieve and rank sentences from training sets given an image. Other works (Elliott and Keller, 2013; Kulkarni et al., 2011; Kuznetsova et al., 2012; Yang et al., 2011; Yao et al., 2010) 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 and Li (2014), which used a combination of a Convolutional Neural Network (for images) and a bi-directional Recurrent Neural Network (for sentences). Subsequent works (Kiros et al., 2014; Lebre et al., 2015; Lin et al., 2015; Mao et al., 2014a) 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 Schuster et al. (2015) are open-ended phrases and the spatial graphs in Elliott and Keller (2013) only represent the spatial relations between objects and regions. Reasoning directly on such structures using known logical reasoning languages (such as Answer Set Programming (Baral, 2003), ProbLog (De Raedt et al., 2007)) 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 Zitnick and Parikh (2013) uses abstraction to discover semantically similar images, Divvala et al. (2014) proposes to learn all variations pertaining to all concepts, and Santofimia et al. (2012) 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 (Antol et al., 2015; Gao et al., 2015; Ma et al., 2016; Malinowski et al., 2015). Interestingly, both Antol et al. (2015) and Gao et al. (2015) have adapted MS-COCO (Lin et al., 2014) images to create an open domain dataset with human generated questions and answers. The works Malinowski et al. (2015) and Gao et al. (2015) 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 Fig. 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 (Aloimonos et al., 1988), the visual detection module is also expected to interact with the reasoning 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 (Bagherinezhad et al., 2016); 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 knowledge<sup>5</sup>

<sup>5</sup> The 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)” (Yee et al., 2013).



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 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 (Bach et al., 2013), and Markov Logic Networks (Richardson and Domingos, 2006).

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 (Fig. 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 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 objects<sup>6</sup> and their *observable* attributes (location, shape, size, color, contour etc.), regions and their *observable* attributes, and actions. To avoid further complexity, we consider only 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 or regions (described in phrases or words) that can be directly grounded in the image.<sup>7</sup> 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.

**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) actions, involving (one or more) objects; and spatial relationships among objects and regions. The action(s) together make up a natural event which can be described by sentence(s), such as: *a person is lying on a bench, in a park; a person 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 text<sup>8</sup>. 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) (Wiriyathamabhum et al., 2016). 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 Girshick et al. (2014). 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 Zhou et al. (2014). 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 about the properties of objects.<sup>9</sup> 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 Krizhevsky et al. (2013) to extract a 4096 dimensional feature vector (using Donahue et al., 2014). We then trained a multi-label SVM to recognize constituents using these deep features.

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

<sup>6</sup> Objects 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.

<sup>7</sup> To 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.

<sup>8</sup> Karpathy and Li (2014)'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).

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

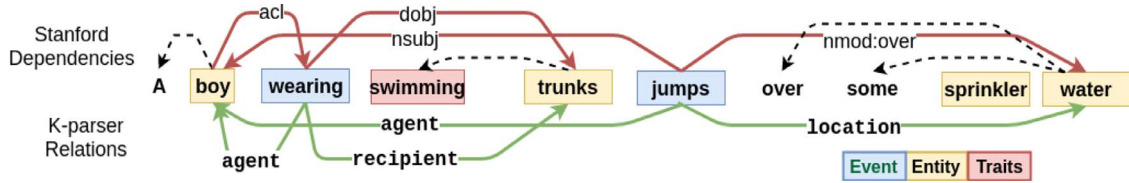


Fig. 3. An example sentence with Stanford Dependency relations and transformed K-parser relations. Only important Stanford Dependencies and K-parser relations are shown. K-parser also adds semantic roles and superclass information for the Entities (not shown in the figure).

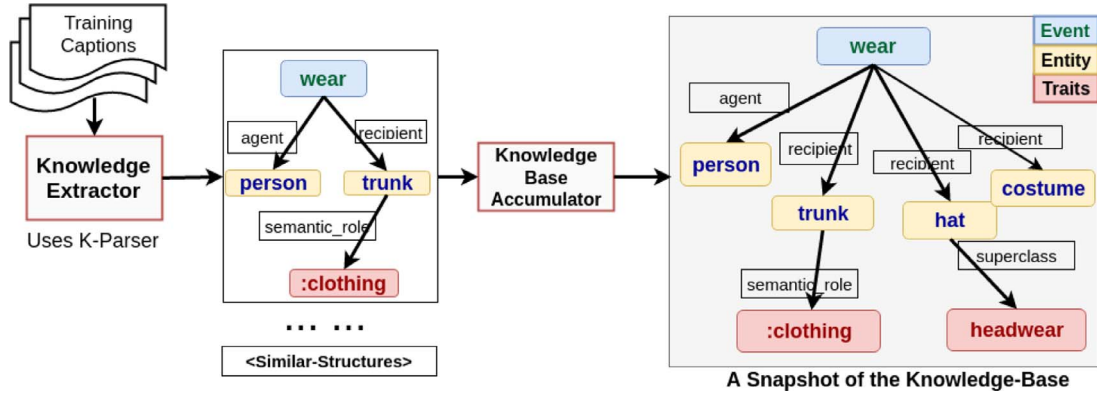


Fig. 4. Knowledge Base Creation using a Semantic Parser.

#### 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 ( $P(\text{isc}|\text{scene})$ ) for each ISC in a scene. For example, for the scene class *airport\_terminal*, we add  $\{\text{waiting room, big glass view, travelers}\}$  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) Recognition module to detect the top 5 scene categories from each training image. We denote the human annotations for all training images as  $\mathcal{S}_{tr}$ .

##### 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 ( $\langle \mathcal{S}_M, \mathcal{S}_T, \mathcal{S}_{tr} \rangle$ ). To extract knowledge from the annotations, we extensively use a semantic parser, called K-parser (Sharma et al., 2015).

**K-Parser:** K-parser ([kparser.org](http://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 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 Fig. 3). It also extracts Traits (attributes) *swimming*, *sprinkler* corresponding to the entities. Internally, K-parser uses the Stanford Parser (Chen and Manning, 2014) to get the syntactic dependency graph from a sentence. The K-parser then uses a rule-based mapping algorithm to map these dependency relations to the set of KM-

Relations (Clark et al., 2004) 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 (Clark et al., 2004) to represent semantic relations (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 Prop-Bank framesets; and v) it has other features such as Co-reference resolution, Word Sense Disambiguation and Named Entity Tagging.<sup>10</sup>

**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  $\mathcal{S}_{tr}$ ), Entity (from  $\mathcal{N}$ ) or a Trait (adjectives, qualitative-nouns from  $\mathcal{S}_{tr}$  or WordNet-superclass of a word). The *relation* comes from a closed set of semantic relations from KM-Ontology.<sup>11</sup> 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. Fig. 4 depicts a snapshot of  $\mathcal{G}$ .

As shown in Fig. 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 similarities. 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*  $(y\text{-superclass-z})$  *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

<sup>10</sup> For more details, please see Sharma et al. (2015).

<sup>11</sup> *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](http://kparser.org).






				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

Fig. 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 Sharma et al. (2015)).

co-occur in a scene. Authors in Kollar and Roy (2009) 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 ( $\mathcal{I}_s$ ), 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  $N = |\mathcal{N}| + |\mathcal{N}_s| + |\mathcal{I}_s|$ ). 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-nlearn package (Scutari, 2010).

To create  $\mathcal{D}$ , we process the annotations for each training image ( $\mathcal{A}_r$ ) 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  $\text{edge}(\text{wear}, \text{agent}, \text{person})$  and  $\text{edge}(\text{wear}, \text{recipient}, \text{shorts})$  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 (Yang et al., 2011). Inspired by these approaches, we use the commonsense knowledge  $\langle \mathcal{N}_b, \mathcal{B}_n, \mathcal{S}_M \rangle$  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 Fig. 5.

**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  $\hat{\mathcal{C}}$ . Initially,  $\mathcal{C}_{inf} = \phi$ , and  $\mathcal{O}_x = \{n | P_r(n|x) > \alpha_n\}$ . We calculate

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

and add  $\mathcal{C}_{max}$  to  $\mathcal{C}_{inf}$ . We iterate while the entropy  $E$  keeps decreasing

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

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

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  $\mathcal{C}_{inf}$ . 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

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

and then add  $\mathcal{O}_{max}$  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 Entities *person* and *swimming trunks* (corresponds to the vertex *trunk* in  $\mathcal{N}_b$ ). 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  $\mathcal{E}_v$ .

For filtering spurious Events, we use the semantics in K-parser edge labels and the superclass (type) of the Entities from  $\mathcal{N}_b$ . 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 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 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  $\mathcal{C}$ .

**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 for each scene based on the

<sup>12</sup> The hyper-parameters ( $T, \alpha_n$ ) are set based on performance on validation data. In our experiments, we have used the values 5, 0.5 respectively.



$P_r(n|x)$ ,  $P_r(s|x)$ ,  $P_r(c|x)$  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, s)` for all ISC  $s$  in  $C_{inf}$ ; 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  $(\mathcal{E}_x \setminus \mathcal{E}_v)$ : if it is an animate Entity, add `edge(o_{im}, location, scene)`; Otherwise, find the shortest path from  $o_{im}$  to the top detected Event in the  $\mathcal{N}_b$  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, 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 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 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 Karpathy and Li (2014) as it was one of the recent (and among the first) 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. (2017) (appeared in IEEE TPAMI 2016) which reported improved quality of captions in comparison to Karpathy and Li (2014). This method uses the latest Inception-V3 architecture to process images and an Long-Short Term Memory (LSTM) 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 (Hodosh et al., 2013). These three datasets have 8092, 31,783 and more than 160K images respectively. Every image from these datasets is annotated with 5 sentences describing the image. For all datasets, we used the train-test splits from Karpathy and Li (2014) 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 using SimpleNLG (Gatt and Reiter, 2009). 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 quantitative<sup>13</sup>) of sentences using simple rules. The other rules used are: i) `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)`

<sup>13</sup> For high-scoring detections, we consider the spatial information from the bounding-boxes. For  $N$  such detections of an object  $obj$ , we generate sentences like  $N$   $obj$ 's are in the scene.

**Table 1**

Sentence generation relevance (R) and thoroughness (T) human evaluation results with gold standard and (Karpathy and Li, 2014) on Flickr 8k, 30k test images and COCO validation images. D: Standard Deviation.

Experiment	Karpathy and Li (2014) BRNN	Our method	Gold Standard
R $\pm$ D(8k)	2.08 $\pm$ 1.35	<b>2.82 <math>\pm</math> 1.56</b>	4.69 $\pm$ 0.78
T $\pm$ D(8k)	2.24 $\pm$ 1.33	<b>2.62 <math>\pm</math> 1.42</b>	4.32 $\pm$ 0.99
R $\pm$ D(30k)	1.93 $\pm$ 1.32	<b>2.43 <math>\pm</math> 1.42</b>	4.78 $\pm$ 0.61
T $\pm$ D(30k)	2.17 $\pm$ 1.34	<b>2.49 <math>\pm</math> 1.42</b>	4.52 $\pm$ 0.93
R $\pm$ D(COCO)	<b>2.69 <math>\pm</math> 1.49</b>	2.14 $\pm$ 1.29	4.71 $\pm$ 0.67
T $\pm$ D(COCO)	<b>2.55 <math>\pm</math> 1.41</b>	2.06 $\pm$ 1.24	4.37 $\pm$ 0.92

is converted to a sentence based on the template “the scene contains B and...”. For BRNN (Karpathy and Li, 2014), we use the implementation provided by the authors to train and generate sentences from an image. To generate captions using (Vinyals et al., 2017), we use the code provided by the authors.<sup>14</sup> 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 metrics: Relevance and Thoroughness, are therefore, proposed as empirical measures. 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 descriptions based on a discrete scale ranging from 1–5 (low relevance/thoroughness to high relevance/thoroughness).<sup>15</sup> 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 Karpathy and Li (2014).

**A supplementary AMT study:** It is often considered a good practice to perform multiple independent AMT studies. In Table 2, we provide the results 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 Karpathy and Li (2014) and Vinyals et al. (2017). As previously stated, we use the 2000 MS-COCO validation images to report the results.

The work in Vinyals et al. (2017) is one of the latest proposed methods using a state-of-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 our experiments with evaluation results using BLEU (Papineni et al., 2002) and Meteor (Denkowski and Lavie, 2014) scores. The BLEU scores are calculated using the original PERL script<sup>16</sup> provided for statistical machine translation tasks. The Meteor scores are calculated using the instructions provided by the original authors.<sup>17</sup> We provide detailed insights about Table 1–3 in the Analysis section.

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

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

<sup>15</sup> We 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>.

<sup>16</sup> BLEU Evaluation Perl Script.

<sup>17</sup> Meteor 1.5.



**Table 2**

Sentence generation relevance (R) and thoroughness (T) human evaluation results with gold standard, [Karpathy and Li \(2014\)](#) and [Vinyals et al. \(2017\)](#) on COCO validation images. D: Standard Deviation.

Experiment	<a href="#">Vinyals et al. (2017)</a> ShowAndTell	<a href="#">Karpathy and Li (2014)</a> 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

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

$$Sim(\mathcal{Q}_q, \mathcal{Q}_x) = \left( \sum_{v_q \in V_q} \max_{v_i \in V_i} sim(v_q, v_i) \right) / |V_q|$$

$$sim(v_q, v_i) = 0.5 * (wnsim(label(v_q), label(v_i)) + Jaccard(neighbors(v_q), neighbors(v_i))).$$

Vertex-similarity is calculated based on word-meaning similarity



**Fig. 6.** We provide some comparative captions generated by our system (in yellow box), by BRNN ([Karpathy and Li, 2014](#)) (top blue box), by ShowAndTell ([Vinyals et al., 2017](#)) (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. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Table 3**

Sentence generation BLEU, Meteor Scores in comparison with some of the Existing Neural Architectures (Karpathy and Li, 2014) and (Vinyals et al., 2017) 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 Karpathy and Li (2014).

	Flickr-8k				Flickr-30k				COCO-2014				
Experiment	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
Vinyals et al. (2017) ShowAndTell	<b>63</b>	<b>41</b>	<b>27</b>	–	<b>66.3</b>	<b>42.3</b>	<b>27.7</b>	<b>18.3</b>	<b>66.6</b>	<b>46.1</b>	<b>32.9</b>	<b>24.6</b>	–
Karpathy and Li (2014) 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	<b>19.5</b>
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 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.

	Flickr8k			
Model	R@1	R@5	R@10	Med r
Karpathy and Li (2014) BRNN	11.8	32.1	44.7	12.4
Vinyals et al. (2017) ShowAndTell	<b>19</b>	–	<b>64</b>	<b>5.0</b>
Our method-SDG	18.1	39.0	50.0	10.5
	Flickr30k			
Karpathy and Li (2014) BRNN	15.2	37.7	50.5	9.2
Vinyals et al. (2017) ShowAndTell	17	–	57	7.0
Our method-SDG	<b>26.5</b>	<b>48.7</b>	<b>59.4</b>	<b>6.0</b>
	MS-COCO			
Karpathy and Li (2014) BRNN (1k)	<b>20.9</b>	<b>52.8</b>	<b>69.2</b>	<b>4.0</b>
Our method-SDG (1k)	19.3	35.5	49.0	11.0
Our method-SDG (2k)	15.4	32.5	42.2	17.0

and neighbor similarity. Here  $wnsim(., .)$  is Lin Similarity (Lin, 1998) between two words and  $Jaccard(., .)$  is the standard Jaccard coefficient similarity. Based on the above measure, we provide the image retrieval results compared with results from Karpathy and Li (2014) in Table 4. Additionally, we provide the results of the Show-and-Tell method (Vinyals et al., 2017) 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

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.

**Comparable Systems:** There are other works in Image Retrieval (Ma et al., 2015) and Caption Generation (Devlin et al., 2015) 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 Schuster et al. (2015) 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 Schuster et al. (2015) are open-ended text. To that end, other related works (Elliott and Keller, 2013; Lan et al., 2012; Yang et al., 2017) 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 (Karpathy and Li, 2014) and (Vinyals et al., 2017).

**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 (Paolacci et al., 2010) 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 Karpathy and Li, 2014 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 Fig. 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 is scored 36.4, 0.0, 0.0, 0.0. In an even worse comparison, for the image in Fig. 6(d), both generated sentences are correct in meaning. Yet, the sentence from the Neural captioning engine 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. This is 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 community<sup>18</sup> and more automatic caption evaluation metrics are proposed. However, 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 reflect that the performing system is worse, as the generated caption can match some other caption written by a different Turker than the Turkers who annotated the image. This is why we perform human evaluations of thoroughness 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 (Bernardi et al., 2016), 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 Bernardi et al. (2016)).

**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  $\mathcal{H}_b$ ) and Bayes Net encodes contextual knowledge, i.e. which *type* of entities and events, or entities and ISCs co-occur in common contexts. In Fig. 6, the information in sentences “the scene contains ...” are obtained from the Bayes Net. Additionally, the

<sup>18</sup> The work in Kilickaya et al. (2017) shows the different automatic image captioning metrics have very little correlation with human judgment. Notably, this work uses our COMPOSITE dataset (captions from SDG, Karpathy and Li (2014) and AMT scores) to show the above result.

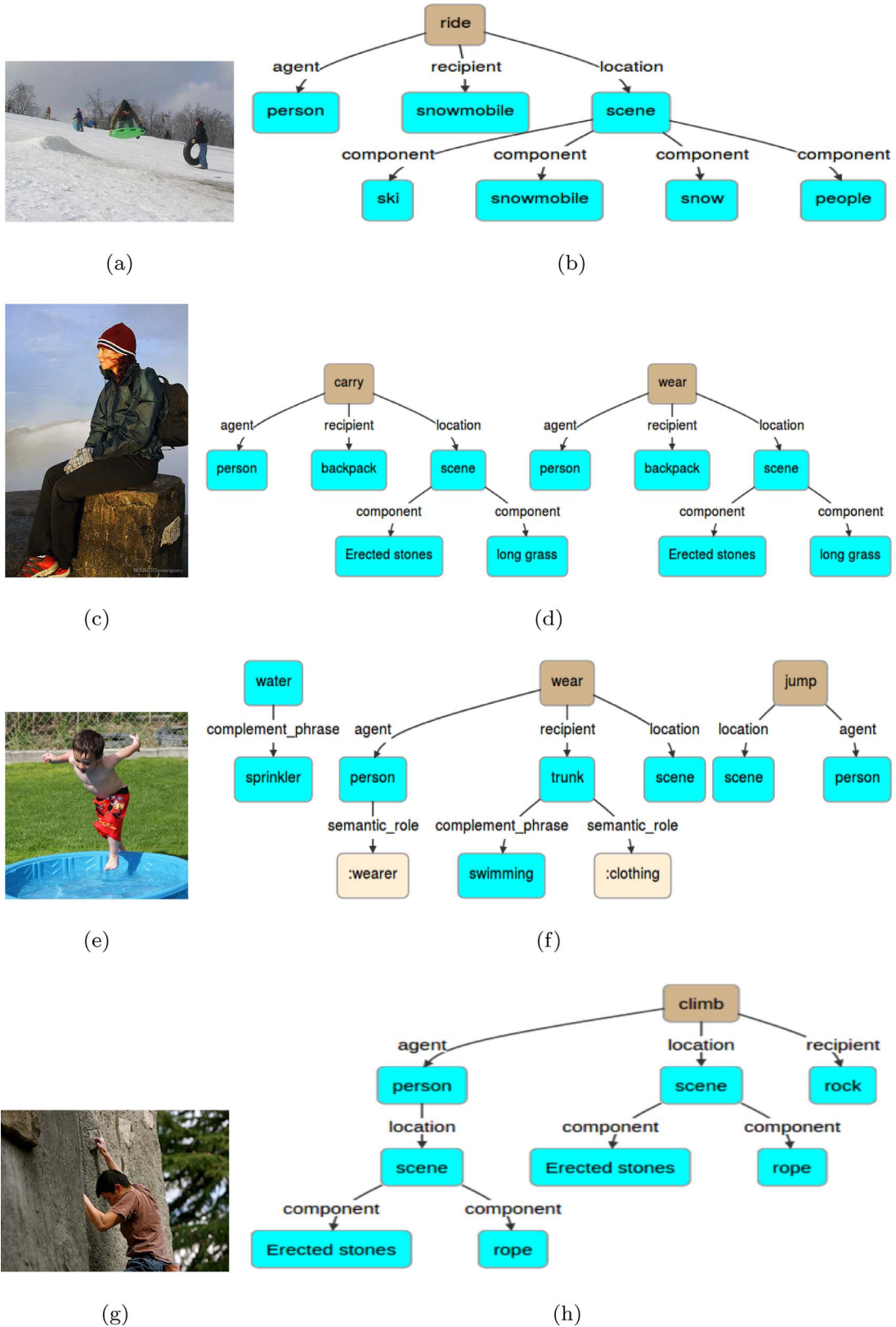


Fig. 7. The SDGs in (b), (d), (f) and (h) corresponds to images (a), (c), (e) and (g) respectively. More examples are at <http://bit.ly/1NJycKO>.

Knowledge base encodes events or actions that occur in context of entities, for example all verbs in Fig. 6 is inferred by the Knowledge Base based on the detected entities.

**Interpretability:** One of the major disadvantages of many end-to-end 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 structure and our architecture. Referring to Fig. 7(g), the initial top object and scene detections are: {person, backpack, artichoke, hat with a wide brim}; {wheat field, cemetery,



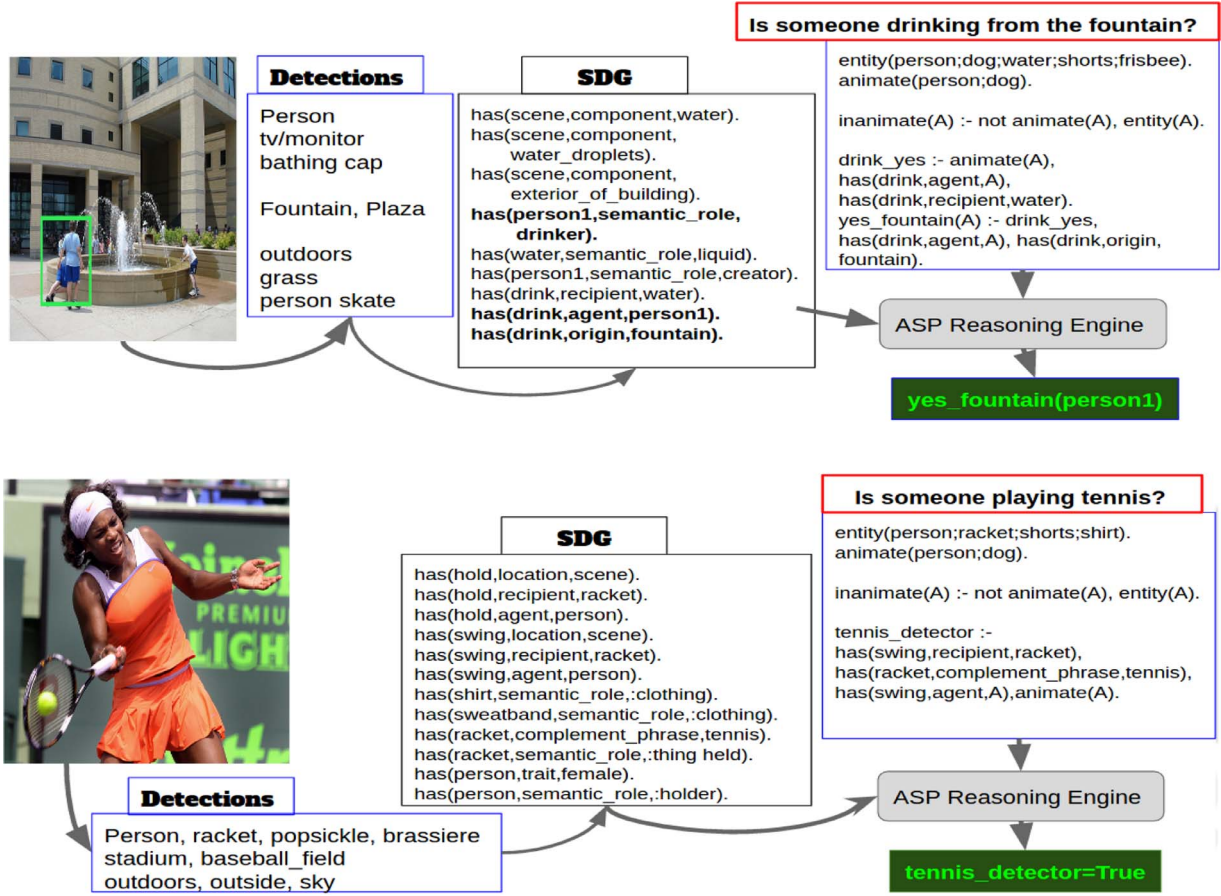


Fig. 8. Two example Images from Flickr 8k. Note that for both the images, the state-of-the-art detections are quite noisy. Still, the current framework is able to detect plausible structured graphs which can be queried upon.

*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 following ways: i) why the SDG in Fig. 7(g) contains *person* and *backpack*? They are detected by object classifiers with high probability; ii) why the SDG in Fig. 7(g) contains *erected stone*? Because scene categories such as *cemetery* co-occurs with erected stone (knowledge from  $\mathcal{S}_M$ ); iii) why the SDG in Fig. 7(g) has verb *carry, wear*? Because it co-occurs with the entities (person, backpack) (knowledge from  $\mathcal{H}_b$ ). In short, explanations for the components in the SDG in Fig. 7(g) can be tracked back to one of the knowledge sources in  $(\mathcal{H}_b, \mathcal{B}_n, \mathcal{S}_M)$  or the 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 question and give an answer. Such mechanisms require further research and development. 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 Fig. 8(a), the Scene Description Graph is represented as 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 (Gelfond and Lifschitz, 1988). 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)*.

For the second image in Fig. 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*. Though the above question is written in ASP without any probabilistic weight, one can rewrite the rules in Probabilistic Soft Logic (Kimmig et al., 2012) 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 to design the weights of the rules (as in Aditya et al., 2016).

## 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 background 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 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).

## Acknowledgement

Yiannis Aloimonos and Cornelia Fermüller acknowledge the support of the National Science Foundation under grants SMA 1540917 and CNS 1544797.

## Supplementary material

Supplementary material associated with this article can be found, in the online version, at [10.1016/j.cviu.2017.12.004](https://doi.org/10.1016/j.cviu.2017.12.004).

## References

- Aditya, S., Yang, Y., Baral, C., Aloimonos, Y., 2016. Answering image riddles using vision and reasoning through probabilistic soft logic. *arXiv preprint arXiv:1611.05896*.
- Aloimonos, J., Weiss, I., Bandyopadhyay, A., 1988. Active vision. *Int. J. Comput. Vis.* 1 (4), 333–356.
- Anderson, P., Fernando, B., Johnson, M., Gould, S., 2016. Spice: semantic propositional image caption evaluation. In: *ECCV*.
- Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D., 2015. Vqa: Visual question answering. *International Conference on Computer Vision (ICCV)*.
- Bach, S.H., Huang, B., London, B., Getoor, L., 2013. Hinge-loss markov random fields: convex inference for structured prediction. *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*. AUAI Press, Arlington, Virginia, United States, pp. 32–41. <http://dl.acm.org/citation.cfm?id=3023638.3023642>.
- Bagherinezhad, H., Hajishirzi, H., Choi, Y., Farhadi, A., 2016. Are elephants bigger than butterflies? Reasoning about sizes of objects. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. AAAI Press, pp. 3449–3456. <http://dl.acm.org/citation.cfm?id=3016387.3016389>.
- Bajcsy, R., Campos, M., 1992. Active and exploratory perception. *CVGIP* 56 (1), 31–40.
- Baral, C., 2003. *Knowledge Representation, Reasoning and Declarative Problem Solving*. Cambridge University Press.
- Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Iklizler-Cinbis, N., Keller, F., Muscat, A., Plank, B., 2016. Automatic description generation from images: a survey of models, datasets, and evaluation measures. *J. Artif. Int. Res.* 55 (1), 409–442. <http://dl.acm.org/citation.cfm?id=3013558.3013571>.
- Chen, D., Manning, C., 2014. A fast and accurate dependency parser using neural networks. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pp. 740–750. <http://www.aclweb.org/anthology/D14-1082>.
- Chen, X., Lawrence Zitnick, C., 2015. Mind's eye: a recurrent visual representation for image caption generation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 2422–2431.
- Clark, P., Porter, B., Works, B.P., 2004. *Km-The Knowledge Machine 2.0: Users Manual*. Department of Computer Science, University of Texas at Austin.
- Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection. *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. 1. IEEE, pp. 886–893.
- Davis, E., Marcus, G., 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM* 58 (9), 92–103. <http://dx.doi.org/10.1145/2701413>.
- De Raedt, L., Kimmig, A., Toivonen, H., 2007. Problog: a probabilistic prolog and its application in link discovery. *Proceedings of the 20th International Joint Conference on Artificial Intelligence*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 2468–2473.
- Denkowski, M., Lavie, A., 2014. Meteor universal: Language specific translation evaluation for any target language. *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.
- Devlin, J., Cheng, H., Fang, H., Gupta, S., Deng, L., He, X., Zweig, G., Mitchell, M., 2015. Language models for image captioning: the quirks and what works. *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. 2. pp. 100–105. <http://aclweb.org/anthology/P/P15/P15-2017.pdf>. Short Papers.
- Divvala, S.K., Farhadi, A., Guestrin, C., 2014. Learning everything about anything: Webly-supervised visual concept learning. *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23–28, 2014*. pp. 3270–3277.
- Donahue, J., Hendricks, L.A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T., 2015. Long-term recurrent convolutional networks for visual recognition and description. In: *CVPR*.
- Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T., 2014. Decaf: a deep convolutional activation feature for generic visual recognition. *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. pp. 647–655.
- Elliott, D., Keller, F., 2013. Image description using visual dependency representations. *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*. pp. 1292–1302. <http://aclweb.org/anthology/D/D13/D13-1128.pdf>.
- Farhadi, A., Endres, I., Hoiem, D., Forsyth, D., 2009. Describing objects by their attributes. *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, pp. 1778–1785.
- Farhadi, A., Hejrati, M., Sadeghi, M.A., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D., 2010. Every picture tells a story: generating sentences from images. *Proceedings of the 11th European Conference on Computer Vision: Part IV*. Springer-Verlag, Berlin, Heidelberg, pp. 15–29. <http://dl.acm.org/citation.cfm?id=1888089>.
- Felzenszwalb, P., McAllester, D., Ramanan, D., 2008. A discriminatively trained, multi-scale, deformable part model. *Computer Vision and Pattern Recognition, 2008. CVPR 2008*. IEEE, pp. 1–8.
- Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., Xu, W., 2015. Are you talking to a machine? Dataset and methods for multilingual image question answering. *Proceedings of the 28th International Conference on Neural Information Processing Systems*. MIT Press, Cambridge, MA, USA, pp. 2296–2304. <http://dl.acm.org/citation.cfm?id=2969442.2969496>.
- Gatt, A., Reiter, E., 2009. Simplenlg: a realisation engine for practical applications. *Proceedings of the 12th European Workshop on Natural Language Generation. Association for Computational Linguistics, Stroudsburg, PA, USA*, pp. 90–93. <http://dl.acm.org/citation.cfm?id=1610195.1610208>.
- Gelfond, M., Lifschitz, V., 1988. *The Stable Model Semantics for Logic Programming*. MIT Press, pp. 1070–1080.
- Girshick, R., Donahue, J., Darrell, T., Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. *Comput. Vis. Pattern Recognit.*
- Gupta, A., Davis, L.S., 2007. Objects in action: An approach for combining action understanding and object perception. *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, pp. 1–8.
- Havasi, C., Speer, R., Alonso, J., 2007. Conceptnet 3: A Flexible, Multilingual Semantic Network for Common Sense Knowledge. *Recent Advances in Natural Language Processing*. Citeseer, pp. 27–29.
- Hodosh, M., Young, P., Hockenmaier, J., 2013. Framing image description as a ranking task: data, models and evaluation metrics. *J. Artif. Intell. Res.* 853–899.
- Karpathy, A., Li, F.-F., 2014. Deep visual-semantic alignments for generating image descriptions. *arXiv preprint arXiv:1412.2306*.
- Kilickaya, M., Erdem, A., Iklizler-Cinbis, N., Erdem, E., 2017. Re-evaluating automatic metrics for image captioning. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. 1. Association for Computational Linguistics, pp. 199–209. Long Papers. <http://www.aclweb.org/anthology/E17-1019>.
- Kimmig, A., Bach, S.H., Broecheler, M., Huang, B., Getoor, L., 2012. A short introduction to probabilistic soft logic. In: *NIPS Workshop on Probabilistic Programming: Foundations and Applications*.
- Kiros, R., Salakhutdinov, R., Zemel, R. S., 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*.
- Kollar, T., Roy, N., 2009. Utilizing object-object and object-scene context when planning to find things. *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*. IEEE, pp. 2168–2173.
- Krizhevsky, A., Sutskever, I., Hinton, G., 2013. Imagenet classification with deep convolutional neural networks. In: *NIPS* 2012.
- Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A.C., Berg, T.L., 2011. Baby talk: understanding and generating image descriptions. *Proceedings of the 24th CVPR*.
- Kuznetsova, P., Ordonez, V., Berg, A.C., Berg, T.L., Choi, Y., 2012. Collective generation of natural image descriptions. *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. 1. Long Papers. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 359–368. <http://dl.acm.org/citation.cfm?id=2390524.2390575>.
- Lampert, C.H., Nickisch, H., Harmeling, S., 2009. Learning to detect unseen object classes by between-class attribute transfer. *Computer Vision and Pattern Recognition (CVPR), 2009. IEEE*. pp. 951–958.
- Lan, T., Yang, W., Wang, Y., Mori, G., 2012. Image retrieval with structured object queries using latent ranking svm. In: *ECCV*.
- Laptev, I., 2005. On space-time interest points. *Int. J. Comput. Vis.* 64 (2–3), 107–123.
- Lebet, R., Pinheiro, P.H., Collobert, R., 2015. Phrase-based image captioning. *International Conference on Machine Learning (ICML)*.
- Lenat, D.B., 1995. Cyc: a large-scale investment in knowledge infrastructure. *Commun. ACM* 38 (11), 33–38. <http://dx.doi.org/10.1145/219717.219745>.
- Lin, D., 1998. An information-theoretic definition of similarity. *ICML*. 98. pp. 296–304.
- Lin, D., Fidler, S., Kong, C., Urtasun, R., 2015. Generating multi-sentence natural language descriptions of indoor scenes. In: *Xianghua Xie, M.W.J., Tam, G.K.L. (Eds.), Proceedings of the British Machine Vision Conference (BMVC)*. BMVA Press, pp. 93.1–93.13. <http://dx.doi.org/10.5244/C.29.93>.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft Coco: Common Objects in Context. *Computer Vision—ECCV 2014*. Springer, pp. 740–755.

- Lowe, D.G., 1999. Object recognition from local scale-invariant features. *Computer vision, 1999. The proceedings of the Seventh IEEE International Conference on*. 2. IEEE, pp. 1150–1157.
- Ma, L., Lu, Z., Li, H., 2016. Learning to answer questions from image using convolutional neural network. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. AAAI Press, pp. 3567–3573. <http://dl.acm.org/citation.cfm?id=3016387.3016405>.
- Ma, L., Lu, Z., Shang, L., Li, H., 2015. Multimodal convolutional neural networks for matching image and sentence. *2015 IEEE International Conference on Computer Vision (ICCV)*. pp. 2623–2631. <http://dx.doi.org/10.1109/ICCV.2015.301>.
- Malinowski, M., Rohrbach, M., Fritz, M., 2015. Ask your neurons: a neural-based approach to answering questions about images. *Proceedings of the IEEE International Conference on Computer Vision*. pp. 1–9.
- Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., Yuille, A., 2014a. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*.
- Mao, J., Xu, W., Yang, Y., Wang, J., Yuille, A. L., 2014b. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*.
- Marr, D., 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Co., Inc., New York, NY, USA.
- Messing, R., Pal, C., Kautz, H., 2009. Activity recognition using the velocity histories of tracked keypoints. *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, pp. 104–111.
- Ogale, A.S., Karapurkar, A., Aloimonos, Y., 2006. View-invariant modeling and recognition of human actions using grammars. In: Vidal, R., Heyden, A., Ma, Y. (Eds.), *WCV. Springer*, pp. 115–126. <http://dblp.uni-trier.de/db/conf/eccv/wdv2006.html#OgaleKA06>.
- Ordonez, V., Kulkarni, G., Berg, T.L., 2011. Im2text: describing images using 1 million captioned photographs. In: Shawe-Taylor, J., Zemel, R.S., Bartlett, P.L., Pereira, F.C.N., Weinberger, K.Q. (Eds.), *NIPS*, pp. 1143–1151. <http://dblp.uni-trier.de/db/conf/nips/nips2011.html#OrdonezKB11>.
- Paolacci, G., Chandler, J., Ipeirotis, P. G., 2010. Running experiments on amazon mechanical turk.
- Papineni, K., Roukos, S., Ward, T., Zhu, W.-J., 2002. Bleu: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pp. 311–318.
- Richardson, M., Domingos, P., 2006. Markov logic networks. *Mach. Learn.* 62 (1–2), 107–136.
- Santofimia, M., Martinez-del Rincon, J., Nebel, J.-C., 2012. Common-Sense Knowledge for a Computer Vision System for Human Action Recognition. In: Bravo, J., Hervás, R., Rodríguez, M. (Eds.), *Ambient Assisted Living and Home Care. Lecture Notes in Computer Science*, vol. 7657. Springer Berlin Heidelberg, pp. 159–166. [https://doi.org/10.1007/978-3-642-35395-6\\_22](https://doi.org/10.1007/978-3-642-35395-6_22).
- Schuster, S., Krishna, R., Chang, A., Fei-Fei, L., Manning, C.D., 2015. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. *Proceedings of the Fourth Workshop on Vision and Language*. Association for Computational Linguistics, pp. 70–80.
- Scutari, M., 2010. Learning Bayesian networks with the bnlearn R package. *J. Stat. Softw.* 35 (3), 1–22.
- Sharma, A., Vo, N.H., Aditya, S., Baral, C., 2015. Towards addressing the winograd schema challenge - building and using a semantic parser and a knowledge hunting module. *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25–31, 2015*. pp. 1319–1325. <http://ijcai.org/papers15/Abstracts/IJCAI15-190.html>.
- Socher, R., Karpathy, A., Le, Q.V., Manning, C.D., Ng, A.Y., 2014. Grounded compositional semantics for finding and describing images with sentences. *TACL* 2, 207–218. <http://www.transacl.org/wp-content/uploads/2014/04/52.pdf>.
- Teo, C.L., Fermüller, C., Aloimonos, Y., 2015. A gestaltist approach to contour-based object recognition: combining bottom-up and top-down cues. *Int. J. Rob. Res.* 0278364914558493.
- Teo, C.L., Myers, A., Fermüller, C., Aloimonos, Y., 2013. Embedding high-level information into low level vision: Efficient object search in clutter. *Robotics and Automation (ICRA)*, 2013 IEEE International Conference on. IEEE, pp. 126–132.
- Vinyals, O., Toshev, A., Bengio, S., Erhan, D., 2015. Show and tell: a neural image caption generator. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 3156–3164.
- Vinyals, O., Toshev, A., Bengio, S., Erhan, D., 2017. Show and tell: lessons learned from the 2015 MSCOCO image captioning challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* 39 (4), 652–663. <http://dx.doi.org/10.1109/TPAMI.2016.2587640>.
- Wang, H., Klaser, A., Schmid, C., Liu, C.-L., 2011. Action recognition by dense trajectories. *Computer Vision and Pattern Recognition (CVPR)*, 2011 IEEE Conference on. IEEE, pp. 3169–3176.
- Wiriyathammabhum, P., Summers-Stay, D., Fermüller, C., Aloimonos, Y., 2016. Computer vision and natural language processing: recent approaches in multimedia and robotics. *ACM Comput. Surv. (CSUR)* 49 (4), 71.
- Xiong, C., Merity, S., Socher, R., 2016. Dynamic memory networks for visual and textual question answering. *International Conference on Machine Learning*. pp. 2397–2406.
- Yang, M.Y., Liao, W., Ackermann, H., Rosenhahn, B., 2017. On support relations and semantic scene graphs. *{ISPRS} J. Photogramm. Remote Sens* 131, 15–25. <http://dx.doi.org/10.1016/j.isprsjprs.2017.07.010>.
- Yang, Y., Fermüller, C., Aloimonos, Y., Guha, A., 2014. A cognitive system for understanding human manipulation actions. *Adv. Cogn. Syst* 3, 67–86.
- Yang, Y., Teo, C.L., Daumé III, H., Aloimonos, Y., 2011. Corpus-guided sentence generation of natural images. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 444–454. <http://dl.acm.org/citation.cfm?id=2145432.2145484>.
- Yao, B.Z., Yang, X., Lin, L., Lee, M.W., Zhu, S.C., 2010. I2t: Image parsing to text description. *Proc. IEEE* 98 (8), 1485–1508. <http://dblp.uni-trier.de/db/journals/pieee/pieee98.html#YaoYLLZ10>.
- Yee, E., Chrysikou, E.G., Thompson-Schill, S.L., 2013. The Cognitive Neuroscience of Semantic Memory.
- You, Q., Jin, H., Wang, Z., Fang, C., Luo, J., 2016. Image captioning with semantic attention. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Yu, X., Aloimonos, Y., 2010. Attribute-based transfer learning for object categorization with zero/one training example. *Computer Vision–ECCV 2010*. Springer Berlin Heidelberg, pp. 127–140.
- Yu, X., Fermüller, C., Teo, C.L., Yang, Y., Aloimonos, Y., 2011. Active scene recognition with vision and language. *Computer Vision (ICCV)*, 2011 IEEE International Conference on. IEEE, pp. 810–817.
- Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A., 2014. Learning deep features for scene recognition using places database. In: *NIPS*.
- Zitnick, C.L., Parikh, D., 2013. Bringing semantics into focus using visual abstraction. *CVPR. IEEE*, pp. 3009–3016. <http://dblp.uni-trier.de/db/conf/cvpr/cvpr2013.html#ZitnickP13>.