Advances in Cognitive Systems 4 (2016) Submitted 4/2016; published 6/2016

DeepIU: An Architecture for Image Understanding

Somak Aditya SADITYA1@ASU.EDU
Chitta Baral CHITTA@ASU.EDU
Computing Science and Engineering, Arizona State University, Tempe, AZ 85287 USA

Yezhou Yang YZYANG@CS.UMD.EDU
Yiannis Aloimonos YIANNIS@CS.UMD.EDU
Cornelia Fermuller FER@UMIACS.UMD.EDU
Computer Science, University of Maryland, College Park, MD, 201740, USA

Abstract

Image Understanding is fundamental to systems that need to extract contents and infer concepts from images. In this paper, we develop an architecture for understanding images, through which a system can recognize the content and the underlying concepts of an image and, reason and answer questions about both using a visual module, a reasoning module, and a commonsense knowledge base. In this architecture, visual data combines with background knowledge and; iterates through visual and reasoning modules to answer questions about an image or to generate a textual description of an image. We first provide motivations of such a Deep Image Understanding architecture and then, we describe the necessary components it should include. We also introduce our own preliminary implementation of this architecture and empirically show how this more generic implementation compares with a recent end-to-end Neural approach on specific applications. We address the knowledge-representation challenge in such an architecture by representing an image using a directed labeled graph (called Scene Description Graph). Our implementation uses generic visual recognition techniques and commonsense reasoning[1] to extract such graphs from images. Our experiments show that the extracted graphs capture the syntactic and semantic content of an image with reasonable accuracy.

1. Introduction and Motivation

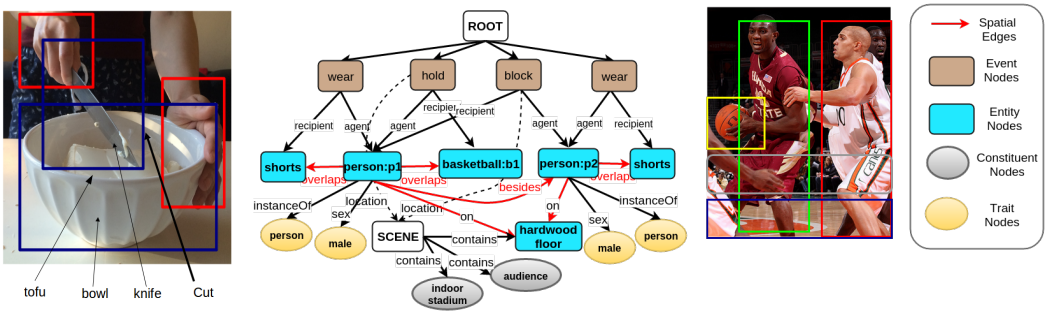
In Artificial Intelligence, the word "understanding" has been used in several contexts such as "Natural Language Understanding", "Image Understanding" etc. The general notion of "understanding" is well-studied in the domain of Education. In the Educational Domain, a student's "understanding" of a concept is evaluated by asking relevant questions about it. Similarly, "understanding" in an automated environment can be tested by asking questions and an intelligent system attempting to "understand" any concept should have the ability to answer them. Natural Language Understanding systems (Katz et al. (2001),Weston et al. (2015)) have applied such philosophy since its incep-

1. Commonsense reasoning and commonsense knowledge can be of many types (Davis & 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 activities.

tion and recently, the Computer Vision community (Gao et al. (2015),Antol et al. (2015)) has also adopted question-answering about images to evaluate systems that intend to understand images.

To achieve a human-level image "understanding" in artificial systems, we should also be able to measure the level or extent of understanding in such systems. In educational domain, we achieve this by modulating the difficulty of the questions. Bloom's taxonomy (Bloom et al. (1956)) classifies such questions into the categories: Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation; each focusing on testing increasingly difficult levels of cognitive thinking in students. In the context of an image such as Figure 1 (a), some example questions corresponding to the above categories could be: i) (Knowledge - demonstrating recall) *list* the objects in the image; ii) (Comprehension - demonstrating understanding) *predict* what the man will do next; iii) (Application - ability to apply the knowledge) *illustrate* how to cut tofu (or something similar to tofu); iv) (Analysis - ability to analyze and identify motives, causes) *why* is the man holding the bowl with his other hand; v) (Synthesis - ability to synthesize the information gathered and compile differently) *can* you propose how else to cut a tofu; vi) (Evaluation - ability to make judgment about information) *is* there a better way to cut a tofu.

The current systems are mainly evaluated using factoid questions (what, where, how many, is there etc.), belonging to the *Knowledge* category. These questions primarily tests the systems' capability to find, locate or collocate objects in the scene. But, these are only the tip of the iceberg and a large area of questions still remain completely unexplored. The primary reason is as the difficulty increases, so does the need of reasoning and inference using commonsense (and other kinds of) knowledge about the world and the current systems do not explicitly model commonsense knowledge. Consider Figure 1(a). It is often very hard to detect "tofu" in the bowl (Aditya et al. (2015b)). It will be near impossible if the image were shot from a lower elevation. In such a case, the system should be able to infer that *the knife might be cutting something inside the bowl, not the bowl itself*, i.e. in a QA setting, the answer to the question *"Is the knife cutting the bowl?"* should be *No*. This requires commonsense knowledge about the entities *knife, bowl* and the action *cut*.



(a) (b)

Figure 1: (a) An image where there could be objects which are indistinguishable even to the naked eye. (b) Example Image and an ideal SDG with spatial relations.

Again, for the image in Figure 1(b), one can ask a series of questions that requires commonsense knowledge about the physical world. Questions can range from the ones that require basic knowledge about the game of basketball ("*are the players in red and white belong to the same team?*") to the questions requiring more deeper knowledge such as the intuition of Physics ("*will the player*

in the right be able to block the player holding the ball?" or "*in which direction should the player holding the ball move?*"). Current state-of-the-art Image Captioning or QA systems however, is far from attempting such questions. Even the traditional definition of Image Understanding (Shapiro (1992)) only concerns itself with the space of systems that "produce descriptions of both the images and the world scenes that the images represent" given a goal or a reason for looking at a particular scene. So, henceforth in this work, we use the phrase **"Deep Image Understanding" (DeepIU)** to denote the study of architectures or systems better equipped to handle such a variety of questions.

A DeepIU system should facilitate i) generating description of the contents of the image, ii) answering factoid questions about the objects, regions and the events involved and iii) reasoning about the events and concepts in the scene using commonsense knowledge; and iv) updating the agent's belief about the image or even the global knowledge.

As many other cognitive systems, we look towards human beings to draw inspiration for the architecture. Human perception is active, selective and exploratory. 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 are interacting with each other. In this paper we propose to model the architecture that can support such an interaction.

Our architecture (DeepIU) essentially consists of the modules that reflect above intuition. The vision module detects objects, its *visible* properties; possible actions involving the objects and probable scenes. Guided by a commonsense knowledge-base and the reasoning module, we can then produce more semantic information hidden in the facts. Given these facts and inferred details, the reasoning module either goes back to vision module for more information or generates a knowledge structure that represents the (required) semantic and information content of the image. This knowledge structure can then be used for other applications such as Sentence Generation (template/statistical model based), Question-Answering systems.

One of the fundamental challenge of such a system is to come up with this knowledge structure that captures the information from the vision and the reasoning modules. This representation should also facilitate a seamless interaction between the commonsense knowledge, reasoning systems and QA systems. To this end, we propose Scene Description Graphs (SDG), a graphical representation of the semantic content and the information in images. In Figure 1, we show a possible SDG for an example image. SDG is a directed labeled graph[2] among entities (nouns), events (verbs) and traits (properties, superclasses of entities). The event nodes are connected to a dummy node *SCENE* by an edge labeled "location". The constituent nodes are separately color-coded to show inferred-concepts which cannot always be grounded to a region of the image. The spatial relations are inspired by Elliott & Keller (2013). In this work, we present a preliminary implementation that abides by the proposed architecture, obtains a Scene Description Graph from an image. This graph can be used to generate sentences or answer questions about the image.

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

2. DeepIU: A Deep Image Understanding Architecture

An image is a vast source of information. For example, in an image containing a tree: one might require the knowledge of leaves (which ones turned yellow), its branches and birds sitting on its branches. On the other end, we might want to know how many trees are there. All these information belong to different levels of granularity and focuses our attention to different regions in an image. In short, it does not seem plausible to gather complete knowledge about an image in one iteration.

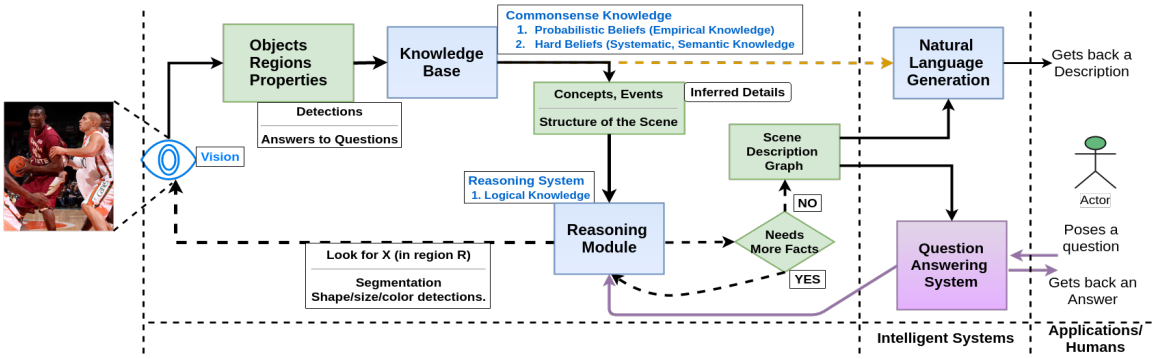


Figure 2: A cognitive architecture for Deep Image Understanding.

One way a human refines his own knowledge is to ask relevant questions based on his current knowledge and previously answered questions, and keep asking till he reaches a "plausible answer"/"enough information". To support such an exploratory behavior, a DeepIU architecture should support a **loop** of *..-reasoning-vision-reasoning-vision...* In Figure 2, we present our architecture supporting such a loop of vision and reasoning. The core of the architecture comprises of the following modules: i) Visual Detection, ii) Knowledge Base and iii) Logical Reasoning system. A system under this design should provide interfaces to: i) Sentence Generation and iii) Question-Answering system modules.

Visual Detection: Similar to the anatomic properties of a human body, a "Visual Detection" module should ideally resemble the functionalities of our eyes i.e. recognition and perception. The properties are as follows: i) (Objects and Regions) it should be able to detect objects and regions (such as human, water-body etc.); ii) (Scenes) for a better understanding of the scene, it should also be equipped to detect properties of the whole scene in view, for example, the scene resembles a platform; iii) (Properties) it should detect different properties (including spatial ones) of objects and regions (such as size, height, color of objects; color, shape of region; relative positions of two objects) etc. In computing terms, such detection generally amounts to different Image Processing techniques such as segmentation, shape-contour detection etc. Smarter techniques are being developed to detect relative sizes of objects (Bagherinezhad et al. (2016)); iv) (Attention) a visual detection module is also expected to interact with a reasoning module and hence, the former should have a proper interface for controlling "which detector to fire over which region of the image". Idealistically, such a Detection module might consist of a large set of Object Detection Classifiers, Scene Detection Classifiers, and Attribute (color, shape, size) Detection and Image Segmentation modules.

Knowledge Base: Different forms of background knowledge is essential to solving intelligent tasks by artificial intelligent systems and hence, a "Knowledge Base" is integral to any cognitive architecture. In a DeepIU architecture, we need commonsense knowledge[3] to especially answer the following questions: i) the probable events that the detected objects are participating in; ii) the past and future events that could be causally connected to such events; iii) ontological information about the probable scenes detected; iv) and lastly, a holistic background (ontological, spatial, commonsense etc.) knowledge pertaining to every object of the scene in view. It should be noted that, on a high level, two kinds of knowledge/beliefs are required i) probabilistic beliefs that are empirical and updated over time and, ii) hard beliefs that are rarely updated.

Reasoning System: A reasoning system in a DeepIU architecture is expected to "guide, predict and advise" (similar to the functionalities of "intelligence" defined in MacFarlane (2013)). In cases where it might need more information to answer a question or find a solution to a given task, it should be able guide and advise the Visual Detection module to search "what" and "where" in the image. Essentially, a logical reasoning system should represent the logical knowledge using a set of rules and should be able to traditionally perform deductive, inductive and abductive reasoning considering both probabilistic and hard beliefs. Traditional formalisms like Answer Set Programming are powerful representation languages; though the inference suffers from intractability problems and the usage of hard rules and facts. Whereas, much of the commonsense knowledge is probabilistic and in most cases, reasoning is performed with incomplete knowledge. Hence, it is important to use a Probabilistic adaptation of such logical systems in which rules and facts are not constrained to be binary and supports the agent's "imperfect" knowledge about the world. Further implementations of this architecture might adopt the languages such as Probabilistic Soft Logic (Bach et al. (2013)), Markov Logic Network (Richardson & Domingos (2006)) etc.

Iterate: Human beings often explore a scene in multiple iterations. In each iteration, our search for a particular "answer" becomes more targeted and focused. In the same way, a DeepIU architecture should have the capability to ask for more focused and targeted information. We can think of such a loop as asking the "right" question to the Visual Detection module or asking it to perform directed Image Processing on restricted regions of images. To achieve this, the reasoning system can output a possible question that guides the Visual Detection module to perform such processing, and the loop goes on till we find enough information. Another purpose that the "loop" can serve is to resolve ambiguity. If the vision module has detected two probable (equally low-confidence) objects, then reasoning with background knowledge might help us resolve this ambiguity.

In Table 1, we show some of the vision-reasoning-vision loop examples to answer questions of different levels of difficulty.

3. Current Architectures and Related Works

The effort to have a general architecture to understand and reason about images can be traced back to Marr (1982), in which D. Marr proposed three different levels: computational theory; representation and algorithm; and hardware implementation. Here, we focus on Marr's representation and algo-

3. The type of commonsense needed here can be compared with 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)).

	Knowledge
Questions	**Loop**
List the objects in the image.	*Vision - detect*: objects
	Comprehension
What will the man do next?	*Vision - detect*: objects, events *Reason - infer*: higher-level concept (e.g.: A kind of Food preparation) *Reason - output*: probable next-event of *cutting*
	Analysis
How will you cut tofu?	*Vision - detect*: objects (hands, tofu, knife, bowl), events (holding bowl, holding knife, cutting) *Reason - suggest*: detect hand-positions *Vision - detect*: hand-position *Reason* - Represent knowledge of the activity *cutting tofu* in terms of the object's relative locations and constituent actions. *Reason - describe*: the activity *cutting tofu*.
	Application
Why is the man holding the bowl with his other hand?	*Vision - detect*: objects (hands, tofu, knife, bowl), events (holding bowl, holding knife, cutting) *Reason - lookup*: background knowledge. *search* causes of *holding a bowl* (or holding an object) or *search* effects of *not holding bowl*.
	Synthesis
Propose an alternative method to cut a tofu.	*Vision - detect*: objects (hands, tofu, knife, bowl), events (holding bowl, holding knife, cutting) *Reason - lookup* background knowledge. *search* other methods of cutting tofu, or *search* for "cutting vegetables" (generalization).

Table 1: A Few Examples of the loop of *..-vision-reasoning-vision-..* to answer different categories of questions. A few black-box methods have been used to describe the action taken by each module: i) detect (fire object, action detectors), ii) suggest (guiding visual module to fire a detector), iii) lookup and search (query the knowledge base), iv) infer (infer (causally related) previous, next events; higher-level concepts), v) describe (natural language generation).

rithm level, and further extend it into three sub-levels: perception algorithm; image representation; and reason beyond appearance. A more general thinking of a cognitive architecture was presented by Langley et al. (2009), in which they argue an unified integration of perception, memory and reasoning is needed to perform cognitive tasks such as image understanding.

The first popular architecture of image understanding treats the problem as a one way feedforward process. The systems under this paradigm aim to extract meaningful information from images and videos. As Karpathy & Li (2014) suggests, some of the categories that these systems belong to are 1) dense image annotations, 2) generating textual descriptions, 3) grounding natural language in images, 4) neural networks in visual and language domains. A part of our work has some commonalities with the works of generating textual descriptions. This includes the works that retrieves and ranks sentences from training sets given an image such as Farhadi et al. (2010); Ordonez et al. (2011); Socher et al. (2014). Kulkarni et al. (2011), Yang et al. (2011)are some of the works that have generated descriptions by stitching together annotations or applying templates on detected image content. The main drawback of these image understanding architecture is the lack of a module to represent, organize and utilize the information extracted from images and videos.

The second architecture of image understanding goes one bit further, in which there have been works to represent the information content in images explicitly (Lan et al. (2012); Elliott & Keller (2013)). Recently, Johnson et al. (2015) introduced scene graphs to describe scenes and Schuster et al. (2015) creates scene graphs from descriptions, and Yang et al. (2014) proposed action

grammar to create activity trees to represent human manipulation actions. However, these representations of the image or video content is not designed for being utilized to do image based reasoning. In this work, our representation (SDGs) is automatically constructed from an image, and due to the event-entity-attribute based representation and meaningful edge-labels (borrowed from KM-ontology (Clark et al. (2004))), SDGs are more equipped to facilitate symbolic-level reasoning.

The third and very new architecture of image understanding goes even further to think about utilizing the extracted representation for image question answering. At reason beyond appearance level, several works have shown promising efforts to acquire and apply commonsense in different aspects of Scene Analysis. Zitnick & Parikh (2013) use abstraction to discover semantically similar images and Santofimia et al. (2012) uses common-sense to learn actions. In the field of **Visual Question Answering**, very recently researchers spent efforts on both creating challenging datasets and proposing new models (Antol et al. (2015); Malinowski et al. (2015); Gao et al. (2015); Ma et al. (2015a)). Both Malinowski et al. (2015) and Gao et al. (2015) use recurrent networks to encode the sentence and output the answer. The work from Ren et al. (2015) formulated the task as a classification problem and the work from Yang et al. (2015) approached image understanding as a continuous questioning and answering process. More recently, Xiong et al. (2016) proposed to use the architecture of Dynamic Memory Networks to model the episodic memory required for answering visual questions. The main drawback of these architectures are the missing of explicit representation of the knowledge. When the system produces wrong results, it is almost impossible to trace back the system and analyze the failure case.

The most related work to our architecture is Aloimonos & Fermüller (2015)) (and Summers-Stay (2013)) in which authors lay down the foundations of the "cognitive dialogue" (the loop) and proposes systems where vision is an active part of a reasoning system.

4. Our Preliminary Implementation

In this work, we provide a preliminary implementation[4] of the above architecture accompanied by experimental results on three popular Image Datasets (Flickr 8k, Flickr 30k and MS-COCO).

To understand images, we map the space of images to the space of text through SDGs. To do that, we first robustly define such mappings between regions of images to (meaningful) segments of text[5]. Let us assume that the fundamental units of an image (say $\mathcal{F}$) are the objects[6] and its *visible* properties (location, shape, size, color, contour etc.), regions and its *visible* properties, and actions. To avoid further complexity, we consider only those images, in which at least one fundamental unit ($f \in \mathcal{F}$) can be detected (by an ideal detector). Now, these units can be roughly mapped to words with the following parts-of-speech (POS) tags: nouns, verbs, adjectives, adverbs and prepositions. Next, we define some mappings explicitly to express the relations between the composition of these units in the space of images and phrases in the space of text.

4. For a more detailed description of this system, please check out the arXiv version Aditya et al. (2015a).
5. Karpathy & 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 phrases. 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 adverserial situations).
6. Objects can consist of visible, partly visible or occluded objects. If the object *person* is detected, occluded objects like organs in a body, can be inferred to be present using commonsense Knowledge Bases such as ConceptNet.

Observed Scene Constituents (OSC) are phrases or words that represent what we actually see in the image[7]. In a phrase, the individual words can identify an object, group of objects, their visible properties, regions or actions. For example: *person wearing shorts, person skateboarding, young person, kid smiling, people playing* etc. are all scene constituents.

Inferred Scene Constituents (ISC) are phrases or words that cannot be directly seen in the image, but can be inferred. For example *waiting room, open space, dark corners* are ISCs.

A **Scene** 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, such as: *a person is lying on a bench, in a park*; *a person is being evicted*; *a bank is being robbed*.

4.1 Visual Detection and Recognition

We use deep Object recognition, deep Scene (category) recognition and deep Observed Scene Constituent recognition as part of the Visual Detection module.

- 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 ILSVRC 2013 dataset.

- 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 $\mathcal{S}$).

- For deep constituent (OSC) recognition, we further augment the Flickr 8K image dataset with human annotation of constituents using Amazon Mechanical Turks. We specifically ask the annotators to annotate not only objects, but what objects are doing or properties of objects. We allow the labelers to use free-form text for describing constituents to reduce 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 frequent phrases (denoted as $\mathcal{C}$). Some of the top phrases are *dog run, dog play, kid play, person wear short* 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 do constituents recognition using these deep features.

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

4.2 Constructing SDGs from Visual Detections

Our Reasoning framework has three phases: i) the pre-processing phase, ii) the knowledge extraction and storage phase and iii) the reasoning module to infer a knowledge structure.

4.2.1 Pre-processing Phase

In this phase, we collect Ontological information (synonyms, hyponyms and hypernyms) about object classes in Object Meta-data table ($\mathcal{O}_T$) and scene categories in Scene Meta-data ($\mathcal{S}_M$). For scene categories, alongwith synonyms, hypernyms and hyponyms, we also collect correspondence

7. 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 bounding box for an action will be union of the bounding boxes of its constituent objects.

between each scene category and ISC. We hand-annotated all the ISCs for each scene category and learnt (by counting) a prior belief for each ISC in a scene from human annotations. For example, for the scene class *airport_terminal*, we add *{waiting room, big glass view, people}* as the list of ISCs and *air terminal* as the synonym; and learn the priors $0.7, 0.6$ and 0.9 respectively for ISCs. We also use scene category detection tuples ($\mathcal{S}_T$) and human annotations ($\mathcal{A}_d$) for all training images. For detections, we use the deep Scene (category) Recognition module from the previous section to detect top 5 scene categories from each training image.

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.

Knowledge Base: We used K-parser (`kparser.org`) for knowledge extraction from each sentence of the Image Annotations. Mainly, for a sentence such as "a man, laying on a bench, is eating cookies" K-parser extracts the events *eat, lay* and their participants as *person, cookie* and *person, bench* respectively, as a set of entity and event-nodes connected by meaningful labels. Internally, K-parser uses Stanford Parser to get the syntactic dependency graph. The K-parser then maps these dependency relations to the set of KM-Relations (Clark et al. (2004)) and some more newly created ones (see `http://bit.ly/1Wd8nGa`). The resulting graph is further augmented using ontological and semantic information from different sources (more details in Algorithm 1 in Sharma et al. (2015)). This K-parser output graph is then generalized i.e. entities are replaced by their superclasses (this creates consistency with the classifiers in $\mathcal{N}$). Then the graphs are merged incrementally based on overlapping entities and events, to create a single knowledge-graph ($\mathcal{K}_b = \langle \mathcal{G}, \mathcal{C} \rangle$). $\mathcal{G} = \langle V, E \rangle$ denoting set of labeled vertices V, set of labeled edges E. Each vertex can be of three types: *events, entities* and *traits*. **Events** correspond to verbs, **entities** correspond to nouns that directly interact with events and **traits** represent all other nouns or adjectives. **Edge labels** in the $\mathcal{K}_b$ are exactly the same as in the K-parser. $\mathcal{C}$ is a set of **scenes** which corresponds to generalized (nouns replaced by super-classes) K-parser graphs of sentences and is essentially a sub-graph of $\mathcal{G}$.

The Bayesian Network ($\mathcal{B}_n$): To capture the knowledge of naturally co-occurring entities ($\mathcal{N}$) and ISCs ($\mathcal{C}_{Is}$), we learn 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 $N = |\mathcal{N}| + |\mathcal{C}_{Is}|$). Each term t_i is binary and denotes 1 if the i^{th} entity (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 (Scutari (2010)).

To create $\mathcal{D}$, we process the annotations for each training image to automatically detect entities and ISCs. We parse the sentences using K-parser and extract entities. We match these entities with entities in ($\mathcal{N}$) 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 use the scene detection tuples $\mathcal{S}_T$ and we look-up all ISCs of the scene category with the highest score ($P_r(s|I_{tr})$), from $\mathcal{S}_M$.

4.2.3 Inference Through Knowledge and Reasoning

Equipped with the model $\langle \mathcal{K}_b, \mathcal{B}_n, \mathcal{S}_M, \mathcal{O}_T \rangle$, we use $\langle P_r(n|img), P_r(s|img), P_r(c|img) \rangle$ for an image ($img \in I$) to construct an SDG in the following way.

I. **Observed Scene Constituents**: We extract entities (nouns) and events (verbs) from top 10 constituents (based on $P_r(c|img)$) and add to the set of detections. For example, the constituent *person wearing sweatshirt* results in an event *wear* with two edges: one labeled *agent* joining the entity *person* and another labeled *recipient* joining the entity *sweatshirt*.

II. **Inferred Scene Constituents**: We look-up the ISCs for top 5 detected scenes (based on $P_r(s|img)$) from $\mathcal{S}_M$, and construct C_{freq}. Initially, $C_{inf} = \phi$, and $\mathcal{O}_{img} = \{n|P_r(n|img) > \alpha_h\}$. We calculate $S_{max} \leftarrow \text{argmax}_{s \in C_{freq}} P(s|C_{inf}, \mathcal{O}_{img})$ and add S_{max} to C_{inf}. We iterate while the entropy ($\sum_{s \in C_{freq}} \{-P(s|C_{inf}, \mathcal{O}_{img}) * \log P(s|C_{inf}, \mathcal{O}_{img})\}$) keeps decreasing (or while number-of-iterations is less than T[8]).

III. **Noisy Objects:** Next, we rectify the low-scoring entities based on $\mathcal{O}_{img}$ 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 $o_{max} = \text{argmax}_{o \in siblings} P(o|C_{inf}, \mathcal{O}_{img})$ and then add o_{max} to the high-scoring entities list ($\mathcal{O}_{img}$).

IV. **Inferring Scenes:** Given the inferred ISCs ($\mathcal{C}_{inf}$) and entities ($\mathcal{O}_{img}$), we find *scenes* that the image describes.

First, we find a co-occurring event for a pair of entities in $\mathcal{O}_{img}$ by considering the event-nodes on the 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 $\mathcal{K}_b$). 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 $\mathcal{E}_v$.

For filtering spurious events, we introduce the notion of *Edge-Compatible Events*. An event (ev) is **edge-compatible** with respect to two entities (e_1 and e_2) if they are connected to the event using edges with compatible labels (l_1 and l_2). Written more formally in horn-clause semantics[9]:

$edgeCompatible(ev, e_1, e_2) \leftarrow edge(ev, l_1, e_1, c_{img}) \wedge edge(ev, l_2, e_2, c_{img}) \wedge labelCompatible(l_1, l_2)$

The labels in K-parser (and in turn our $\mathcal{K}_b$) are interpreted based on definitions from KM-ontology and the label-compatibility is understood based on such interpretations. For example, *(agent, recipient)* is a compatible pair and only an animate entity can be an *agent*. Thus, the event *wear* is edge-compatible with respect to entities *person* and *trunk*.

To filter events such as *climb* etc, we consult the knowledge in $\mathcal{O}_T$ and the set of Scenes $\mathcal{C}$ and we retain only those events (ev) that are connected to an entity (e_1) which is of the same superclass (cl_1), in some Scene (c_1) in $\mathcal{C}$. Expressed in a formal way:

$valid(ev) \leftarrow edge(e_1, instanceOf, cl_1, c_1) \wedge edge(ev, l_1, e_1, c_1) \wedge edge(e_1, instanceOf, cl_1, query)$

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 using the remaining entities in ($\mathcal{O}_{img} \setminus \mathcal{O}_{ev}$) and ISCs; i.e., increase a counter if an entity or ISC occurs in the graph. We also calculate a joint confidence-score for each scene based on the $P_r(n|I), P_r(s|I), P_r(c|I)$ values

8. The hyper-parameters (T, α_h) are set based on performance on validation data.
9. In simplistic terms, the rule $A \leftarrow B \wedge C$ implies that A is true if B and C are both true. $edge(e_1, l, e_2, c)$ denotes there is an edge between nodes e_1 and e_2, labeled l_2 in Scene c.

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.

V. **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 $(\mathcal{O}_{img} \setminus \mathcal{O}_{ev})$: if it is an animate entity, add *edge(o_{im}, location, scene)*; Otherwise, find the shortest path from o_{im} to the top detected event in the $\mathcal{K}_b$ and add the edges on the path to the SDG.

VI. **Template Based Sentence Generation**: We generate sentences from the SDG using SimpleNLG (Gatt & 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[10]) of sentences using simple rules.

5. Preliminary Experiments and Results

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 8,092, 31,783 and more than 160K images respectively. All the images from these datasets are accompanied with 5 annotated sentences that describe the image. For all datasets, we used the train-test splits from Karpathy & Li (2014) and the 4000 testing images (1000 each from Flickr 8k and 30k; 2000 from MS-COCO validation set) serve as the testing bed for our experiments. On these datasets, 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 & Li (2014) as it was one of the recent (and among the first) neural approaches which produced best results over all the previous works.

Amazon Mechanical Turk (AMT) Evaluation of Generated Sentences: Since image description generation is innately a creative process, a good metric is to ask humans to evaluate these sentences. The evaluation metrics: Relevance and Thoroughness, are therefore proposed as empirical measures of how much the description conveys the image content (relevance) and how much of the image content is conveyed by the description (thoroughness). 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). The average of the scores and their deviation are summarized in Table 2. For comparison, we asked the AMTs to also judge one gold-standard description and the output from Karpathy & Li (2014).

Image-Sentence Alignment Evaluation: We evaluate the image-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 $\mathcal{G}_q = (V_q, E_q)$ using the same procedure by which we construct $\mathcal{K}_b$. For each image, we take the SDG $\mathcal{G}_{img} = (V_i, E_i)$ and calculate similarity between the SDG and the query using the formula:

$$Sim(\mathcal{G}_q, \mathcal{G}_{img}) = (\sum_{v_q \in V_q} max_{v_i \in V_i}(sim(v_q, v_i)))/|V_q|$$

$$sim(v_q, v_i) = (wnsim(label(v_q), label(v_i)) + Jaccard(neighbors(v_q), neighbors(v_i)))/2.$$

10. 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*.

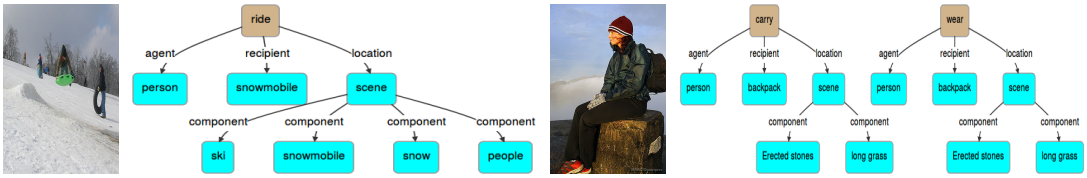
Experiment	Karpathy & Li (2014)	Our Method	Gold Standard
R $\pm$ D(8k)	2.08 ± 1.35	$\mathbf{2.82 \pm 1.56}$	4.69 ± 0.78
T $\pm$ D(8k)	2.24 ± 1.33	$\mathbf{2.62 \pm 1.42}$	4.32 ± 0.99
R $\pm$ D(30k)	1.93 ± 1.32	$\mathbf{2.43 \pm 1.42}$	4.78 ± 0.61
T $\pm$ D(30k)	2.17 ± 1.34	$\mathbf{2.49 \pm 1.42}$	4.52 ± 0.93
R$\pm$D(COCO)	$\mathbf{2.69 \pm 1.49}$	2.14 ± 1.29	4.71 ± 0.67
T$\pm$D(COCO)	$\mathbf{2.55 \pm 1.41}$	2.06 ± 1.24	4.37 ± 0.92

Table 2: Sentence generation relevance (R) and thoroughness (T) human evaluation results with gold standard and Karpathy & Li (2014) on Flickr 8k, 30k test images and COCO validation images. D: Standard Deviation.

Vertex-similarity is calculated based on their word-meaning similarity and neighbor similarity. Here $wnsim(.,.)$ is WordNet-Lin Similarity Lin (1998) between two words and $Jaccard(.,.)$ is the standard Jaccard coefficient similarity. Based on the above measure, we give the image retrieval results compared with results from Karpathy & Li (2014) in Table 3.

	Flickr8k			
Model	**R@1**	**R@5**	**R@10**	**Med r**
Karpathy & Li (2014) BRNN	11.8	32.1	44.7	12.4
Our Method-SDG	**18.1**	**39.0**	**50.0**	**10.5**
	Flickr30k			
Karpathy & Li (2014) BRNN	15.2	37.7	50.5	9.2
Our Method-SDG	**26.5**	**48.7**	**59.4**	**6.0**
	MS-COCO			
Karpathy & Li (2014) 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 3: 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.



(a) (b) (c) (d)

(e) (f)

Figure 3: The SDGs in (b), (d) and (f) corresponds to images (a), (c) and (e) respectively. **More examples: http://bit.ly/1NJycKO.**

Analysis: There are other works in Image Retrieval (Ma et al. (2015b)) and Caption Generation (Devlin et al. (2015)) which achieve better results than shown in Table 1 and 2. We believe that from motivational standpoint, our work is not directly comparable with such systems. To the best of our

knowledge, there are only two works Lan et al. (2012); Elliott & Keller (2013) which proposes automatic construction of semantic representation of images. The results of Karpathy & Li (2014) are better than these approaches and this is why we take one of the neural approaches for comparison.

5.1 Question-Answering (QA) Case Studies

In this section, we give a brief overview of the QA system that is currently under development. In this paper, we only provide the intuitions behind the system using some example images.

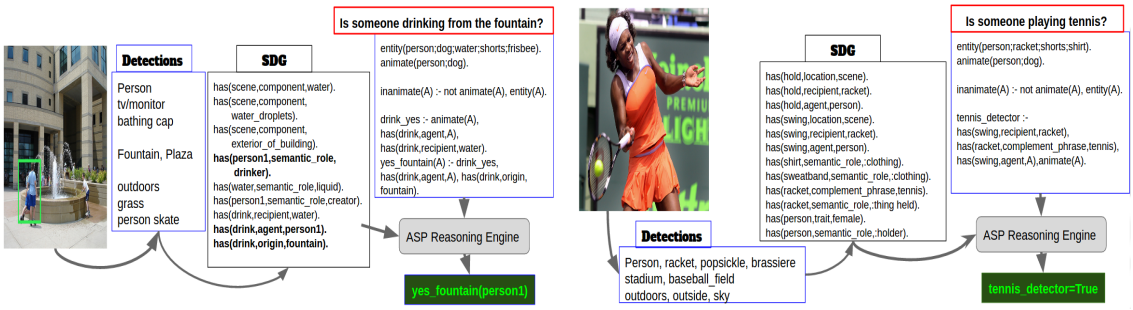


Figure 4: Two example Images from Flickr 8k. Note that for both the images, the state-of-the-art detections are quite noisy. Still, the current framework is able to detect "explainable" stories which can be queried upon.

For the image in Figure 4(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. If we pose the question that "Is someone drinking from the fountain?" in ASP (as shown in the figure), we can execute the above program in Clingo-3 and we get the answer as *yes_fountain(person1)*.

For the second image in Figure 4(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. 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".

For future work, we plan to extend this top-down preliminary implementation to support the loop of reasoning and vision, mainly starting with an interface where the visual module can be guided to detect specific objects (regions, properties) in specific locations in the image. For QA task, it is also important to guide the visual module from the beginning. For example, for the image in Figure 4(b), if the question is posed "is the woman wearing a headband", relevant information might not be directly obtained from a generated caption (or even an annotated one). Hence, it is important that vision is guided by the natural language question itself to generate relevant information and the crux of our QA system is such guidance. The details of this work is out of the scope of this paper.

6. Conclusion

This paper introduced an architecture to facilitate deep understanding of Images. We provide motivation about the necessities of such an architecture, followed by the necessary components it should include. In this work, we also elaborate on a preliminary implementation of this architecture and provide empirical results to show that this system is able to perform comparably to one of the recent Neural approaches. We identify the fundamental challenge of realization of such an architecture

as the representation of the knowledge in image and automatic derivation of such a representation. We introduce a novel intermediate semantic representation of scenes, namely the Scene Description Graph (SDG). The SDG is a representation of the scene that integrates direct visual knowledge (objects and their locations in the scene) with background commonsense knowledge. In addition, the SDGs have a structure similar to semantic representations of sentences, thus facilitating the interaction between Vision and Natural Language. Here we used the SDG for the automatic creation of sentences describing the scene; but, equipped with background knowledge, it also allows reasoning and question/answering about the scene.

References

Aditya, S., Yang, Y., Baral, C., Fermuller, C., & Aloimonos, Y. (2015a). From images to sentences through scene description graphs using commonsense reasoning and knowledge. *arXiv preprint arXiv:1511.03292*.

Aditya, S., Yang, Y., Baral, C., Fermuller, C., & Aloimonos, Y. (2015b). Visual common-sense for scene understanding using perception, semantic parsing and reasoning. *2015 AAAI Spring Symposium Series*.

Aloimonos, Y., & Fermüller, C. (2015). The cognitive dialogue: A new model for vision implementing common sense reasoning. *Image and Vision Computing*, *34*, 42–44.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). Vqa: Visual question answering. *ICCV*.

Bach, S., Huang, B., London, B., & Getoor, L. (2013). Hinge-loss markov random fields: Convex inference for structured prediction. *arXiv preprint arXiv:1309.6813*.

Bagherinezhad, H., Hajishirzi, H., Choi, Y., & Farhadi, A. (2016). Are elephants bigger than butterflies? reasoning about sizes of objects. *CoRR*, *abs/1602.00753*. URL `http://arxiv.org/abs/1602.00753`.

Bloom, B. S., et al. (1956). Taxonomy of educational objectives. vol. 1: Cognitive domain. *New York: McKay*, (pp. 20–24).

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*.

Davis, E., & Marcus, G. (2015). Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, *58*, 92–103.

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. *CoRR*, *abs/1505.01809*. URL `http://arxiv.org/abs/1505.01809`.

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).

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* (pp. 15–29).

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. *arXiv preprint arXiv:1505.05612*.

Gatt, A., & Reiter, E. (2009). Simplenlg: A realisation engine for practical applications. *Proceedings of the 12th European Workshop on Natural Language Generation* (pp. 90–93). Stroudsburg, PA, USA: Association for Computational Linguistics.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *Computer Vision and Pattern Recognition*.

Havasi, C., Speer, R., & Alonso, J. (2007). Conceptnet 3: a flexible, multilingual semantic network for common sense knowledge. *Recent advances in natural language processing* (pp. 27–29). Citeseer.

Hodosh, M., Young, P., & Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, (pp. 853–899).

Johnson, J., Krishna, R., Stark, M., Li, J., Bernstein, M., & Fei-Fei, L. (2015). Image retrieval using scene graphs. *IEEE CVPR*.

Karpathy, A., & Li, F.-F. (2014). Deep visual-semantic alignments for generating image descriptions. *arXiv preprint arXiv:1412.2306*.

Katz, B., Lin, J., & Felshin, S. (2001). Gathering knowledge for a question answering system from heterogeneous information sources. *Proceedings of the workshop on Human Language Technology and Knowledge Management-Volume 2001* (p. 9). ACL.

Kimmig, A., Bach, S. H., Broecheler, M., Huang, B., & Getoor, L. (2012). A short introduction to probabilistic soft logic. *NIPS Workshop on Probabilistic Programming: Foundations and Applications*.

Krizhevsky, A., Sutskever, I., & Hinton, G. (2013). Imagenet classification with deep convolutional neural networks. *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*.

Lan, T., Yang, W., Wang, Y., & Mori, G. (2012). Image retrieval with structured object queries using latent ranking svm. *ECCV*.

Langley, P., Laird, J. E., & Rogers, S. (2009). Cognitive architectures: Research issues and challenges. *Cogn. Syst. Res.*, *10*, 141–160.

Lenat, D. B. (1995). Cyc: A large-scale investment in knowledge infrastructure. *Commun. ACM*, *38*, 33–38.

Lin, D. (1998). An information-theoretic definition of similarity. *ICML* (pp. 296–304).

Ma, L., Lu, Z., & Li, H. (2015a). Learning to answer questions from image using convolutional neural network. *arXiv preprint arXiv:1506.00333*.

Ma, L., Lu, Z., Shang, L., & Li, H. (2015b). Multimodal convolutional neural networks for matching image and sentence. *CoRR*, *abs/1504.06063*. URL `http://arxiv.org/abs/1504.06063`.

MacFarlane, A. (2013). Information, knowledge and intelligence. URL `\url{https://philosophynow.org/issues/98/Information_Knowledge_and_Intelligence}`.

Malinowski, M., Rohrbach, M., & Fritz, M. (2015). Ask your neurons: A neural-based approach to answering questions about images. *arXiv preprint arXiv:1505.01121*.

Marr, D. (1982). A computational investigation into the human representation and processing of visual information.

Ordonez, V., Kulkarni, G., & Berg, T. L. (2011). Im2text: Describing images using 1 million captioned photographs. *NIPS* (pp. 1143–1151). URL `http://dblp.uni-trier.de/db/conf/nips/`

nips2011.html#OrdonezKB11.

Ren, M., Kiros, R., & Zemel, R. (2015). Exploring models and data for image question answering. *arXiv preprint arXiv:1505.02074*.

Richardson, M., & Domingos, P. (2006). Markov logic networks. *Machine learning*, *62*, 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 J. Bravo, R. Hervás, & M. Rodríguez (Eds.), *Ambient Assisted Living and Home Care*, volume 7657 of *Lecture Notes in Computer Science*, (pp. 159–166). Springer Berlin Heidelberg.

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* (pp. 70–80). Lisbon, Portugal: Association for Computational Linguistics.

Scutari, M. (2010). Learning bayesian networks with the bnlearn R package. *Journal of Statistical Software*, *35*, 1–22.

Shapiro, S. C. (1992). *Encyclopedia of Artificial Intelligence*. New York, NY, USA: John Wiley & Sons, Inc., 2nd edition.

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).

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.

Summers-Stay, D. A. (2013). Productive vision: Methods for automatic image comprehension.

Weston, J., Bordes, A., Chopra, S., & Mikolov, T. (2015). Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

Xiong, C., Merity, S., & Socher, R. (2016). Dynamic memory networks for visual and textual question answering. *arXiv preprint arXiv:1603.01417*.

Yang, Y., Fermüller, C., Aloimonos, Y., & Guha, A. (2014). A cognitive system for understanding human manipulation actions. *Advances in Cognitive Systems*, *3*, 67–86.

Yang, Y., Li, Y., Fermuller, C., & Aloimonos, Y. (2015). Neural self talk: Image understanding via continuous questioning and answering. *arXiv preprint arXiv:1512.03460*.

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* (pp. 444–454). Stroudsburg, PA, USA: ACL.

Yee, E., Chrysikou, E. G., & Thompson-Schill, S. L. (2013). The cognitive neuroscience of semantic memory.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., & Oliva, A. (2014). Learning deep features for scene recognition using places database. *NIPS*.

Zitnick, C. L., & Parikh, D. (2013). Bringing semantics into focus using visual abstraction. *CVPR* (pp. 3009–3016). IEEE.