Explainable Image Understanding Using Vision and Reasoning

Somak Aditya
Department of Computer Science
Arizona State University
saditya1@asu.edu

Introduction

Image Understanding is fundamental to intelligent agents. Researchers have explored Caption Generation and Visual Question Answering as independent aspects of Image Understanding (Johnson et al. 2015; Xiong, Merity, and Socher 2016). Common to most of the successful approaches, are the learning of end-to-end signal mapping (image-to-caption, image and question to answer). The accuracy is impressive. However, t is also important to explain a decision to end-user (justify the results, and rectify based on feedback). Very recently, there has been some focus (Hendricks et al. 2016; Liu et al.) on explaining some aspects of the learning systems. In my research, I look towards building explainable Image Understanding systems that can be used to generate captions and answer questions. Humans learn both from examples (*learning*) and by reading (*knowledge*). Inspired by such an intuition, researchers have constructed Knowledge-Bases that encode (probabilistic) commonsense and background knowledge. In this work, we look towards efficiently using this probabilistic knowledge on top of machine learning capabilities, to rectify noise in visual detections and generate captions or answers to posed questions.

Current Research

My current work on Image Understanding spans from outlining broader pictures to developing systems for specific applications. Here, I briefly outline the projects[1].

Visual Common-sense for Scene Understanding using Perception, Semantic Parsing and Reasoning: In this work (Aditya et al. 2015), we combine visual processing with techniques from natural language understanding (especially semantic parsing), common-sense reasoning and knowledge representation and reasoning to improve visual perception to reason about finer aspects of activities. We show how from a video about "making a line" shot in a constrained setting, we are able to answer questions such as (a) Which hand is being used in aligning the ruler? (b) Is the ruler aligned when the pen is drawing on the plank? We leverage the concept of actions and fluents in Answer Set Programming language to answer the above questions. In another constrained setting, we formulate how to detect higher level activity such as "making a sandwich" or "robbery" from fine-grained detailed actions. In several such scenarios, one might not have enough data to train on and might be completely constrained on one or two examples and background knowledge.

Image Understanding Through Scene Description Graphs: A popular formulation of the problem of Image Understanding is in terms of the task of generating meaningful textual description from images. Current deep-learning based image region-to-text grounding based methods (Karpathy and Li 2014; Xu et al. 2015) have shown amazing results. But, we are still far from "understanding" or answering deep questions about an image. Motivated by the preliminary reasoning-based approaches in (Aditya et al. 2015), we developed a combined Perception and reasoning module to retrieve a knowledge-structure (a graph that represents knowledge) from a static image. The visual perception module produces an initial set of objects, scenes and constituent detections. Our reasoning module then rectifies the object detections, predicts correct constituents and then creates a knowledge-structure. To predict how the objects interact in the scene, we build a common-sense knowledge base from image annotations along with co-occurrence frequency table of commonly occurring objects and abstract concepts. With these two precomputed knowledge-sources, we infer the following: 1) the correct set of correlated objects; 2) the most probable events (*verbs*) that these objects participate in; 3) the roles that the objects play in this event; and 4) given the events, objects and constituents, the concept that emerges from such information. Based on these inferences, we output a Scene Description Graph (SDG) (Figure 1) that depicts how these different entities and events interact. SDG is essentially a directed labeled graph among entities and events that enables an array of possibilities to do further analysis beyond visual appearance, such as event-entity based analysis, question answering about the scene and flexible caption generation[2]. Our initial experiments on the sentences generated from SDG shows that we achieve

[1]Most of these works are jointly done with Dr. Yezhou Yang (currently Assistant Professor, ASU) and Prof. Yiannis Aloimonos of University of Maryland, College Park.

[2]Such structures are also generated by Semantic parsers such as K-parser (www.kparser.org).

comparable accuracy in describing images as (Karpathy and Li 2014).

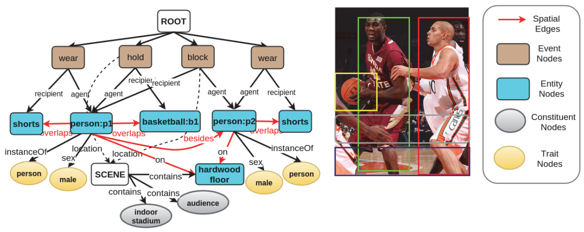

Figure 1: Example Image and an ideal SDG.

DeepIU: An Architecture for Image Understanding: In this work (Aditya et al. 2016a), we extend the above implementation and propose an architecture for understanding images. 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. The architecture is depicted in Figure 2. A human often refines his own

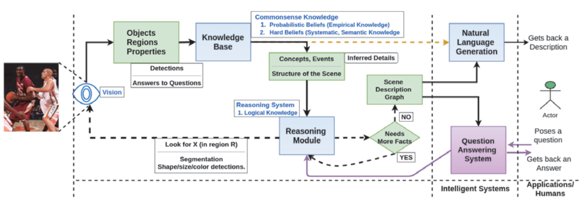

Figure 2: An architecture for Deep Image Understanding.

knowledge by asking relevant questions based on his current knowledge. 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.

Image Riddles: In this work, we explore a genre of puzzles where a set of images are provided and the task is to answer "what word connects these images?". We call these problems "Image Riddles". One such example is provided in Figure 3. In this example, the answer is "fall" as the word

Figure 3: Question: "What word connects these images?" (See our results in (Aditya et al. 2016b)).

logically connects all four images (the fall season, waterfall, rainfall and statue falling). We discuss the capabilities a system should possess to provide a logical answer to the riddles. We argue that similar capabilities are expected from a Visual Question Answering System. We develop a probabilistic approach that utilizes (probabilistic) commonsense knowledge about words and phrases to answer these riddles with a reasonable accuracy. We collected a dataset of over 3k riddles where each riddle consists of 4 images which are annotated with a groundtruth answer. The annotations are validated using crowd-sourced evaluation. We validate the results of our approach on these riddles using both automatic and human evaluations. Our approach based on Probabilistic Soft Logic achieves a few interesting results on riddles that are apparently harder for humans.

Future Research

In future, I wish to explore explainable Visual Question Answering and Caption Generation by employing a combination of Deep Learning and Probabilistic Soft Logic (PSL)-based Reasoning modules. We combine the pattern learning capability of the Neural approaches and background knowledge to obtain Knowledge Structures. For Question Understanding, we will explore different learning techniques to learn Semantic Parses of questions, *consistent* with the Knowledge Structures of images. These structures then can be used to generate short captions or fed into a PSL rule-base to answer questions. Weights of these rules can be learnt or calculated using additional knowledge.

References

Aditya, S.; Yang, Y.; Baral, C.; Fermuller, C.; and Aloimonos, Y. 2015. Visual Commonsense for Scene Understanding Using Perception, Semantic Parsing and Reasoning. In *2015 AAAI Spring Symposium Series*.

Aditya, S.; Baral, C.; Yang, Y.; Aloimonos, Y.; and Fermuller, C. 2016a. DeepIU: An Architecture for Image Understanding. In *Advances of Cognitive Systems*.

Aditya, S.; Yang, Y.; Baral, C.; and Aloimonos, Y. 2016b. Answering Image Riddles using Vision and Reasoning through Probabilistic Soft Logic. *arXiv preprint arXiv:1611.05896*.

Hendricks, L. A.; Akata, Z.; Rohrbach, M.; Donahue, J.; Schiele, B.; and Darrell, T. 2016. Generating visual explanations. *CoRR* abs/1603.08507.

Johnson, J.; Krishna, R.; Stark, M.; Li, J.; Bernstein, M.; and Fei-Fei, L. 2015. Image Retrieval using Scene Graphs. In *IEEE CVPR*.

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

Liu, J.; Cheng, H.; Javed, O.; Yu, Q.; Chakraborty, I.; Zhang, W.; Divakaran, A.; Sawhney, H. S.; Allan, J.; Manmatha, R.; et al. Multimedia event detection and recounting.

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

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2048–2057.