Integrating Knowledge and Reasoning in Image Understanding

Somak Aditya[1] , **Yezhou Yang**[2] , **Chitta Baral**[2]
[1]Bigdata Experience Lab, Adobe Research; [2]CIDSE, Arizona State University, USA
saditya@adobe.com, {yz.yang, chitta}@asu.edu

Abstract

Deep learning based data-driven approaches have been successfully applied in various image understanding applications ranging from object recognition, semantic segmentation to visual question answering. However, the lack of knowledge integration as well as higher-level reasoning capabilities with the methods still pose a hindrance. In this work, we present a brief survey of a few representative reasoning mechanisms, knowledge integration methods and their corresponding image understanding applications developed by various groups of researchers, approaching the problem from a variety of angles. Furthermore, we discuss upon key efforts on integrating external knowledge with neural networks. Taking cues from these efforts, we conclude by discussing potential pathways to improve reasoning capabilities.

1 Introduction

From the early years of computer vision research, many researchers realized that prior knowledge could help in different tasks ranging from low-level to high-level image understanding. For example, knowledge about the shape of an object can act as a strong prior in segmentation tasks [13; 43], or knowledge about the most probable action given a subject and the object can aid in action recognition tasks [14; 36; 41]. In this recent era of data-driven techniques, most of this knowledge is hoped to be learned from training data. While this is a promising approach, annotated data can be scarce in certain situations, and many domains have a vast amount of knowledge curated in form of text (structured or unstructured) that can be utilized in such cases. Utilization of background knowledge in data-scarce situations is one of the reasons that necessitate the development of approaches that can utilize such knowledge (from structured or unstructured text) and reason on that knowledge. Additionally, the lack of reasoning and inference capabilities (such as counterfactual, causal queries) of deep learning systems recently started to resurface in various forums [1; 17]. Motivated by these challenges, our goal in this paper is to present a survey of recent works (including a few of our works) in image understanding where knowledge and reasoning plays an important role. While discussing these interesting applications, we introduce corresponding reasoning mechanisms, knowledge sources, and argue the rationale behind their choice. *Lastly, we discuss different mechanisms that integrate external knowledge sources directly with deep neural networks.*

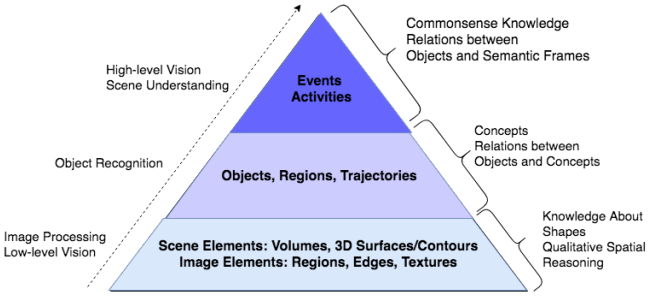

Figure 1: Diagram shows the information hierarchy for images and the knowledge associated with each level of information.

To understand what knowledge is meaningful in images, we can look at the different types of knowledge that relate to different levels of the semantic hierarchy induced by a natural image. Natural images are compositional. A natural image is composed of objects, and regions. Each object is composed of parts that could be objects themselves and regions can be composed of semantically meaningful sub-regions. This compositionality induces a natural hierarchy from pixels to objects (and higher level concepts). We show a diagram representing the information hierarchy induced by a natural image in Figure 1. Different types of knowledge might be relevant in the context of low-level information (objects and their parts) to higher-level semantics (abstract concepts, actions). Essentially, in this survey, we will study how knowledge and reasoning are applicable to these following levels of semantics: i) objects, regions and their attributes, ii) object-object or object-region interactions, relations and actions; iii) high-level commonsense knowledge (about events, activities).

In this work, we revolve the stories around different image applications ranging from object classification to question answering. As each new reasoning engine and knowledge source is encountered, we introduce them in individual

boxes. We provide brief critique as to why the chosen mechanisms were appropriate for the corresponding application. We discuss different ways to integrate knowledge in the deep learning era. The following sections then summarize the reasoning mechanisms. We conclude by shedding light on how the research in high-level reasoning and utilization of commonsense knowledge in Computer Vision can progress.

2 Use of Knowledge in Computer Vision

Here we describe applications that utilized relevant background knowledge beyond annotated data. Applications are categorized according to the levels of hierarchy in Fig. 1.

2.1 Knowledge about Objects, Regions, Actions

Image Classification: Various groups of researchers demonstrated the use of knowledge bases (generic and application-specific) in object, scene and action recognition; or reason about their properties.

Markov Logic Network: MLN ([28]) is a popular framework that uses weighted First Order Logical formulas to encode an undirected grounded probabilistic graphical model. Unlike PSL, the MLN is targeted to use the full expressiveness of First Order Logic and induce uncertainty in reasoning by modeling it using a graphical model. Formally, an MLN L is a set of pairs $\langle F, w \rangle$, where F is a first order formula and w is either a real number or a symbol α denoting hard weight. Together with a finite set of constants C, a Markov Network $M_{L,C}$ is defined where: i) $M_{L,C}$ contains one binary node for each grounding of each predicate appearing in L; ii) $M_{L,C}$ contains one feature for each grounding of each formula F_i in L. The value of feature is 1 if ground formula is true otherwise 0. The probability distribution over possible worlds x specified by the ground Markov Network $M_{L,C}$ is given by:

$$P(X = x) = \frac{1}{Z} \exp(\sum_i w_i n_i(x)) = \frac{1}{Z} \prod_i \phi_i(x_i)^{n_i(x)},$$

where $n_i(x)$ is the number of true groundings of the formula F_i in the world x. The MLN inference is equivalent to finding the maximum probable world according to the above formula. Weights are learnt using maximum likelihood methods.

Authors in [45] successfully used Markov Logic Network (MLN) in the context of reasoning about object affordances in images. An example of affordance is *fruit is edible*. Authors collect such assertions from textual cues and image sources, and complete the knowledge base using weighted rules in MLN. Example of collected assertions are *basketball is rollable and round, apple is rollable, edible and a fruit, pear is edible and a fruit* etc. The weights of the grounded rules are learnt by maximizing the pseudo-likelihood given the evidence collected from textual and image sources. Few such weighted rules are:

$$0.82\ \text{hasVisualAttribute}(x, Saddle) \implies \text{hasAffordance}(x, SitOn).$$
$$0.75\ \text{hasVisualAttribute}(x, Pedal) \implies \text{hasAffordance}(x, Lift).$$
$$\text{isA}(x, Animal) \wedge \text{locate}(x, Below).$$

This knowledge base encoded in MLN is used to infer the affordance relationships given the detected objects (and their confidence scores) in an image. Besides performing well over pure SVM-based methods, authors also observe that the knowledge-based approach is more robust against removal of visual attributes.

Appropriateness: For modeling object affordances, the authors in [45] faced the following challenges: i) uncertainty to account for noise and practical variability (if a *pedal* is dis-functional, it cannot be *lifted*); ii) expressive beyond IF-THEN horn-clauses, iii) relational knowledge. Encoding relational knowledge and modeling uncertainty warranted the use of a probabilistic logical mechanisms. As Probabilistic Soft Logic (PSL) can not express beyond horn clauses, and Problog [12] solvers are comparatively slow - MLN was a logical choice for this application.

Authors in [32] proposes Logic Tensor Network (LTN) that combines logical symbolism and automatic learning capabilities of neural network. For object-type classification and PartOf relation detection on PASCAL-PART-DATASET, LTNs with prior knowledge is shown to improve over Fast-RCNN. This work has started off many contributions in the area of neuro-symbolic reasoning such as DeepProbLog [24], end-to-end neural networks using prolog [29].

Logic Tensor Network: In LTN, soft logic is used to represent each concept as a predicate, for example $apple(x)$ is used to represent *apple*. The first-order formula $\forall x\ apple(x) \wedge red(x) \rightarrow sweet(x)$ represents that all red apples are sweet. Here, the truth values of each ground predicates are between 0 to 1, and truth values of conjunctive or disjunctive formulas are computed using combinations functions such as Lukasiewicz's T-norm. To combine this idea of soft logic with end-to-end learning, each concept or predicate is represented by a neural network and objects are represented by points in a vector space. The neural network for "apple" takes a point in the feature space and outputs its confidence about the input being a member of the "apple" concept. The weights in the neural networks are optimized to abide by rules such as $\forall x\ apple(x) \wedge red(x) \rightarrow sweet(x)$. These symbolic rules are added as constraints in the final optimization function.

The authors in [21] uses the commonsense knowledge encoded in ConceptNet to enhance the language model and apply this knowledge to two recognition scenarios: action recognition and object prediction. The authors also carried out a detailed study of how different language models (window-based model topic model, distributional memory) are compatible with the knowledge represented in images. For action recognition, authors detect the human, the object and scenes from static images, and then predict the most likely verb using the language model. They use object-scene, verb-scene and verb-object dependencies learnt from the language models to predict the final action in the scene. Examples of relations extracted from ConceptNet are: `Oil-Located near-Car, Horse-Related to-Zebra`. The conditional probabilities are computed using the frequency counts of these relations. To jointly predict the action i.e. $\langle subject, verb, object \rangle$ triplet the from object, the scene probability and the conditional probabilities

from language model, an energy based model is used that jointly reasons on the image (observed variable), object, verb and the scene.

ConceptNet Some well-known large-scale commonsense knowledge bases about the natural domain are ConceptNet, ([15]), WordNet ([26]), YAGO ([35]) and Cyc. Among these, ConceptNet is a semi-curated multilingual Knowledge Graph, that encodes commonsense knowledge about the world and is built primarily to assist systems that attempts to understand natural language text. The nodes (called concepts) in the graph are words or short phrases written in natural language. The nodes are connected by edges which are labeled with meaningful relations, such as $\langle reptile, IsA, animal \rangle$, $\langle reptile, HasProperty, coldblood \rangle$. Each edge has an associated confidence score. Being semi-curated, ConceptNet has the advantage of having a large coverage yet less noise.

Probabilistic Soft Logic (PSL): Similar to MLN, PSL ([7]) uses a set of weighted First Order Logical rules of the form $w_j : \vee_{i \in I_j^+} y_i \leftarrow \wedge_{i \in I_j^-} y_i$, where each y_i and its negation is a literal. The set of grounded rules is used to declare a Markov Random Field, where the confidence scores of the literal is treated a continuous valued random variable. Specifically, a PSL rule-base is used declare Hinge-Loss MRF, which is defined as follows: Let $\boldsymbol{y}$ and $\boldsymbol{x}$ be two vectors of n and n' random variables respectively, over $D = [0,1]^{n+n'}$. Let $\tilde{D} \subset D$, which satisfies a set of inequality constraints over the random variables. A *Hinge-Loss MRF* $\mathbb{P}$ is a probability density over D, defined as: if $(\boldsymbol{y}, \boldsymbol{x}) \notin \tilde{D}$, then $\mathbb{P}(\boldsymbol{y}|\boldsymbol{x}) = 0$; if $(\boldsymbol{y}, \boldsymbol{x}) \in \tilde{D}$, then:

$$\mathbb{P}(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{w})} exp(-f_{\boldsymbol{w}}(\boldsymbol{y}, \boldsymbol{x})).$$

Using PSL, the hinge-loss energy function $f_{\boldsymbol{w}}$ is defined as:

$$f_{\boldsymbol{w}}(\boldsymbol{y}) = \sum_{C_j \in \boldsymbol{C}} w_j \, max\{1 - \sum_{i \in I_j^+} V(y_i) - \sum_{i \in I_j^-} (1 - V(y_i)), 0\},$$

where $max\{1 - \sum_{i \in I_j^+} V(y_i) - \sum_{i \in I_j^-} (1 - V(y_i)), 0\}$ is distance to satisfaction of a grounded rule C_j. MPE inference in HL-MRFs is equivalent to finding a feasible minimizer for the convex energy function, and in PSL it is equivalent to $\arg\min_{\boldsymbol{y} \in [0,1]^n} f_{\boldsymbol{w}}(\boldsymbol{y})$. To learn the parameters $\boldsymbol{w}$ of an HL-MRF from the training data, maximum likelihood estimation (and its alternatives) is used.

Actions and Activities: Authors in [23] uses PSL to detect collective activities (i.e. activity of a group of people) such as *crossing, queuing, waiting* and *dancing* in videos. This task is treated as a high-level vision task, whereby detection modules are employed to extract information from the frames of the videos and such information (class labels and confidence scores of predicates) is input to the joint PSL model for reasoning. To obtain frame-level and person-level activity beliefs, human figures are represented using HOG features and Action Context (AC) descriptors. Then an SVM is used to obtain activity beliefs using these AC descriptors. Next, a collection of PSL rules is used to declare the ground HL-MRF to perform global reasoning about collective activities:

$$\text{LOCAL}(B, a) \implies \text{DOING}(B, a).$$
$$\text{FRAME(B, F)} \wedge \text{FRAMELBL(F, A)} \implies \text{DOING(B, A)}.$$

The intuitions behind the two rules are: Rule R1 corresponds to beliefs about local predictions using HOG features, and R2 expresses the belief that if many actors in the current frame are doing a particular action, then perhaps everyone is doing that action. To implement this, a *FrameLbl* predicate for each frame is computed by accumulating and normalizing the *Local* activity beliefs for all actors in the frame. Similarly, there are other rules that captures the intuition about these activities. Using PSL inference, final confidence scores are obtained for each collective activity, and authors observe that using PSL over baseline HOG features achieves significant increase (10%) in accuracy.

Appropriateness: For modeling group activities, the authors in [23] faced the following challenges: i) uncertainty to account for noise in real-world data and noisy predictions form machine learning models; ii) a fast scalable mechanism to predict activity classes. In the presence of large data, it can be debated that deep learning models are the de facto standards for any kind of classification. However, smaller datasets and the requirement of interpretability warranted a logical reasoning language. Requirement of robustness to noise and scalability warranted the use of PSL, as its underlying optimization problem is convex and its found to be faster.

Infrequently Used Logical Languages: The reasoning mechanisms discussed in this survey are chosen based on the following considerations: i) plausible inference, ii) learning capability, iii) expressiveness, and iv) speed of inference [10]. Several other logical languages have factored in useful aspects such as uncertainty, spatio-temporal reasoning etc. Qualitative Spatial Reasoning [27] languages and description logic [6] are noteworthy among them for image understanding. A popular representation formalism in QSR is Region Connection Calculus (RCC) introduced in [27]. The RCC-8 is a subset of the original RCC. It consists of the eight base relations: disconnected (DC), externally connected (EC), partial overlap (PO), equal (EQ), tangential proper-part (TPP), non-tangential proper-part (NTTP), tangential proper-part inverse (TPP^{-1}), and non-tangential properpart inverse (NTPP^{-1}). Extensions of RCC-8 is used to successfully reason about visuo-spatial dynamics, and (eye-tracking based) visual perception of the moving image in cognitive film studies ([33]) and reason about actions in meeting environments. ([34]).

Description Logics ([6]) model relationships between entities in a particular domain. In DL, three kind of entities are considered, concepts, roles and individual names. Concepts represent classes (or sets) of individuals, roles represent binary relations between individuals and individual names represent individuals (instances of the class) in the domain. Fuzzy DLs extend the model theoretic semantics of classical DLs to fuzzy sets. In [9], Fuzzy DL is used to reason and check consistency on object-level and scene-level classification systems.

2.2 High-level Common-sense Knowledge

Several researchers employed commonsense knowledge to enrich high-level understanding tasks such as visual question answering, zero-shot object detection, relationship detection. Answering questions beyond scene information, detecting objects in data-scarce or partial observable situation are natural candidates for employing reasoning with knowledge.

Graph-Gated Neural Network (GGNN): Given a graph of N nodes, at each time-step GGNN produces some output for each node $o_1, o_2, ...o_N$ or global output o_G. The propagation model is similar to an LSTM. For each node v in the graph, there is a corresponding hidden state $h_v^{(t)}$ at every step t. At $t = 0$, they are initialized with initial state x_v, for example for a graph of object-object interactions, it is initialized as one bit activation representing whether an object is present in an image. Next, the structure of the graph is used (encoded in adjacency matrix A) along with the gated update module to update hidden states. The following equations summarize the update for each timesteps:

$$
\begin{aligned}
h_v^{(1)} &= [x_v^T, 0]^T. \\
a_v^{(t)} &= A_v^T [h_1^{(t-1)}, \ldots, h_N^{(t-1)}]^T + b. \\
z_v^t &= \sigma(W^z a_v^{(t)} + U^z h_v^{(t-1)}). \\
r_v^t &= \sigma(W^r a_v^{(t)} + U^r h_v^{(t-1)}). \\
\tilde{h_v^t} &= tanh(W a_v^{(t)} + U(r_v^t \odot h_v^{(t-1)})). \\
h_v^{(t)} &= (1 - z_v^t) \odot h_v^{(t-1)} + z_v^t \odot \tilde{h_v^t},
\end{aligned}
$$

where h_v^t is the hidden state of node v at timestep t and x_v is the initial specific annotation. After T timesteps, node-level outputs can be computed as: $o_v = g(h_v^T, x_v)$.

Image Classification

Authors in [25] employed the structured prior knowledge of similar objects and their relationships to improve end-to-end object classification task. Authors utilize the notion that humans can understand a definition of an object written in text and leverage such understanding to identify objects in an image. The authors introduce Graph Search Neural Network to utilize a knowledge graph about objects to aid in (zero-shot) object detection. This network uses image features to efficiently annotate the graph, select a relevant subset of the input graph and predict outputs on nodes representing visual concepts. GSNN learns a propagation model which reasons about different types of relationships and concepts to produce outputs on the nodes which are then used for image classification. The knowledge graph is created from Visual Genome by considering object-object and object-attribute relationships.

For GSNN, the authors propose that rather than performing recurrent updates over entire graph like GGNN, only a few initial nodes are chosen and nodes are expanded if they are useful for the final output. For example, initial nodes are chosen based on the confidence from an object detector (using a threshold). Next, the neighbors are added to the active set. After each propagation step, for every node in our current graph, authors predict an importance score using the importance network:

$$i_v^t = g_i(h_v, x_v).$$

This importance network is also learnt. Based on the score, only top P scoring non-expanded nodes are selected and added to the active set. The structure (nodes and edges) of the GSNN can be initialized according to ConceptNet or other knowledge graphs, thereby directly incorporating external knowledge. As GSNN can be trained in an end-to-end manner, this approach provides distinct advantages over sequential architectures. On Visual Genome multi-label classification, the authors achieve significant accuracy over a VGG-baseline using combined knowledge from visual genome and WordNet.

Appropriateness: Authors in [25] utilizes GSNN purely for knowledge integration purpose, i.e. to enhance an object classifier by using ontological knowledge about objects. An alternative would be use object classifier first and then use PSL or MLN to reason i.e. refine the final output - this method however cannot backpropagate the errors to the classifier.

High-Level Tasks

Figure 2: (a) Example of questions that require explicit external knowledge [37], (b) Example where knowledge helps [39].

Knowledge in Question-Answering: Authors in [37; 30] observed that popular datasets do not emphasize on questions that require access to external knowledge. The authors [37] proposed a new dataset named Fact-based VQA (or FVQA) where all questions require access to external (factual or commonsense) knowledge that is absent in the input image and the question. A popular example from their dataset is presented in Figure 2. The questions are generated using common-sense facts about visual knowledge which is extracted from ConceptNet, DBPedia, WebChild. In the proposed approach, structured predicates are predicted using LSTM from the question. For the question *Which animal in the image is able to climb trees*, the generated query example is $\{?X, ?Y\} = Query("Img1", "CapableOf", "Object")$. Then a set of object detector, scene detectors and attribute classifiers are used to extract objects, scenes and attributes from the image. This query is fired against the knowledge base of RDF triplets, and the answers are matched against the information extracted from the image.

Authors in [39] use knowledge in web-sources to answer visual questions. They propose to use fixed-length vector representations of external textual description paragraphs about

objects present in the image in an end-to-end fashion. For example, for an image about a dog, a Multi-label CNN classifier is used to extract top 5 attributes, which are then used to form a SPARQL query against DBPedia to extract the definition paragraph about relevant objects. The Doc2vec representation of this paragraph is then used to initialize the hidden state at the initial time-step of the LSTM that ultimately processes the question-words in their end-to-end question-answering architecture. The example of an image, question and relevant external knowledge is provided in the Fig. 2. On Toronto Coco-QA dataset, the authors achieve a sharp 11% increase in accuracy after using knowledge sources.

Visual Relationship Detection: Knowledge distillation has been effective in integrating external knowledge (rules, additional supervision etc.) in natural language processing applications. Authors in [42] incorporate subject-object correlations from ConceptNet using knowledge distillation. Authors use linguistic knowledge from ConceptNet to predict conditional probabilities ($P(pred|subj, obj)$) to detect visual relationships from image. [2] uses knowledge distillation to integrate additional supervision about objects such as their properties, relationships to answer questions.

Knowledge in Image Retrieval: Authors in [11] observed the semantic gap between high-level natural language query and low-level sensor data (images), and proposed to bridge the gap using semantic rules and knowledge graph such as ConceptNet. They proposed a semantic search engine, where the dependency graph of a query is enhanced using hand-written rules and ConceptNet to match scene elements. The enhanced graph is used rank and retrieve images.

3 Discussion: Knowledge Integration in the Deep Learning Era

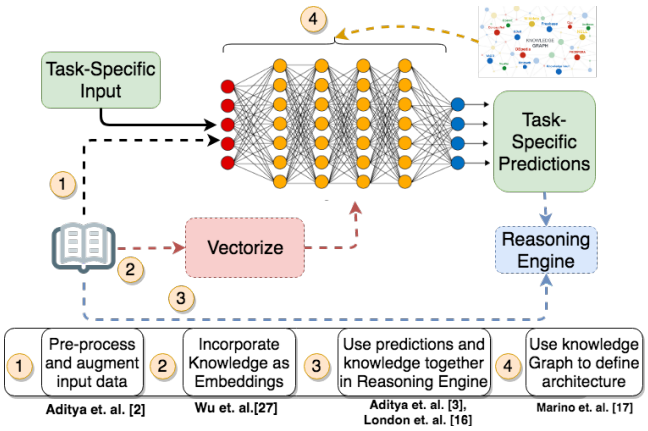

Figure 3: Ways to integrate background knowledge: i) Pre-process knowledge and augment input [2]; ii) Incorporate knowledge as embeddings [38]; iii) Post-processing using explicit reasoning mechanism [3]; iv) Using knowledge graph to influence NN architecture.

In this era where differentiable neural modules are dominating the state-of-the-art, a natural question arises as to how to integrate background or commonsense knowledge with deep neural networks. The machine learning community ([40]) have explored this in terms of constraining the search space of an optimization (or machine learning) algorithm. Four ways are adopted to use prior domain knowledge: 1) preparing training examples; 2) initiating the hypothesis or hypothesis space; 3) altering the search objective; and 4) augmenting search process. Along similar lines, we discuss four primary ways to integrate knowledge into deep neural networks: i) pre-process domain knowledge and augment training samples, ii) vectorize parts of knowledge base and input to intermediate layers, iii) inspire neural network architecture from an underlying knowledge graph, iv) post-process and reason with external knowledge. For each type, we provide a few recent works that have shown success along the line of increased accuracy or interpretability.

Knowledge Distillation ([16]) is a generic framework where there are two networks, namely the teacher and the student network. There are two traditional settings, i) teacher with additional computing layers, ii) teacher with additional knowledge. In the first setting, the teacher network is a much deeper (and/or wider) network with more layers. The teacher is trained using ground-truth supervision where in the last layer `softmax` is applied with a higher temperature (ensuring smoothness of values, while keeping the relative order). The student network, is a smaller network that aims to compress the knowledge learnt by the teacher network by emulating the teacher's predictions. In the second setting popularized in natural language processing and computer vision, the teacher network is a similar-sized network which has access to external knowledge, so that it learns both from ground-truth supervision and the external knowledge. The student network, in turn, learns from ground-truth data and teacher's soft prediction vector. The student network's loss is weighted according to an imitation parameter that signifies how much the student can trust the teacher's predictions over groundtruth.

Relational Reasoning Layer: Authors in [31] defined a relational reasoning layer that can be used as a module in an end-to-end deep neural network and trained using traditional gradient descent optimization methods. This module takes as input a set of objects, learns the relationship between each pair of objects, and infer a joint probability based on these relationships (with or without the context of a condition such as a question). Mathematically, the layer (without the appended condition vector) can be expressed as: $RN(O) = f_\phi\Big(\sum_{i,j} g_\theta(o_i, o_j)\Big)$, where O denote the list of input objects $o_1, \ldots, o_n$. In this work, the relation between a pair of objects (i.e. g_θ) and the final function over this collection of relationships i.e. f_ϕ are modeled using multilayer perceptrons and are learnt using gradient descent.

Pre-processing knowledge and using it as input to starting or intermediate layers is a popular intuitive way of integrating knowledge. For example, authors in [39] have vectorized relevant external textual documents and used the knowledge to answer knowledge-based visual questions. Similarly, other authors have used similar techniques in deep reinforcement learning. However, additional fact-based knowledge can be noisy and may bias the learning procedure. The knowledge distillation framework is effective for learning to balance be-

tween ground-truth annotations and external knowledge.

Often question-answering datasets (CLEVR [18], Sort-of-Clver, Visual Genome [20]) have additional annotations such as properties, labels and bounding box information of the objects, and the spatial relations among the objects. Authors in [2] utilized this *knowledge* in a framework that also considers the non-availability of such information during inference time. As *preprocessing*, the authors use PSL engine to reason with the structured source of knowledge about objects and their spatial relations, and the question, and construct a preprocessed attention mask for training image-question pairs. For an image-question pair, this attention mask blocks out the objects (and regions) not referred to in the question. As *reasoning mechanism*, authors use the combination of knowledge distillation and relational reasoning [31] which achieves a 13% increase in accuracy over the baseline. Relational Reasoning mechanism is also used for visual reasoning in [31], VQA in [8], temporal reasoning in [44].

Appropriateness: Authors in [2; 31] used relational reasoning for image question answering. They model relationships as functions between objects and use these functions together to answer questions about an image. Despite several efforts of creating scene graphs [19; 5], defining a closed set of complete relationships between objects is nontrivial. Hence modeling the relationships as functions is equally acceptable practice. Modeling the reasoning as a function on these triplets is similar to that of PSL, MLN. But learning objects, relationships and the function together is why the RR is a popular choice. Though, semantics associated with this function (f_ϕ) is hardly understood, which makes it a cautionary tale for out-of-the-box reasoning alternative.

Reasoning with outputs from deep neural networks and utilizing structured knowledge predicates is another natural alternative. As most off-the-shelf reasoning engines suffer because of issues of scalability, and uncertaintly modeling; authors in [4; 3] develop a PSL engine that achieves fast inference on weighted rules with open-ended structured predicates and applied it to solve image puzzles and visual question-answering. For both tasks, authors use ConceptNet and prelearnt word2vec embeddings as knowledge sources. In the image puzzle solving, the task is to find a common meaningful concept among multiple images. Authors use an off-the-shelf image classifier algorithm to predict concepts (or objects) present in each image; followed by a set of simple propositional rules in PSL such as $w_{ij} : s_i \rightarrow t_j$, where s_i is a predicted class, t_j is a target concept from ConceptNet vocabulary. The weight of the rule w_{ij} is computed by considering the (ConceptNet-based) similarity of the predicted class (s_i) and the target concept (t_j), and the popularity of the predicted class in ConceptNet. Reasoning with the rules of this form, authors predict the most probable set of targets from a larger vocabulary given class-predictions (and their scores). Using a similar rule-base, authors then jointly predict the most probable common targets for all images, which provides the final ranking of concepts. Additionally for the VQA task in [3], authors first obtain textual information from images using dense captioning methods. Then they parse the question and the captions using a rule-based semantic parser to create semantic graphs; and use these two knowledge structures in the reasoning engine to answer the question. To understand open-ended relations, ConceptNet and word2vec is used. The solution is shown to increase accuracy over state-of-the-art for "what" and "which" questions. The reasoning engine can be used to predict structured predicates as evidence along-with the answer, aiding in increased interpretability of the system.

Another important contribution is to utilize the nodes and connections of publicly available knowledge-graphs (such as ConceptNet) to build a Neural Network. As explained before, authors in [25] have used this technique for a more robust image classification. Authors have improved upon GGNN to propose Graph-search Neural Network that lazily expands the nodes when they are encountered during training. However, the approach is only shown to works on sub-graphs of a large knowledge graph, and does not have explicit consideration for handling incompleteness in the graph.

4 Summary and Future Works

In this paper we have discussed several reasoning mechanisms such as PSL, MLN, LTN, relational reasoning layers and their use in various image understanding applications. Here we give a quick summary of our assessment of these reasoning mechanisms. Early researchers in AI realized the importance of knowledge representation and reasoning and also realized that classical logics (such as first-order logic) may not be suitable for reasoning in where one may have to retract an earlier conclusion, when presented with new knowledge. This led to the development of various non-monotonic logics such as Answer Set Programming (ASP). Recent extensions of it, such as P-log, Problog [12] and LP-MLN [22] allow expression of probabilistic uncertainty, weights and contradictory information to various degrees. There are also recent works that extend Inductive Logic Programming techniques to learn ASP rules and also to learn weights. However, like MLN, which can be thought of as extension of first order logic, ASP has high computational complexity, even when the set of ground atoms are finite. PSL, uses a restricted syntax for its rules (thus is less expressive than the others), does not have non-monotonic features, requires its ground atoms to have continuous truth values and uses characterizations of logical operations so that its space of interpretations with nonzero density forms a convex polytope. This makes inference in PSL a convex optimization problem in continuous space, which in turn allows efficient inference. Many description logics are decidable fragments of first-order logic (FOL) with focus on reasoning concepts, roles and individuals, and their relationships. Relational reasoning layers (in the deep learning framework), on the other hand, lose expressiveness as it is hard to comprehend what rules are being learnt. An important need for building real-world AI applications, is to support counterfactual and causal queries. Reasoning mechanisms such as MLN and ProbLog can take cues from P-log (lite) to accommodate such reasoning.

For human beings, image understanding is a cognitive process that identifies concepts from previous encounters or knowledge. The process goes beyond data-driven pattern matching processes, and delves into the long-standing quest on integrating bottom-up signal processing with top-down

knowledge retrieval and reasoning. In this work, we discussed various types of reasoning mechanisms used by researchers in computer vision to aid a variety of image understanding tasks, ranging from segmentation to QA. To conclude, we suggest the following further research pathways to address the observed limitations: i) speeding up of inference (in MLN, ProbLog, etc. and integrating rule-learning (such as in ILP) will accelerate adoption in vision, ii) scalable reasoning on large common-sense knowledge graphs; iii) probabilistic logical mechanisms supporting counterfactual, causal and arithmetic queries, enhancing possibilities for higher-level reasoning on real-world datasets.

5 Acknowledgements

The support of the National Science Foundation under the Robust Intelligence Program (1816039 and 1750082), research gifts from Verisk AI, and support from Adobe Research (for the first author) are gratefully acknowledged.

References

[1] Machine-learning maestro michael jordan on the delusions of big data and other huge engineering efforts. https://bit.ly/2GNejBx, 2014.

[2] Somak Aditya, Rudra Saha, Yezhou Yang, and Chitta Baral. Spatial knowledge distillation to aid visual reasoning. *IEEE WACV*, 2019.

[3] Somak Aditya, Yezhou Yang, and Chitta Baral. Explicit Reasoning over End-to-End Neural Architectures for Visual Question Answering. In *AAAI*, 2018.

[4] Somak Aditya, Yezhou Yang, Chitta Baral, and Yiannis Aloimonos. Combining knowledge and reasoning through probabilistic soft logic for image puzzle solving. In *UAI 2018*, volume 1, pages 238–248. Association For Uncertainty in Artificial Intelligence (AUAI), 1 2018.

[5] Somak Aditya, Yezhou Yang, Chitta Baral, Yiannis Aloimonos, and Cornelia Fermüller. Image understanding using vision and reasoning through scene description graph. *Computer Vision and Image Understanding*, 2017.

[6] Franz Baader, Diego Calvanese, Deborah McGuinness, Peter Patel-Schneider, and Daniele Nardi. *The description logic handbook: Theory, implementation and applications*. Cambridge university press, 2003.

[7] Stephen H Bach, Matthias Broecheler, Bert Huang, and Lise Getoor. Hinge-loss markov random fields and probabilistic soft logic. *Journal of Machine Learning Research*, 18:1–67, 2017.

[8] Remi Cadene, Hedi Ben-Younes, Nicolas Thome, and Matthieu Cord. Murel: Multimodal Relational Reasoning for Visual Question Answering. In *IEEE Conference on Computer Vision and Pattern Recognition CVPR*, 2019.

[9] Stamatia Dasiopoulou, Ioannis Kompatsiaris, and Michael G Strintzis. Applying fuzzy dls in the extraction of image semantics. In *Journal on Data Semantics XIV*, pages 105–132. Springer, 2009.

[10] Ernest Davis and Gary Marcus. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, 58(9):92–103, August 2015.

[11] Maaike de Boer, Laura Daniele, Paul Brandt, and Maya Sappelli. Applying semantic reasoning in image retrieval. *Proc. ALLDATA*, 2015.

[12] Luc De Raedt, Angelika Kimmig, and Hannu Toivonen. Problog: A probabilistic prolog and its application in link discovery. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 2468–2473, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

[13] Daniel Freedman and Tao Zhang. Interactive graph cut based segmentation with shape priors. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 755–762. IEEE, 2005.

[14] Abhinav Gupta, Aniruddha Kembhavi, and Larry S Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1775–1789, 2009.

[15] Catherine Havasi, Robert Speer, and Jason Alonso. Conceptnet 3: a flexible, multilingual semantic network for common sense knowledge. In *Recent advances in natural language processing*, pages 27–29. Citeseer, 2007.

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *stat*, 1050:9, 2015.

[17] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[18] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017.

[19] Justin Johnson, Ranjay Krishna, Michael Stark, Jia Li, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[20] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vision*, 123(1):32–73, May 2017.

[21] Dieu-Thu Le, Jasper Uijlings, and Raffaella Bernardi. Exploiting language models for visual recognition. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 769–779, 2013.

[22] Joohyung Lee, Samidh Talsania, and Yi Wang. Computing lp mln using asp and mln solvers. *Theory and Practice of Logic Programming*, 17(5-6):942–960, 2017.

[23] Ben London, Sameh Khamis, Stephen Bach, Bert Huang, Lise Getoor, and Larry Davis. Collective activity detection using hinge-loss markov random fields. In *Proceedings of the IEEE CVPR Workshops*, pages 566–571, 2013.

[24] Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. Deepproblog: Neural probabilistic logic programming. In *Advances in Neural Information Processing Systems*, pages 3753–3763, 2018.

[25] Kenneth Marino, Ruslan Salakhutdinov, and Abhinav Gupta. The more you know: Using knowledge graphs for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2673–2681, 2017.

[26] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, November 1995.

[27] David A. Randell, Zhan Cui, and Anthony G. Cohn. A spatial logic based on regions and connection. In *Proceedings 3rd International Conference ON Knowledge Representation And Reasoning*, 1992.

[28] Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine learning*, 62(1-2):107–136, 2006.

[29] Tim Rocktäschel and Sebastian Riedel. End-to-end differentiable proving. In *Advances in Neural Information Processing Systems*, pages 3788–3800, 2017.

[30] Naganand Yadati Sanket Shah, Anand Mishra and Partha Pratim Talukdar. Kvqa: Knowledge-aware visual question answering. In *AAAI*, 2019.

[31] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *NIPS, pages=4967–4976, year=2017*.

[32] Luciano Serafini and Artur d'Avila Garcez. Logic tensor networks: Deep learning and logical reasoning from data and knowledge. *arXiv preprint arXiv:1606.04422*, 2016.

[33] Jakob Suchan and Mehul Bhatt. The geometry of a scene: On deep semantics for visual perception driven cognitive film, studies. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016.

[34] Jakob Suchan, Mehul Bhatt, and Harshita Jhavar. Talking about the moving image: A declarative model for image schema based embodied perception grounding and language generation. *CoRR*, abs/1508.03276, 2015.

[35] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 697–706, New York, NY, USA, 2007. ACM.

[36] Douglas Summers-Stay, Ching L Teo, Yezhou Yang, Cornelia Fermüller, and Yiannis Aloimonos. Using a minimal action grammar for activity understanding in the real world. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4104–4111. IEEE, 2012.

[37] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Fvqa: fact-based visual question answering. *IEEE TPAMI*, 2017.

[38] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163:21–40, 2017.

[39] Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *IEEE Conference on Computer Vision and Pattern Recognition CVPR*, pages 4622–4630, 2016.

[40] Markus Wulfmeier, Dushyant Rao, and Ingmar Posner. Incorporating human domain knowledge into large scale cost function learning. *arXiv preprint arXiv:1612.04318*, 2016.

[41] Yezhou Yang, Yi Li, Cornelia Fermuller, and Yiannis Aloimonos. Robot learning manipulation action plans by" watching" unconstrained videos from the world wide web. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[42] Ruichi Yu, Ang Li, Vlad I. Morariu, and Larry S. Davis. Visual Relationship Detection with Internal and External Linguistic Knowledge Distillation. *ICCV*, 2017.

[43] Bo Zheng, Yibiao Zhao, Joey Yu, Katsushi Ikeuchi, and Song-Chun Zhu. Scene understanding by reasoning stability and safety. *International Journal of Computer Vision*, 112(2):221–238, 2015.

[44] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *ECCV*, September 2018.

[45] Yuke Zhu, Alireza Fathi, and Li Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *ECCV (2)*, pages 408–424. Springer, 2014.