Knowledge and Reasoning in Image Understanding

Somak Aditya
Ph.D. Candidate, Computer Science
Dissertation Defense, June 2018

Advisors:
Prof. Chitta Baral
Dr. Yezhou Yang

Good Morning and Thank You for Joining

Thanks especially to the committee members:

Dr. Chitta Baral (Prof., CIDSE), Dr. Yezhou Yang (Asst. Prof., CIDSE),

Dr. Yiannis Aloimonos (Prof., UMD College Park), Dr. Joohyung Lee(Prof., CIDSE), Dr. Baoxin Li (Prof., CIDSE)

Motivation

Image Understanding in Recent Literature

- Image Understanding through text gained huge popularity recently:
  - Why: Innate Compositionality of text [According to ACM Survey]
  - Somewhat Easier to obtain structured information.
- Primarily two tasks were targeted:
  - Caption Generation
  - Visual Question Answering.
- Many more auxiliary tasks:
  - Dense Captioning
  - Visual Relationship Detection
  - Scene Graph Generation

Our Ideal view of "Image Understanding"

- What is understanding?
  - Evaluated in (Educational Domain/NLP) using Question-Answering.
  - Increasingly difficult questions qualitatively evaluate understanding.
- According to Bloom's Taxonomy, categories of questions:
  - Knowledge - recall
  - Comprehension - understanding
  - Application - the ability to apply the knowledge
  - Analysis - the ability to analyze and identify motives, causes
  - Synthesis - the ability to synthesize the information gathered and compile differently
  - Evaluation - the ability to make judgment about information

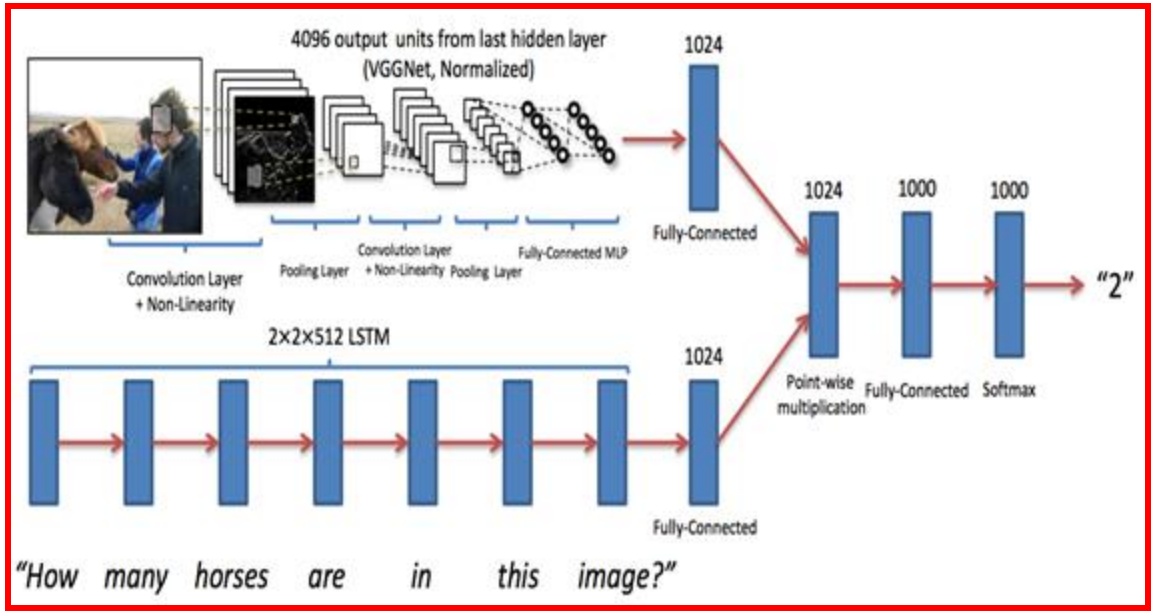
Requires Knowledge and Reasoning

- We require:
  - Vision
  - Knowledge
  - Reasoning

Tasks

Caption Generation: Generate a caption given an image

1. Huge Improvement after Deep Learning based architectures (CNN+RNN)
2. New models are proposed every month (days?)
    a. Focusing on attention



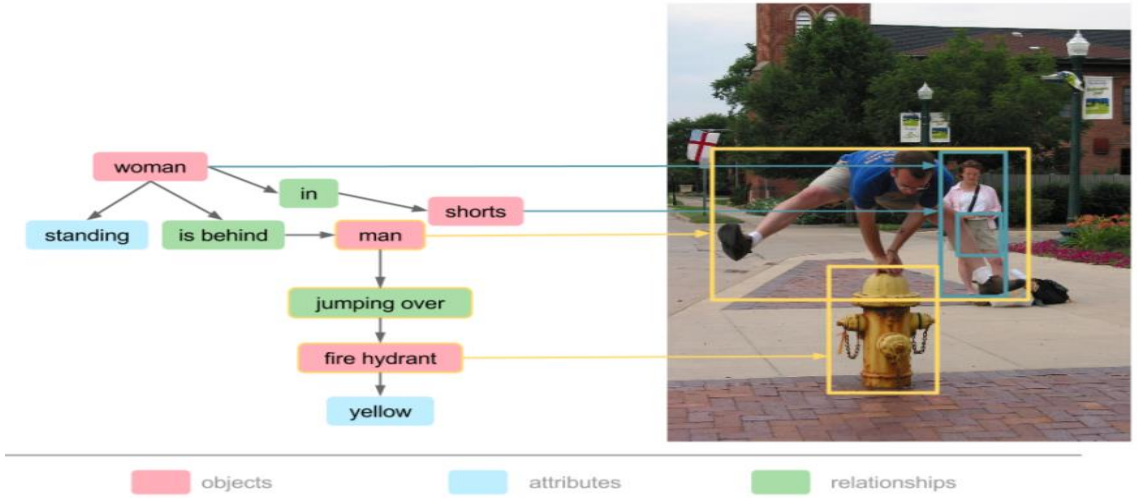
Visual Question Answering: Answer a question about an image.

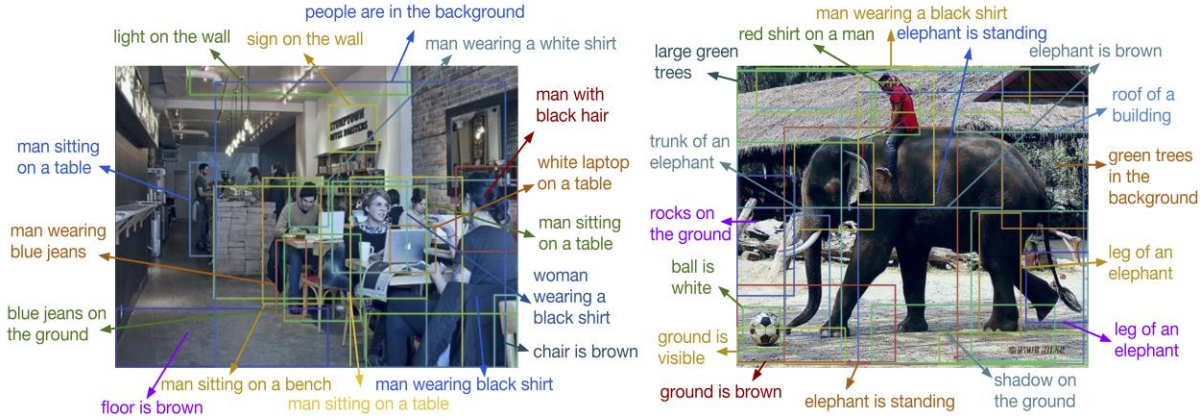
1. Textual QA was already popular.
2. Gained popularity in Vision Community after large datasets (VQA) and NN-based methods.

Auxiliary Tasks

Visual Relationship Detection:

- Detect relationships between objects in the scene





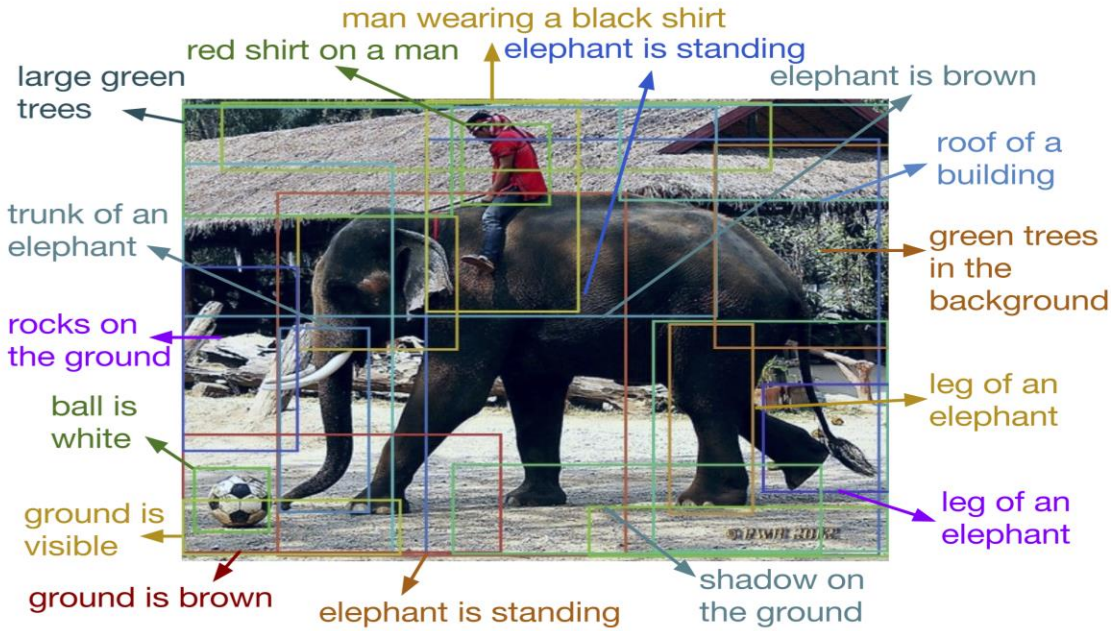
Dense Captioning:

- Generate short descriptive captions for all important (salient) regions of an image.

Scene Graph Generation:

- Generate descriptive semantic graph capturing semantic and spatial relations between objects, regions and attributes

Drawbacks (An Example)



Currently we already can:

- Detect Objects/Properties: Elephant, man, ball, hut
- Detect Properties: Color, shape, size of objects
- Detect relations (spatial)

Long-term Goals

- ❑ Is the elephant playing football?
- ❑ Is the elephant a pet? Is it a village?
- ❑ Will the man fall (without any external influence?)
- ❑ Intermediate Structure Prediction for High-level Reasoning
- ❑ Explain our inferences

Common-sense Inference

Using Background Knowledge

Knowledge Representation

Explanations

What did this thesis achieve

- ❑ Common-sense inferences (using external knowledge) ✓
- ❑ Explain our inferences for "what" and "which" ✓
- ❑ Predict Intermediate structures. ✓

K&R: A Community Necessity

" Well, humans are able to deal with cluttered scenes. They are able to deal with huge numbers of categories. They can deal with inferences about the scene: "What if I sit down on that?" "What if I put something on top of something?" These are far beyond the capability of today's machines. Deep learning is good at certain kinds of image classification. "What object is in this scene?"" – Michael Jordan, UC BERKELEY, 2014 to *IEEE Spectrum*

"With face recognition, it's been clear for a while now that it can be solved. Beyond faces, you can also talk about other categories of objects: "There's a cup in the scene." "There's a dog in the scene." But it's still a hard problem to talk about many kinds of different objects in the same scene and how they relate to each other, or how a person or a robot would interact with that scene. There are many, many hard problems that are far from solved." – Michael Jordan, UC BERKELEY , 2014 to *IEEE Spectrum*

K&R: A Community Necessity

"We also see results that show how narrow and brittle these systems are," Etzioni says. "What we would naturally mean by reading, or language understanding, or vision is really much richer or broader." – Oren Etzioni

"To achieve human-level performance in domains such as NLP, vision and robotics- basic knowledge of the commonsense world – time, space, physical interactions, people and so on, will be necessary" - Ernest Davis et. Al.

"To make real progress in A.I., we have to overcome the big challenges in the area of common sense"- Paul Allen, AI2

" Humans (and many animals) construct complex predictive models of the world that give them "common sense", ……. a major obstacle towards significant progress in AI" - Yann Lecun

"AI and machine learning is constrained by what you can learn" - Peter Norvig

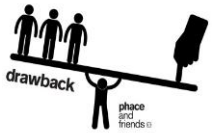
K&R: A Community Necessity

On Common-sense Reasoning in Computer Vision – Ernest Davis et. Al.

Similar issues arise in computer vision. Consider the photograph of Julia Child's kitchen

Many of the objects that are small or partially seen, such as the metal bowls in the shelf on the left, the cold water knob for the faucet, the round metal knobs on the cabinets, the dishwasher, and the chairs at the table seen from the side, are only recognizable in context; the isolated image would be difficult to identify. The top of the chair on the far side of the table is only identifiable because it matches the partial view of the chair on the near side of the table.

The viewer infers the existence of objects that are not in the image at all. There is a table under the yellow tablecloth. The scissors and other items hanging on the board in the back are presumably supported by pegs or hooks. There is presumably also a hot water knob for the faucet occluded by the dish rack. The viewer also infers how the objects can be used (sometimes called their "affordances"); for example, the cabinets and shelves can be opened by pulling on the handles. (Cabinets, which rotate on joints, have the handle on one side; shelves, which pull out straight, have the handle in the center.)



Drawbacks

We identify three fundamental drawbacks (in the Deep Learning systems and the problem setting):

- Lack of Interpretability
  - ✓ Fix errors? Readable Explanations?
- Lack of knowledge and Reasoning
  - ✓ Ontological, commonsense, background, physical ...
- Lack of Representative datasets
  - ✓ Enforce competing systems to model above.

Our Intention: Higher-level reasoning on visual data

Represent Knowledge?

```
entity(person;racket;shorts;shirt).
animate(person;dog).

inanimate(A) :- not
animate(A),  entity(A).

tennis_detector :-
has(swing,recipient,racket),
has(racket,complement_phrase,tenni
s), has(swing,agent,A),animate(A).
```

Integrate Knowledge?

Is the woman playing tennis?

Knowledge: Swinging racket and hitting ball => playing tennis

The Fundamental Questions

- *How to represent knowledge in images?*
- *What knowledge is required?*
- *Where and how to get the knowledge?*
- *Which Reasoning Mechanism to use?*

Outline of the Contribution

Other Inputs

Question for QA

Control System

Question-Answering
Captioning

Inputs

Vision

Outputs with Probabilities

Dense Captions
Object Classes
Action Classes ...

Representation Module

Intermediate Knowledge Representation

?

Knowledge Base

Reasoning Module

?

?

?

Outputs

Answer to question for QA.
Caption for Captioning.

Outline of the Contribution

Other Inputs

Question for QA

Control System

Question-Answering
Captioning

Inputs

Vision

Outputs with Probabilities

Dense Captions
Object Classes
Action Classes ...

Representation Module

Intermediate Knowledge Representation

Knowledge Base

ConceptNet
Application-specific

Reasoning Module

ASP.
PSL.
Graph-based Reasoning.

SDG
Graph, Scen

Outputs

Answer to question for QA.
Caption for Captioning.

Contributions
[KR] SDG, Scene Graph
[KB] Automatic Knowledge Acquisition
[Reasoning] An Efficient Prob. Reasoning Engine

Current Work

Outline…

1. **Representation and Reasoning in Images**
   - **[Image Captioning]** Image Understanding Through Scene Description Graphs ***IJCAI '15, CVIU '17***
2. **PSL Applications**
   - **[VQA]** Visual Question Answering using a reasoning wrapper. ***AAAI '18 (24.5% acceptance)***
   - **[Puzzles/New Challenge]** Image Riddles using Vision Reasoning **UAI '18 (30% acceptance), *CVPR '17 Workshop***
3. **An End-to-End Endeavor:**
   - **[Visual Reasoning]** VQA using a Spatial Knowledge Distillation. ***Finally fully end-to-end !!!***
4. **Overview of the rest and Conclusion**

Outline…

1. **Representation and Reasoning in Images**
   - **[Image Captioning]** Image Understanding Through Scene Description Graphs ***IJCAI '15, CVIU '17***
2. **PSL Applications**
   - [VQA] Visual Question Answering using a reasoning wrapper. ***AAAI '18 (24.5% acceptance)***
   - **[Puzzles/New Challenge]** Image Riddles using Vision Reasoning UAI '18 (30% acceptance), ***CVPR '17 Workshop***
3. **An End-to-End Endeavor:**
   - **[Visual Reasoning]** VQA using a Spatial Knowledge Distillation. ***Finally fully end-to-end !!!***
4. **Overview of the rest and Conclusion**

Image Understanding though Scene Description Graph

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

Image Understanding using vision and reasoning through Scene Description Graph.", **Somak Aditya**, Yezhou Yang, Chitta Baral, Yiannis Aloimonos, and Cornelia Fermüller. ***Computer Vision and Image Understanding* (2017).**

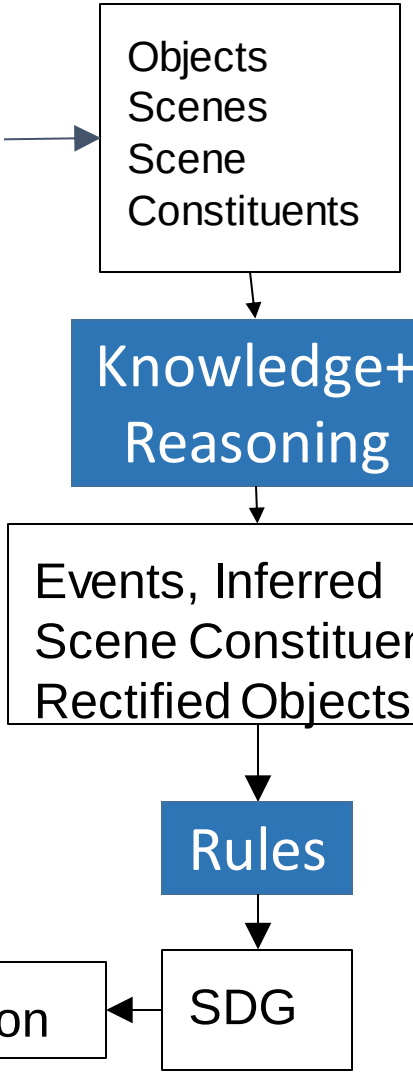
Image Captioning

Current captioning system gives great results.

Problems:

- ❑ In Perception Module: Erroneous Recognitions.
  - Partially/fully occluded objects.
  - Unknown or OOV objects.
  - Recognition models trained on natural photographic images.
- ❑ Others: Statistical Biases to frequent words, phrases etc.

We might not always get a correct caption.
➔ we need (some) explanations.



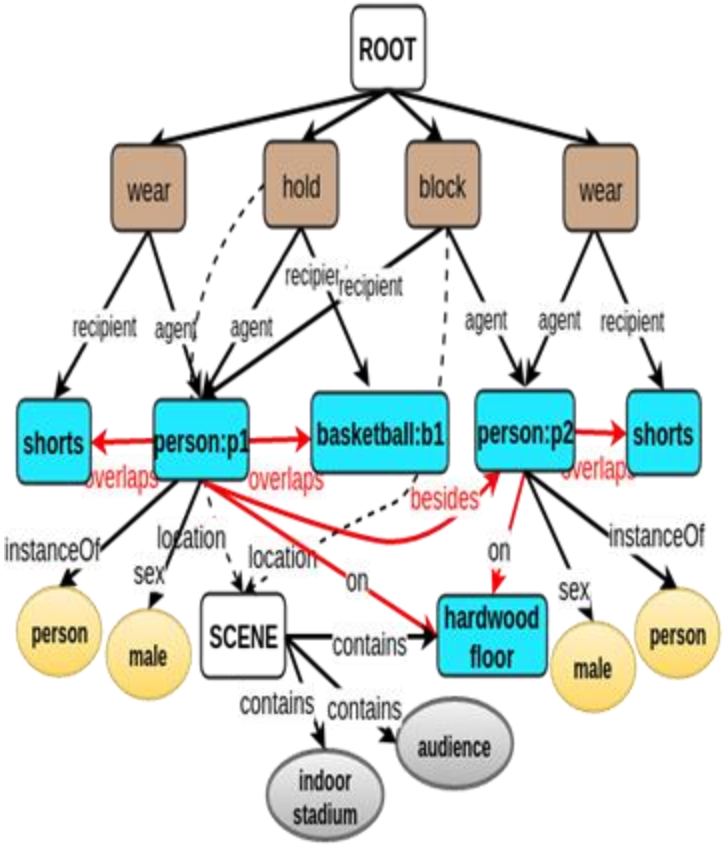
Solution: Predict an Intermediate Structure.

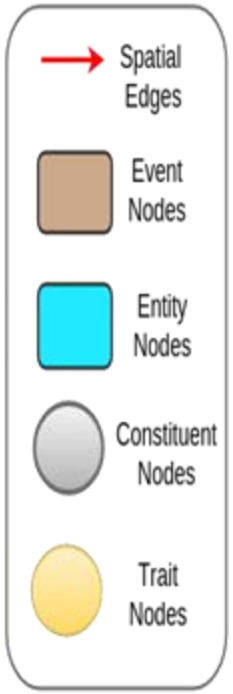
- Each component stems from different modules.
- Which component is wrong → the culprit module
- Use NLG modules *(Java SimpleNLG API)* to generate captions

SDG: Scene Description Graph

Directed Labeled Graph:

- Nodes: are verbs (actions), nouns (objects, regions, attributes). ***(Person, shorts, wear)***
- Verbs connected to concrete nouns with semantic roles. ***(Person-agent-wear)***
- Nouns (objects/regions) are connected spatially. ***(Person-overlaps-shorts)***
- Inferred Aspects connected to dummy node SCENE





i. Supports Event-based, Spatial Reasoning, Factoid QA!
ii. Same Representation as Semantic Parser.

SDG (Vision+Knowledge+Reasoning)

- CNN-ILSVRC trained Object Recognition [Girshick et.al.]
- Scene Classification [Zhou et. Al.]
- Scene Constituent: Pre-trained CNN (VGG-16) +SVM multi-class

Vision

Scene Classes: **Erroneous!!**
- Dete**ct Frequent Phrases**

Type of Knowledge Used:
- A Knowledge Graph: Combined semantic parses of captions.
- Bayesian Network: Dependencies among objects and scene constituents.

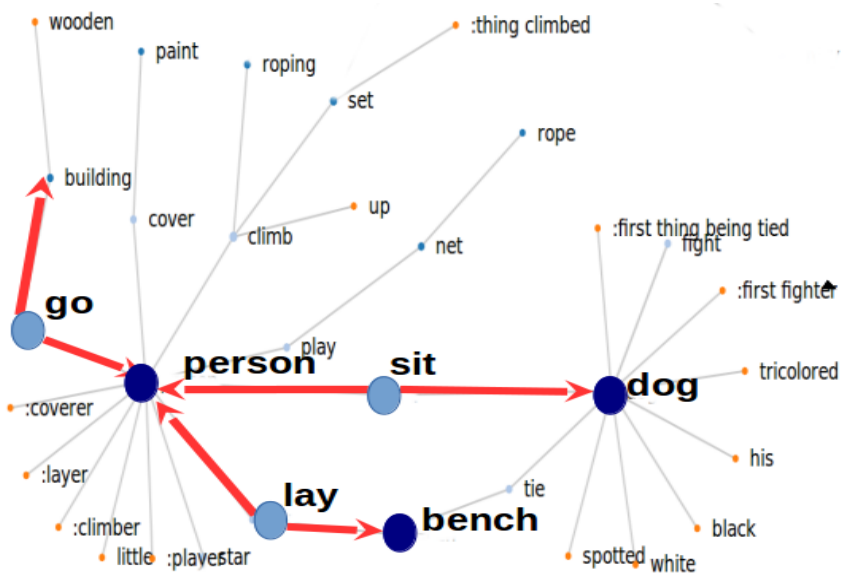
How did we store the knowledge:
- Graph on File-system with self-built Query-Engine

Knowledge

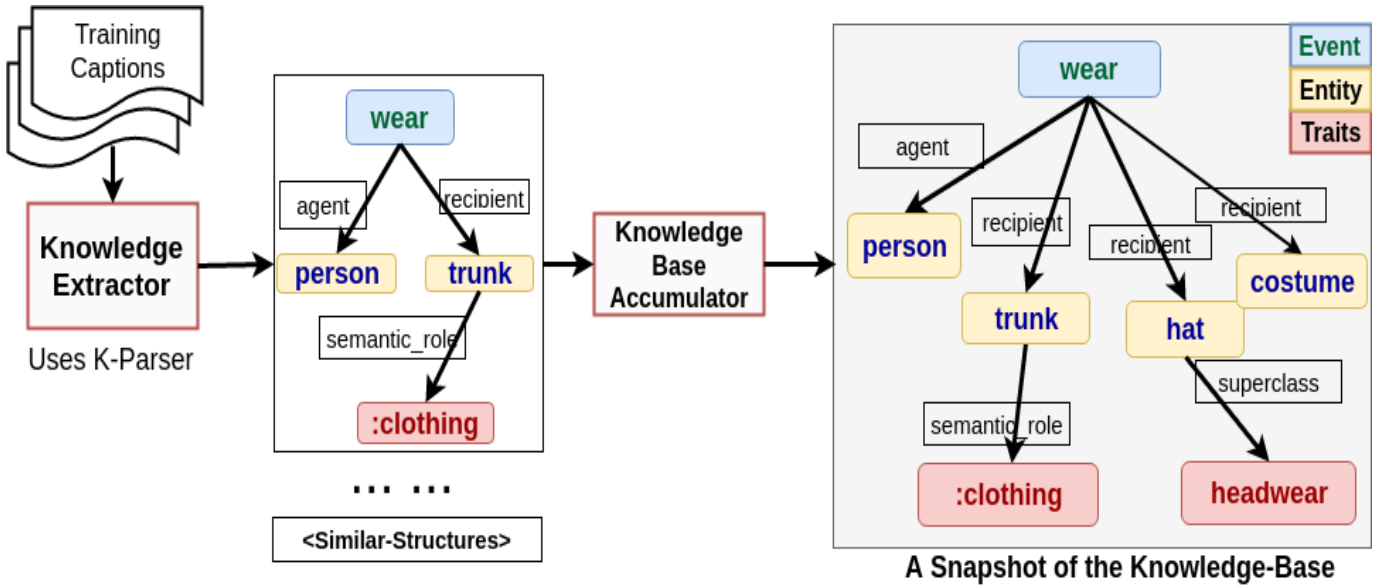
Probabilistic Reasoning using Bayesian Network, and IF-THEN reasoning using Constructed Knowledge-Base

Reasoning

SDG (Knowledge Acquisition)



Knowledge Base (from training Captions)

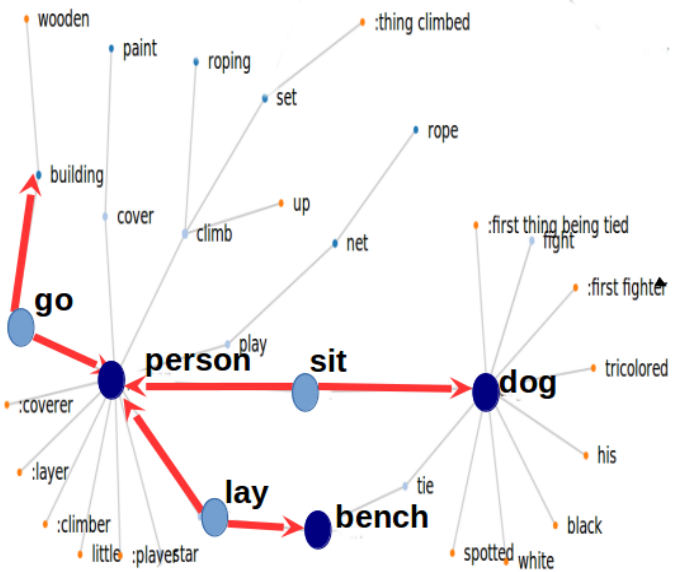


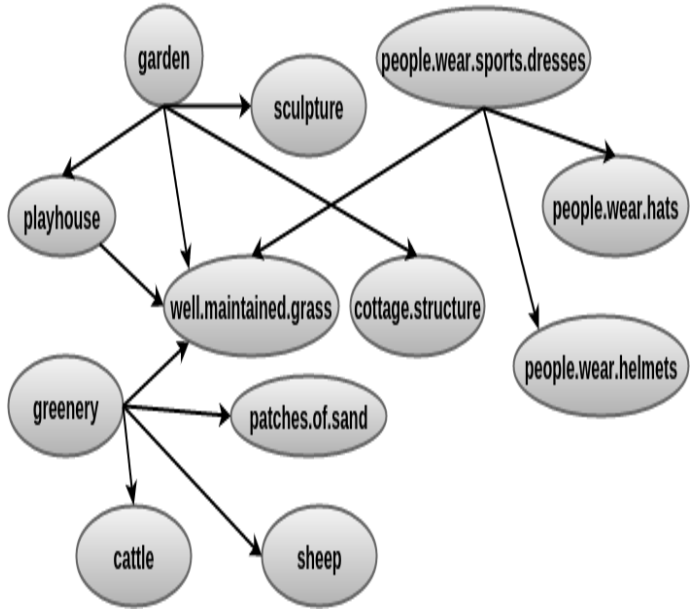
A Snapshot of the Knowledge-Base

Automatic Knowledge Acquisition:

- ❑ Parse each training caption using K-Parser.
- ❑ Merge parsed-graphs using overlapping entities and events
- ❑ Store all the parses

SDG (Knowledge and Reasoning)





Knowledge Base (from training Captions)

Bayes Net (objects and Inferred SCs)

Knowledge

Input:

- ❑ Objects.
- ❑ Candidate ISCs || Scene -> ISC mapping ||

Reasoning:

- ❑ Iteratively select most probable ISCs || Use Bayes Net ||
- ❑ Rectify noisy (low-scoring) objects. || Use Bayes Net and WordNet ||
  - choose the most probable sibling from WordNet Hierarchy.
  - One level: superclass(bathing cap) = cap, **children(cap) = ski cap, basketball cap** etc…..
- ❑ Search Events and then Scenes ||Use Knowledge Base ||

An Example

Objects, Scenes, Scene Constituents

Person
Bird
Rock
…

Rope-bridge
Cemetery
…

Trees in background

Knowledge

Parsing: Entities, Events
Scene-to-ISC-map: frequent ISCs

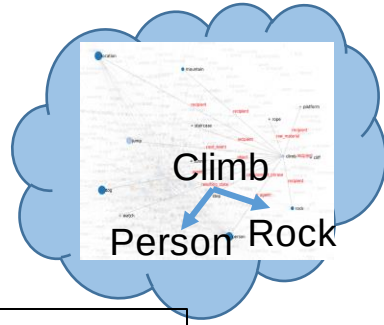
Person, Rock, Trees, **Bird**,
Erected Stone, Rope, **Bridge**

Reasoning

Bayes Net: Most Probable ISCs Rectified Objects

Person, Rock, Trees, **Bird**
Erected Stone, Rope, ~~**Bridge**~~

Knowledge + Reasoning



Climb:: has(climb, location, Rock), has(climb, agent, person)

From KB

confidenceOnScene=0.182,
Edges::[
has(person,semantic_role,:climber),
has(mountain,trait,flat),
has(mountain,trait,tall),
has(mountain,semantic_role,:thing climbed),
has(climb,agent,person),
has(climb,recipient,mountain),
has(climb,next_event,hold),
has(rope,complement_phrase,safety)]

Person is climbing rock in the
scene and holdin
g a safety rope.

Experiments (Test-bed)

Datasets:

- ❑Flickr 8k. Total: 8k images, 5 captions per image, 1k test
- ❑Flickr 30k Total: 31k images, 5 captions per image, 1k test
- ❑MS-COCO Total: >160k images, 5 captions per image, 2k test

Baselines:

- ❑Deep Visual Semantic Alignments [Karpathy and Fei-Fei Li, 2014]
- ❑Show, Attend and Tell [Vinyals et. Al. 2017]

Generated Sentence Evaluations (AMT)

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

Table 1: Sentence generation relevance (R) and thoroughness (T) human evaluation results with gold standard and BRNN-Karpathy on Flickr 8k, 30k and MS-COCO datasets. D: Standard Deviation.

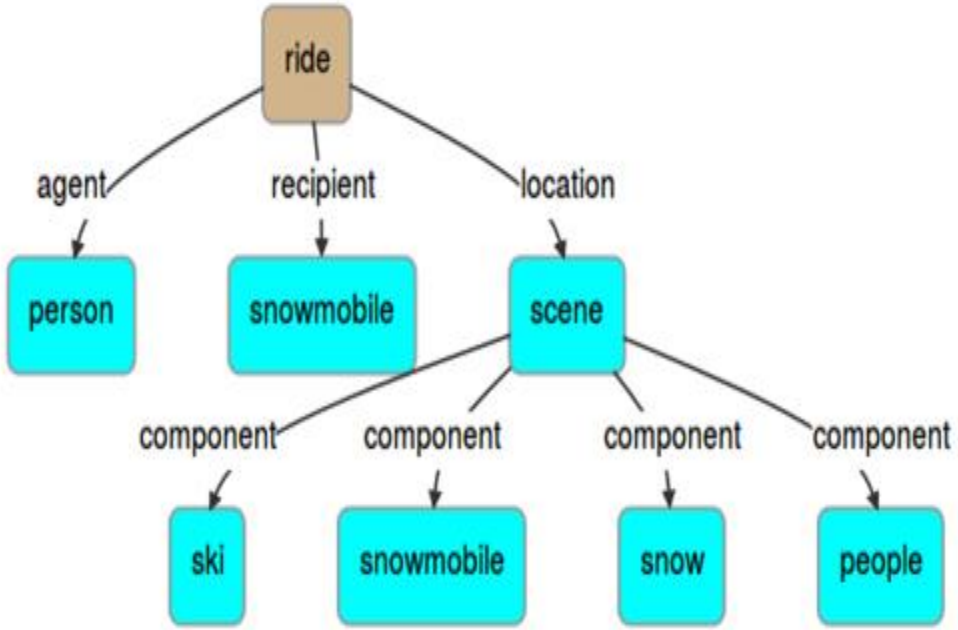
Evaluations and SDG Inspired the SPICE metric, ECCV 2016!!
Find more Evaluations in CVIU Version.

Image Search Experiment

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

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

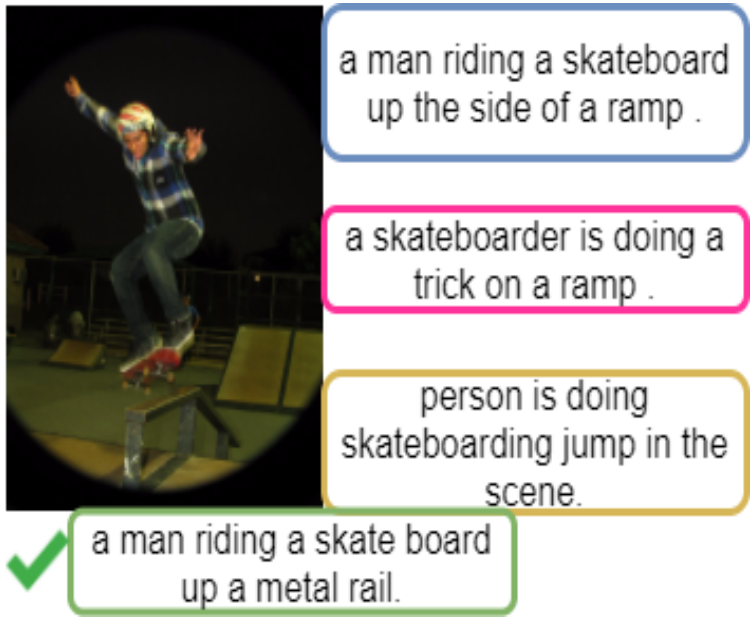
Results: SDG



Person is riding snowmobile in the scene.
The scene contains people and ski and snow and snowmobile.

* More Examples in imagesdg.wordpress.com/image-to-scene-description-graph/.

Results: Captions



We provide some comparative captions generated by **our system (in yellow box),** by BRNN [24] (top blue box), by ShowAndTell [38] (in pink box). The Ground-truth captions are given in lower green boxes.

Third Figure: BLEU 1-4 metric the sentence from the Neural captioning engine is rated 90.0, 83.7, 80.7, 78.3, while the caption from our system is rated 20.0, 0.0.0.0, 0.0.

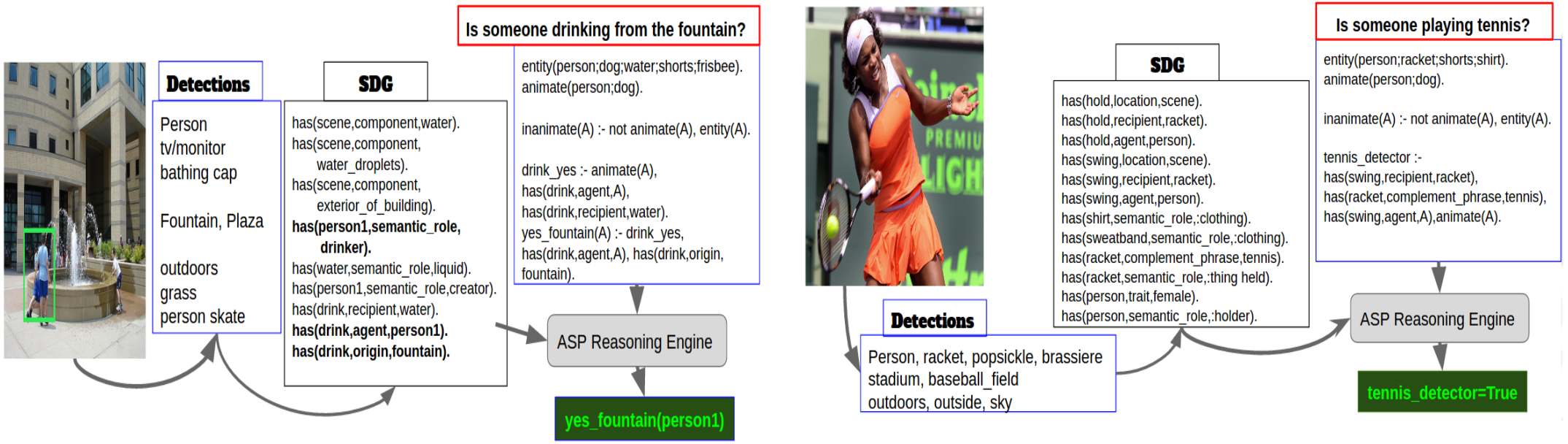
* More Examples in imagesdg.wordpress.com/image-to-scene-description-graph/.

Explain Why?

- Why person, snowmobile? → Detected From CNN (object detectors)
- Why ride? → Best "Logical" Choice from Knowledge-Base.
  - Logical: Both Statistics (frequently) and Semantics (<ride, agent, person> , <ride, recipient, snowmobile>)
- Why ski, snow, snowmobile → Frequently occurs together in Graphs of the Knowledge-Base with "person" and "snow"

- Perception
- Knowledge-Base
- Reasoning Module

Results: QA Case Study



Some hard-coded examples of QA with SDG. These motivated us to pursue further!

Outline…

1. **Representation and Reasoning in Images**
   - **[Image Captioning]** Image Understanding Through Scene Description Graphs *IJCAI '15, CVIU '17*
2. **PSL Applications**
   - **[VQA]** Visual Question Answering using a reasoning wrapper. *AAAI '18 (24.5% acceptance)*
   - **[Puzzles/New Challenge]** Image Riddles using Vision Reasoning **UAI '18 (30% acceptance), *CVPR '17 Workshop***
3. **An End-to-End Endeavor:**
   - **[Visual Reasoning]** VQA using a Spatial Knowledge Distillation. ***Finally fully end-to-end !!! :)***
4. **Overview of the rest and Conclusion**

Probabilistic Soft Logic: A Primer

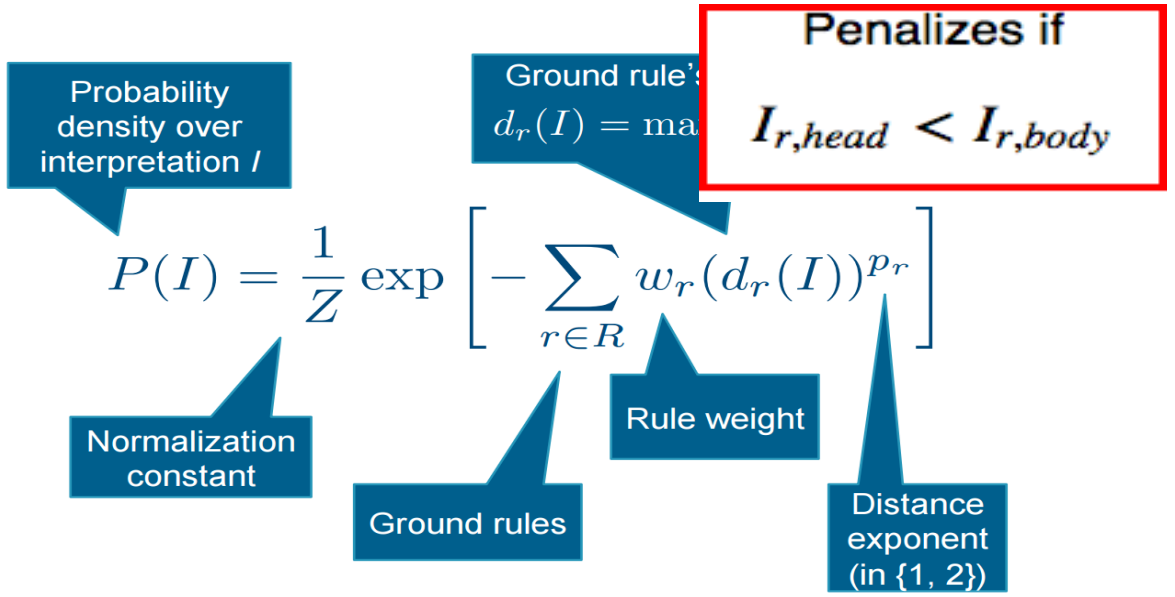
$$H_1 \bigvee H_2 \ldots \bigvee H_m \leftarrow B_1 \bigwedge B_2 \ldots \bigwedge B_n$$

Combination Functions (Lukasiewicz T-Norm):

- $A \bigvee B = min(1, A + B)$
- $A \bigwedge B = max(0, A + B - 1)$

- Uses a subset of (weighted) First-order Logic rules.
- Each ground predicate is an r.v.
- Distribution modeled using an MRF.
- Grounded rule ---> used as a template to define a clique potential.

- Probability of an interpretation (grounding of facts and rules) defined by:



- Rules (clauses), functions are chosen so that the MAP optimization becomes a convex one.

Features of In-house PSL

- Previously available engine had disadvantages.
  - **A toy PSL-Groovy implementation of Image Riddles infers all targets as 1.0.**
  - In most cases, it gave out error.
- Coded PSL from scratch using python-gurobi optimization.
- Several Advantages:
  - It's Python!
  - Using Gurobi Optimization speedups
  - Manipulate underlying Optimization Formulation.
  - Phrasal Similarities and External Functions
    - ✓ **Has(man, wears, shirt) ≈ Has(man, is wearing, shirt)**
  - Integration of word2vec, ConceptNet
- **Download from** https://github.com/adityaSomak/PSLQA

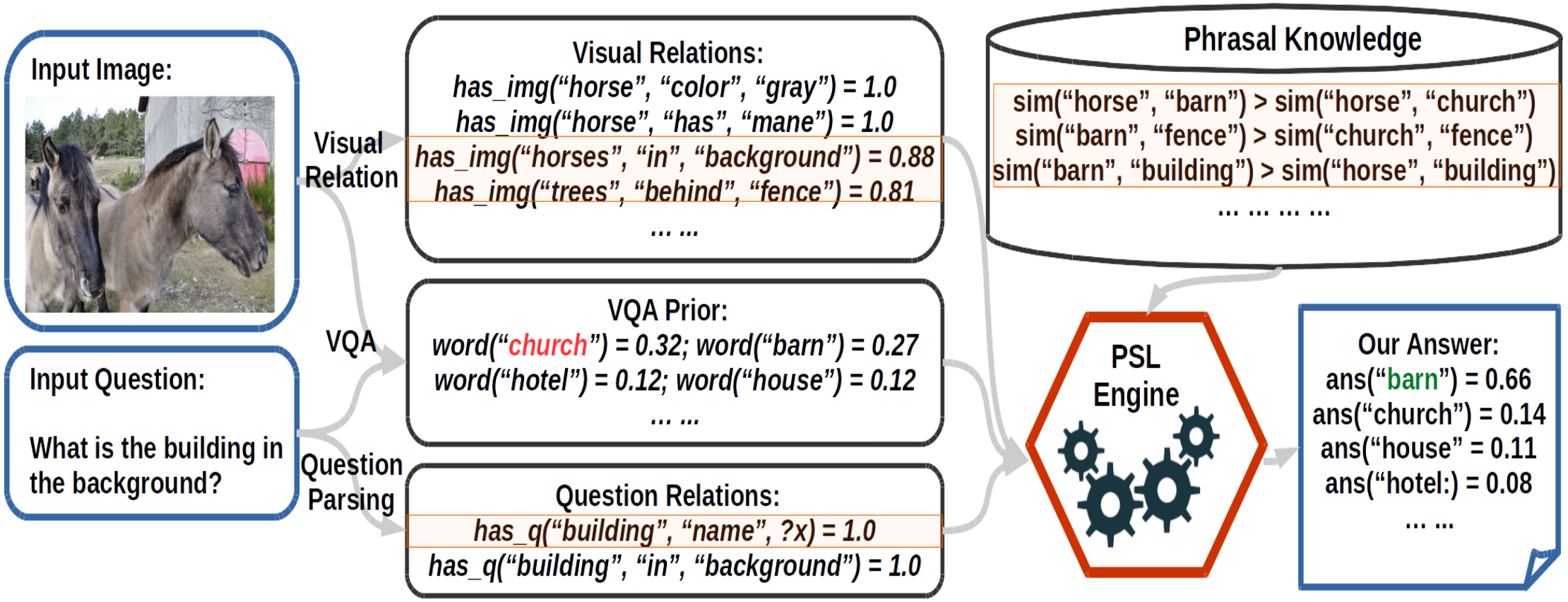
Visual Question Answering

Explicit Reasoning over End-to-end Neural Architectures for Answering Visual Questions, **S Aditya**, Y Yang, C Baral **AAAI 2018** (acceptance rate **24.5%**)

Visual Question Answering

- Answer a natural-language question about an image.
  - Requires Image Understanding, NLU and commonsense reasoning
- We focus on:
  - Modeling knowledge and reasoning.
  - Interpretability.
- Challenges:
  - What knowledge is required?
  - Where and how to acquire them?
  - Which reasoning mechanism to use?

Architecture



We also provide evidence-predicates, as highlighted in yellow.

Image and Question Parsing

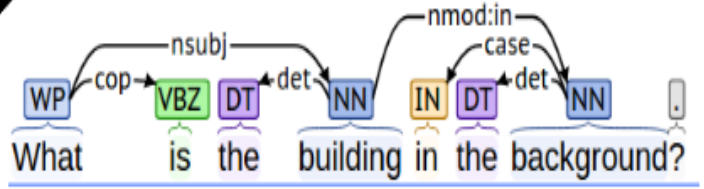
two horses standing in a field. brown horse with white mane. a brown horse. a cross on top of a building. two horses standing in a field. white clouds in blue sky. a brown horse. two horses on the road. white clouds in blue sky. white clouds in blue sky. a steeple on top of a building. white clouds in blue sky. a building with a roof.

We color-code the predicates to show a direct correspondence between the captions and the parsed predicates. The greyed out predicates come from low-confidence captions.

has_story(horses-2,in standing,field-6) = 1.0
has_story(horse-2,color,brown-1) = 1.0
has_story(horse-2,with,mane-5) = 1.0
has_story(mane-5,color,white-4) = 1.0
has_story(horse-3,color,brown-2) = 1.0
has_story(cross-2,on,top-4) = 1.0
has_story(cross-2,on top of,building-7) = 1.0
has_story(horses-2,in standing,field-6) = 1.0
has_story(clouds-2,color,white-1) = 1.0
has_story(clouds-2,in,sky-5) = 1.0
has_story(sky-5,color,blue-4) = 1.0
has_story(horse-3,color,brown-2) = 1.0
has_story(horses-2,on,road-5) = 0.8460
has_story(clouds-2,color,white-1) = 1.0
has_story(clouds-2,in,sky-5) = 1.0
has_story(sky-5,color,blue-4) = 1.0
has_story(clouds-2,color,white-1) = 1.0
has_story(clouds-2,in,sky-5) = 1.0
has_story(sky-5,color,blue-4) = 1.0
has_story(steeple-2,on,top-4) = 1.0
has_story(steeple-2,on top of,building-7) = 1.0
has_story(clouds-2,color,white-1) = 1.0
has_story(clouds-2,in,sky-5) = 1.0
has_story(sky-5,color,blue-4) = 1.0
has_story(building-2,with,roof-5) = 1.0
has_story(clouds-2,color,white-1) = 1.0
has_story(clouds-2,in,sky-5) = 1.0
has_story(sky-5,color,blue-4) = 1.0

What is the building in the background?

has_q(building,is,?x) = 1.0
has_q(building,is,what) = 1.0
has_q(building,in,background) = 0.87
has_q(background,building,what) = 1.0



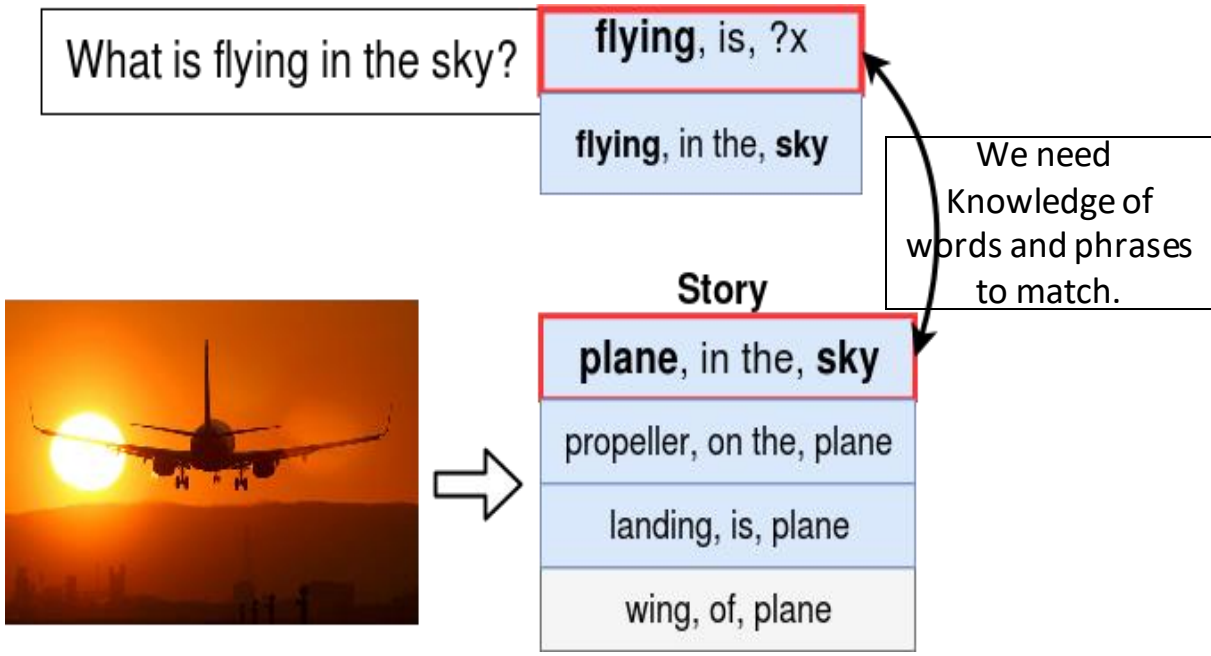
Use Dense-Captioning to generate captions for all important regions

Parse the captions/question: Extract noun-pairs and use phrasal similarity to choose an open-ome Relations

Pushing complexity to reasoning: Understand open-ended phrases using knowledge, reason on predicates and find answer.

Glimpse into the PSL Engine

- Inspired by Structured QA:
  - *Question maps to a semantic-graph with edges rooted with "?x", the missing label.*
  - *Candidate answer has a semantic-graph.*
  - *Match the graph and find "?x".*
  - *Best match gives the answer.*



Glimpse into the PSL Engine

- Get prior confidence of answer Z with respect to the image:
  - $has\_story\_ans(Z,X,R,Y1) \rightarrow word(Z) \wedge has\_story(X,R1,Y1) \wedge Z \approx X \wedge Z \approx Y1.$
- Enumerate Answer Candidates:
  - $ans\_candidate(Z) \rightarrow word(Z) \wedge has\_q(Y,R,X) \wedge has\_story\_ans(Z,\mathbf{X1,R1,Y1}) \wedge R \approx R1 \wedge Y \approx Y1 \wedge X \approx X1.$
- Enumerate Answers:
  - $ans(Z) \rightarrow has\_q(X,R,?x) \wedge has\_story(Z,R,X) \wedge ans\_candidate(Z)$
  - $ans(Z) \rightarrow has\_q(X,R,?x) \wedge has\_story(Z1,R,X) \wedge ans\_candidate(Z) \wedge Z \approx Z1.$
  - $ans(Z) :- has\_q(X,R,?x) \wedge has\_story(Z1,R1,X) \wedge ans\_candidate(Z) \wedge Z \approx Z1 \wedge R \approx R1.$

- $\tilde{P}(Z| < X,R,Y >)$: factor in the image triplet
- $\tilde{P}(Z|I,Q_{other})$: factors in the question triplets other than the root triplets.
- $\tilde{P}(Z|I,Q)$: factors in the root triplet.

Weights of rules are learnt using Maximum Likelihood

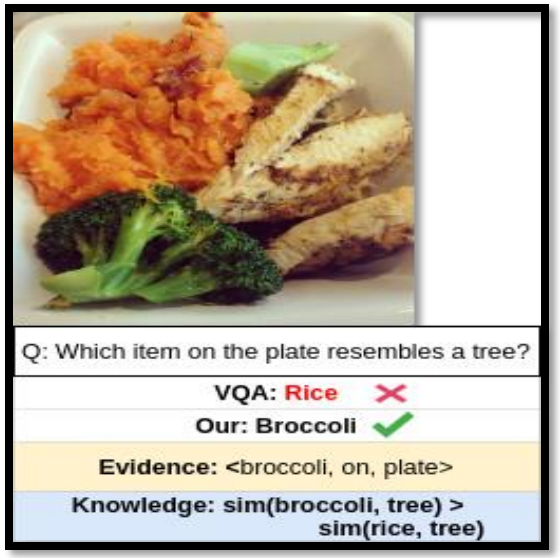
Results

Q: What is the name of the utensil ?

VQA: Spoon ✗

Our: Fork ✓

Evidence: <fork, on, plate>



Q: Which item on the plate resembles a tree?

VQA: Rice ✗

Our: Broccoli ✓

Evidence: <broccoli, on, plate>

Knowledge: sim(broccoli, tree) > sim(rice, tree)

Q: What are the white areas on the water called?

VQA: Rocks ✗

Our: Waves ✓

Evidence: <wave, in, ocean>
<waves, ocean water, white>

Q: What is the lady holding?

VQA: Teddy Bear ✗

Our: Umbrella ✓

Evidence: <woman, holding large, umbrella>

Q: What is the weather like?

VQA: Cloudy ✓

Our: Clouds ✗

Evidence: <clouds, in, sky>, <sky, is,cloudy>

Q: What room is this?

VQA: Toilet ✓

Our: Bathroom ✗

Evidence:<urinal,color,white>, <toilet,color,white>

Knowledge: sim(bathroom, room) > sim(toilet, room)

Q: What is the the bear sitting on?

VQA: log ✓

Our: Rock ✗

Evidence: <bear,sitting on,tree>, <rock, on,ground>, <log,on,ground>, <branch,on,ground>

VQA Prior: tree (0.02), rock (0.33), log (0.53)

Q: What is the device for?

VQA: computer ✗

Our: Decoration ✗

GT: lighting, concert, lights

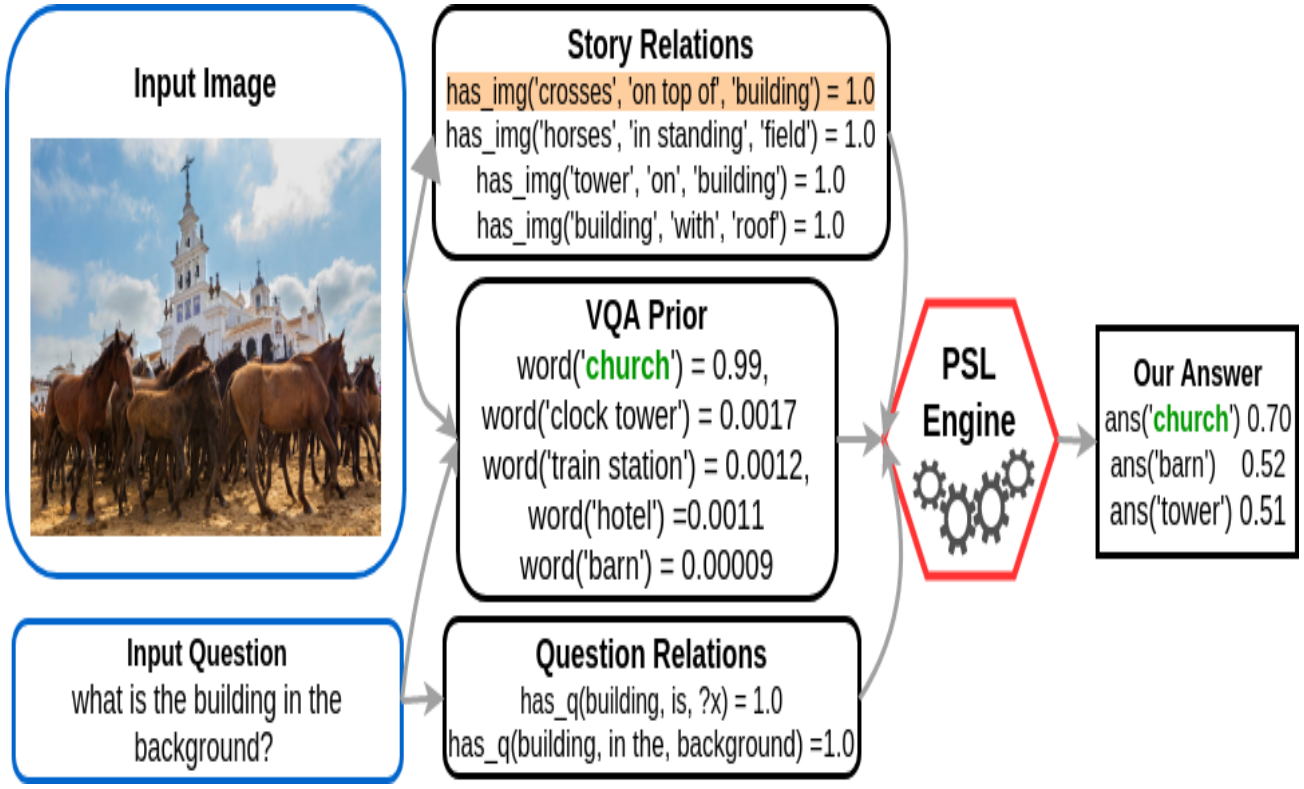
Evidence: <keyboard,color,white>, <table,is,wooden>, <wall,behind,table>

VQA Prior: Computer (0.17), Decoration (0.169)

Does additional commonsense add bias?

We pose the same question against an image where:

1. There is a **Church** in the background of horses.
2. Commonsense still dictates "**barn**" is more related.



Q: What is the building in the background?
Commonsense Knowledge: Barn is more related than church.
Our Answer: **Church**

Result

- Good results for "what" and "which" questions.
- Improvement in raw accuracy is low.
- The structured evidential predicate shows clear avenues, where the improvements can be.

Contributions:

1. A novel generic reasoning component that successfully infers answers from various (noisy) knowledge sources.
2. We provide structured evidence along-with answers.
3. We show additional knowledge helps in the VQA task.

	Categories	CoAttn	PslDVQ	PslDVQ-+CN
Specific	what animal is (516)	65	**66.22**	**66.36**
	what brand (526)	38.14	37.51	37.55
	what is the man (1493)	54.82	**55.01**	54.66
	what is the name (433)	8.57	8.2	7.74
	what is the person (500)	54.84	**54.98**	54.2
	what is the woman (497)	45.84	**46.52**	45.41
	what number is (375)	4.05	**4.51**	**4.67**
	what room is (472)	88.07	87.86	**88.28**
	what sport is (665)	89.1	**89.1**	89.04
	what time (1006)	22.55	22.24	22.54
Summary	**Other**	57.49	**57.59**	57.37
	Number	2.51	**2.58**	**2.7**
	Total	48.49	**48.58**	48.42
Color Related	what color (791)	48.14	47.51	47.07
	what color are the (1806)	56.2	55.07	54.38
	what color is (711)	61.01	58.33	57.37
	what color is the (8193)	62.44	61.39	60.37
	what is the color of the (467)	70.92	67.39	64.03
General	what (9123)	39.49	39.12	38.97
	what are (857)	51.65	**52.71**	**52.71**
	what are the (1859)	40.92	40.52	40.49
	what does the (1133)	21.87	21.51	21.49
	what is (3605)	32.88	**33.08**	32.65
	what is in the (981)	41.54	40.8	40.49
	what is on the (1213)	36.94	35.72	35.8
	what is the (6455)	41.68	41.22	41.4
	what is this (928)	57.18	56.4	56.25
	what kind of (3301)	49.85	49.81	49.84
	what type of (2259)	48.68	48.53	**48.77**
	where are the (788)	31	29.94	29.06
	where is the (2263)	28.4	28.09	27.69
	which (1421)	40.91	**41.2**	40.73
	who is (640)	27.16	24.11	21.91
	why (930)	16.78	16.54	16.08
	why is the (347)	16.65	16.53	**16.74**

Comparative results on VQA dev split for question categories.

Image Riddles

Combining Knowledge and Reasoning through Probabilistic Soft Logic for Image Puzzle Solving; ***S Aditya***, *Y Yang, C Baral, Y Aloimonos* - arXiv preprint arXiv:1611.05896, 2016, ***Accepted in UAI 2018***

Extended Abstract published in VQA Challenge workshop, CVPR 2017 and Vision Meets Cognition workshop, CVPR 2017

Given images, find the Connecting Word

Can you guess the word?

i) First image depicts the season Fall.

ii) Second one has water-fall.

iii) Third one has rainfall.

iv) Fourth one depicts a statue is “fall”ing.

- Have to detect action (verb), concept, noun, region!!!
- Have to logically connect!!!
- If detections are imperfect, can you still guess?

Image Riddles ≈Image IQ test!

What Prompted Us:

1. VQA: Task requires explicit model of commonsense reasoning. However a big percentage of the dataset concentrates on “what” and “where”, “how many” questions.
   - Riddles require ontological Reasoning
   - We need answers, but explanations too. Connections are important!
2. Additional Constraints make the problem difficult for end-to-end classification
   - Image Riddles: target answers in train and test are mostly unique. (Zero-shot)
   - One-to-Many: Same image, different riddle → different answer.
3. Lastly, puzzles are fun.

Similar to Familiar IQ tests
Word Analogy Tasks Sequence Filling Tasks

Image Riddles (Vision + Knowledge+ Reasoning)

Visual Detections Used:
- ❑ Residual Network (ResNet-200)
- ❑ Clarifai API (commerical, finetuned)

Vision

Type of Knowledge Used:
- ❑ ConceptNet
- ❑ Word2vec

How did we store the knowledge:
- Publicly Available

Knowledge

Reasoning Module Used: Probabilistic Soft Logic (Uses FOL syntax to define Markov Random Field Potentials)

Reasoning

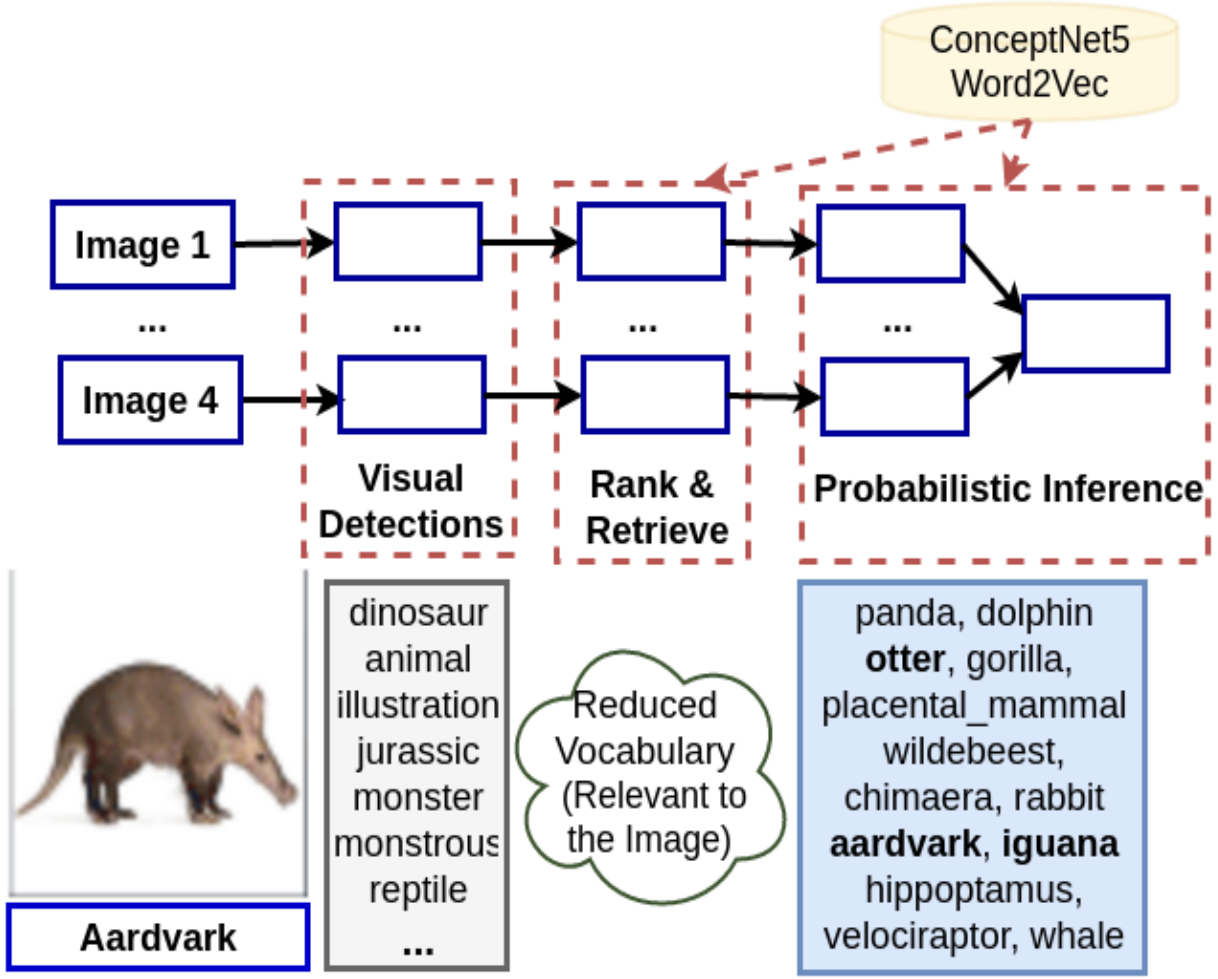
Connecting Vision, Knowledge and Probabilistic Reasoning

For Each Image:

1. Get possible object classes from an Image Classifier.
2. Use a Hinge-Loss MRF to predict all related concepts (words and phrases) from ConceptNet (~0.2 mil English words).

Jointly Predict:

1. Common words from all predicted concepts using similar HL-MRF model.



Mapping from 1000 classes to 0.2 mil English Words using HL-MRF model!

PSL Reasoning (Phase I)

For each image k, for each seed s_{ik} and target t_{jk}, add:

$$wt_{ij} : s_{ik} \rightarrow t_{jk}$$

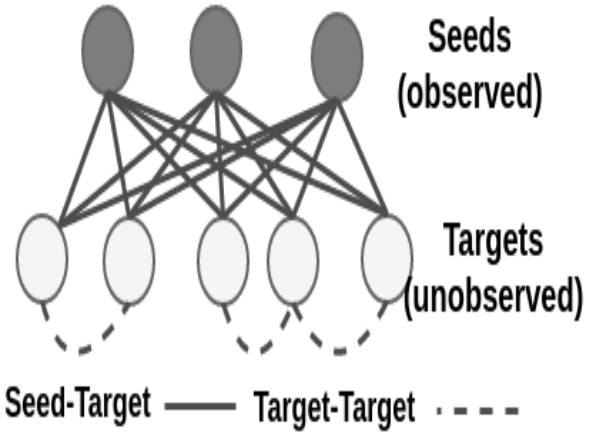
For each target t_{jk}, get the most similar word and add:

$$wt_{jm} : t_{jk} \rightarrow t_{mk}$$

$$wt_{jm} : t_{mk} \rightarrow t_{jk}$$

The weights are set using word2vec and ConceptNet similarities, and centrality score (popularity) of the target:

$$wt_{ij} = \alpha_1 cn(s_{ik}, t_{jk}) + \alpha_2 w2v(s_{ik}, t_{jk}) + 1/C(t_{jk})$$



MAP Inference equivalent to:

$$\underset{I(T_k)}{\operatorname{argmin}} \sum_{\substack{s_{ik} \in S_k \\ t_{jk} \in T_k}} wt_{ij} \max\{I(s_{ik}) - I(t_{jk}), 0\} +$$

$$\sum_{\substack{t_{jk} \in T_k \\ t_{mk} \in T_j}} wt_{jm} * \{\max\{I(t_{jk}) - I(t_{mk}), 0\} + \{\max(I(t_{mk}) - I(t_{jk}), 0)\}$$

Reasoning Phase II

- For all images jointly:
  - We add similar rule from each seed to target.

For each image k, for each seed s_{ik} and target t_{jk}, add:

$$wt_{ij} : s_{ik} \rightarrow t_{jk}$$

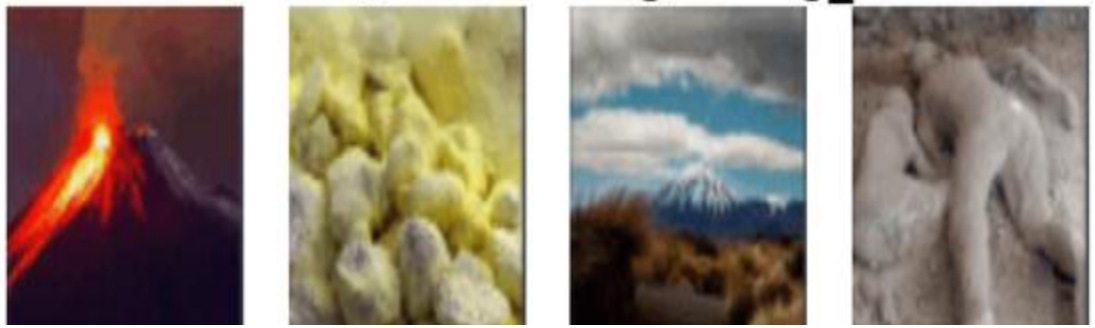
- We add a summation constraint over the targets so that most "infer"-ed targets get higher value.

$$\sum_{j,k} I(t_{jk}) \leq S$$

Results

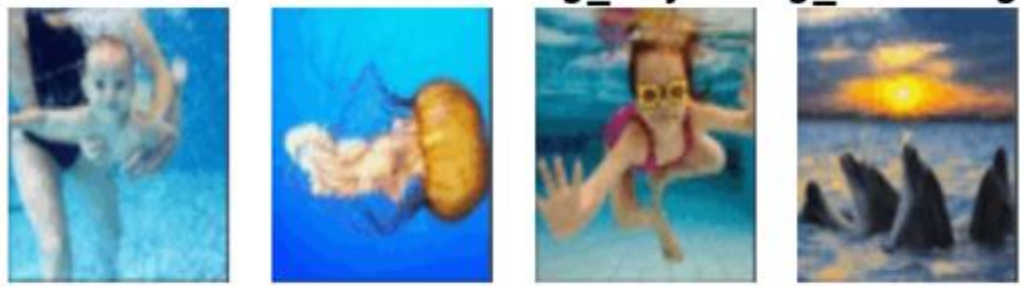
GroundTruth: atlantic
GUR: atlantic_ocean, Baseline: breathtaking_vista

GroundTruth: volcano
GUR: volcano, Baseline: gushing_waterfalls



GroundTruth: moon
GUR: moon, Baseline: Partial_lunar_eclipse

GroundTruth: swim
GUR: swim, Baseline: Diving_kayaking_trekking



* More Examples in imageriddle.wordpress.com/imageriddle/.

Some Negative Ones:

GroundTruth: bird
GUR: penguin, Baseline: bird

GroundTruth: chilled
GUR: soft_drink, Baseline: Serve_chilled

GroundTruth: flower
GUR: tulip, Baseline: flower

GroundTruth: outdoors
GUR: daylight, Baseline: outdoors

Automatic Evaluation

			3.3k		2.8k	
			W2V	WN	W2V	WN
VQA	VB	UR †	59.6	15.7	59.7	15.6
		GUR	62.59	17.7	62.5	17.7
Clarifai	VB	UR †	65	26.2	65.3	26.4
		GUR	65.3	26.2	65.36	26.2
	RR	UR	65.9	34.9	65.7	34.8
		GUR	65.9	36.6	65.73	36.4
	All	UR	68.5	**40.3**	68.57	**40.4***
		GUR	**68.8***	**40.3**	**68.7**	**40.4***
Resnet	VB	UR †	68.3	35	68	33.5
		GUR	66.8	33.1	66.4	32.6
	RR	UR	66.7	38.5	66.7	38.2
		GUR	66.3	38.1	66.2	37.6
	All	UR	68.53	39.9	68.2	**40.2**
		GUR	68.2	39.5	68.2	39.6

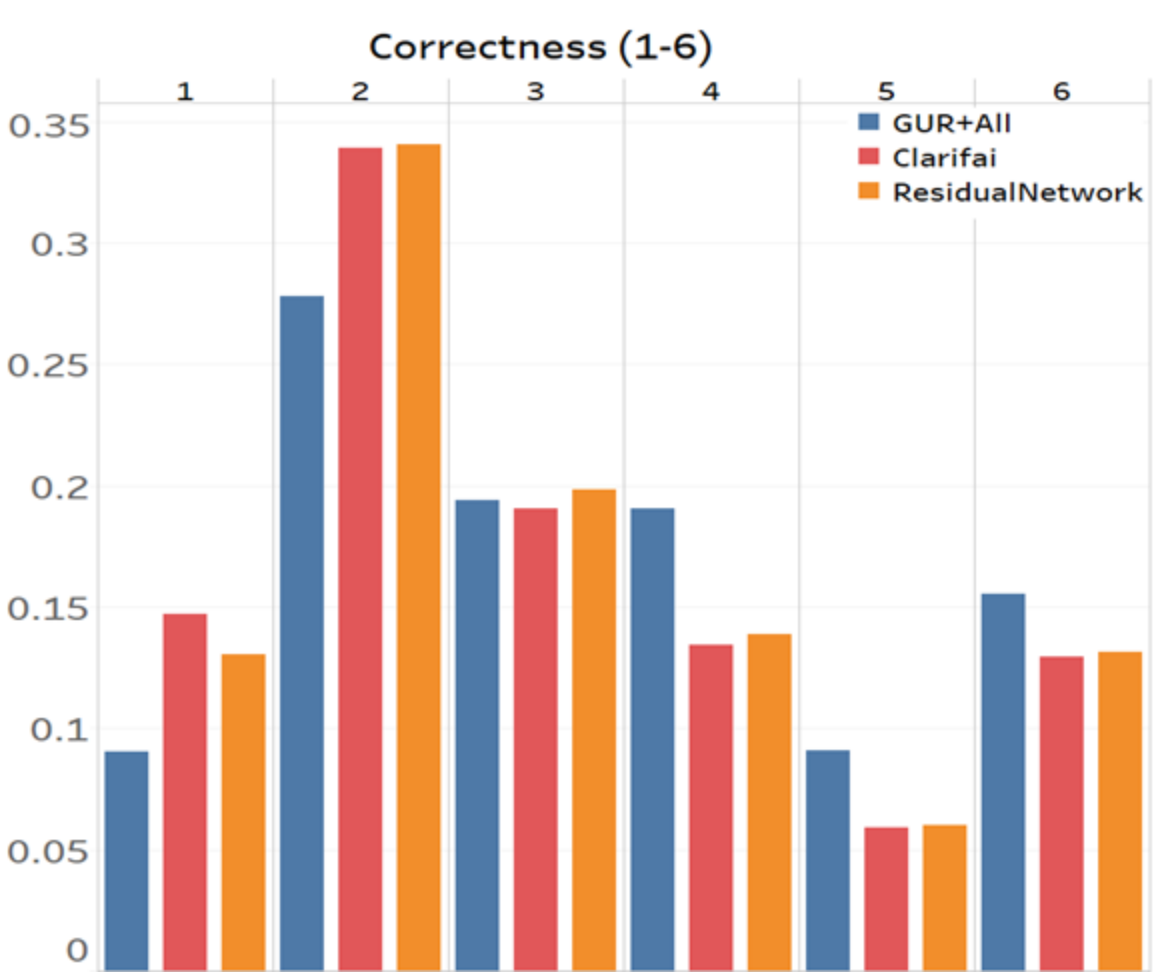
Table 2: Accuracy (in percentage) on the Image Riddle Dataset. Pipeline variants (VB, RR and All) are combined with Bias-Correction stage variants (GUR, UR). We show both word2vec and WordNet-based (WN) accuracies. (*- Best, † - Baselines).

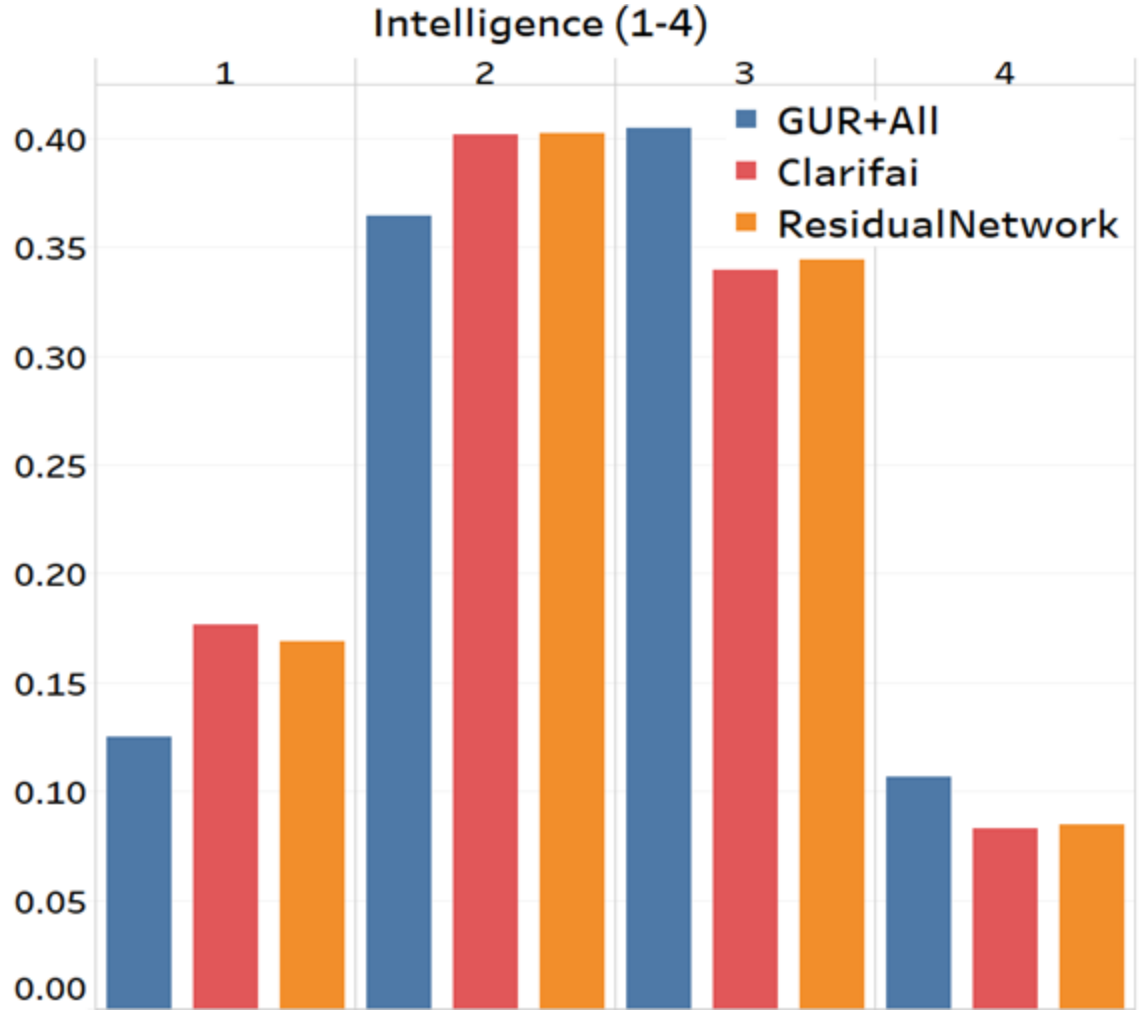
VB: Vision-only greedy baselines.

Note, VQA (attention-based) fails as the question does not contain a clue.

According to WordNet-based accuracy, it shows a **14% jump over greedy baseline!**

AMT Experiment: Comparative Correctness





Our Model is the GUR+All (Blue One). Higher heights on right means more accurate and intelligent (less gibberish) results.

Contributions

1. We introduce the 3K Image Riddles Dataset.

2. We present a probabilistic reasoning approach to solve the riddles with reasonable accuracy.

3. Reasoning module inputs detected words (a closed set of class-labels) and ***logically infers*** **all relevant concepts (belonging to a much larger vocabulary)**.

Outline…

1. **Representation and Reasoning in Images**
   - **[Image Captioning]** Image Understanding Through Scene Description Graphs *IJCAI '15, CVIU '17*
2. **PSL Applications**
   - **[VQA]** Visual Question Answering using a reasoning wrapper. *AAAI '18 (24.5% acceptance)*
   - **[Puzzles/New Challenge]** Image Riddles using Vision Reasoning UAI '18 (30% acceptance), *CVPR '17 Workshop*
3. **An End-to-End Endeavor:**
   - **[Visual Reasoning]** VQA using a Spatial Knowledge Distillation. ***Finally fully end-to-end !!! :)***
4. **Summary of Contribution and Conclusion**

Visual Reasoning Using Spatial Knowledge Distillation

Using Spatial Knowledge to Aid Visual Reasoning, **S Aditya**, R Saha, Y Yang, C Baral.

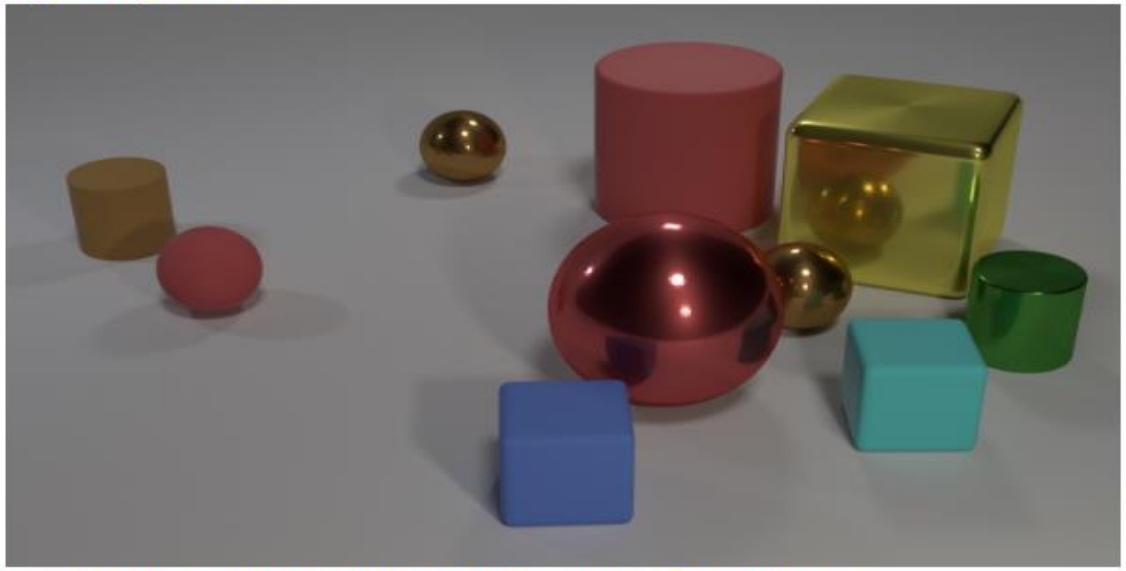
Spatial Knowledge in Visual Reasoning

- ❑ VQA requires knowledge and reasoning.
- ❑ How else to integrate knowledge into Deep Learning Mechanisms
  - The knowledge do not bias/confuse the results
  - Additional supervision does not become a bottleneck in test time
- ❑ **Solution**: Knowledge Distillation and supplying teacher with Spatial knowledge in form of a mask over image
  - What is Knowledge Distillation*?
    - ✓ Teacher has access to additional supervision (knowledge).
    - ✓ Student learns from ground-truth and teacher's soft prediction (wt.ed loss).

* KD: First proposed by Hinton et. Al. 2015 and Vapnik et. Al. 2015.

Dataset

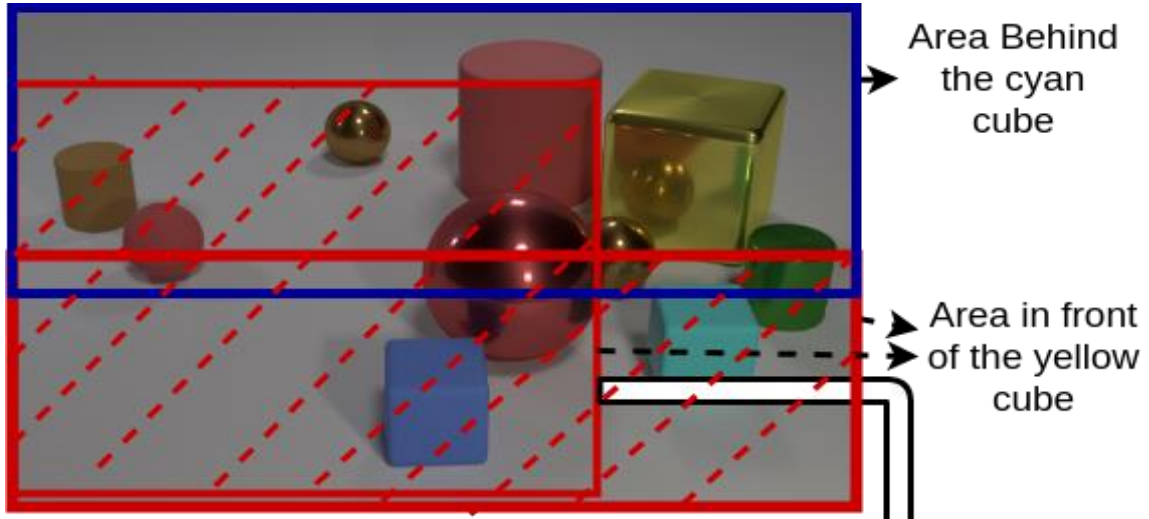
Questions in CLEVR test various aspects of visual reasoning including attribute identification, counting, comparison, spatial relationships, and logical operations.



Q: Are there an **equal number** of **large things** and **metal spheres**?
Q: **What size** is the **cylinder** **that is left of** the **brown metal** thing **that is left of** the **big sphere**?
Q: There is a **sphere** with the **same size as** the **metal cube**; is it **made of the same material as** the **small red sphere**?
Q: **How many** objects are **either** **small cylinders** or **red** things?

CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning

Example Knowledge

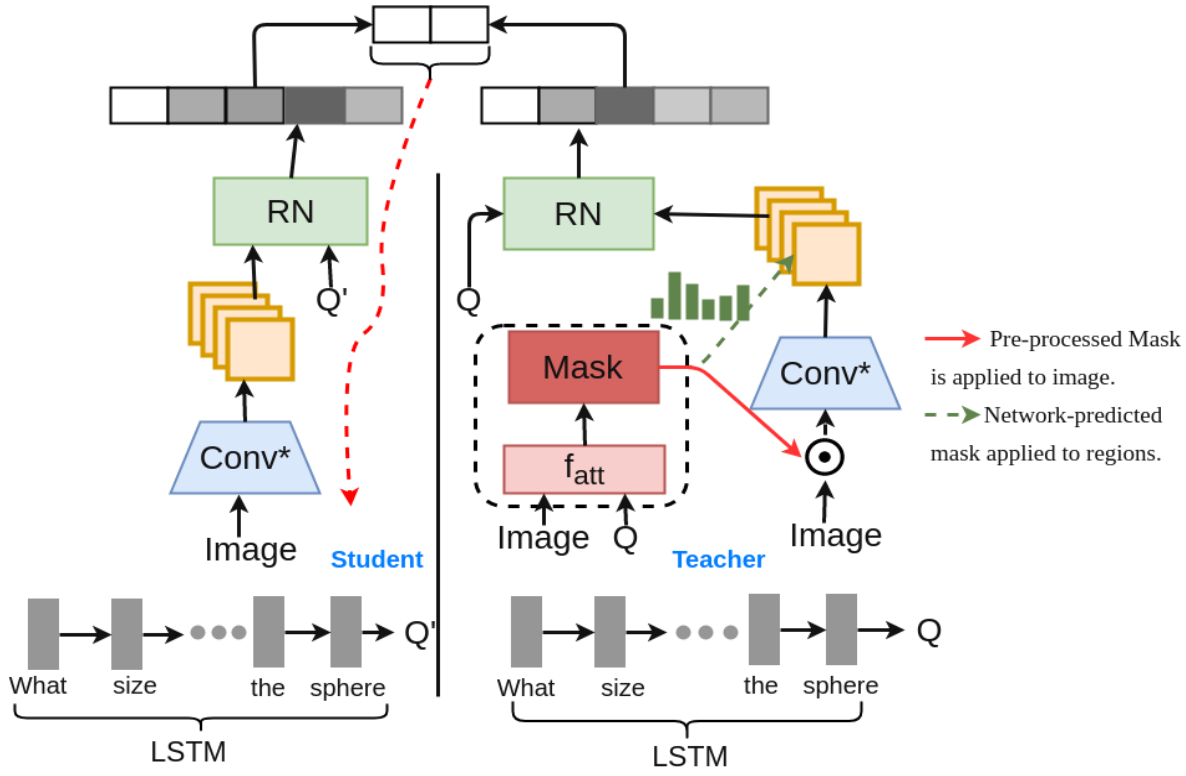


A failure case for the RN paper. Complex knowledge required to understand this question.

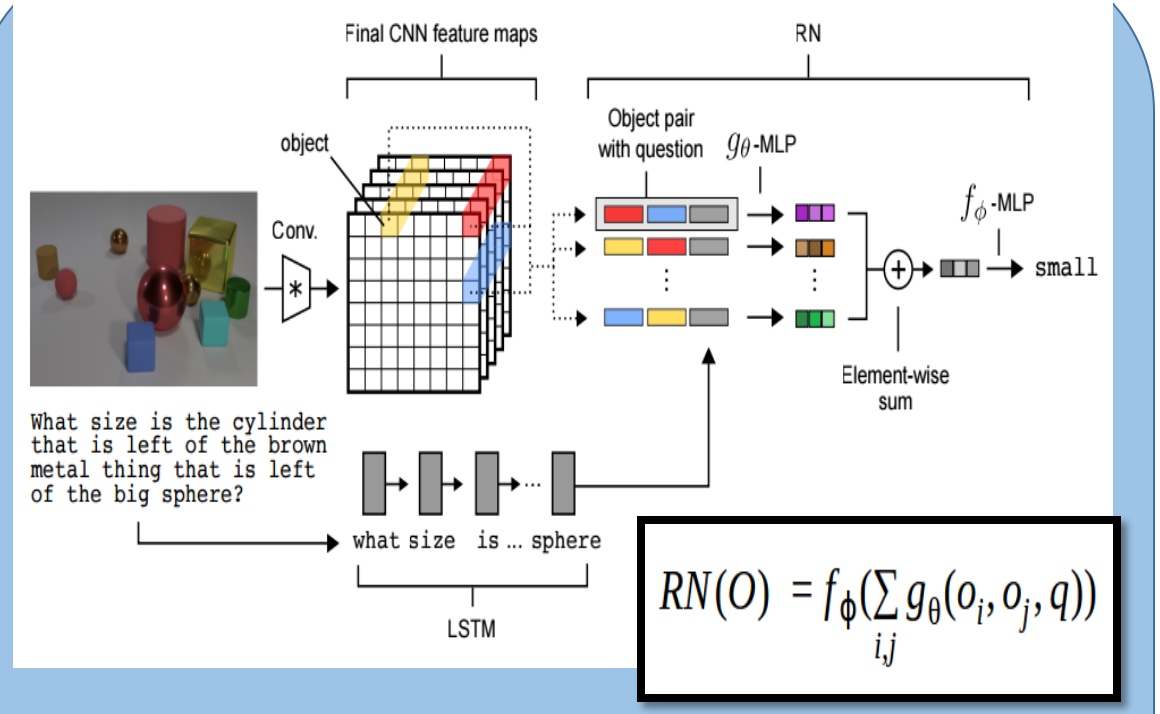
So, we pre-process question to disambiguate and provide a spatial mask that masks out unimportant objects.

Central Idea:
Knowledge can be encoded as spatial masks over the image for a teacher and then distilled to a student network.

Our Architecture and the Baseline



Knowledge Distillation with an attention mask provided or predicted in-network.



[Santoro et. Al. 2017]: CNN, LSTM and the Relational Reasoning Layer

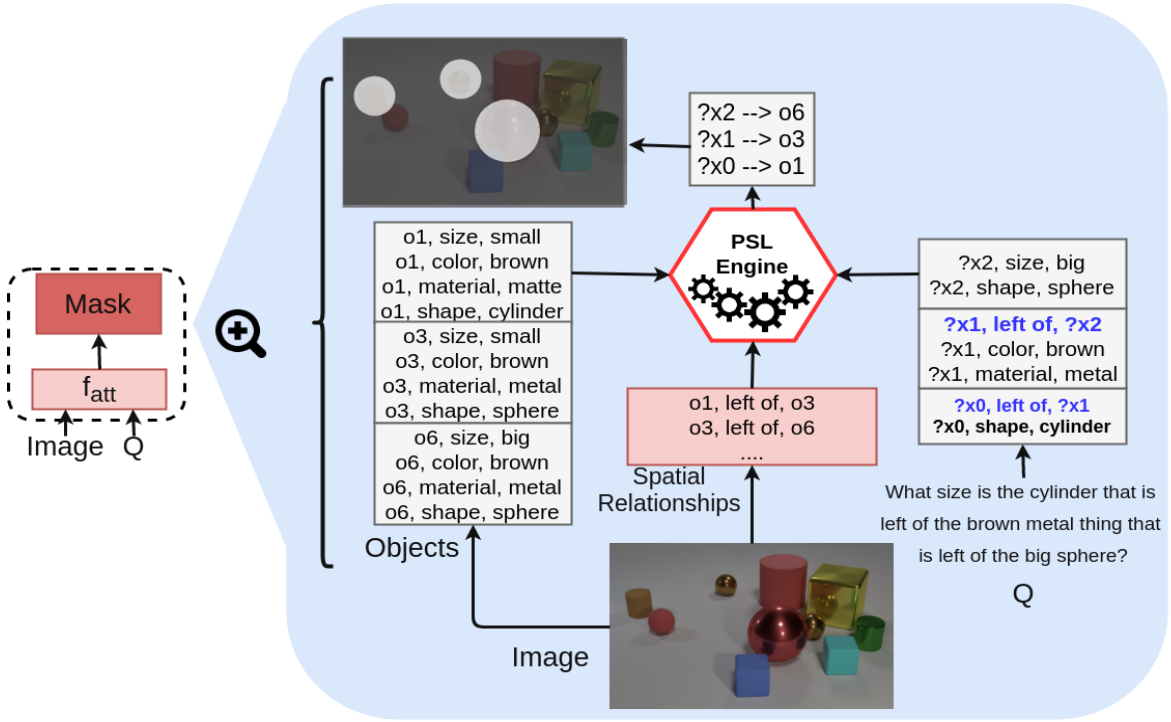
* RN: Relational Reasoning Network proposed by Santoro et. Al. 2017

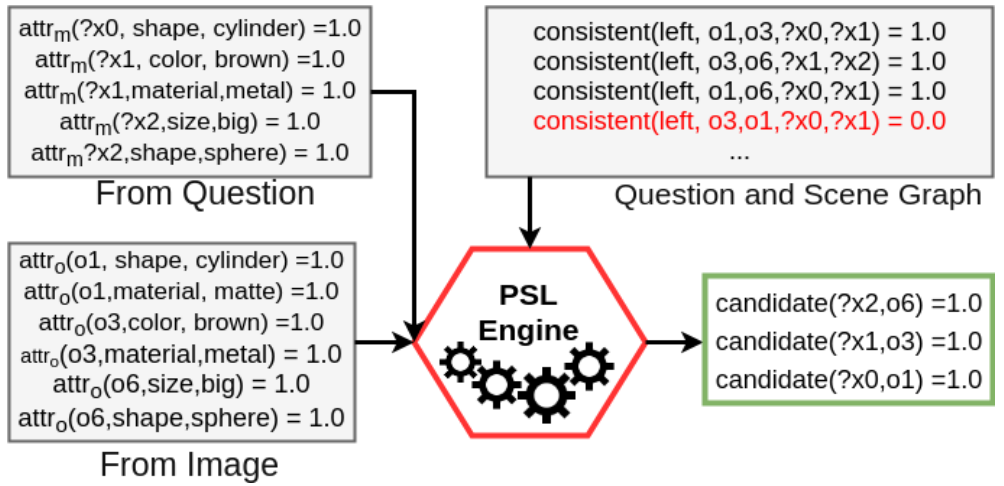
(I) Reasoning and Knowledge

- A simple rule-base.
- Objective: identify matches of textual mentions and actual objects.

$$w_1 : candidate(M, O) \leftarrow object(O) \wedge mention(M) \wedge attr_o(O, P, V) \wedge attr_m(M, P, V).$$

$$w_2 : candidate(M, O) \leftarrow object(O) \wedge mention(M) \wedge candidate(M, O) \wedge candidate(M_1, O_1) \wedge consistent(A, O, O_1, M, M_1).$$





The calculated PSL predicates for the example image and question.

(II) Modeling In-Network Attention

- Attention mechanism is popular.
- We formulate a hybrid question and image based attention over image-regions.

We formulate the problem as attention mask generation over image regions using the image ($\boldsymbol{x_I} \in \mathbb{R}^{64\times64\times3}$) and the question ($\boldsymbol{x_q} \in \mathbb{R}^{w\times d}$). The calculation can be summarized by the following equations:

$$
\begin{aligned}
r_I &= conv^*(\boldsymbol{x_I}).\\
q_{emb} &= LSTM(\boldsymbol{x_q}).\\
v &= tanh(W_I r_I + W_q q_{emb} + b).\\
\alpha &= \exp(v) / \sum_{r=1}^{x*y} \exp(v_r),
\end{aligned}
$$

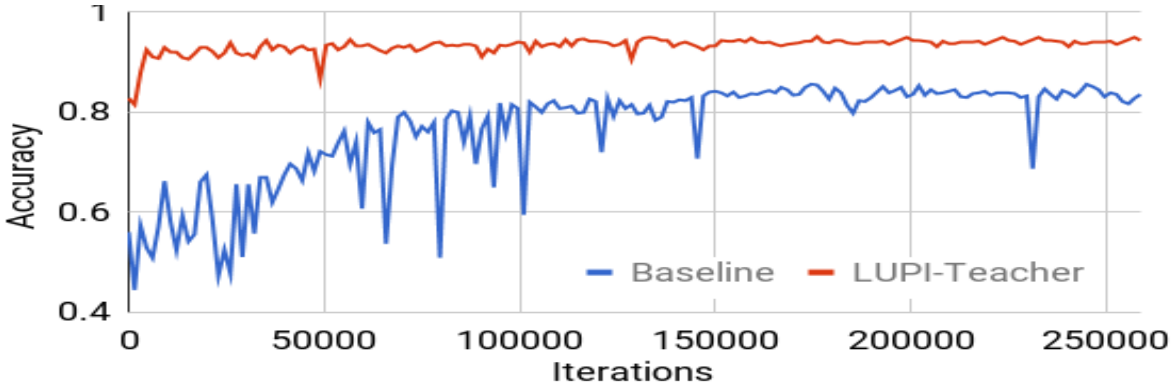
where r_I is $x \times y$ regions with o_c output channels, $q_{emb} \in \mathbb{R}^h$ is the final hidden state output from $LSTM$ (hidden state size is h); $W_I (\in \mathbb{R}^{xyo_c \times xy})$ and $W_q (\in \mathbb{R}^{xy\times h})$ are the weights and b is the corresponding bias.

Results

- Sort-of-Clevr Dataset
  - Teacher's performance over baseline RN improved by large margin (83% --> 95%).
  - Student's performance improved (88.2%, 6% improvement) after iterative knowledge distillation.
- For CLEVR, 5% performance boost is observed

	Baseline	Baseline (Reported)	External Mask Teacher	External Mask Student	In-Network Mask Teacher	In-Network Mask Student	Performance Boost Δ
Sort-of-Clevr	82% ([4])	94% (1-hot questions[4])	**95.7%**	88.2%	87.5%	82.8%	13.7%
CLEVR	53% ([25])	61% ([25])	**58%**	55%	-	-	5%

Table 1: Test set accuracies of different architectures for the Sort-of-Clevr (with natural language questions) and CLEVR dataset. For CLEVR, we have used the Stacked Attention Network (SAN) [25] as baseline and only conducted the external-mask setting experiment as it already calculates in-network attention. Our re-implementation of SAN achieves 53% accuracy on CLEVR. Accuracy reported by [4] on SAN is 61%. The reported best accuracy for Sort-of-Clevr and CLEVR are 94% (one-hot questions [4]) and 97.8% ([33]).



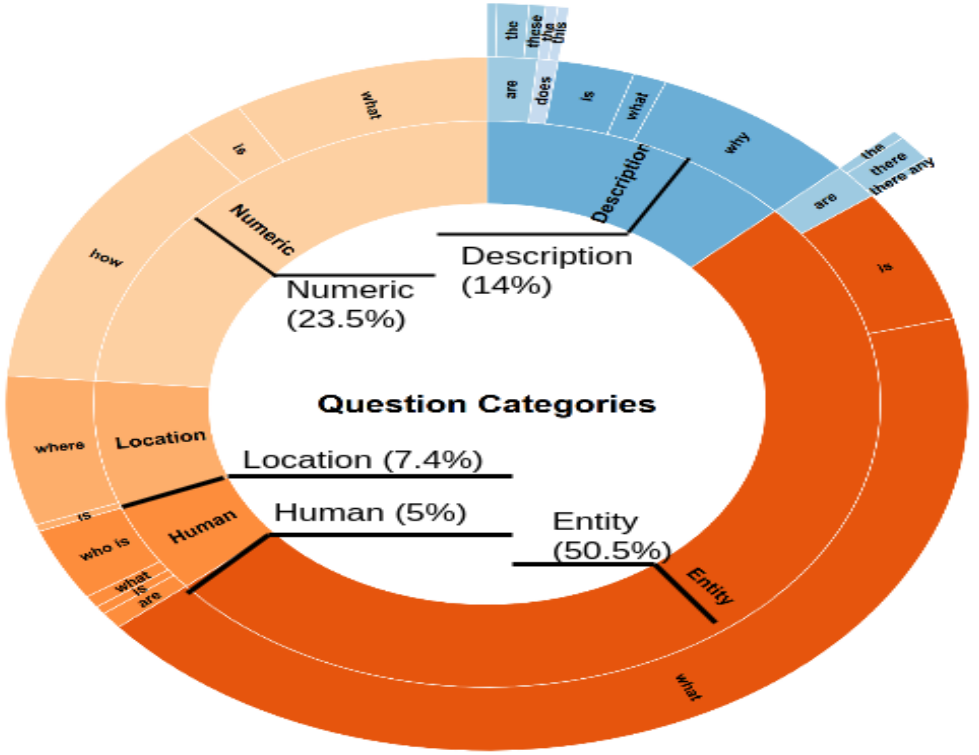
Quick Overview of Other Related Research

Visual Question Categorization

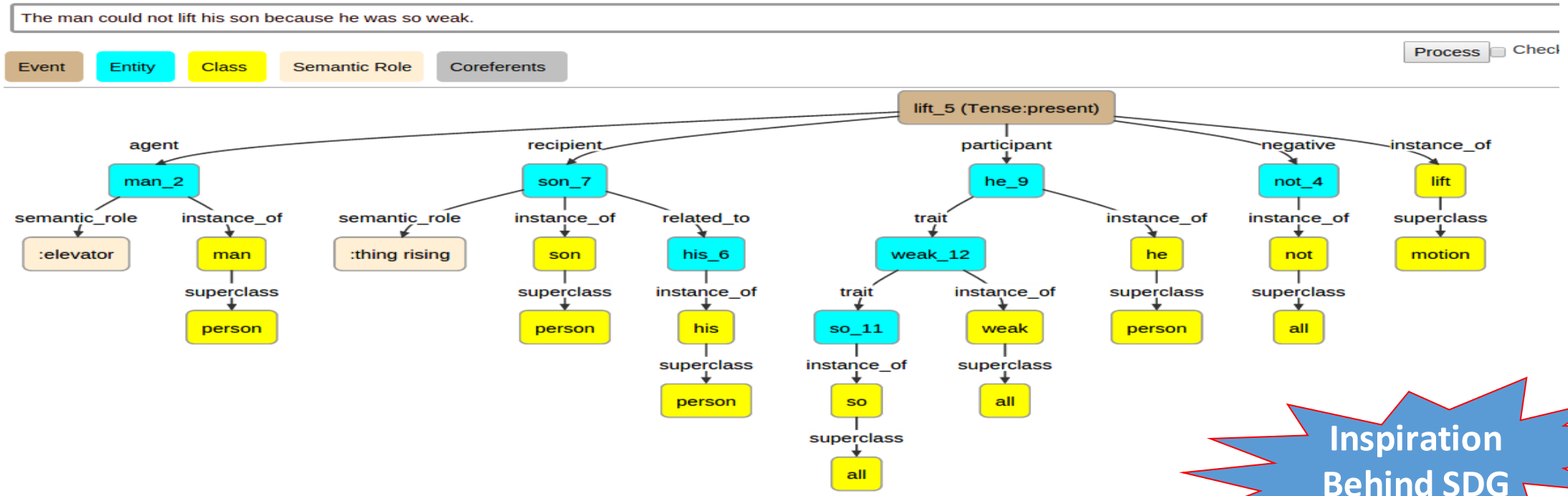
Visual Question Categorization,

_D_ Bandil*, S Aditya*, Y Yang, C Baral,

- We adopt TREC QA categories to categorize visual questions
- Why? Each category requires different knowledge and reasoning.



Related Previous Work (NLP)



Inspiration Behind SDG

Semantic Parsing (A Knowledge Parser):

Towards Addressing the Winograd Schema Challenge-Building and Using a Semantic Parser and a Knowledge Hunting Module. A. Sharma, N. H. Vo, **S. Aditya,** & C. Baral, In ***IJCAI* 2015** (pp. 1319-1325).

Outline…

1. **Representation and Reasoning in Images**
   - **[Image Captioning]** Image Understanding Through Scene Description Graphs *IJCAI '15, CVIU '17*
2. **PSL Applications**
   - **[VQA]** Visual Question Answering using a reasoning wrapper. *AAAI '18 (24.5% acceptance)*
   - **[Puzzles/New Challenge]** Image Riddles using Vision Reasoning UAI '18 (30% acceptance), *CVPR '17 Workshop*
3. **An End-to-End Endeavor:**
   - **[Visual Reasoning]** VQA using a Spatial Knowledge Distillation. ***Finally fully end-to-end !!! :)***
4. **Summary of Contribution and Conclusion**

Summary of Contribution

Summary of Contribution

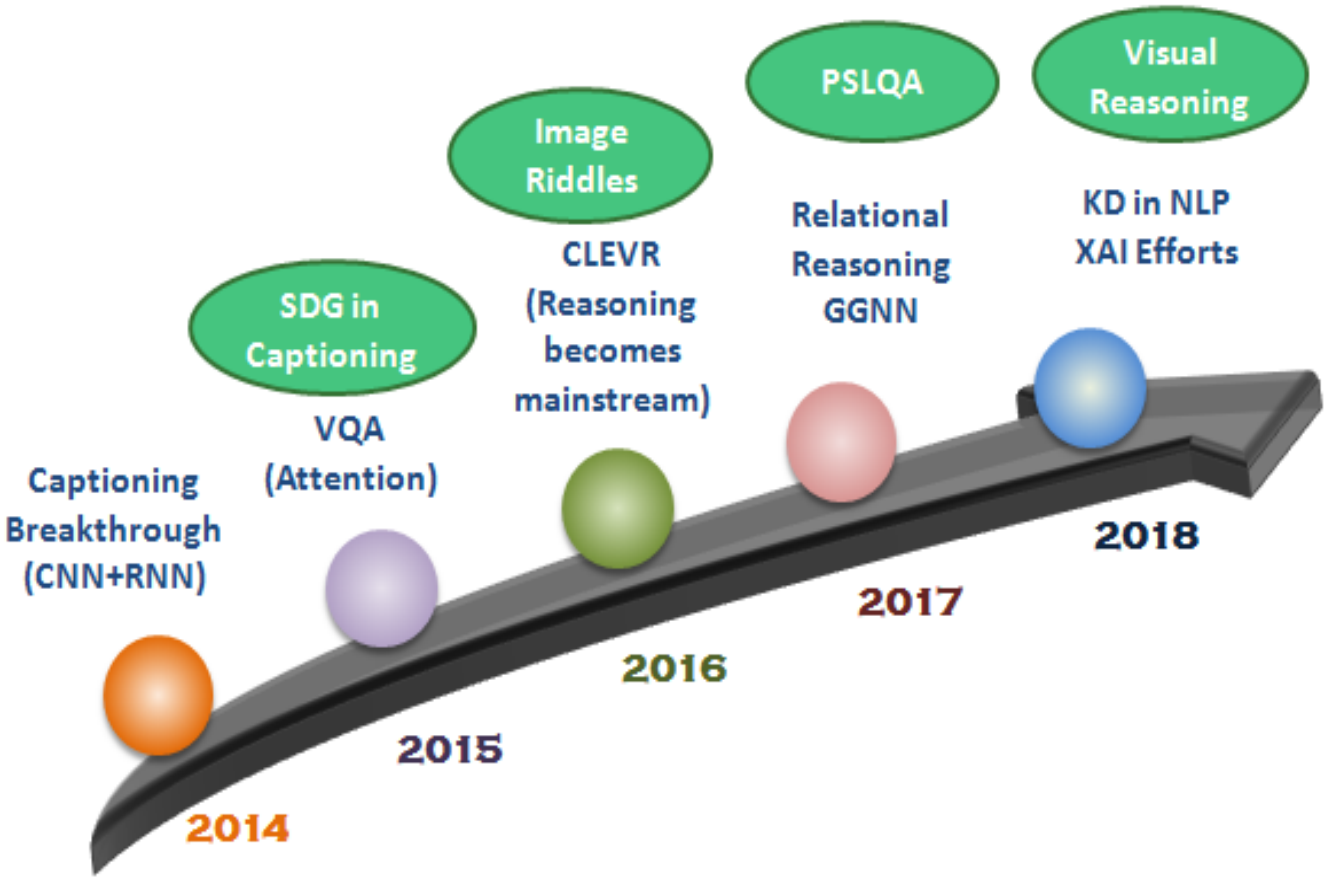
Technical:

- ❑ **[Representation Module]** Two generic knowledge representation.
  - SDG (captioning), Prob. Scene Graph (VQA)
- ❑ **[Knowledge Module]** An automatic knowledge acquisition method
- ❑ **[Reasoning Module]** A fast efficient implementation of PSL (publicly available)

Dataset:

- ❑Image Riddle task and corresponding dataset
- ❑Extensions to Flickr8k, Visual Genome

Judgement of Impact



Milestones in AI (A vision and language perspective):

- RNN+CNN – for V&L datasets
- VQA, CLEVR – Large Datasets
- Attention Mechanism
- Reasoning Alternatives: Relational Reasoning Layer, GGNN

Our Work:

- Combines Prob. Logical Reasoning and Deep Learning
- Utilizes New datasets: VQA, CLEVR
- Proposes New Dataset
- Shows Explicit Reasoning still rocks!

Individual Contribution

- ❑ Image to SDG:
  - Coding (Reasoning + NLP part), Algorithm – 100%
  - Experiments (first AMT experiment helped by Yezhou)
- ❑ Image Riddles:
  - Coding, Algorithm, Experiments – 100%
- ❑ PSL Reasoning Engine
  - Coding – 100%
- ❑ VQA:(from scratch)
  - Coding, algorithm – 100%
  - Experiments – 99%
  - Experimental Setup partially helped by Divyanshu
- ❑ Spatial Reasoning
  - Algorithm – 100%
  - Coding – 90%
  - Coding and Experiments – Partially helped by Rudra

Future Directions

Concluding Claim: For higher-level reasoning and reasoning using knowledge explicit reasoning frameworks will still be important.

- ❑ Knowledge Base Completion
  - Completing ConceptNet based in application-specific manner (closed-world).
- ❑ Extending PSL
  - How-why questions.
  - Extend to include full FOL syntax.
  - Probabilistic ILP to learn PSL rules
- ❑ Large-scale Dataset Compilation
  - More large-scale collaboraitve efforts for multiple Dataset based on different types of knowledge

Acknowledgements

I thank all my collaborators and peers from Dr. Baral's lab, ASU APG and Prof. Yiannis' lab@UMD.

- ❑ Advisors and Mentors
  - Prof. Chitta Baral, CIDSE, ASU
  - Dr. Yezhou Yang, CIDSE, ASU
  - Prof. Yiannis Aloimonos, UMIACS, UMD College Park
  - Dr. Cornelia Fermüller, UMIACS, UMD College Park
  - Dr. Maneesh Singh, Director, AI, Verisk Analytics
- ❑ Fellow Peers/Collaborators @ASU
  - Arpit Sharma, Kazuaki
  - Trideep Rath, Divyanshu Bandil, Rudra Saha

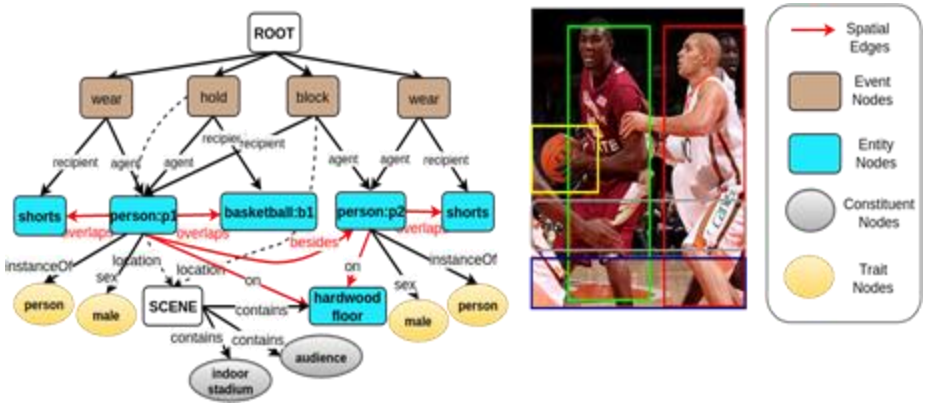
Important References

1. "Vqa: Visual question answering.“, Antol, Stanislaw, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2425-2433. 2015.
2. "Image retrieval using scene graphs." Johnson, Justin, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3668-3678. 2015.
3. "Scene Graph Generation by Iterative Message Passing." , Xu, Danfei, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. *arXiv preprint arXiv:1701.02426* (2017).
4. “Generating visual explanations “, Hendricks, L.A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B. and Darrell, T., 2016, October. In *European Conference on Computer Vision* (pp. 3-19). Springer International Publishing.
5. “Rationalizing neural predictions.” Lei, T., Barzilay, R. and Jaakkola, T., 2016. *arXiv preprint arXiv:1606.04155*.

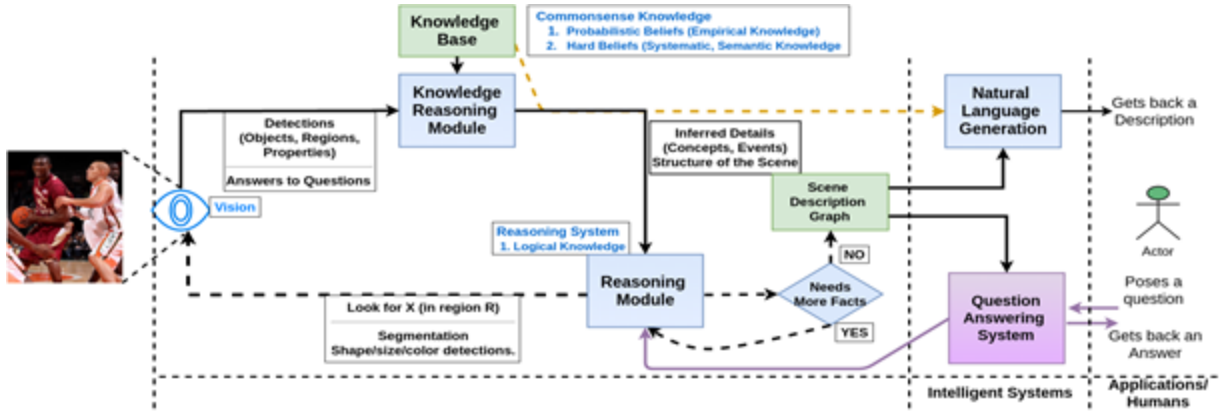
Selected References

- Answering Image Riddles using Vision and Reasoning through Probabilistic Soft Logic, **S Aditya**, Y Yang, C Baral, Y Aloimonos - arXiv preprint arXiv:1611.05896, 2016
- "Visual common-sense for scene understanding using perception, semantic parsing and reasoning." **Somak Aditya**, Yezhou Yang, Chitta Baral, Cornelia Fermuller, and Yiannis Aloimonos. In *2015 **AAAI** Spring Symposium Series*. 2015.
- "Towards Addressing the Winograd Schema Challenge-Building and Using a Semantic Parser and a Knowledge Hunting Module." Sharma, Arpit, Nguyen Ha Vo, **Somak Aditya**, and Chitta Baral. In ***IJCAI***, pp. 1319-1325. 2015.
- "From images to sentences through scene description graphs using commonsense reasoning and knowledge.", **Somak Aditya**, Yezhou Yang, Chitta Baral, Cornelia Fermuller, and Yiannis Aloimonos. *arXiv preprint arXiv:1511.03292* (2015).
- "DeepIU: An Architecture for Image Understanding.", **Aditya, Somak**, Yezhou Yang, Chitta Baral, Yiannis Aloimonos, and Cornelia Fermuller, Advances of Cognitive Systems 2016
- "Image Understanding using vision and reasoning through Scene Description Graph.", **Aditya, Somak**, Yezhou Yang, Chitta Baral, Yiannis Aloimonos, and Cornelia Fermuller, **Computer Vision and Image Understanding 2017**
- "Explicit Reasoning over End-to-End Neural Architectures for Visual Question Answering.", **Aditya, Somak**, Yezhou Yang, Chitta Baral, **AAAI 2018**

Questions?

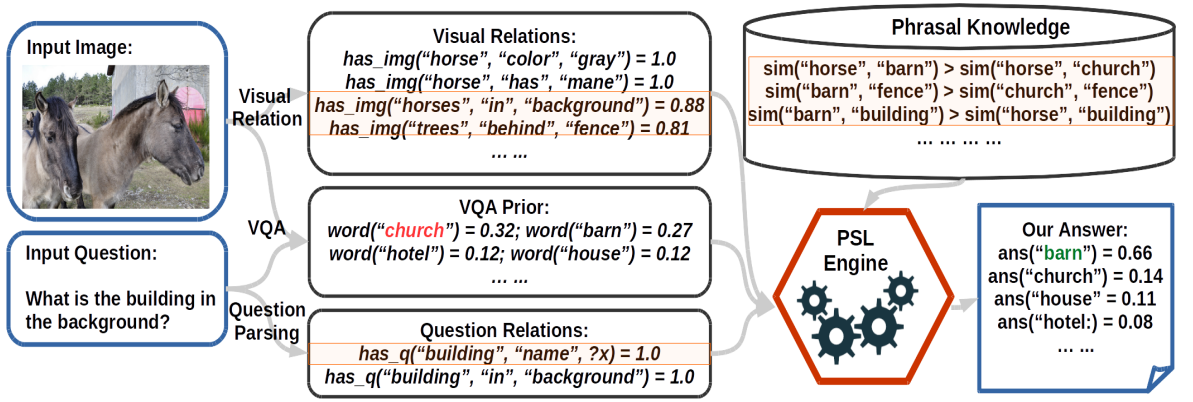


Scene Description Graph



An Image Understanding Architecture

Image Riddles



Visual Question Answering