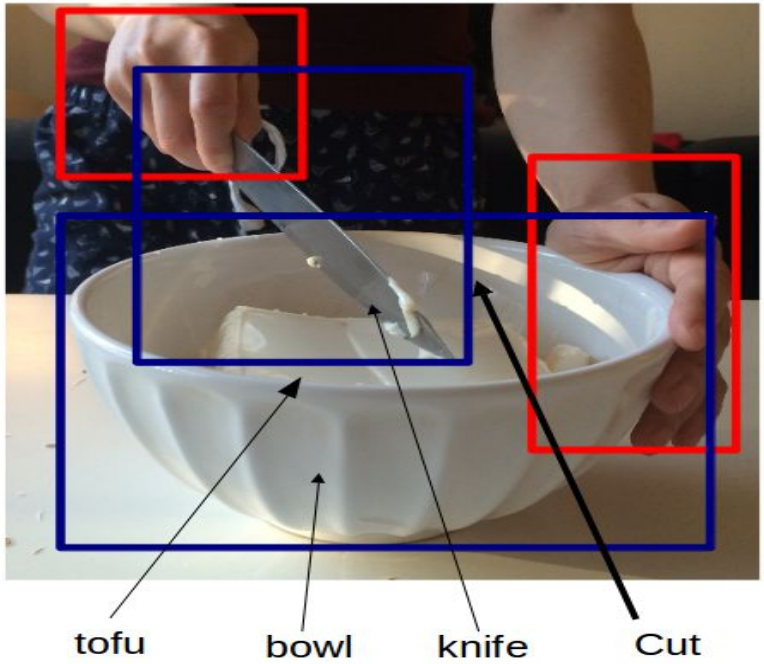
DEEPIU: AN ARCHITECTURE TO UNDERSTAND IMAGES

Somak Aditya, *CIDSE, Arizona State University*
Chitta Baral, *CIDSE, Arizona State University*
Yezhou Yang, *University of Maryland College Park*
Yiannis Aloimonos, *University of Maryland College Park*
Cornelia Fermuller, *University of Maryland College Park*

"UNDERSTANDING"

- What is Understanding? (do you understand)
  - Well-studied in Educational Domain.
  - Ask students questions about a subject, if the student can answer then he/she "understands" *it.*
  - UNDERSTANDING here *is equivalent to* Question-Answering
- Quality of understanding (how much do you understand)
  - Increase difficulty of questions.
  - According to Bloom's Taxonomy **[1]**, they are:
    - 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

Image Understanding



Example From [2]

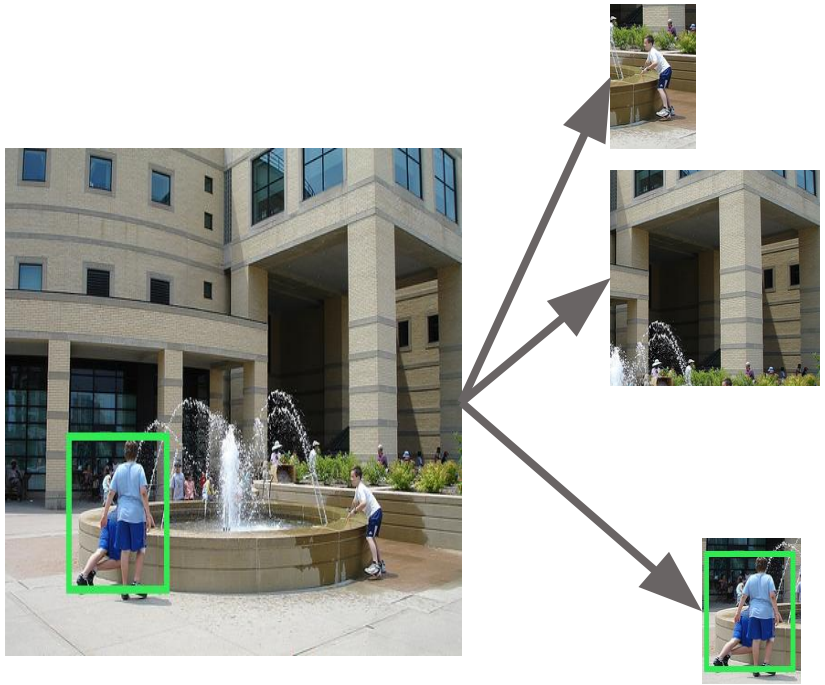
i. (Knowledge) List the objects in the image.
ii. (Comprehension) what will the man do next?
iii. (Application) how to cut tofu?
iv. (Analysis) Why is the man holding the bowl with his other hand?
v. (Synthesis) Can you propose how else to cut a tofu?
vi. (Evaluation) Is there a better way to cut a tofu?

STATE-OF-THE-ART

- Current Systems:
  - Question-Answering **[3,4]**
    - Focusses to finding, locating or co-locating objects.
  - Image (or video)-Captioning **[5,6,7]**
    - End-to-end mapping from image space to text space.
    - Quiet hard to evaluate.
  - Semantic Representation of Images **[8,9]**
    - Uses intermediate representation for caption generation. But, mainly restricted to spatial relations among objects.
- Drawback:
  - Overall Goal: find, locate objects, regions or their properties.
  - All these systems aim to find the answers to “Knowledge” questions.
    - *What color*, *how many*, *is there* etc.
  - ***What about the other categories?***

WHAT DO WE NEED?

- We need Reasoning module.
- Establish “Vision” as an active process.
  - Guided by Reasoning module.
- Reasoning module uses background knowledge to:
  - Rectify Noisy detections.
  - Or, guide Vision Module to detect or gather more information.



Different Information in Different Granularities.
Difficult to get in one shot.

- Is there a flower in the small garden? (Patch1)
- What type of shoe is the man in the left wearing?

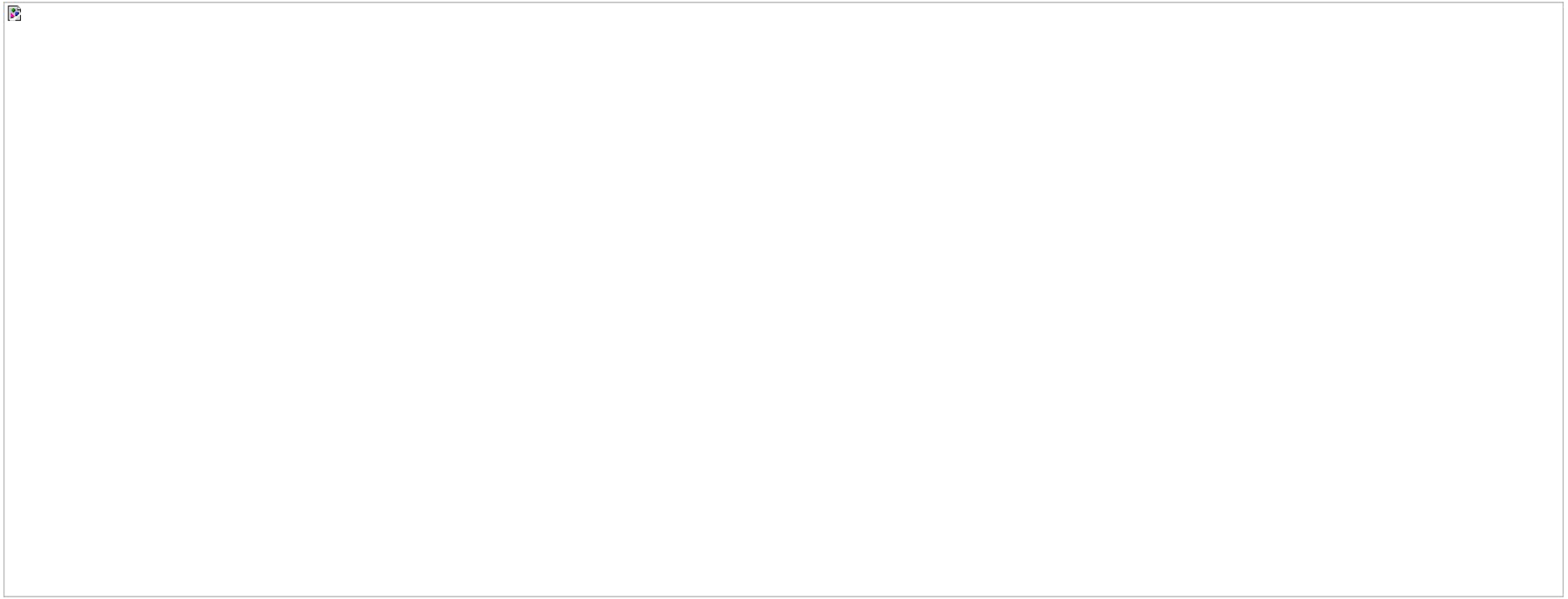
DeepIU: The Architecture

DEEP IMAGE UNDERSTANDING

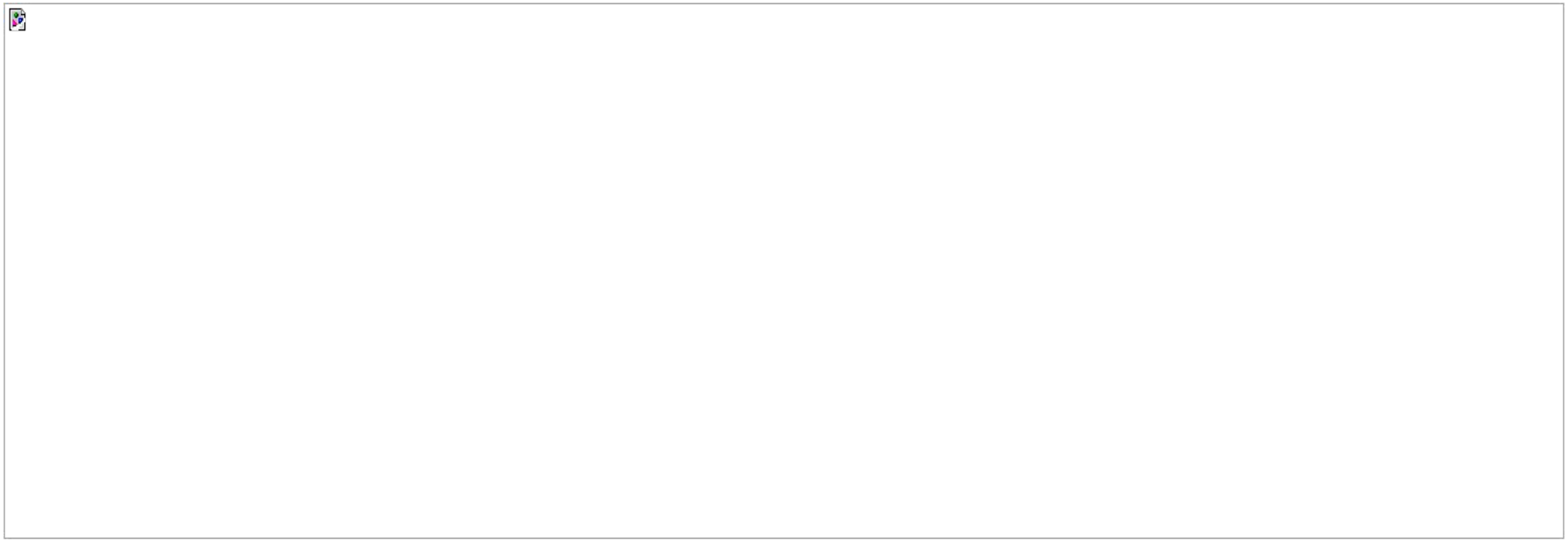
So, what modules do we need?

- Vision Module:
  - “Eyes” to see.
- Reasoning Module:
  - A “brain” to reason and advise.
- Knowledge Base
  - “Books” to read
    - Humans uses knowledge to reason about his/her surroundings.
    - We gather knowledge from daily experiences, by reading (also by seeing etc.).

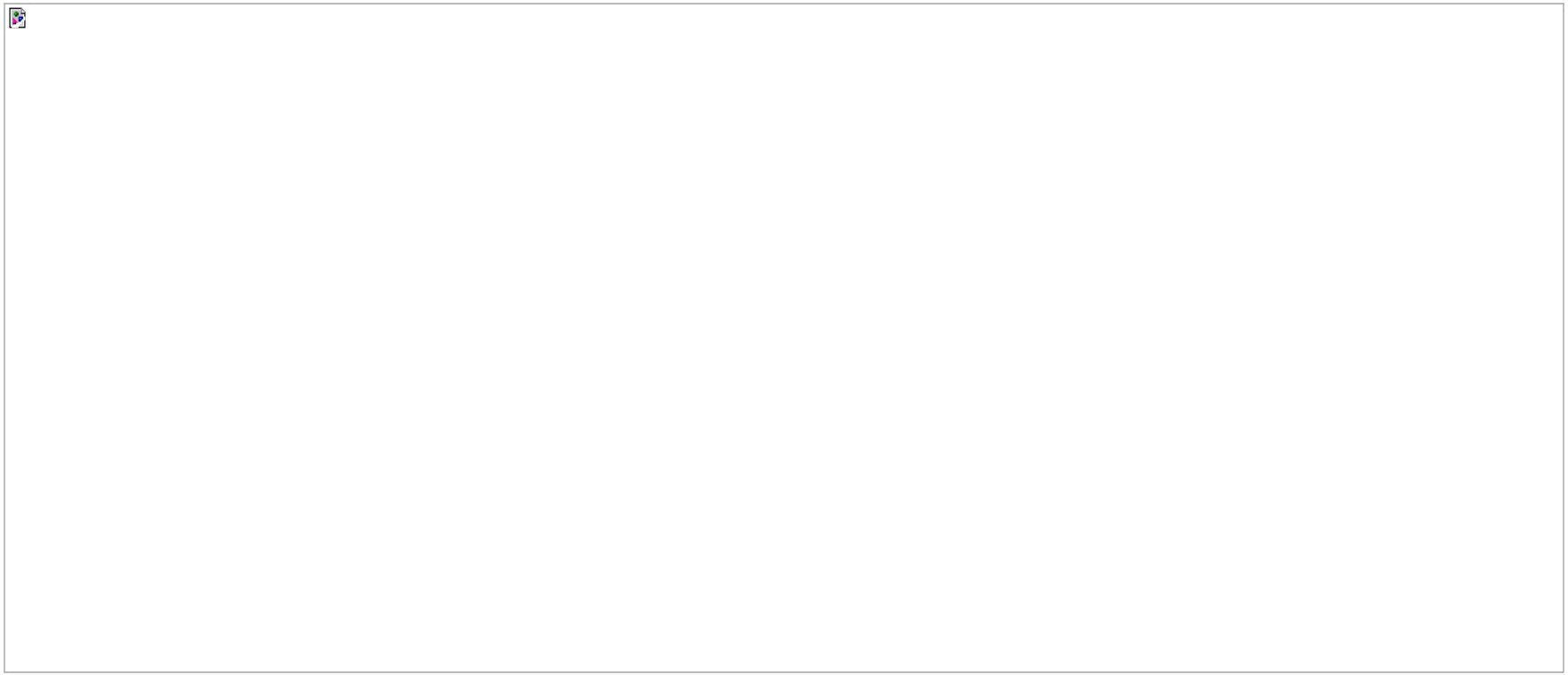
DEEP IMAGE UNDERSTANDING



EXAMPLE OF THE LOOP



EXAMPLE OF THE LOOP



Implementation

PRELIMINARY IMPLEMENTATION

A preliminary implementation and first set of experiments on Flickr 8k, 30K and MS-COCO:

1. Motivation: Representing the Knowledge.
2. Description of the Framework
3. Results

A MEANINGFUL SEMANTIC REPRESENTATION

Image Space

Scene Description Graph

Graph = set of edge(u,v,label) triples

NLG

A man is standing near a fountain outside a building...

Text Space

(Probabilistic) Logical Languages

Is a person drinking from the fountain?

Yes

QA/Reasoning space

Image Retrieval

1. MOTIVATION

INTUITION : SEMANTIC PARSERS

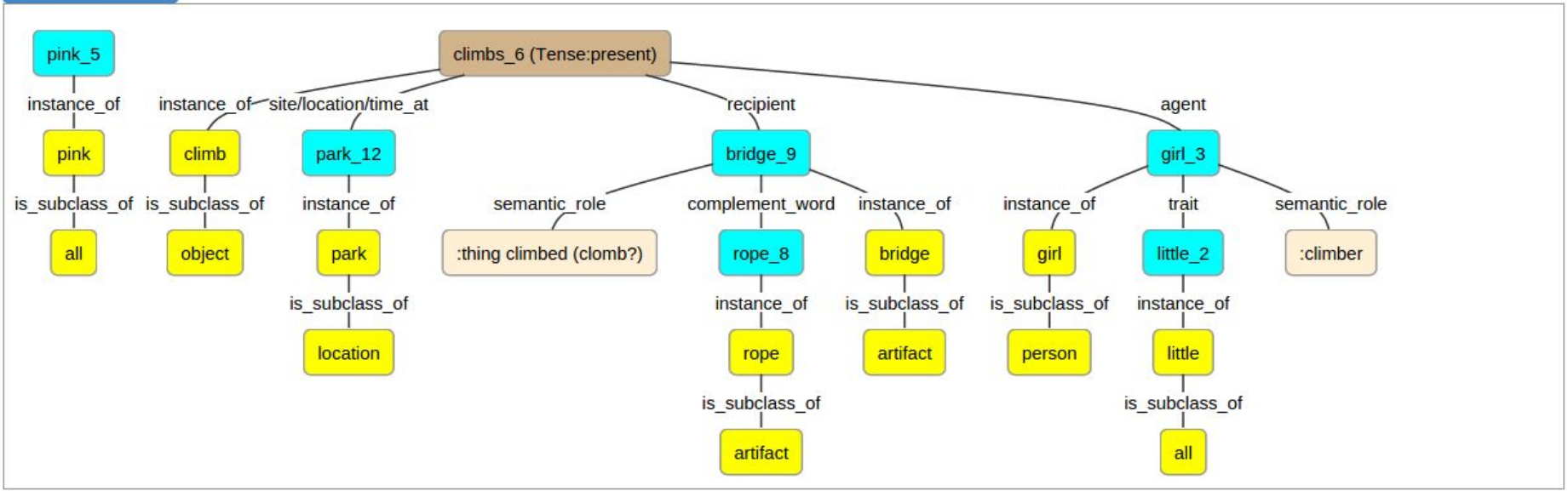
A little girl in pink climbs a rope bridge at the park .

Check Coreference

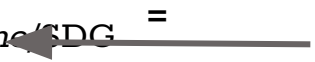
Clear Submit

Event Entity Class Semantic Role Co-referents

Show Full Screen



OUTLINE

1. Image  *Scene*/SDG  Text (Sentence/Set of sentences)
2. **Definitions:**
   a. Observed Scene Constituents: constituents of a scene, that we actually see in an image.
   b. Inferred Scene Constituents: constituents of a scene, that has to be inferred, cannot be scene directly
   c. Scene: made up of entities (nouns), events (verbs) and Inferred Scene Constituents (ISCs)
3. **Overall framework:**
   a. Detect objects, (“scene” categories, “scene constituent”s).
   b. Infer Events and ISCs.
   c. Search for or construct *Scene*/SDG.

FRAMEWORK OUTLINE

1. Perception System:
   a. (200) Object classes : *accordion, airplane, ant, person….*
   b. (205) Scene Classes: *abbey, airport_terminal, amphitheater…..*
   c. (1000) Observed Scene Constituents: *person ride bike, dog wear collar….*
      i. *Created from (enhanced) Flickr8k Phrase annotations.*
      ii. *Lemmatized, stop-words removed and then top 1000 frequent OSCs chosen.*
2. Pre-processing:
      i. Create Scene Classes-to-ISC mappings (with pri
      ii. Collect annotations, scene class detection tuples.
      iii. Knowledge Extraction and Storage
3. Reasoning Framework:
      i. Inferring SDGs through Reasoning.

boxing_ring ::
ring lines, people wear shorts,
people wear boxing shorts
Auditorium ::
People, staircase like structure, people sit,
people sit in rows

Perception System Details

Object Recognition:

- we use the trained bottom-up region proposals and convolutional neural networks (CNN) object detection method from (Girshick et. al. 2014) [11].
- It considers 200 common object classes (denoted as N) and it is trained on ILSVRC 2013 dataset.

Scene (category) recognition:

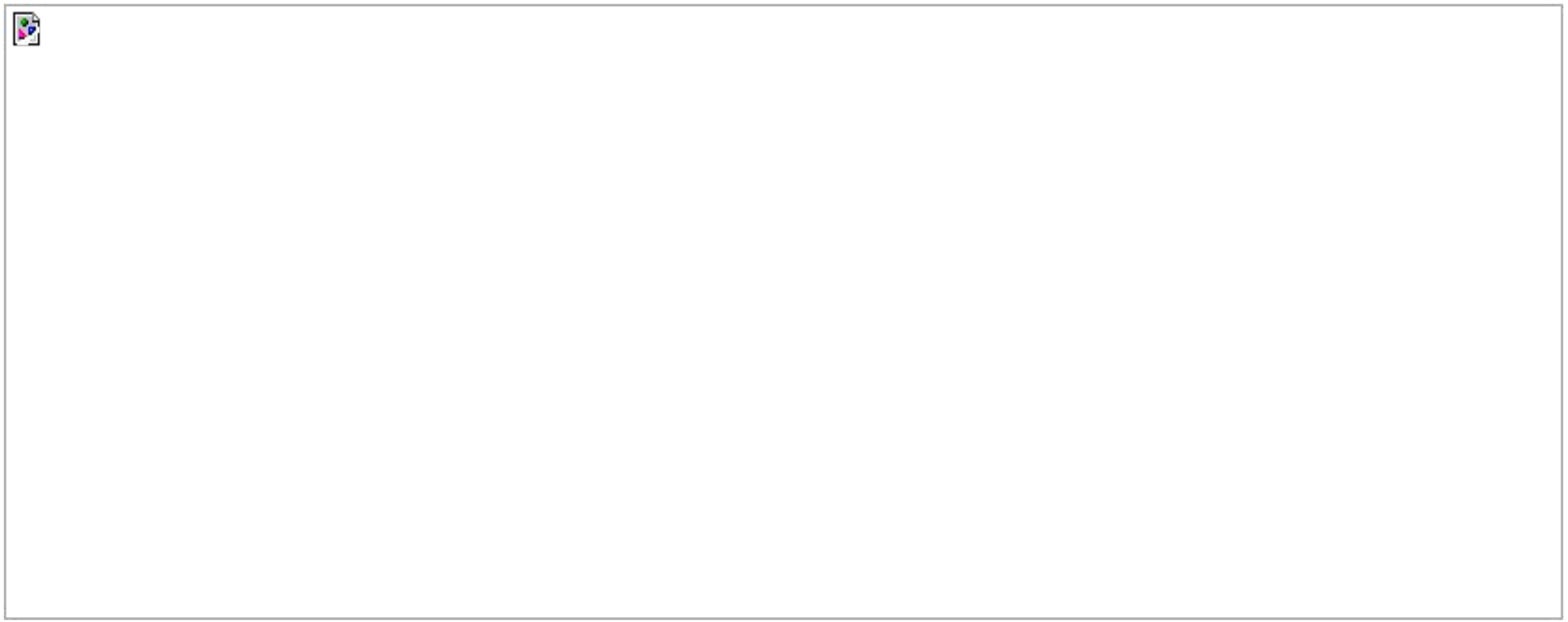
- we use the trained CNN scene classification method from (Zhou et al. 2014) [12].
- The classification model is trained on 205 scene categories (denoted as S).

Perception System Details

Scene Constituent (OSC) recognition:

- Augment the Flickr 8K image dataset with human annotation of constituents using Amazon Mechanical Turks.
  - We ask the annotators to annotate what objects are doing or properties of objects.
  - We allow the labelers to use free-form text for describing constituents to reduce annotation effort.
  - We obtain a standardized set of constituents by performing stop-words removal, parts-of-speech processing to retain nouns, adjectives and verbs. We use the top 1000 frequent phrases (denoted as C).
  - Example: dog run, dog play, kid play, person wear short etc.
- For each image, we use the pre-trained CNN model from (Krizhevsky et al. 2013)**[13]** to extract a 4096 dimensional feature vector (using Donahue et al. 2014)**[14]**.
- We then trained a multi-label SVM to do constituents recognition using these deep features.

Perception System Output



Pre-Processing

We perform a one-time pre-processing to store various kinds of knowledge:

1. Store Scene Classes-to-ISC mappings (with priors).
   a. In this implementations, mappings are manually annotated.
   b. Priors are learnt from training images.
2. Collect annotations, scene class detection tuples.
   a. Training annotations provided alongwith each training image.
3. Knowledge Extraction and Storage
   a. Knowledge- Base
      i. Stores the knowledge of how commonly occurring entities and events interact.
   b. Bayes Network
      i. Stores the knowledge of co-occurrence of entities and inferred-scene-constituents (ISCs).

KNOWLEDGE BASE

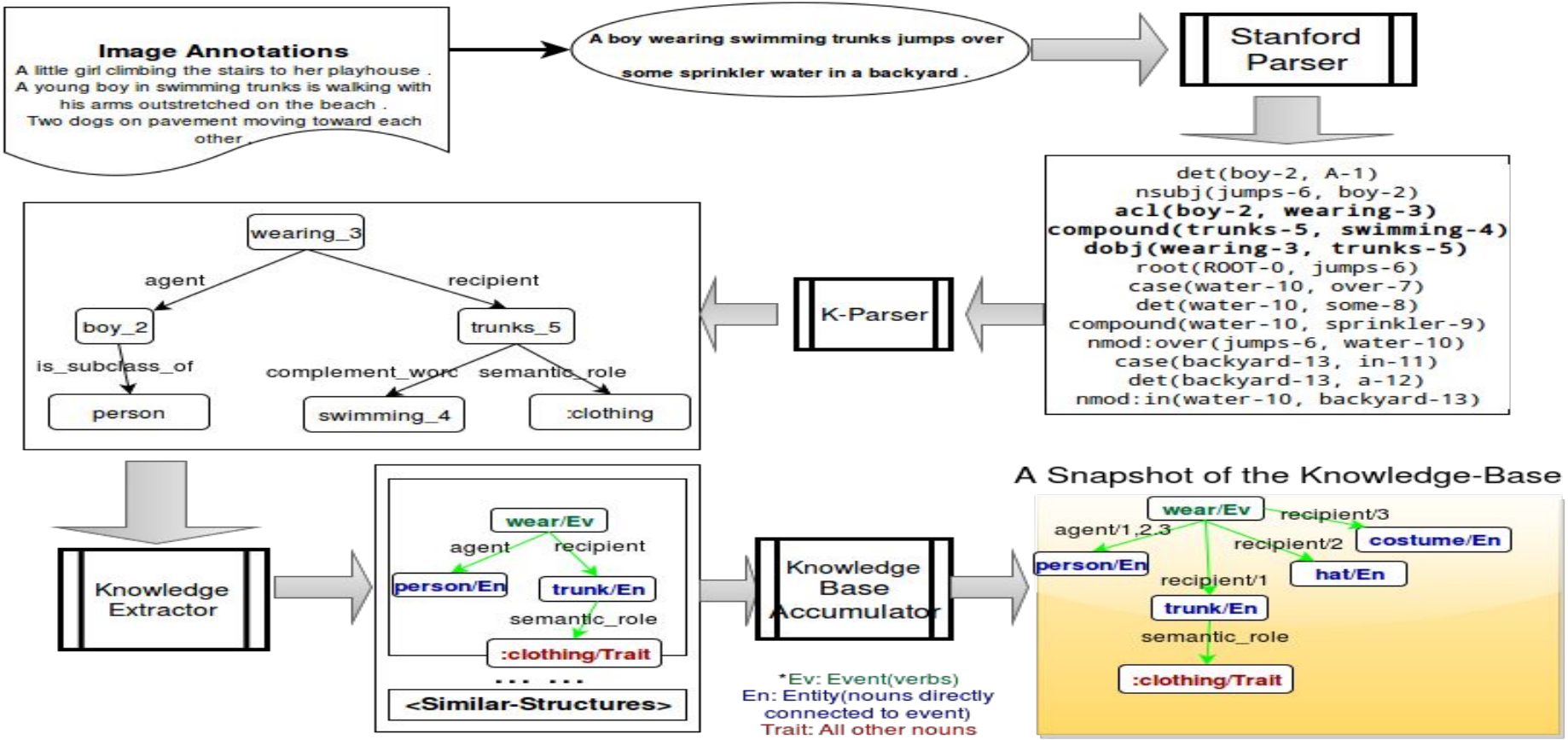
Knowledge Base Construction: (KB = G,C)

a. Parse each sentence using K-parser.
   - We get a knowledge graph.
b. Merge them using overlapping entities (nouns), events (verbs).
c. Retain the individual graphs.

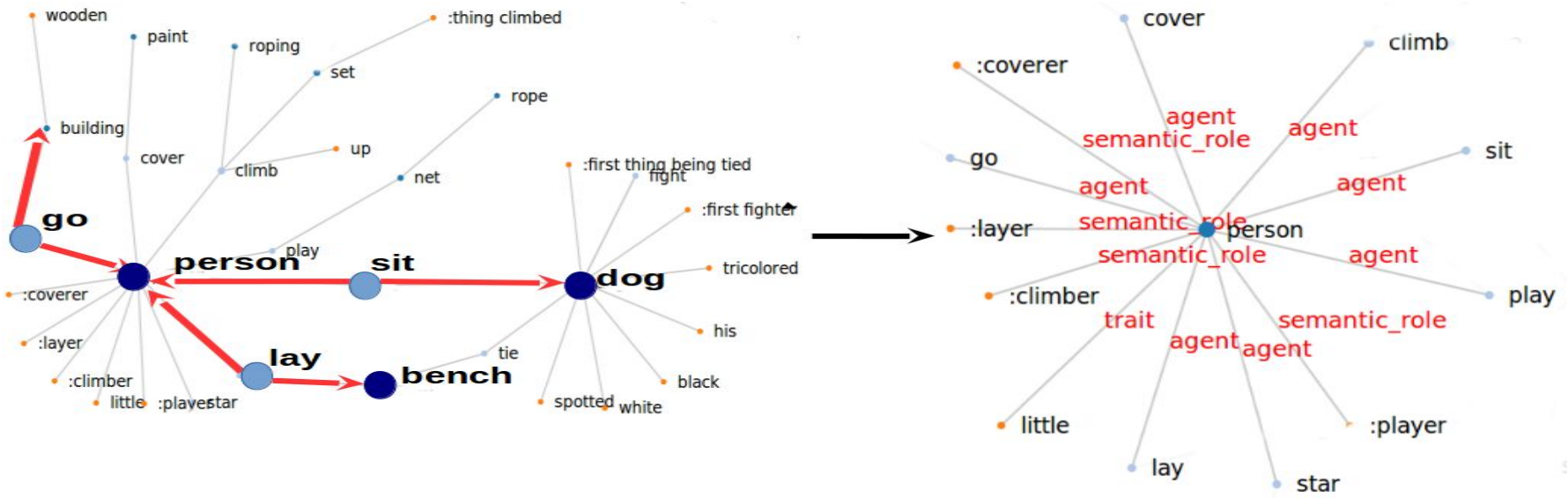
Nodes In the Knowledge-Graph:

a. Events: sit, walk, climb, wear...
b. Entities: person, dog, bench, trunk, tree, bird....
c. Concepts: (processed) K-Parser graph of a sentence
d. Edges: event-event, event-entity edges as assigned by K-parser.

KB CONSTRUCTION



KB EXAMPLE



Event: For “person” and “bench”: “lay” .
candidate(V) <- edge(V,agent,person)^edge(V,recipient,bench).

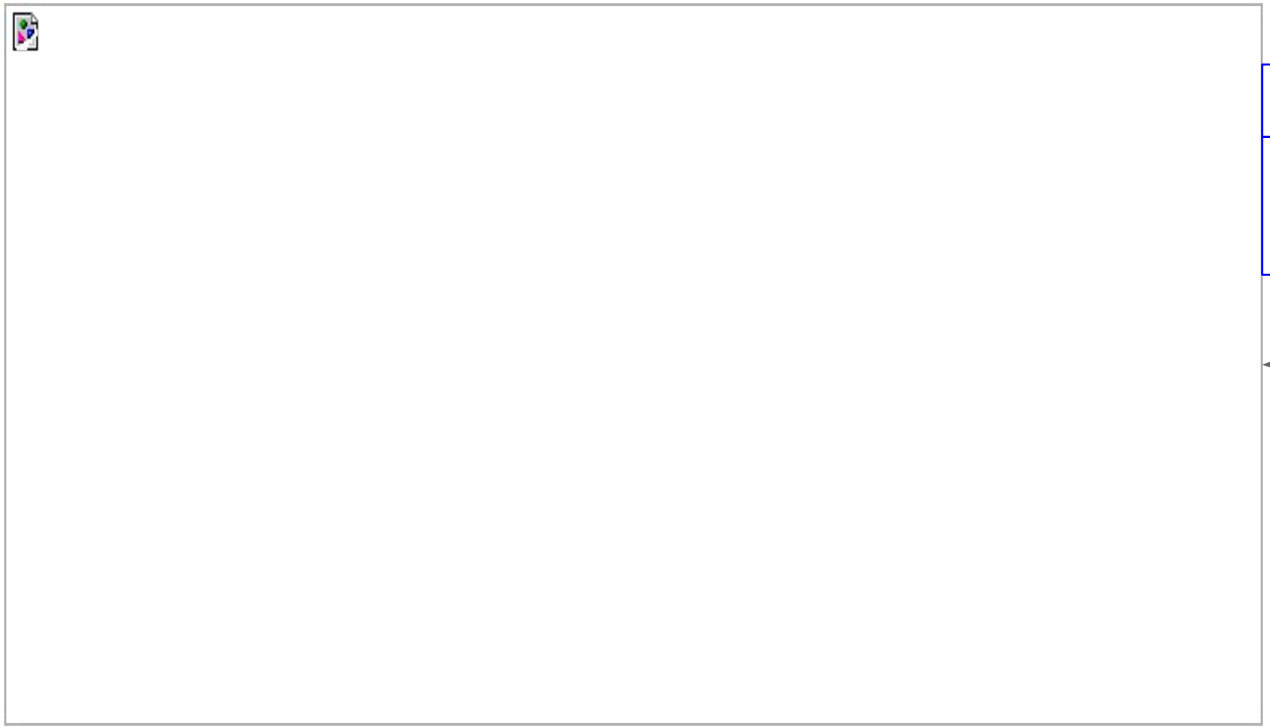
Scene: A subgraph of KB. Based on entities and events, search and rank the subgraphs.

BAYES NETWORK (ENTITIES, ISCS)

Bayes Network Construction:

1. Use Training Image Annotations (from Flickr8k) ⇒ Get objects, Scene Constituent mentions.
2. Use Training Image Scene class Detection Tuples.
   a. Lookup Scene-class-to-ISC table.
   b. Get all ISCs for the top scene.
3. These makes the tuple of observed (entities, ISCs).
4. Use the Tabu-Search **[10]** algorithm to learn the Network.

BAYES NETWORK (ENTITIES, ISCS)



Probabilistic Queries

1. P(ISC|entities,ISCs).
2. P(entity|entities,ISCs)

REASONING FRAMEWORK

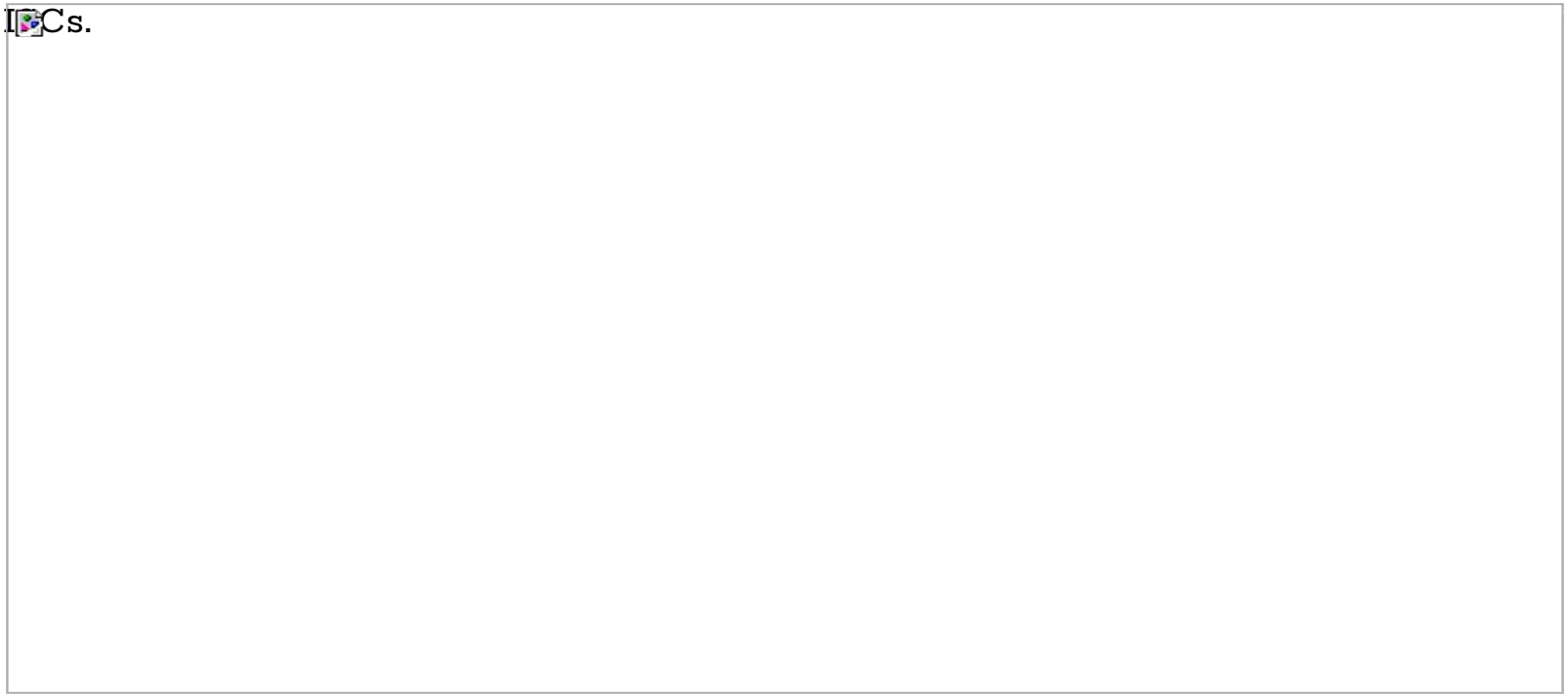
- We have: object detections, scene detections and constituent detections.
- We get:
  a. Entities: basically object classes.
  b. Entities and Events: Extract nouns and verbs from constituents (using K-parser).
     - "Person rock climb" → Entities: person, rock and Events: climb
  c. Probable ISCs:
     - For all detected scene classes,
       - Use Scene-class-to-ISC mapping table to get probable ISCs.

REASONING FRAMEWORK INPUT

Object Classes

Scene Constituents

- We have: (noisy) entities, (probable) events, (probable) ICs.



REASONING FRAMEWORK

Object Classes

Scene Constituents

Scene Classes+Mapping table (S_M)

We have: (noisy) entities, (probable) events, (probable) ISCs.

1. Entity Detection:
   a. Assumption: high-Scoring (score > α) Entities (E_H) are corre
2. Iteratively select most probable ISCs ‖ Use Bayes Net ‖
   a. Get all ISCs (C_I) from top-5 scenes detected using $\mathbf{S_M}$
   b. $C_{inf} = \Phi$
   c. Iterate while entropy decreases:
      - c1 = max over all c ϵC_I
      - Add c1 to C_{inf}. $\left[P(c|E_H, C_{inf}) \right]$

2.3. Reasoning Framework

CONTD...

3. Rectify noisy (low-scoring) objects. || Use Bayes Net ||
   a. choose the most probable sibling from WordNet Hierarchy.
   b. For each noise object:
      - Get the possible siblings.
      - For example: superclass(bathing cap) = cap, **siblings(cap) = ski cap, basketball cap** etc.....
      - $e = \text{argmax}_{o \in siblings} P(o|C_{inf}, E_H)$ and add e to E_H
4. Search Connecting Events and then Scenes || Use Knowledge Base||
   c. Get all connecting events between pairs of entities in E_H.
   d. Filter the events using rule-based techniques.
   e. Construct concept using the ISCs C_{inf}, E_H and filtered events.

After Rectifying Noisy Objects and Inferring ISCs

In this Example:

- No low-confidence objects

Inferred ISCs:

- Algorithm eliminate “bridge”, “large trees”

CONNECTING EVENTS

- For each entity-pair:
  - Perform a DFS over the knowledge-base.
  - Get the connecting event-nodes.
- For example:
  - person-climb-trunk, person-wear-trunk.
  - First trunk: tree trunk, second : swimming trunks.

FILTERING EVENTS (RULE-BASED)

Filter Noisy Events using Knowledge:

- Motivation: Several works [7] have tried to predict events based on statistical models. But they not provide an explanation as to "why" certain co-occurring event is the most suitable choice.
  - So, we wanted to "infer'' events using knowledge.
- Process:
  - Find events that connect entities (in the image) in the KB.
    - Say "wear" for *person, swimming trunks*.
  - Filter using Edge-Compatibility.
    - (Wear, agent, person); (wear,recipient,trunk) $\Rightarrow$ wear is compatible for (person,trunk).
  - Filter using superclass/semantic-role information.
    - In concepts in KB, "wear" connects to "trunk" which have semantic role "clothing".

FILTERING EVENTS (Alternative)

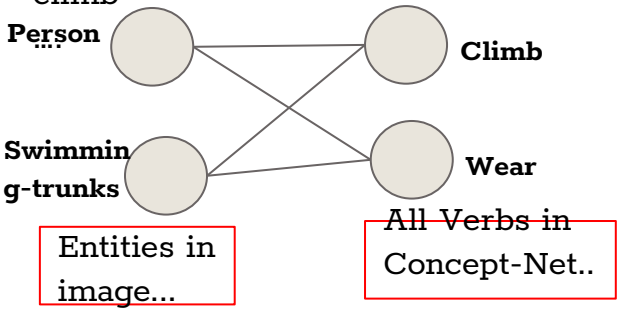
Drawbacks:

1. How many more rules do we need?
2. Superclass/semantic-role information in K-parser has errors.

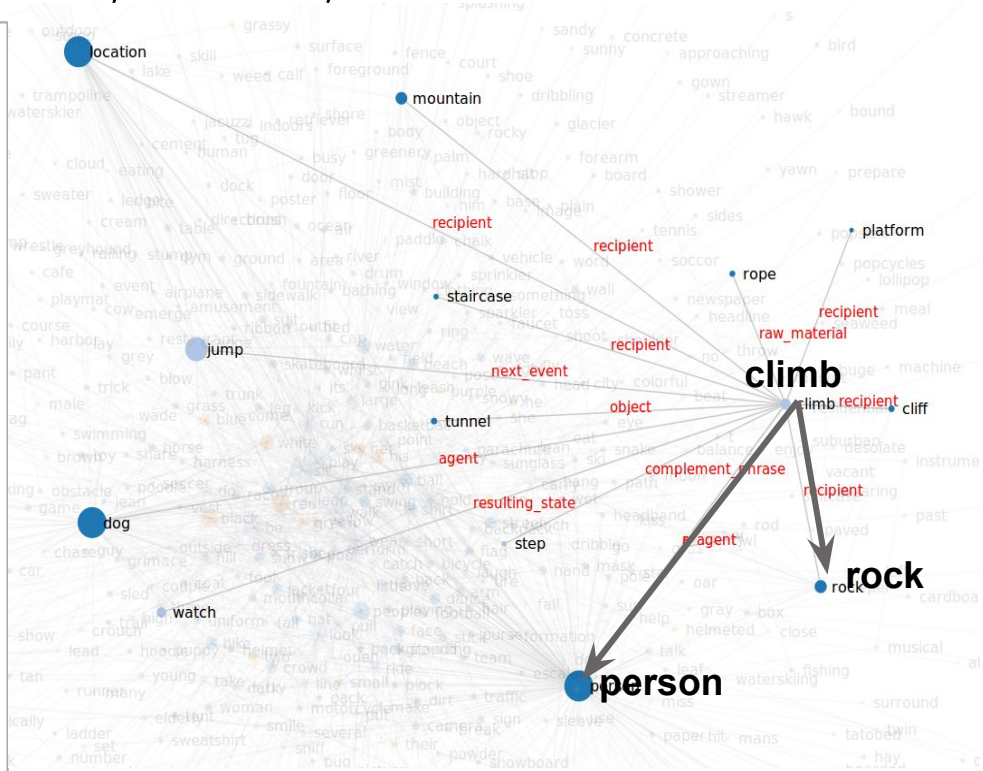
An Alternative to Rule-Based Inference of verbs:

1. We will use Concept-Net as our knowledge-base for entities and events.
2. Use a framework which combines the advantages of logical semantics and probabilistic modelling, such as Probabilistic Soft Logic:
   a. Soft rules of the form **wt: noun → verb** can be used for all nouns (in image) to all verbs (in concept-net).
   b. The weight (**wt**) can be plugged in from similarity measures like word2vec similarity.
   c. PSL represents them as Markov Random Field.
3. Similar rules can be used to infer ISCs too, eliminating the mapping tables.

W1: person → wear
W2: swimming-trunks → wear
W3: person → climb
W4: swimming-trunks → climb



Current Set of Entities, ISCs, Events



A snapshot of the KB built from 1000 sentences

SEARCH SCENES

- Search Scenes related to individual Objects.
- Filter the Scenes by other objects (if the object or its synonyms occur in the concept-graph)
- (Count-based) Weight the Scenes by events (E_H) and ISCs (C_{INF}).
  - Increase weight by 1 if an event or ISC occurs.
- (Probability-based) Weight the Scenes by objects, scene-categories and scene-constituents that are detected from the perception system.
- Sort the Scenes according to the weight

CONSTRUCT SCENES

we construct an SDG using the following set of rules:

1. Add edge(*scene*, *component*, s) for all ISC s $\in C_{inf}$
2. Add edge(event, *location*, *scene*) for the top detected events;
3. Add all compatible edges related to the set of Compatible Events such as edge(wear,agent,person) and edge(wear,recipient,trunk);
4. for all entities o_{im} in ($O_{img} \setminus O_{ev}$), do the following:
   a. If it is an animate entity, add edge(o_{im}, location, scene);
   b. Otherwise, find the shortest path from o_{im} to the top detected event in the Knowledge-base and add the edges on the path to the SDG.

CONSTRUCT SCENES

Results

SENTENCE GENERATION

Sentence Evaluation Metrics are tricky.

- Most previously used metrics (BLEU [14]) do not always agree with human evaluations.
- We used only human evaluations. Two metrics.
  - Relevance (1-5):
    - how much the description conveys the image content.
      - 1- no relevance, 2- weak relevance, 3- some relevance, 4-relates closely, 5-relates perfectly
  - Thoroughness (1-5):
    - how much of the image content is conveyed by the description
      - 1- cover nothing, 2- covers minor aspects, 3- covers some aspects, 4- covers many aspects, 5- covers almost every aspect.

SENTENCE GENERATION

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.

3. Results

IMAGE RETRIEVAL

Image-sentence alignment quality is tested using ranking experiments:

- We withhold the testing images and use the generated sentences as queries.
- We process the textual query and construct $G_q=(V_q,E_q)$, using K-Parser.
- For each image, we take the SDG $G_{img} = (V_i,E_i)$ and calculate the similarity using the following formula:

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

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

- Vertex-similarity is calculated based on their word-meaning similarity and neighbor similarity.
  - wnsim(.,.) is WordNet-Lin Similarity between two words and
  - Jaccard(.,.) is the standard Jaccard coefficient similarity.

IMAGE RETRIEVAL

	Flickr8k			
Model	**R@1**	**R@5**	**R@10**	**Med r**
BRNN-Karpathy	11.8	32.1	44.7	12.4
Our Method-SDG	**18.1**	**39.0**	**50.0**	**10.5**
	Flickr30k			
BRNN-Karpathy	15.2	37.7	50.5	9.2
Our Method-SDG	**26.5**	**48.7**	**59.4**	**6.0**
	MS-COCO			
BRNN-Karpathy (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 2 : 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.

MOTIVATING QA EXAMPLE:

Is someone drinking from the fountain?

Detections

Person
tv/monitor
bathing cap

Fountain, Plaza

outdoors
grass
person skate

SDG

has(scene,component,water).
has(scene,component,
water_droplets).
has(scene,component,
exterior_of_building).
**has(person1,semantic_role,
drinker).**
has(water,semantic_role,liquid).
has(person1,semantic_role,creator).
has(drink,recipient,water).
has(drink,agent,person1).
has(drink,origin,fountain).

```
entity(person;dog;water;shorts;frisbee).
animate(person;dog).

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

drink_yes :- animate(A),
has(drink,agent,A),
has(drink,recipient,water).
yes_fountain(A) :- drink_yes,
has(drink,agent,A), has(drink,origin,
fountain).
```

ASP Reasoning Engine

yes_fountain(person1)

MOTIVATING QA EXAMPLE:

Detections

Person, racket, popsickle, brassiere
stadium, baseball_field
outdoors, outside, sky

SDG

```
has(hold,location,scene).
has(hold,recipient,racket).
has(hold,agent,person).
has(swing,location,scene).
has(swing,recipient,racket).
has(swing,agent,person).
has(shirt,semantic_role,:clothing).
has(sweatband,semantic_role,:clothing).
has(racket,complement_phrase,tennis).
has(racket,semantic_role,:thing held).
has(person,trait,female).
has(person,semantic_role,:holder).
```

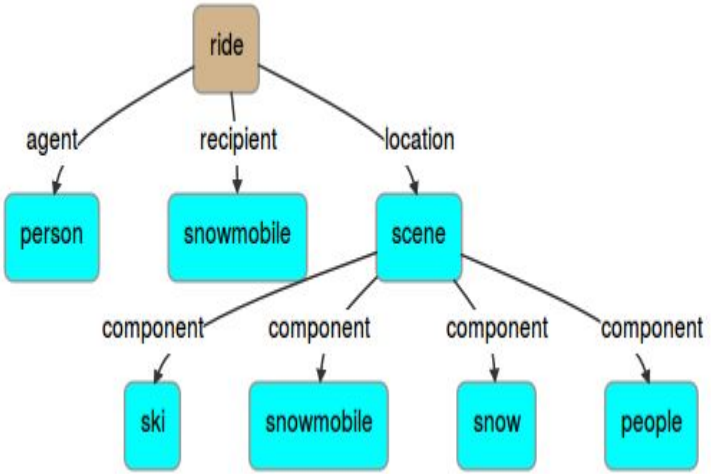
Is someone playing tennis?

```
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,tennis),h
as(swing,agent,A),animate(A).
```

ASP Reasoning Engine

tennis_detector=True

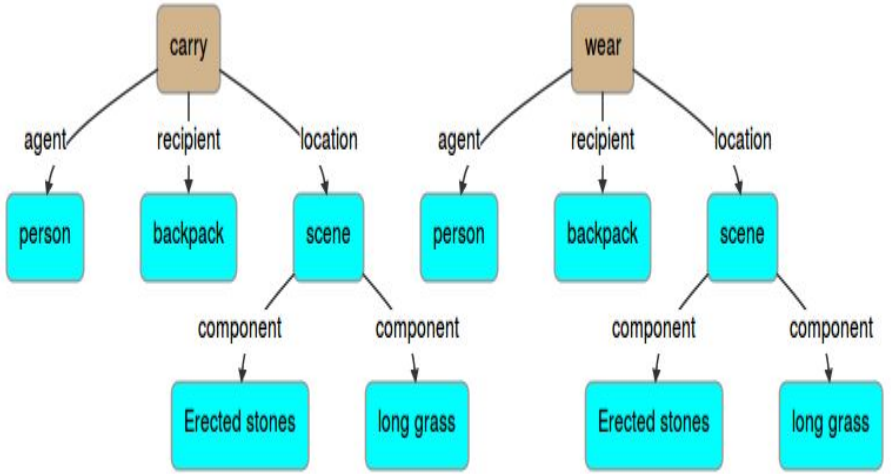
RESULTS (SDGs)



Person is riding snowmobile in the scene.

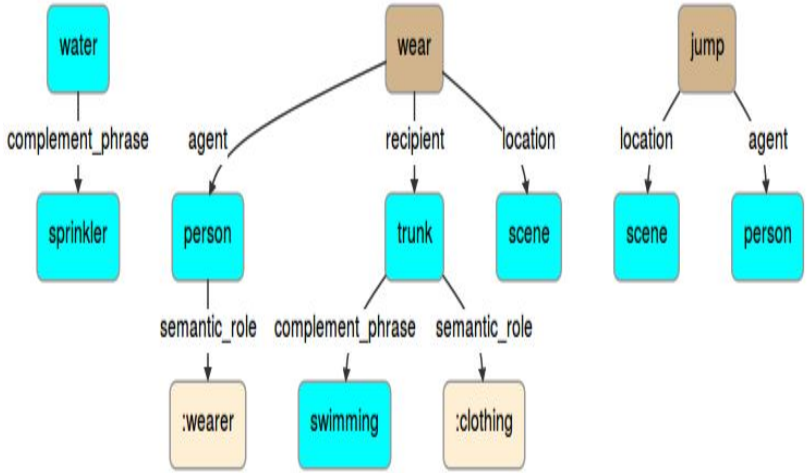
The scene contains people and ski and snow and snowmobile.

RESULTS (SDGs)



A person might be holding backpack in the scene. A person is carrying backpack in the scene. A person might be wearing backpack in the scene. The scene contains long grass and Erected stones.

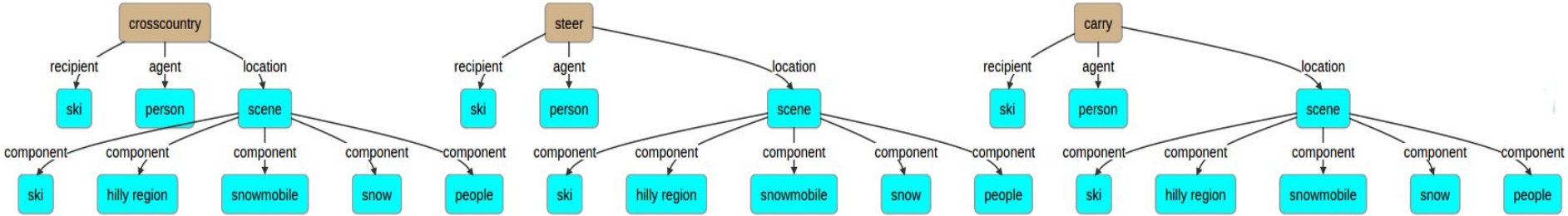
RESULTS (SDGs)



A person might be wearing swimming trunks in the scene. A person is jumping.

RESULTS (SDGs)

A person is crosscountry skiing in the scene.
A person is steering ski in the scene. A person is carrying ski in the scene.
The scene contains people and snow and ski and snowmobile and hilly region.



RESULTS (Sentences)

A person is driving car in the scene.

The scene contains street and people walk and concrete roads and booths and vehicles.

A person is surfing at ocean. The scene contains large waterbody and sand and water and ocean.

RESULTS (Sentences)

A person might be wearing sunglass in the scene.
The scene contains big size recording instruments and people wearing headphones.

A person is pulling cart in the scene.
A person is pushing person at cart.

CONTRIBUTIONS

- We proposed an architecture for image understanding where a system can answer questions (of varying difficulty) regarding the image.
- We provide a preliminary implementation, which combines state-of-the-art Deep Recognition with NLP techniques for knowledge acquisition.
- To solve the challenge of Knowledge Representation, we propose a novel representation of an image, called the Scene Description Graph (SDG), which combines visual data with background knowledge.
- Our primary implementation achieves comparable results with a recent Deep End-to-End Neural Approach.
- We provide some preliminary examples of how question-answering using the SDG can be achieved.

REFERENCES I

[1] Bloom BS. Taxonomy of educational objectives. Vol. 1: Cognitive domain. New York: McKay. 1956:20-4.

[2] Aditya S, Yang Y, Baral C, Fermuller C, Aloimonos Y. Visual common-sense for scene understanding using perception, semantic parsing and reasoning. In2015 AAAI Spring Symposium Series 2015 Mar 12.

[3] Gao H, Mao J, Zhou J, Huang Z, Wang L, Xu W. Are you talking to a machine? dataset and methods for multilingual image question answering. arXiv preprint arXiv:1505.05612. 2015 May 21.

[4] Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Lawrence Zitnick C, Parikh D. VQA: Visual question answering. InProceedings of the IEEE International Conference on Computer Vision 2015 (pp. 2425-2433).

[5] Karpathy A, Fei-Fei L. Deep visual-semantic alignments for generating image descriptions. InProceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2015 (pp. 3128-3137).

[6] Ordonez V, Kulkarni G, Berg TL. Im2text: Describing images using 1 million captioned photographs. In Advances in Neural Information Processing Systems 2011 (pp. 1143-1151).

[7] Yang, Y., Teo, C.L., Daumé III, H. and Aloimonos, Y., 2011, July. Corpus-guided sentence generation of natural images. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 444-454). Association for Computational Linguistics.

[8] Elliott, D. and Keller, F., 2013, October. Image Description using Visual Dependency Representations. In *EMNLP* (Vol. 13, pp. 1292-1302).

REFERENCES II

[9] Johnson, J., Krishna, R., Stark, M., Li, L.J., Shamma, D.A., Bernstein, M.S. and Fei-Fei, L., 2015, June. Image retrieval using scene graphs. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on* (pp. 3668-3678). IEEE.

[10] Glover, F., Kelly, J.P. and Laguna, M., 1995. Genetic algorithms and tabu search: hybrids for optimization. *Computers & Operations Research*, *22*(1), pp.111-134.

[11] Girshick, R., Donahue, J., Darrell, T. and Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In*Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580-587).

[12] Zhou, B., Lapedriza, A., Xiao, J., Torralba, A. and Oliva, A., 2014. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems* (pp. 487-495).

[13]Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

[14] Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E. and Darrell, T., 2013. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*.

[15] Papineni, K., Roukos, S., Ward, T. and Zhu, W.J., 2002, July. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311-318). Association for Computational Linguistics.