VisKE: Visual Knowledge Extraction and Question Answering by Visual Verification of Relation Phrases

FERESHTEH SADEGHI,
SANTOSH K DIVVALA,
ALI FARHADI
UNIVERSITY OF WASHINGTON
THE ALLEN INSTITUTE FOR AI

PRESENTED BY JAVIER LORES
(JLORES@KNIGHTS.UCF.EDU)

Outline

- Motivation
- Problem Statement
- Main Contributions
- Approach
- Implementation
- Results

Do Dogs Eat Ice Cream?

Image credits: https://www.pinterest.com/pin/200621358375219357/

Motivation

- Primary modality for learning and reason
  - Vision
  - Text
- Relation questions
  - Do horses eat hay?
  - Do butterflies flap wings?

Problem Statement

- How to estimate the confidence of a mentions-relation predicate by reasoning with images
  - Focusing on verb-based relations between common nouns

Main Contributions

- Presents an unsupervised approach for verifying relationships by analyzing the spatial consistency of the relative configurations of the entities and the relation involved
- Verified 12,000 relation phrases and doubled the size of the ConceptNet knowledge base at a precision of 0.85

Approach-Overview

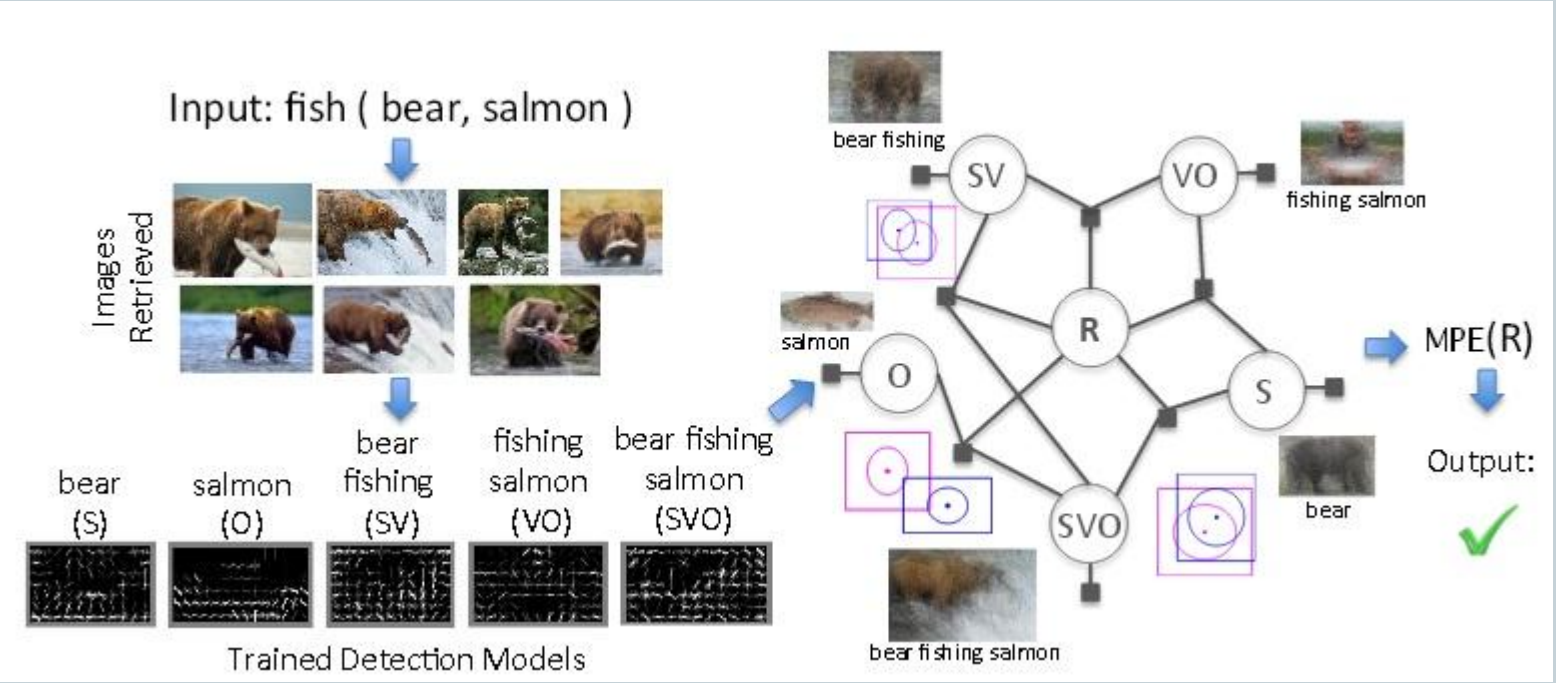

Image credit: *[1]*

Approach-Relation Model

- How to model visual composites?
  - e.g. A person riding a horse
- Independently?
  - Detect person
  - Detect horse
  - Describe relation
- Problem
  - Appearances change with relation

Person Riding a Horse

Image credit: horse, person, person riding horse

Approach-Relation Model

- Relation
  - Verb (Subject, Object)
- Meaningful patterns (R)
  - (SVO, SV)
  - (S, VO)
  - (VO, SV)
  - (SVO, O)
  - (SVO, S)
  - V, SO variables
    - Ignored due to visual ambiguity
  - Patterns required to have V, S, and O

Problem Formulation

$$P(\mathcal{R},\mathcal{S},\mathcal{O},\mathcal{SV},\mathcal{VO},\mathcal{SVO}) \propto \prod_{x\in\{\mathcal{O},\mathcal{S},\mathcal{SV}\}} \Phi(\mathcal{R},\mathcal{SVO},x) * \prod_{y\in\{\mathcal{SV},\mathcal{S}\}} \Phi(\mathcal{R},\mathcal{VO},y) * \prod_{z\in\{\mathcal{S},\mathcal{O},\mathcal{SV},\mathcal{VO},\mathcal{SVO}\}} \Psi(z), \quad (1)$$

$$\Phi^i(\mathcal{R},x,y) - \begin{cases} \max_\theta \mathcal{L}(x,y,\bar{I};\theta) & \mathcal{R}=i \\ 1 & \text{otherwise} \end{cases} \quad (2)$$

Image credit: *[1]*

Feature Representation

- Translation between detections
- Intersection over union of two detection boxes
- Ratio of intersection over area of bounding box x
- Ratio of intersection over area of bounding box y
- Height and width of bounding box x
- Height and width of bounding box y
- Unary potential
  - Height of bounding box
  - Width of bounding box
  - (x, y) mid-point of bounding box

Feature Representation

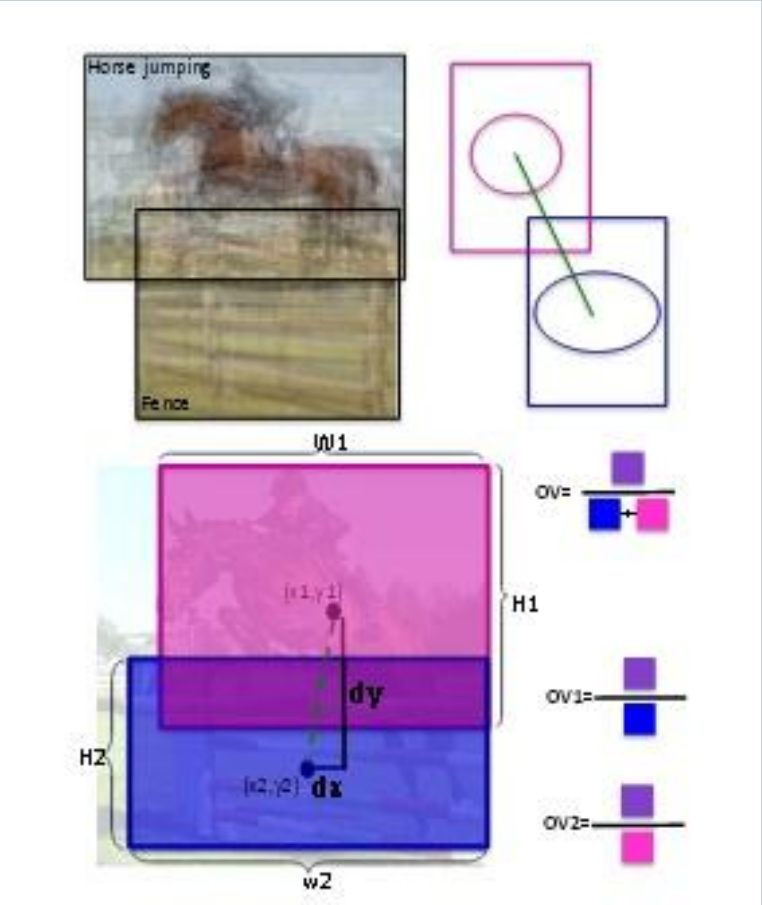

Image credit: *[1]*

Learning Model Detectors-Webly Learning

- Webly-unsupervised learning
  - Leverages image-search engines
- System
  - Feed n-gram
  - Image-search
  - Prune images
  - Train Deformable Part Model (DPM)
    - Various parts of the image are used separately to determine if object of interests exists

Relation Phrase Dataset

- Google Books Ngram (English 2012)
  - <noun, verb, noun>
- Base Set
  - 6093 relations
    - 45 subjects
    - 1158 verbs
    - 1839 objects
- Permute Set (Permuted S,V,O from Base Set)
  - 6500 relations

Results

- Mean Average Precision

	Base Set	Permute Set	Combined Set
Visual Phrase [32]	49.67	14.12	42.49
Co-detection Model	49.24	14.65	43.14
Google Ngram Model [1]	46.17	NA	NA
Language Model [22]	56.20	**22.68**	50.23
VisKE	**62.11**	20.93	**54.67**

Image credit: *[1]*

Results Per Subject

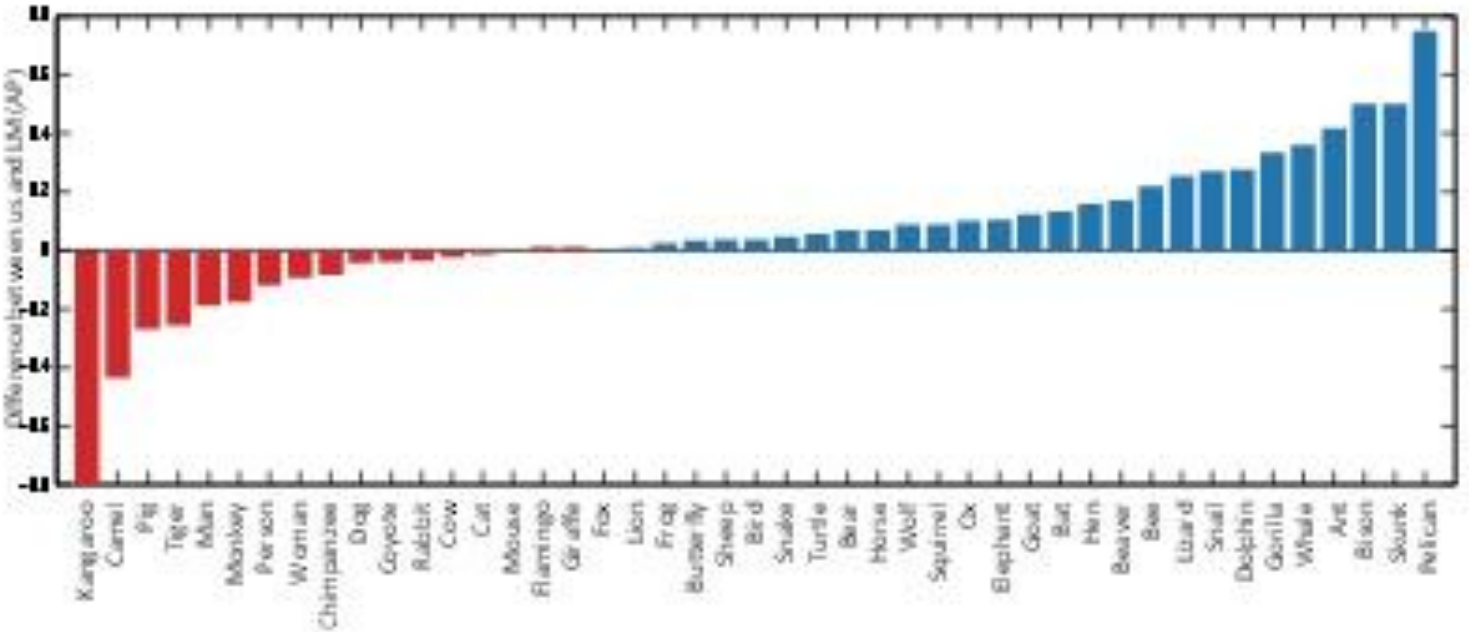

Image credit: *[1]*

Relation Examples

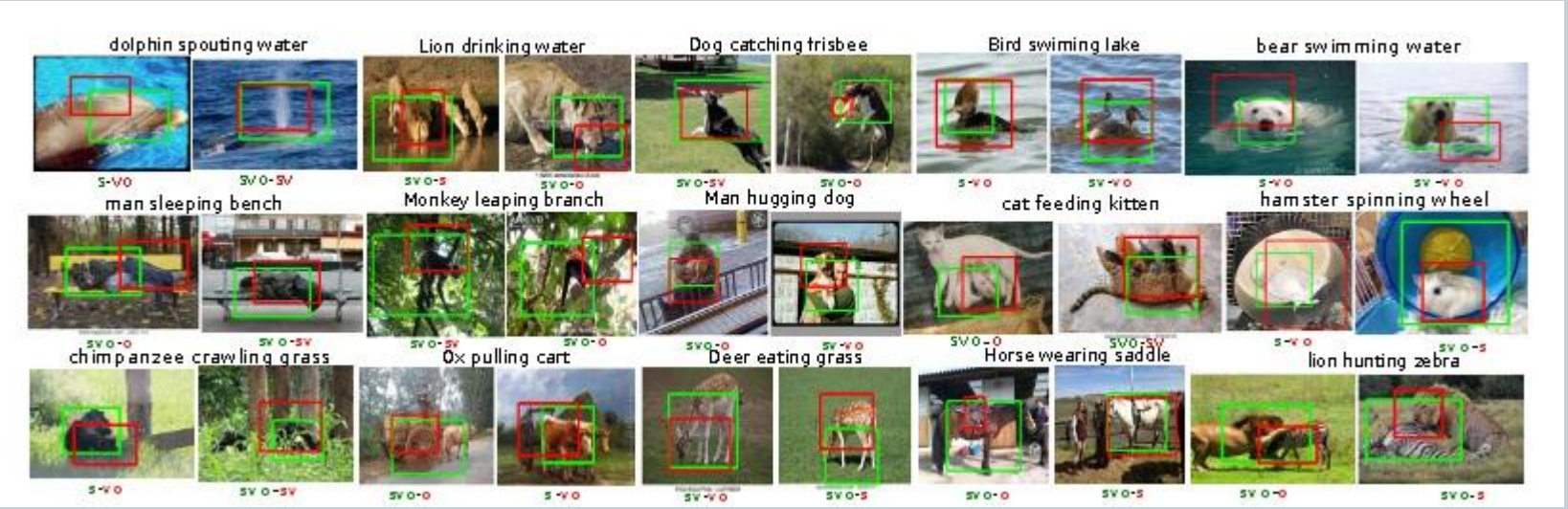

Image credit: *[1]*

Ablation Analysis

Model	M.A.P.
VisKE (All Factors)	62.11
Without $\Phi(\mathcal{R}, \mathcal{VO}, \mathcal{SV})$	60.41
Without $\Phi(\mathcal{R}, \mathcal{VO}, \mathcal{S})$	61.16
Without $\Phi(\mathcal{R}, \mathcal{SVO}, \mathcal{S})$	60.40
Without $\Phi(\mathcal{R}, \mathcal{SVO}, \mathcal{O})$	59.55
Without $\Phi(\mathcal{R}, \mathcal{SVO}, \mathcal{SV})$	59.55
Without binary terms	60.61
Without unary terms	58.52
CRF	58.01

Image credit: *[1]*

Application: Enriching Knowledge Bases

- ConceptNet
  - A Semantic network containing common sense knowledge
  - Contains very little relational facts
    - 300 across 45 concepts
  - Added relationships for
    - IsA
    - PartOf
    - HasA
    - MemberOf
    - CapableOf

ConceptNet Addition Results

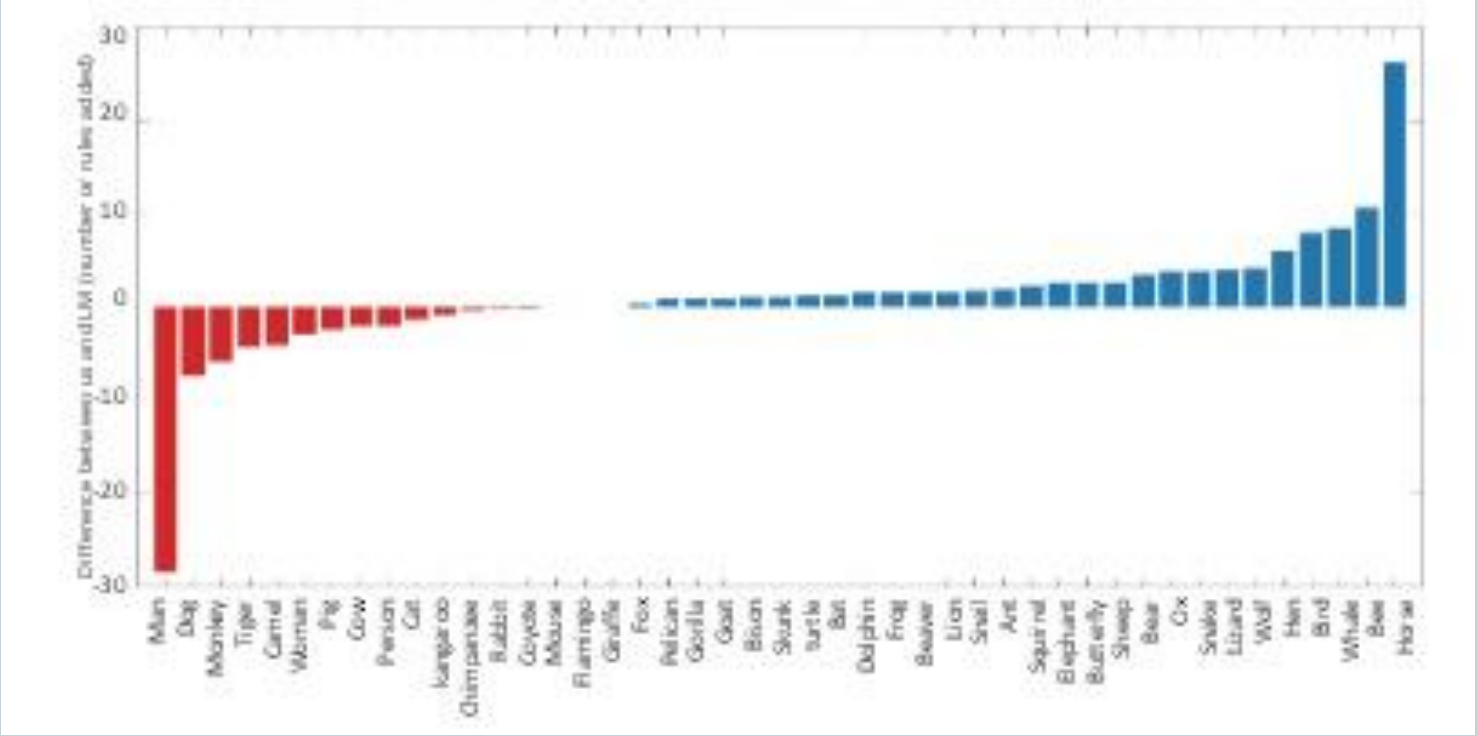

Image credit: *[1]*

Question Answer Reasoning

- What do lions devour?
  - Verb(subject, ?)
    - E.g. prey
- What eats ice cream?
  - Verb(?, object)
    - E.g. dog
- What do men do with sofa?
  - ?(subject, object)
    - E.g. sit

Question Answer Reasoning

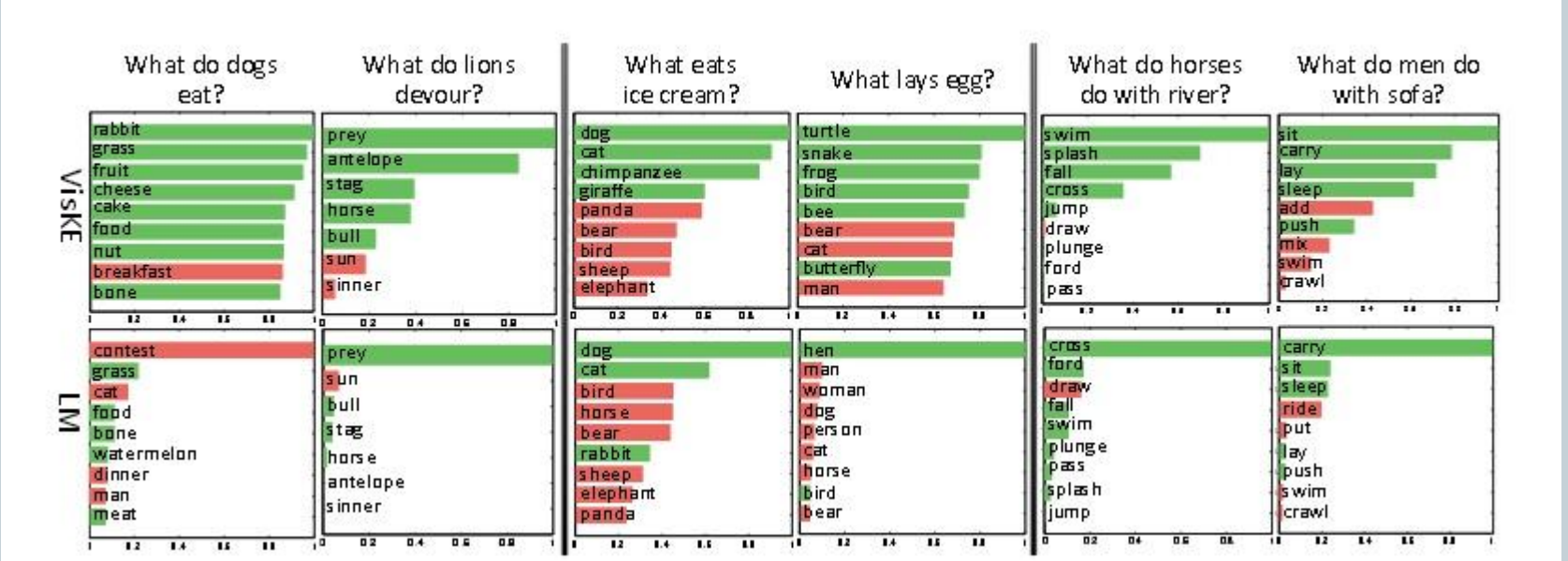

Image credit: *[1]*

Question Answer Reasoning

- Multiple Choice
  - Elementary-level general science questions
    - What part of a plant produces seeds?
      - (a) Flower
      - (b) Leaves
      - (c) Stem
      - (d) Roots
- System accuracy
  - VisKE: 85.7%
  - Text Based Reasoning: 71.4%

References

[1] “VisKE: Visual Knowledge Extraction and Question Answering by Visual Verification of Relation Phrases”, Fereshteh Sadeghi, Santosh K Divvala, Ali Farhadi, CVPR, 2015