IIIT, HYDERABAD

Choosing Linguistics over Vision to Describe Images

?

An ocean boat is travelling in narrow water

INPUT

OUTPUT

Ankush Gupta, Yashaswi Verma, C.V. Jawahar
IIIT Hyderabad, India

What is present in an Image?

Keywords : { boeing 747, airport, parked }

Phrase : “boeing 747 at the airport”

Sentence : “A boeing 747 is parked at the airport.”

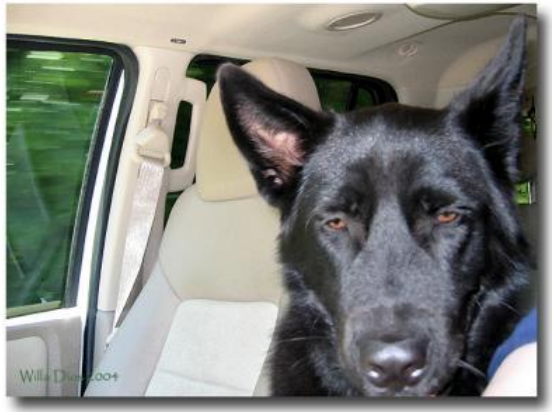
Keyword-based Annotation

IIIT, HYDERABAD

Advantages

Useful in Image Indexing and Retrieval

Limitations



{ black, dog, car }

"A **black dog** is sitting inside a car."

{ black, dog, car }

"A dog is sitting inside a **black car**."

Problem Statement

IIIT, HYDERABAD

TRAINING DATASET

A cyclist relaxes on a bench and gazes toward the ocean.

A man on a bicycle with a racing suit.

A silver bicycle is parked in a living room.

Two bicyclists pass spectators on the road near a field.

UNSEEN IMAGE

?

Key Challenges

IIIT, HYDERABAD

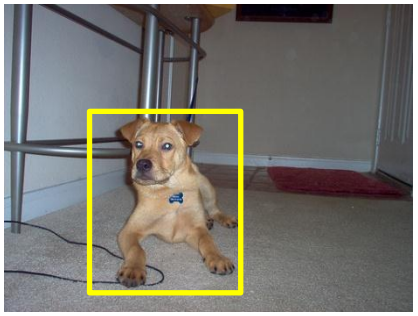
- Detecting Image content (objects, attributes, scene, action) is tough
- How they interact with each other?
- Determining correct order of words

This is a photograph of one person and one sky. The white person is by the blue sky. (Kulkarni et al. CVPR 2011)

Previous Approaches

IIIT, HYDERABAD

- Use trained object detectors and scene/attribute classifiers



Object : Dog
Attribute : Brown
Scene : Room

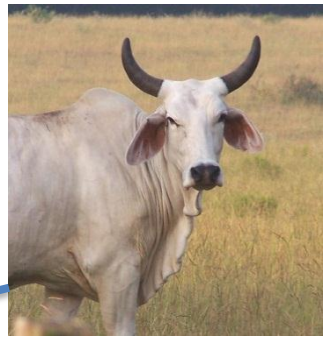
- Use corpus statistics to smooth noisy vision predictions
- Predict associated verb/preposition
- Sentence generation
  - Language Model based
  - Template based

Kulkarni et al. (CVPR 2011), Yang et al. (EMNLP 2011), Li et al. (CoNLL 2011), Ordonez et al. (NIPS 2011), Yao et al. (IEEE 2008), Aker et al. (ACL 2010), Farhadi et al. (ECCV 2010), Feng et al. (ACL 2010), Sadeghi et al. (CVPR 2011), Kuznetsova et al. (ACL 2012), Mitchell et al. (EACL 2012)

Hypothesis

- Image inherits characteristics of similar images

(white, cow), (cow, with, ear), (cow, in, field), (grassy, field)



(white, cow), (bull, standing), (bull, in, field), (grassy, field)

TEST IMAGE

(brown, cow), (white, cow), (young, cow), (cow, in, field), (grassy, field), (cow, with, ear)

(brown, cow), (young, calf), (cow, on, grass), (cow, in, field)

(brown, cow), (cow on farm), (cow, with, tag), (tag, in, ear)

Overview of our Approach

IIIT, HYDERABAD

Input Image

Neighbouring Images with Extracted Phrases

(boat),
(ocean, boat),
(boat, travel),
(travel, in, water)

(boat),
(beach),
(narrow, channel),
(come, into, port),

(boat),
(shore),
(yellow, boat),
(boat, on, beach)

Ranked List of Predicted Triples

((ocean, boat), travel), (travel, in, (narrow, water)), ((ocean, boat), in, (narrow, water))
((ocean, boat), beach), (beach, on, (ocean, shore)), ((ocean, boat), on, (ocean, shore))
((ocean, boat), dock), (dock, at, (large, port)), ((ocean, boat), at, (large, port))

An ocean boat is travelling in narrow water.

Phrase Extraction

IIIT, HYDERABAD

- 9 distinct types of phrases are extracted from training image descriptions
  - E.g. *(subject)*, *(attribute, subject)*, *(subject, verb)*, *(verb, prep, object)*, *(attribute, object)* etc.
- Each noun is expanded up to at most 3 hyponym levels using WordNet

Sentence	Synonym	A black and white pug is looking at the camera.
Phrases	No	(pug$_{(subj)}$), (black$_{(attr)}$, pug$_{(subj)}$), (white$_{(attr)}$, pug$_{(subj)}$), (pug$_{(subj)}$, look$_{(verb)}$), (camera$_{(obj)}$), (look$_{(verb)}$,at$_{(prep)}$,camera$_{(obj)}$)
Phrases	Yes	(dog$_{(subj)}$), (black$_{(attr)}$, dog$_{(subj)}$), (white$_{(attr)}$, dog$_{(subj)}$), (dog$_{(subj)}$, look$_{(verb)}$), (camera$_{(obj)}$), (look$_{(verb)}$,at$_{(prep)}$,camera$_{(obj)}$)

Step-1 : Predicting Phrase Relevance

IIIT, HYDERABAD

- Given an unseen image *I*, its *k* most similar images (τ^k) are computed based on weighted distance between image features
  - Color, Texture, Scene (GIST) and Shape (SIFT) Descriptors

- Joint Probability of associating phrase y_i with test image *I* is determined based on
  - Presence / Absence of y_i in neighboring image $J \in \tau^k$
  - Corpus count of y_i
  - Feature similarity between image *I* and *J*

Step-2 : Parameter Learning

IIIT, HYDERABAD

- Two types of parameters in phrase prediction model
  - For Optimal combination of different image features
  - For smoothing of phrase presence/absence probability using corpus
- Error function is defined such that
  - Probability of predicting any phrase not present in image *I* should be minimized
  - Probability of predicting any phrase present in image *I* should be greater than any other phrase
- Use Gradient Descent method to estimate parameters

Step-3 : Phrase Integration

($object_1$)	($attribute_1$, $object_1$)	($object_1$, verb)	(verb, prep, $object_2$)	($object_2$)	($attribute_2$, $object_2$)	($object_1$, prep, $object_2$)
boat	ocean ship	boat come	travel in water	port	mountain village	boat at port
ocean	ocean boat	boat travel	sit on water	sea	tropical beach	ship near coast
channel	lone boat	ship position	come into port	water	ocean shore	boat in water
riverboat	canal boat	boat beach	sail in water	ocean	narrow water	boat in city
ship	vintage ship	boat sail	park at harbor	coast	dirty shore	boat in river

((ocean, boat), travel), (travel, in, (narrow, water)), ((ocean, boat), in, (narrow, water))

t = { ((attr$_1$, obj$_1$), verb), (verb, prep, (attr$_2$, obj$_2$)), (obj$_1$, prep, obj$_2$) }

Step-4 : Sentence Generation

IIIT, HYDERABAD

- Natural Language Generation
  - Content Selection
    - Kulkarni et al. CVPR 2011 : <person, under, road>
    - Our Approach : <person, on, road>
  - Aggregation
    - *Syntactic Aggregation* : E.g. “A dark-skinned person is climbing with a pick-axis and posing with a green chile.”
  - Surface Realization
    - SimpleNLG (Gatt and Reiter, ENLG 2009)

Data sets

1. **UIUC PASCAL Sentence dataset**

- 1,000 images

A girl riding a brown horse.

A girl wearing a red blouse riding a brown horse.

A woman riding a brown horse.

A young girl riding a brown horse.

A young girl wearing a helmet riding a pony.

---

2. **IAPR TC-12 Benchmark**

- 20,000 images

A grey and brown terraced house with a large, ornate entrance and two cars that are parked in front of it.

Qualitative Results (PASCAL)

IIIT, HYDERABAD

A black ferrari is parked in front of a green tree.

A sporty car is parked on a concrete driveway.

An adult hound is laying on an orange couch.

A sweet cat is curling on a pink blanket.

A blond woman is posing with an elvis impersonator.

An orange fixture is hanging in a messy kitchen.

Qualitative Results (IAPR TC-12)

IIIT, HYDERABAD

In this image, a dark-skinned person is climbing with a pick-axis and posing with a green chile. A green garden is surrounded with a striped chair.

In this image, a child and a woman are sitting at an orange table. A child is sitting with tangerine.

In this image, a king-size bed is made with a white bedcover. A black cabinet is made with a brown telephone and standing on left.

Application-1 (Multiple Interesting Descriptions)

IIIT, HYDERABAD

- A white toddler is playing with a condiment package.
- A wide-eyed baby is holding in a highchair.
- A female child is smiling in a floral shirt.

- A panama railway is sitting on an off-track.
- An idle train is heading toward a double-decker station.
- A young teenager is dancing near a long bus.

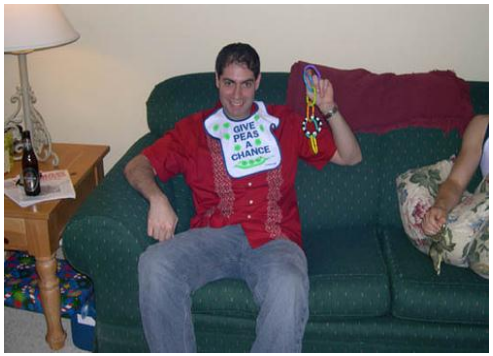
- A black lamb is standing in a dry bush.
- A shaggy sheep is walking through a wild grass.
- A gray wolf is standing in a wild grass.

Application-2 (Phrase Annotation)

IIIT, HYDERABAD

“aeroplane at airport”

“sit on sofa”



“graffiti-covered bus”

Evaluation Measures

IIIT, HYDERABAD

- Automatic Evaluation
  - BLEU score
  - Rouge score
- Human Evaluation
  - Readability
  - Relevance

Quantitative Results(Automatic Evaluation)

DATASET	SYNONYM	BLEU-1	BLEU-2	BLEU-3	ROUGE-1
PASCAL	Yes	0.41	0.11	0.02	0.28
PASCAL	No	0.36	0.09	0.01	0.21
PASCAL Human (std.)	-	0.64	0.42	0.24	0.50
IAPR TC-12	Yes	0.21	0.07	0.01	0.14
IAPR TC-12	No	0.15	0.06	0.01	0.11

Higher Score means Better Performance

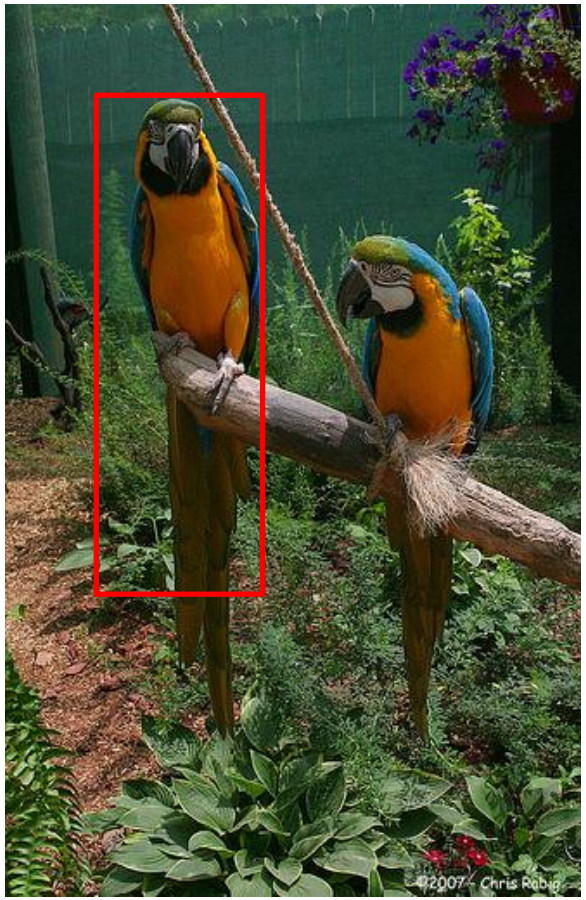
Quantitative Results (Human Evaluation)

- Likert scale of {1,2,3} where 1 is good, 2 is ok and 3 is bad

DATASET	SYNONYM	READABILITY	RELEVANCE
PASCAL	Yes	1.19	1.57
PASCAL	No	1.24	1.76
IAPR TC-12	Yes	1.38	2.32
IAPR TC-12	No	1.41	2.55

Lower Score means Better Performance

Qualitative Results (Comparison)



Kulkarni et al. 2011 : “This is a picture of one tree, one road and one person. The rusty tree is under the red road. The colorful person is near the rusty tree, and under the road.”

Yang et al. 2011 : “The person is showing the bird on the street.”

Li et al. 2011 : “Black women hanging from a black tree. Colored man in the tree.”

Ours : **“An American eagle is perching on a thick rope.”**

Quantitative Results (Comparison)

IIIT, HYDERABAD

APPROACH	BLEU-1	BLEU-2	BLEU-3	ROUGE-1
Baby Talk (Kulkarni et al. 2011)	0.30	-	-	-
Ours	**0.47**	0.19	0.06	0.33
Corpus Guided (Yang et al. 2011)	0.41	0.13	0.03	0.31
Ours	**0.54**	**0.23**	**0.07**	**0.41**

Higher Score means Better Performance

Advantages over Previous Approaches

IIIT, HYDERABAD

- Use of available image descriptions
- Unlimited Vocabulary
- Grammatically correct sentences
- Scalable
  - No use of trained detectors/classifiers
  - Performance likely to improve as the collection of available annotated images grows
  - Domain-Independent sentence generation approach

Limitations

IIIT, HYDERABAD

- Requires a collection of images with their corresponding descriptions

- Needs sufficient number of images of different categories
  - Trained detectors/classifiers might work well for categories with very low frequency in a dataset

Conclusion and Future Work

IIIT, HYDERABAD

- A generic method which simultaneously benefits from all 3 diverse sources of information
  - Visual clues
  - Corpus statistics
  - Available descriptions

- Superior performance on the UIUC PASCAL sentence data set and baseline for IAPR TC-12 benchmark.

- Future Work : Incorporating more visual clues from detectors and classifiers

Acknowledgment

- Prashanth Mannem from LTRC, IIIT Hyderabad for helpful discussions and pointers

- Thanks to Indian Association for Research in Computing Science (IARCS) and Microsoft Research India for travel support

THANK YOU