Understanding and Predicting Importance in Images

Alexander C. Berg, Tamara L. Berg, Hal Daume III´, Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Aneesh Sood, Karl Stratos, Kota Yamaguchi

Presented by: Niveda Krishnamoorthy

CS395T Visual Recognition

Oct 26th 2012

Agenda

- Problem Overview
- Motivation
- Approach
- Experiments
- Discussion

- Problem Overview
- Motivation
- Approach
- Experiments
- Discussion

Problem Overview

- Predicting the importance of visual content of images in natural language sentences using
  - Compositional factors
  - Semantic factors
  - Contextual factors

- Problem Overview
- Motivation
- Approach
- Experiments
- Discussion

Why is this important?

"A raft with 3 adults and two children in a river."
"Four people in a canoe paddling in a river lined with cliffs."
"Several people in a canoe in the river."

Applications

- Image/Video search
- Generating human-like descriptions for images and videos

- Problem Overview
- Motivation
- Approach
- Experiments
- Discussion

Approach

1. Gathering data, content labels and descriptions
2. Mapping from content to description
3. Exploring importance factors
4. Building and evaluating models to predict importance

Data

ImageCLEF Dataset

- Collection of 20K images covering various aspects of contemporary life, such as sports, cities, animals, people, and landscapes.
- IAPR TC-12 Benchmark includes a free-text description for each image.
- Each image is also segmented into constituent objects and labeled according to a set of (275) labels

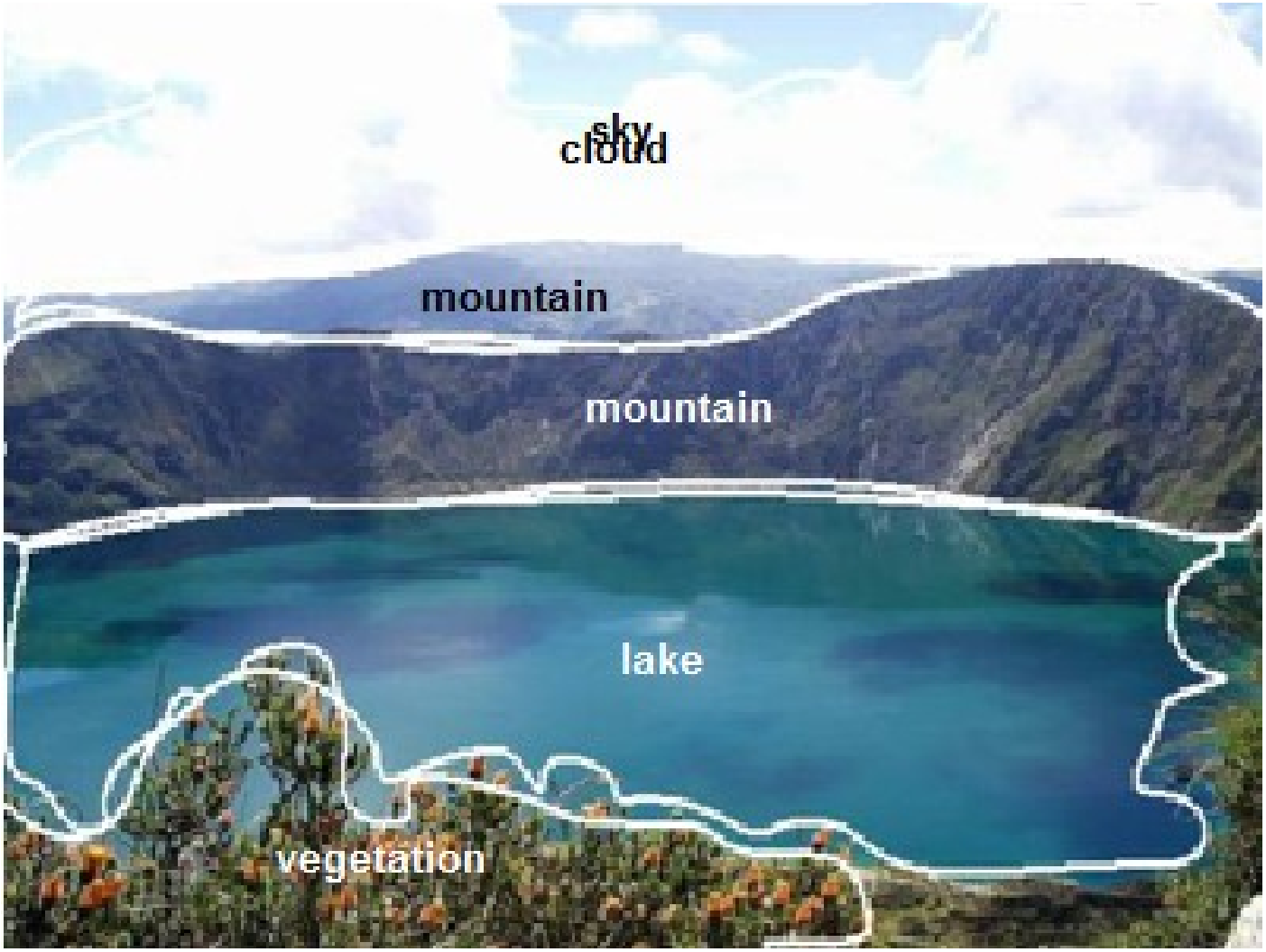
UIUC Pascal Sentence Dataset

- Consists of 1K images sub-sampled from the Pascal Challenge
- 5 descriptions written by humans for each image
- Annotated with bounding box localizations for 20 object categories

UIUC example

- One jet lands at an airport while another takes off next to it.
- Two airplanes parked in an airport.
- Two jets taxi past each other.
- Two parked jet airplanes facing opposite directions.
- Two passenger planes on a grassy plain

ImageCLEF example

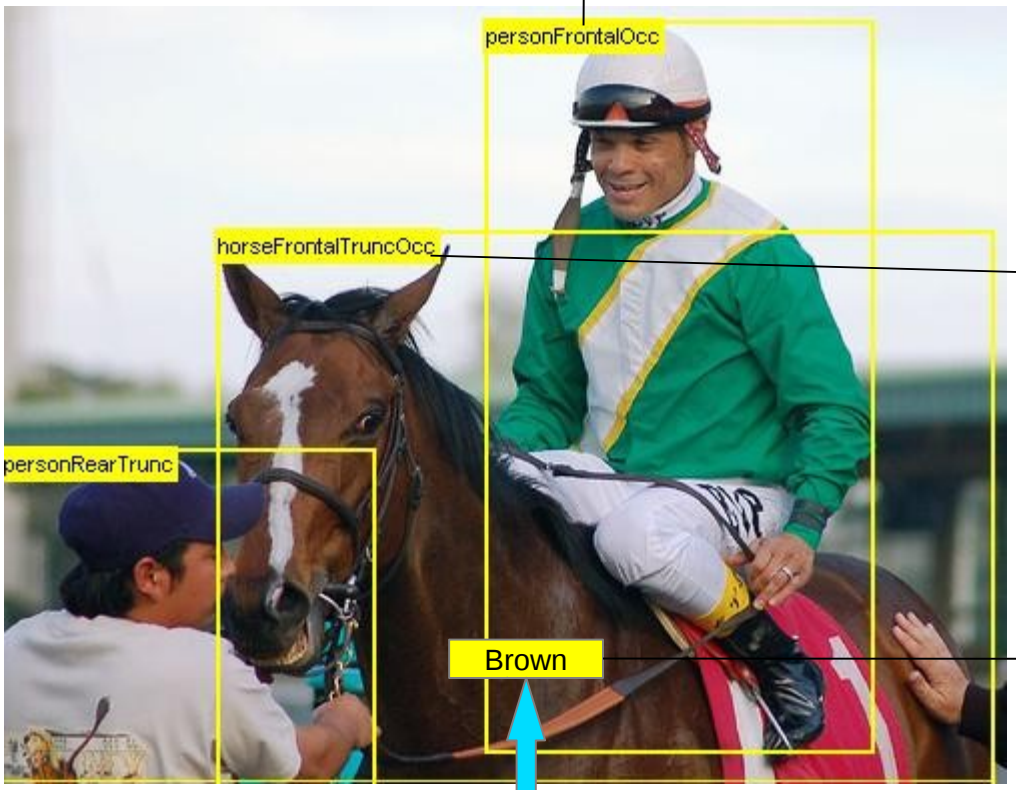


Discussion

- The ImageCLEF dataset contains images with only one sentence description per image while UIUC Pascal dataset has 5.

Is it fair to use only a single annotation to represent what most humans perceive as important?

Collecting Content Labels and Mapping them to Descriptions



Human description:
A man is riding a brown horse

Amazon Mechanical Turk

Discussion – Handling complex cases

Objects: man, woman, white, dress, church,...

Human annotation: Royal wedding

Exploring Importance Factors

Compositional factors:

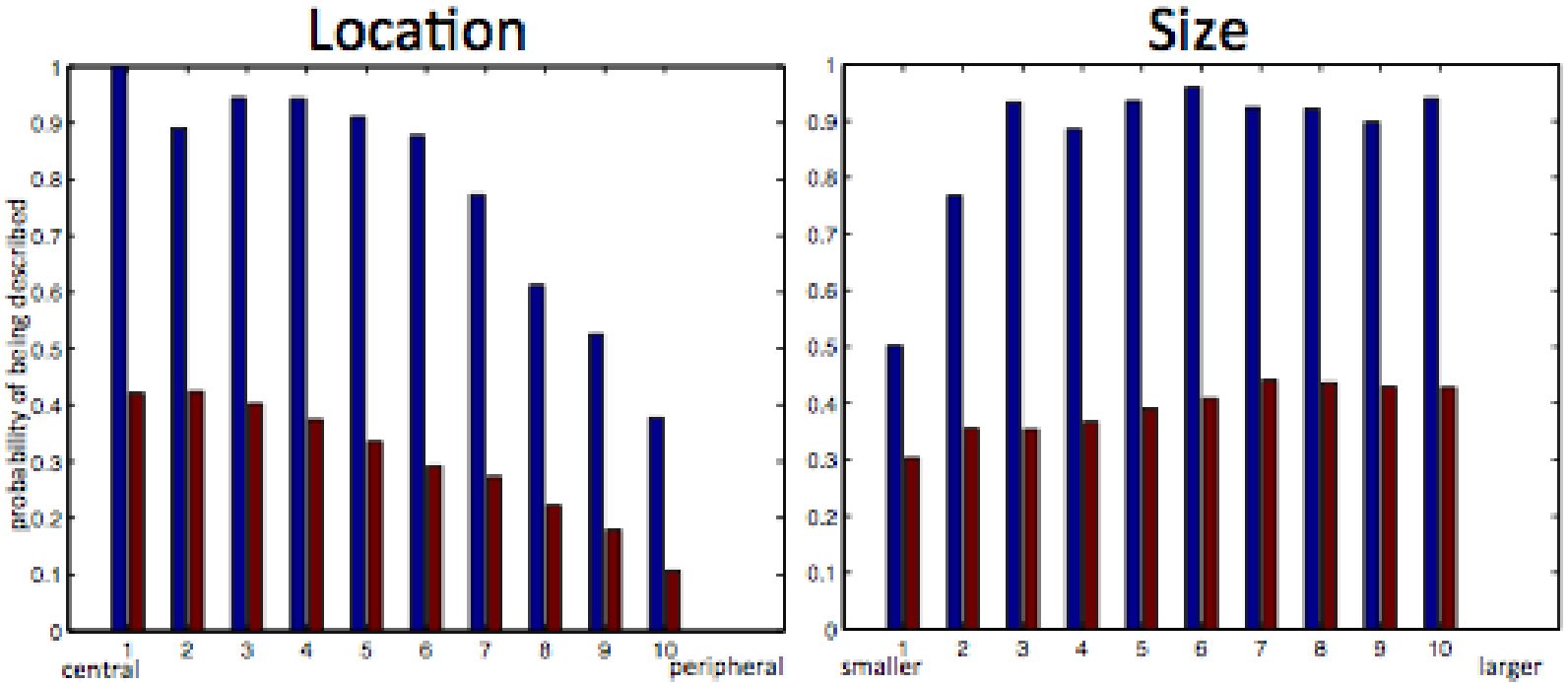
Size

"A sail boat on the ocean."

Location

"Two men standing on beach."

Influence of Compositional Factors



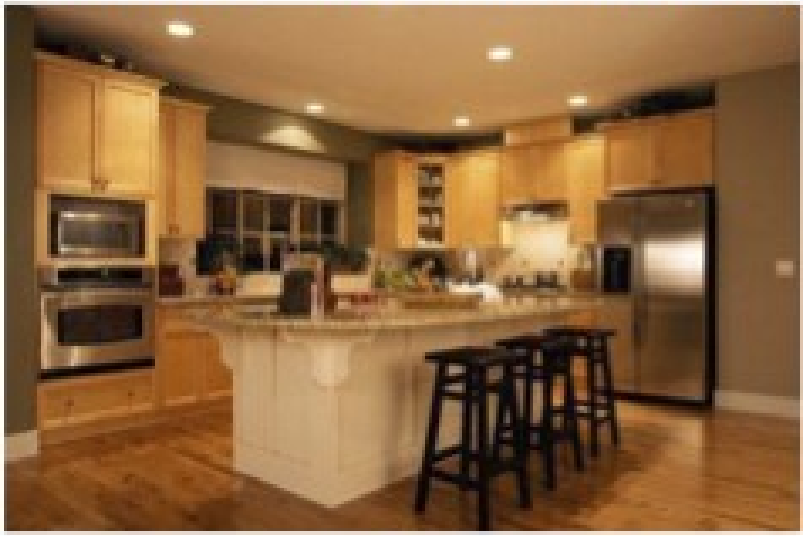
Exploring Importance Factors

Semantic factors:

Object Type

"Girl in the street"

Scene Type & Depiction Strength



"kitchen in house"

Influence of Semantic Factors (Object)

Top10	Prob	Last10	Prob
firework	1.00	hand	0.15
turtle	0.97	cloth	0.15
horse	0.97	paper	0.13
pool	0.94	umbrella	0.13
airplane	0.94	grass	0.13
bed	0.92	sidewalk	0.11
person	0.92	tire	0.11
whale	0.91	smoke	0.09
fountain	0.89	instrument	0.07
flag	0.88	fabric	0.07

ImageCLEF

	Prob-ImageCLEF	Prob-Pascal
Animate	0.91	0.84
Inanimate	0.53	0.55

UIUC

Influence of Semantic Factors (Scene)

office	airport	kitchen	dining room	field	living room	street	river	restaurant	sky	forest	mountain
0.29	0.13	0.36	0.21	0.16	0.13	0.18	0.1	0.28	0.18	0.0	0.07

Probability of description for each scene type

Rating	1	2	3	4	5
Prob	0.15	0.21	0.21	0.22	0.26

Probability of scene term given scene depiction strength

Discussion: Is scene description strength a good indicator of scene importance in human descriptions?

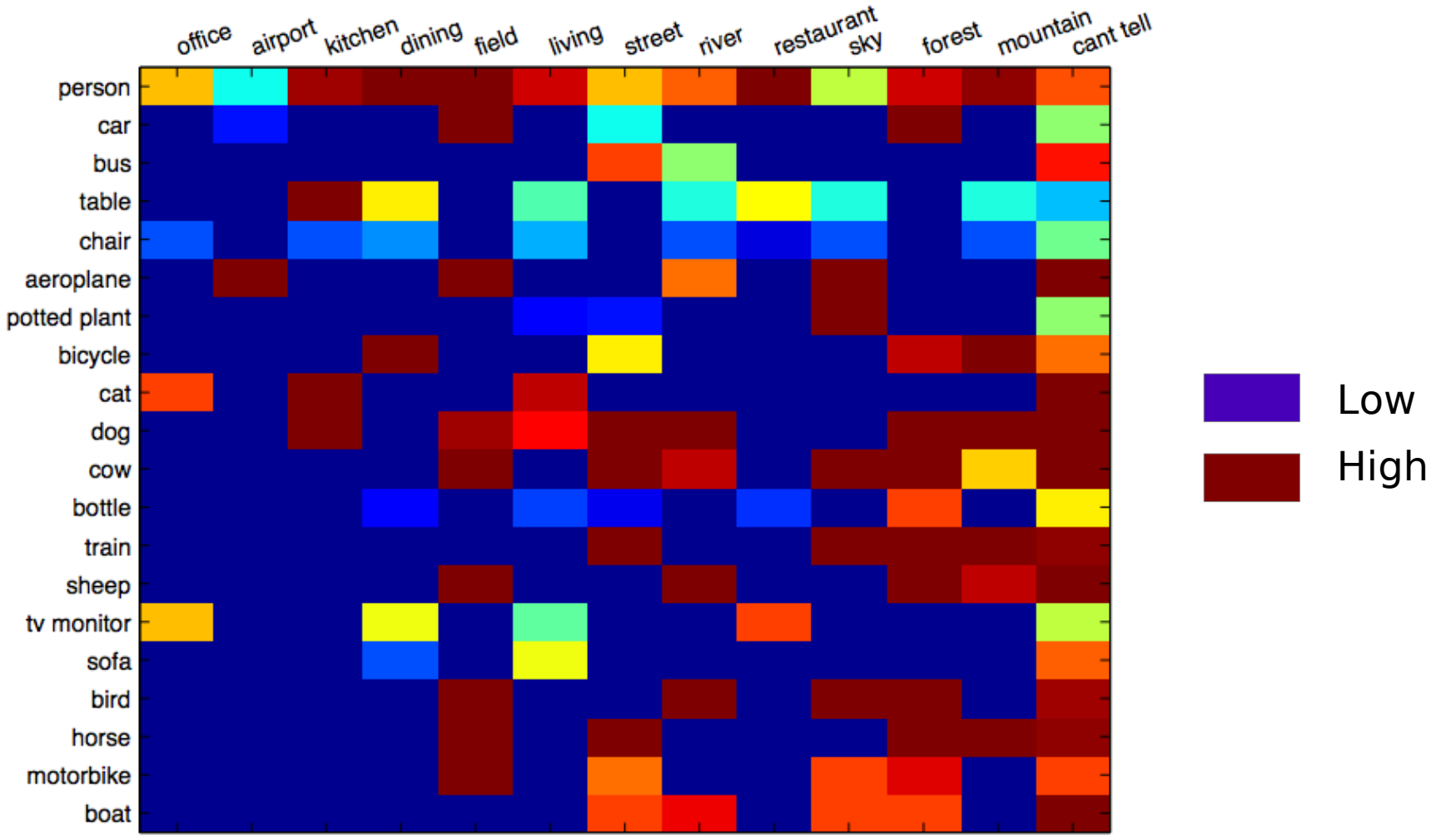
Exploring Importance Factors

Context factors:

Unusual object-scene Pair

"A tree in water and a boy with a beard"

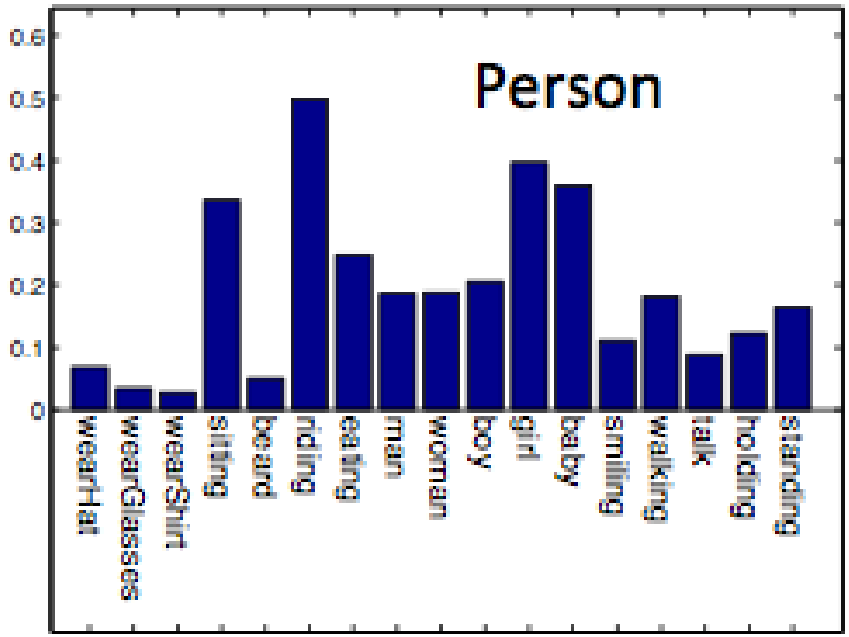
Influence of Context Factors (Obiect-Scene)

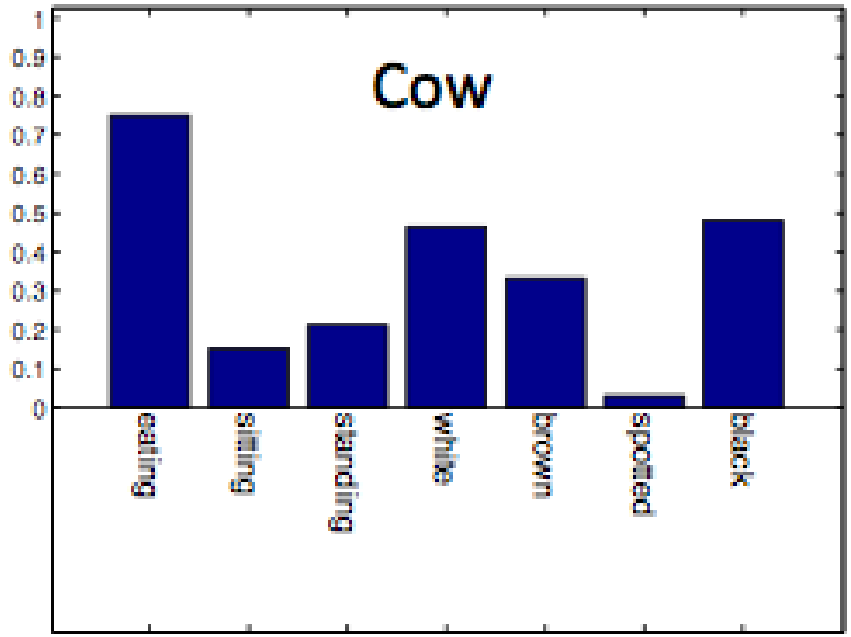


Discussion

- How do the authors handle cases when there are no images or relatively few images for the object-scene combination being considered?

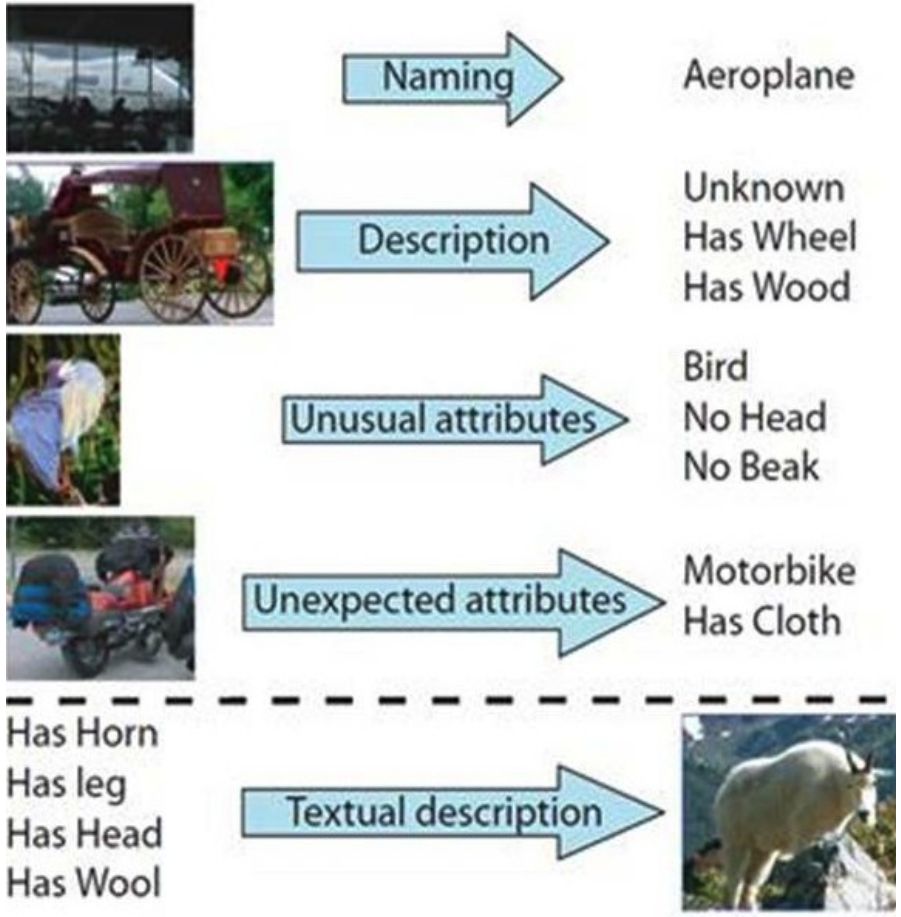
Influence of Context Factors (Attribute-Object)





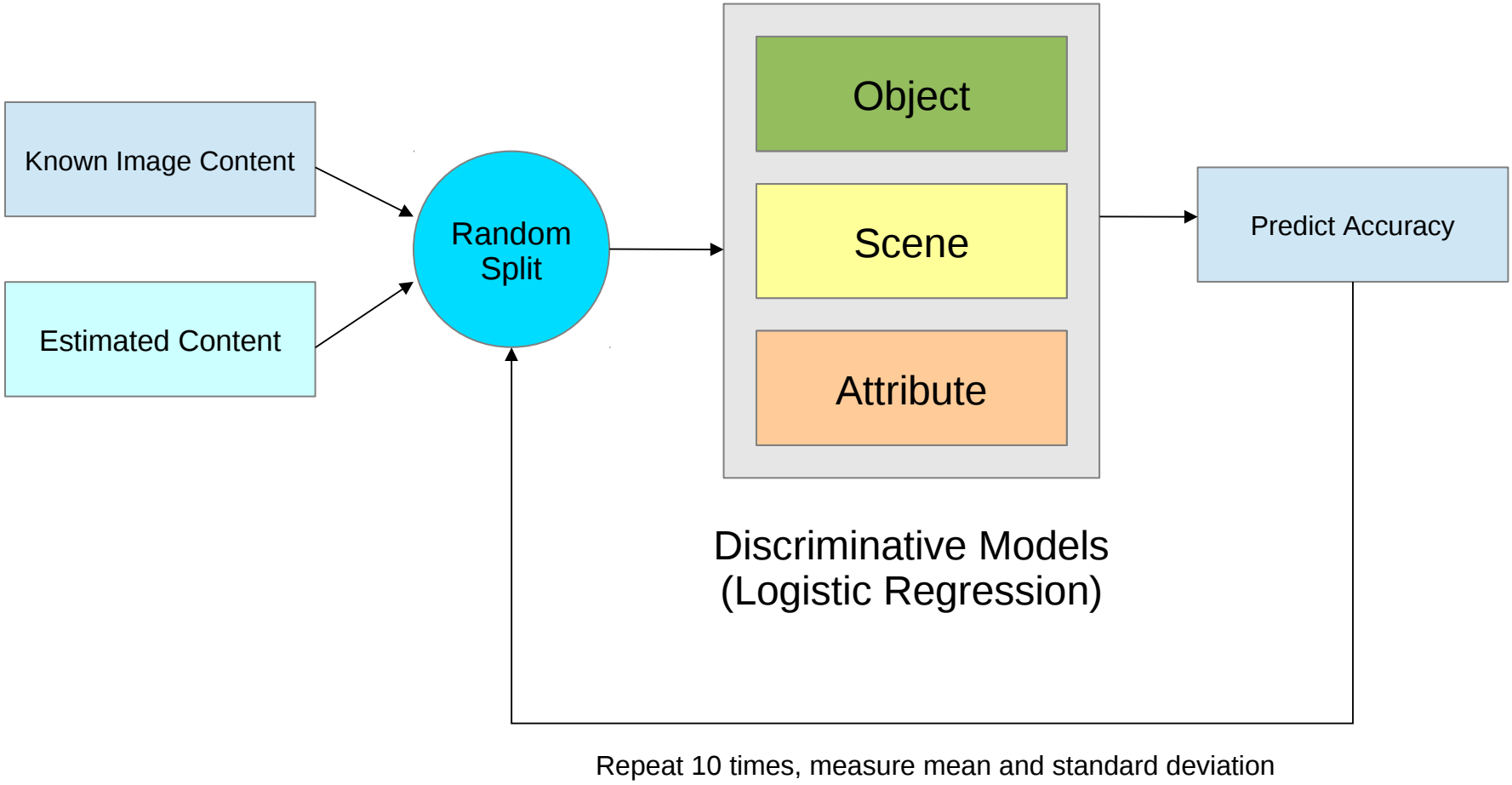
Discussion

- Humans tend to describe unexpected objects in scenes or unexpected attributes in objects.
- Can the technique proposed by Farhadi et al. for detecting unexpected attributes be used for predicting context factors instead?

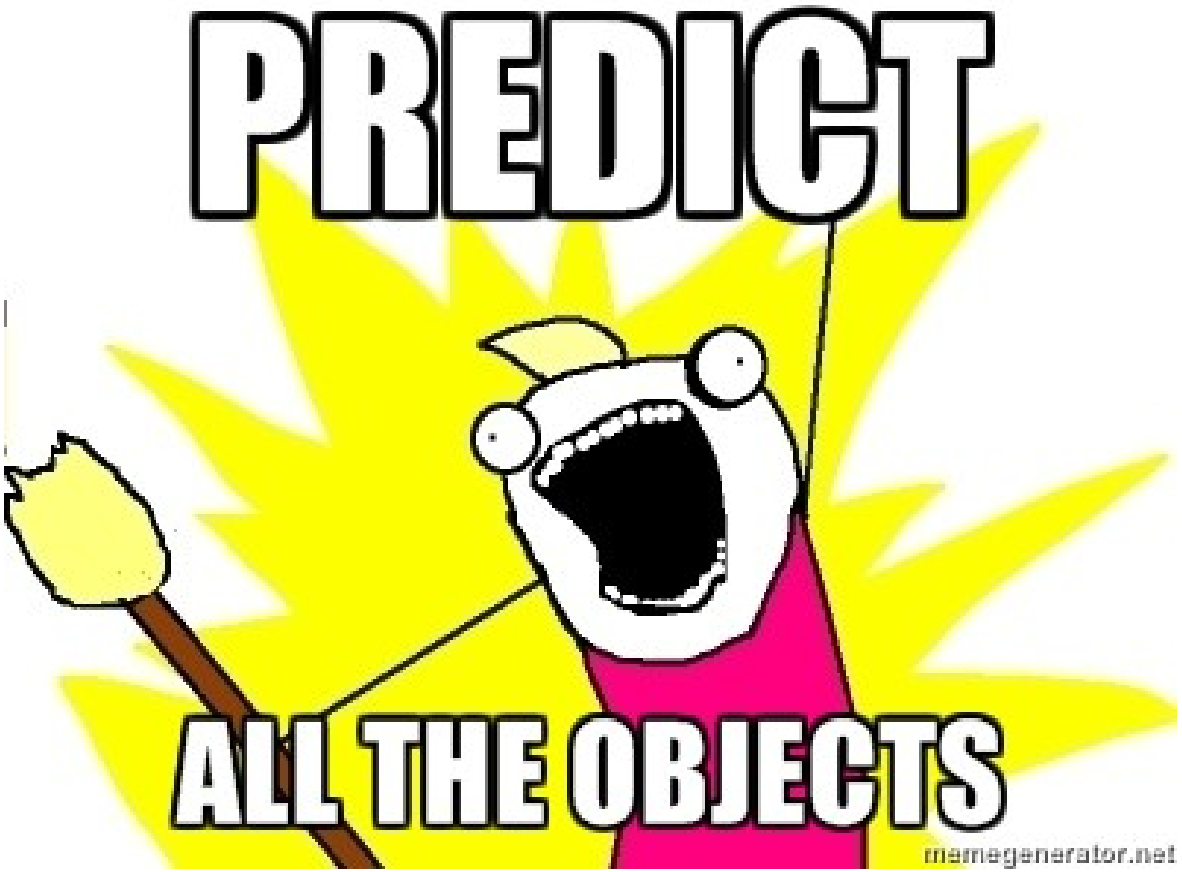


- Problem Overview
- Motivation
- Approach
- Experiments
- Discussion

Predicting Importance



Object Prediction: Baseline



Results: Predicting objects in sentences (UIUC+ImageCLEF)

Model	Features	Accuracy% (std)
Baseline (ImageCLEF)		57.5 (0.2)
Log Reg (ImageCLEF)	$K_o^s + K_o^l$	60.0 (0.1)
Log Reg (ImageCLEF)	K_o^c	68.0 (0.1)
Log Reg (ImageCLEF)	$K_o^c + K_o^s + K_o^l$	69.2 (1.4)
Baseline (UIUC-Kn)		69.7 (1.3)
Log Reg (UIUC-Kn)	$K_o^s + K_o^l$	69.9 (0.6)
Log Reg (UIUC-Kn)	K_o^c	79.8 (1.4)
Log Reg (UIUC-Kn)	$K_o^c + K_o^s + K_o^l$	82.0 (0.9)
Baseline (UIUC-Est)		76.5 (1.0)
Log Reg (UIUC-Est)	$E_o^s + E_o^l$	76.9 (1.1)
Log Reg (UIUC-Est)	E_o^c	78.9 (1.4)
Log Reg (UIUC-Est)	$E_o^c + E_o^s + E_o^l$	79.52 (1.2)

Results: Predicting scenes in sentences (UIUC)

Model	Features	Accuracy% (std)
Baseline (UIUC-Kn)		86.0 (0.2)
Log Reg (UIUC-Kn)	K_s^c+K_s^r	96.6 (0.2)
Log Reg (UIUC-Est)	E_s^d	87.4 (1.3)

Discussion: It would have been useful to know how the model performed based on scene category only.

Results: Predicting attributes in sentences (UIUC)

Model	Features	Accuracy% (std)
Baseline (UIUC-Kn)		96.3 (.01)
Log Reg (UIUC-Kn)	$K_a^c + K_o^c$	97.0 (.01)
Log Reg (UIUC-Est)	$E_a^d + E_o^c$	96.7 (.01)

Experimental weakness: Accuracy is known to not be a good evaluation measure when there are a high number of false positives. Precision would be a better measure in such cases as we are measuring the fraction of true positives instead.

Strengths

- This paper evaluates the importance of scene and attribute information in human descriptions
- They define importance as the probability of image content appearing in natural language descriptions
- Importance factors are examined on a larger scale (1K-20K images)

- Problem Overview
- Motivation
- Approach
- Experiments
- Discussion

Memorability and Predicting Importance

- Some images are memorable, while some are not. However, while predicting importance, all images are assumed to have some important element(s)
- Both tend to give importance to animate objects. Memorability features can definitely be used for predicting importance

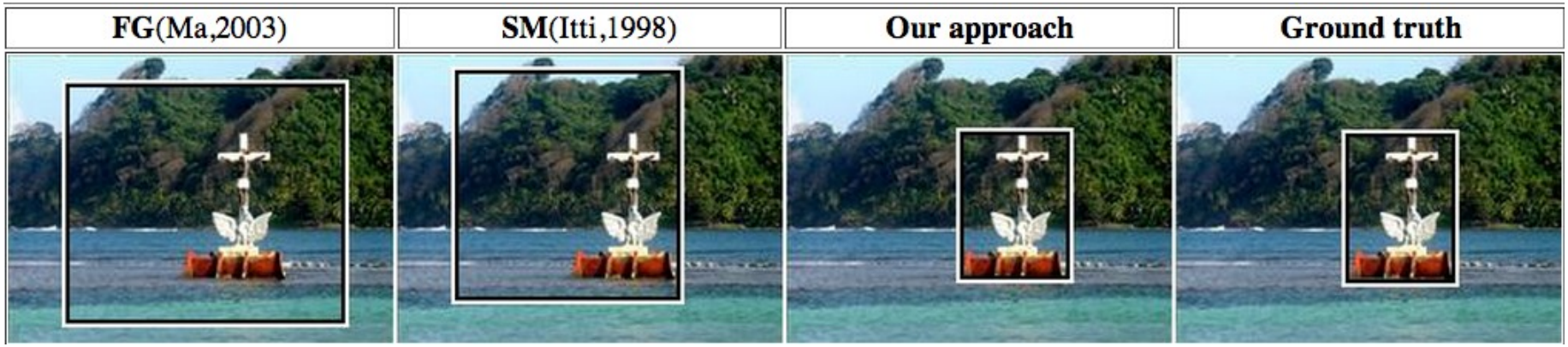
a) Most memorable images (86%)

b) Typical images (74%)

c) Least memorable images (34%)

Saliency and Predicting Importance

- Saliency: what grabs visual attention in an image?
- Humans tend to talk about salient objects and hence they are important
- While scene elements may be predicted as important, they are rarely considered salient on their own.



Questions?