Inferring the Why in Images

[Pirsiavash et al]

CSC2523 Winter 2015: Paper Presentation

Micha Livne

Goals

(a)

Sitting

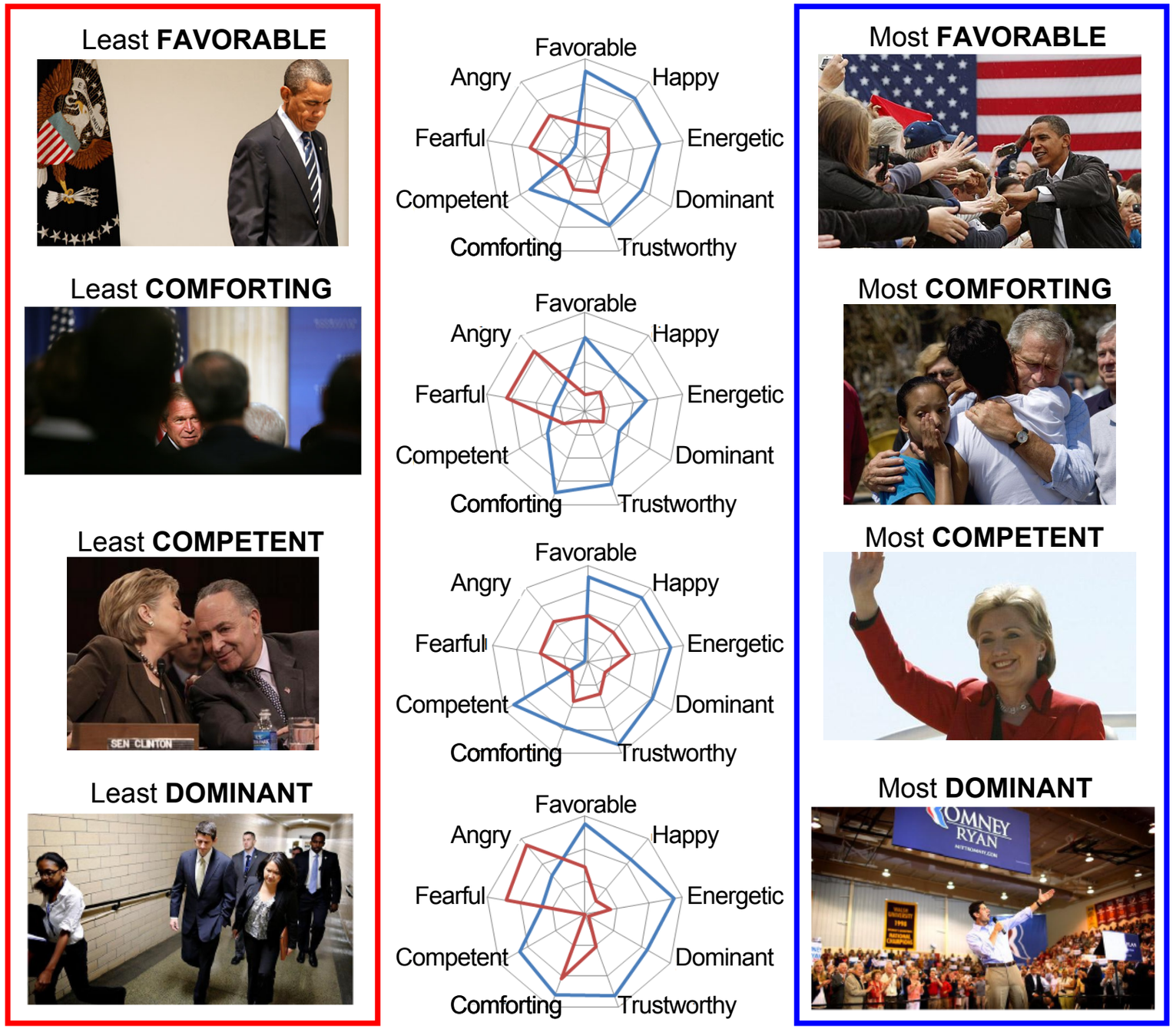
because he wants
to watch television

(b)

Sitting

because she intends
to see the doctor

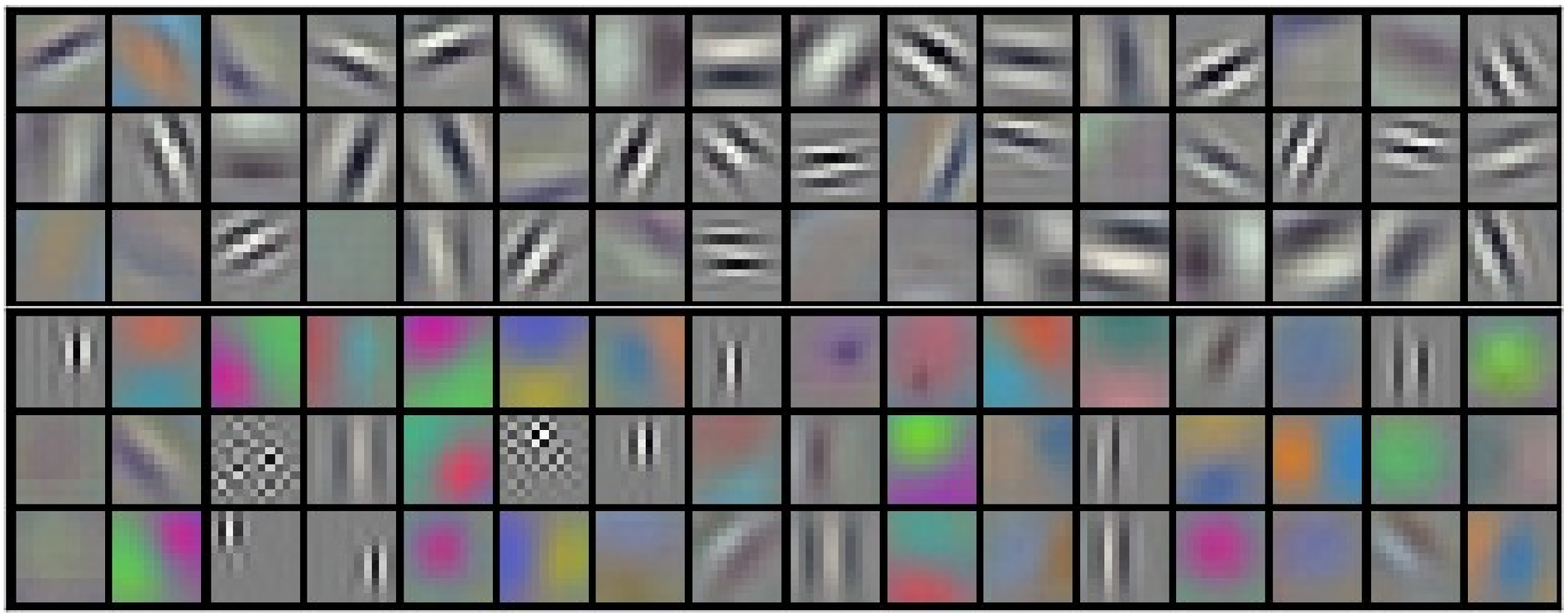
Related Work



(d) Example images and predicted intents

Visual Persuasion: Inferring Communicative Intents of Images [Joo et al 2014]

Proposed Solution: Vision Only



$\phi(x)$

[Krizhevsky et al 2012]

Visual classifier $\underset{y \in \{1,\ldots,M\}}{\operatorname{argmax}} \; w_y^T \phi(x)$

Proposed Solution: Full Solution

Language Potentials

Relationship	Query to Language Model

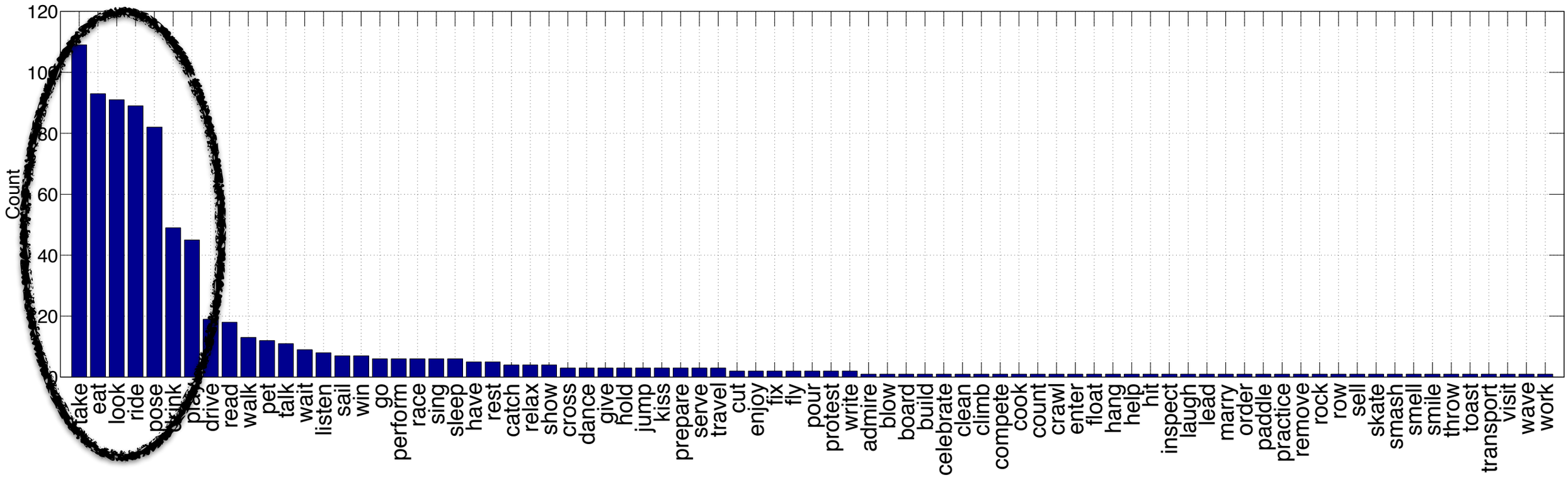
action + object + motivation	`action` the `object` in order to `motivation`
	`action` the `object` to `motivation`
	`action` the `object` because `pronoun` wants to `motivation`

action + object + scene	`action` the `object` in a `scene` in a `scene`, `action` the `object`

action + scene + motivation	`action` in a `scene` in order to `motivation`
	`action` in order to `motivation` in a `scene`
	`action` because `pronoun` wants to `motivation` in a `scene`

$$L_{ij}(y_i, y_j)$$

Dataset



Statistics of Motivations

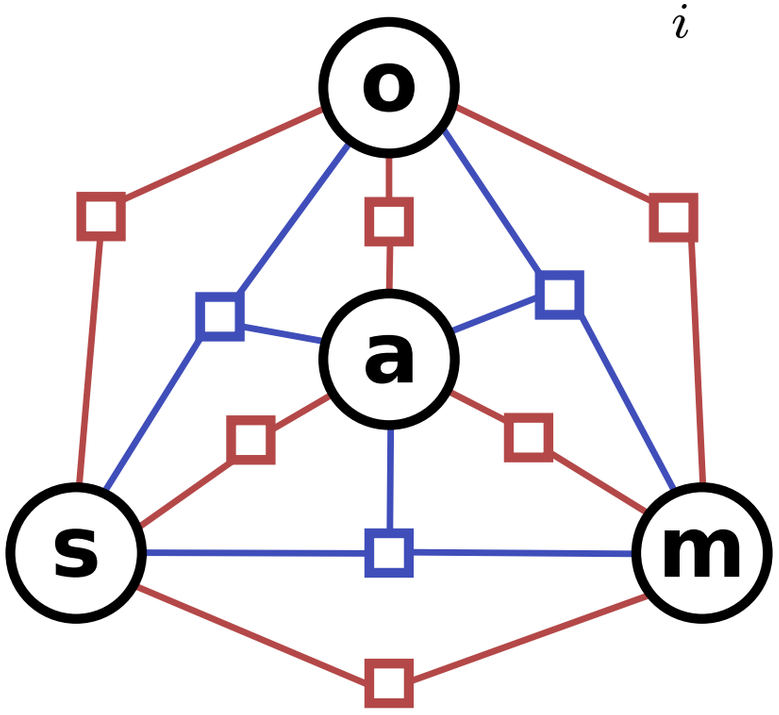
- Based on PASCAL VOC 2012.
- Only images with a person.
- Annotation of: action, object, scene , and motivation (79).

Proposed Solution: Full Solution

Scoring Function

$$\Omega(y; w, u, x, L) = \sum_{i}^{N} w_{y_i}^T \phi_i(x)$$

$$+ \sum_{i}^{N} u_i L_i(y_i) + \sum_{i<j}^{N} u_{ij} L_{ij}(y_i, y_j) + \sum_{i<j<k}^{N} u_{ijk} L_{ijk}(y_i, y_j, y_k)$$



Proposed Solution: Full Solution

Learning

$$\underset{\theta,\xi^n \geq 0}{\text{argmin}} \ \frac{1}{2}||\theta||^2 + C\sum_n \xi^n$$

$$\text{s.t.} \quad \theta^T\psi(y^n, x^n) - \theta^T\psi(h, x^n) \geq \Delta(y^n, h) - \xi^n \quad \forall_n, \forall_h$$

Inference

$$y^* = \underset{y}{\text{argmax}} \ \Omega(y; w, u, x, L)$$

Results

Success

Human Label: sitting on bench in a train station because he is waiting

Top Predictions: 1. sitting on bench in a park because he is waiting
2. holding a tv in a park because he wants to take
3. holding a seal in a park because he wants to protest
4. holding a guitar in a park because he wants to play

Human Label: sitting on chair in a dining room because she wants to eat

Top Predictions: 1. sitting near table in dining room because she wants to eat
2. sitting on a sofa in a dining room because she wants to eat
3. holding a cup in a dining room because she wants to eat
4. sitting on a cup in a dining room because she wants to eat

Results

Failure

Human Label: holding a person in a living room because she wants to show

Top Predictions:
1. sitting on sofa in living room because she wants to pet
2. sitting on sofa in living room because she wants to look
3. sitting on sofa in living room because she wants to read
4. sitting on chair in living room because she wants to pet

Human Label: standing next to table because she wants to prepare

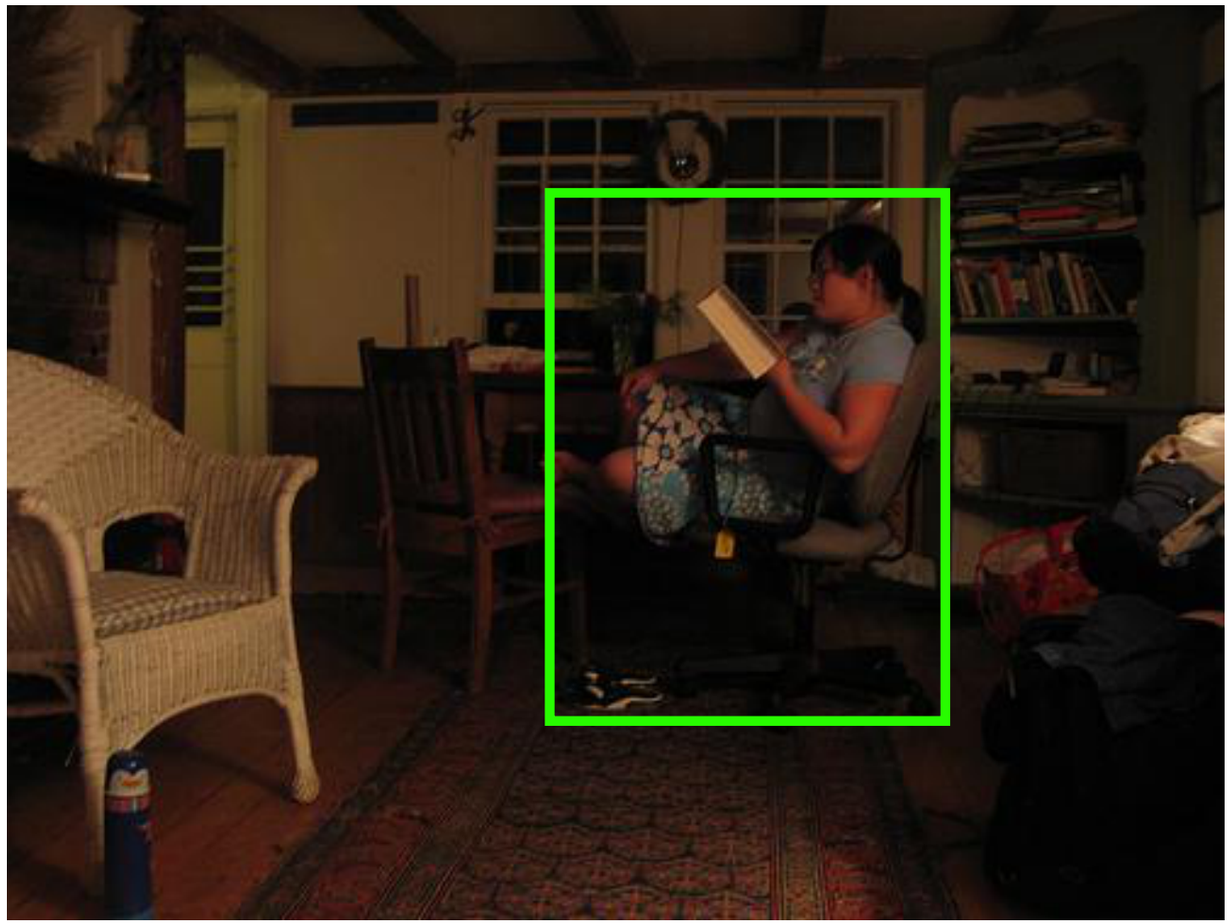
Top Predictions:
1. talking to person in dining because she wants to eat
2. standing next to table in dining room because she wants to eat
3. sitting next to table in dining because she wants to eat
4. talking to person in kitchen because she wants to eat

Results

Failure: Vision Only

Human Label: sitting on a bus in a parking lot because he wants to drive

Top Predictions:
1. because he wants to look
2. because he wants to ride
3. because he wants to drive
4. because he wants to eat



Human Label: sitting on chair in living room because she wants to read

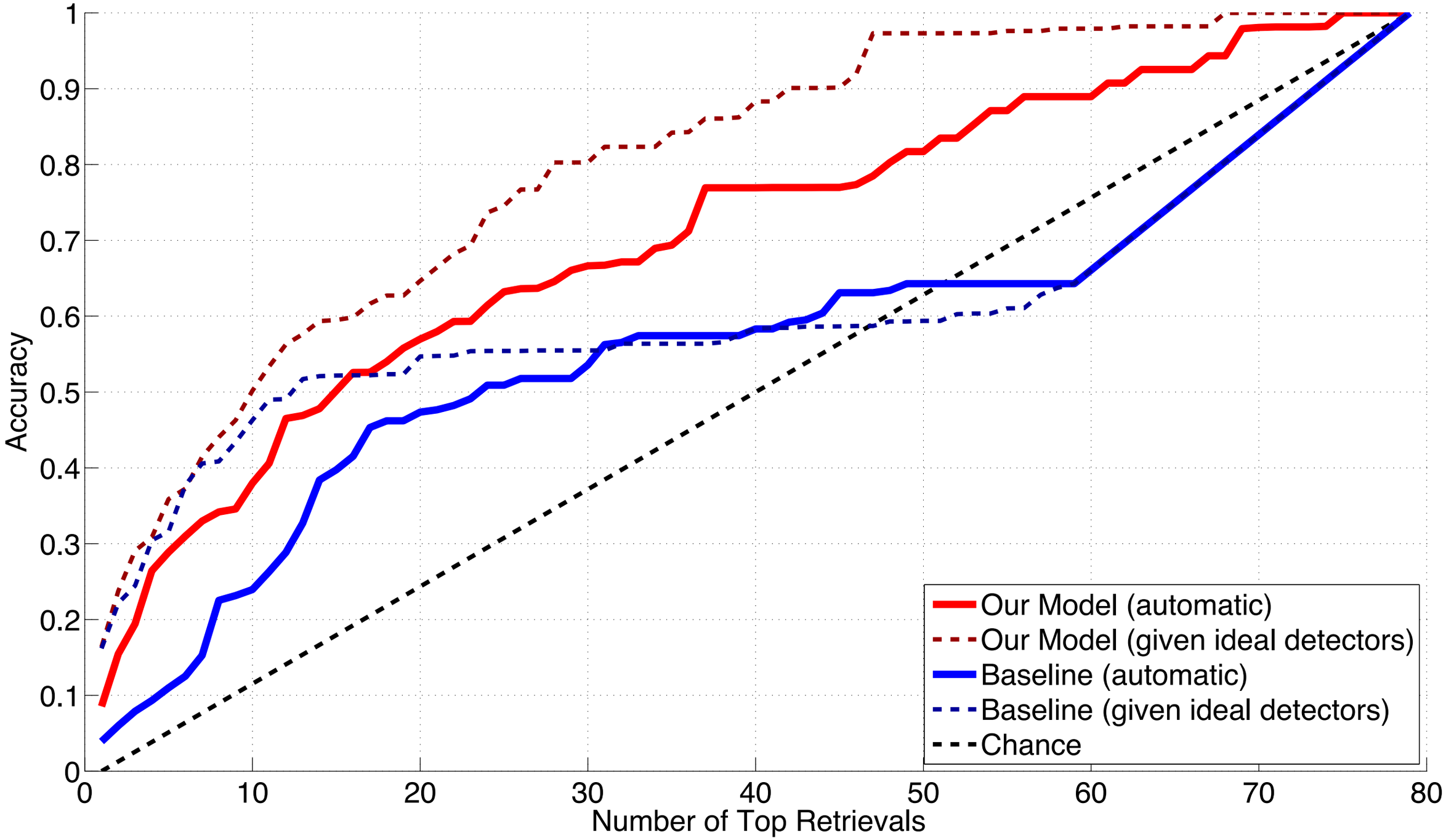
Top Predictions:
1. because she wants to eat
2. because she wants to look
3. because she wants to drink
4. because she wants to ride

Results

		Baseline (Vision Only)	Our Method (With Language)
Given Ideal Detectors for:	Action+Object+Scene	13	10
	Action+Object	12	11
	Object+Scene	15	12
	Action+Scene	19	13
	Object	19	13
	Action	18	15
	Scene[1]	37	18
Fully Automatic		23[2]	**15**

Chance has rank of 39

Results


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Accuracy
0
10
20
30
40
50
60
70
80
Number of Top Retrievals
Our Model (automatic)
Our Model (given ideal detectors)
Baseline (automatic)
Baseline (given ideal detectors)
Chance

Point of Strength

- Novel and important problem
- Simple model - easy to understand
- Augmenting image with text through data mining was proven to be effective

Point of Weakness

- Results are only ok (qualitatively, failure of vision-only model does not make much more sense)
- Model is linear - too simple
- Language queries are simple as well

Contributions

- Introducing the problem of inferring motivation behind people's actions to the computer vision community.

- Propose to use common knowledge mined from web to improve computer vision systems.

Conclusion

- Interesting problem
- The proposed method is more of a baseline
- Future research can extend prediction model, and language model

Thanks!

Questions?