

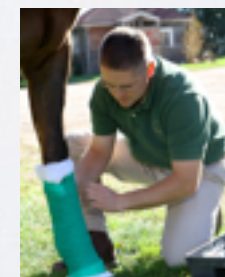
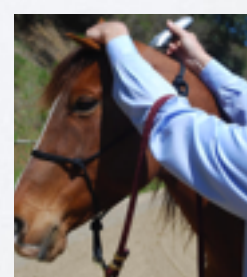
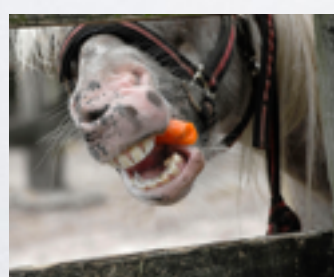
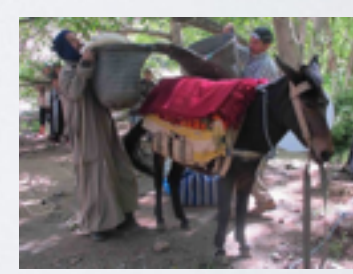
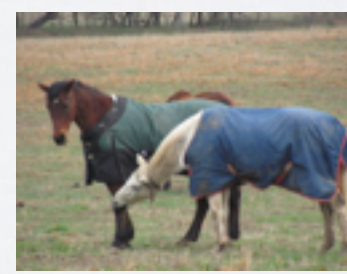
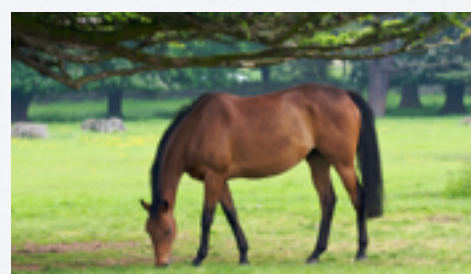
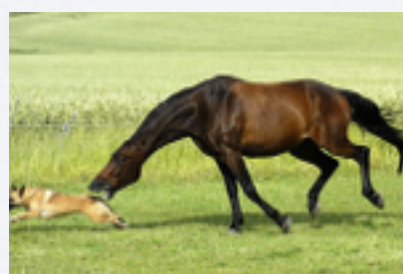
Situation Recognition

Visual Semantic Role Labeling for Image Understanding

Mark Yatskar



in collaboration w/ Luke Zettlemoyer, Ali Farhadi



How can we summarize what is happening in an image?

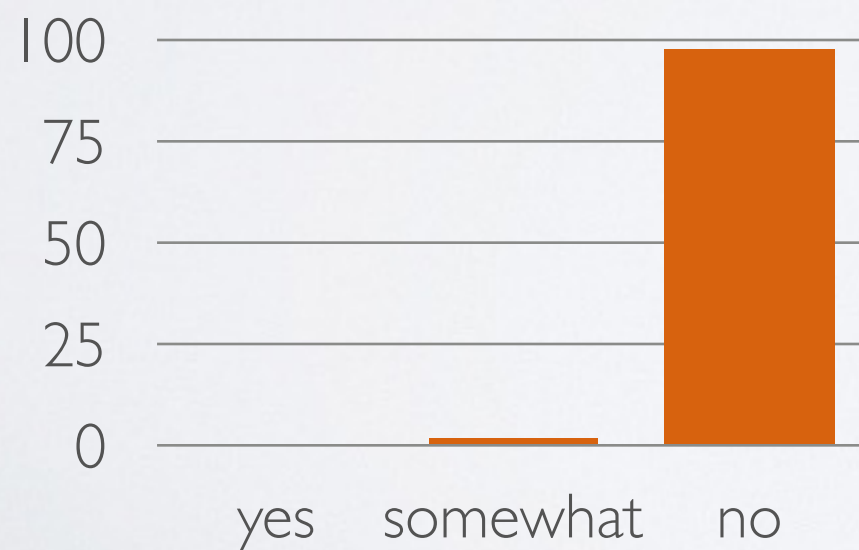


LOADING				
AGENT	ITEM	DESTINATION	TOOL	PLACE
WOMAN	HORSE	TRAILER	ROPE	OUTDOORS

Is the same thing happening in two images?



turkers say...

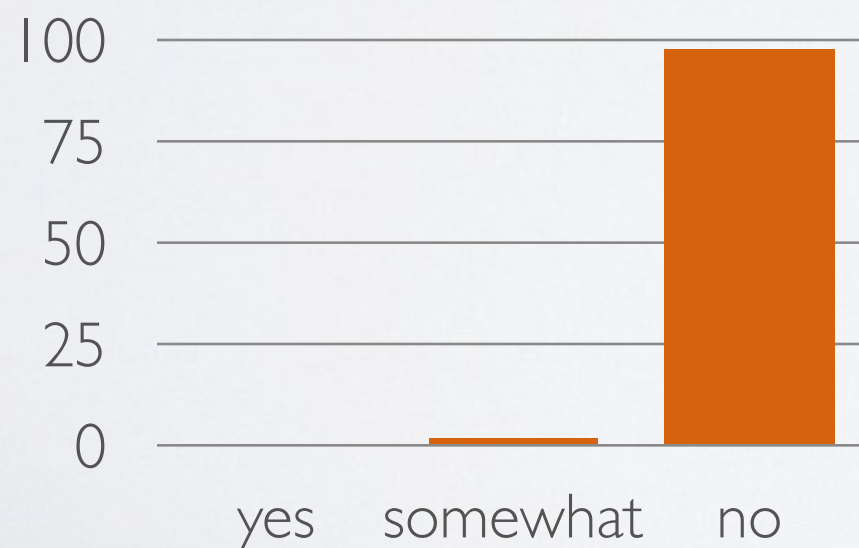


Is the same thing happening in two images?

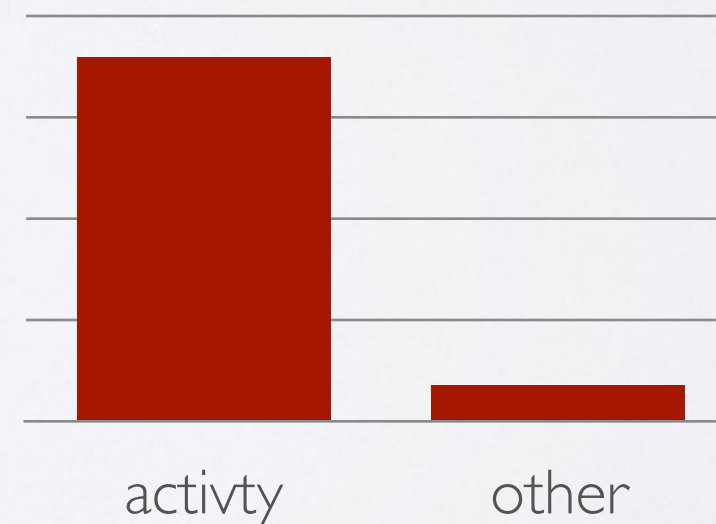


Activity

turkers say...



why no?

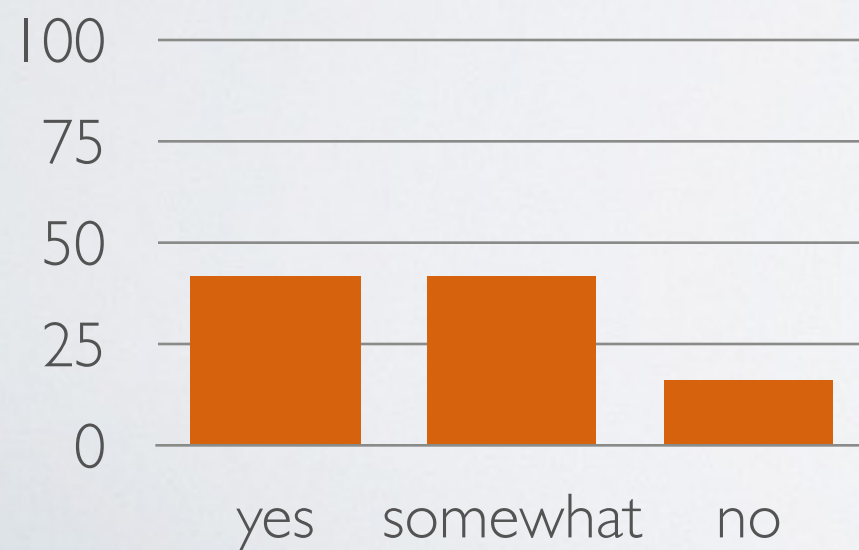


Is the same thing happening in two images?

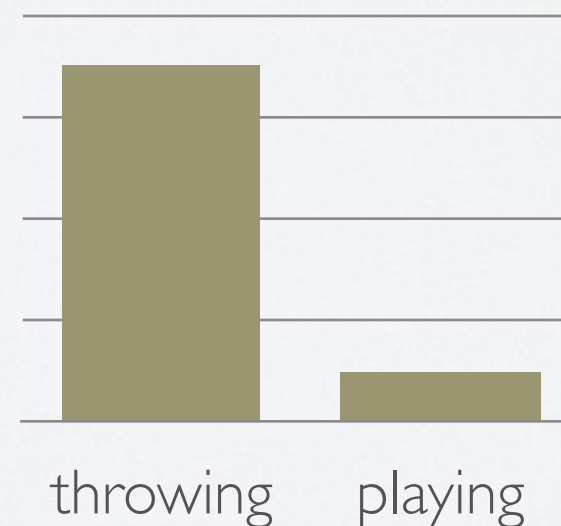


Activity

turkers say...



why yes?

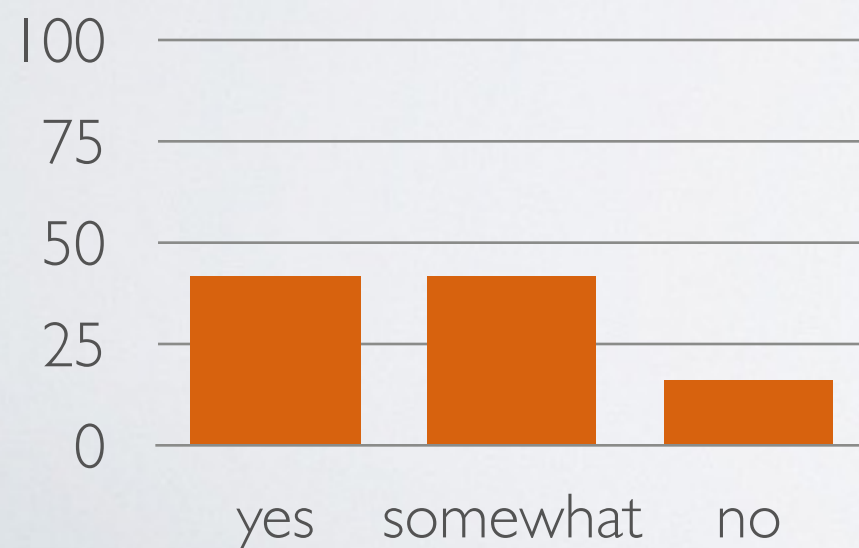


Is the same thing happening in two images?

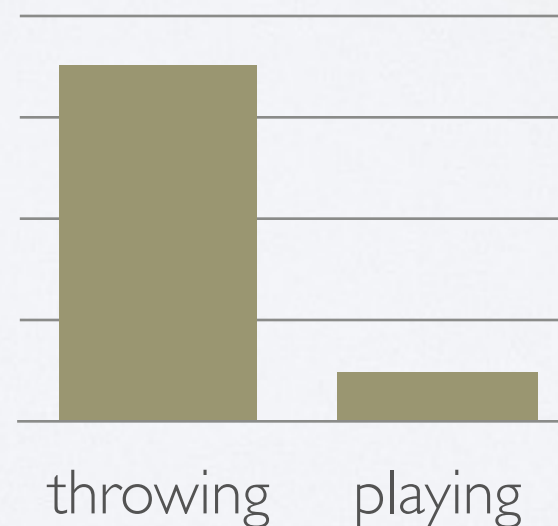


Activity
Object

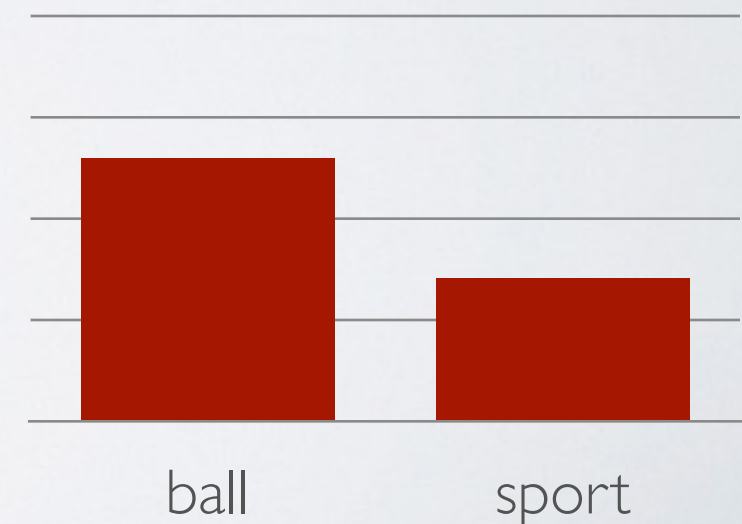
turkers say...



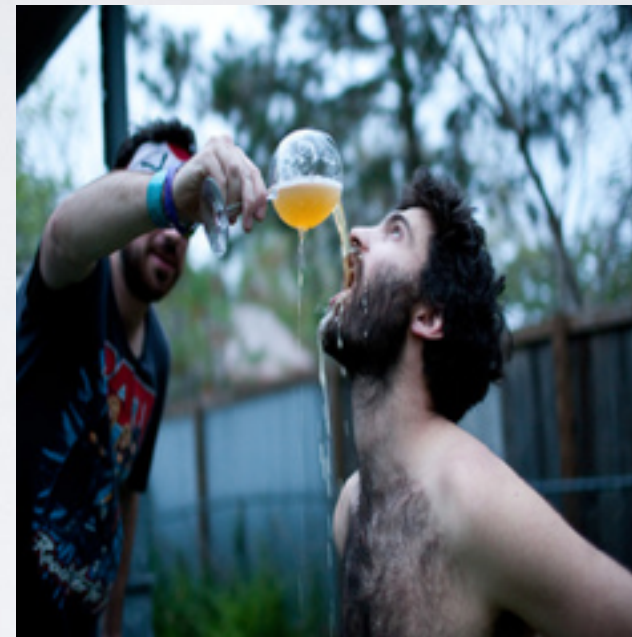
why yes?



why no?

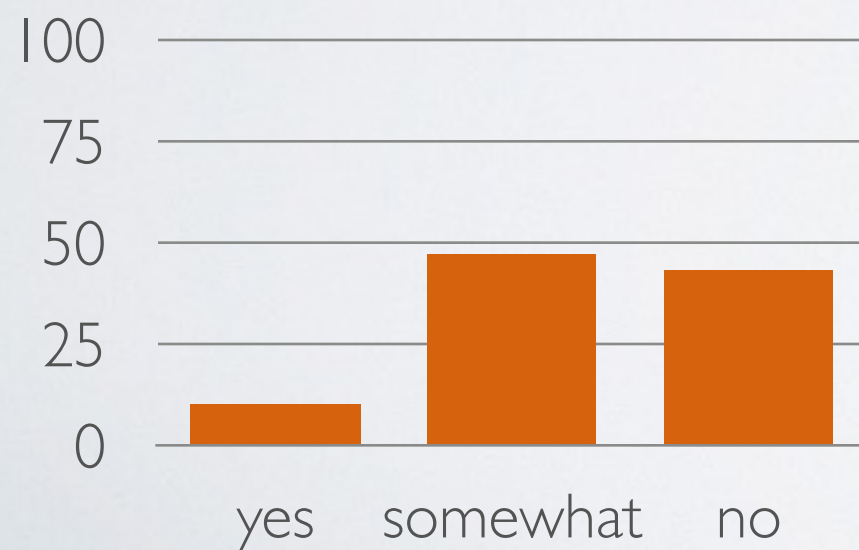


Is the same thing happening in two images?

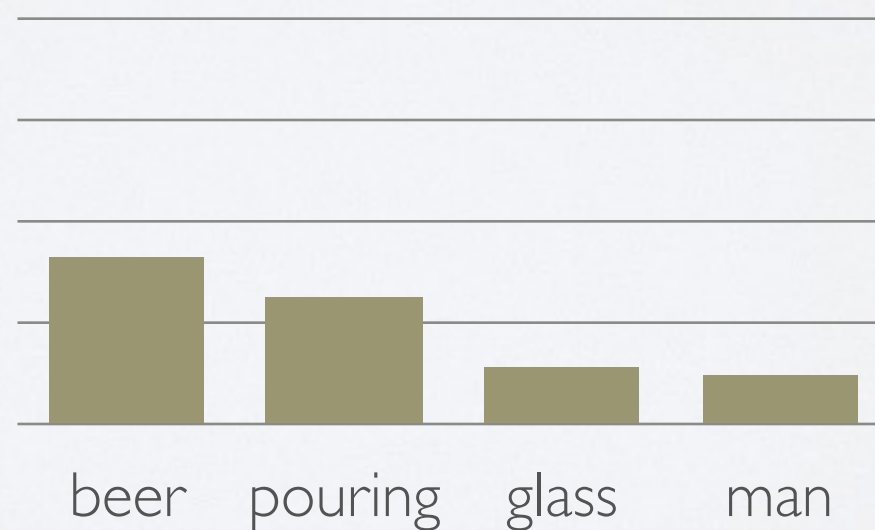


Activity
Object

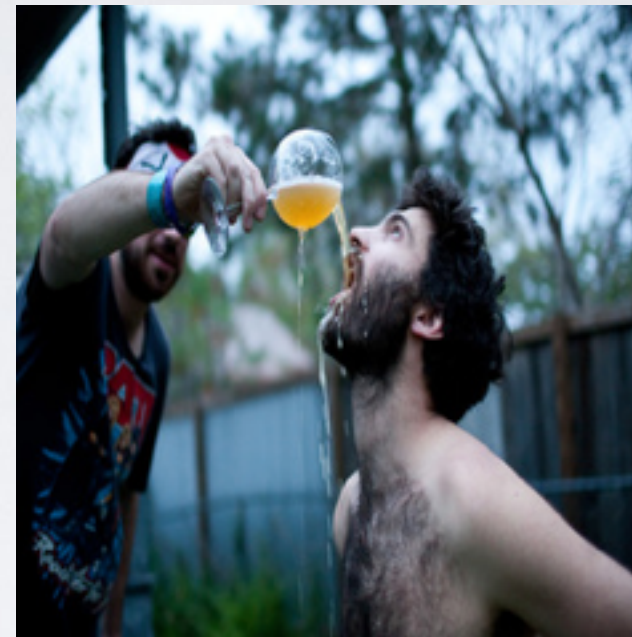
turkers say...



why yes?

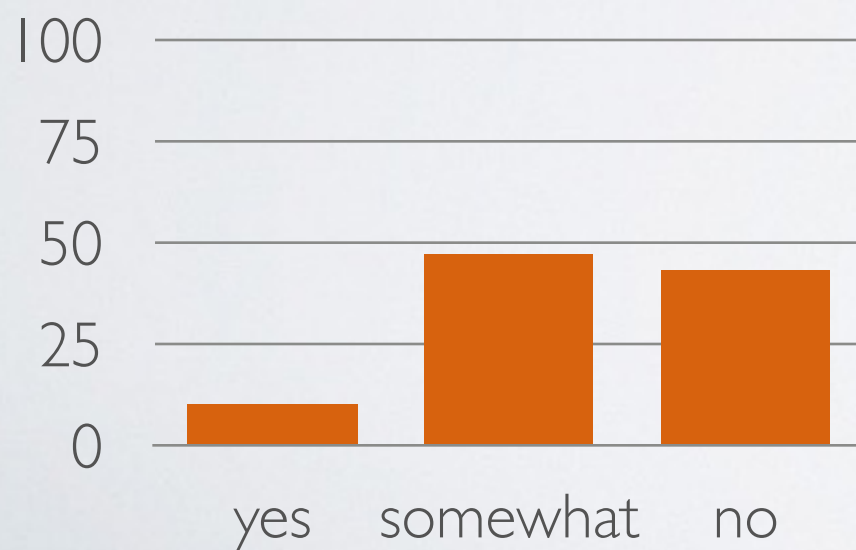


Is the same thing happening in two images?

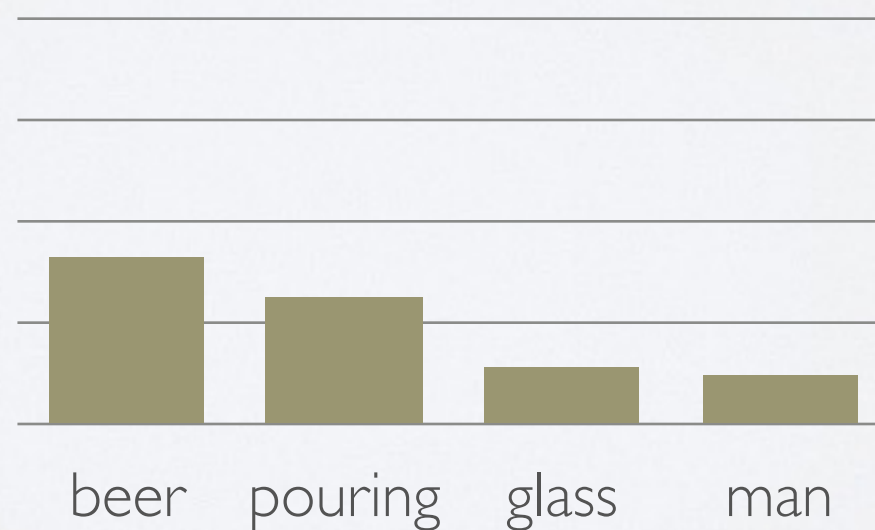


Activity
Object
Role

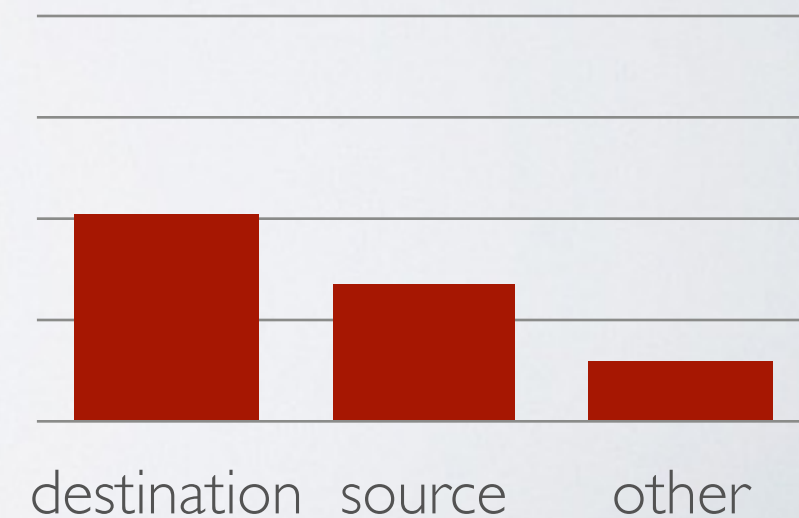
turkers say...



why yes?



why no?



Systematically describe how objects participate in activities through **roles**



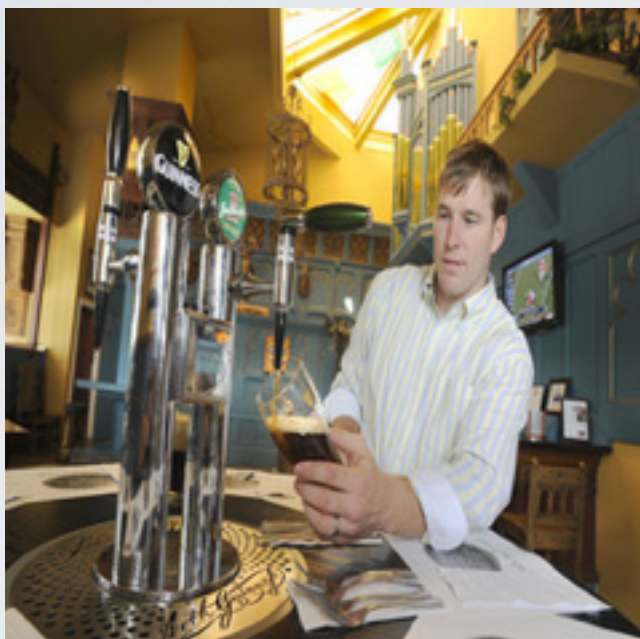
LOADING				
AGENT	ITEM	DESTINATION	TOOL	PLACE
WOMAN	HORSE	TRAILER	ROPE	OUTDOORS

Situation Recognition

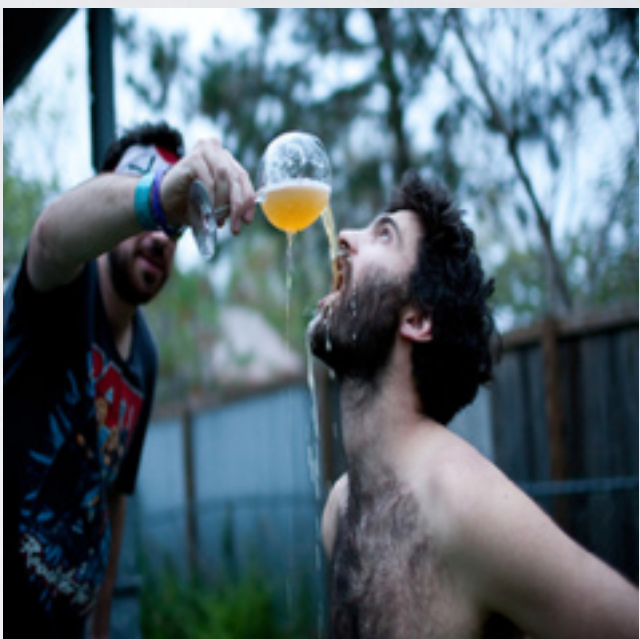


FIXING				
AGENT	OBJECT	PART	TOOL	PLACE
BOY	CAR	TIRE	TIRE IRON	OUTDOORS

Situation Recognition



POURING	
AGENT	MAN
SUBSTANCE	BEER
SOURCE	TAP
DESTINATION	GLASS
PLACE	BARROOM



POURING	
AGENT	MAN
SUBSTANCE	BEER
SOURCE	GLASS
DESTINATION	MOUTH
PLACE	BACKYARD

Same

Different

Situation Recognition



What is the space of possible situations?

POURING	
AGENT	MAN
SUBSTANCE	BEER
SOURCE	TAP
DESTINATION	GLASS
PLACE	BARROOM

Same

Different

POURING	
AGENT	MAN
SUBSTANCE	BEER
SOURCE	GLASS
DESTINATION	MOUTH
PLACE	BACKYARD

imSitu

A Large Scale Situation Dataset

120k+ images, 500+ verbs, 100k+ situations

Natural Language Processing: Semantic Role Labeling



A boy is fixing a car tire with a tire iron outdoors.

Natural Language Processing: Semantic Role Labeling



A boy is **fixing** a car tire with a tire iron outdoors.

Natural Language Processing: Semantic Role Labeling



A boy is **fixing** a car tire with a tire iron outdoors.

Natural Language Processing: Semantic Role Labeling



A boy is **fixing** a car tire with a tire iron outdoors.

FIXING				
AGENT	OBJECT	PART	TOOL	PLACE
BOY	CAR	TIRE	TIRE IRON	OUTDOORS

Natural Language Processing: Semantic Role Labeling

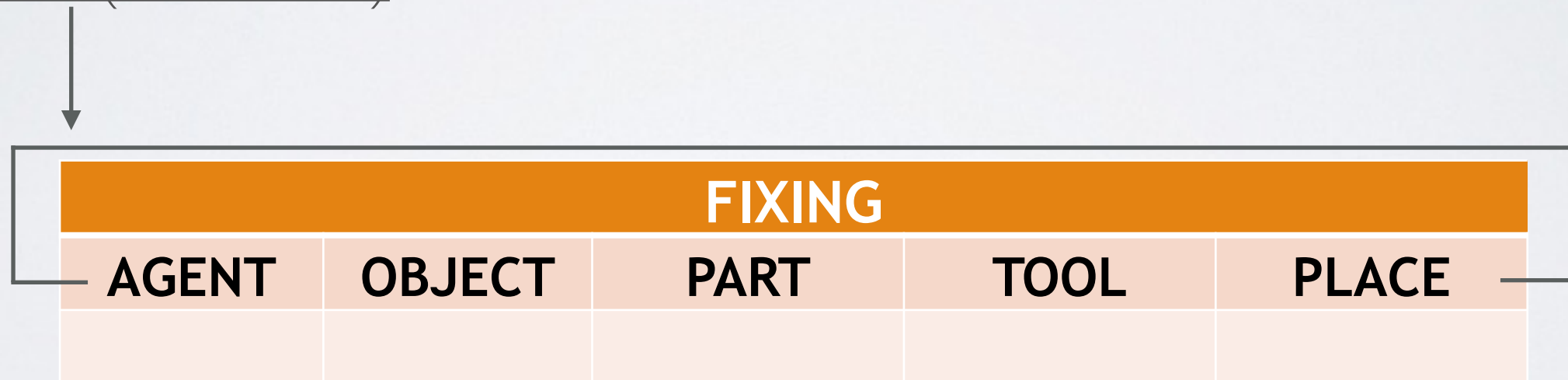


A jockey **falling** from a horse onto the ground at a racetrack.

FALLING			
AGENT	SOURCE	DESTINATION	PLACE
JOCKEY	HORSE	GROUND	RACETRACK

FrameNet for Verb and Role Inventory

semantic role labeling
ontology:
FrameNet (8000 verbs)



The diagram illustrates the mapping from the FrameNet ontology to a specific frame structure. A vertical arrow points from the text 'FrameNet (8000 verbs)' to a table. The table has a header row with the word 'FIXING' in white text on an orange background. Below this header are five columns with the roles 'AGENT', 'OBJECT', 'PART', 'TOOL', and 'PLACE' in black text on a light orange background. Each role column has a corresponding empty cell below it. A bracket on the left side of the table groups the header and the first two columns, and a bracket on the right side groups the last two columns and the empty cells below them.

FIXING				
AGENT	OBJECT	PART	TOOL	PLACE

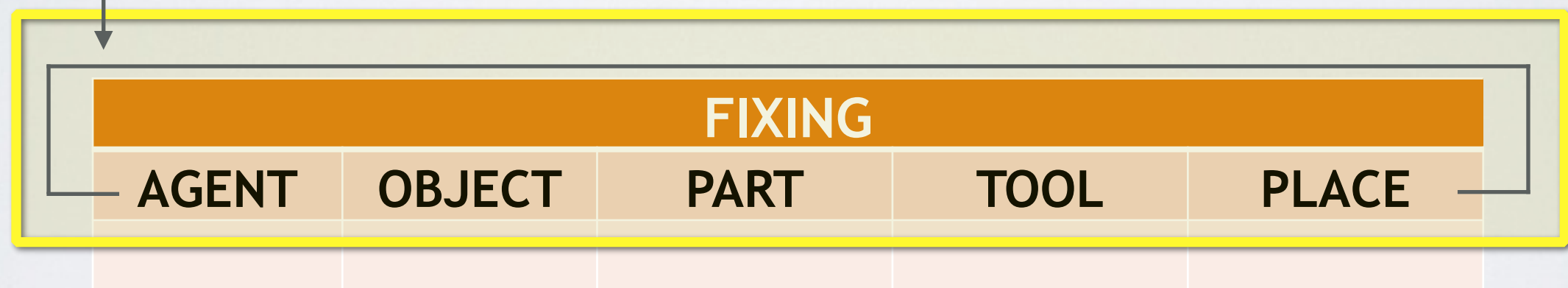
Visualness

filter verbs, semantic roles

~1000 visual verbs
~3.5 roles/verb

semantic role labeling
ontology:

FrameNet (8000 verbs)



WordNet for Noun Inventory

semantic role labeling
ontology:
FrameNet (8000 verbs)

FIXING				
AGENT	OBJECT	PART	TOOL	PLACE

values from noun ontology: WordNet (80,000 nouns)

FrameNet
Visualness
WordNet

Filter Images



Web N-grams
Google Images Search

semantic role labeling
ontology:
FrameNet (8000 verbs)

FIXING				
AGENT	OBJECT	PART	TOOL	PLACE

values from noun ontology: WordNet (80,000 nouns)

FrameNet
Visualness
WordNet
Filter Images

Fill Values



semantic role labeling
ontology:
FrameNet (8000 verbs)

FIXING				
AGENT	OBJECT	PART	TOOL	PLACE
BOY	CAR	TIRE	TIRE IRON	OUTDOORS

values from noun ontology: WordNet (80,000 nouns)

imSitu: Dataset Statistics

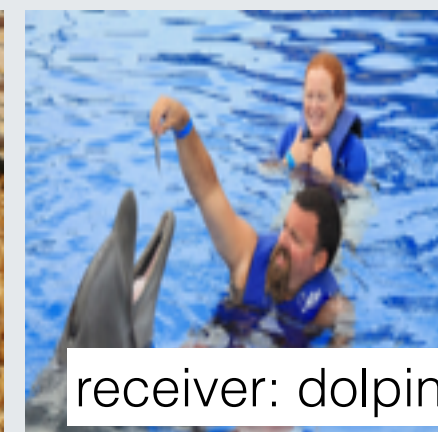
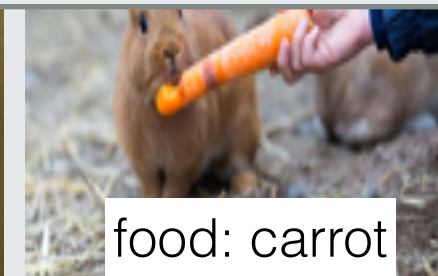
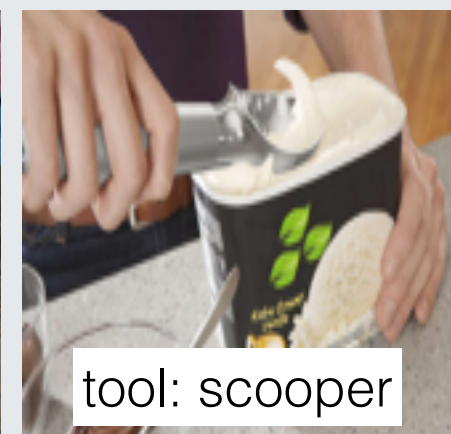
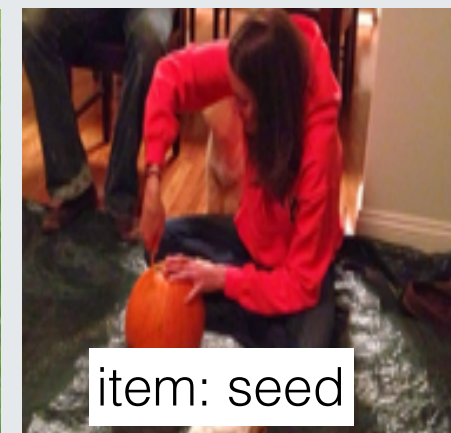
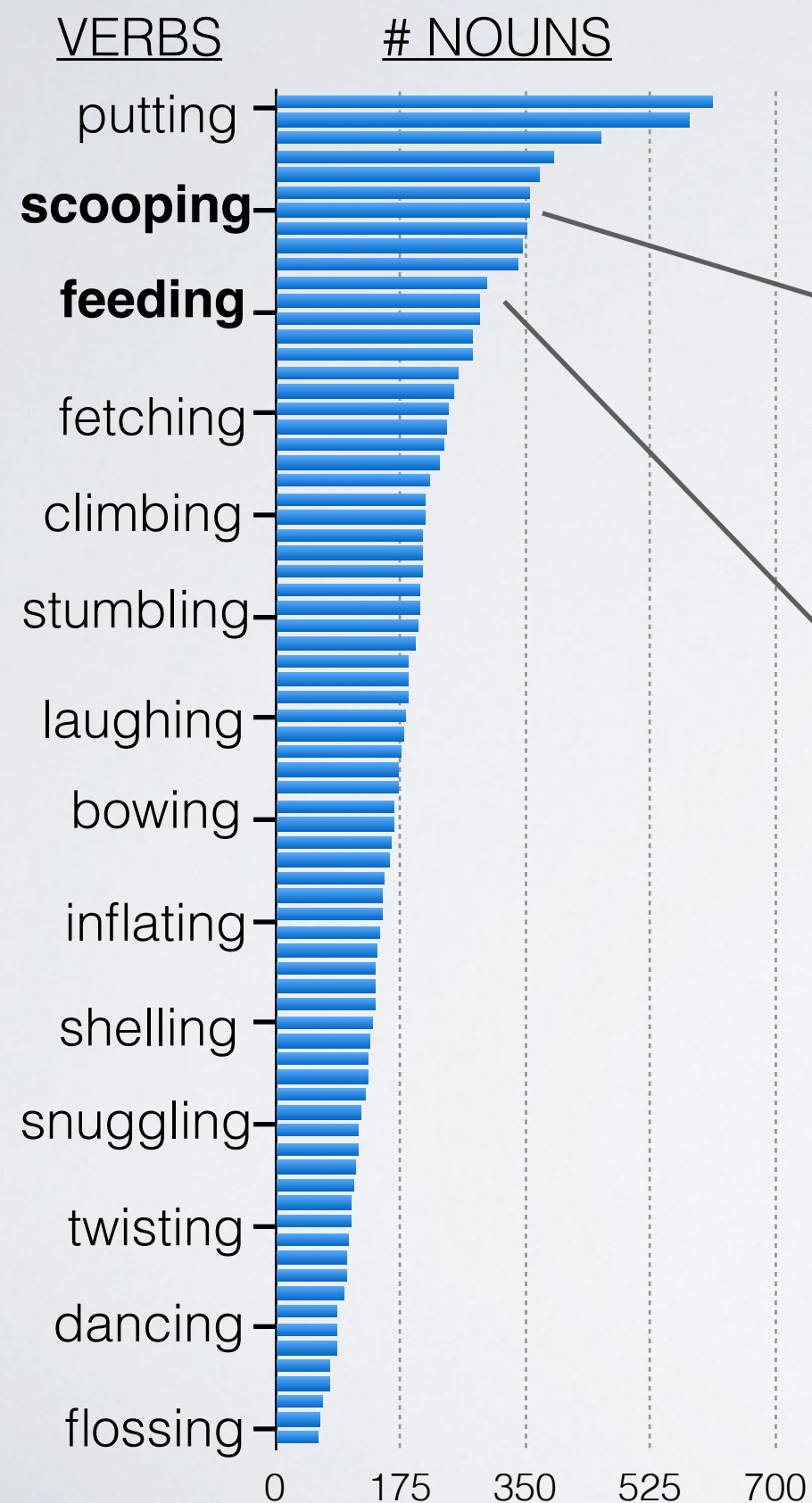


FrameNet
Visualness
WordNet
Filter Images
Fill Values

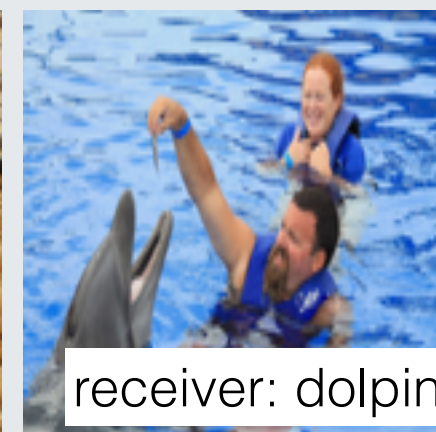
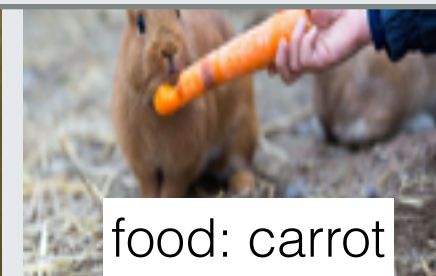
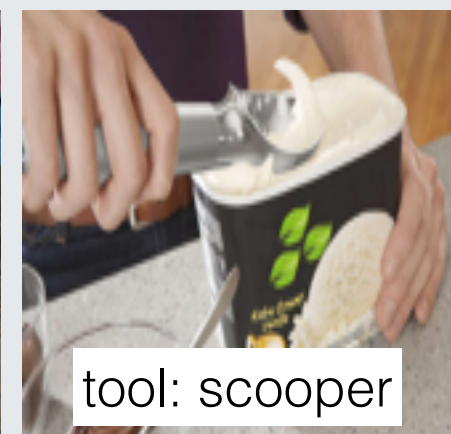
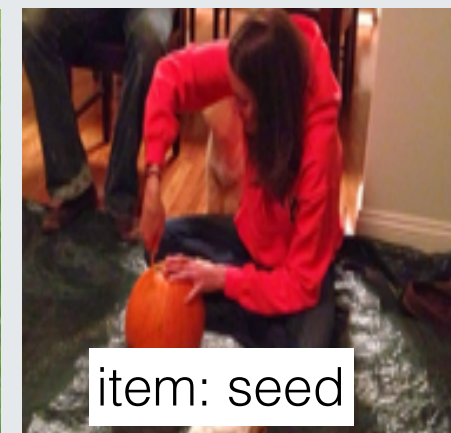
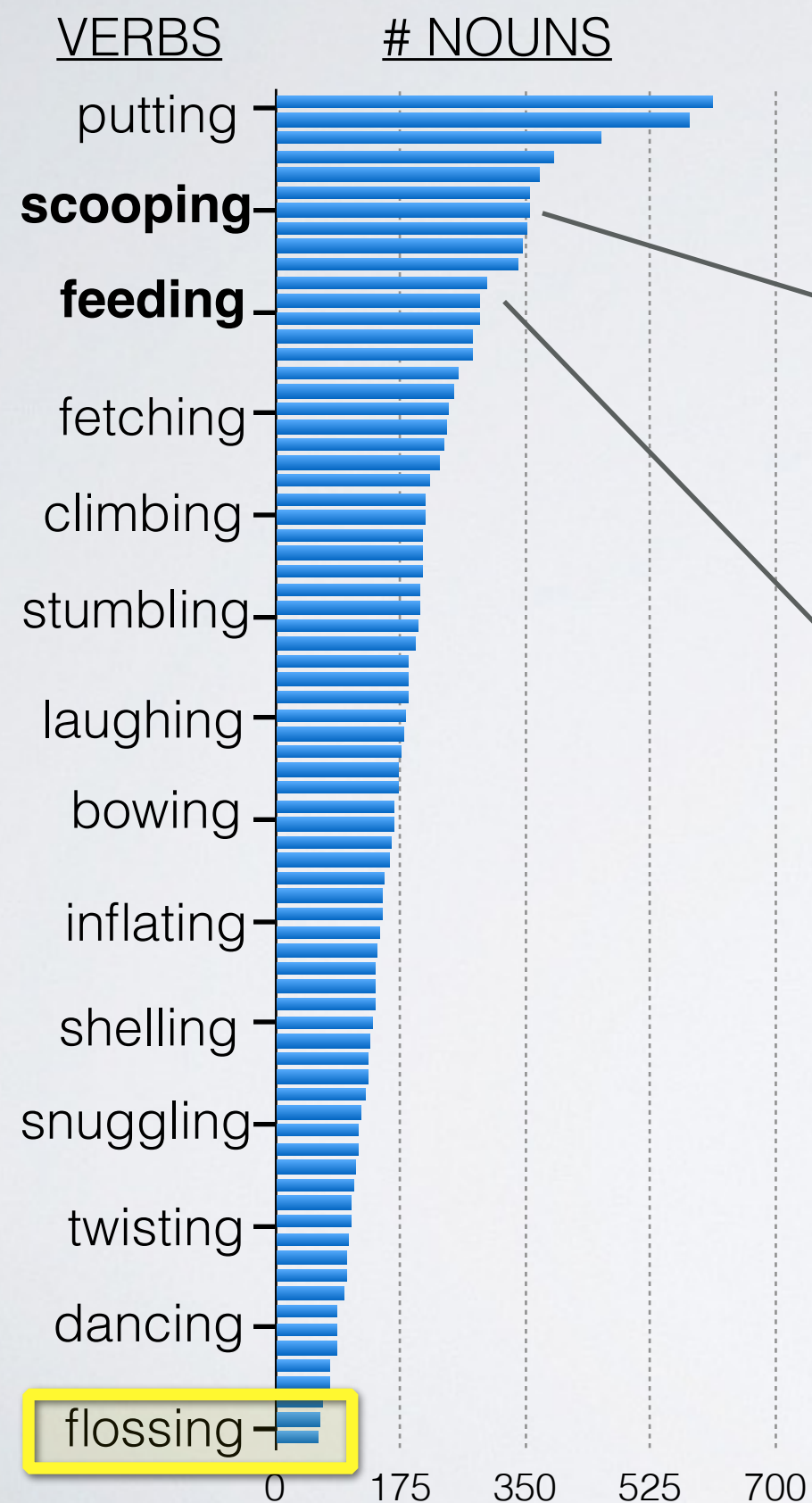
Verbs	504
Images	126,102
Situation / Image	3
Roles (types)	1,788 (190)
Nouns (≥ 3)	11,538 (6,794)
Annotations	1,481,851
Images / Verb	200-400
Uniq. situations (≥ 3)	205,095 (21,505)

Despite 80,000 possible values, 2 / 3 annotators on 76.8% of role-value

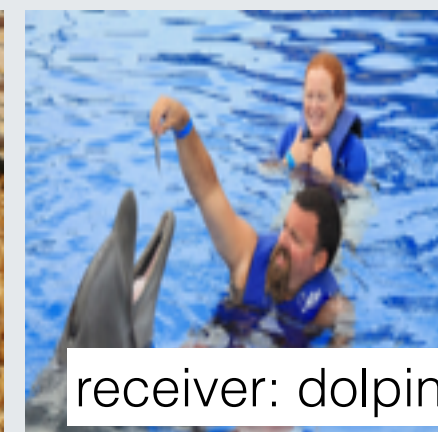
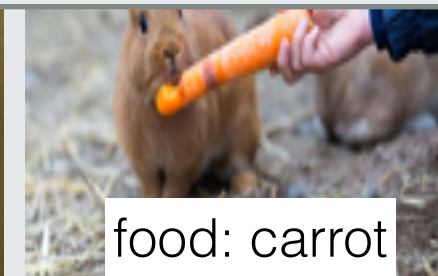
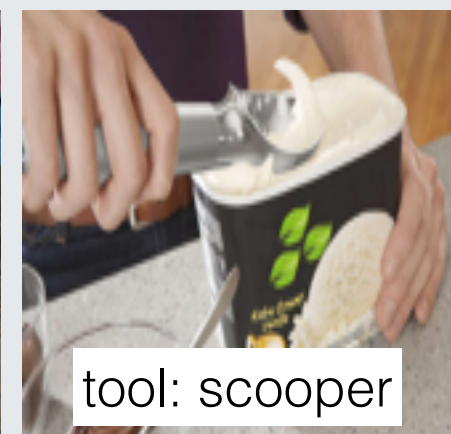
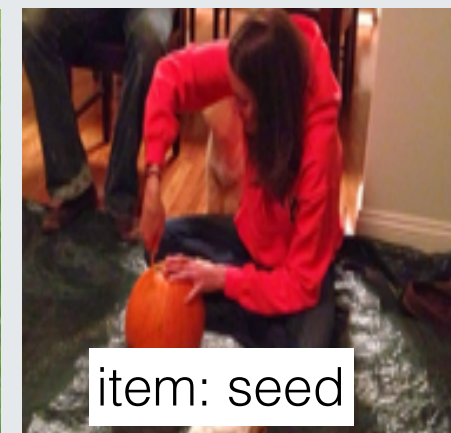
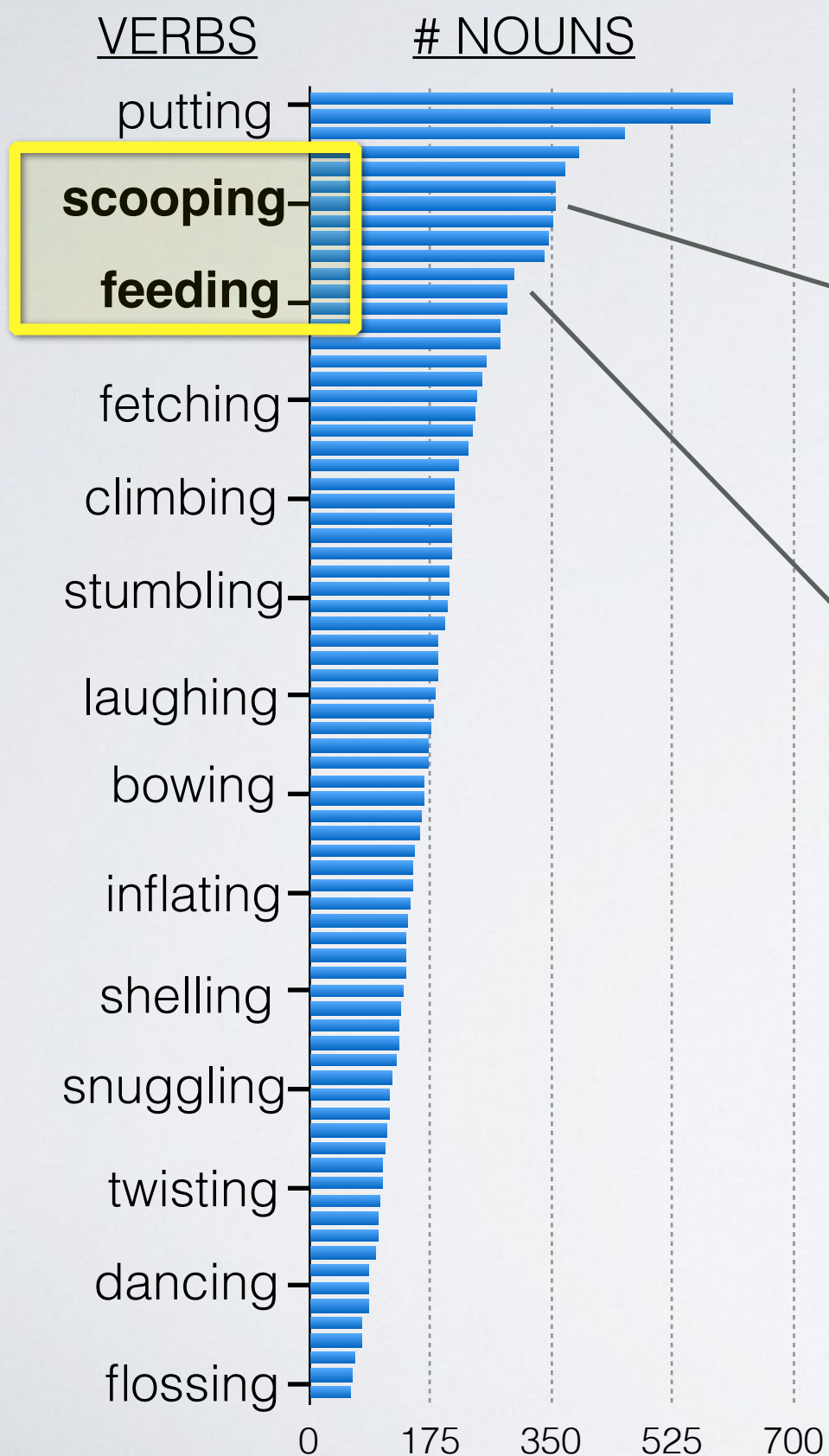
Skew - not all verbs are equal



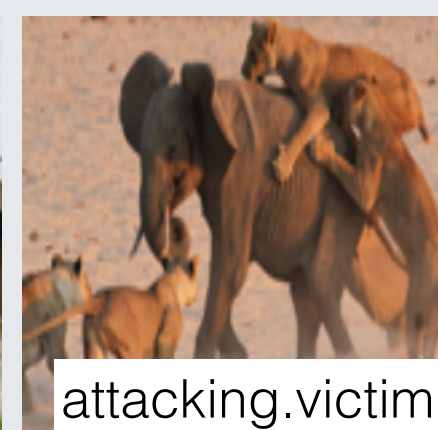
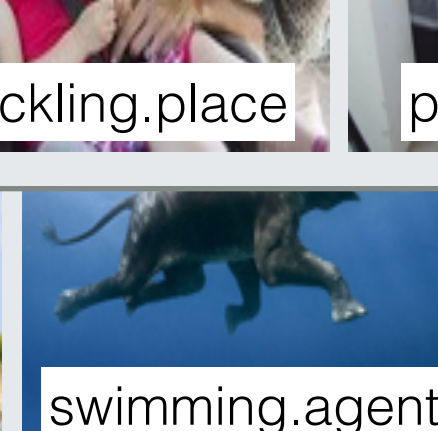
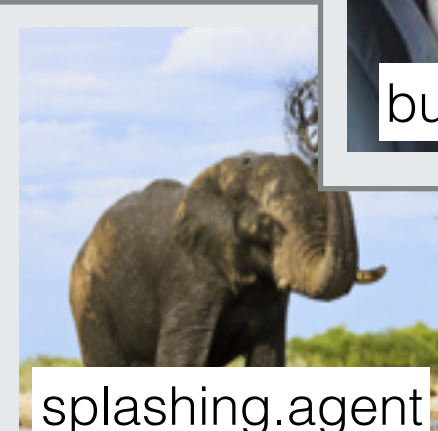
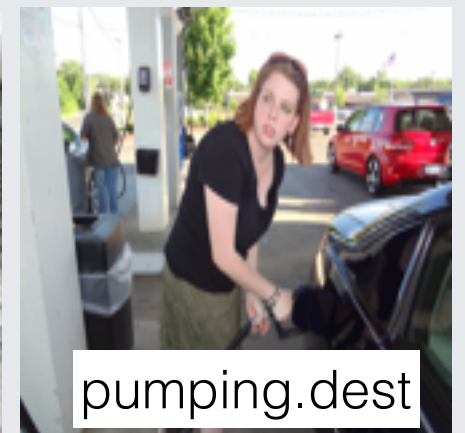
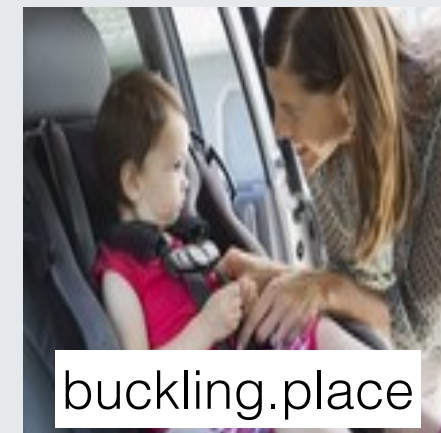
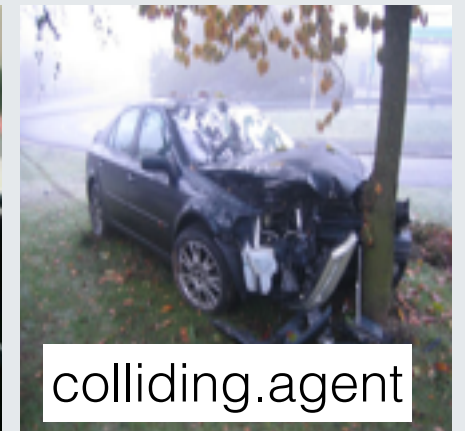
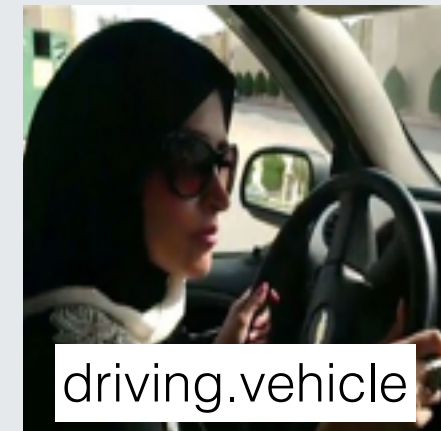
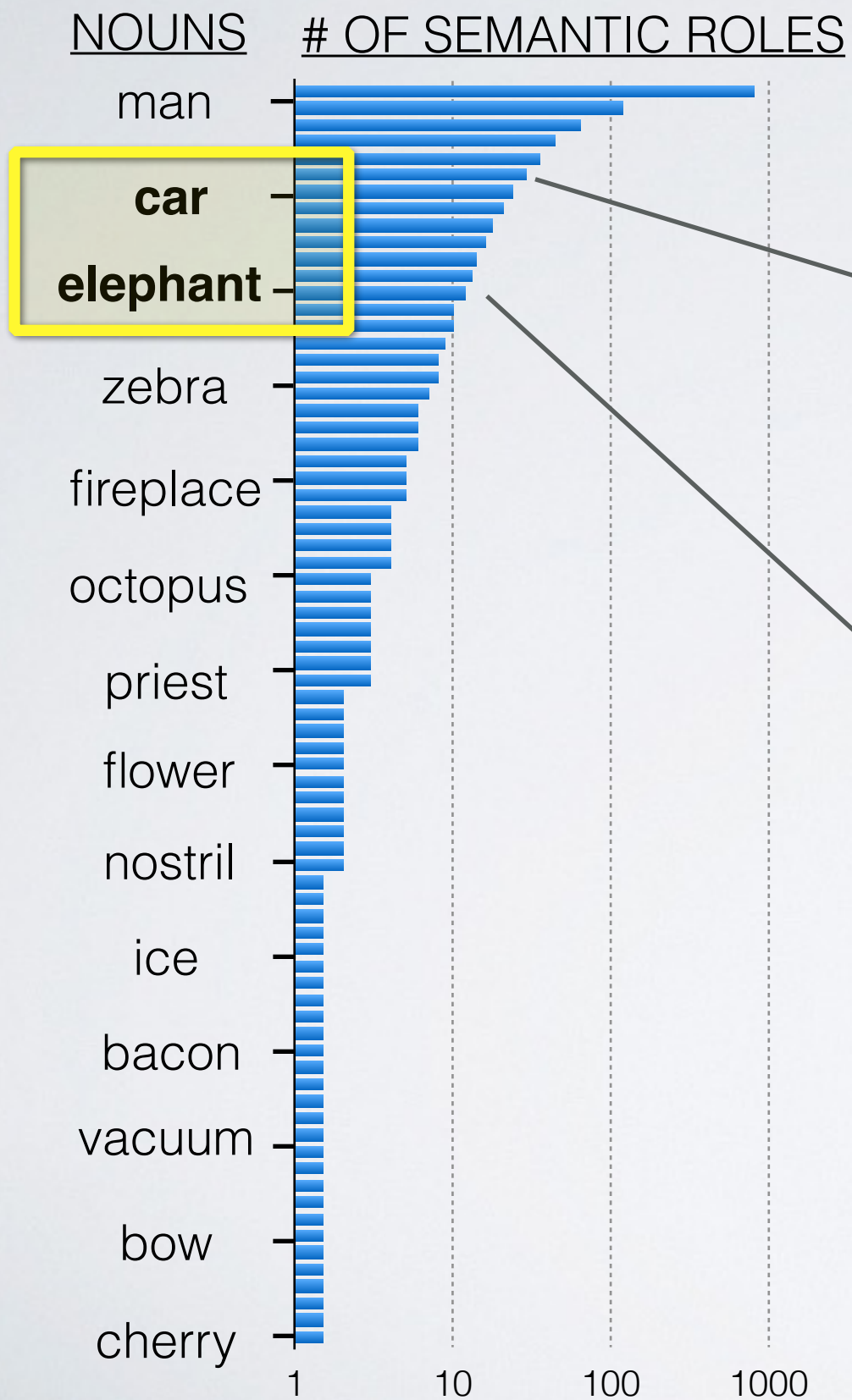
Skew - not all verbs are equal



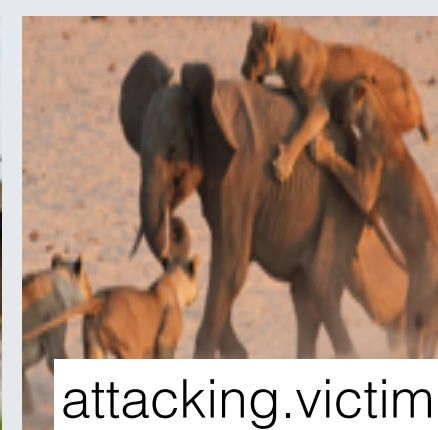
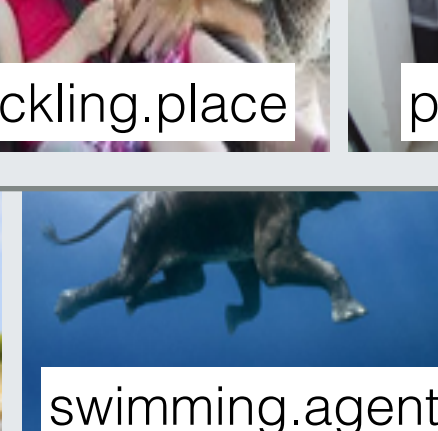
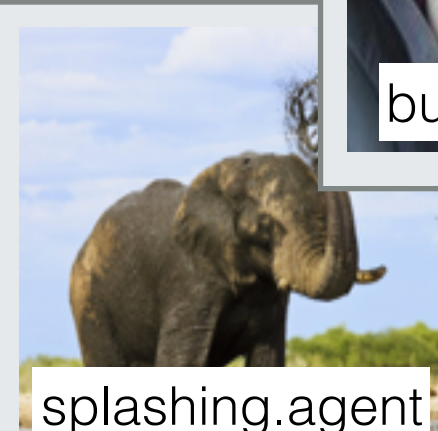
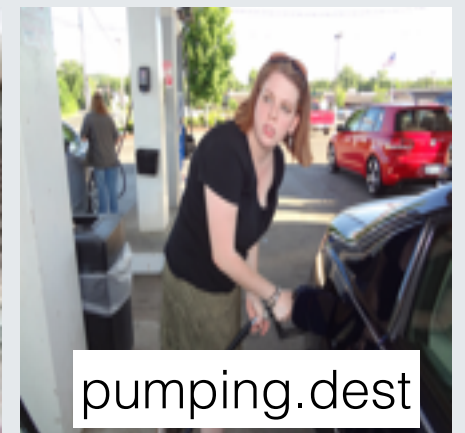
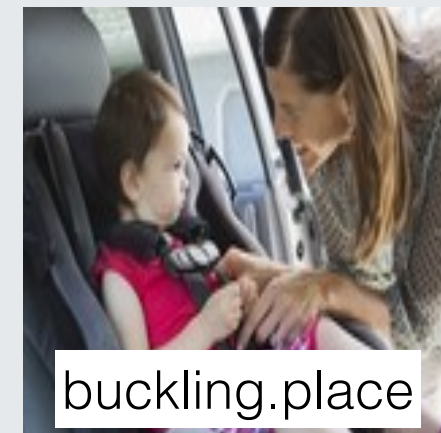
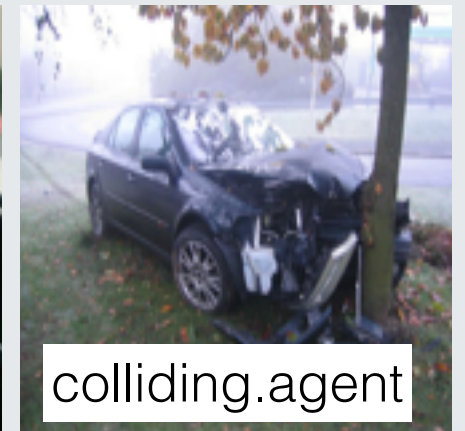
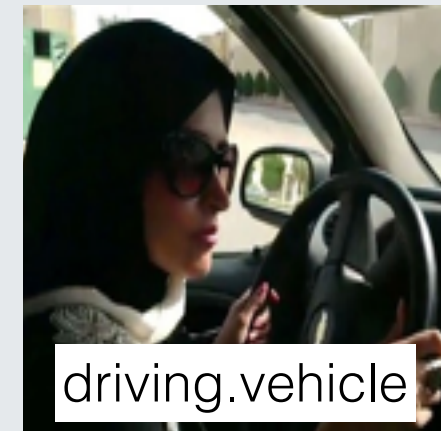
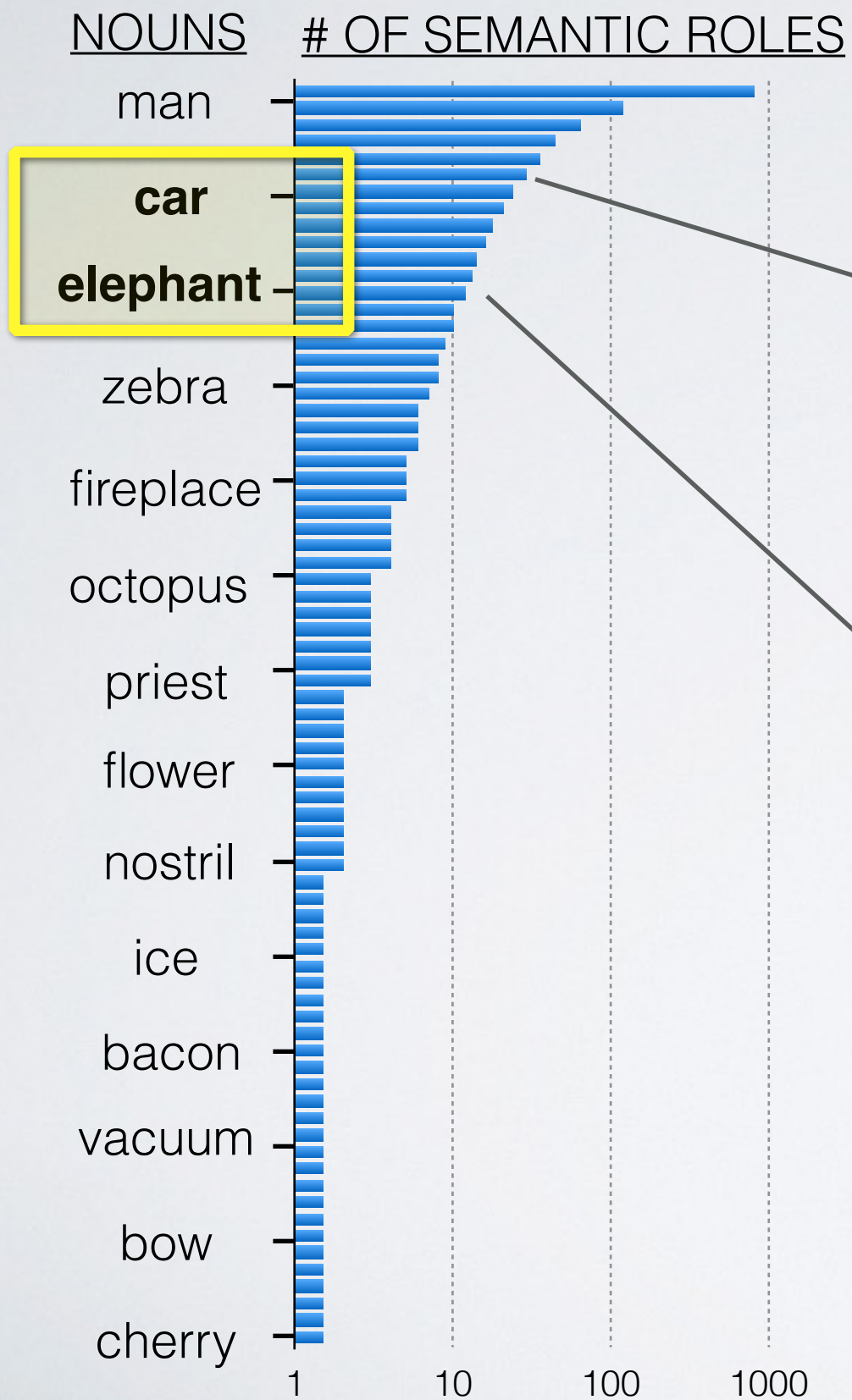
Skew - not all verbs are equal



Skew - not all nouns are equal



Skew - not all nouns are equal



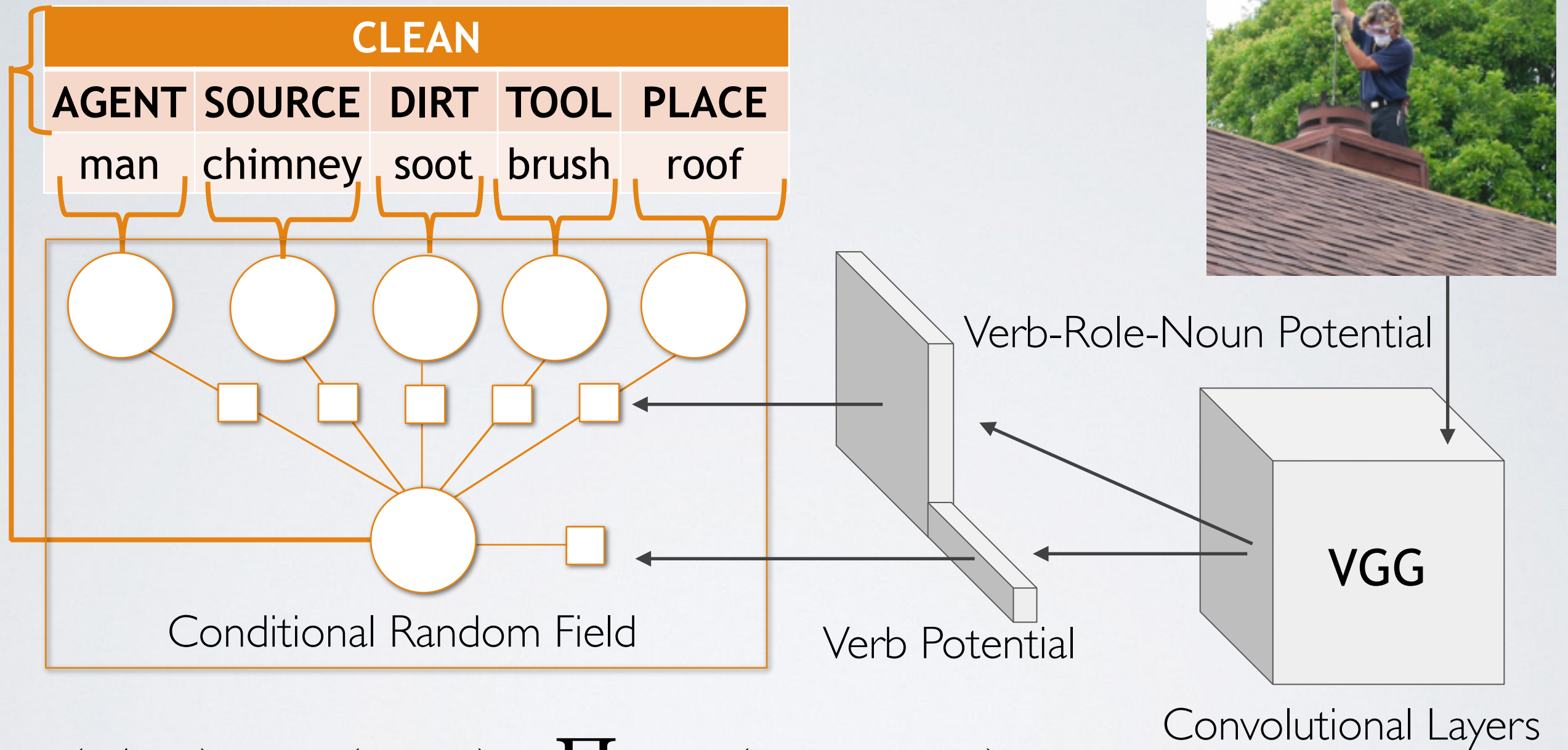
Situation Recognition

Models, Evaluation and Basic Results

structure matters

situation recognition improves object and activity recognition

Neural Conditional Random Field



$$p(S|i; \theta) \propto \psi_v(v, i; \theta) \prod_{(r, n_r) \in F} \psi_e(v, r, n_r, i; \theta)$$

Backpropagate CRF loss through VGG

Qualitative Examples

Gold Correct Incorrect



SWIMMING		
AGENT	SNAKE	SNAKE
PLACE	OCEAN	OCEAN

SPEARING		
AGENT	PERSON	PERSON
VICTIM	FISH	FISH
PLACE	OCEAN	OCEAN

FALLING		
AGENT	PERSON	PERSON
SOURCE	HORSE	HORSE
DEST.	GRND.	GRND.
PLACE	FIELD	FIELD

Qualitative Examples

Gold Correct Incorrect



SHAVING		
AGENT	MAN	PERSON
CO-AGENT	MAN	MAN
BODYPART	HEAD	HEAD
SUBSTANCE		S. CREAM
TOOL	RAZOR	RAZOR
PLACE	INSIDE	INSIDE

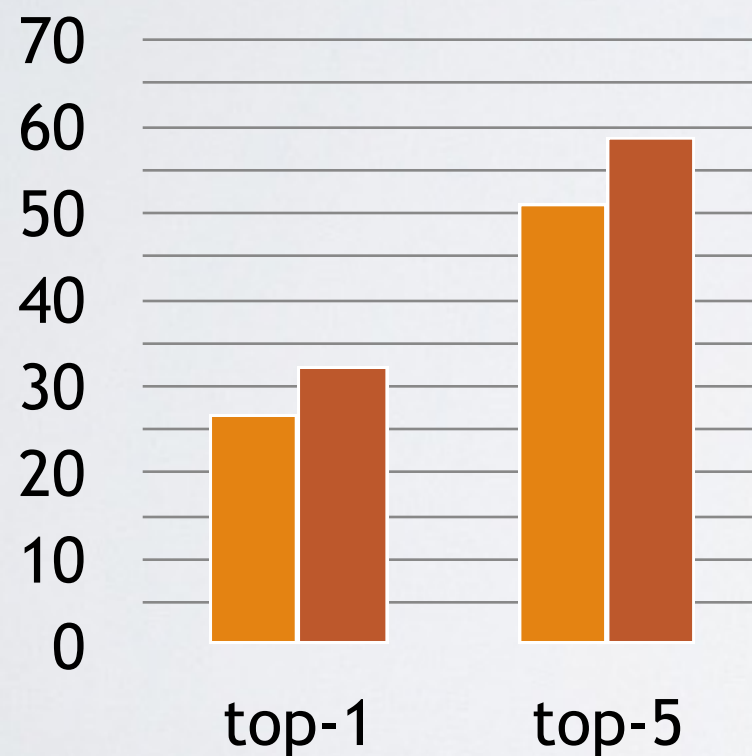
DETAINING	
AGENT	SOLDIER
VICTIM	MAN
PLACE	OUTSIDE

GIVING	
AGENT	SOLDIER
RECIPIENT	GIRL
ITEM	BAG
PLACE	OUTSIDE

Quantitative : Structured Prediction Crucial

FIXING				
AGENT	OBJECT	PART	TOOL	PLACE
BOY	CAR	TIRE	TIRE IRON	OUTDOORS

Verb



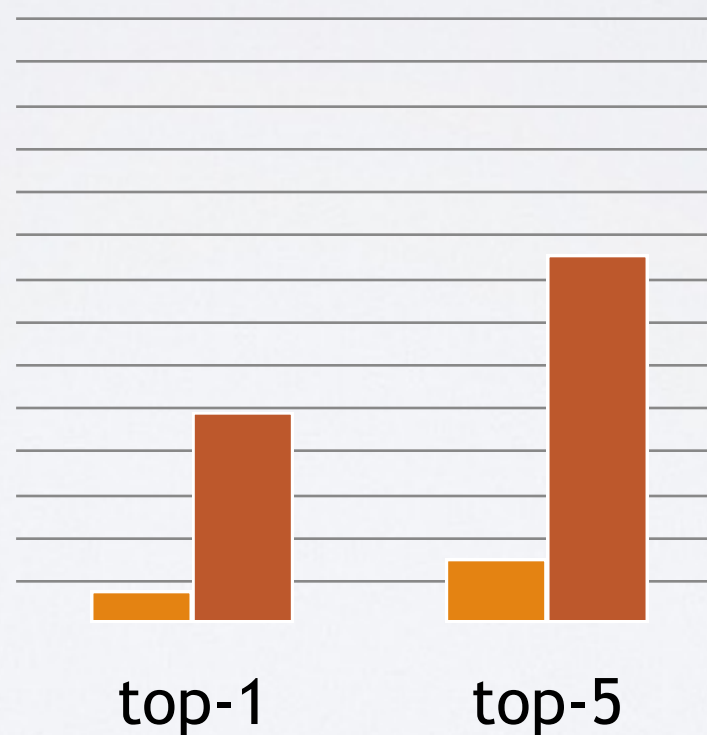
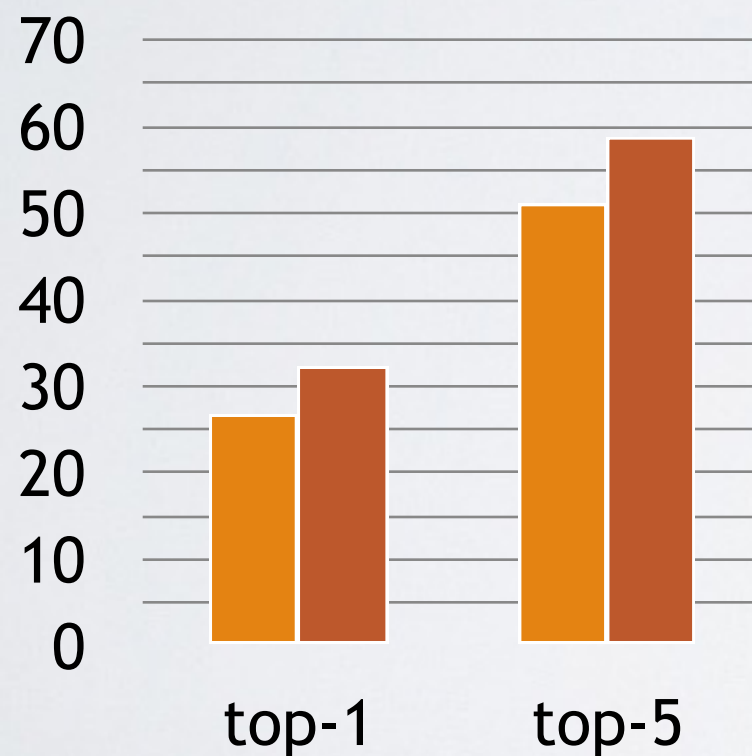
- Baseline: 5040-way CNN Predictor (10 most frequent situation/verb)
- Situation CRF

Quantitative : Structured Prediction Crucial

FIXING				
AGENT	OBJECT	PART	TOOL	PLACE
BOY	CAR	TIRE	TIRE IRON	OUTDOORS

Verb

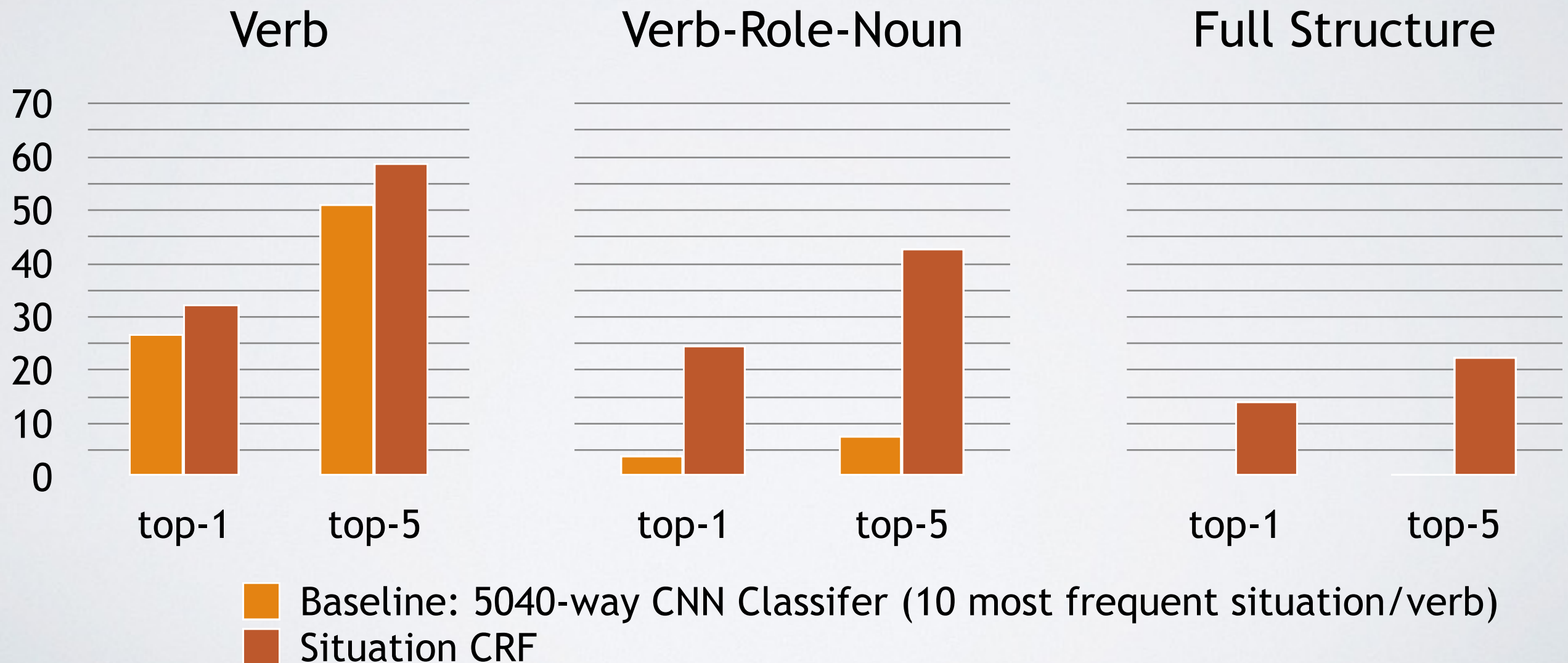
Verb-Role-Noun



- Baseline: 5040-way CNN Predictor (10 most frequent situation/verb)
- Situation CRF

Quantitative : Structured Prediction Crucial

FIXING				
AGENT	OBJECT	PART	TOOL	PLACE
BOY	CAR	TIRE	TIRE IRON	OUTDOORS



Generalize to Unseen Combinations

Train



FEEDING	
AGENT	MAN
EATER	BABY
FOOD	MILK
SOURCE	BOTTLE
PLACE	ROOM

Instances in train : 35



FEEDING	
AGENT	GIRL
EATER	HORSE
FOOD	CARROT
SOURCE	HAND
PLACE	PEN

Instances in train : 7

Test

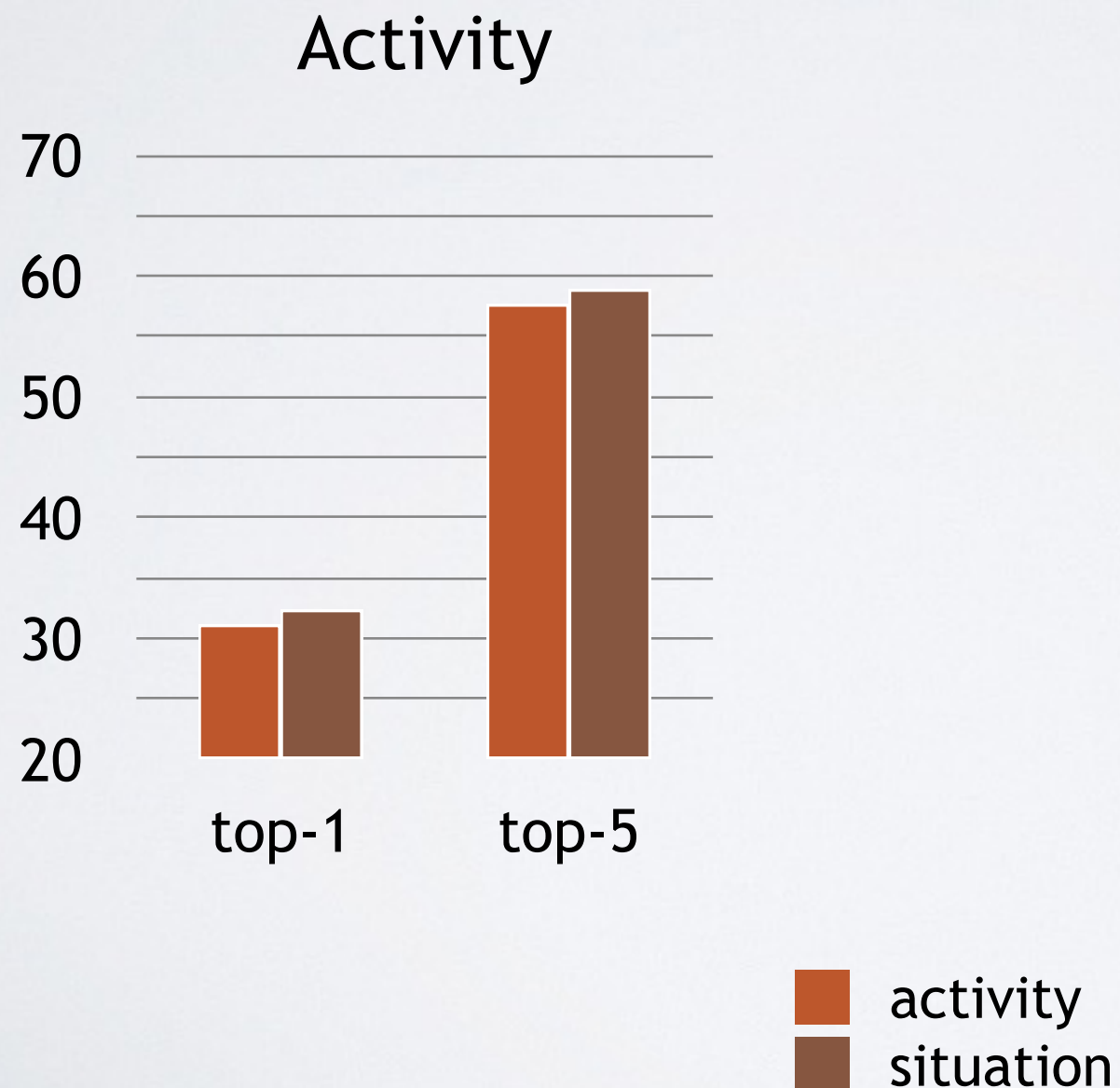


FEEDING	
AGENT	WOMAN
EATER	HORSE
FOOD	MILK
SOURCE	BOTTLE
PLACE	BARN

Instances in train : 0

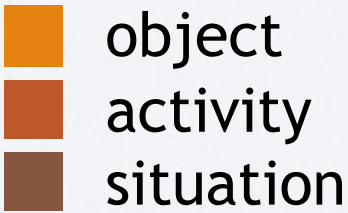
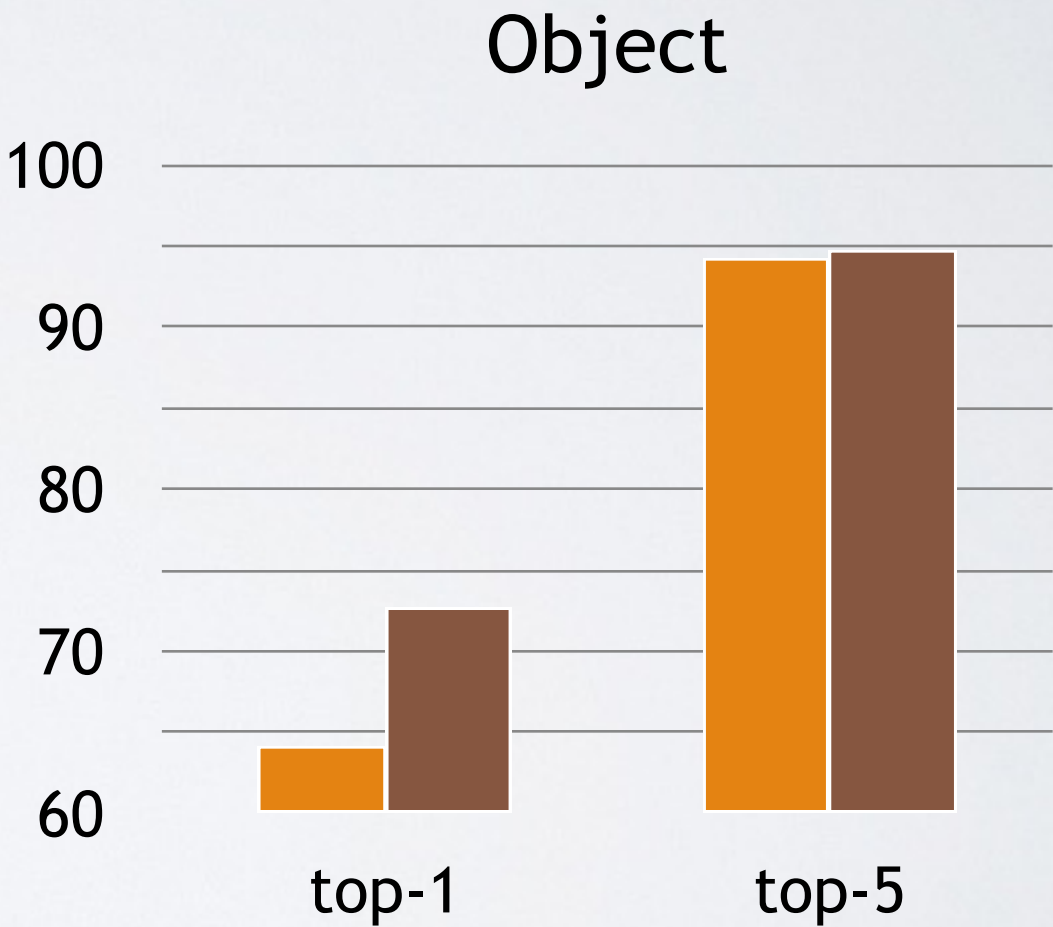
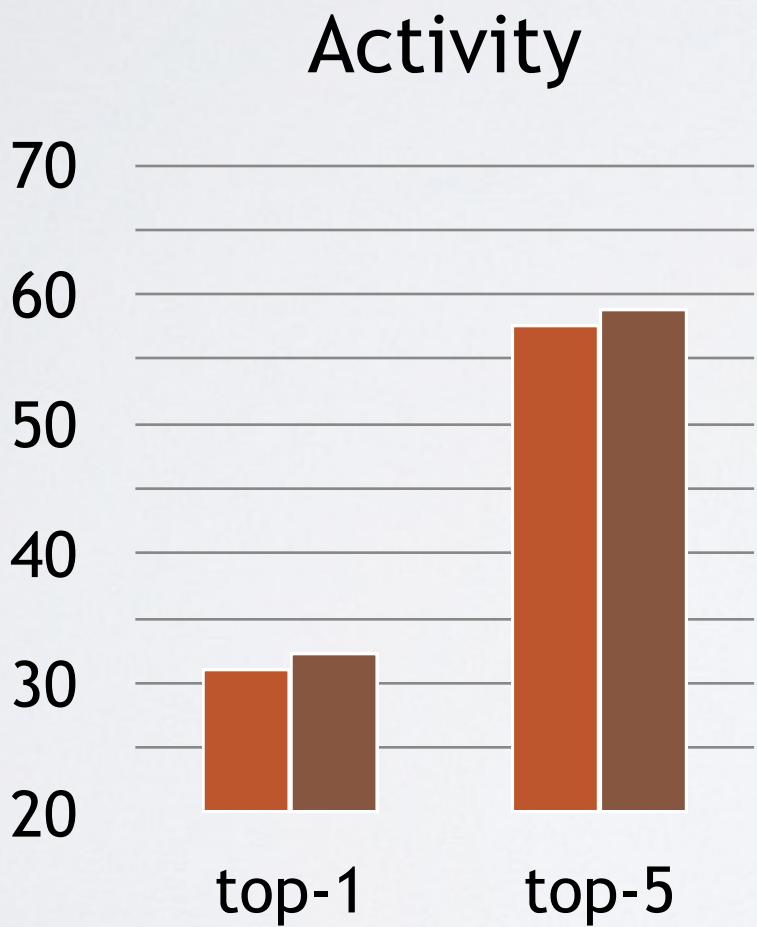
Situations Improves Object and Activity Recognition

FIXING				
AGENT	OBJECT	PART	TOOL	PLACE
BOY	CAR	TIRE	TIRE IRON	OUTDOORS



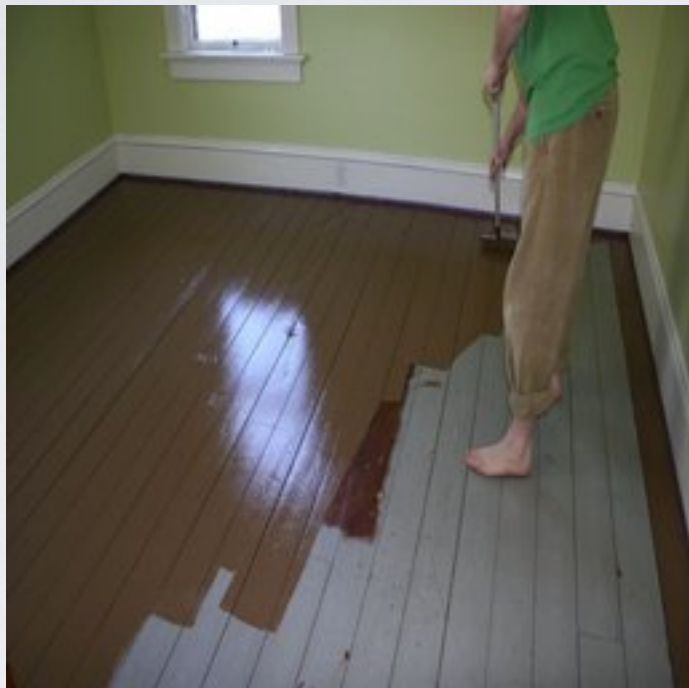
Situations Improves Object and Activity Recognition

FIXING				
AGENT	OBJECT	PART	TOOL	PLACE
BOY	CAR	TIRE	TIRE IRON	OUTDOORS



Errors

PAINTING



PRYING

AGENT	PERSON
ITEM	WOOD
SOURCE	FLOOR
TOOL	CROWBAR
PLACE	ROOM

SPRAYING



PUMPING

AGENT	PERSON
ITEM	AIR
SOURCE	AIR
DESTINATION	WHEEL
TOOL	PUMP
PLACE	OUTSIDE

imsitu.org

data/browsing/demo/code

Instructions

In the form below, you can search for images that match by verbs, nouns or both. All boxes are dropdowns that contain words from imSitu. As you select words, the dropdowns will update based on the number of remaining matching images. Click [show results](#) to retrieve images. Matched values are in red. Results are retrieved grouped, indicated by green. Click on images to cycle through their annotations, or click [more](#) to drill deeper.

Try these examples

what do horses do?
what do babies do?
what can be carried on the head?
what body parts get shaved?
how do babies and men interact?

what do cars do?
what gets chased?
what tools get used when fixing?
what got poured into a glass?
where does one end after jumping?

Search

Verb

chasing		
agent	chasee	place
noun	noun	noun

group

ungroup

group

Noun

noun
noun
noun

Clear Form

Show Results

281 matching images, 75 groups grouped by chasee

Results

Retrieved 281 chasing images, 75 groups grouped by chasee (showing 50)

chasee : dog

3 of 24 (more chasee : dog grouped by agent)



chasing		
agent	chasee	place
dog	dog	beach

chasee : man

3 of 23 (more chasee : man grouped by agent)



chasing		
agent	chasee	place
dog	man	outside

Recognize Situations

This demo used the imSitu dataset to train a neural CRF to predict situations, as described in this [paper](#). Try these examples...



Classify URL

Or upload an image to recognize:

Choose File

No file chosen



Was image in the imSitu train set?
Similar images from train



feeding					0.02935
agent	food	source	eater	place	
man	milk	bottle	lamb	o	
tickling					0.00039
agent	tickled		object	place	
man	child		hand	sofa	
pating					0.00009
agent	item	tool		place	
woman	dog	hand		outside	
checking					0.00006
agent	patient	aspect	tool	place	
woman	dog	o	stethoscope	o	
encouraging					0.00003
agent	reciever			place	
man	male child			outdoors	
rubbing					0.00003
agent	item	agent/part		place	
woman	dog	hand		outside	
giving					0.00002
agent	item	recipient		place	
woman	bottle	baby		outdoors	

Conclusion

Introduced **situation recognition**

role-centric structured representation of whats happening

Collected **imSitu**

120k+ images, 500+ verbs, 100k+ situations

Introduced simple model **neural CRF** for situation
structure matters

provides strong context for activity and object recognition

data/browsing/demo/code

imsitu.org