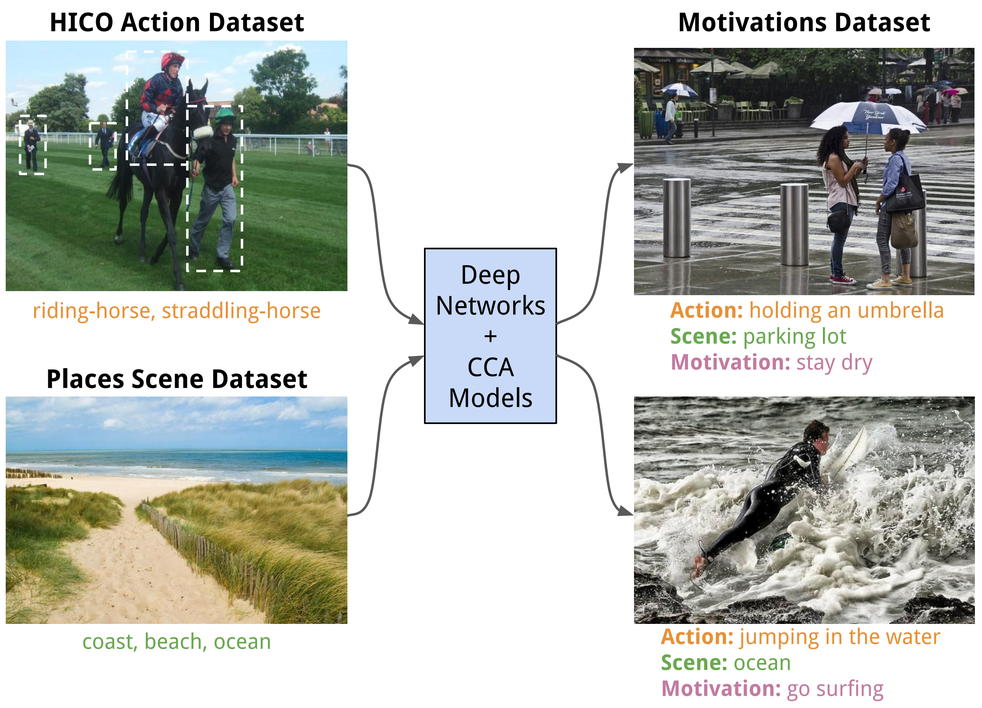
High-level Cues for Predicting Motivations

Arun Mallya and Svetlana Lazebnik
University of Illinois at Urbana-Champaign

vision.cs.illinois.edu/go/motivations

Overview



- We propose a framework that utilizes **action** and **scene** cues with Canonical Correlation Analysis (CCA) models to predict high-level "**motivations**" of humans in images [1].
- The task is to retrieve a suitable Action, Scene, and Motivation sentence for each image, from the pool of corresponding sentences present in the training set.
- Prior work [1] trained a Structured-SVM with generic VGG-16 fc7 cues from the whole image and person box, as well as large-scale Language Model features.

Our Method:

Action Predictions (600-d) are obtained from the Fusion model [3] trained on the HICO person-object interaction dataset [2].

Scene Predictions (365-d) are obtained from a VGG-16 network trained on the Places scene recognition dataset [4].

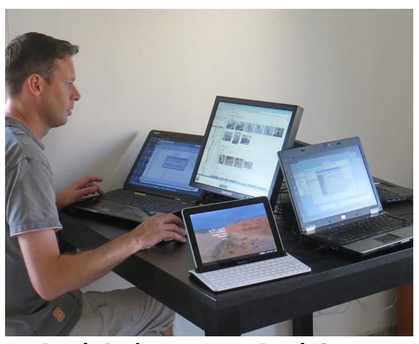
Image Embeddings are (600+365=**965-d**) vectors v/s **8192-d** vectors in [1].

Sentence Embeddings are (4800-d) skip-thought vectors.

3 separate Image-Sentence CCA models with 300-d embeddings are trained, one for each type of sentence, and choices are ranked on cosine similarity in the embedding space.

Train/Val/Test set of size 6133/1532/2526: Small dataset ⇒ good performance by simple CCA

Results


Ground Truth
Action: looking at computer screen
Scene: home office
Motivation: do work

Prediction
Action: typing on a computer
Scene: meeting
Motivation: work

Pred. Actions: read-laptop, type-on-laptop, hold-laptop
Pred. Scenes: computer-room, home-office, physics-lab


Ground Truth
Action: bottle
Scene: zoo
Motivation: feeding a baby giraffe

Prediction
Action: feeding a giraffe
Scene: zoo
Motivation: care for giraffe

Pred. Actions: watch-giraffe, feed-giraffe, pet-giraffe
Pred. Scenes: stable, rodeo-arena, barndoor


Ground Truth
Action: skiing
Scene: ski resort
Motivation: get down the hill

Prediction
Action: skiing
Scene: resort
Motivation: have fun

Pred. Actions: wear-skis, stand-on-skis, ride-skis
Pred. Scenes: ski-slope, igloo, ski-resort


Ground Truth
Action: using a computer
Scene: room
Motivation: entertain himself

Prediction
Action: looking at the laptop
Scene: living room
Motivation: read the screen

Pred. Actions: read-laptop, type-on-laptop, hold-laptop
Pred. Scenes: computer-room, restaurant, office


Ground Truth
Action: riding a bike
Scene: city
Motivation: go to work

Prediction
Action: riding a bike
Scene: city
Motivation: go to work

Pred. Actions: straddle-bicycle, ride-bicycle, hold-bicycle
Pred. Scenes: crosswalk, street, promenade


Ground Truth
Action: standing on skateboard
Scene: parking lot
Motivation: ride the skateboard

Prediction
Action: riding a skateboard
Scene: parking lot
Motivation: learn to do jumps

Pred. Actions: ride-skateboard, stand-on-skateboard, throw-frisbee
Pred. Scenes: skating-outdoor, skating-indoor, parking-lot

Mistakes in Motivation / Cue Prediction


Ground Truth
Action: jumping
Scene: city sidewalk
Motivation: click their heels

Prediction
Action: pulling luggage
Scene: city
Motivation: walk

Pred. Actions: carry-suitcase, flip-skateboard, hold-suitcase
Pred. Scenes: outdoor, plaza, ice-skating-rink


Ground Truth
Action: holding an orange
Scene: room
Motivation: look at it

Prediction
Action: holding a donut
Scene: indoor area
Motivation: eat it

Pred. Actions: hold-orange, hold-apple, eat-apple
Pred. Scenes: aquarium, underwater, sky


Ground Truth
Action: preparing food
Scene: kitchen
Motivation: serve food

Prediction
Action: stirring a pot
Scene: kitchen
Motivation: cook food

Pred. Actions: no-interaction-oven, cook-pizza, open-microwave
Pred. Scenes: restaurant-kitchen, galley, kitchen


Ground Truth
Action: on a laptop
Scene: living room
Motivation: pass her time

Prediction
Action: typing on laptop
Scene: computer store
Motivation: show off a laptop

Pred. Actions: no-interaction-oven, cook-pizza, open-microwave
Pred. Scenes: restaurant-kitchen, galley, kitchen

Method	Motivation			Action			Scene		
	Median Rank	Recall @1	Recall @10	Median Rank	Recall @1	Recall @10	Median Rank	Recall @1	Recall @10
Sentences replaced with cluster centers — #Clusters (Motivation, Action, Scene): (256, 100, 100)									
SSVM - Image fc7 [1]	39	—	—	17	—	—	4	—	—
SSVM - Image + Person fc7 [1]	42	—	—	18	—	—	4	—	—
SSVM - Image + Person fc7 + Language Model [1]	28	—	—	14	—	—	**3**	—	—
CCA - ImageNet VGG fc7	19	9.0	38.8	11	14.5	49.0	3	35.2	72.2
CCA - Action & Scene cues	**14**	**12.0**	**44.9**	**8**	**18.6**	**56.3**	**3**	**33.4**	**73.8**
Sentences used as-is — #Sentences: 2526									
CCA - ImageNet VGG fc7	171	0.9	7.7	187	1.3	9.5	116	**1.1**	7.9
CCA - Action & Scene cues	**130**	**1.4**	**12.0**	**117**	**1.8**	**13.2**	**113**	1.0	**8.8**

Results for retrieval of correct sentence on the test set of the Motivations dataset [1].
Lower scores are better for Median Rank, higher scores are better for Recall@x.

Key Takeaways:

- High-level cues such as actions and scenes provide a compact image representation which is more informative than generic VGG features and useful for solving high-level tasks.
- High-level cues increase model interpretability and allow us to diagnose and understand errors.
- In spite of advances in architectures and datasets for action and scene recognition, there is clearly scope for improving predictions.

References

1. C. Vondrick, D. Oktay, H. Pirsiavash, and A. Torralba, Predicting motivations of actions by leveraging text. In CVPR, 2016.
2. Y.-W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng, Hico: A benchmark for recognizing human-object interactions in images. In ICCV, 2015.
3. A. Mallya and S. Lazebnik, Learning models for actions and person-object interactions with transfer to question answering. In ECCV, 2016
4. B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, Learning deep features for scene recognition using places database. In NIPS, 2014.