Action-Induced Object Detection

NRI: FND: Towards Scalable and Self-Aware Robotic Perception

PI: Nuno Vasconcelos, UC San Diego

https://www.nsf.gov/awardsearch/showAward?AWD_ID=1924937

Introduction

- **We define action-inducing object detection**

- simultaneous prediction of allowable actions and detection of objects that constrain those actions

- **Given a complex scene**

- rather than all objects in the scene, detect only those important to driving decisions

- these are action-inducing objects: pedestrians, traffic lights, cars on the road, etc.

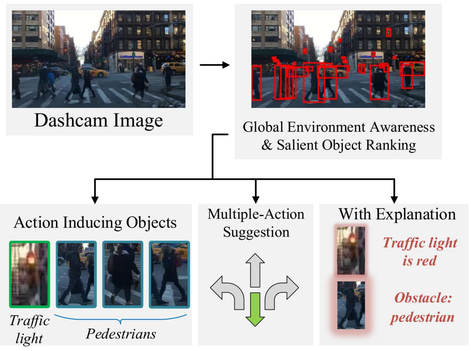
- also, predict the actions allowable

- **Benefits**

- less complexity

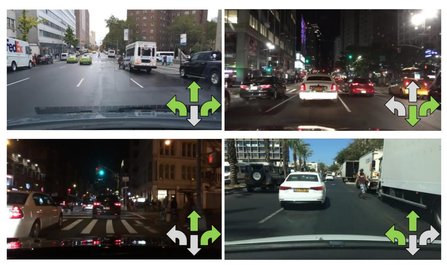
- higher prediction accuracy

- finite explanation vocabulary, e.g. “slow down because the light is red and there are pedestrians crossing”



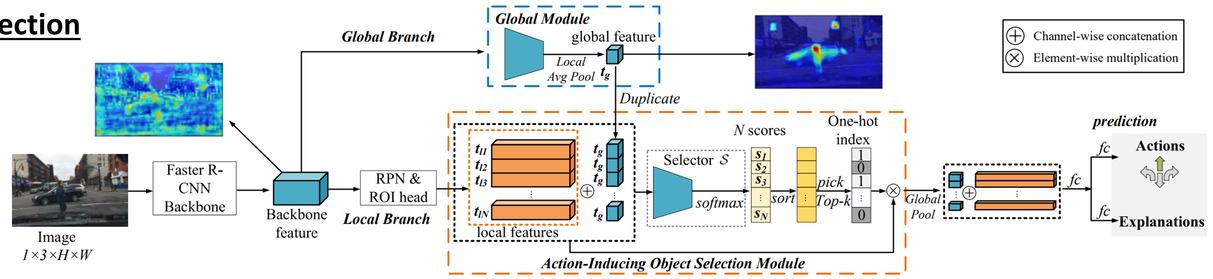
BDD-OIA Dataset

- based on BDD100K
- only complex scenes
- 8 pedestrians, 12 vehicles per scene on avg
- 4 action categories
- 21 possible explanations
- labeled for feasible actions, and their explanations



Deep Learning Architecture for AIO Detection

- Faster R-CNN backbone
- Global processing branch to model context
- Localized processing branch to detect objects
- Selection module to identify AIOs according to
- scene context
- Prediction heads for actions and explanations



Learning Details

- **Training with a multi-task loss for action and explanation prediction**

$$\mathcal{L} = \mathcal{L}_A + \lambda \mathcal{L}_E$$

- **Global Module**

- generates global features t_g from the Faster R-CNN backbone features

- It is composed of two convolutional layers with ReLU activation functions plus a local average pooling operation

- **Action-Inducing Object Selection Module**

- pick action-inducing objects from all object proposals produced by the Faster R-CNN

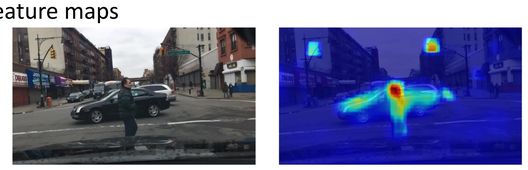
- N local feature tensors t_{l_i} are first extracted from the proposal locations and concatenated with the global feature t_g to form an object-scene tensor per object

- A selector S then chooses the action-inducing objects from these tensors

- **Behind these**

- The combination of local and global features and end-to-end supervision enables the network to reason about scene-object relationships

- produce a global feature map more selective of action-inducing objects than the backbone feature maps



Experiments

- **Interplay between actions and explanations**

- Action prediction performance improves with the quality of the explanations.

- The requirement to justify why actions can be taken makes the system more accurate in the prediction of which actions to take.

- Conversely explanations also benefit from action prediction.

λ	F	S	L	R	action mF1	action F1$_{all}$	explanation F1$_{all}$
0	0.783	0.758	0.419	0.568	0.632	0.675	-
0.01	0.819	0.760	0.504	0.605	0.672	0.696	0.329
0.1	0.784	0.769	0.562	0.627	0.686	0.709	0.371
1.0	**0.829**	**0.781**	**0.630**	**0.634**	**0.718**	**0.734**	**0.422**
∞	-	-	-	-	-	-	0.418

Experiments

- **Interplay between local and global branch**

- Global context is important to reason actions (A) and explanations (X)

- The two-branch model performs better even with random object selection

- The proposed AIO module highly improves the performance

models	F	S	L	R	A mF1	A F1$_{all}$	X mF1	X F1$_{all}$
only local branch	0.760	0.649	0.413	0.473	0.574	0.605	0.139	0.351
only global branch	0.820	0.777	0.499	0.621	0.679	0.704	0.206	0.419
random selection	0.823	0.778	0.499	0.637	0.685	0.709	0.197	0.413
select top-5	0.821	0.768	0.617	0.625	0.708	0.720	**0.212**	0.416
select top-10	**0.829**	**0.781**	**0.630**	**0.634**	**0.718**	**0.734**	0.208	**0.422**

Visualization

Boxes yellow: Faster R-CNN detected, red: action-inducing objects

Explanations green: correct, red: incorrect, gray: missing