Embodied Amodal Recognition: Learning to Move to Perceive Objects

Presenter: Karthik Velayutham

9/30/2021

Motivation

- State-of-the-art computer vision technologies are mature and effective in object recognition, segmentation, etc.
- Current robot systems train on static images and apply this model to the real world
- However, a robot's vision is drastically different from images used to train
  - There are a variety of factors that can affect how an image looks from a different angle

Main Problem

- Amodal recognition is intuitive to humans, not intuitive to machines
- Visual recognition tasks train on single images - difficult to recognize objects when there is a lot of collusion in a given image
- Current work has not approached the problem in the same angle - we already have some info about the object, we need to get more

Key Contributions

- A new task - Embodied Amodal Recognition where agent can move in a 3D space to recognize 2D objects
- New dataset and new model, Embodied Mask R-CNN (remember this?) used for amodal recognition
- Evaluate the methodology and show that agents with movement outperform passive systems
    - Learn some cool behavior about the optimal path chosen

Problem Setting

- **Three sub-tasks**. In EAR, we aim to recover both semantics and shape for the target object. EAR consists of three sub-tasks:
  - Object recognition
  - 2D amodal localization (a 2D bounding box enclosing the full extent of the object),
  - 2D amodal segmentation (a 2D mask enclosing the full shape of the object).
- **Single target object**. The agent's goal is to then move to perceive this single target object within bounding box
- **Predict for the first frame**. The agent performs amodal recognition for the target object observed at the spawning point. Both passive and embodied algorithms are trained using the same amount of supervision and evaluated on the same set of image.

Related Work

- Object Recognition:
  - T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick. Microsoft COCO: Com- ´ mon objects in context. In Proceedings of the European Conference on Computer Vision (ECCV), 2014
- Amodal Perception:
  - K. Ehsani, R. Mottaghi, and A. Farhadi. Segan: Segmenting and generating the invisible. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
- Active Vision:
  - J. Denzler and C. M. Brown. Information theoretic sensor data selection for active object recognition and state estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2002.
- Embodiment:
  - P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sunderhauf, I. Reid, S. Gould, and A. van den Hen- ¨ gel. Vision-and-Language Navigation: Interpreting visuallygrounded navigation instructions in real environments. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

Approach

Authors propose a new model called Embodied Mask R-CNN

- Amodal Recognition Module
  - Responsible for predicting the object category, amodal bounding box, and amodal mask at each navigational time step
- Learning To Move
  - Goal of the policy network is to propose the next moves in order to acquire useful information for amodal recognition

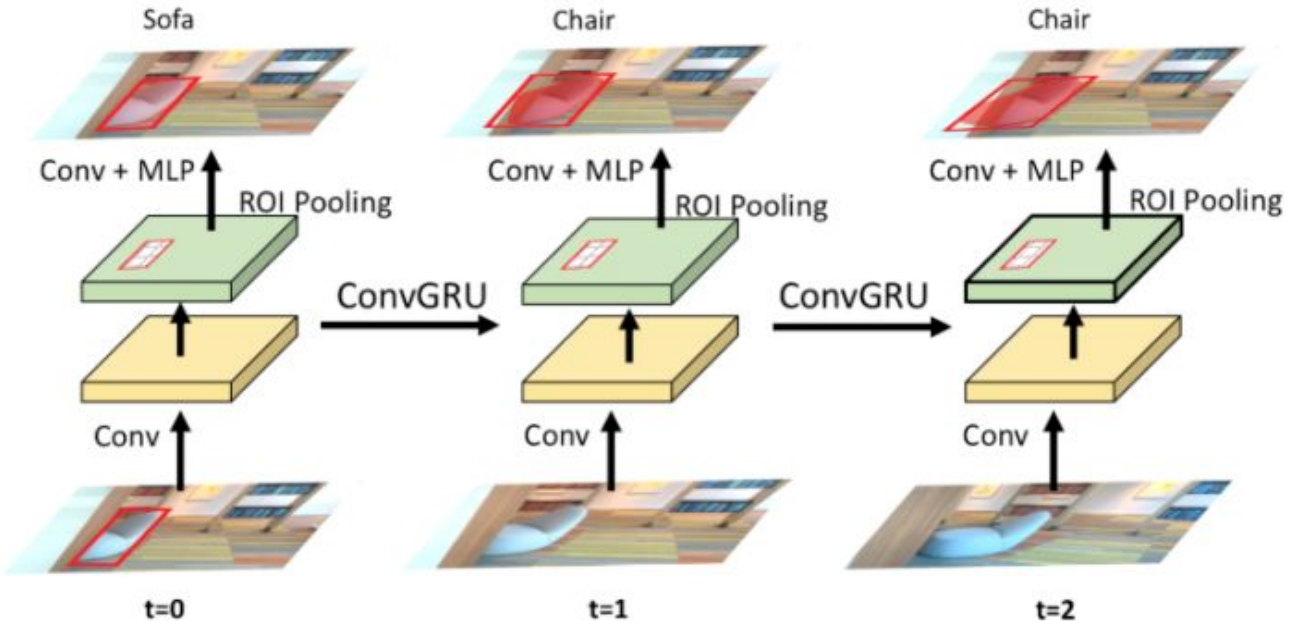
Amodal Recognition Module

$$\boldsymbol{y}_t = f(\boldsymbol{b}_0, I_0, I_1, ..., I_t). \tag{1}$$

$$\boldsymbol{y}_t = f_{\text{head}}(\boldsymbol{b}_0, \hat{\boldsymbol{x}}_t). \tag{2}$$

$$L^p = \frac{1}{T}\sum_{t=1}^{T}\Big[L_c^p(c_t, c^*) + L_b^p(\boldsymbol{b}_t, \boldsymbol{b}^*) + L_m^p(\boldsymbol{m}_t, \boldsymbol{m}^*)\Big], \tag{3}$$

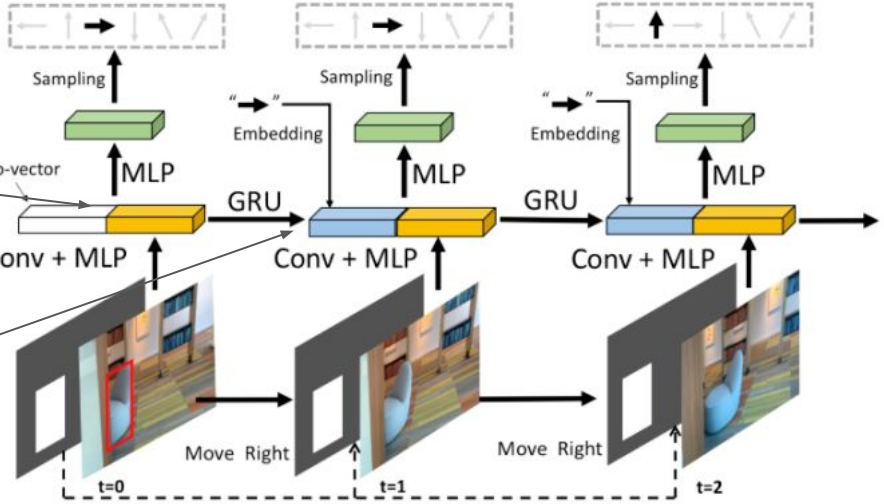
Amodal Recognition Module



Learning To Move

$$z_t^{\text{act}} = f_{\text{actEnc}}(a_{t-1})$$

$$z_t = f_{\text{act}}\left([z_t^{\text{img}}, z_t^{\text{act}}], h_{t-1}\right)$$



$$r_t = \lambda_c Acc_t^c + \lambda_b IoU_t^b + \lambda_b IoU_t^m, \quad (6)$$

$$R_t = r_t - r_{t-1}, \quad (7)$$

Experimental Setup

Dataset

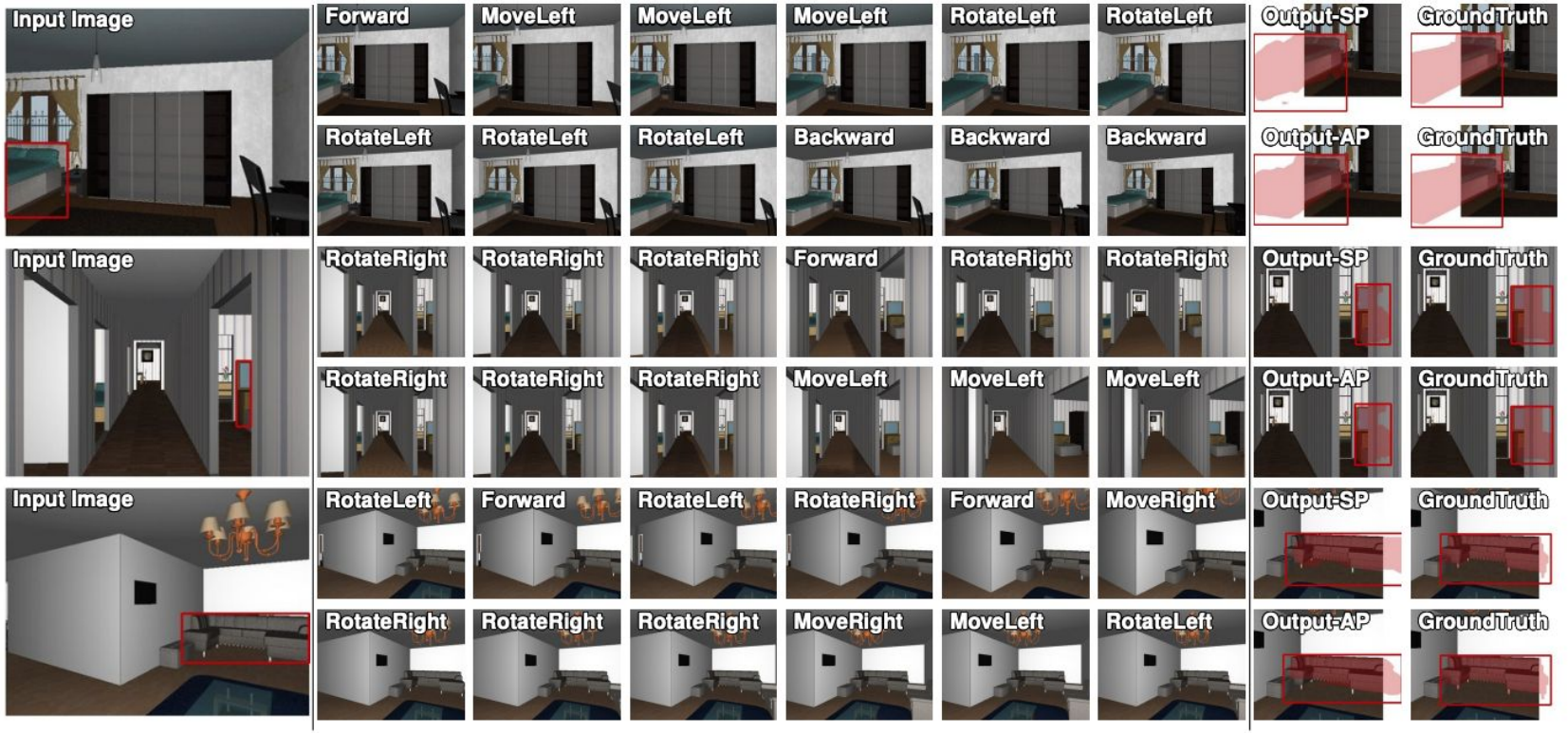
- Indoor simulation environment is used - use 550 houses that are the right size so, 400, 50, 100 for training, validation and test
- Render 640x800 images, and generate ground truth annotations for object category, amodal bounding boxes, and amodal masks
- 859/123/349 unique object instances (i.e., shapes) in the train/val/test set respectively, and 235 are shared by train and test sets
- Randomly sample spawning locations and viewpoints for the agent - 8940 instances in training set, 1113 in validation set, and 2170 in test set
- Agent can move forward, backward, left, and right. Can turn right and left by 2 degrees

Experimental Setup

Metrics and Baselines

- Evaluate the amodal recognition performance on the first frame in the moving path.
  - Object classification accuracy (Clss-Acc)
  - The IoU scores for amodal box (ABox-IoU)
  - Amodal mask (AMask-IoU)
  - Amodal segmentation only on the occluded region of the target object (AMask-Occ-IoU)
- Passive/Passive (PP/PP), ShortestPath/Passive (SP/PP), etc. compared with ShortestPath/ActivePath (SP/AP) and ActivePath/ActivePath (AP/AP)

Experimental Setup



Experimental Setup

Training Details

- Training amodal recognition
  - Temporal Mask R-CNN, based on the PyTorch implementation of Mask R-CNN
  - We use ResNet50 pre-trained from ImageNet and crop RoI features with a C4 head
  - SGD with learning rate 0.0025, batch size 8, momentum 0.99, and weight decay 0.0005
- Training action policy
  - RMSProp as optimizer, initial learning rate 0.00004, and set ϙ“0.00005. The agent moves 10 steps in total.
- Fine-tuning amodal recognition.
  - We use SGD, with learning rate 0.0005.

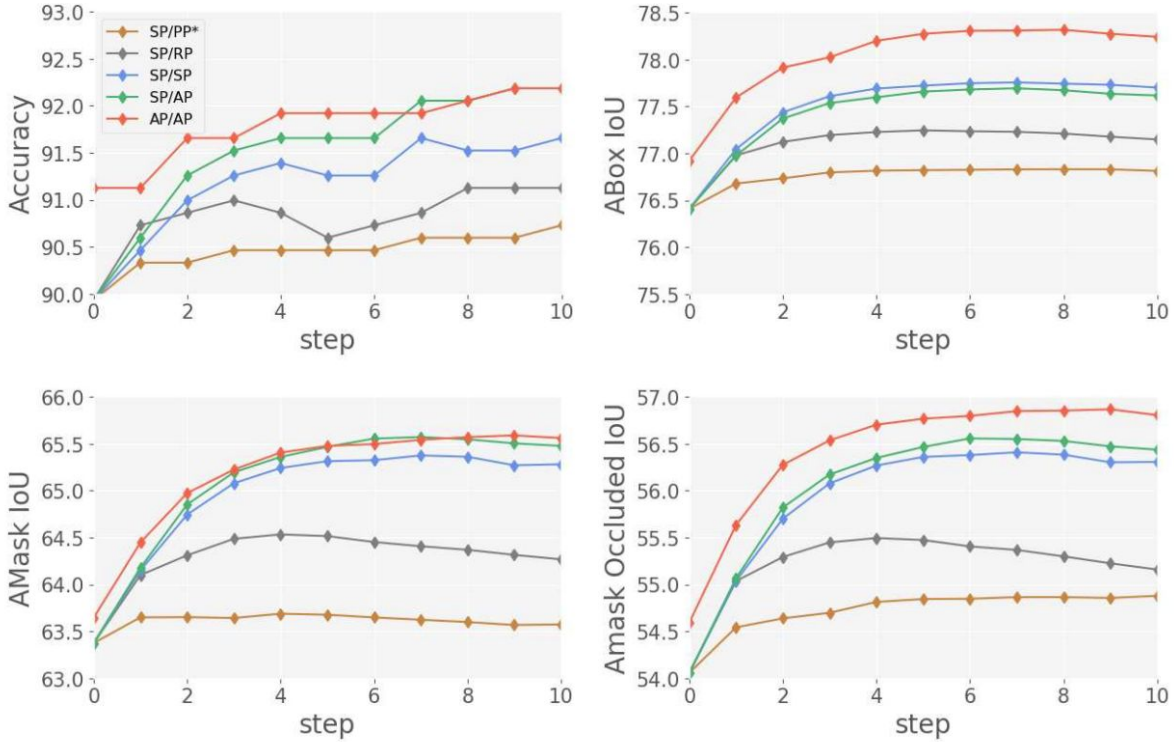
Experimental Results

Moving Path		Clss-Acc			ABox-IoU			AMask-IoU			AMask-Occ-IoU		
Train	Test	all	easy	hard	all	easy	hard	all	easy	hard	all	easy	hard
Passive	Passive	92.9	94.1	90.9	81.3	83.9	76.5	67.6	69.6	63.9	49.0	46.0	54.6
ShortestPath	Passive	92.8	94.3	89.9	81.2	83.8	76.4	67.4	69.6	63.4	48.6	45.8	54.1
ShortestPath	Passive*	93.0	94.3	90.7	80.9	83.1	76.8	66.7	68.4	63.6	48.4	44.9	54.9
ShortestPath	RandomPath	93.1	94.1	91.1	81.6	83.9	77.1	67.8	69.7	64.3	49.0	45.8	55.2
ShortestPath	ShortestPath	93.2	94.1	91.7	82.0	84.3	77.7	68.6	70.4	65.3	**50.2**	**46.9**	56.3
ShortestPath	ActivePath	93.3	93.9	**92.2**	82.0	**84.4**	77.6	**68.8**	**70.5**	65.5	**50.2**	**46.9**	56.4
ActivePath	ActivePath	**93.7**	**94.6**	**92.2**	**82.2**	84.3	**78.2**	68.7	70.3	**65.6**	**50.2**	46.8	**56.7**

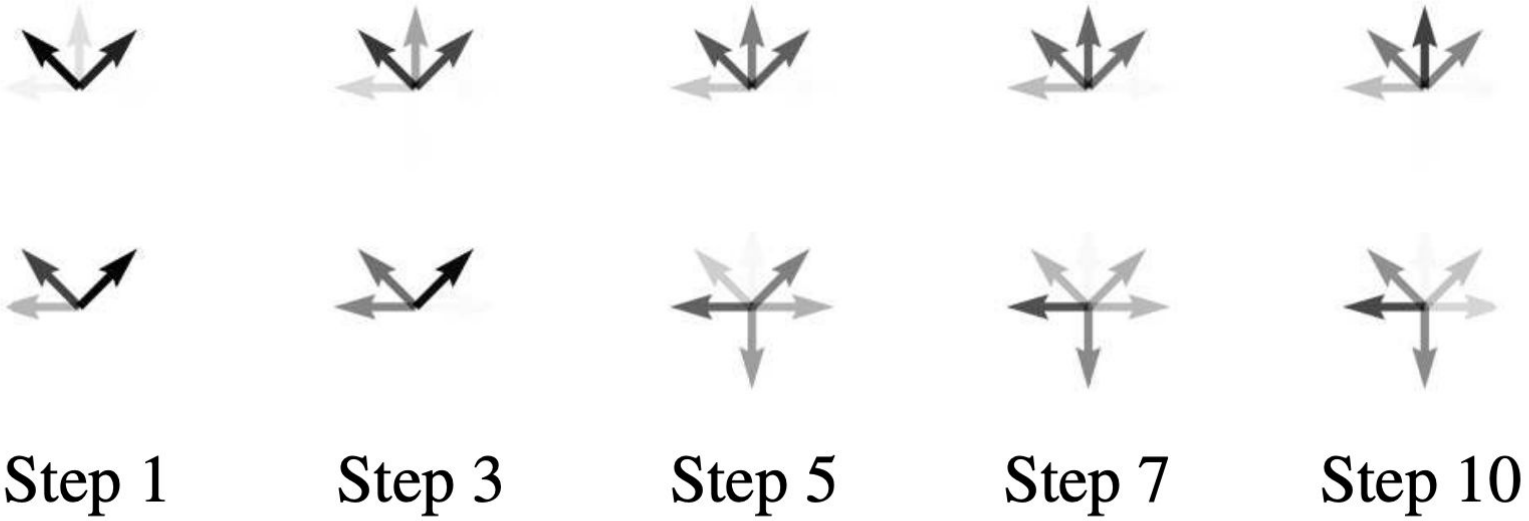
Discussion of Results

- **Shortest path move does not help passive amodal recognition**
  - Both ShortestPath/Passive and ShortestPath/Passive* are slightly inferior to Passive/Passive
- **Embodiment helps amodal recognition**
  - Agents that move at test time consistently outperform agents that stay still
- **New model learns a better moving policy**
  - ShortestPath/ActivePath finds a better moving policy, and the performance is on par or slightly better than ShortestPath/ShortestPath

Additional Experimental Results



Additional Experimental Results



More Discussion of Results

- **Improvements over action step**
  - In general, the performance improves as more steps are taken and more information aggregated, but eventually saturates.
- **Moving strategies**
  - Agent rarely moves forward, learns to occasionally move backward instead.
  - This comparison indicates the shortest path may not be the optimal path for EAR

Critique / Limitations / Open Issues

- Work only conducted in simulations - authors haven't tested it on actual robots in real world situations.
- Experimental setup is kind of limited - we already know what we are looking for
  - Agent is not exploring an open world
  - Would be difficult to transfer the knowledge learned in this model to more real world scenarios

Future Work for Paper / Reading

- It would be interesting to see this applied to real-world scenarios with robots instead of through just simulation
- Paper could be extended to world exploration with more semantic information. Maybe robot could catch a glimpse of something and try to find the optimal path to detecting the object.
- Testing in the simulated environment could benefit from varied objects with more complex shapes, colors, perhaps lighting (different angles could affect recognition).

Summary

- **Problem**: real world scenarios have objects that are obstructed in view, making it difficult for classical CV algorithms to recognizing objects. Amodal recognition is easy for humans, not for machines.
  - Learning via training is great, but is not exercised as well during real world scenarios
- **Previous limitations**: despite many advances, visual systems still fail to recognize objects in the presence of significant occlusion and unusual poses.
- **Key insights:** using Embodied Amodal Recognition we learn that agents with movement consistently perform better than passive agents. Agents developed strategic movements that were different from shortest path, to recover the semantics and shape of occluded objects