Learning to Perform Described Actions in a VirtualHome

TINGWU WANG, ML GROUP, U OF T

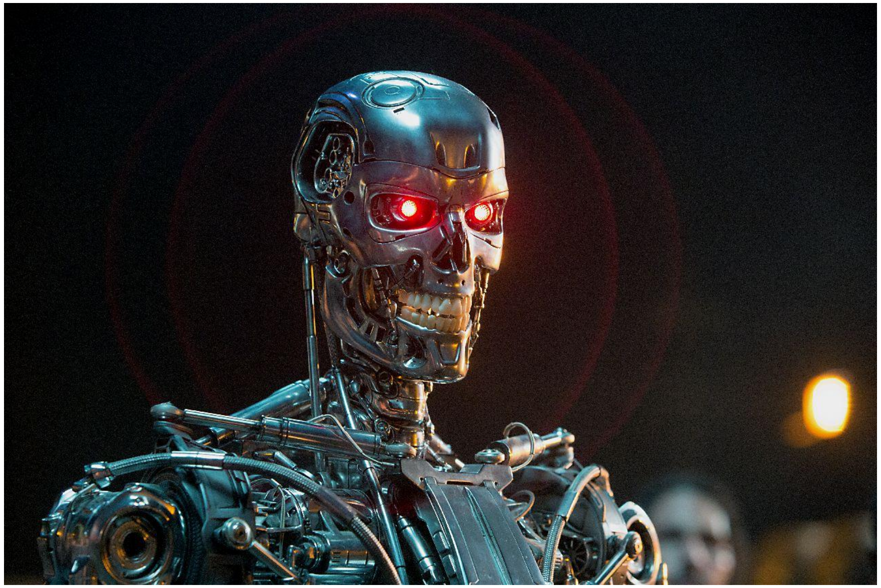
JOINT WORK WITH
XAVIER PUIG,
KEVIN RA,
MARKO BOBEN,
ANTONIO TORRALBA,
AND SANJA FIDLER

Contents

1. Introduction
2. Dataset and Platform
3. Script Generation from Described Actions
4. Results
5. Concluding Remarks

Introduction

1. Background
   1. Target: autonomous agents (domestic robots)
   2. Understand human instructions
   3. Able to execute them correctly



say we have this
cute red-eyed robot
in our house.

Introduction

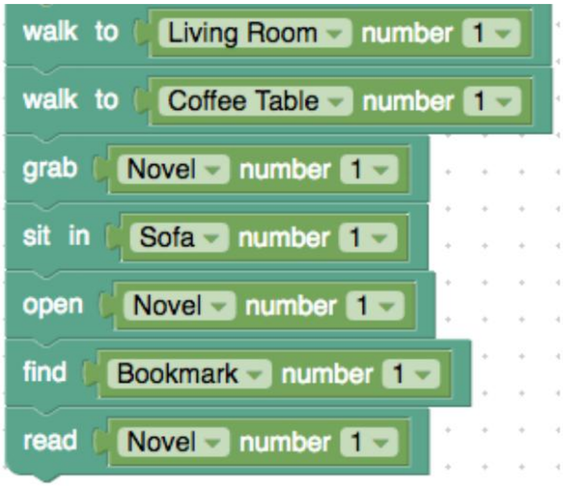
1. How robots "do" things?
    1. Robot uses executable pseudo-code
    2. Stupid robots acts according to
        1. predefined "Atomic Action Triplets"
        2. if smarter, download new sequences
    3. Clever robots learn and predict new sequences
        1. understand natural language

           "find a book and start to read"

           "give me a beer"

           "tell the salesman I am not here

           in the house!"
        2. understand teaching videos

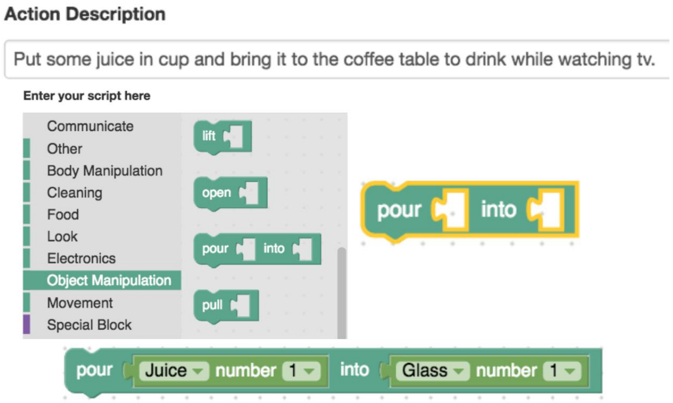


```
walk to Living Room number 1
walk to Coffee Table number 1
grab Novel number 1
sit in Sofa number 1
open Novel number 1
find Bookmark number 1
read Novel number 1
```

Dataset and Platform

1. Crowd-sourcing the Scripts for Tasks
    1. We crowd-source the scripts on AMT, and have them rechecked with a high-quality annotator via Upwork

2. Creating the Virtual Environment
    1. We exploit the Unity3D game engine to create our VirtualHome
        1. provide video ground truth data
        2. independent of the real robot platform
        3. excecute the predicted actions



Action Description

Put some juice in cup and bring it to the coffee table to drink while watching tv.

Enter your script here

Dataset and Platform

1. Data statistics
   1. five rooms, three 'robots'
   2. more than 70 actions and 260 objects to interact
2. Robots able to act according to the predicted atomic actions

Script Generation from Described Actions

1. Sequence to sequence baseline
    1. each atomic triplet is a token.
2. Attention decoder with minimum number of parameters
    1. treat the transition from human natural language to atomic action sequences as language translation
    2. directly using w2v embedding as attentions?
3. w2v pretrained embedding
    1. note that atomic triplet consists of one action and two objects
    2. limited data
4. Video model?

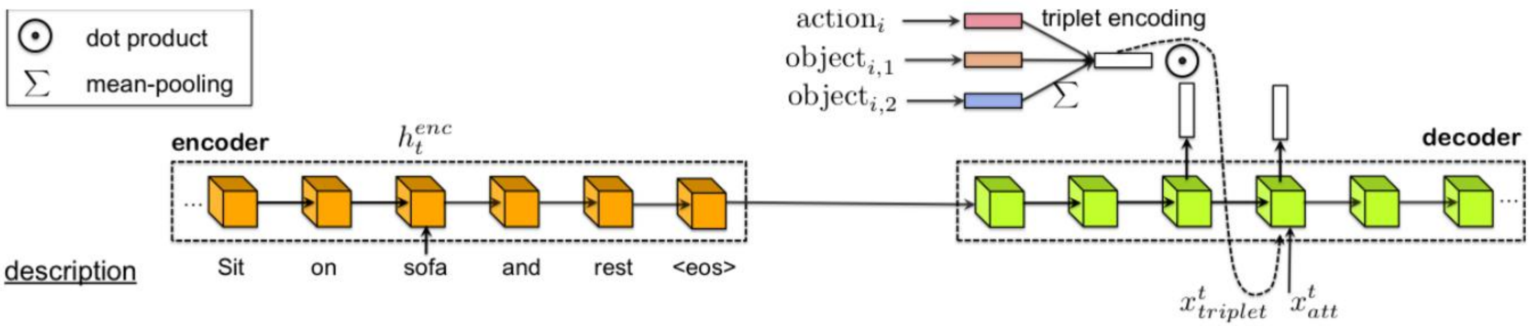
Script Generation from Described Actions

1. Proposed Model

$$\tilde{a}_i = W_a a_i, \quad \tilde{o}_{i,1} = W_o o_{i,1}, \quad \tilde{o}_{i,2} = W_o o_{i,2}$$

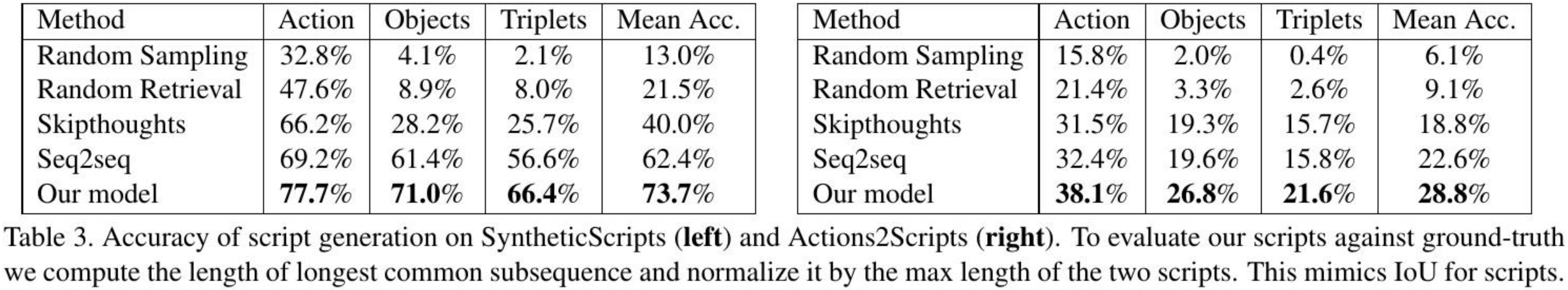
$$v_i = \text{mean}(\tilde{a}_i, \tilde{o}_{i,1}, \tilde{o}_{i,2})$$

$$p_i^t = \text{softmax}_i(v_i^T \cdot h^t)$$



Results

1. Text model



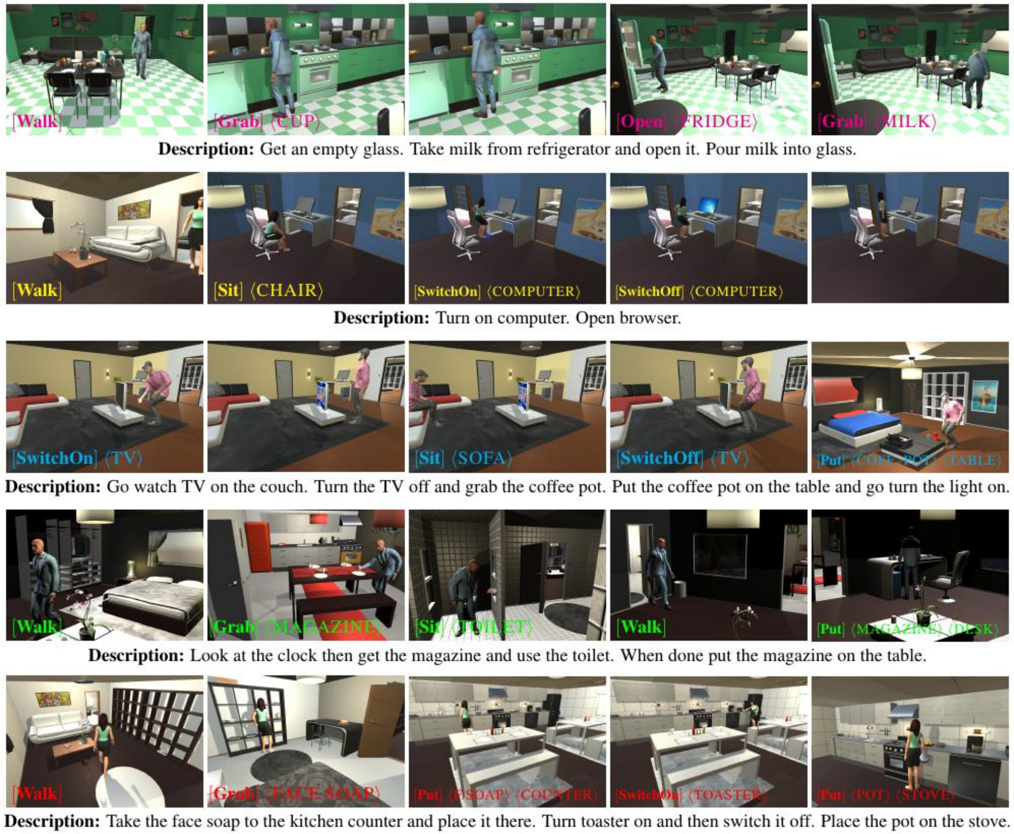
Method	Action	Objects	Triplets	Mean Acc.
Random Sampling	32.8%	4.1%	2.1%	13.0%
Random Retrieval	47.6%	8.9%	8.0%	21.5%
Skipthoughts	66.2%	28.2%	25.7%	40.0%
Seq2seq	69.2%	61.4%	56.6%	62.4%
Our model	**77.7%**	**71.0%**	**66.4%**	**73.7%**

Method	Action	Objects	Triplets	Mean Acc.
Random Sampling	15.8%	2.0%	0.4%	6.1%
Random Retrieval	21.4%	3.3%	2.6%	9.1%
Skipthoughts	31.5%	19.3%	15.7%	18.8%
Seq2seq	32.4%	19.6%	15.8%	22.6%
Our model	**38.1%**	**26.8%**	**21.6%**	**28.8%**

Table 3. Accuracy of script generation on SyntheticScripts (**left**) and Actions2Scripts (**right**). To evaluate our scripts against ground-truth we compute the length of longest common subsequence and normalize it by the max length of the two scripts. This mimics IoU for scripts.

Results

1. Generating videos according to action prediction (limited time)



Description: Get an empty glass. Take milk from refrigerator and open it. Pour milk into glass.

Description: Turn on computer. Open browser.

Description: Go watch TV on the couch. Turn the TV off and grab the coffee pot. Put the coffee pot on the table and go turn the light on.

Description: Look at the clock then get the magazine and use the toilet. When done put the magazine on the table.

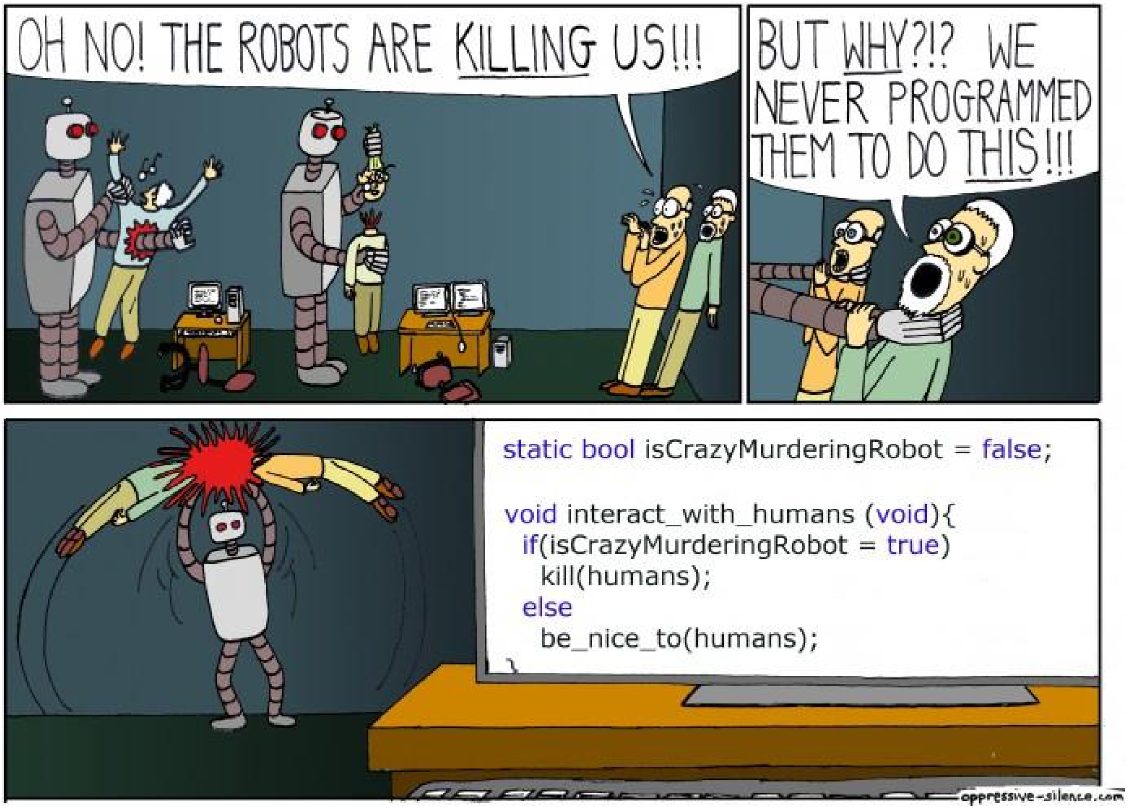
Description: Take the face soap to the kitchen counter and place it there. Turn toaster on and then switch it off. Place the pot on the stove.

Concluding Remarks

1. Work submitted to CVPR 17'
2. Future work
   1. Reinforcement Learning?
   2. Video Teaching?
   3. Zero-shot Learning?

Concluding Remarks

1. Q & A


OH NO! THE ROBOTS ARE KILLING US!!!
BUT WHY?!? WE NEVER PROGRAMMED THEM TO DO THIS!!!

```
static bool isCrazyMurderingRobot = false;

void interact_with_humans (void){
  if(isCrazyMurderingRobot = true)
    kill(humans);
  else
    be_nice_to(humans);
}
```

oppressive-silence.com