1707.03374 - Imitation from Observation: Learning to Imitate Behaviors from Raw Video via Context Translation

- **Yunqiu Xu**
- Other reference
    - https://zhuanlan.zhihu.com/p/27935902
    - https://www.youtube.com/watch?v=kJBRDhInbmU

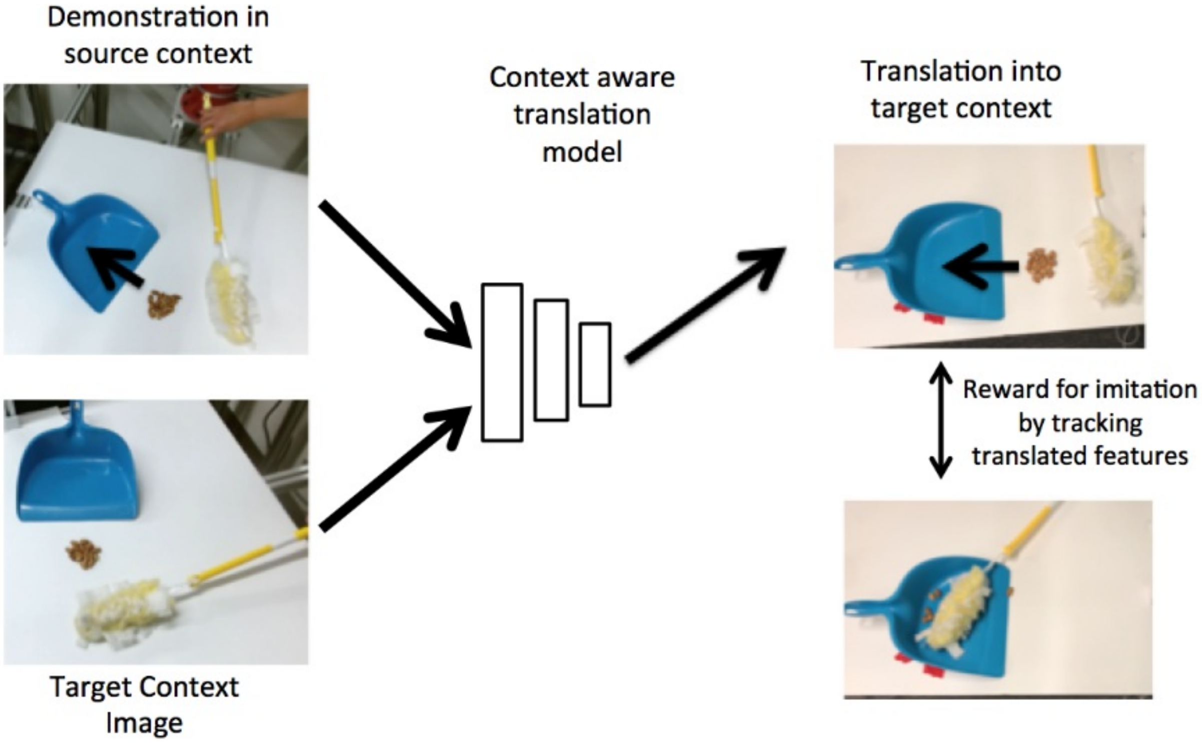
---

1. Introduction

- Comparison of reinforcement learning and imitation learning:
    - Reward function based RL:
        - The goal is described in high-level via reward function
        - The agent can learn skills through trail and error
        - Limitation: reward function is difficult to specify by hand, particularly when the success of the task can only be determined from complex observations such as camera images
    - Imitation learning:
        - Bypass this issue by examples of successful behavior
        - Typical example:
            - BC: off-policy, supervised
            - Reward function learning through inverse reinforcement learning
        - An agent receives examples with sequences of observation-action tuples
        - Then learn a function that maps observations to actions
        - Limitation: this type of imitation is quite different from human and animals
- How do human and animals imitate → imitation-from-observation
    - The observations are obtained from an alternate viewpoint and the actions are not known
    - Humans can learn from not only live observations of demonstrated behavior,

but also video recordings of behavior

- Our work:
    - A context translation model : convert a demonstration from one context (e.g., a third person viewpoint and a human demonstrator) to another context (e.g., a first person viewpoint and a robot)


Demonstration in source context
Context aware translation model
Translation into target context
Reward for imitation by tracking translated features
Target Context Image

2. Related Work

- Behavioral cloning
    - Casts the problem of imitation learning as supervised learning
    - The policy is learned from state-action tuples provided by the expert
    - If we only have observations of the state in a different context and the actions are unknown, we can not use BC directly
- Inverse reinforcement learning
    - Learn a reward function from the expert demonstrations
    - This reward function can then be used to recover a policy by running standard reinforcement learning
    - Difficult in practice in high-dimensional observations e.g. images

- Some recent work which will be compared
    - Ho & Ermon 1606.03476 - Generative Adversarial Imitation Learning
        - GAIL, which has been mentioned in other notes
        - Generator: generate action sequence
        - Discrimitor: similar to reward function, check whether this sequence is expert demonstration
        - Do not need reward setting, but GAN is hard to train
    - Sermant 1612.06699 - Unsupervised Perceptual Rewards for Imitation Learning
        - Address differences in context by using pretrained visual features
        - Rely on the inherent invariance of visual features for learning
        - Do not provide mechanism for context translation
    - Stadie 1703.01703 - Third-Person Imitation Learning
        - This can be seen as an early exploration of our work
        - Some restrictive requirments
        - Performs poorly on the more complex manipulation tasks

3. Imitation-from-observation Problem

- An agent observes demonstrations of a task in a variety of contexts, then executes the demonstrated behavior in its own context
- Contexts: properties of the environment and agent
    - The viewpoint
    - The agent’ s embodiment
    - The positions and identities of objects in the environment
- Demonstrations

  $$\{D_1, D_2, \dots, D_n\} = \{[o_0^1, o_1^1, \dots, o_T^1], [o_0^2, o_1^2, \dots, o_T^2], \dots, [o_0^n, o_1^n, \dots, o_T^n]\}$$

    - o_t : observation produced by partially observed Markov process
    - Observations are governed by
        - w : context sampled independently from $p(\omega)$ for each demonstration
        - $p(o_t|s_t, w)$: observation distribution
        - $p(s_{t+1}|s_t, a_t, \omega)$: dynamics
        - $p(a_t|s_t, w)$: expert's policy

- In this work we simplify the problem as : the context can vary between the demonstrations and the learner, but the learner’ s context still comes from the same distribution
- The challenges of imitation-from-observation:
    - Be able to determine what information from the observations to track
    - Be able to determine wich actions will allow it to track the demonstrated observations → **solved by RL**
        - Use distance of demonstration as reward function
        - Learn a policy that takes actions to minimize this distance

4. Model

- Translation between contexts

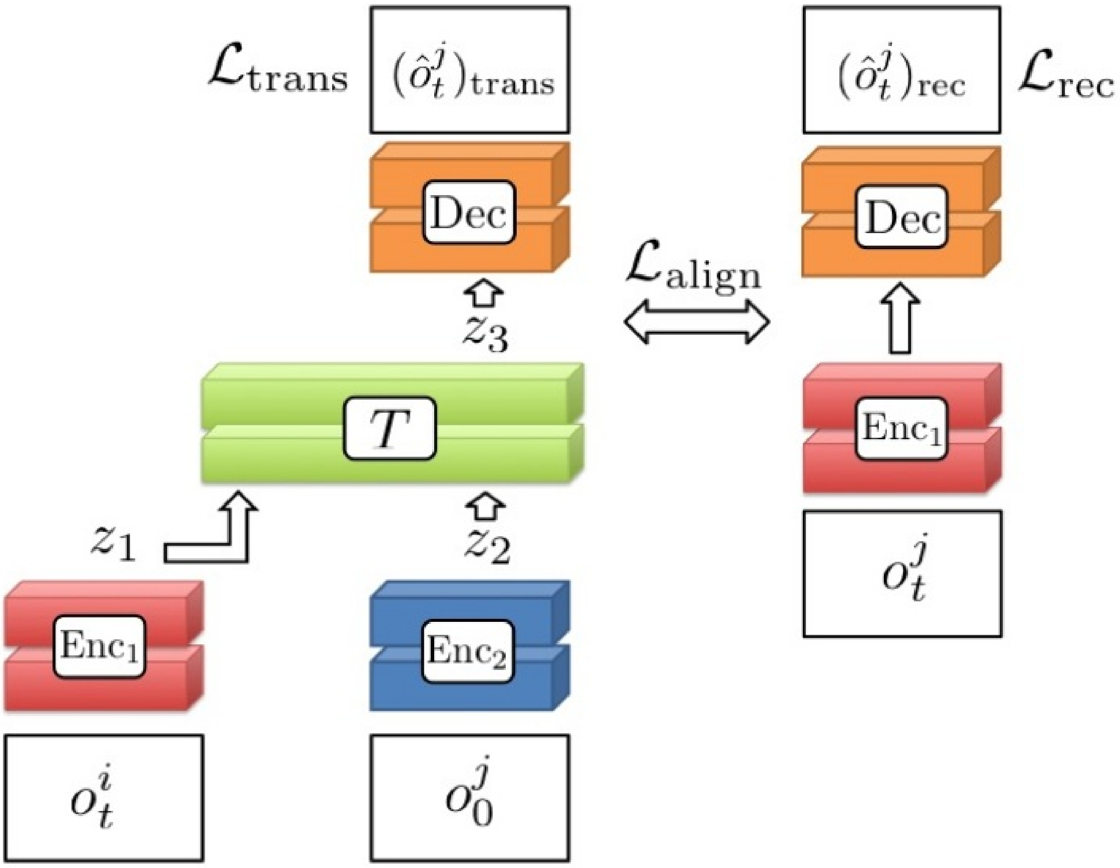

$\mathcal{L}_{\text{trans}}$
$(\hat{o}_t^j)_{\text{trans}}$
Dec
$\mathcal{L}_{\text{align}}$
z_3
T
z_1
z_2
Enc_1
Enc_2
o_t^i
o_0^j
$(\hat{o}_t^j)_{\text{rec}}$
$\mathcal{L}_{\text{rec}}$
Dec
Enc_1
o_t^j

Figure 2: Context translation model: The source observation o_t^i is translated to give the prediction of the observation in the target context $(\hat{o}_t^j)_{\text{trans}}$, given the context image o_0^j from the target context. The convolutional encoders are Enc_1 and Enc_2, while the deconvolutional decoder Dec decodes features back into observations.

- Reward function for RL:

$$\hat{R}_{img}(o_t^l) = -||o_t^l - \frac{1}{n}\sum_{i}^{n} M(o_t^i, o_0^l)||_2^2$$

$$\hat{R}(o_t^l) = \hat{R}_{feat}(o_t^l) + w_{rec}\hat{R}_{img}(o_t^l)$$

5. Experiments

- Questions:

  - Can our context translation model handle raw image observations, changes in viewpoint, and changes in the appearance and positions of objects between contexts
  - How well do prior imitation learning methods perform in the presence of such variation, in comparison to our approach
  - How well does our method perform on real-world images, and can it enable a real-world robotic system to learn manipulation skills
- We compare the performance of our model with some work mentioned in 2

- The video can be seen at https://www.youtube.com/watch?v=kJBRDhInbmU

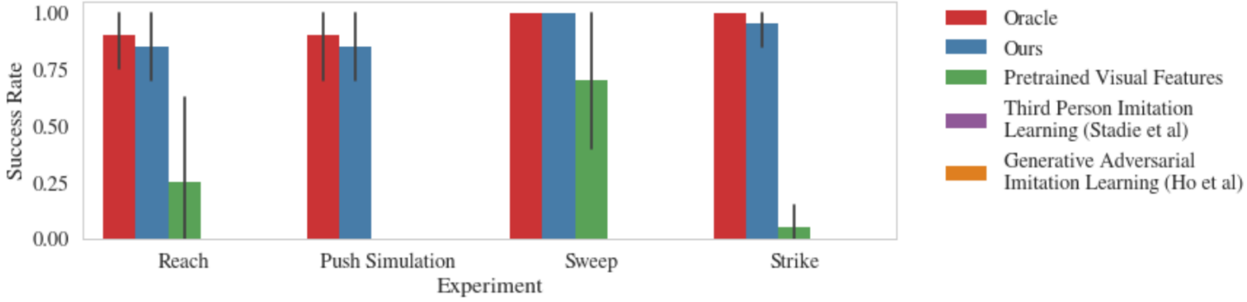


Figure 5: Comparisons with several prior methods on the reaching, pushing, sweeping, and striking tasks. The results show that our method successfully learned each task, while the prior methods were unable to perform the reaching, pushing and striking tasks, and only the pretrained visual features approach was able to improve well on the sweeping task. Third person Imitation Learning and Generative Adversarial Imitation Learning are both at 0% success rate on the graph.

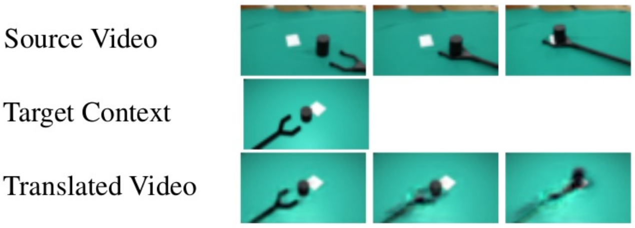


Figure 6: Translations from a video in the holdout set for the pushing task on real demonstrations to a context in the simulated world.

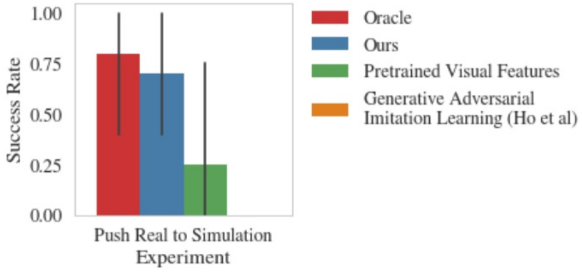


Figure 7: Performance of our method versus other methods on the pushing task with real world demonstrations and policy learning in simulation.

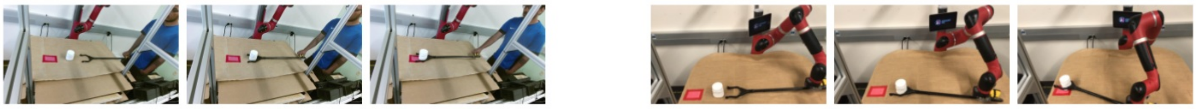


Figure 8: Video of our method successfully pushing the object onto the target from arbitrary viewpoint with the Sawyer robot. Left: One of the demonstrations provided by human. Right: Imitation learned by the robot

6. Conclusion

- This work focus on imitation-from-observation problem
- Translate demonstration observation sequences (e.g., videos) between different contexts, such as differences in viewpoint
- The translation model is trained by translating between the different contexts observed in the training set, and generalizes to the unseen context of the learner.
- Limitations:
    - Require a substantial number of demonstrations to learn the translation model → Training maybe inefficient
    - Require observations of demonstrations from multiple contexts to learn to translate
    - We can explore explicit handling of domain shift in future work
        - Large differences in embodiment
        - Learn robotic skills directly from videos of human demonstrators