Procedure Planning in Instructional Videos via Contextual Modeling and Model-based Policy Learning

Jing Bi Jiebo Luo Chenliang Xu
University of Rochester
jing.bi@rochester.edu jiel@cs.rochester.edu chenliang.xu@rochester.edu

1. Implementation Details

As we only use image-based observations, the Inference model is DCGAN [4], where the encoder is fixed to a three-layer convolutional neural network with 64 as the hidden vector size for context variable z_c, and the decoder has the similar structure as encoder but with transposed convolutions. In the generation model, the policy network π_θ is a two-headed multi-layer perceptron parameterized by θ for computing actions and estimating the expected return values. The transition model in Ext-MGAIL is the Gaussian distribution with mean and variance parameterized by a multi-layer perceptron. We represent the discriminator as a multi-layer perceptron with parameters ω that takes as input a state-action pair and outputs a value between 0 and 1. All network consisted of 2 hidden layers with [64,32] units in each layer and Leaky-ReLU as non-linearity function. During training, all model are optimized by Adam optimizer [3] with the starting learning rate of $1e^{-4}$.

2. Experiment on a Second Dataset

To further illustrate the effectiveness of our method, we experiment on a second dataset [1]. This dataset contains narrated instruction videos that described five diverse tasks: 1. Change tire, 2. Perform CPR, 3. Repot plant, 4. Make coffee, and 5. Jump car, where each task has 30 videos with the dense instructional caption. Table. 1 shows the statistical comparison of two datasets. Compared with CrossTask, this dataset has fewer tasks, but the average trajectory length is three steps longer. We choose this dataset over other instructional video datasets because it is challenging for our method as it contains fewer samples with longer trajectory.

Table 1: Comparsion between two datasets.

	Tasks	Videos	Actions	Avg. length
Crosstask	18	2,750	105	6.5
2^{nd} Dataset	5	150	58	9.5

Table 2: **Results of Procedure Planning.** Both of our models outperformed the DDN [2], but the performance increase is smaller than that obtained on the CrossTask; this shows that for longer sequence, optimizing on the whole trajectory can marginally improve the performance.

		Uniform	DDN	Int.	Ext.
T=3	Succ. rate	2.21	18.41	20.19	**22.11**
	Accuracy	4.07	32.54	39.02	**42.20**
	mIoU	6.09	56.56	60.65	**65.93**
T=3	Succ. rate	1.12	15.97	**19.91**	17.47
	Accuracy	2.73	27.09	36.31	**37.89**
	mIoU	5.84	47.32	53.84	**55.52**

Table 3: **Results of Walk-through Planning**. Our model outperforms the baseline models by modelling the transition over whole sequence.

		Uniform	DDN	Int	Ext
T=3	Hamming	0.89	0.62	0.31	**0.24**
	Pair acc.	62.60	88.61	90.29	**95.59**
T=4	Hamming	1.12	0.87	0.64	**0.56**
	Pair acc.	58.55	85.43	87.65	**93.76**

2.1. Walk-through Planning

The key to successfully perform walk-through planning is to construct the rank matrix, capturing the transition probability between the two observations. The results are shown in Table 3, our two methods outperform the baseline model even we only have marginal performance in procedure planning. Explicitly modelling the transition makes our method focus on capturing the transition between the two adjacent observations; thus, both models have a better performance than the walk-through planning experiment on CrossTask.

2.2. Procedure Planning

As shown in Table 2, both of our models outperformed the DDN [2]. However, the increase of the performance is

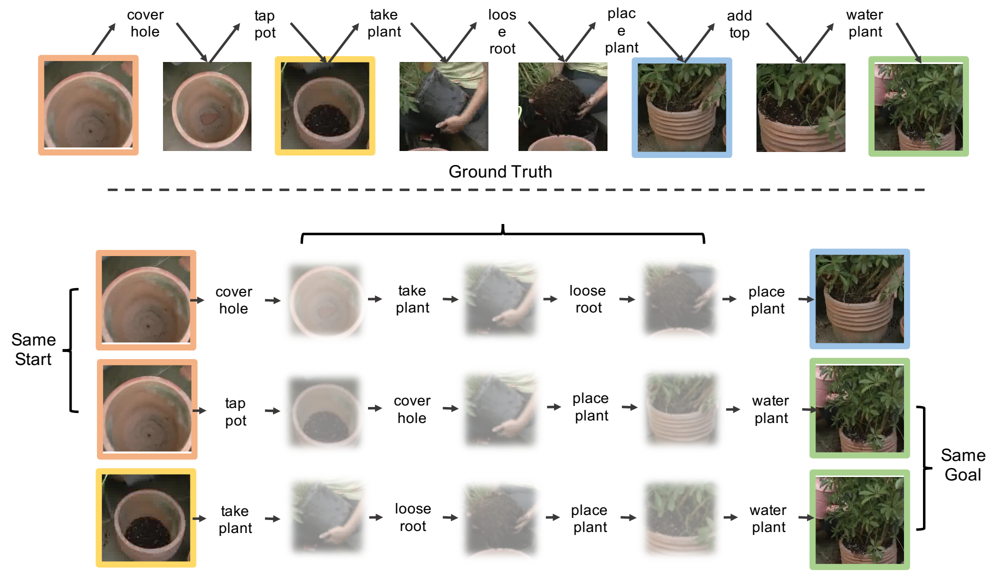

Figure 1: **Procedure Planning qualitative results**. Procedure Planning qualitative results for *Repot Plant*. The top row describes the correct action sequence required to repot the plant. To examine our mode's robustness, We vary the start and goal observations to evaluate our method. The results show that our approach is robust to perform planning within different stages in the video

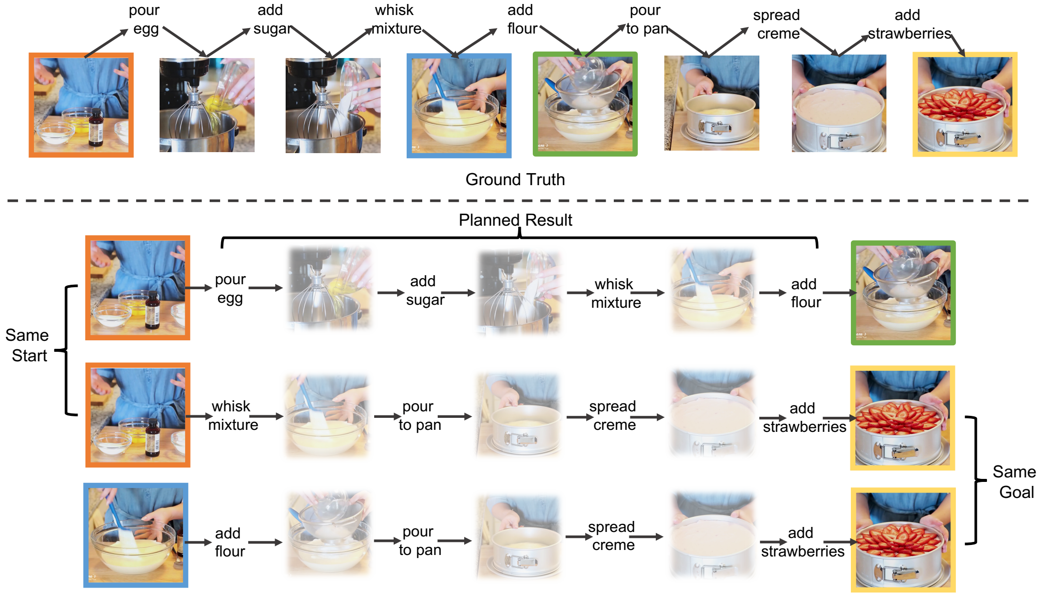

Figure 2: **Procedure Planning qualitative results** for *Make French Strawberry Cake*. The top row describes the correct action sequence required to make a strawberry cake. We evaluate our method by varying the start and goal observations, respectively. The results show that our approach is robust to perform planning within different stages in the video.

smaller than one obtained on the CrossTask, and the performance between Int-MGAIL and Ext-MGAIL are very similar. This is due to the fact that optimizing the whole trajectory is similar to the Monte Carlo tree search– the algorithm needs more samples to find the optimal path for longer sequence length. However, the new dataset does not provide sufficient samples to improve the performance over the longer trajectory. In this case, the performance primarily depends on the network's ability to capture the one-step transition. Therefore, using a more robust transition model (Ext.) will not help here because the small number of samples can be easily modelled even with the interior model (Int.). The transition model will degenerate to the first case (w/o seq) in the ablation study.

References

[1] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4575–4583, 2016. 1

[2] Chien-Yi Chang, De-An Huang, Danfei Xu, Ehsan Adeli, Li Fei-Fei, and Juan Carlos Niebles. Procedure planning in instructional videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1

[3] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference for Learning Representations*, 2015. 1

[4] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 1