Induce, Edit, Retrieve: Language-grounded Multimodal Schemata for instructional Video Retrieval

Yue Yang, Joongwon Kim, Artemis Panagopoulou, Mark Yatskar, Chris Callison-Burch

Penn Engineering, University of Pennsylvania

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

Abstract

Schemata are structured representations of complex tasks that can aid artificial intelligence by allowing models to break down complex tasks into intermediate steps. We propose a novel system that induces schemata from web videos and generalizes them to capture unseen tasks with the goal of improving video retrieval performance.

Methods

Our system proceeds in 3 major phases:
1) Schema Induction; 2) Schema Editing;
3) Schema-Guided Video Retrieval.
See figure in the middle for details.

Experiments

Datasets:

- Howto100M: 1.2M videos, 23K tasks. We induce 21,299 schemata using it.
- Howto-GEN: 3,365 from Howto100M. Train/Val/Test: 500/500/1088.
- COIN: 11,827 videos for 180 tasks.
- Youcook2: 2,000 videos for 89 tasks.

Baselines (Other induction methods):

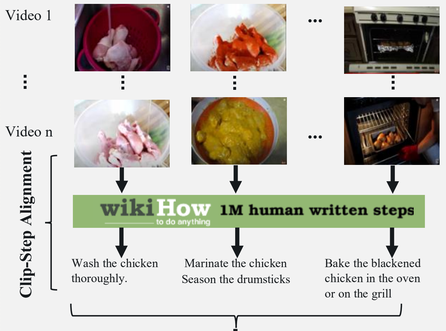
- Language Models: T5, GPT-2, GPT-3.
- **G**oal **O**riented **S**cript **C**onstruction.
- Human written: wikiHow and Oracle.

Schema Induction

Objective: Construct schemata on tasks.

Task: Bake Chicken

YouTube Videos (Learning Data)

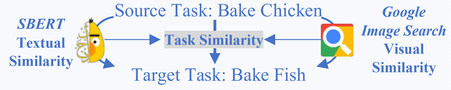


Schema of *Bake Chicken*

- Wash the chicken thoroughly
- Season your drumsticks
- Marinate the chicken
- Insert a roasting thermometer into the thigh
- Sprinkle the ginger on the chicken
- Bake the blackened chicken in the oven
- ... ...

Schema Editing

Objective: Edit existing schema for unseen task.



Object Replacement

Source Step: Wash the chicken thoroughly

Target Step: Wash the fish thoroughly

Step Deletion

Source Step 1: Insert a roasting thermometer into the thigh
P(Source Step 1 | Bake Fish) << P(Source Step 1 | Bake Chicken)
Delete this step.

Source Step 2: Bake the blackened chicken/fish in the oven
P(Source Step 2 | Bake Fish) ≈ P(Source Step 2 | Bake Chicken)
Include this step.

Token Replacement

Source Step: Sprinkle the ginger on the fish
Mask the token: Sprinkle the <mask> on the fish
Use LM predict a new token:
Target Step: Sprinkle the sauce on the fish

Predicted Schema of *Bake Fish*

- Wash the fish thoroughly
- Season your fish
- Marinate the fish
- Sprinkle the sauce on the fish
- Bake the blackened fish in the oven
- ... ...

Schema-Guided Video Retrieval

Objective: Use schema to improve retrieval performance.

Retrieve videos of "Bake Fish".
Use the task name "Bake Fish" as query.

Global Matching: Lack of Intermediate Information.

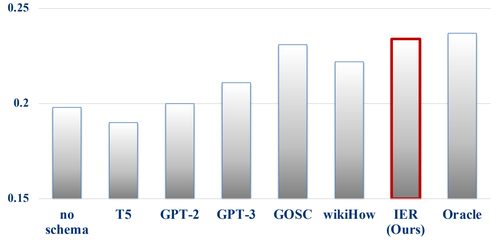
With Schema:

Predicted Schema of *Bake Fish*

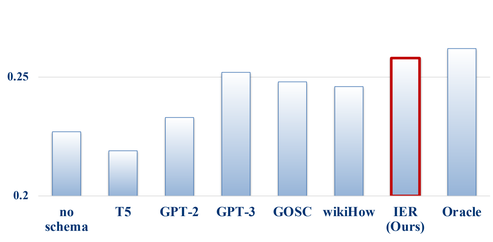
- Wash the fish thoroughly
- Season your fish
- Marinate the fish
- Sprinkle the sauce on the fish
- Bake the blackened fish in the oven
- ... ...

Wash the fish thoroughly | Sprinkle the sauce on the fish | Bake the blackened fish in the oven

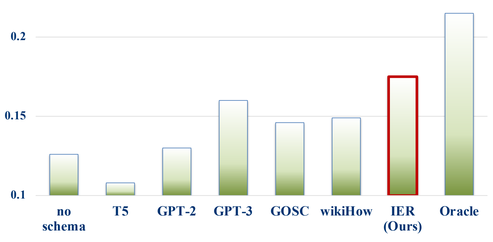
Howto-GEN (MRR)



COIN (MRR)

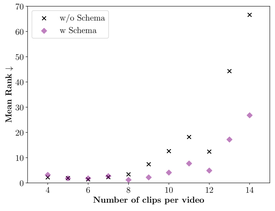


Youcook2 (MRR)

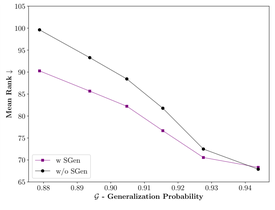


Results

Schema helps long videos:



Editing helps unseen tasks:



All three editing modules are beneficial:

	Method	P@1↑	R@5↑	R@10↑	Med r↓	MRR↑
Howto-GEN	full	**54.4**	**37.3**	**50.1**	**10.0**	**.231**
	− mask	53.7	36.3	49.3	11.0	.229
	− deletion	53.6	36.9	49.8	11.0	.230
	− replacement	51.5	34.9	47.3	12.0	.220
	− all	45.5	31.0	43.1	15.0	.199

Schemata can be used by other vision-text models:

Model	P@1↑	R@5↑	R@10↑	Med r↓	MRR↑
MIL-NCE	48.3	37.1	52.8	9.5	.227
+schema	57.2	42.2	57.8	7.0	.256
CLIP [38]	58.9	44.9	58.8	6.0	.264
+schema	65.0	47.4	60.8	5.5	.282

Conclusion

We propose a schema induction and generalization system that improves instructional video retrieval performance. We demonstrate that the induced schemata benefit video retrieval on unseen tasks, and our IER system outperforms other methods. In the future, we plan to investigate the structure of our schemata, such as the temporal order, and discover other applications of schemata.