Look for the Change: Learning Object States and State-Modifying Actions from Untrimmed Web Videos

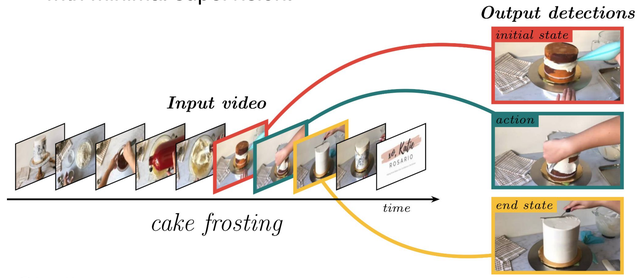
Tomáš Souček[1], Jean-Baptiste Alayrac[2], Antoine Miech[2], Ivan Laptev[3], Josef Sivic[1]

[1]Czech Technical University [2]DeepMind [3]ENS/INRIA

OVERVIEW

Goal

- Temporally localize object states together with the corresponding state-modifying actions in videos.
- An example: the initial state *plain corpus*, the end state *frosted cake*, and the action *cake frosting*.
- Use only uncurated videos from the internet with minimal supervision.



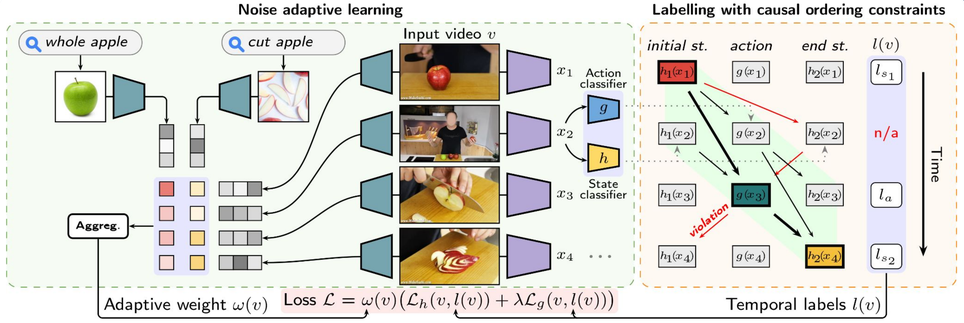
Challenges

- Visual variability of the objects and its states.
- Thousands of objects with many more states, annotating is both difficult and expensive.
- In-the-wild, noisy, uncurated, and long videos.

Contributions

- Self-supervised model for learning object states and state-modifying actions from long uncurated web videos.
- Causal ordering signal (*initial state* → *action* → *end state*) is used as the supervision.
- New uncurated dataset with 2600+ hours of video and 34 thousand changes of object states.

METHOD



Learning step

- compute loss $\mathcal{L}$ and perform a gradient update on the classifiers using the computed labels $l(v)$

Noise adaptive weighting $\omega(v)$

training videos may be unrelated → we compute similarity between video frames and example images $\mathcal{E}_1$, $\mathcal{E}_2$ and use normalized similarity score $\omega(v) = \sigma\left(\frac{r_v - \theta}{\tau}\right)$ to weight contribution of the video in the loss

$$r_v = \max_{t<t'} \sum_{e_1 \in \mathcal{E}_1} \mathrm{sim}(e_1, v_t) \sum_{e_2 \in \mathcal{E}_2} \mathrm{sim}(e_2, v_{t'})$$

Labelling step

- compute labels $l(v)$ for a video $v = \{x_t\}_{t=1}^{T_v}$ using the *state* h and *action* g classifiers:

$$l(v) = \arg\max_{l \in \mathcal{D}_v} h_1(x_{l_{s_1}}) \cdot g(x_{l_a}) \cdot h_2(x_{l_{s_2}})$$

Computing Labels $l(v)$

labels $l(v)$ is a triplet of the *initial state* l_{s_1}, *action* l_a and *end state* l_{s_2} positions.

The labels satisfy the ***causal ordering constraint***:

$$\mathcal{D}_v = \{l : 1 \le l_{s_1} < l_a < l_{s_2} \le T_v\}$$

CHANGEIT DATASET

- Gathered by searching Youtube for terms such as *"How to cut an apple?"*
- 44 state-changing action categories such as apple cutting, ball inflating, etc.
- 34,428 in-the-wild videos, in total 2,642 hours, average video length 4.6 minutes
- 667 videos per-frame annotated with labels: *background*, *initial state*, *action*, *end state*

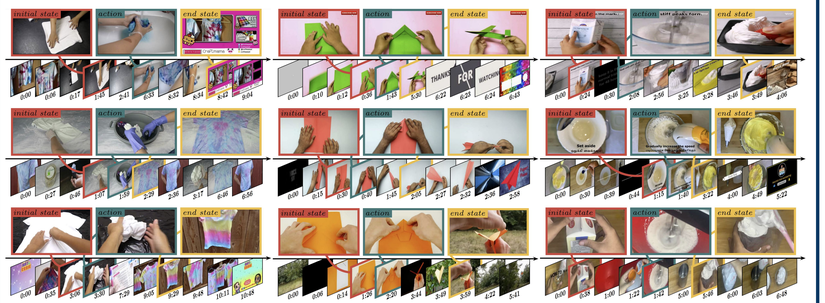
RESULTS

	St prec	Ac prec
Random w/ constraint	0.15	0.41
MIL-NCE S3D [1]	0.27	0.50
CLIP ViT-L/14 [2]	0.30	0.63
Alayrac '17 [3]	0.30	0.59
Ours	**0.35**	**0.68**

- Training on 44 categories of the ChangeIt dataset
- A separate model trained for each category
- A single frame predicted as the initial state, the action and the end state per video

Qualitative results, see the consistency

Task: T-shirt dying | Task: Cream whisking | Task: Paper plane folding



[1] A. Miech et al. End-to-End Learning of Visual Representations from Uncurated Instructional Videos. In CVPR, 2019.
[2] A. Radford et al. Learning transferable visual models from natural language supervision, 2021.
[3] J.-B. Alayrac et al. Joint discovery of object states and manipulation actions. In ICCV, 2017.

ABLATIONS

	St prec	Ac prec
w/o noise adapt.	0.34	0.64
only a single layer	0.34	0.61
w/o data augment.	0.33	0.66
Ours	**0.35**	**0.68**

- clear improvement with increasing dataset size
- model improves even when trained on more noisy low-ranking videos

Project page
data.ciirc.cvut.cz/public/projects/2022LookForTheChange