Discovering Actions by Jointly Clustering Video and Narration Streams Across Tasks

Minttu Alakuijala, Julien Mairal, Jean Ponce, Cordelia Schmid minttu@google.com

Introduction

Cross-task action discovery without task labels

- Using narrated instruction videos, we discover actions in that may be shared across different tasks
  - e.g. cooking steps in different recipes, construction steps in home DIY tutorials, ...
- Weakly supervised setting: **use only the narration and its timing as supervision**
- We consider an unlabeled set of videos, without assuming a known or underlying grouping into distinct tasks, or that videos depicting the same task consist of an identical sequence of steps
- Prior work mostly looks at clustering for a single activity
  - Alternatively, compositional action models have been considered [2], but this relies on task labels and a global script to be known for each task

- We treat the order of actions in each video as independent of other videos: **allows out-of-order execution, repeated and missing steps**
  - Only the actions shown and the actions described verbally in one video should be consistent

Method

Discriminative clustering

- Learn P (resp. Q) which assigns each frame (resp. word) to one of k action classes or background
  - should be recoverable by linear classifiers $XU + a$ and $YV + b$
  - X: frame features, Y: word features

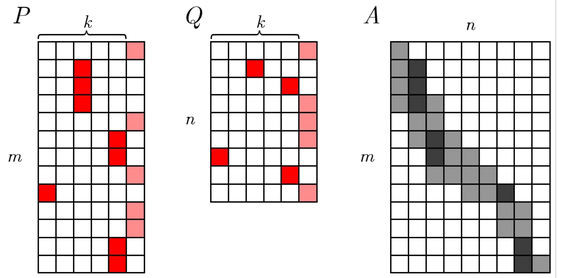
$$h(P,Q) = \min_{\substack{U\in\mathbb{R}^{p\times K},\\ V\in\mathbb{R}^{q\times K}\\ a,b\in\mathbb{R}^K}} \frac{1}{2m}\left\|P - XU - \mathbb{1}_m a^T\right\|_F^2 + \frac{\alpha}{2}\|U\|_F^2 + \frac{1}{2n}\left\|Q - YV - \mathbb{1}_n b^T\right\|_F^2 + \frac{\beta}{2}\|V\|_F^2 .$$

- Minimize $h(P,Q)$ subject to constraints
  - separate constraints: non-overlapping actions
  - joint constraints
    - alignment: for each action present in P, there should be at least one word labeled with the same action in its alignment window
    - order: the order of actions in P and Q should be the same
- Optimize w.r.t P and Q using Frank-Wolfe with block-coordinate descent
  - replace joint constraints by a penalty l(P, Q) representing **distance between sequences**

Further objective terms

$$h(P,Q) = f(P,U^*) + \alpha g(Q,V^*) - \rho_P H(P) - \rho_Q H(Q) + \gamma_P bg(P) + \gamma_Q bg(Q) + \lambda_t l(P,Q)$$

Entropy *H* on the distribution of labels and **a cost for assigning to background** are added to balance sizes of clusters. λ_t is increased over time.



A **frame-to-action assignment *P*, word-to-action assignment *Q*,** and an **approximate temporal alignment *A*** between frames and words, for a single video. The matching pairs in *A* depicting the same action (in dark grey) are a priori unknown.

Sequence alignment penalty for a fixed (P, Q) pair for a single video

		A	B	-	C
	0	1	2	2	3
B	1	2	1	1	2
A	2	1	2	2	3
-	2	1	2	2	3
B	3	2	1	1	2
B	4	3	1	1	2
C	5	4	2	2	1

Results

Experimental evaluation

- Evaluation: intersection over union (IoU) of discovered video segmentation; F1 or recall of single-step predictions (using middle time step of the predicted interval)
- Comparable performance to existing methods, with much less supervision
  - Additionally **recover full intervals** and rich **textual descriptions** of actions instead of single time steps only
  - Allowing out-of-order execution is particularly important in Perform CPR, which includes many repetitions

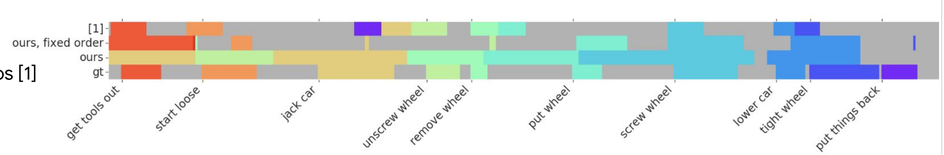
IoU Instruction Videos [1]

	Change tire	Make coffee	Perform CPR	Jump-start car	Repot plant
random	0.037	0.027	0.036	0.016	0.030
k-means	0.120	0.093	0.127	0.034	0.070
Ours, init. from k-means	0.125	**0.097**	0.127	0.038	0.070
[1], fixed length intervals	0.126	0.065	0.076	**0.042**	0.064
Ours, from [1]'s init.	**0.162**	0.092	**0.238**	0.036	**0.121**

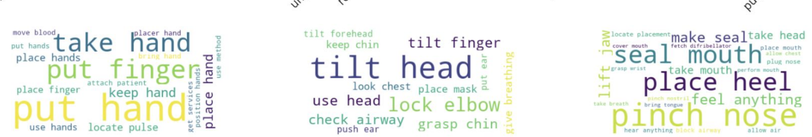
Mean Recall (%) CrossTask [2]

Uniform	9.7
Ours	12.3
Alayrac et al.	13.3
Zhukov et al.	22.4

Example segmentation Instruction Videos [1]



Example word clusters Instruction Videos [1]



Conclusion

Takeaways and further work

- Shortage of datasets to study cross-task sharing
  - task descriptions often very specific
- Expressiveness of features is key
- Detecting background is difficult with discriminative clustering

Limitations

- Relies on quality of transcribed narration: manually corrected in Instruction Videos
- May be difficult to find suitable hyperparameters: important to have a validation procedure

Extensions

- If task labels are available, may want to encode a prior on a shared sequence while allowing for cross-task sharing
- Investigate features learned on instruction videos, such as HowTo100M [3]
- Use free-form narration
- Other definitions of conflict, e.g. penalize each inserted segment not present in the other modality, instead of each time step

[1] Alayrac, Jean-Baptiste, et al. "Unsupervised learning from narrated instruction videos." CVPR. 2016.
[2] Zhukov, Dimitry, et al. "Cross-task weakly supervised learning from instructional videos." CVPR. 2019.
[3] Miech, Antoine, et al. "Howto100M: Learning a text-video embedding by watching hundred million narrated video clips." CVPR. 2019.