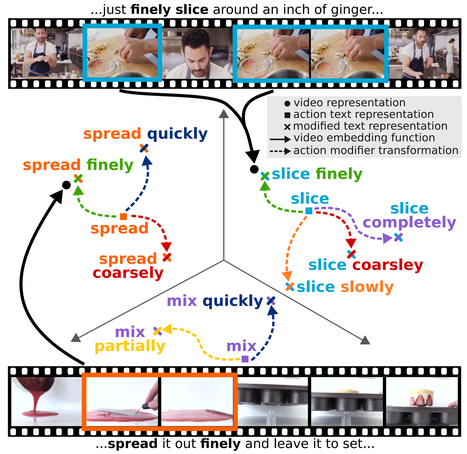
# Action Modifiers: Learning from Adverbs in Instructional Videos

Hazel Doughty[1], Ivan Laptev[2], Walterio Mayol Cuevas[1,3] and Dima Damen[1]

[1]University of Bristol, [2]Inria, École Normale Supérieure, [3]Amazon

CVPR SEATTLE WASHINGTON JUNE 16-18 2020

## Overview

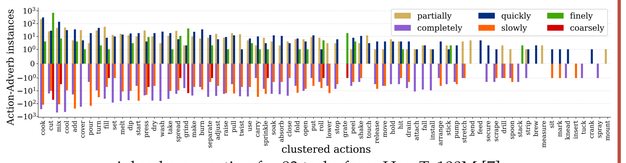



**Motivation**: Recent effort in instructional videos learn the key steps to complete a task [1] or assess overall task performance [2]. They do not identify whether individual actions have been performed as recommended by, say, a recipe. Many steps need to be performed in a certain way to achieve the desired outcome. These are often identified by *adverbs*.

**Novelty:** We learn adverbs in a manner that generalizes across actions and tasks. We propose a video-text embedding space, learned from weakly-supervised action-adverb pairs in narrations of instructional videos. In this space, adverbs are learned as action modifiers – transformations which modify the action's embedding. Our approach addresses two main challenges:

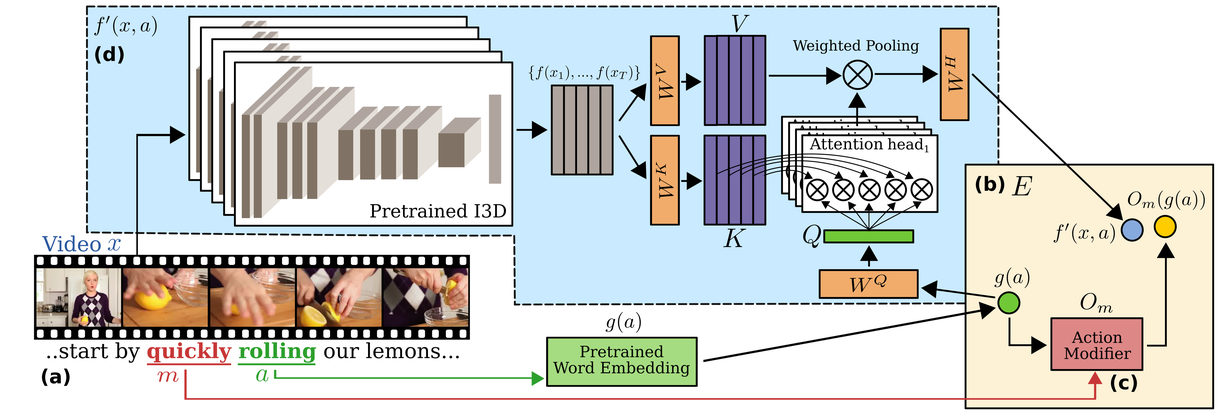
- Disentangling the action from the adverb. Learning adverbs as action modifiers allows us to learn how the same adverb applies across actions.
- Learning from the relevant parts of the video with weak supervision from the narration. Our weakly-supervised embedding uses the action as a query in scaled-dot product attention.

## Data




Adverbs per action for 83 tasks from HowTo100M [7].


## Method



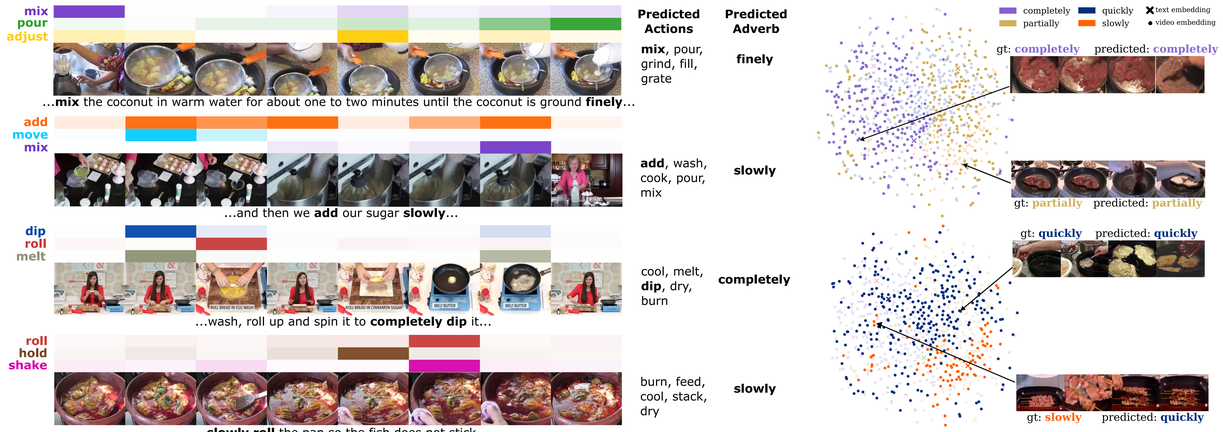

(a) The input is a video $x$ with weak label $(a, m)$ for the action and adverb.

(b) We aim to learn a space $E$ where the embeddedings of the video (blue) and action-adverb (yellow) are close.

(c) Adverbs are learned as action modifiers $O_m$, which are transformations in $E$.

(d) We embed relevant video parts $f'(x, a)$ using multi-head scaled dot-product attention where the action embedding $g(a)$ is the query.

## Qualitative Results




Temporal attention values for several action queries, indicated by the intensity of the color. Using the narrated action as a query, we display to top-5 predicted actions and the correctly predicted adverb.




The embedding space after training for videos narrated with the action 'cook'. Pairs 'completely'/'partially' and 'quickly'/'slowly' are highlighted separately.

## References

[1] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *CVPR*, 2019.
[2] Hazel Doughty, Walterio Mayol-Cuevas, and Damen Damen. The pros and cons: Rank-aware temporal attention for skill determination in long videos. In *CVPR*, 2019.
[3] Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. In *CVPR*, 2017.
[4] Tushar Nagarajan and Kristen Grauman. Attributes as operators: factorizing unseen attribute-object compositions. In *ECCV*, 2018.
[5] Sujoy Paul, Sourya Roy, and Amit K Roy-Chowdhury. W-talc: Weakly-supervised temporal activity localization and classification. In *ECCV*, 2018.
[6] Daochang Liu, Tingtin Jiang, and Yizhou Wang. Completeness modeling and context separation for weakly supervised temporal action localization. In *CVPR*, 2019.
[7] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019.

## Results

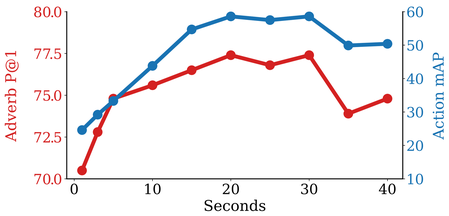
| Method | video-to-adverb | | adverb-to-video | |
|---|---|---|---|---|
| | Antonym | All | Antonym | All |
| Chance | 50.0 | 40.8 | 51.1 | 17.0 |
| Classifier-SVM | 60.5 | 53.2 | 56.3 | 26.4 |
| Classifier-MLP | 68.5 | 60.2 | 60.3 | 30.4 |
| RedWine [3] | 69.3 | 59.4 | 59.5 | 29.0 |
| LabelEmbed [3] | 71.7 | <u>62.1</u> | <u>61.8</u> | 29.7 |
| AttributeOp [4] | <u>72.8</u> | 61.2 | 59.7 | **35.0** |
| Ours | **80.8** | **71.9** | **65.7** | <u>32.9</u> |

Comparison with methods designed to learn object-attribute pairs. The best performance for each metric is highlighted in **bold**, second best is <u>underlined</u>.

## Ablation Study

| Method | Action | Adverb |
|---|---|---|
| Single | 24.6 | 70.5 |
| Average | 25.7 | 71.6 |
| Class-agnostic attention | 23.5 | 70.8 |
| Class-specific Attention | 40.1 | 72.8 |
| Ours | **69.2** | **80.8** |

Comparison of temporal attention methods




Performance as the amount of video used around the narrated timestamp increases. With a larger window, videos are more likely to contain the relevant action, but also other actions.

| Method | Attention | Adverb Rep | P@1 |
|---|---|---|---|
| W-TALC [5] | Avg | Classifier-MLP | 70.5 |
| | Avg | Action Modifiers | 73.9 |
| | SDP | Action Modifiers | 76.8 |
| CMCS [6] | Avg | Classifier-MLP | 69.6 |
| | Avg | Action Modifiers | 69.9 |
| | SDP | Action Modifiers | 70.5 |
| Ours | SDP | Action Modifiers | **80.8** |

Comparison of our method to weakly supervised action localization methods, with and without our scaled dot-product (SDP) attention and action modifier representations.

More info at: `hazeldoughty.github.io/Papers/ActionModifiers`