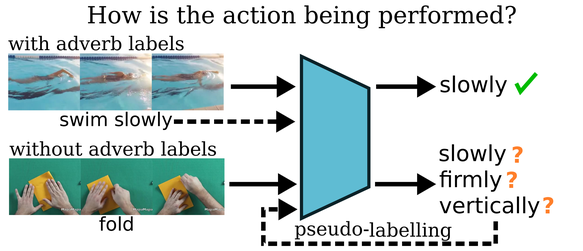
How Do You Do It? Fine-Grained Action Understanding with Pseudo-Adverbs

Hazel Doughty and Cees G. M. Snoek University of Amsterdam

University of Amsterdam

CVPR June 19-24 2022 New Orleans • Louisiana

Overview

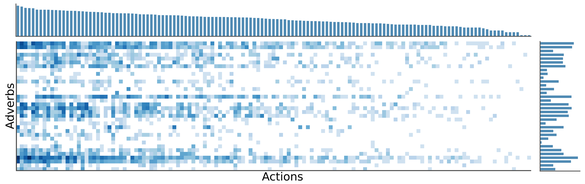


Idea: We identify subtle differences between actions by learning to recognize adverbs in a semi-supervised manner where we use action-only videos with multi-adverb pseudo-labeling.

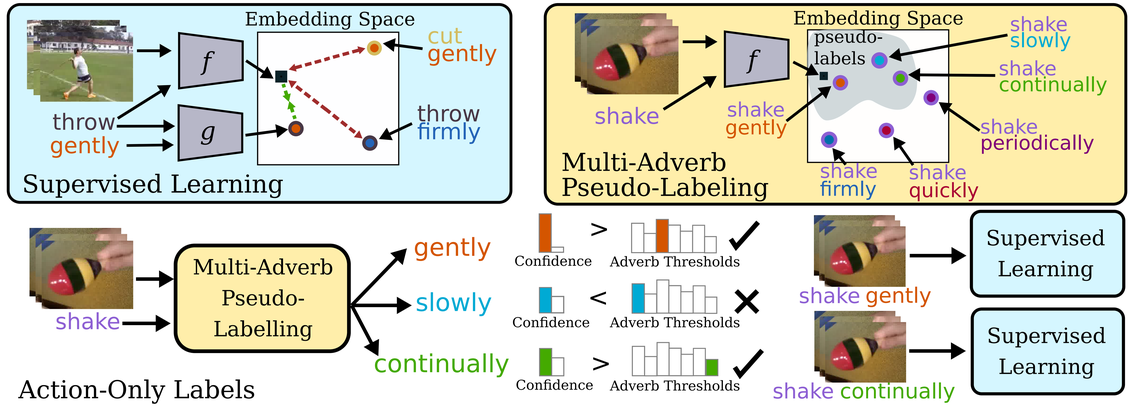
Three New Adverb Datasets

Dataset	Adverbs & Actions: Adverbs	Actions	Pairs	Videos: Clips	Length
HowTo100M Adverbs [1]	6	72	263	5,824	20.0
VATEX Adverbs	34	135	1,550	14,617	10.0
MSR-VTT Adverbs	18	106	464	1,824	15.7
ActivityNet Adverbs	20	114	643	3,099	37.3

Three new adverb datasets available at:
hazeldoughty.github.io/Papers/PseudoAdverbs



Semi-Supervised Learning of Adverbs



Supervised Learning
Video parts relevant to the action are embedded close to the ground-truth action-adverb text embedding.

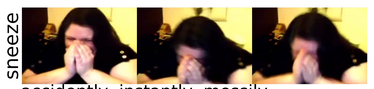
Action-Only Labels
For videos without adverb labels we create adverb pseudo-labels and use these in supervised learning

Multi-Adverb Pseudo-Labelling
Multiple adverbs can apply to an action, thus we take the top-k adverbs as pseudo-labels.

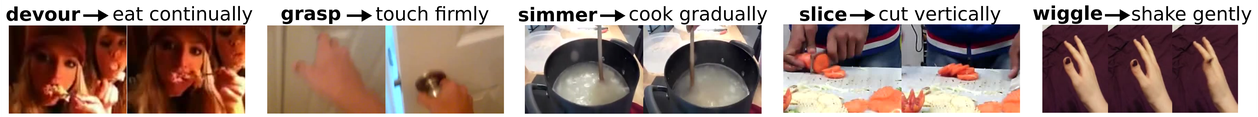
Adverb Thresholds
To cope with the long-tail, we use per-adverb thresholds to select which pseudo-labels we use.

Task I: Seen Compositions

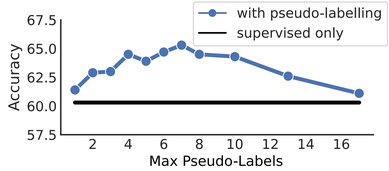
Method	VATEX Adverbs: 1%	2%	5%	10%	20%	Av.	HowTo100M Adverbs: 1%	2%	5%	10%	20%	Av.
Supervised only [1]	54.0	54.5	60.3	64.7	64.2	59.5	67.3	**68.5**	67.9	73.4	74.8	70.4
Pseudo-Label [2]	55.1	54.4	60.4	63.5	64.1	59.5	**69.3**	66.5	67.3	74.5	70.5	69.6
FixMatch [3]	**55.4**	52.3	61.2	62.8	64.8	59.3	68.2	67.9	67.3	74.5	75.9	70.7
TCL [4]	51.6	**56.6**	58.3	58.0	64.8	57.9	67.6	65.9	68.2	74.3	76.2	70.4
Ours	55.0	**56.6**	**63.9**	**65.3**	**67.5**	**61.7**	67.0	66.8	**69.9**	**77.1**	**79.1**	**72.0**



Describing Relationships Between Actions



Ablation Study



Multi-adverb pseudo-labeling improves results.

Threshold	Acc.
None	61.1
Fixed	61.4
Adaptive	63.9

Adaptive thresholds give better pseudo-labels

With multiple pseudo-labels and adaptive thresholding the distribution of pseudo-labels is more balanced.

Task II: Unseen Compositions

Method	Accuracy
Supervised only	52.2
Ours	56.1
Training with full labels	65.1

Our method improves generalization to unseen action-adverb compositions.

Task III: Unseen Domains

Method	MSR-VTT Adverbs	ActivityNet Adverbs
Source only	62.9	67.2
Pseudo-Label	63.9	66.4
Ours	65.0	66.6
Source + Target	67.5	71.6
Target only	70.5	71.8

Our method aids generalization to similar domains, but struggles with larger shifts.

References

[1] Hazel Doughty, Ivan Laptev, Walterio Mayol-Cuevas, and Dima Damen. Action modifiers: Learning from adverbs in instructional videos. In *CVPR*, 2020.
[2] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshops*, 2013.
[3] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *NeurIPS*, 2020.
[4] Ankit Singh, Omprakash Chakraborty, Ashutosh Varshney, Rameswar Panda, Rogerio Feris, Kate Saenko, and Abir Das. Semi-supervised action recognition with temporal contrastive learning. In *CVPR*, 2021.