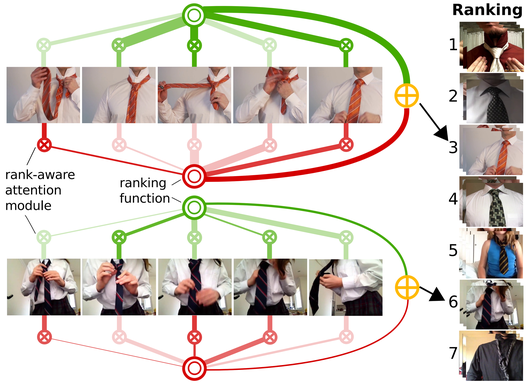
The Pros and Cons: Rank-aware Temporal Attention for Skill Determination in Long Videos

Hazel Doughty, Walterio Mayol Cuevas and Dima Damen University of Bristol, UK

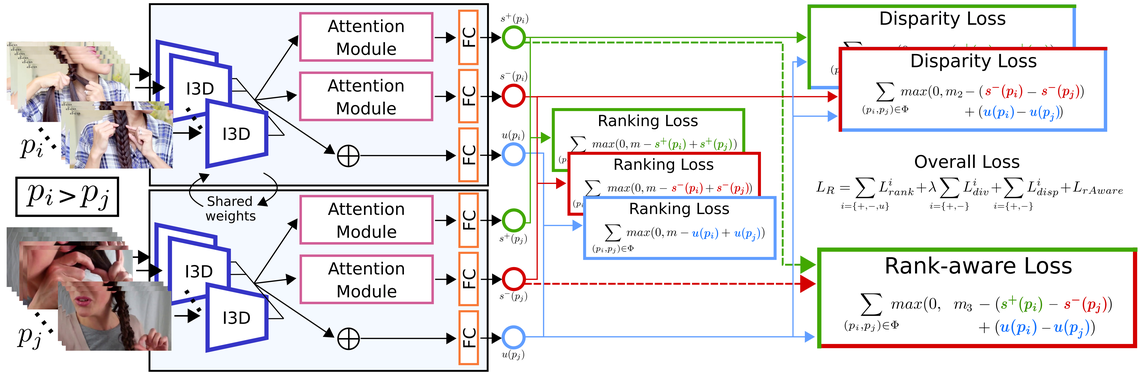
CVPR LONG BEACH CALIFORNIA June 16-20, 2019

Overview



Idea: Temporal attention modules trained to separately attend to video parts which are indicative or higher skill (pros) and lower skill (cons). Modules are fused (orange) for an overall skill assessment of the video.

Method

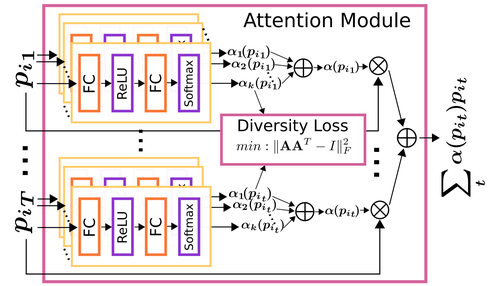


Ranking Loss
Standard ranking loss to ensure p_i is ranked higher than p_j.

Disparity Loss
Explicitly encourages each attention branch to outperform uniform weighting.

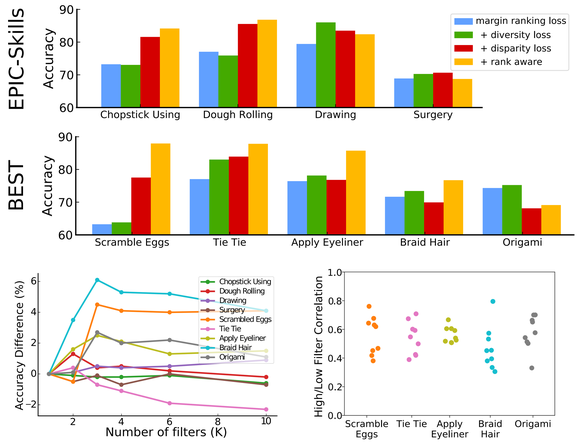
Rank-aware Loss
Optimizes the two attention modules to become rank-aware and attend to parts of the video indicating high or low skill.

Attention Module



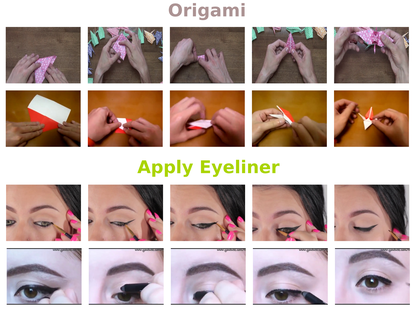
Quantitative Results

Method	EPIC Skills	BEST
Who's Better [1]	76.0	75.8
Last Segment	76.8	61.0
Uniform Weighting	78.8	73.6
Softmax Attention	74.5	72.3
STPN [2]	74.3	70.0
Ours (Rank Aware Attention)	**80.3**	**81.2**



BEST Dataset

	Task	#Videos	#Pairs	%Pairs	Av. Length (s)
EPIC-Skills	Chopstick Using	40	536	69%	46 ± 17
	Dough Rolling	33	181	34%	102 ± 29
	Drawing	40	247	65%	101 ± 47
	Surgery	103	1659	95%	92 ± 41
BEST	Scramble Eggs	100	2112	43%	170 ± 113
	Tie Tie	100	3843	77%	81 ± 47
	Apply Eyeliner	100	3743	76%	122 ± 105
	Braid Hair	100	3847	78%	179 ± 91
	Origami	100	3237	65%	386 ± 193



Dataset and Code available at: hazeldoughty.github.io

Qualitative Results

References

[1] Hazel Doughty, Dima Damen, and Walterio Mayol-Cuevas. Who's Better? Who's Best? Pairwise Deep Ranking for Skill Determination. In *CVPR*, 2018.
[2] Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han. Weakly Supervised Action Localization by Sparse Temporal Pooling Network. In *CVPR*, 2018.