Weakly-Supervised Representation Learning for Video Alignment and Analysis

Guy Bar-Shalom[1,2], George Leifman[2], Michael Elad[2]

Technion[1], Verily[2]

Introduction

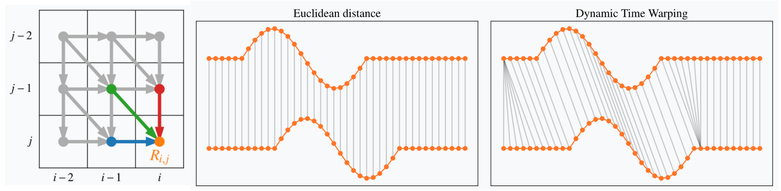
This paper presents *Learning Representations by position PROPagation* (**LRProp**), a novel weakly-supervised method for frame-based feature learning in video analysis, using a transformer encoder and a variant of the DTW algorithm for temporal alignment of video pairs, improving performance in various downstream applications.

Preliminaries

Weak-supervision. Weakly supervised learning involves cases in which the videos of interest consist of the same action category sequence. In these cases, we are given an ordered list of actions during training, but the exact temporal boundaries or paste of each action are not provided.

DTW. Dynamic Time Warping (DTW) is a well-known technique for measuring the distance between two sequences which may vary in speed,

$$\text{DTW}(i,j) = d(i,j) + \min\{\text{DTW(i,j-1)}, \text{DTW(i-1,j)}, \text{DTW(i-1,j-1)}\}$$



SoftDTW. SoftDTW [2] is a differentiable DTW variant, which employs a soft, differentiable minimum function regulated by a parameter γ:

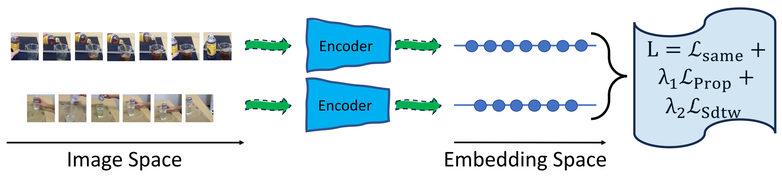
$$\min\{a_1, a_2, \ldots, a_n\} \xrightarrow{\gamma} \min^{\gamma}\{a_1, a_2, \ldots, a_n\} = -\gamma \cdot \log\left(\sum_{i=1}^{n} \exp\left(-\frac{a_i}{\gamma}\right)\right).$$

Contributions

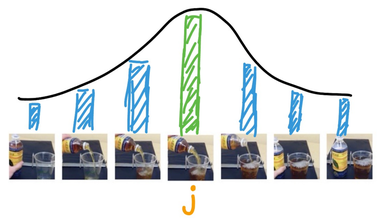
1. We present a general weakly-supervised framework for learning frame-wise representations with a focus on video alignment.
2. The proposed *pair-wise position propagation* is shown to result in features that offer better temporal awareness compared to prior work.
3. Our approach achieves superior performance to the state-of-the-art on various temporal understanding tasks on the Pouring and PennAction datasets, setting a new performance benchmark for downstream tasks.

LRProp

General idea. We consider a pair of weakly supervised videos. We pose the question: what distribution should their embeddings follow? We enforce this distribution over the embedding space using D-KL divergence, via three loss functions, $\mathcal{L}_{\text{Same}}, \mathcal{L}_{\text{Prop}}, \mathcal{L}_{\text{Sdtw}}$.



$\mathcal{L}_{\text{Same}}$ - Prior Distribution for Frames in the Same Video.



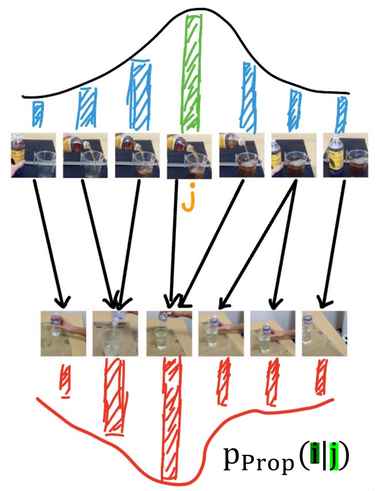
$$p(\text{i}|\text{j}) \propto \exp\left\{(\text{j}-\text{i})^2/2\sigma^2\right\}$$

$$q_\theta(\text{i}|\text{j}) \propto \exp\left\{\text{sim}(z_{\text{j}}, z_{\text{i}})/\tau\right\}$$

$$\mathcal{L}^{\text{j}}_{\text{Same}} = D_{\text{KL}}\left[q_\theta(\text{i}|\text{j}) \parallel p(\text{i}|\text{j})\right]$$

$$\mathcal{L}_{\text{Same}} = \sum_{\text{j}=1}^{n} \mathcal{L}^{\text{j}}_{\text{Same}}$$

$\mathcal{L}_{\text{Prop}}$ - Prior Distribution for Frames in Different Videos.



Incorporating the previously established $p(\text{i}|\text{j})$ and the DTW algorithm, our methodology constructs a prior distribution for a second video, informed by the first. This is represented as $p_{\text{Prop}}(\text{i}|\text{j})$, as illustrated at the bottom of the Figure. Specifically, our process involves the following steps:

- **Alignment Path Extraction**: We derive the alignment path between the two videos using the DTW algorithm. This path defines the similarity between frames across the videos.
- **Pair-Wise Position Propogation**: Based on the DTW alignment path, we propagate the frame bins from the first video to the second. This step is crucial for mapping similarities between corresponding frames.
- **Prior Distribution Definition**: The result of this process is formalized as $p_{\text{Prop}}(\text{i}|\text{j})$. This function acts as a prior distribution, quantifying the similarity between frame i in the second video and frame j in the first video – the similarity is visualized by the height of the bin.

Finally,

$$\mathcal{L}^{\text{j}}_{\text{Prop}} = D_{\text{KL}}\left[q_\theta(\text{i}|\text{j}) \parallel p_{\text{Prop}}(\text{i}|\text{j})\right]$$

$$\mathcal{L}_{\text{Prop}} = \sum_{\text{j}=1}^{n} \mathcal{L}^{\text{j}}_{\text{Prop}}$$

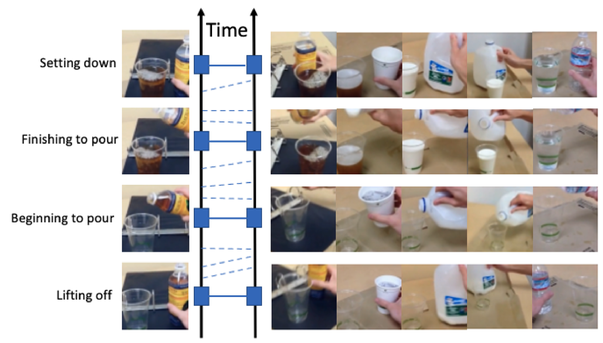
$\mathcal{L}_{\text{Sdtw}}$ - Can we Trust the alignment path?

Lets learn it! The final component of our loss function is the SoftDtw loss component, which implicitly learns a better alignment path. Given that the length of the two videos is n, m, it can be formulated as follows,

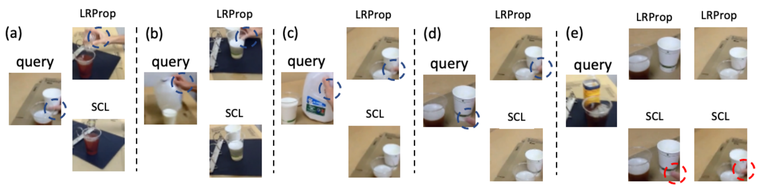
$$\mathcal{L}_{\text{Sdtw}} = \text{DTW}^{\gamma}(n,m) = d(n,m) + \min^{\gamma}\{\text{DTW}^{\gamma}(n,m-1), \text{DTW}^{\gamma}(n-1,m), \text{DTW}^{\gamma}(n-1,m-1)\}$$

Experiments

Video alignment on the Pouring dataset with **LRProp** features and the DTW algorithm shows successful capture of key events across videos.



Aligned frames in the Pouring dataset using DTW algorithm demonstrate **LRProp**'s superior accuracy in capturing actions, compared to SCL [1], as shown by blue and red circles.



Results on the Pouring dataset.

Evaluation metrics for video frame representations include Phase Classification Accuracy, Phase Progression, Average Precision@K, Kendall's Tau, and DTW Accuracy. Our method surpasses others, achieving 93.88% Phase Classification Accuracy with just 25% of labels, and shows significant improvements in semantic frame identification as depicted in the Average Precision@K column. We also see gains in Kendall's tau and Phase progression metrics compared to SCL [1], which has already shown a phenomenal improvement of more than 10% over all previous methods. Finally, a notable 6% increase in DTW Accuracy, indicating that using our frame-wise features will result in a more precise video alignment.

Method	τ	*Progress*	*AP@K*			*Classification@*				*DTW A*
			K=5	*K=10*	*K=15*	*10*	*25*	*50*	*100*	
SCL	<u>99.2</u>	<u>93.5</u>	<u>90.04</u>†	<u>89.69</u>†	88.92†	85.78†	<u>87.14</u>†	89.45†	<u>93.73</u>	<u>84.68</u>†
SAL	79.61	77.28	84.05	83.77	83.79	87.63	-	87.58	88.81	-
TCN	85.12	80.44	83.56	83.31	83.01	89.67	-	87.32	89.53	-
TCC	86.36	83.73	87.16	86.68	86.54	90.65	-	91.11	91.53	-
LAV	85.61	80.54	89.13	89.13	<u>89.22</u>	91.61	-	<u>92.82</u>	92.84	-
VAVA	87.55	83.61	-	-	-	<u>91.65</u>	-	91.79	92.84	-
LRProp	99.46	94.09	92.41	90.33	90.86	92.7	93.88	94.44	94.36	90.22

Scan for paper

References.

[1] Minghao Chen, Fangyun Wei, Chong Li, and Deng Cai. Frame-wise action representations for long videos via sequence contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13801–13810, 2022.

[2] Marco Cuturi and Mathieu Blondel. Soft-dtw: a differentiable loss function for time-series. In *International conference on machine learning*, pages 894–903. PMLR, 2017.