Supplementary Materials: Self-Supervised Multi-Task Procedure Learning from Complex Activity Videos via DNNs

Ehsan Elhamifar[0000−1111−2222−3333] and Dat Huynh[1111−2222−3333−4444]

Khoury College of Computer Sciences, Northeastern University, Boston, USA
{e.elhamifar,huynh.dat}@northeastern.edu

Key-step Localization Network. To build the key-step localization network (KLN), we make connection between our goal, which is to take a T_ℓ input vectors and output T_ℓ outputs each of dimension M, and semantic image segmentation whose goal is to take an input image and produce an output image where each pixel takes one of few discrete values corresponding to a category. Thus, we take the network in [**?**] and, given that we are working with sequential data instead of 2D images, we convert 2D convolutions, 2D pooling and 2D deconvolutions to 1D temporal convolutions, 1D pooling and 1D deconvolutions, respectively, see Figure 1.

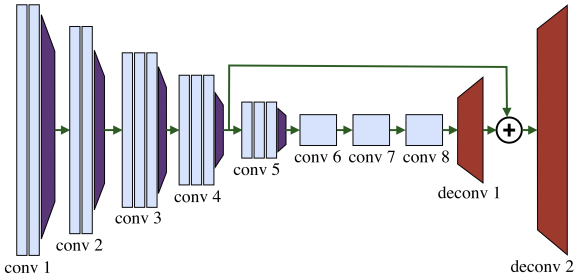

Fig. 1: Our key-step localization network (KLN) takes T_ℓ input vectors, corresponding to feature vectors from T_ℓ segments/frames of video ℓ, and generates T_ℓ outputs each of dimension M, where the t-th output encodes the assignment probabilities of segment/frame t to each of the M latent states obtained from videos.

As the figure shows, our KLN consists of multiple consecutive layers of 1D convolutions and 1D pooling, forming the encoding part of the network, followed by 1D deconvolutions, forming the decoding part of the network. The KDN is organized similar to [**?**]. The first five convolutional subnetworks (conv1 to conv5) each consist of multiple temporal convolutional layers, where each temporal convolution is followed by batch normalization and ReLU activation. Each convolutional subnetwork is followed by a temporal max-pooling. The subnetworks conv6 and conv7, each consists of a temporal convolution followed by ReLU and dropout. The subnetwork conv8 consists of 1×1 convolution with batch normalization. We apply deconvolution along the time axis on the output of conv8. We also apply a 1×1 convolution and batch normalization to the output of pool4 and add (element-wise) the result with deconv1 features.

Fig. 2: Visualization of the self-supervised learned attention model on two videos from the task 'assemble clarinet' (left) and 'perform CPR' (right) from ProceL. Notice that our method successfully learns to focus on important region of each frame. For example, for clarinet, it focused on cork, ligature, screws, lower and upper joints in the associated key-steps.

	$k=7$		$k=10$		$k=12$		$k=15$	
	Recall	Jaccard	Recall	Jaccard	Recall	Jaccard	Recall	Jaccard
Random	14.8	3.3	10.8	2.9	9.2	2.8	7.5	2.6
JseqFL	29.6	6.5	26.6	6.8	23.0	6.6	21.3	5.9
TEC	**34.0**	**7.4**	28.5	6.9	27.2	**7.1**	25.5	6.7
Ours	33.3	7.1	**31.6**	**7.2**	**29.8**	6.9	**32.2**	**7.6**
Ours (multi-task)	*41.7*	*8.9*	*41.7*	*8.9*	*41.7*	*8.9*	*41.6*	*8.9*

Table 1: MoF and Jaccard (%) on CrossTask for different number of key-steps, k.

		$k=7$		$k=10$		$k=12$		$k=15$	
		F1	Acc	F1	Acc	F1	Acc	F1	Acc
Attention	Cls	9.3	89.2	9.3	89.2	9.3	89.2	9.3	89.2
	Cls+Self-Sup	11.3	88.3	**12.6**	89.6	**12.9**	88.3	10.4	89.2
Learnable Pooling	Cls	11.0	91.7	11.0	91.7	11.0	91.7	11.0	91.7
	Cls+Self-Sup	13.3	91.7	12.1	90.8	12.0	90.8	**12.5**	90.8
Attention + Learnable Pooling	Cls	7.3	89.2	7.3	89.2	7.3	89.2	7.3	89.2
	Cls+Self-Sup	**14.0**	**92.6**	12.4	**91.7**	12.8	**93.3**	11.8	**93.8**

Table 2: Localization score (MoF) and classification accuracy (Acc), in precent, of different algorithms on the ProceL dataset for different number of key-steps, k.

This skip connection, used in semantic segmentation to produce better visual features, is also useful in key-step discovery, as it helps to recover temporal information for key-step and video classification. Finally, we apply a temporal deconvolution and obtain the final predictions, which are T_ℓ outputs each of dimension M.

More Results from Visual Attention. Figure 2 shows more results for the visualization of our self-supervised learned attention model on two videos from the task 'assemble clarinet' (left) and 'perform CPR' (right) from ProceL. Notice that our method successfully learns to focus on important region of each frame. For example, for clarinet, it focuses on cork, ligature, screws, lower and upper joints in the associated key-steps. This is particularly vital for localization of key-steps and recognition of the task, as also demonstrated by our quantitative results.

		$k=7$		$k=10$		$k=12$		$k=15$	
		F1	Acc	F1	Acc	F1	Acc	F1	Acc
Attention	Cls	4.3	75.5	4.3	75.5	4.3	75.5	4.3	75.5
	Cls+Self-Sup	16.0	74.3	11.1	71.7	11.4	71.7	13.8	73.8
Learnable Pooling	Cls	7.1	72.3	7.1	72.3	7.1	72.3	7.1	72.3
	Cls+Self-Sup	13.8	70.1	10.4	66.6	12.5	72.9	10.6	72.8
Attention	Cls	12.8	**79.9**	12.8	79.9	12.8	**79.9**	12.8	**79.9**
+ Learnable Pooling	Cls+Self-Sup	**16.2**	79.1	**16.3**	**80.4**	**16.2**	79.5	**16.3**	77.9

Table 3: Localization score (MoF) and classification accuracy (Acc), in precent, of different algorithms on the CrossTask dataset for different number of key-steps, k.