Supplementary Material for Temporal Feature Enhancement Dilated Convolution Network for Weakly-supervised Temporal Action Localization

In this supplementary material, we provide more implementation details of the proposed Temporal Feature Enhancement Dilated Convolution Network (TFE-DCN). We also conduct additional experiments on hyper-parameters to analyze their effect on TFE-DCN.

1. More Implementation Details

In our experiments, we use the network backbone I3D [1], which is pre-trained on Kinetics [1], to extract features from RGB frames and optical flow. Following standard pipeline [1], we use a fixed FPS of 25 for I3D rather than the original FPS. The optical flow is estimated via the TV-L1 algorithm [2].

In our implementation, we use a set of temporal 1D convolution layers to transform the original features to enhanced features. In TFE-DC Module and Modality Enhancement Module, each temporal 1D convolution layer consists of 1024 convolutional kernels with a temporal kernel size of 3. Each convolution layer is followed by a leaky ReLU layer with ratio of 0.2 and a dropout layer with a dropout rate of 0.5. The filtering module consists of three temporal 1D convolution layers and one sigmoid function. The classifier consists of two temporal 1D convolution layers. A temporal 1D convolution layer and two dropout layers with a dropout rate of 0.7 are added before the classification layer as the fusion layer.

2. Experiments on Hyper-parameters a_i

We conduct additional experiments on hyper-parameters to analyze their effect on TFE-DCN. Ablation studies are conducted on the testing set of the THUMOS'14 dataset. In our paper, we have demonstrated the effectiveness of each component in the final loss function by setting corresponding hyper-parameter λ_i to be 0. Here we want to study the effect of hyper-parameters a_i $(i = 1, 2, 3)$ in Equation. 6 in our paper. The temporal attention weights A_n^{Flow} is the weighted average of $\{A_{n,k}^{Flow}\}_{k=1}^{K}$, where K is the number of dilated convolution layers and $A_{n,k}^{Flow}$ is the attention weights generated by the k-th layer.

In our implementation, we set $K = 3$ and $a_1 = a_2 = a_3 = \frac{1}{3}$ so that A_n^{Flow} can equally contain temporal attention weights from all layers. To study the effect of each a_i, we leave one set the other two to be 0. Table 1 shows the comparison between different settings of a_i. The results reveal that attention weights generated by each layer do have effect on TFE-DCN. It can be viewed that the attention weights generated by the third layer can achieve better results than the other two. This is because the third layer has a larger receptive field, which can cover more temporal information and capture dependencies between snippets. And using all layers' attention weights to obtain A_n^{Flow} can help our model achieve the best performance.

Table 1: Comparison of different settings of a_i on THUMOS'14 testing set

a_1	a_2	a_3	mAP@IoU (%)		
			0.1:0.1:0.5	0.1:0.1:0.7	0.1:0.1:0.9
1	0	0	55.0	44.8	35.8
0	1	0	55.5	45.2	36.2
0	0	1	56.0	45.5	36.3
1/3	1/3	1/3	**57.5**	**46.9**	**37.5**

3. Qualitative Results with Video Demo

To give reviewers a direct view of the performance of our model, we make a video demo to show the qualitative results of two samples. The first sample is video_test_0000569.mp4 from the testing set of THUMOS'14, which contains category 'Cricket Bowling' and 'Cricket Shot' and each action instance of these two categories is extremely short. The second sample is video_test_0000896.mp4 from the testing set of THUMOS'14, which contains the category 'High Jump', and each action instance of this class is relatively long.

We use a moving red-brown line to represent the current timestamp. The five boxes represent the activation scores of the baseline (Bas-Net [3]), the proposals of the baseline, the activation scores of TFE-DCN, the proposals of TFE-DCN and ground truth in order from top to bottom. This video demo demonstrates the outstanding performance of our proposed TFE-DCN.

4. Questions and Answers about Ablation Study

i. Why the dilated module should be applied to enhance flow features? What would happen when it is applied to RGB features?

Ans: We apply dilated module to enhance flow features instead of RGB features because we think optical flow reflects temporal information and RGB reflects spatial information. In Table.5 of our paper, we tested “Exchange Modalities”. We applied TFE-DC module to enhance RGB features and Modality Enhancement Module to enhance optical flow features. The performance has dropped a lot, from 46.9% to 43.2% (AVG 0.1:0.1:0.7).

ii. What if we use conventional convolutions rather than dilated convolutions in TFE-DC Module?

Ans: Dilated convolutions can enlarge the receptive field effectively and are appropriate for temporal modeling. If we use conventional convolutions rather than dilated convolutions in TFE-DC Module (K=3), the performance decreases from 57.5% to 56.9% (AVG 0.1:0.1:0.5).

Table 2: Comparison between dilated convolutions and conventional convolutions (K=3).

	mAP@IoU (%)				AVG	
	0.1	0.3	0.5	0.7	0.1:0.1:0.5	0.1:0.1:0.7
Dilated Convolutions	**72.3**	**58.6**	**40.7**	**13.7**	**57.5**	**46.9**
Conventional Convolutions	72.2	58.1	39.7	13.4	56.9	46.4

iii. How about the issue of adjoining actions being misunderstood as a long action instance?

Ans: Up to now, many works fail to handle this issue, but our model can localize each adjoining action instance well without mixing (on the right side of Fig.5 in our paper). TFE-DC Module not only enlarges the receptive field, but also captures dependencies between snippets with different temporal scales. As seen in Fig.3 in our paper, each scale sends their features to influence the final temporal attention A_n^{Flow}. If common multi-layer dilated convolution is used here, only the final layer influences A_n^{Flow} and then the performance decreases (see Table.1 in supplementary material).

References

[1] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4724-4733, doi: 10.1109/CVPR.2017.502.

[2] C. Zach, T. Pock, and H. Bischof. 2007. A duality based approach for realtime TV-L1 optical flow. In Proceedings of the 29th DAGM conference on Pattern recognition. Springer-Verlag, Berlin, Heidelberg, 214–223.

[3] Lee, P., Uh, Y., & Byun, H. (2020). Background Suppression Network for Weakly-Supervised Temporal Action Localization. Proceedings of the AAAI Conference on Artificial Intelligence, 34(07), 11320-11327.