Supplementary
PDAN: Pyramid Dilated Attention Network for Action Detection

Rui Dai[1,2], Srijan Das[1,2], Luca Minciullo[3], Lorenzo Garattoni[3], Gianpiero Francesca[3], François Bremond[1,2]

[1]Inria [2]Université Côte d'Azur [3]Toyota Motor Europe

{name.surname}@inria.fr {name.surname}@toyota-europe.com

1. NL ablation baseline structure

In Fig. 1, we compare the structure of PDAN (STCL) with the one of PDAN (STCL) + NL-T1, PDAN (STCL) + NL-T2 and PDAN (DAL).

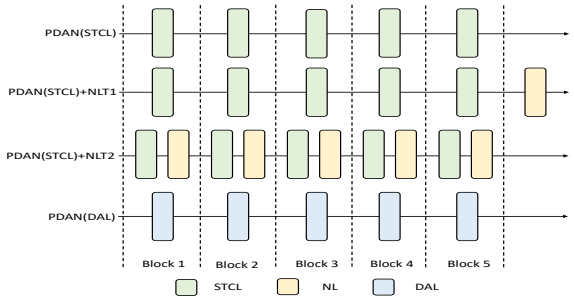

Figure 1. Structure of PDAN (STCL), PDAN (STCL) + NL-T1, PDAN (STCL) + NL-T2 and PDAN (DAL). For simplicity, the residual links are not drawn in this figure.

2. Timeception for action detection

Inspired by Inception [10], Timeception [5] utilizes several convolution layers in a parallel manner. This structure enables Timeception to explore multi-temporal scales in a single block. However, Timeception is designed for action classification, it has a max pooling layer in every block to halve the temporal resolution and aggregate the temporal information. For comparison with Timeception on the action detection task (as in this paper), we adopt the Timeception version [5], which employs dilated convolutions and which has multiple convolution layers with same kernel size and different dilation rates. Moreover, we removed the max pooling layer in the Timeception layer to keep the temporal resolution for action detection task. To confirm the effectiveness of our DAL, we replace the standard temporal convolution by DAL to obtain Timeception (DAL). The structure of Timeception (DAL) is shown on Fig. 2. We stack three adapted Timeception layers along with a classifier layer for the action detection task.

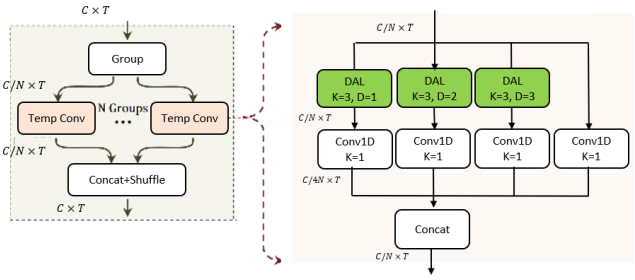

Figure 2. Timeception layer. K, D indicate kernel and dilation rates respectively. Depending whether the blocks in green are standard temporal convolution layers (STCLs) or DALs, we have Timeception (STCL) and Timeception (DAL) respectively.

3. Ablation study: two modalities

In MultiTHUMOS and Charades, we use two-stream structure. Table 1 reports the result of different streams (RGB/Flow) for PDAN with or without attention layer. First, we observe that attention layer can improve performance for both RGB and Flow streams on both datasets (about 2% improvement w.r.t. RD-TCN). Second, for sport actions in MultiTHUMOS, Flow stream yields better performance than RGB stream (+4.3% for PDAN). Third, for object-based actions with low motion in Charades, RGB stream achieves better performance (+3.8% w.r.t. Flow stream for PDAN), which indicates that RGB can better model the object appearance information, especially for low motion frames.

4. Baseline implementation

Six activity detection methods are evaluated on TSU dataset [1], namely, I3D baseline, LSTM [3], Dilated-

Table 1. Performance of different streams for RD-TCN (i.e. PDAN without attention) and PDAN on MultiTHUMOS and Charades datasets.

	MultiTHUMOS		Charades	
	RD-TCN	PDAN	RD-TCN	PDAN
RGB stream	38.5	40.2	21.4	**23.7**
Flow stream	43.9	**44.5**	17.4	19.9
Fusion	46.6	**47.6**	24.1	**26.5**

TCN [6], Super event [8], TGM [9] and MS-TCN [2]. Different from [1], We fine-tune their parameters on TSU to optimize their performance. Here are the optimal settings: I3D Baseline has only one dropout layer (with dropout probability 0.5) before the classifier. LSTM [4] has one LSTM layer with 512 hidden units and one dropout layer (with dropout probability 0.5). Similarly, for For TGM [9], on TSU dataset, we add one layer to have a 4-layer structure. For Dilated-TCN [6], we adopt a 5 block structure, each block has 3 layers. A residual link connects the input of the first layer and the output of the last layer in each block. We increase the number of filters to 256 per layer to model complex temporal relations. Similarly, for MS-TCN [2], we keep the structure of 5 stages and 10 layers/stage, increasing the number of filters to 256 per layer. All the baselines use the same video encoding as the proposed method and they are trained with binary cross-entropy loss with sigmoid activation [7]. The unspecified parameters are similar to the original papers.

References

[1] Rui Dai, Srijan Das, Saurav Sharma, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome untrimmed: Real-world untrimmed videos for activity detection. *arXiv preprint arXiv:2010.14982*, 2020.

[2] Yazan Abu Farha and Jurgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3575–3584, 2019.

[3] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, Nov. 1997.

[4] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[5] Noureldien Hussein, Efstratios Gavves, and Arnold WM Smeulders. Timeception for complex action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 254–263, 2019.

[6] Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165, 2017.

[7] Jinseok Nam, Jungi Kim, Eneldo Loza Mencía, Iryna Gurevych, and Johannes Fürnkranz. Large-scale multi-label text classification—revisiting neural networks. In *Joint european conference on machine learning and knowledge discovery in databases*, pages 437–452. Springer, 2014.

[8] AJ Piergiovanni and Michael S Ryoo. Learning latent super-events to detect multiple activities in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5304–5313, 2018.

[9] AJ Piergiovanni and Michael S. Ryoo. Temporal gaussian mixture layer for videos. In *International Conference on Machine Learning (ICML)*, 2019.

[10] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.