Weakly-Supervised Temporal Action Localization Through Local-Global Background Modeling

Xiang Wang[1,2] Zhiwu Qing[1,2] Ziyuan Huang[2] Yutong Feng[2] Shiwei Zhang[2*]
Jianwen Jiang[2] Mingqian Tang[2] Yuanjie Shao[1] Nong Sang[1*]
[1] Key Laboratory of Image Processing and Intelligent Control
School of Artificial Intelligence and Automation, Huazhong University of Science and Technology
[2]Alibaba Group
{wxiang, qzw, shaoyuanjie, nsang}@hust.edu.cn
{pishi.hzy, fengyutong.fyt, zhangjin.zsw, jianwen.jjw, mingqian.tmq}@alibaba-inc.com

Abstract

*Weakly-Supervised Temporal Action Localization (WS-TAL) task aims to recognize and localize temporal starts and ends of action instances in an untrimmed video with only video-level label supervision. Due to lack of negative samples of background category, it is difficult for the network to separate foreground and background, resulting in poor detection performance. In this report, we present our 2021 HACS Challenge - Weakly-supervised Learning Track [24] solution that based on BaSNet [10] to address above problem. Specifically, we first adopt pre-trained CSN [17], Slowfast [5], TDN [19], and ViViT [1] as feature extractors to get feature sequences. Then our proposed Local-Global Background Modeling Network (LGBM-Net) is trained to localize instances by using only video-level labels based on Multi-Instance Learning (MIL). Finally, we ensemble multiple models to get the final detection results and reach **22.45%** mAP on the test set.*

1. Introduction

Video understanding is an important area in computer vision, including many sub-research directions, such as Action Recognition [20, 5, 7], Temporal Action Detection [23, 11, 15, 22], Spatio-Temporal Action Detection [16, 9], *etc*. In this report, we introduce our method for the temporal action detection task with only video-level supervision, termed weakly-supervised temporal action localization (WS-TAL).

Since the setting of weak supervision is more in line with real needs, WS-TAL has attracted more and more attention. Recently, several methods [12, 8, 14, 13] were developed to localize instances in untrimmed videos using the video-level labels. However, though these methods have achieved significant performance, there is still a performance gab between fully-supervised methods [11, 15, 23]. We attribute this to that there are plenty of foreground-background confusions. It is challenging based on only video-level labels to separate action and background. To address this issue, we improve the mainstream approach BaSNet [10] and propose Local-Global Background Modeling Network (LGBM-Net), which integrates two-branch weight sharing local-global sub-net and a local-global attention module to suppress background and improve the discrimination of actions.

2. Feature Extractor

Following recent WS-TAL methods [14, 13], given an untrimmed video V, we first divide it into multiple snippets based on a pre-defined sampling ratio, and then apply pre-trained networks to extract snippet-level features. Next, we briefly introduce the feature extraction networks we used in the competition.

2.1. Channel-separated convolutional network

Inspired by the accuracy gains and good computational savings demonstrated by 2D separable convolutions in image classification. Du *et al*. [17] propose a set of architectures for video classification – 3D Channel-Separated Networks (CSN) – in which all convolutional operations are separated into either pointwise $1 \times 1 \times 1$ or depthwise $3 \times 3 \times 3$ convolutions. CSN shows that excellent accu-

∗ Corresponding authors.
This work is supported by Alibaba Group through Alibaba Research Intern Program.
This work is done when X. Wang (Huazhong University of Science and Technology), Z. Qing (Huazhong University of Science and Technology), Z. Huang (National University of Singapore) and Y. Feng (Tsinghua University) are interns at Alibaba Group.

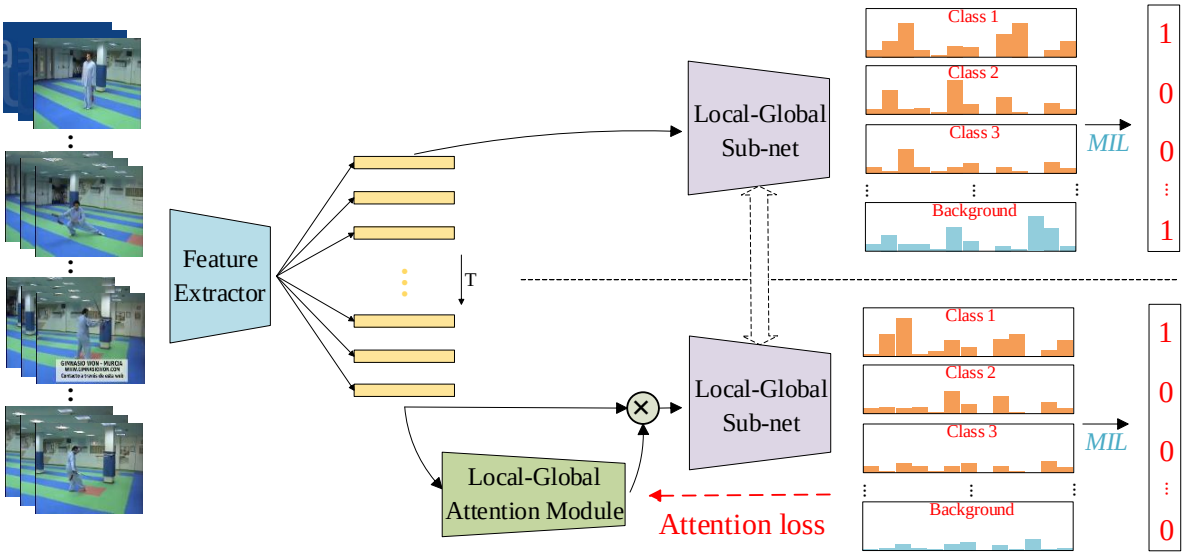

Figure 1. The overall architecture of our Local-Global Background Modeling Network (LGBM-Net). Using a pre-trained model, we extract the clip features for the input video, which are then fed into two branches. Both branches, sharing local-global sub-net weights, produce class activation sequence (CAS) to predict video-level labels. Note that the ground-truth background category of the upper branch is always 1, and the ground-truth background category of the lower branch is always 0. For some other details, you can refer to BaSNet [10].

racy/computational cost balances can be obtained by leveraging channel separation to reduce FLOPs and parameters as long as high values of channel interaction are retained. Due to its excellent performance in action recognition, we use Kinetics400 [3] pre-trained CSN as one of our feature extractors.

2.2. Slowfast

Slowfast [5] model involves two pathways operating at different frame rates. One path-way is designed to capture semantic information that can be given by images or a few sparse frames, and it operates at low frame rates and slow refreshing speed. In contrast, the other pathway is responsible for capturing rapidly changing motion, by operating at fast refreshing speed and high temporal resolution. Note that despite its high temporal rate, this pathway is made very lightweight. This is because this pathway is designed to have fewer channels and weaker ability to process spatial information, while such information can be provided by the first pathway in a less redundant manner. In the competition, We use Slowfast101 pre-trained on Kinetics400 dataset and Slowfast152 pre-trained on Kinetics700 [2] dataset as backbone.

2.3. Temporal Difference Network

Temporal Difference Network (TDN [19]) focuses on capturing multi-scale temporal information for action recognition. The core of TDN is to devise an efficient temporal module by explicitly leveraging a temporal difference operator, and systematically assess its effect on short-term and long-term motion modeling. Meanwhile, TDN is established with a two-level difference modeling paradigm to fully capture temporal information over the entire video. Specifically, for local motion modeling, temporal difference over consecutive frames is used to supply 2D CNNs with finer motion pattern, while temporal difference across segments is incorporated to capture long-range structure for motion feature excitation. TDN provides a simple and principled temporal modeling framework and is selected as our backbone. In the competition, we pre-train TDN on Kinetics700 dataset.

2.4. ViViT

Inspired by the large-scale application and good effects of transformer [18] in the field of vision [4], ViViT [1] proposes to use Transformer as basic block to model the relations between temporal and space separately. ViViT is a pure Transformer based model for action recognition. We apply the ViViT-B/16x2 version with factorised encoder, which initialized from imagenet pretrained Vit [4], and then pre-train it on Kinetics700 dataset. Specifically, we use AdamW as our optimizer and set the base learning rate to 0.0001. The weight decay is set to 0.1. The training is warmed up with 2.5 epochs, with the start learning rate as 1e-6.

3. Local-Global Background Modeling Network

In this section, we will introduce the process of our Local-Global Background Modeling Network (LGBM-

Method	Feature	Pre-train	mAP@IoU (%)										Average mAP
			0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	
BaSNet	Slowfast101	K400	27.4	24.9	22.6	20.2	17.9	15.6	13.3	10.8	8.0	5.1	16.6
LGBM-Net	Slowfast101	K400	29.5	27.1	24.8	22.4	20.2	17.9	15.3	12.6	9.7	6.1	**18.6**
BaSNet	Slowfast152	K700	28.2	25.5	23.2	20.9	18.6	16.4	13.8	11.1	8.4	5.1	17.1
LGBM-Net	Slowfast152	K700	32.8	29.8	26.9	24.1	21.1	18.5	15.7	12.8	9.6	5.6	**19.7**
BaSNet	CSN	K400	29.9	27.2	24.9	22.7	20.2	17.8	14.9	12.4	9.5	5.4	18.5
LGBM-Net	CSN	K400	33.9	30.6	27.6	24.8	22.0	19.1	16.2	13.1	9.6	5.4	**20.2**
BaSNet	TDN	K700	27.5	24.5	22.0	19.7	17.5	15.2	12.8	10.5	7.9	4.7	16.2
LGBM-Net	TDN	K700	29.1	26.8	24.4	22.0	19.6	17.1	14.6	11.8	8.8	5.3	**18.0**
BaSNet	ViViT	K700	27.9	25.4	23.0	20.9	18.5	16.0	13.6	10.9	8.0	4.8	16.9
LGBM-Net	ViViT	K700	29.2	26.8	24.4	22.2	19.7	17.6	15.2	12.8	9.9	5.9	**18.4**
Ensemble	-	-	37.0	33.9	30.9	28.0	24.9	22.0	18.9	15.7	12.1	7.2	**23.0 (test: 22.45)**

Table 1. Performance comparison for different features on validation set of HACS [24] dataset in terms of mAP (%).

Net), and then give the experimental results. The architecture of LGBM-Net is showed in Figure 1.

3.1. Local-Global Attention Module

The goal of local-global attention module is to suppress background frames/segments by the opposite training objective for the background class. Local-global attention module consists of local operation (Conv) and global operation (LSTM [6]). Note that the two operations are trained in parallel and merged through two convolutional layers followed by sigmoid function. The output of the module is foreground weights which range from 0 to 1. At the same time, in order to train the attention of a specific category, we use the activation of the highest category in CAS to supervise the attention output in local-global attention module.

3.2. Local-Global Sub-net

Local-global sub-net is used to generate CAS, which can be used to predict segment-level class scores. Like local-global attention module, Local-global sub-net also contains local operation (Conv) and global operation (LSTM). Note that we also tried other global operations (*e.g.*, non-local [21] and global pooling) for the final ensemble. Afterwards, we aggregate segment-level class scores to derive a video-level class score. Here, we adopt top-k mean technique for training.

3.3. Detection Results

After getting the CAS, we can use the watershed algorithm to get the detection result. Note that we only use the CAS of the lower branch in Figure 1 to generate detection results because of separating foreground and background.

From Table 1, we can draw the following conclusions: (1) It can be seen from the results of Slowfast101 and Slowfast152 that large-scale pre-training and deeper models play a great role in improving performance. (2) The Transformer-based method (ViViT) generally has a slightly worse detection performance than the CNN methods (*e.g.*, CSN and Slowfast). (3) From the results of ensemble, it can be seen that there is complementarity between the models.

4. Conclusion

In this report, we propose LGBM-Net, which is based on BaSNet and can well separate the foreground and background. We conduct experiments on multiple features (*e.g.*, CSN, Slowfast, TDN and ViViT) to show the effectiveness of LGBM-Net. Particularly, through ensemble strategy, the detection performance can be further improved.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. *arXiv preprint arXiv:2103.15691*, 2021. 1, 2

[2] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 2

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Syl-

vain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[5] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6202–6211, 2019. 1, 2

[6] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232, 2016. 3

[7] Ziyuan Huang, Shiwei Zhang, Jianwen Jiang, Mingqian Tang, Rong Jin, and Marcelo Ang. Self-supervised motion learning from static images. *arXiv preprint arXiv:2104.00240*, 2021. 1

[8] Ashraful Islam, Chengjiang Long, and Richard Radke. A hybrid attention mechanism for weakly-supervised temporal action localization. *arXiv preprint arXiv:2101.00545*, 2021. 1

[9] Jianwen Jiang, Yu Cao, Lin Song, Shiwei Zhang, Yunkai Li, Ziyao Xu, Qian Wu, Chuang Gan, Chi Zhang, and Gang Yu. Human centric spatio-temporal action localization. In *ActivityNet Workshop on CVPR*, 2018. 1

[10] Pilhyeon Lee, Youngjung Uh, and Hyeran Byun. Background suppression network for weakly-supervised temporal action localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11320–11327, 2020. 1, 2

[11] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3889–3898, 2019. 1

[12] Zhekun Luo, Devin Guillory, Baifeng Shi, Wei Ke, Fang Wan, Trevor Darrell, and Huijuan Xu. Weakly-supervised action localization with expectation-maximization multi-instance learning. In *European Conference on Computer Vision*, pages 729–745. Springer, 2020. 1

[13] Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han. Weakly supervised action localization by sparse temporal pooling network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6752–6761, 2018. 1

[14] Sujoy Paul, Sourya Roy, and Amit K Roy-Chowdhury. W-talc: Weakly-supervised temporal activity localization and classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 563–579, 2018. 1

[15] Zhiwu Qing, Haisheng Su, Weihao Gan, Dongliang Wang, Wei Wu, Xiang Wang, Yu Qiao, Junjie Yan, Changxin Gao, and Nong Sang. Temporal context aggregation network for temporal action proposal refinement. *arXiv preprint arXiv:2103.13141*, 2021. 1

[16] Lin Song, Shiwei Zhang, Gang Yu, and Hongbin Sun. Tacnet: Transition-aware context network for spatio-temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11987–11995, 2019. 1

[17] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. Video classification with channel-separated convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5552–5561, 2019. 1

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 2

[19] Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. Tdn: Temporal difference networks for efficient action recognition. *arXiv preprint arXiv:2012.10071*, 2020. 1, 2

[20] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2740–2755, 2018. 1

[21] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 3

[22] Xiang Wang, Shiwei Zhang, Zhiwu Qing, Yuanjie Shao, Changxin Gao, and Nong Sang. Self-supervised learning for semi-supervised temporal action proposal. *arXiv preprint arXiv:2104.03214*, 2021. 1

[23] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10156–10165, 2020. 1

[24] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8668–8678, 2019. 1, 3