Research Proposal For Temporal Action Detection

Zhen-Ying Fang[1]
[1] https://zhenyingfang.github.io/

Abstract

Temporal action detection (TAD) is extensively studied in the video understanding community by following the object detection pipelines in images. It aims to determine the semantic label and the boundaries of every action instance in an untrimmed video. Depending on different labels, TAD can be divided into two forms of full supervision and weak supervision. Although previous researchers have done lots of work in TAD, the current mainstream methods still have some problems. For example, low performance, process complex, non-end-to-end, etc. So, it is still an important research topic for the development of more efficient and lightweight action detector.

1. Introduction

With the rapid growth of video media, video understanding has become an important area of computer vision. Temporal action detection (TAD) is particularly important for video understanding, which aims to determine the action label and temporal interval of every action instance in an untrimmed video. For its wide range of applications, including security video highlight generation [9], video surveillance [7], and so on, temporal action detection has gained increasing attention from the community in recent years.

As show in Figure 1. According to different training label, TAD can be devided into two forms of fully supervised and weakly supervised. During model training, fully supervised temporal action detection (FS-TAD) has both action class and temporal interval labels. However, weakly supervised temporal action detecton (WS-TAD) only has action class labels. In the inference phase, FS-TAD need to predict the action class and temporal interval of each action instance and the same can be said about WS-TAD.

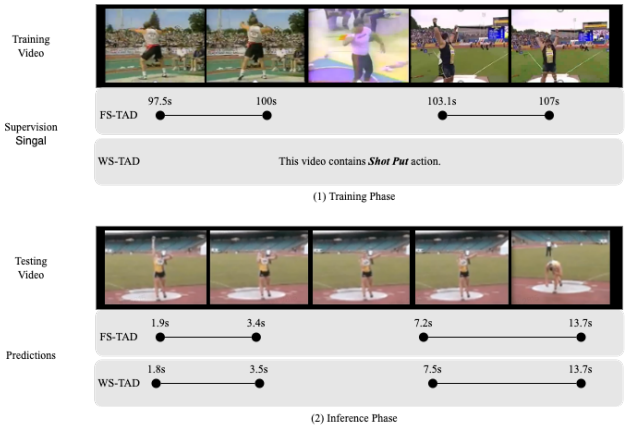

图 1. Compare FS-TAD and WS-TAD. Inspired by Le Yang's talk for BackTAL in VALSE.

In the past several years, TAD has developed rapidly with the attention of more and more researchers. Whereas, existing methods still have some shortcomings. For example, most FW-TAD methods are not end-to-end, the inference process is not clear, and manual annotations of temporal boundaries is laborious. Although WS-TAD can

save annotation time and manpower, its performance lags far behind FS-TAD. Therefore, it is still an important research topic for the development of more efficient and lightweight action detector.

2. Related Work

Action Recognition. Action recognition is a fundamental task in video understanding. With the success of deep learning, a vast array of methods utilize 2D and 3D CNNs [35, 36, 38, 16, 8, 23, 37] to achieve impressive performance on standard action recognition benchmarks [34, 3]. Due to the large scale pretraining on action recognition datasets, the pretrained models from this domain are widely used in TAD as a feature extractor. The parameters of the feature extraction layer are usually frozen. Therefore the weights cannot be adjusted according to the training set of TAD. This limits feature extraction capabilities.

Multi-stage FS-TAD. (a) multi-stage methods [44, 19, 17, 1, 31] first train a binary classifier to generate action proposals. Those proposals are fed to a multi-class classifier to classify the action labels. BMN [17] using Boundary-Matching (BM) mechanism to generate precise action proposals. TAPG [44] using a general auxiliary Background Constraint idea to suppress low-quality proposals. PcmNet [31] propose a temporal-position-sensitive context modeling approach to incorporate both positional and semantic information. RCL [39] learns a fully continuous anchoring representation to replace manually designed anchors.

Two-stage FS-TAD. (b) two-stage methods [42, 4, 32, 5, 43] merges action proposal generation and action classification. These methods need to manually set multiple anchor scales, which restricts the flexibility. ContextLoc [50] enrich both local and global contexts. It enhance the representation of each proposal via the local context, global context and context-aware inter-proposal relations. RefactorNet [41] generate a new feature representation with more salient action information by decoupling action features and the co-occurrence features.

One-stage FS-TAD. (c) one-stage methods [18, 12, 24, 46, 22, 6, 21, 40, 14] detect the boundaries and category of action segments in a single shot without using action proposals. PBRNet [21] build two feature pyramids to perform the anchor-based detection via coarse pyramidal detection and refined pyramidal detection. Then, fined-grained detection module is used to augment the frame-level features and update the confidence scores of action instances. [15] design a anchor-free method and refine boundary use the way of coarse-to-fine. BREM [11] estimate boundary quality and improve contextual information of proposal feature representation, using Boundary Evaluate Module (BEM) and Region Evaluate Module (REM), respectively.

Transformer-based FS-TAD. Recently, transformer have achieved great success in natural language processing and computer vision. Some papers start using transformers in FS-TAD. TadTR [22] and ActionFormer [48] design an end-to-end action detector which predict action through learnable action queries followed by DETR [2]. TALLFormer [6] provide long-term feature for transformer-based TAD model through the benefit of long-term memory mechanism.

Weakly-spuervised Temporal Action Detection. WS-TAD [28, 30, 27, 47, 29, 26, 33] detects action class and temporal interval using only action class labels. STPN [28] generate action interval using Temporal Class Activation Sequence (T-

CAS) followed by Class Activation Mapping (CAM) [49]. CMCS [20] use multi-branch complementary CAS. $CO_2 - Net$ [10] explore FS-TAD task-specific feature by cross-modal consensus module (CCM). Although the above methods achieve good performance, none of them can model background information. [29, 13] improve performance by modeling background information. Recently, some researchers explore extra labeled information. SF-Net [25] using single-frame action label to generate more precise action instance. However, SF-Net can't separate background when two action instances are are close. BackTAL [45] separate closely action instances through single-frame background label. WS-TAD has received more and more attention from researchers in recent years. However, Its performance is still far behind FS-TAD. How to improve the performance of the weakly-supervised detector is still a research focus in the future.

3. Research Objectives

The long term goal of the research is to improve performance. However, the existing models are too complex for practical application. It is still necessary to research more efficient detector. Most of the existing detectors are aimed at coarse granularity of action classes. These models may overfit in high-level context information. On the other hand, with the rapid development of media, there are a large number of weakly tagged videos on the Internet. In the future, how to train better detectors based on weakly supervised labels is a very important. More specifically, the following research questions need to be addressed:

- How to develop a more efficient fully-supervised detector?
- How detect finer motion without being strongly dependent on the background?
- Can a weekly-supervised detector with stronger performance be developed?

In order to solve the above problems, exploring more efficient video feature extraction methods is very important, so that feature extraction and detector head can be efficiently combined to achieve end-to-end training. How to better perform video frame sampling is also a problem that needs to be studied. And, the temporal feature pyramid will have better detection performance for shot-term action instance.

Weakly-supervised Temporal Action Detection (WS-TAD) is also a very important research topic. It can help us make better use of a large number of weakly labeled video resources on the Internet, thereby helping to solve the problem of manual labeling effort. The existing WS-TAD method has a complicated process, It needs to extract video features first, and then perform weakly supervised detection. Therefore, a end-to-end WS-TAD method is necessary to develop. On the other hand, improving the performance of WS-TAD is also an important topic. How to better model the background and how to better separate the action instances with short distances in temporal are the keys to improve the performance of WS-TAD.

4. Conclusion

In this proposal, we first review some fundamental methods for FS-TAD and WS-TAD. And, analyze their advantages and disadvantages. Then, we put forward some problems that need to be solved in the field of TAD. Finally, in response to the above problems, we propose some possible solutions.

参考文献

[1] S. Buch, V. Escorcia, C. Shen, B. Ghanem, and J. Carlos Niebles. Sst: Single-stream temporal action proposals. In Proceedings of the IEEE conference on Com-

puter Vision and Pattern Recognition, pages 2911–2920, 2017. 2

[2] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In European conference on computer vision, pages 213–229. Springer, 2020. 2

[3] J. Carreira and A. 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] Y.-W. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng, and R. Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In proceedings of the IEEE conference on computer vision and pattern recognition, pages 1130–1139, 2018. 2

[5] G. Chen, Y.-D. Zheng, L. Wang, and T. Lu. Dcan: Improving temporal action detection via dual context aggregation. arXiv preprint arXiv:2112.03612, 2021. 2

[6] F. Cheng and G. Bertasius. Tallformer: Temporal action localization with long-memory transformer. arXiv preprint arXiv:2204.01680, 2022. 2

[7] M. Cristani, M. Bicego, and V. Murino. Audio-visual event recognition in surveillance video sequences. IEEE Transactions on Multimedia, 9(2):257–267, 2007. 1

[8] C. Feichtenhofer, H. Fan, J. Malik, and K. He. Slowfast networks for video recognition. In Proceedings of the IEEE/CVF international conference on computer vision, pages 6202–6211, 2019. 2

[9] S. Giancola, M. Amine, T. Dghaily, and B. Ghanem. Soccernet: A scalable dataset for action spotting in soccer videos. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pages 1711–1721, 2018. 1

[10] F.-T. Hong, J.-C. Feng, D. Xu, Y. Shan, and W.-S. Zheng. Cross-modal consensus network for weakly supervised temporal action localization. In Proceedings of the 29th ACM International Conference on Multimedia, pages 1591–1599, 2021. 3

[11] J. Hu, L. Zhuang, B. Wang, T. Ge, Y. Jiang, H. Li, et al. Estimation of reliable proposal quality for temporal action detection. arXiv preprint arXiv:2204.11695, 2022. 2

[12] Y. Huang, Q. Dai, and Y. Lu. Decoupling localization and classification in single shot temporal action detection. In 2019 IEEE International Conference on Multimedia and Expo (ICME), pages 1288–1293. IEEE, 2019. 2

[13] P. Lee, J. Wang, Y. Lu, and H. Byun. Weakly-supervised temporal action localization by uncertainty modeling. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 1854–1862, 2021. 3

[14] X. Li, T. Lin, X. Liu, W. Zuo, C. Li, X. Long, D. He, F. Li, S. Wen, and C. Gan. Deep concept-wise temporal convolutional networks for action localization. In Proceedings of the 28th ACM International Conference on Multimedia, pages 4004–4012, 2020. 2

[15] C. Lin, C. Xu, D. Luo, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, and Y. Fu. Learning salient boundary feature for anchor-free temporal action localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3320–3329, 2021. 2

[16] J. Lin, C. Gan, and S. Han. Tsm: Temporal shift module for efficient video understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 7083–7093, 2019. 2

[17] T. Lin, X. Liu, X. Li, E. Ding, and S. 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. 2

[18] T. Lin, X. Zhao, and Z. Shou. Single shot temporal action detection. In Proceedings of the 25th ACM international conference on Multimedia, pages 988–996, 2017. 2

[19] T. Lin, X. Zhao, H. Su, C. Wang, and M. Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In Proceedings of the European conference on computer vision (ECCV), pages 3–19, 2018. 2

[20] D. Liu, T. Jiang, and Y. Wang. Completeness modeling and context separation for weakly supervised temporal action localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1298–1307, 2019. 3

[21] Q. Liu and Z. Wang. Progressive boundary refinement network for temporal action detection. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 11612–11619, 2020. 2

[22] X. Liu, Q. Wang, Y. Hu, X. Tang, S. Bai, and X. Bai. End-to-end temporal action detection with transformer. arXiv preprint arXiv:2106.10271, 2021. 2

[23] Z. Liu, L. Wang, W. Wu, C. Qian, and T. Lu. Tam: Temporal adaptive module for video recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 13708–13718, 2021. 2

[24] F. Long, T. Yao, Z. Qiu, X. Tian, J. Luo, and T. Mei. Gaussian temporal awareness networks for action localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 344–353, 2019. 2

[25] F. Ma, L. Zhu, Y. Yang, S. Zha, G. Kundu, M. Feiszli, and Z. Shou. Sf-net: Single-frame supervision for temporal action localization. In European conference on computer vision, pages 420–437. Springer, 2020. 3

[26] M. Moniruzzaman, Z. Yin, Z. He, R. Qin, and M. C. Leu. Action completeness modeling with background aware networks for weakly-supervised temporal action localization. In Proceedings of the 28th ACM International Conference on Multimedia, pages 2166–2174, 2020. 2

[27] S. Narayan, H. Cholakkal, F. S. Khan, and L. Shao. 3c-net: Category count and center loss for weakly-supervised action localization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 8679–8687, 2019. 2

[28] P. Nguyen, T. Liu, G. Prasad, and B. 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. 2

[29] P. X. Nguyen, D. Ramanan, and C. C. Fowlkes. Weakly-supervised action localization with background modeling. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 5502–5511, 2019. 2, 3

[30] S. Paul, S. Roy, and A. 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. 2

[31] X. Qin, H. Zhao, G. Lin, H. Zeng, S. Xu, and X. Li. Pcmnet: Position-sensitive context modeling network for temporal action localization. arXiv preprint arXiv:2103.05270, 2021. 2

[32] M. Santhi and L. E. Sunny. Contextual multi-scale region convolutional 3d network for anomalous activity detection in videos. In International Conference On Computational Vision and Bio Inspired Computing, pages 98–108. Springer, 2019. 2

[33] B. Shi, Q. Dai, Y. Mu, and J. Wang. Weakly-supervised action localization by generative attention modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1009–1019, 2020. 2

[34] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402, 2012. 2

[35] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE international conference on computer vision, pages 4489–4497, 2015. 2

[36] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 6450–6459, 2018. 2

[37] L. Wang, Z. Tong, B. Ji, and G. Wu. Tdn: Temporal difference networks for efficient action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1895–1904, 2021. 2

[38] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks for action recognition in videos. IEEE transactions on pattern analysis and machine intelligence, 41(11):2740–2755, 2018. 2

[39] Q. Wang, Y. Zhang, Y. Zheng, and P. Pan. Rcl: Recurrent continuous localization for temporal action detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13566–13575, 2022. 2

[40] X. Wang, C. Gao, S. Zhang, and N. Sang. Multi-level temporal pyramid network for action detection. In Chinese Conference on Pattern Recognition and Computer Vision (PRCV), pages 41–54. Springer, 2020. 2

[41] K. Xia, L. Wang, S. Zhou, N. Zheng, and W. Tang. Learning to refactor action and co-occurrence features for temporal action localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13884–13893, 2022. 2

[42] H. Xu, A. Das, and K. Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In Proceedings of the IEEE international conference on computer vision, pages 5783–5792, 2017. 2

[43] M. Xu, C. Zhao, D. S. Rojas, A. Thabet, and B. 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. 2

[44] H. Yang, W. Wu, L. Wang, S. Jin, B. Xia, H. Yao, and H. Huang. Temporal action proposal generation with background constraint. arXiv preprint arXiv:2112.07984, 2021. 2

[45] L. Yang, J. Han, T. Zhao, T. Lin, D. Zhang, and J. Chen. Background-click supervision for temporal action localization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021. 3

[46] L. Yang, H. Peng, D. Zhang, J. Fu, and J. Han. Revisiting anchor mechanisms for temporal action localization. IEEE Transactions on Image Processing, 29:8535–8548, 2020. 2

[47] T. Yu, Z. Ren, Y. Li, E. Yan, N. Xu, and J. Yuan. Temporal structure mining for weakly supervised action detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 5522–5531, 2019. 2

[48] C. Zhang, J. Wu, and Y. Li. Actionformer: Localizing moments of actions with transformers. arXiv preprint arXiv:2202.07925, 2022. 2

[49] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2921–2929, 2016. 3

[50] Z. Zhu, W. Tang, L. Wang, N. Zheng, and G. Hua. Enriching local and global contexts for temporal action localization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 13516–13525, 2021. 2