Multiple Attempts for AVA-Kinetics challenge 2020

Jin Xia[1]*, Wei Li[1]*, Jie Shao[1], Zehuan Yuan[1], Jiajun Tang[2], Cewu Lu[2], Changhu Wang[1]
ByteDance AI Lab[1], Shanghai Jiao Tong University[2]

Abstract

In this work, we present our solution for the AVA-Kinetics challenge in Intertional Challenge on Activity Recognition at CVPR 2020. We attempt multiple solutions for this task. The results from multiple models are merged for the final submission. Our final submission achieves 32.91 mAP on AVA-Kinetics challenge, which is the second place solution. We got the good result without using some common training techniques. For example, we didn't use the whole AVA-Kinetics dataset to train the model and we didn't use a very strong model for Kinetics. These show the efficiency of the models used in this work and indicate easy future works to enhance the performance.

1. Introduction

Video understanding has made considerable progress in recent years thanks to the evolution of deep learning and available large-scale datasets. The task of action detection is one of the most important and difficult tasks among multiple techniques. It aims to detect and recognize actions in space and time.

Previous works attack the action detection problem using different methods. Various 3D CNN [8, 11, 2, 3] extend 2D convolution to 3D convolution. Two stream networks [12, 4] leverage different kinds of visual information such as RGB stream and optical flow. Other than that, recently proposed methods such as Nonlocal network [14] play on high-level feature map extracted from backbone to enhance the performance.

Large-scale datasets push forward the progress of action detection. AVA [6] is a large video dataset of spatio-temporally localized atomic visual actions, containing 430 15-minutes video clips. Spatio-temporal labels are provided for one frame per second, with every person annotated with a bounding box and multiple actions, resulting in 1.58M action labels in total. the Kinetics 700 [1] is the largest video classification dataset which contains 700 action classes. In 2020, the AVA-Kinetics localized human actions video dataset [9] is proposed. The dataset is collected by annotating videos from the Kinetics-700 dataset using the AVA annotation protocol, and extending the original AVA dataset with these new AVA annotated Kinetics clips.

In this challenge, we attempt different methods for the action detection task. For the origin AVA dataset, we adopt Asynchronous Interaction Aggregation (AIA) network and LGN network. For the Kinetics part of the dataset, we apply simply a SlowFast Network [3]. The network used for Kinetics dataset is relatively simple, a stronger model such as AIA can further the performance. What's more, we trained models on AVA dataset and Kinetics dataset separately. However, transfering the knowledge from one dataset to the other can enhance the performance for both datasets. We leave them as future work.

2. Methods

Backbone. We select SlowFast network with ResNet-101 structure [7] as our backbone model. The backbone is pretrained on Kinetics-700 video classification dataset.

AIA Network. Asynchronous Interaction Aggregation network [13] is the recently proposed network for action detection. It achieves the state of the art performance on the AVA dataset. We select it as our solution for this challenge due to its high performance. On top of the backbone model, AIA has an Interaction Aggregation structure to model multiple types of interactions. The interactions are aggregated in a deep structure to accurately catch the attention between persons and the context. What's more, AIA proposes the Asynchronous Memory Update algorithm that enables us model long-term interaction dynamically without huge computation cost. For AIA dataset, we only train the network on origin AVA dataset. Adding training samples from Kinetics dataset could massively improve the performance according to [9].

LGN. Long-term Gating Network focuses on the temporal modelling of feature banks and utilizes gating mechanism to emphasize relevant features. Same as AIA, we only train the network on origin AVA dataset.

[*] Both authors contributed equally to this work.

Entry	AVA Kinetics	Kinetics	AVA
Official Baseline	22.70	19.74	21.23
YH Technologies[17]	-	-	19,69
Tsinghua University	-	-	21.03
LFB [15]	-	-	27.20
Action Transformer [5]	-	-	24.93
ByteDance [10]	-	-	30.20
Three Branch[16]	-	-	32.49
SlowFast [3] (7ens)	-	-	34.25
Ours	**32.91**	**25.57**	**35.50**

Table 1: **Test Results on AVA Kinetics**

Model for Kinetics. Our solution for Kinetics part is simple. We finetune the pretrained backbone on the Kinetics part of AVA-Kinetics. The ground true human box is used during training. During inference, a Faster-RCNN model is used for human box detection. These human boxes are used to extract human features from video feature. These features are fed to the action head for the final classification. The model for Kinetics is simple, thus more progresses could be made for Kinetics. First, AIA network could be applied to Kinetics for a huge performance gain. Second, the knowledge learnt from AVA dataset could be transfered to Kinetics. This knowledge could be especially important for the difficult classes. Third, the ground truth human box of Kinetics could also be used to train a stronger human detector.

3. Conclusion and Future Work

In this report, we present our solution for the AVA-Kinetics challenge at CVPR 2020. The test result is shown in Table. 1. Our solution is simple and there are several evident improvements that could further largely improve the performance. First, we could use strong model such as AIA and LGN for the Kinetics part of AVA-Kinetics. Second, we could jointly train model using both AVA and Kinetics dataset to have more training data. Third, the ground truth human boxes from both AVA and Kinetics could be used to train a better human detector.

References

[1] 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. 1

[2] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. *CoRR*, abs/1705.07750, 2017. 1

[3] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. *CoRR*, abs/1812.03982, 2018. 1, 2

[4] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. *CoRR*, abs/1604.06573, 2016. 1

[5] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 244–253, 2019. 2

[6] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018. 1

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. 1

[8] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, Jan 2013. 1

[9] Ang Li, Meghana Thotakuri, David A Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. The ava-kinetics localized human actions video dataset. *arXiv preprint arXiv:2005.00214*, 2020. 1

[10] Wei Li, Zehuan Yuan, An Zhao, Jie Shao, and Changhu Wang. Bytedance ai lab ava challenge 2019 technical report. 2

[11] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. *CoRR*, abs/1711.10305, 2017. 1

[12] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *CoRR*, abs/1406.2199, 2014. 1

[13] Jiajun Tang, Jin Xia, Xinzhi Mu, Bo Pang, and Cewu Lu. Asynchronous interaction aggregation for action detection. *arXiv preprint arXiv:2004.07485*, 2020. 1

[14] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. *CoRR*, abs/1711.07971, 2017. 1

[15] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krähenbühl, and Ross B. Girshick. Long-term feature banks for detailed video understanding. *CoRR*, abs/1812.05038, 2018. 2

[16] Jin Xia, Jiajun Tang, and Cewu Lu. Three branches: Detecting actions with richer features, 2019. 2

[17] Ting Yao and Xue Li. YH technologies at activitynet challenge 2018. *CoRR*, abs/1807.00686, 2018. 2