Learning Unbiased Transformer for Long-Tail Sports Action Classification

Yijun Qian, Lijun Yu, Wenhe Liu, Alexander G. Hauptmann
Language Technologies Institute, Carnegie Mellon University
{yijunqian,lijun}@cmu.edu,{wenhel,alex}@cs.cmu.edu

ABSTRACT

The Sports Video Task in MediaEval 2021 Challenge contains two subtasks, detection and classification. The classification subtask aims to classify different strokes in table tennis segments. These strokes are fine grained actions and difficult to distinguish. To solve this challenge, we, the INF Team, proposed a fine grained action classification pipeline with SWIN-Transformer and a combination of optimization techniques. According to the evaluation results, our best submission ranks first with 74.21% accuracy and significantly outperforms the runner-up (74.21% v.s. 68.78%).

1 INTRODUCTION

Action classification has been a heated topic in computer vision and can be widely implemented in real-world applications. Recent years have witnessed many successful works on action classification[6, 9, 12]. The recent improvements of these methods can be highly attributed to the advancement of temporal modeling capacity. Different from previous series of 2D-Stream CNN works or 3D-CNN methods, [12] factorizes the 3D spatial-temporal convolution to a 2D spatial convolution and a 1D temporal convolution. TRM [9] directly replaces convolution operation with temporal relocation operation to enable the 2D CNNs the capability of spatial-temporal modeling with an equivalent temporal receptive field of the whole input video clip. Given the recent success of implementing transformer [13] based methods in image-level computer vision tasks (*i.e. ViT [3] for image classification*), Video SWIN-Transformer (VST) [6] proposed a transformer based video feature extractor model and surpassed previous CNN based SOTAs with noticeable margins on multiple action recognition benchmarks. However, directly implementing the VST model on the dataset of sports video classification task in the 2021 Mediaeval Challenge won't be the optimal solution. Different from the other action classification benchmarks [4, 7, 11], the Sports Video Classification Task [7] of 2021 Mediaeval Challenge specifically focused on strokes within table tennis segments. These strokes are fine-grained actions that are visually similar and take place in limited scenes. Meanwhile, the samples for training are pretty limited, and the dataset is severely long-tail distributed. Without specially-designed techniques, the model will be easily overfitted and biased to strokes of head classes. To solve this, we implemented Background Erasing [14] which prevents the model from overfitting to background regions. We also proposed a sample-balanced cross entropy loss for model optimization on the long-tail distributed dataset.

MediaEval'21, December 13-15 2021, Online

2 APPROACH

2.1 Implementation of VST Model

Unless otherwise mentioned, all our reported results use VST-B [6] as the backbone extractor. Specifically, the channel number of the hidden layers in the first stage is 128. The window size is set to $P = 8$ and $M = 7$. The query dimension of each head is $d = 32$, and the expansion layer of each MLP is set to $\alpha = 4$. The layer numbers of the four stages are $\{2, 2, 18, 2\}$. The model is initialized with weights pretrained on Kinetics600 [1]. We employ an SGD optimizer with plateau scheduler and train the model for 30 epochs. We use rank1 accuracy as the monitor metric of plat scheduler, and the patience is set as 1. During training stage, the input frames are firstly resized to 256×256, then randomly cropped to 224×224 for data augmentation. In evaluation stage, the input frames are firstly resized to 256×256, then center cropped to 224×224. For each segment, 32 frames are evenly sampled as the input instance. Therefore, for each segment, the size of input sample V_{be} is $32 \times 224 \times 224$.

2.2 Implementation of Background Erasing

After analyzing the training set videos, we find the scenes are quite similar, e.g., many videos are recorded in the same scene. As a result, the model may easily become background biased as reported in [5, 16–18] and experiments in [2]. To solve this issue, we followed [14] to apply a background erasing algorithm in training. To be specific, one static frame is randomly sampled from each input segment and added to every other frames within the segment to construct a distracting sample. Then, an MSE loss is implemented to force the features extracted from the original clip to be similar to those extracted from the distracting sample.

$$\mathcal{L}_{mse} = \|\mathcal{N}(V_{org}) - \mathcal{N}(V_{be})\|^2 \tag{1}$$

where $\mathcal{N}$ represents the backbone VST extractor, V_{org} represents the original input clip, and V_{be} represents the background erased clip.

2.3 Implementation of Balanced Loss

As is shown in Figure 1, the training dataset is severely long-tail distributed. If all samples are evenly weighted, the model may easily become biased to the head classes (i.e. the classes with much more samples than others in the training set). Thus, we use a class-wise weight $W_s = \{w_s^1, w_a^2, ..., w_s^n\}$ to balance samples of different strokes.

$$\hat{w}_s^i = \frac{1}{N^i} \tag{2}$$

$$w_s^i = n \times \frac{\hat{w}_s^i}{\sum_i \hat{w}_s^i} \tag{3}$$

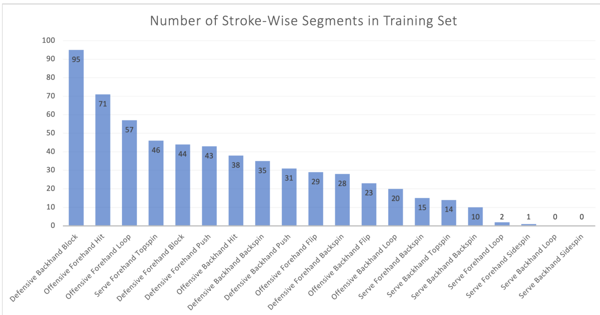

Figure 1: The number of segments for training varies among different strokes. Especially, there are no samples of Serve Backhand Loop and Serve Backhand Sidespin for training.

Table 1: Results of CMU INF Team in Sports Classification Task of 2021 Mediaeval Challenge

Run ID	System Spec	Val Acc %	Test Acc %
Run1	swin-transformer	63.40	63.35
Run3	Run1 + balanced loss	67.81	66.06
Run2	Run3 + background erasing	**75.25**	**74.21**

where N^i represents the i^{th} stroke's number of samples for training, and n represents the number of strokes (20 here). The overall loss function for optimization becomes:

$$\mathcal{L}_{xe}^i = -w_s^i \log\left(\frac{\exp(\phi(\mathcal{N}(x_n^i)))}{\sum_j \exp(\phi(\mathcal{N}(x_n^j)))}\right) \tag{4}$$

$$\mathcal{L}_{xe} = \sum_i \frac{\mathcal{L}_{xe}^i}{n} \tag{5}$$

$$\mathcal{L} = \alpha \mathcal{L}_{mse} + \beta \mathcal{L}_{xe} \tag{6}$$

where ϕ represents the MLP classifier with dropout layers that projects extracted video feature to vector of probabilities. Unless specially mentioned, we set $\alpha = 1$ and $\beta = 1$ for all our results in this report.

3 RESULTS AND ANALYSIS

As is shown in Table 1, we report the performance of our three submissions on both self-evaluated validation set and official hidden test set. Through comparing Run1 and Run3, we can find that the implementation of balanced loss brings 3.41% improvements on validation set and 2.71% improvements on test set. It shows that balanced sampling can improve the final performance through forcing the model pay more attention on tail classes and less attention on head classes. It may also work for similar tasks[8, 10, 15**?**]. Through comparing Run2 and Run3, we can find that the usage of background erasing significantly improves the performance on both validation set (7.44%) and test set (8.15%).

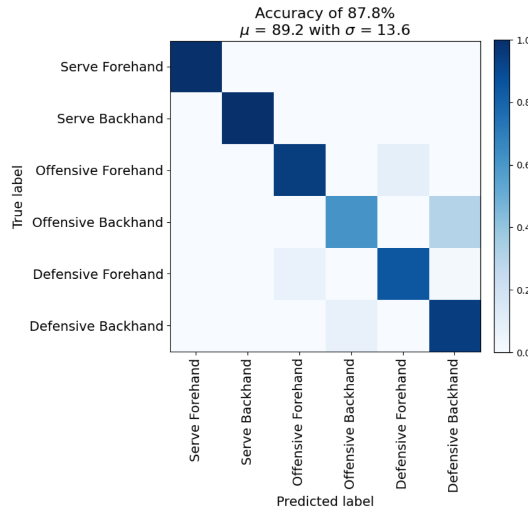

Figure 2: Confusion matrix among sub-group attributes of Run2 submission.

4 DISCUSSION AND OUTLOOK

The strokes in the sports classification task have several sub-group attribute pairs (*i.e., Defensive v.s. Offensive and Forehand v.s. Backhand*). So besides comparing the global accuracy performance, we also analyze the confusion matrix of these sub-group attributes. As is shown in Figure 2, we can find our system can successfully distinguish similar attribute pairs such as forehand v.s. backhand, server v.s. offensive, and server v.s. defensive. However, it doesn't perform as well when encountering offensive v.s. defensive. We suggest the 0-1 classification of sub-group attributes can be included in next year's challenge as extra metric. Meanwhile, we find several strokes (*i.e. Serve Backhand Loop and Serve Backhand Sidespin*) never appear in training or validation sets. Although the balanced loss can relieve the classifier bias to head classes to some extent, the number of samples for several strokes (*i.e. Serve Forehand Loop*) is still too small to train a robust model. Thus, we hope the dataset can be re-split or augmented for next year's challenge. Finally, we didn't use both train and val samples for final submission, we will have a try next year to see if the performance get improved. Meanwhile, we also assume initializing with weights pretrained on large fine-grained action recognition datasets may also improvements.

ACKNOWLEDGMENTS

This research is supported in part by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DOI/IBC) contract number D17PC00340. This research is supported in part through the financial assistance award 60NANB17D156 from U.S. Department of Commerce, National Institute of Standards and Technology. This project is funded in part by Carnegie Mellon University's Mobility21 National University Transportation Center, which is sponsored by the US Department of Transportation.

REFERENCES

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

[2] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. 2019. Why Can't I Dance in the Mall? Learning to Mitigate Scene Bias in Action Recognition. *arXiv preprint arXiv:1912.05534* (2019).

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR* (2021).

[4] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. 2011. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*.

[5] Wenhe Liu, Guoliang Kang, Po-Yao Huang, Xiaojun Chang, Yijun Qian, Junwei Liang, Liangke Gui, Jing Wen, and Peng Chen. 2020. Argus: Efficient activity detection system for extended video analysis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*. 126–133.

[6] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. 2021. Video swin transformer. *arXiv preprint arXiv:2106.13230* (2021).

[7] Pierre-Etienne Martin, Jordan Calandre, Boris Mansencal, Jenny Benois-Pineau, Renaud Péteri, Laurent Mascarilla, and Julien Morlier. 2021. Sports Video: Fine-Grained Action Detection and Classification of Table Tennis Strokes from videos for MediaEval 2021. (2021).

[8] Yijun Qian, Lijun Yu, Wenhe Liu, and Alexander G Hauptmann. 2020. Electricity: An efficient multi-camera vehicle tracking system for intelligent city. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 588–589.

[9] Yijun Qian, Lijun Yu, Wenhe Liu, and Alexander G. Hauptmann. 2022. TRM: Temporal Relocation Module for Video Recognition. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision Workshops*.

[10] Yijun Qian, Lijun Yu, Wenhe Liu, Guoliang Kang, and Alexander G Hauptmann. 2020. Adaptive feature aggregation for video object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*. 143–147.

[11] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).

[12] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 6450–6459.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[14] Jinpeng Wang, Yuting Gao, Ke Li, Yiqi Lin, Andy J Ma, Hao Cheng, Pai Peng, Feiyue Huang, Rongrong Ji, and Xing Sun. 2021. Removing the Background by Adding the Background: Towards Background Robust Self-supervised Video Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11804–11813.

[15] Lijun Yu, Qianyu Feng, Yijun Qian, Wenhe Liu, and Alexander G. Hauptmann. 2020. Zero-VIRUS: Zero-Shot VehIcle Route Understanding System for Intelligent Transportation. 594–595.

[16] Lijun Yu, Yijun Qian, Wenhe Liu, and Alexander G Hauptmann. CMU Informedia at TRECVID 2020: Activity Detection with Dense Spatio-temporal Proposals. In *TRECVID 2020*.

[17] Lijun Yu, Yijun Qian, Wenhe Liu, and Alexander G Hauptmann. CMU Informedia at TRECVID 2021: Activity Detection with Argus++. In *TRECVID 2021*.

[18] Lijun Yu, Yijun Qian, Wenhe Liu, and Alexander G. Hauptmann. 2022. Argus++: Robust Real-time Activity Detection for Unconstrained Video Streams with Overlapping Cube Proposals. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision Workshops*.