HAVPR Submission to Robustness in Sequential Data Challenge 2022

Zitao Gao
Wuhan University
Wuhan
gaozitao@whu.edu.cn

Yuwei Yin
Wuhan University
Wuhan
lichang0115yyw@whu.edu.cn

Yuanzhong Liu
Wuhan University
Wuhan
yzliu.me@whu.edu.cn

Zhigang Tu
Wuhan University
Wuhan
tuzhigang@whu.edu.cnn

Juedfeng Xiao
Wuhan University
Wuhan
1023897360@qq.com

Xiangyue Zhang
Central South University
Wuhan
812343664@qq.com

Abstract

This report describes the details of our solution to robustness in sequential data challenge 2022 which is focused on developing solutions that reduce the gap in performance between training set and real-world testing scenario. To handle the various types of perturbations and corruptions observed in real-world data, we try some classical algorithms. Single-frame wavelet denoising are adopted before feature extraction in the inference phase, which leads to a comparable result. Finally, we achieved accuracy of 76.24 on Kinetics-400P, 85.68 on UCF-101P and 66.08 on HMDB-51P.

1. Introduction

Improving the robustness of the model in real-world testing scenario is an important topic in the research of action recognition. The existing approaches addressing this issue perform their experiments on artificially created datasets with perturbed and corrupted samples. On this basis, we further process the datasets to improve the recognition effect of the models.

In this paper, we will share our solution to Robustness in Sequential Data Challenge. The key points of our solution can be summarized as follows:

1. Extensive experiments of different algorithms.
2. Suitable augmentation in the testing phase.
3. The usage of ensemble methods.

2. Our solution

2.1. Data Analysis

We need to evaluate our approach on three datasets, of which HMDB51 and UCF101 are relatively small datasets with 51 and 101 categories, respectively, and a small number of videos. kinetics has a large amount of data and more categories, with 400 action categories. The resolution of the videos in these datasets also varies.

Our task is that the model can make a correct category judgment for the videos with natural perturbations. We randomly selected some videos from the test set and counted the types of perturbations applied in the test set, which were divided into 11 kinds of perturbations. For example, like salt and pepper noise, blur, jitter and so on. As shown in Figure 1, We found that the most proportion of them are some noise additions, frame repetition and frame missing and other timing interference account for less, after comprehensive consideration we decided to perform single-frame wavelet denoising before feature extraction.

2.2. Models

For the action recognition algorithm, we try Video Swin Transformer [4, 5], TSM Non-Local [3, 7], SlowFast [2], TANet [6], TimeSformer [1]. TSM Non-Local and TANet are action recognition models based on 2D convolution, which can exchange information between frames and extract action features with fewer parameters. SlowFast is based on 3D convolution, which use the fast and slow pathway to extract temporal and spatial feature. TimeSformer and Video Swin Transformer apply transformer to action recognition task, leading to a better speed-accuracy trade-off compared to previous approaches which compute self-attention globally even with spatial-temporal factorization. Among all the models mentioned above, Video Swin Transformer gave us a surprise.

2.3. Augmentation

In the training phase, we use RandomResizedCrop to resize the image to $224 * 224$, and the random erasing is also

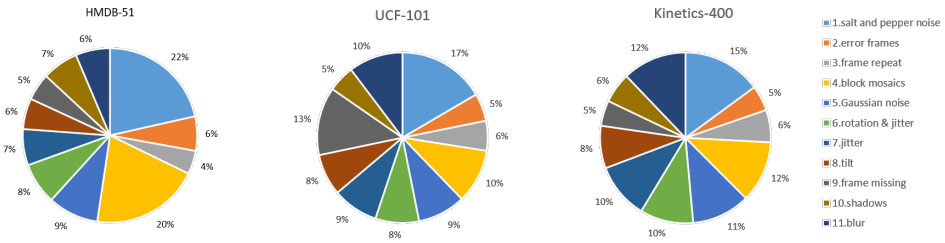

Figure 1. Perturbations in each dataset

Dataset	Accuracy
HMDB-51P(mini)	58.43
UCF-101P(mini)	86.41
Kinetics-400P(mini)	76.56
HMDB-51P(full)	66.08
UCF-101P(full)	85.68
Kinetics-400P(full)	76.24

Table 1. Results on the datasets.

adopted.

In the testing phase, we perform single-frame wavelet denoising according to Figure 1 before feature extraction, making the prediction result more robust.

2.4. Ensemble

For boosting the model performance, ensemble is necessary. We use all the models mentioned above to test the datasets respectively, average and sort the scores of each action category, and take the category with the highest score as the output category.

3. Experiments

Our experimental setup for the training phase mainly based on Video Swin Transformer [4, 5]. We choose AdamW as optimizer and warmup is applied. The weight decay equals 0.05, learning rate equals 10^{-3}, and batch size is 64. Moreover, we also applied stochastic depth and set the ratio 0.4.

The experimental results are shown in Table 1. Our solution achieves the best on the Kinetics-400 mini test set and the second on the full set.

4. Conclusion

Our experimental procedure and results show that model integration is a necessary means for accurate results in action recognition tasks. In addition, a suitable pre-processing process somewhat helps to enhance the input data during inference.

References

[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding?, 2021. 1

[2] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 6202–6211, 2019. 1

[3] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 1

[4] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021. 1, 2

[5] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *arXiv preprint arXiv:2106.13230*, 2021. 1, 2

[6] Zhaoyang Liu, Limin Wang, Wayne Wu, Chen Qian, and Tong Lu. Tam: Temporal adaptive module for video recognition. *arXiv preprint arXiv:2005.06803*, 2020. 1

[7] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. *CVPR*, 2018. 1