Data-Efficient Action Recognition via Temporal Pyramid Network and Spatio-Temporal Consistency Inference

Jie Wu* Yuxi Ren* Xuefeng Xiao
ByteDance Inc.
{wujie.10, renyuxi.20190622, xiaoxuefeng.ailab}@bytedance.com

Abstract

Data-efficient Action Recognition is an exceedingly favorable paradigm since few data and annotations are more readily available on the internet. In this paper, we employ the Slowonly network to capture both fast-tempo and slow-tempo via encoding the features at different depths. Then we build a temporal pyramid network (TPN) to aggregates the information of various visual tempos at the feature level. And TPN can be regarded as the complementary comment to capture multi-granularity and task-oriented cues in the data-efficient setting. In the test procedure, we formulate the inference process as a Spatio-temporal consistency prediction. Experimental results on 2^{nd} *"Visual Inductive Priors for Data-Efficient Computer Vision" demonstrate that our method can achieve competitive results.*

1. Introduction

Action Recognition [6] is a fundamental computer vision task and plays a critical role in video structure analysis and potential down-stream applications. Although it has attracted intense attention in recent years, it remains a very challenging problem due to video fuzziness and instability, and complex temporal relationships within the videos. Albeit with varying degrees of progress, most of its recent successes are involved in a massive data setting, i.e., each category of video data is sufficient to model video features. It is still arduous to acquire such numerous data that require a huge amount of manual effort. To alleviate such expensive and unwieldy annotations, "Visual Inductive Priors for Data-Efficient Computer Vision" challenge proposes to address this task in the data-efficient setting that learns to perform action recognition with few data. This is an exceedingly favorable scheme since few data and annotations are more readily available on the internet. In our work, we focus on this data-efficient paradigm.

In the data-efficient setting, how to adopt a small amount of data to model action abstractions determines the upper limit of the performance. For action recognition, it is essential to capture multi-stage action tempo to model more fine-grained action concepts and make accurate and robust predictions. Visual tempo characterizes the dynamics and the temporal scale of action, which helps to capture multi-granularity and task-oriented cues in the data-efficient setting. In this paper, we employ a stronger backbone to capture both fast-tempo and slow-tempo via encoding the features at different depths. Then we build a temporal pyramid network (TPN) to aggregates the information of various visual tempos at the feature level. By leveraging the feature hierarchy formed inside the network, the proposed TPN is able to work with input frames fed at a single rate. In this paper, we employ the Slowonly network to replace the original resnet network in TPN for feature extraction. Slowonly network contributes to capturing semantic information that can be given by images or a few sparse frames, and it operates at low frame rates and slow refreshing speed. And TPN can be regarded as the complementary comment to encode multi-granularity motion cues, by operating via three modules,i.e., spatial semantic modulation, temporal rate modulation and parallel information flow.

In the test procedure, we formulate the inference process as a Spatio-temporal consistency prediction. Specifically, we employ multiple crops technique to obtain diverse region clips for inference in the spatial dimension. In the temporal dimension, we obtain 24 video clips to predict the action category. The spatial and temporal predictions are combined to make the final consistency prediction.

Experimental results on 2nd "Visual Inductive Priors for Data-Efficient Computer Vision" demonstrate that our method can achieve competitive results.

2. Methodology

In this section, we first introduce the fundamental network we choose to encode the semantical information and the visual tempo of actions in section 2.1 and section 2.2. Then, we illustrate the proposed spatio-temporal consis-

---

[1] These authors contributed equally to this work.

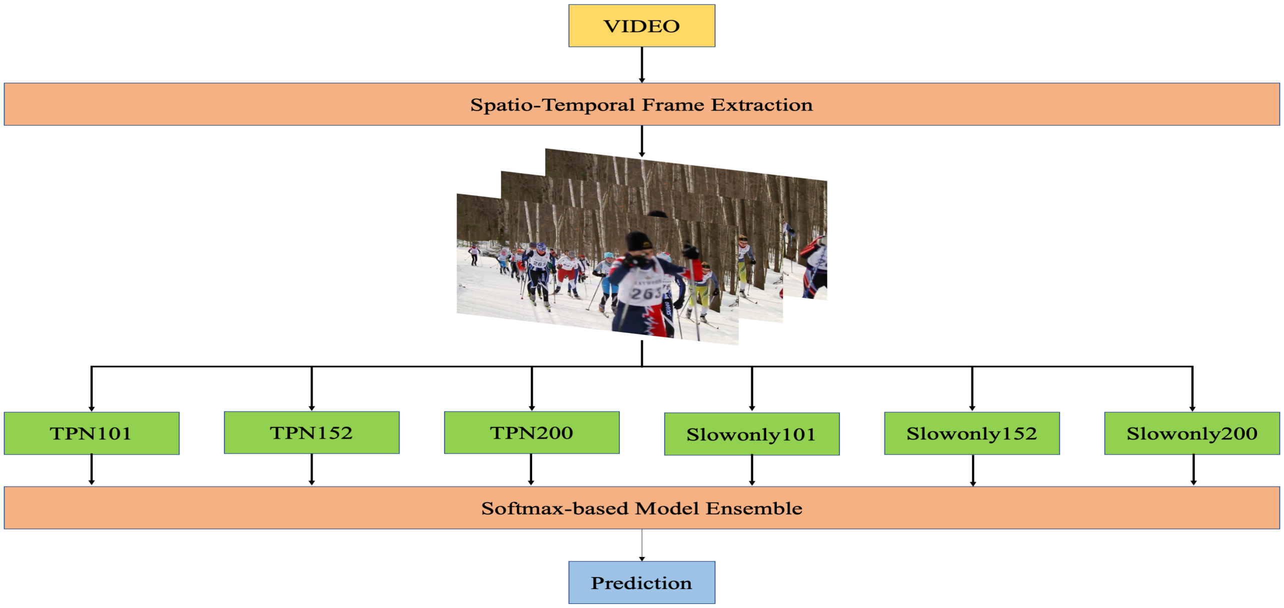

Figure 1: The pipeline of the proposed framework.

tency prediction procedure in section 2.3. Moreover, we design an ensemble strategy to capture the complementation from different models to make robust predictions in section 2.4.

2.1. Temporal Pyramid Network

The two most essential components in action recognition are the spatial semantics and the visual tempo of videos, and the action tempo is especially vital to deal with this particular task. Previous works [5, 1, 8] proposed to capture the visual tempo at the input level that requires multiple network backbones for different frame pyramid levels, which causes time-consuming training and testing procedures.

Temporal pyramid network (TPN) [7] encodes the visual tempo at the feature-level. Namely, since the visual tempo is covered by features at multiple depths, TPN leverages this feature-level visual tempo in a single network. TPN consists of an encoding backbone to extract features from different levels of features and a spatial modulation to align spatial semantics. To realize the flexibility of various sample rates similar to the input-level networks in [5, 1, 8], TPN utilizes a temporal modulation to downsample features according to a given set of hyper-parameters $\{\alpha_i\}_{i=1}^{M}$, which denotes different sample rates. Then, a parallel information flow module is designed to aggregate features in different visual tempos for the final prediction. To sum up, TPN contains three modules to extract semantic information, encoded action temporal clues and aggregate features sequentially.

2.2. Slowonly Network

TPN operates the feature-level visual tempo and can be complemented by an input-level oriented network. And an auxiliary model is able to provide another view of the input action and helps to increase the recognition accuracy. Therefore, we utilize a slowonly network as a supplement, which includes a slow path processing low frame rate samples and a fast path processing high frame rate to formulate a multi-granularity visual tempo. However, we found that under the data-efficient setting, the fast pathway brings minor performance improvement but much slower training speed, so we predigest the network by truncation the fast path to formulate a slowonly network.

2.3. Spatio-Temporal Inference

Maintaining the consistency of prediction for the disturbance both spatially and temporally is an essential key to boost the performance of models in action recognition. For the perspective of the spatial dimension, we incorporate the three-crop and the ten-crop method. Specifically, in the three-crop setting, three random patches are cropped from the original frame, and it is used as the approximation of spatially fully convolutional testing as in [2, 4, 1]. Under the ten-crop setting, 5 patches are extracted then flipped following the procedure in [3]. For the perspective of the temporal dimension, we randomly take 10 or 24 clips from a video to obtain temporal predictions. For the final prediction, we develop an ensemble strategy to combine all of the proposed inferencing granularities as is illustrated in section 2.4.

2.4. Model Ensemble

We conduct a predictions ensemble to generate the final results, the components of our fusion include two complementary networks as are mentioned in section 2.1 and section 2.2. For each network, in addition to the spatially and temporally consistency inference, we found that composing models from different epochs during a training process also helps to boost the performance. Denote the output vector under a specific test setting as $a_{i,j}$ for TPN or $b_{i,j}$ for slowonly network from the j_{th} epoch. And the total sampled epoch number is M_a and M_b, the test setting number is N_a and N_b respectively, the prediction of our model ensemble strategy is:

$$pred = \frac{\sum_{i=1}^{N_a}\sum_{j=1}^{M_a} s(a_{i,j}) + \sum_{i=1}^{N_b}\sum_{j=1}^{M_b} s(b_{i,j})}{M_a \times N_a + M_b \times N_b}, \quad (1)$$

where $s()$ denotes the softmax operation.

3. Experiments

3.1. Experimental Settings

Models. We use TPN and slowonly network as the fundamental models in our experiments. Noting that larger backbones perform better under the data-efficient setting, we exploited to use 3DResNet101, 3DResNet152, and 3DResNet200 as the backbones of the networks respectively. Therefore, we can denote the models utilized during our experiments as TPN101, TPN152, TPN200 and Slowonly101, Slowonly152, Slowonly200.

Datasets. The committee of 2^{nd} "Visual Inductive Priors for Data-Efficient Computer Vision" provided the Kinetics400ViPriors, a modification of the official Kinetics400 dataset. The Kinetics400ViPriors contains 400 types of actions, which is split to 40k, 10k and 20k for training, validation and testing respectively.

Implementation datails. We conduct all of the experiments on Kinetics400ViPriors, in detail, we train our model on a mixture of training and validation set without any extra datasets or pre-trained models. And data augmentation operations include random crop, horizontal flip, cutmix and mixup are employed during our training process.

3.2. Experimental Results

3.2.1 Comparison with different settings

In this section, we investigate the model performance under different settings, include cutmix/mixup during the training process, diverse spatio-temporal inference during the testing process, and the depths of backbones in the networks.

Table 1: Results of different train augmentations.

Network	Train augmentation	Acc(%)
TPN101	Horizontal Flip	59.0
	Horizontal Flip+Cutmix	59.4
	Horizontal Flip+Mixup	59.7
Slowonly101	Horizontal Flip	59.9
	Horizontal Flip+Cutmix	60.5
	Horizontal Flip+Mixup	60.3

Table 2: Results of different test augmentations.

Network	Test augmentation	Clips	Acc(%)
Slowonly101	Center Crop	10	58.6
	Random Crop	10	58.4
	Horizontal Flip	10	57.7
	Three Crop	10	59.9
	Three Crop	24	59.2
	Ten Crop	10	60.3
	Ten Crop	24	59.5

Table 3: Results of different backbones.

Network	Backbone	Acc(%)
TPN	3DResNet 50	53.1
	3DResNet 101	59.0
	3DResNet 152	60.3
	3DResNet 200	63.1
Slowonly	3DResNet 50	52.5
	3DResNet 101	59.9
	3DResNet 152	61.0
	3DResNet 200	63.8

Training setting. We applied cutmix and mixup on the basis of the original data augmentations in the training procedure of both TPN and slowonly networks. As is shown in Table 1, the modification on train augmentation achieved 0.7% and 0.6% accuracy improvement on TPN101 and Slowonly101 respectively, which demonstrates the positive effects of cutmix/mixup in action recognition.

Testing setting. According to the spatio-temporal inference analysis in section 2.3, the prediction consistency on multiple dimensions can boost the model performance. Practically, a sophisticated crop scheme upon the original input frames will bring a spatial difference and flexible video splits can maintain a temporal variation. As summarized in Table 2, compared to the simple data augmentation (center crop, random crop or horizontal flip), "three crop" increases the performance from 57.7% to 59.9%, and "ten crop" can further bring 0.4% improvement. For the modification in clips number, the 24 clips augmentation shows no noticeable advancement, however, a combination of different video segment strategies can achieve better accuracy in section 3.2.2.

Multiple scale backbones. In the data-efficient setting of action recognition, a larger backbone can enhance the generalization ability of the network. As is shown in Table 3, we tried a bunch of backbones of different depth for both

Table 4: Results on CodaLab Platform.

Method	Backbones	Acc(%)
(Slowonly)Multi-epochs+Three Crop	3DResNet 50	54.2
(TPN+Slowonly)Multi-epochs+Three Crop	3DResNet 50	56.4
(TPN+Slowonly)Multi-epochs+Three Crop	3DResNet 50, 3DResNet 101	60.5
(TPN+Slowonly)Multi-epochs+Three Crop+Mixup/Cutmix	3DResNet 50, 3DResNet 101	60.9
(TPN+Slowonly)Multi-epochs+Three Crop+Mixup/Cutmix	3DResNet 101, 3DResNet 152	62.6
(TPN+Slowonly)Multi-epochs+Ten Crop+Mixup/Cutmix	3DResNet 101, 3DResNet 152	62.9
(TPN+Slowonly)Multi-epochs+Ten Crop+Mixup/Cutmix+10/24 Clips	3DResNet 101, 3DResNet 152, 3DResNet 200	66.0

TPN and Slowonly network, and the deeper model achieves higher accuracy. For example, slowonly model constructed by 3DResNet 152 improves the performance by 8.5% compared to 3DResNet 50.

3.2.2 Results on CodaLab Platform

For the final prediction, we incorporate the above-mentioned multi-granularity temporal and spatial consistency, complementary network structures and multi-epochs models by the softmax-based ensemble strategy. Table 4 expatiates 7 different settings we have submitted to the CodaLab Platform and our optimal result is 66.0% on the testing set of Kinetics400ViPriors.

4. Conclusion

In our work, we focus on this data-efficient action recognition paradigm in this paper. We employ the Slowonly network to capture both fast-tempo and slow-tempo via encoding the features at different depths. Then we build a temporal pyramid network (TPN) to aggregates the information of various visual tempos at the feature level. The spatial and temporal predictions are combined to make the final consistency prediction. Experimental results on 2nd “Visual Inductive Priors for Data-Efficient Computer Vision” demonstrate that our method can achieve competitive results.

References

[1] 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.

[2] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2015.

[3] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. *ArXiv*, abs/1608.00859, 2016.

[4] X. Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

[5] Yunbo Wang, Mingsheng Long, Jianmin Wang, and Philip S. Yu. Spatiotemporal pyramid network for video action recognition. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2097–2106, 2017.

[6] Jie Wu, Yingying Li, Wei Zhang, Yi Wu, Xiao Tan, Hongwu Zhang, Shilei Wen, Errui Ding, and Guanbin Li. Modularized framework with category-sensitive abnormal filter for city anomaly detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4669–4673, 2020.

[7] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[8] Da Zhang, Xiyang Dai, and Yuan fang Wang. Dynamic temporal pyramid network: A closer look at multi-scale modeling for activity detection. In *ACCV*, 2018.