Challenge Report: 2nd VIPriors Action Recognition Challenge

1st Zihan Gao
Xidian University, Xi'an, Shaanxi Province, China
21171213889@stu.xidian.edu.cn

2nd Tianzhi Ma
Xidian University, Xi'an, Shaanxi Province, China
21171213975@stu.xidian.edu.cn

3th Jiaxuan Zhao
Xidian University, Xi'an, Shaanxi Province, China
jiaxuanzhao@stu.xidian.edu.cn

4th Licheng Jiao
Xidian University, Xi'an, Shaanxi Province, China
lchjiao@mail.xidian.edu.cn

5th Fang Liu
Xidian University, Xi'an, Shaanxi Province, China
f63liu@163.com

***Abstract*—In order to solve the training problem of the VIPriors Action Recognition Challenge with small sample datasets. In this paper, we propose a multi-network dynamic fusion model which combines self-attention mechanism with local perception. We acquire the predicted log likelihood by sending the prediction of each model through a softmax layer and weight their predictions by the score they achieve.This method achieved accuracy of 0.73 in this challenge. On large-scale datasets, our method achieves satisfying results without pre-trained model. The code will be released soon.**

Keywords—self-attention, multi-network, action recognition, no pre-trained

I. Introduction

The vision community is witnessing a modeling shift from CNNs to Transformers, where pure Transformer architecture-s have attained top accuracy on the major video recognition benchmarks [1]. Convolution-based backbone architectures have long dominated visual modeling in computer vision [2,3,4,5,6,7]. However, a modeling shift is currently underway on backbone architectures for video recognition, from Convolutional Neural Networks (CNNs) [18] to Transformers. The trend began with the introduction of Vision Transformer (ViT) [8], which globally models spatial relationships on non-overlapping image patches with the standard Transformer encoder. The great success of ViT on images has led to investigation of Transformer-based architectures for video-based recognition tasks. However, because of transformers' lacking inductive deviations of convolution (such as translation equivalence), it seems larger datasets and better regularization method is reqired to achieve state of art results.

Our approach is a multi-network dynamic fusion model which combines self-attention mechanism with local perception. While training, we enhance the data of the same target and send it to multiple robust networks to obtain different output feature maps, which is dynamically fused after each stage is finished. While testing a single video, We make predictions after multiple data augmentations for the same test object, and vote on the predicted results to select the most labeled result for better generalization.

II. Proposed Method

The schematic diagram of the proposed method is depicted in Fig 1. We adopt a multi-network structure, which dynamically integrates VIT and 3Dconvnets [9]. This architecture further consists of an ensemble of either different backbone architectures, which can combine self attention mechanism with local perception to obtain better performance.

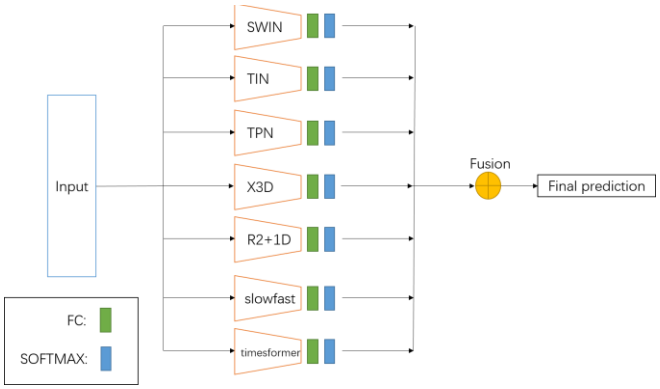

Fig. 1. Schematic Diagram of the proposed method (inference mode)

The Video Swin Transformer is majorly composed of the Video Swin Transformer block, which is built by replacing the multihead self-attention (MSA) [19] module in the standard Transformer layer with the 3D shifted window based multi-head self-attention module and keeping the other components unchanged. Specifically, a video transformer block consists of a 3D shifted window based MSA module followed by a feed-forward network, specifically a 2-layer MLP, with GELU[20] non-linearity in between. Layer Normalization (LN) [21] is applied before each MSA module and FFN, and a residual connection is also applied. In doing so, Video Swin Transformer is able to capture long range memory and thus achieves SOTA in many action recoginition benchmarks.

In order to integrate temporal information at different times, TIN provides different frames with a unique interlacing offset. Instead of habitually assigning each channel with a separately learnable offset, TIN adopt distinctive offsets for different channel groups. As observed in previous experiment, human perception on object motion focuses on different temporal resolutions. To maintain temporal fidelity and recognizes patial semantics jointly,different groups of temporal receptive fields pursuit a thorough separation of expertise convolution. Besides, groups of offsets also reduce the model complexity as well as stabilize the training procedure across heavy backbone architectures.

TABLE I. TRAINING CONFIG

models	config				
	backbone	***lr***	***epochs***	***lr schedule***	***warm up***
slowfast	3DResnet152	0.1	256	Cosine decay	34 epoch
TIN	Resnet50	0.05	200	Cosine decay	20 epoch
TPN	Resnet50	0.1	150	Cosine decay	10 epoch
VideoSwin Transformer	Swin-B 3D	0.001	300	Cosine decay	20 epoch
r2plus1d	Resnet34 2 plus 1d	0.1	180	Cosine decay	None
TimesFormer	ViT	0.05	300	Cosine decay	20 epoch
X3D-M	X3D-M	0.05	180	Cosine decay	10 epoch

Inspired by the observation that features at multiple depths in a single network already cover various visual tempos, we propose a feature-level temporal pyramid network (TPN) for modeling the visual tempo. TPN could operate on only a single network no matter how many levels are included in it. Moreover, TPN could be applied to different architectures in a plug-and-play manner.

SlowFast present a two-pathway SlowFast model for video recognition. One pathway is designed to capture semantic information that can be given by images or a few sparse frames, and it operates at low frame rates and slow refreshing speed. In contrast, the other pathway is responsible for capturing rapidly changing motion, by operating at fast refreshing speed and high temporal resolution. Despite its high temporal rate, this pathway is made very lightweight, e.g., ∼20% of total computation. This is because this pathway is designed to have fewer channels and weaker ability to process spatial information, while such information can be provided by the first pathway in a less redundant manner. We call the first a Slow pathway and the second a Fast pathway, driven by their different temporal speeds. The two pathways are fused by lateral connections.

TimesFormer propose a more efficient architecture for spatiotemporal attention, named “Divided Space-Time Attention” (denoted with T+S), where temporal attention and spatial attention are separately applied one after the other. For Divided Attention, within each block, TimesFormer compute temporal attention by comparing each patch (p, t) with all the patches at the same spatial location in the other frames.

Our intuition is to take advantage of each network to achieve more generalization. For each subnetworks, there will be randomly enhanced input with the same datafor training, such as Slowfast[10], Timesformer[11], TIN[12], TPN[13], Video Swin Transformers, R2plus1d[14], X3D[15], etc. There are three main ways of data augmentations: (1) spatial augmentations: horizontal flip (2) appearance transformation: random gray and color jitter (3) time augmentations: a uniform interval, random number of jump frames and random start frames. While testing a single video, We make predictions after multiple data augmentations for the same test object, and vote on the predicted results of different augmentations for the same object to select the most labeled result,which helps in mitigating common generalization, errors as well as decreasing the variance in neural network predictions.

III. EXPERIMENTS

This section covers the dataset, results, and.

A. Dataset

In this particular challenge, the task is Action Recognition and the dataset is Kinetics400ViPriors, a modification of the official Kinetics400[16] dataset. The Kinetics400ViPriors dataset is a high-quality dataset for human action recognition in videos. The dataset consists of around 70000 video clips covering 400 human action classes with at least 400 video clips for each action class. Each video clip lasts around 10 seconds and is labeled with a single action class. The videos are collected from YouTube.

B. Implantation details

For all networks except TimesFormer, we clip each vedio into 32 frames with 2 frame interval, then resize frames to (224, 224) . Also, data augmentations such as random flip and random colorjitter is adopted.

For testing, TTA(Test Time augmentation) [22] is adopted and we use the same clip length of each vedio, only randomly clip 10 times and ensemble these ten predictions to acquire final results, which we believe would help in achieving better generalization, as well as decreasing the variance caused by random vedio clipping. More details on the training methods for each model are provided in Table1.

C. Results

We make predictions after multiple data augmentations for the same test object, and vote on the predicted results of different augmentations for the same object to select the most labeled result,which helps in mitigating common generalization, errors as well as decreasing the variance in neural network predictions[17]. More details on the training and testing methods for each model are provided in Table1.

We performed our initial experiments on Kinetics400 to observe the performance of our ensemble and model selection purposes. The proposed method achieves a Top-1 accuracy of 84.5% on Kinetics400 (split 1) test set without using any pre-trained weights in our training. We performed our training and

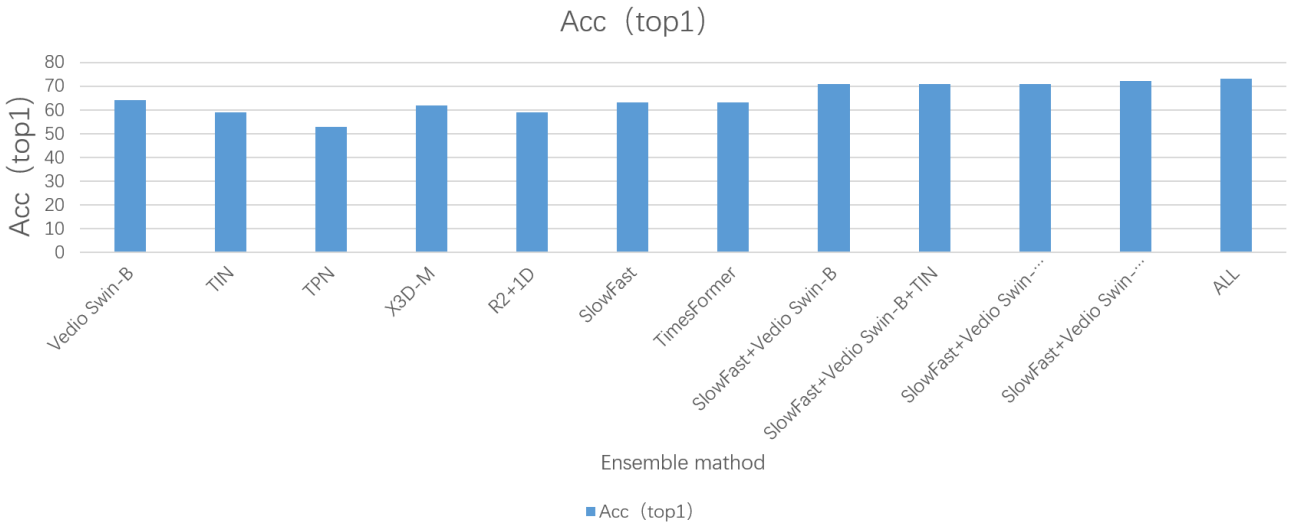

Fig. 2. Results of differbet ensemble metho

model selection on the VIPriors Action Recognition dataset in the same manner. The final fused model (row-9) of Table-1 is used to predict the output of the competition test set and achieves a Top-1 accuracy of 73%.

TABLE II. THE RESULTS OF DIFFERENT ENSEMBLE METHOD

Ensemble mathod	**Acc（Top 1）**
Vedio Swin-B	64
TIN	59
TPN	53
X3D-M	62
R2+1D	59
SlowFast	63
TimesFormer	63
SlowFast+Vedio Swin-B	71
SlowFast+Vedio Swin-B+TIN	71
SlowFast+Vedio Swin-B+TIN+ TPN	71
SlowFast+Vedio Swin-B+TIN+ TPN+X3D-M+ TimesFormer	72
ALL	**73**

CONCLUSION

In this paper, we mainly focused on obtaining reliable predictions on small sampled datasets. We adopt several models to take advantage of each and vote on which, we acquire our final result.

REFERENCES

[1] Liu Z, Ning J, Cao Y, et al. Video swin transformer[J]. arXiv preprint arXiv:2106.13230, 2021.

[2] LeCun, Y ., Bottou, L., Bengio, Y ., Haffner, P ., et al. (1998). Gradient-based learning applied to

[3] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. InAdvances in neural information processing systems, pages

[4] Simonyan, K. and Zisserman, A. (2015). V ery deep convolutional networks for large-scale image recognition. InInternational Conference on Learning Representations.

[5] Szegedy, C., Liu, W., Jia, Y ., Sermanet, P ., Reed, S., Anguelov, D., Erhan, D., V anhoucke, V ., and Rabinovich, A. (2015). Going deeper with convolutions. InProceedings of the IEEE conference on computer vision and pattern recognition, pages 1–9.

[6] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778.

[7] Huang, G., Liu, Z., V an Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. InProceedings of the IEEE conference on computer vision and pattern recognition, pages 4700–4708.

[8] Arnab A, Dehghani M, Heigold G, et al. Vivit: A video vision transformer[J]. arXiv preprint arXiv:2103.15691, 2021.

[9] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3d convolutional networks[C]//Proceedings of the IEEE international conference on computer vision. 2015: 4489-4497.

[10] Feichtenhofer C, Fan H, Malik J, et al. Slowfast networks for video recognition[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 6202-6211.

[11] Bertasius G, Wang H, Torresani L. Is Space-Time Attention All You Need for Video Understanding?[J]. arXiv preprint arXiv:2102.05095, 2021.

[12] Shao H, Qian S, Liu Y. Temporal interlacing network[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(07): 11966-11973.

[13] Yang C, Xu Y, Shi J, et al. Temporal pyramid network for action recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 591-600.

[14] Tran D, Wang H, Torresani L, et al. A closer look at spatiotemporal convolutions for action recognition[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2018: 6450-6459.

[15] Feichtenhofer C. X3d: Expanding architectures for efficient video recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 203-213.

[16] Kay W, Carreira J, Simonyan K, et al. The kinetics human action video dataset[J]. arXiv preprint arXiv:1705.06950, 2017.

[17] Perez L, Wang J. The effectiveness of data augmentation in image classification using deep learning[J]. arXiv preprint arXiv:1712.04621, 2017.

[18] Vedaldi A, Lenc K. Matconvnet: Convolutional neural networks for matlab[C]//Proceedings of the 23rd ACM international conference on Multimedia. 2015: 689-692.

[19] Vedaldi A, Lenc K. Matconvnet: Convolutional neural networks for matlab[C]//Proceedings of the 23rd ACM international conference on Multimedia. 2015: 689-692.

[20] Hendrycks D, Gimpel K. Gaussian error linear units (gelus)[J]. arXiv preprint arXiv:1606.08415, 2016.

[21] Ba J L, Kiros J R, Hinton G E. Layer normalization[J]. arXiv preprint arXiv:1607.06450, 2016.

[22] Wang G, Li W, Aertsen M, et al. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks[J]. Neurocomputing, 2019, 338: 34-45.