Understanding the Behaviors of Modern Video Models Through Signal Perception

Author: Yifan Chen; Supervisor: Dr Lei Wang

Australian National University

GRIFFITH UNIVERSITY

TIME LAB

Background & Objective

Look into Background:

- Video action recognition is widely applied in surveillance, sports analytics, etc.
- It decodes human actions by aggregating spatial features and temporal features.
- Transformer backbones now dominate key benchmarks in this field.

Challenges Remains?

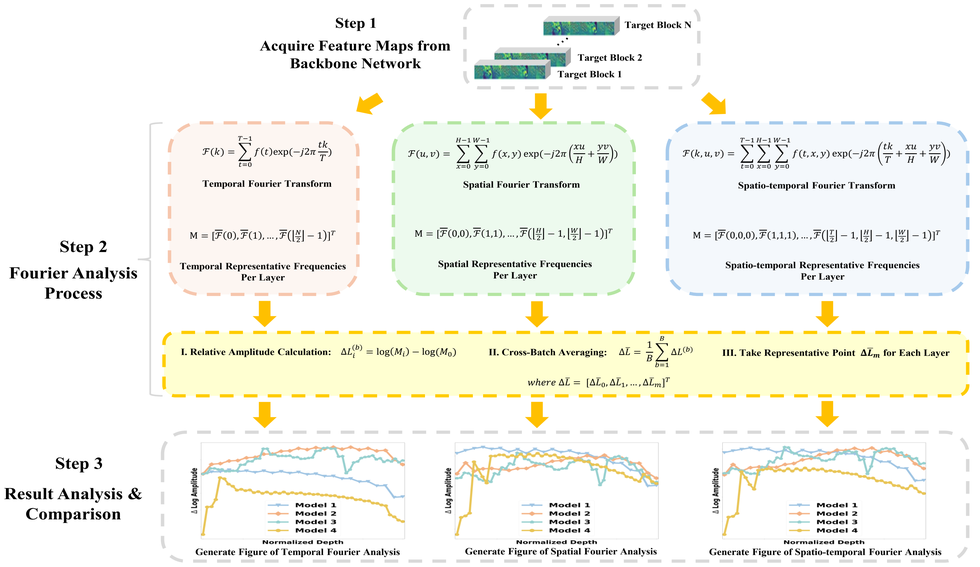
- How do transformer-based models deal with spatio-temporal features?
- How does fine-tuning adapt these models to diverse datasets?
- Which components drive robustness, and how can visualization expose bottlenecks?

Our Objectives Focus on:

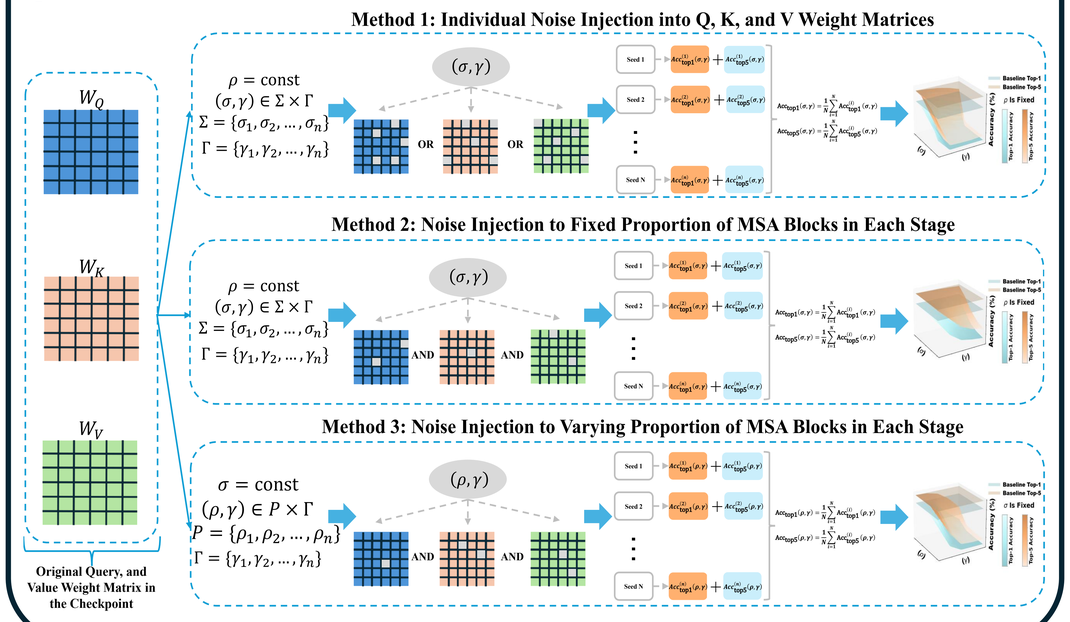
- Explore spatio-temporal feature processing of modern models through analysis of key intermediate blocks.
- Analyze fine-tuning's impact on cross-dataset generalization based on differing dataset characteristics.
- Evaluate the model's robustness to Gaussian Noise in key components under different noise settings and impact ranges.

Method Design

The proposed analysis framework consists of Fourier analysis and noise hacking techniques. The first flowchart below illustrates that the first tool employs one-dimensional to three-dimensional Fourier transforms to decompose feature maps, comparing how architectures capture information from a frequency perspective and the impact of fine-tuning.



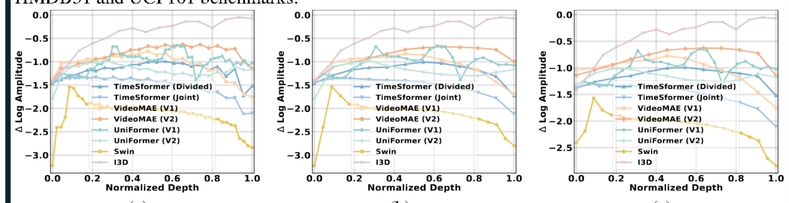
The second flowchart below explains how noise hacking techniques are implemented through three methods to test the transformer-based models' internal robustness and potential bottlenecks.



Experiments

Backbone	Pretraining	Fine-tuning/testing	Top-1	Top-5	Reference Top-1	Reference Top-5
I3D	Kinetics-400	HMDB51	71.44	93.20	74.30	–
		UCF101	94.92	99.50	95.10	–
TimeSformer (Divided)		HMDB51	71.37	93.86	–	–
		UCF101	95.82	99.68	–	–
TimeSformer (Joint)		HMDB51	68.17	93.33	–	–
		UCF101	93.71	99.58	–	–
VideoMAE (V1)		HMDB51	72.22	92.48	73.30	–
		UCF101	95.32	99.37	96.10	–
VideoMAE (V2)	Kinetics-710	HMDB51	79.35	96.86	–	–
		UCF101	98.31	99.92	–	–
UniFormer (V1)	Kinetics-400	HMDB51	76.41	95.10	–	–
		UCF101	97.17	99.71	–	–
UniFormer (V2)	CLIP-400M + Kinetics-710	HMDB51	75.23	95.42	–	–
		UCF101	96.51	99.76	–	–
Swin	Kinetics-400	HMDB51	75.16	93.46	–	–
		UCF101	96.11	99.50	–	–

Table 1: Comparison of different backbones pretrained on various datasets, and then fine-tuned on HMDB51 and UCF101 benchmarks.



(a) (b) (c)

Figure 1: Temporal Fourier analysis of the outputs from intermediate core modules during testing on the HMDB51 dataset, with models fine-tuned on split 1. From left to right: (a) the analysis of the outputs from all MSA and MLP blocks of the transformer-based models, and from all Conv blocks of the CNN-based model; (b) the analysis of the outputs from all MSA blocks of the transformer-based models, and from all Conv blocks of the CNN-based model; (c) the analysis of the outputs from all MLP blocks of the transformer-based models, and from all Conv blocks of the CNN-based model.

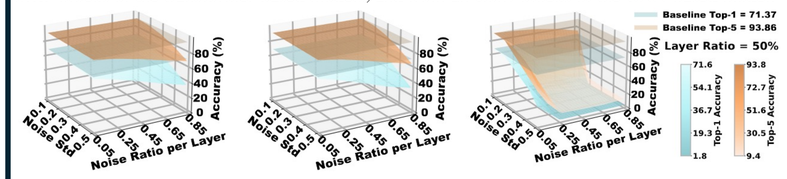


Figure 2: Random gaussian noise with a consistent layer ratio is respectively applied to Q, K, and V Weight Matrix of the spatial attention layers in Divided Space-Time TimeSformer, visualized from left to right.

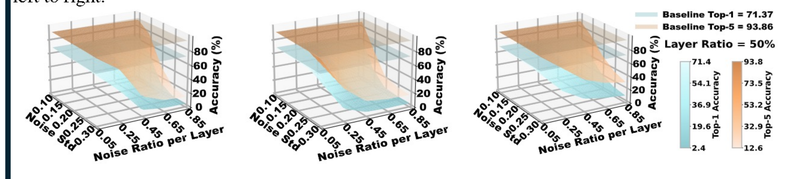


Figure 3: Random Gaussian noise with a consistent layer ratio is applied to the spatial attention layers of Divided Space-Time TimeSformer. The model's layers are sequentially divided into three equal parts (front, middle, and rear), and selected from each part respectively, visualized from left to right.

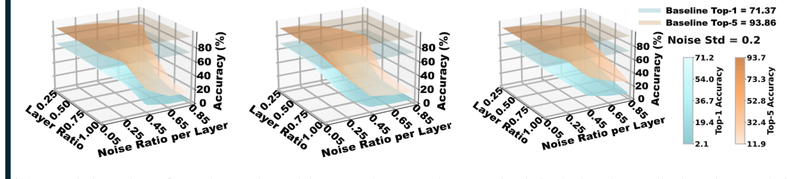


Figure 4: Random Gaussian noise with a consistent noise standard deviation is applied to the spatial attention layers of Divided Space-Time TimeSformer. The model's layers are sequentially divided into three equal parts (front, middle, and rear), and selected from each part respectively, visualized from left to right.