A Deeper Dive Into What Deep Spatiotemporal Networks Encode: Quantifying Static vs. Dynamic Information

Matthew Kowal[1,2] Mennatullah Siam[1] Md Amirul Islam[2,3]

Neil D. B. Bruce[2,5] Richard P. Wildes[1,4] Konstantinos G. Derpanis[1,2,4]

[1]York University [2]Vector Institute [3]Ryerson University [4]Samsung AI Centre Toronto [5]University of Guelph

CVPR JUNE 19-24 2022 NEW ORLEANS LOUISIANA

Project Page: yorkucvil.github.io/Static-Dynamic-Interpretability

Motivation and Research Questions

Video understanding models are known to be biased toward static information and there is a lack of techniques to quantify this bias in relation to dynamic information. Towards a solution to this problem, we propose a method for measuring static vs. dynamic bias in deep video models with the goal of answering the following four questions:

1. What effect does model architecture have on static and dynamic biases?
2. How does the training dataset affect static and dynamic biases?
3. What role do units jointly encoding static and dynamic play in relation to architecture and dataset?
4. What impact do static and dynamic units have on performance?

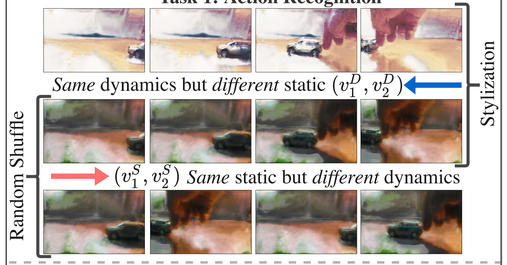
Method

Step 1 - Sampling Static and Dynamic Video Pairs

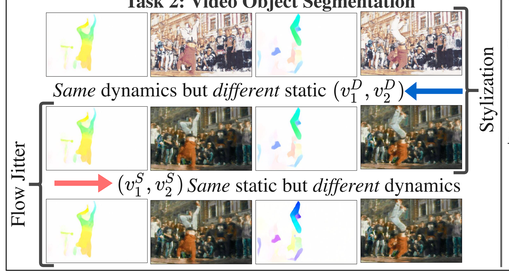
We first sample a video, v, and a factor, F (i.e., static or dynamic).

We then generate static and dynamic video pairs for two tasks:

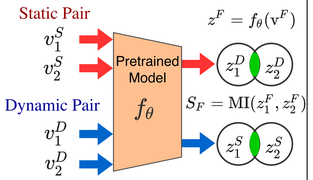
Task 1: Action Recognition



Task 2: Video Object Segmentation



Step 2 - Estimating Static and Dynamic Information



(i) Layerwise metric

$$N_F = \frac{\exp(S_F)}{\sum_{k=0}^{K}\exp(S_k)} \cdot N^l$$

Where: $K = \mathbf{card}(F)$, $N^l = \mathbf{card}(z^F)$

$$S_F = \sum_{i=1}^{N^l} \frac{\mathrm{Cov}(z_1^F(i), z_2^F(i))}{\sqrt{\mathrm{Var}(z_1^F(i))\mathrm{Var}(z_2^F(i))}}$$

(ii) Unitwise metric $(\lambda = 0.5)$

$$s_F^i = \frac{\mathrm{Cov}(z_1^F(i), z_2^F(i))}{\sqrt{\mathrm{Var}(z_1^F(i))\mathrm{Var}(z_2^F(i))}}$$

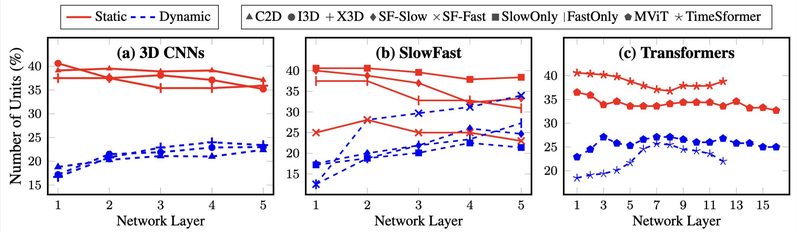
$$N_F = \sum_{i=1}^{N^l} 1[s_F^i > \lambda \wedge s_k^i < \lambda \forall k \in K, k \neq F]$$

$$N_{\text{Joint}} = \sum_{i=1}^{N^l} 1[s_F^i > \lambda \forall F \in K]$$

$$N_{\text{Residual}} = \sum_{i=1}^{N^l} 1[s_F^i < \lambda \forall F \in K]$$

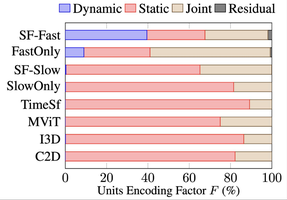
Architecture Analysis

Action Recognition

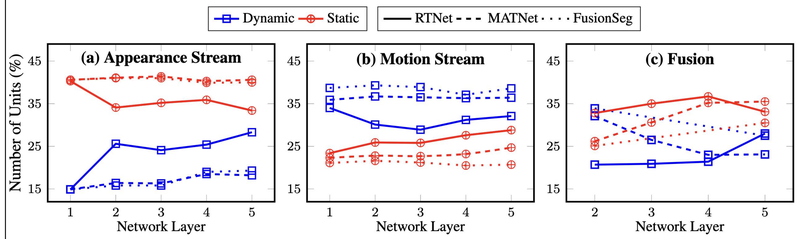


Top: (Layerwise) All single stream architectures are static biased. The Fast branch in the SlowFast model captures significantly more dynamics than statics.

Right: (Unitwise) Only static and joint units are produced in all architectures other than the FastOnly model and SlowFast-Fast branch.

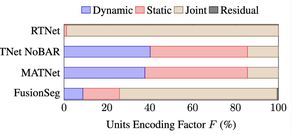


Video Object Segmentation



Top: (Layerwise) For all models, each stream is biased toward its corresponding factor while in the fusion layers all models are generally static biased.

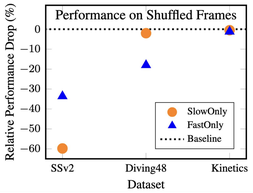
Right: (Unitwise) Joint units are dominantly produced except in MATNet, which shows a balance between static, dynamic and joint units.



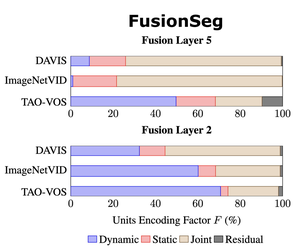
Dataset Analysis

Action Recognition

Dataset	SlowOnly		FastOnly	
	Dyn.(%)	Stat.(%)	Dyn.(%)	Stat.(%)
Kinetics	21.4	38.4	27.3	30.9
Diving48	23.1	34.0	23.8	27.3
SSv2	28.2	30.7	31.6	21.9



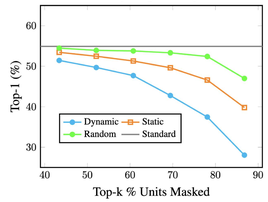
Video Object Segmentation



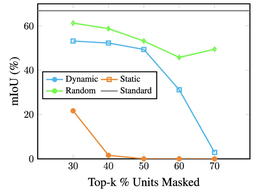
Right: TAO-VOS results in the most dynamics.

Left: SSv2 results in the most dynamics and Diving48 contains similar dynamics as Kinetics.

Effect on Performance - Neuron Masking



Left: Dynamic units harm performance the most.



Right: Static units harm performance the most.

Conclusion

1. We have advanced the understandability of learned deep spatiotemporal models for video understanding, particularly for action recognition and video object segmentation.
2. We introduced a general method for analyzing the extent to which architectures capture static vs. dynamic information.
3. We demonstrated how this method can be applied to investigate biases in architectures, datasets and impact overall performance.