Benchmarking Data and Computational Efficiency of ActionFormer on Temporal Action Localization Tasks

Author: Jan Warchocki
Supervisors: Dr. Jan van Gemert, Robert-Jan Bruintjes, Attila Lengyel, Ombretta Strafforello

1. Introduction

In a video, temporal action localization (TAL) is:

- predicting the start and the end of an action,
- predicting the class of the action.

Current state-of-the-art (SOTA) models are trained on large datasets such as ActivityNet [1] or THUMOS'14 [2]. On top of that, training these models is typically computationally expensive.

➢ Would be desirable to explore how SOTA models work in settings with limited data or limited computational power available.

Data efficient models have been proposed [3], but they are incompatible with current SOTA. Computational analysis of the state-of-the-art could also be expanded.

2. Research question

How well does ActionFormer perform and generalize in limited data and compute settings?

3. Model

Choice of ActionFormer [4], due to it showing one of the first uses of transformers in TAL.

Additionally, newer models, such as TriDet [5], are inspired by the architecture of ActionFormer.

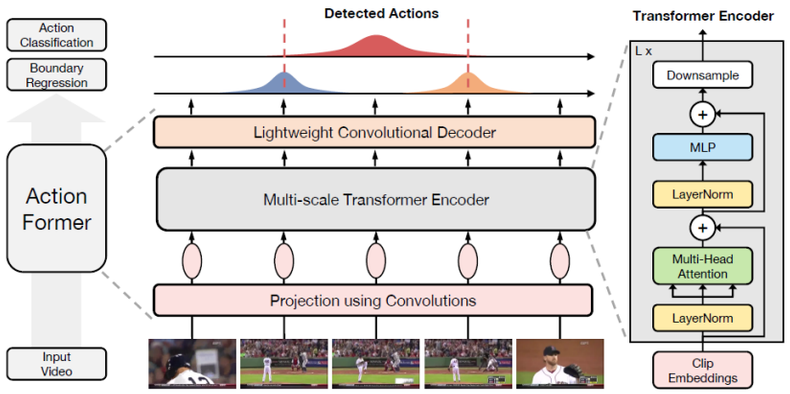

Figure 1. The ActionFormer model. Image taken from [4].

4. Methodology

Data efficiency

Train on a percentage p of the training set and report performance on the test set. Repeat multiple times to understand the variance in results.

Computational efficiency

Training: train the model on a dataset and report the time it took alongside the reached mean average precision.

Inference: pass videos of increasing lengths and note down the time it took, memory consumption, and the number of floating point operations (MACs).

5. Results

Data efficiency

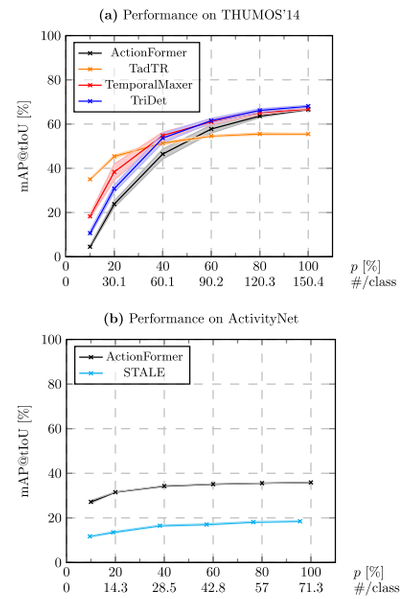

Figure 2. Data efficiency results.

➢ The TemporalMaxer and TriDet models should be chosen in favour of ActionFormer when training data is limited.

Computational efficiency

Tables 1 & 2. Training efficiency results on THUMOS'14 and ActivityNet.

Model	Time [s]	Avg. mAP [%]
ActionFormer	886.8 ± 54.3	65.89 ± 0.09
TadTR	425.7 ± 3.5	55.3 ± 0.63
TemporalMaxer	2956.6 ± 1660	66.96 ± 0.37
TriDet	646.2 ± 26.1	68.07 ± 0.42

Model	Time [s]	Avg. mAP [%]
ActionFormer	1944.9 ± 60.6	35.9 ± 0.14
STALE	400.7 ± 5.8	19.37 ± 0.16

➢ The ActionFormer model is unlikely to be selected in scenarios with limited training time available.

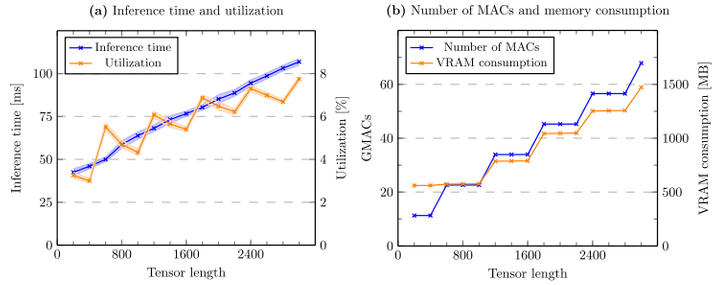

Figure 3. Inference efficiency results.

➢ Most importantly, the inference time, complexity, and memory consumption all increase linearly.

6. Conclusion

➢ Other models offer better data or computing efficiencies.
➢ The model was found to scale linearly with input length. This thus matches the theory from the original paper [4].
➢ Future work could involve attempting to improve data or computational efficiencies of the model.

References

[1] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, 'ActivityNet: A large-scale video benchmark for human activity understanding', in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 961–970.
[2] A. Gorban et al. "THUMOS Challenge: Action Recognition with a Large Number of Classes." http://crcv.ucf.edu/THUMOS14/ (accessed 20/06/2023).
[3] S. Nag, X. Zhu, and T. Xiang, 'Few-Shot Temporal Action Localization with Query Adaptive Transformer'. p. arXiv:2110.10552, October 01, 2021.
[4] C. Zhang, J. Wu, and Y. Li, 'ActionFormer: Localizing Moments of Actions with Transformers'. p. arXiv:2202.07925, February 01, 2022.
[5] D. Shi, Y. Zhong, Q. Cao, L. Ma, J. Li, and D. Tao, 'TriDet: Temporal Action Detection with Relative Boundary Modeling'. p. arXiv:2303.07347, March 01, 2023.

j.z.warchocki-1@student.tudelft.nl

TU Delft