Efficient Video Action Recognition

How well does the TriDet model perform and generalize in a limited compute power and data setting?

Author: Alexandru Damacus (A.Damacu@student.tudelft.nl)
Responsible Professor: Dr. Jan van Gemert
Supervisors: Ombretta Strafforello, Robert-Jan Bruintjes, Attila Lengyel

1 Introduction

Temporal Action Localization (TAL) refers to the process of understanding actions that are happening in an input video. It consists of two main parts:

- predicting the temporal boundaries of the action.
- recognizing the action present in the interval denoted by those temporal boundaries.

It has a large number of applications in action detection, surveillance, video summarising, and more.

2 Objective

Novel TAL models are increasingly resource heavy in order to achieve good results on popular benchmarks, and researchers rarely perform extensive studies on the data and compute efficiency of their models [2]. The goal of this study is to determine how well current state-of-the-art models perform in a limited data and computation power setting.

3 Research Question

How well does TriDet [1] perform and generalize in a limited data and computation power environment?

4 Methodology

Data efficiency

- Sample the original training set from THUMOS'14 [3] according to 5 different percentages p.
- Train the model and evaluate it five times on the validation set for each percentage p.
- Measure mean average precision (mAP) at each temporal intersection over union (tIoU) from [0.3, 0.4, 0.5, 0.6, 0.7].
- Report mean and standard deviation for mAP at each tIoU, and at average tIoU.

Compute efficiency

1. **Training performance**

- Train and evaluate the model on the full dataset several times.
- Measure training time.
- Report the mean and standard deviation for both training time and mAP.

2. **Inference performance**

- Pass random tensors of different lengths through the model.
- Measure inference time, giga multiply-accumulate operations (GMACs), and memory usage.
- Report mean and standard deviation for those metrics.
- Additionally, report GPU utilization.

5 Results

Data efficiency

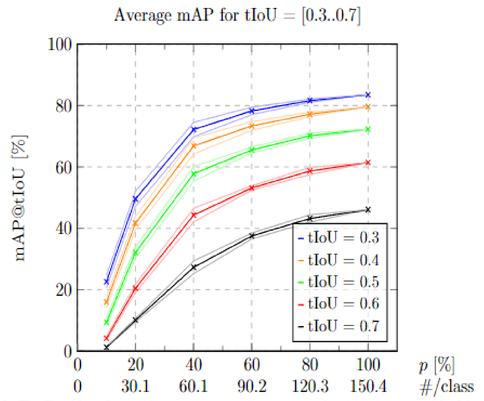

Figure 1: TriDet's [1] data efficiency results for each tIoU in [0.3 .. 0.7].

Compute efficiency - Training

Model	Time [s]	Avg. mAP [%]
TriDet	646.17 ± 26.12	68.07 ± 0.42
ActionFormer	866.22 ± 26.97	66.5 ± 0.31
TadTR	425.72 ± 3.469	55.3 ± 0.63
TemporalMaxer	2955.64 ± 1659.98	66.96 ± 0.37

Figure 2: The research group's models compared by training time and mAP on THUMOS'14 [3].

Compute efficiency - Inference

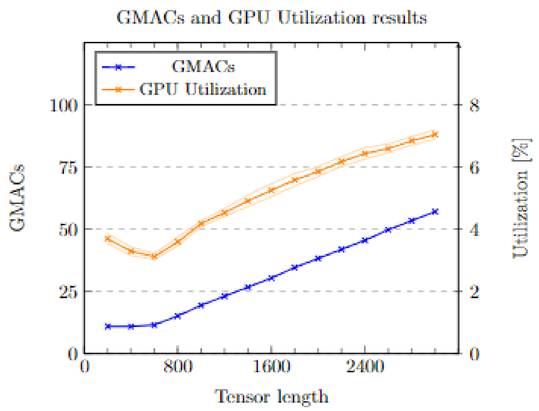

Figure 3: TriDet's [1] GMACs and GPU utilization for each tensor length.

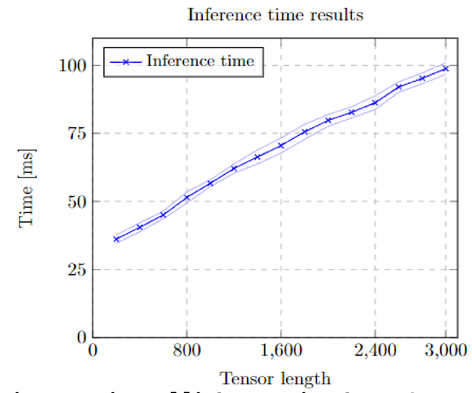

Figure 4: TriDet's [1] inference time for each tensor length.

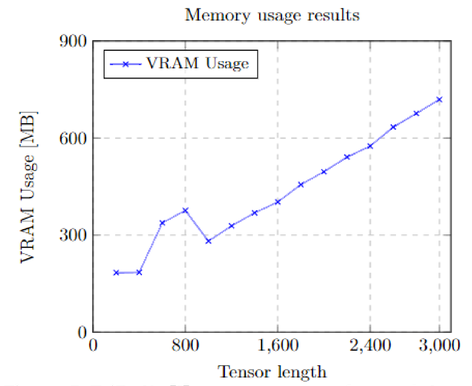

Figure 5: TriDet's [1] memory usage for each tensor length.

6 Conclusion

- Data efficiency results show that TriDet [1] achieves close to SOTA results with only 60% of the training data.
- Data efficiency results show a distinguishable curve pattern, which could allow for infering full capabilities of a model from partial training sets.
- Compute efficiency results show that all of the metrics increase linearly along with tensor (video) lengths passed through the model.
- There are several metrics that would provide interesting insight into novel TAL models, such as transfer learning or zero-shot learning capabilities for data efficiency, and model size or energy consumption for compute efficiency.

References

[1] Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. Tridet: Temporal action detection with relative boundary modeling, 2023.
[2] Huifen Xia and Yongzhao Zhan. A survey on temporal action localization. IEEE Access, 8:70477–70487, 2020.
[3] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. http://crcv.ucf.edu/THUMOS14/, 2014.

Acknowledgements

Special thanks to **Jan Warchocki** for sharing the table shown in Figure 2 with the research group.