EVALUATION OF THE SUM-GAN-AAE METHOD FOR VIDEO SUMMARIZATION

Authors
Georgi Trevnenski (G.D.Trevnenski@student.tudelft.nl)
Supervisors
Ombretta Strafforello, Dr. Seyran Khademi

Affiliations
TU Delft

INTRODUCTION

SUM-GAN-AAE is an unsupervised deep learning model for video summarization developed by Apostolidis et al. This study tests the algorithm on the Breakfast dataset and evaluates it with the rank correlation coefficients: Kendall's τ and Spearman's ρ. The Breakfast dataset is developed for action localization algorithms and the research assumes that actions in a video contain key information. Rank correlation coefficients are a statistical tool used for evaluating the correlation between rankings, in our case - ranked frames by importance.

OBJECTIVES

- Evaluate the performance of the algorithm on the Breakfast dataset
- Compare the F-Score and rank correlation coefficients metrics in evaluating video summaries.
- Perform hyperparameter optimization

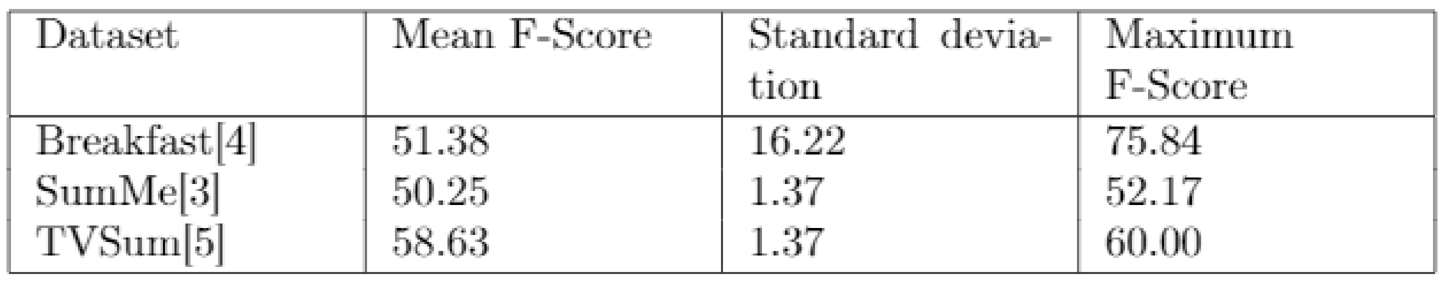
METHODOLOGY

The SUM-GAN-AAE model is trained and evaluated on the video summarization datasets and on the Breakfast dataset using k-fold cross-validation. The rank correlation coefficients are calculated and compared to the F-Score. The importance scores per frame are plotted and compared to the ground truth.

RESULTS

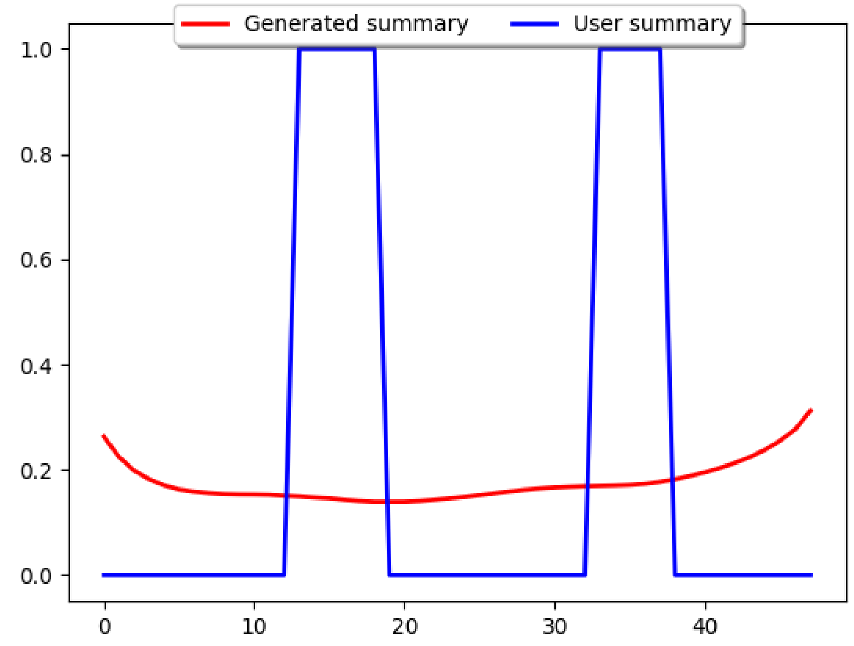
- **Finding 1:** F-Score and rank correlation coefficients differ significantly
- **Finding 2:** The average rank correlation coefficients are close to 0
- **Finding 3:** Generated importance scores follow the same pattern
- **Finding 4**: The optimal learning rate for the Breakfast dataset is around 0.000001

F-SCORE RESULTS

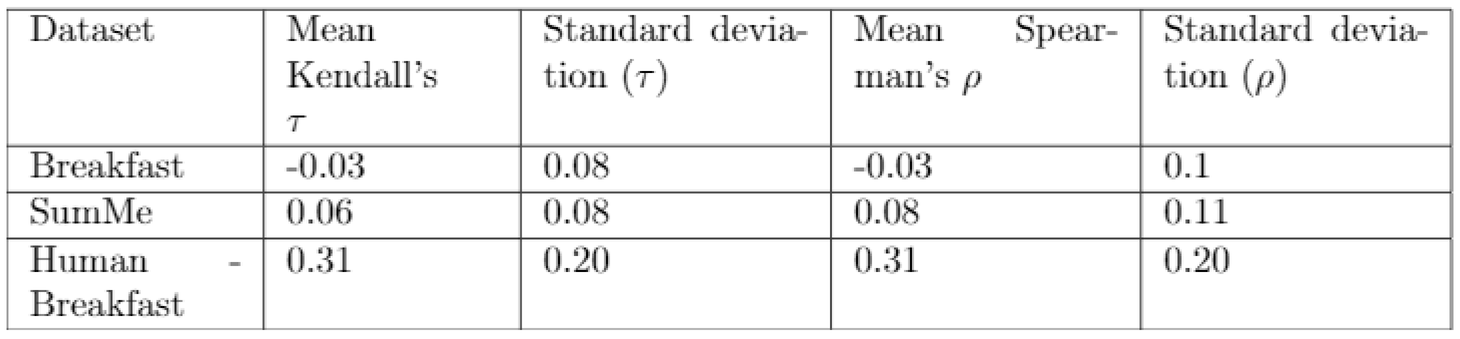


Dataset	Mean F-Score	Standard deviation	Maximum F-Score
Breakfast[4]	51.38	16.22	75.84
SumMe[3]	50.25	1.37	52.17
TVSum[5]	58.63	1.37	60.00

IMPORTANCE SCORES



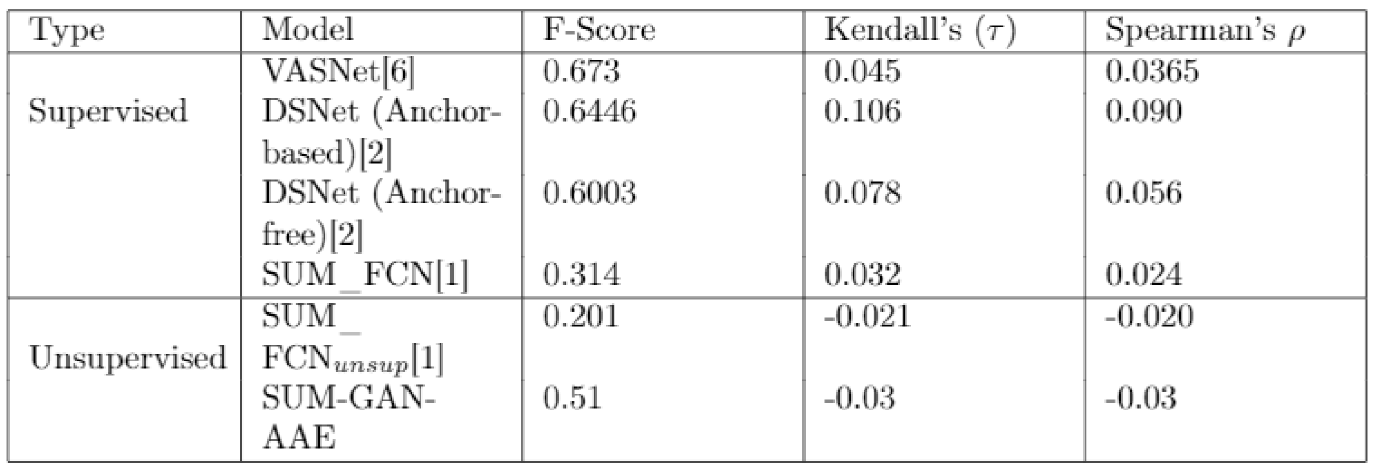
RANK CORRELATION COEFFICIENTS



Dataset	Mean Kendall's τ	Standard deviation (τ)	Mean Spearman's ρ	Standard deviation (ρ)
Breakfast	-0.03	0.08	-0.03	0.1
SumMe	0.06	0.08	0.08	0.11
Human - Breakfast	0.31	0.20	0.31	0.20

RESEARCH GROUP WORK

SUPERVISED COMPARED TO UNSUPERVISED ALGORITHMS



Type	Model	F-Score	Kendall's (τ)	Spearman's ρ
Supervised	VASNet[6]	0.673	0.045	0.0365
	DSNet (Anchor-based)[2]	0.6446	0.106	0.090
	DSNet (Anchor-free)[2]	0.6003	0.078	0.056
	SUM_FCN[1]	0.314	0.032	0.024
Unsupervised	SUM_FCN$_{unsup}$[1]	0.201	-0.021	-0.020
	SUM-GAN-AAE	0.51	-0.03	-0.03

CONCLUSION

- A high F-Score does not mean high similarity with the ground truth
- The rank correlation calculated are similar to those of a random frame selection
- The generated importance scores are very similar for all videos
- Results for unsupervised algorithms indicate randomness in the generated summaries
- Results for supervised algorithms still fall short of those of human-generated summaries.

REFERENCES

- [1] Paul Frölke. “Evaluation of Video Summarization UsingFully Convolutional Sequence Networks on Action Local-ization Datasets”. In: (2021).
- [2] Daan Groenewegen. “Evaluation of Video SummarizationUsing DSNet and Action Localization Datasets”. In: (2021).
- [3] Michael Gygli et al. “Creating Summaries from User Videos”.In:ECCV. 2014.
- [4] Hilde Kuehne, Juergen Gall, and Thomas Serre. “An end-to-end generative framework for video segmentation andrecognition”. In:Proc. IEEE Winter Applications of Com-puter Vision Conference (WACV 16). Lake Placid, Mar.2016.
- [5] Yale Song et al. “TVSum: Summarizing Web Videos UsingTitles”. In:Proceedings of the IEEE Conference on Com-puter Vision and Pattern Recognition (CVPR). June 2015.
- [6] Felicia Elfrida Tjhai. “Evaluating the Supervised VideoSummarization Model VASNet on an Action LocalizationDataset”. In: (2021).