Evaluating the supervised video summarization model VASNet on an action localization dataset

Author: Felicia Elfrida Tjhai (FeliciaElfridaTjhai@student.tudelft.nl)
Supervisor: Ombretta Strafforello
Responsible professor: dr. Seyran Khademi

1 INTRODUCTION

Video summarization is the task of creating a shorter version of a video while preserving that video's main storyline. This task suffers from the problem of subjectivity because, for the same video, different human annotators can create different summaries. Supervised models are especially affected because they rely on human-generated summaries when learning to build summaries. There may also be the reason behind the observation that unsupervised models sometimes outperform supervised models.

This leads us to investigate the effect of action localization on the task of video summarization. To do that, we used the Breakfast Actions datset.

VASNet [1] is a supervised video summarization model that uses "soft, self-attention".

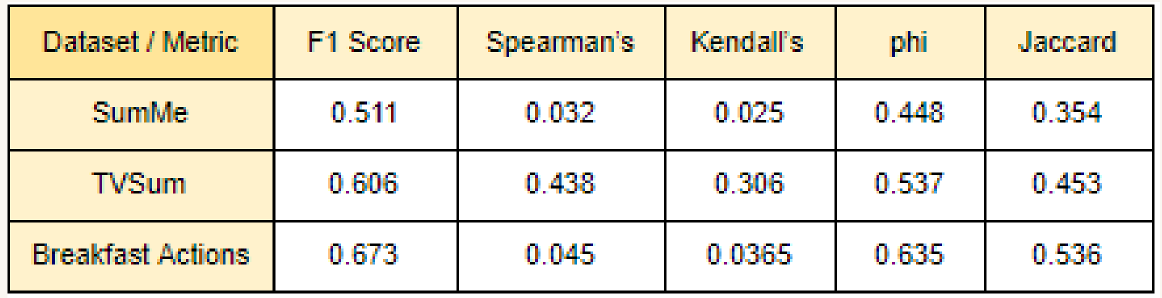
2 RESEARCH QUESTION

How well can a supervised model (VASNet) trained with ground-truth importance scores based on action localization learn representations for video summarization?

3 METHOD

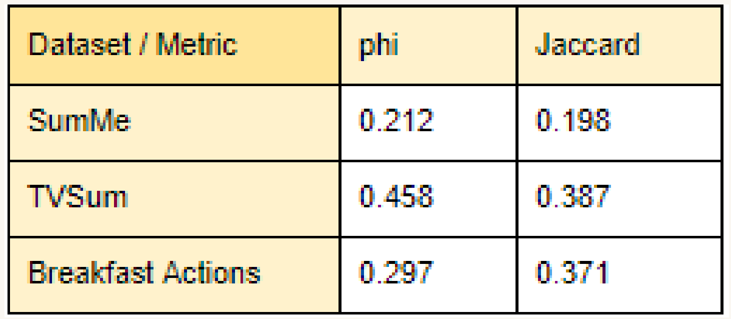
1. Train VASNet on the Breakfast Actions dataset
2. Evaluate performance using F1 score
3. Evaluate performance using correlation metrics:
   a. Spearman's rho
   b. Kendall's tau
   c. phi (Matthews Correlation Coefficient)
   d. Jaccard
4. Compare performance with SumMe and TVSum
5. Compare performance with other supervised and unsupervised models

4 RESULTS



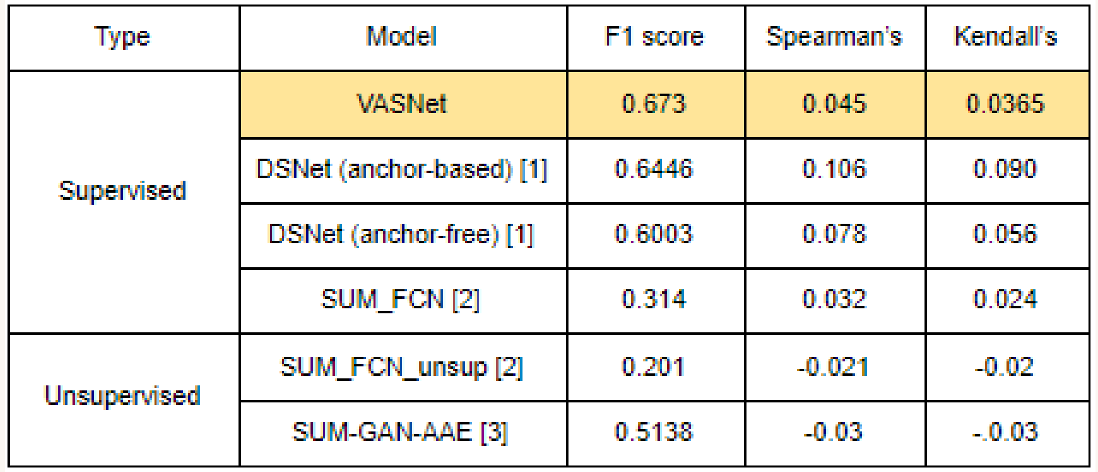
Dataset / Metric	F1 Score	Spearman's	Kendall's	phi	Jaccard
SumMe	0.511	0.032	0.025	0.448	0.354
TVSum	0.606	0.438	0.306	0.537	0.453
Breakfast Actions	0.673	0.045	0.0365	0.635	0.536

Figure 1: VASNet's performance on the SumMe, TVSum and Breakfast Actions dataset



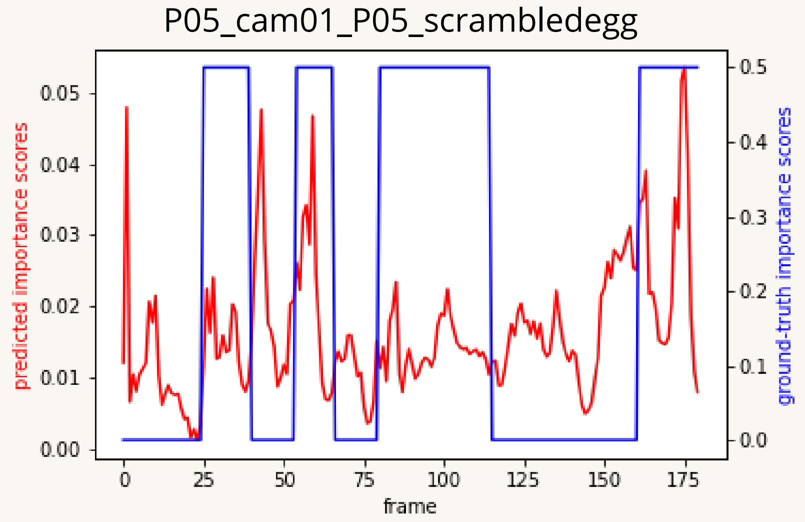
Dataset / Metric	phi	Jaccard
SumMe	0.212	0.198
TVSum	0.458	0.387
Breakfast Actions	0.297	0.371

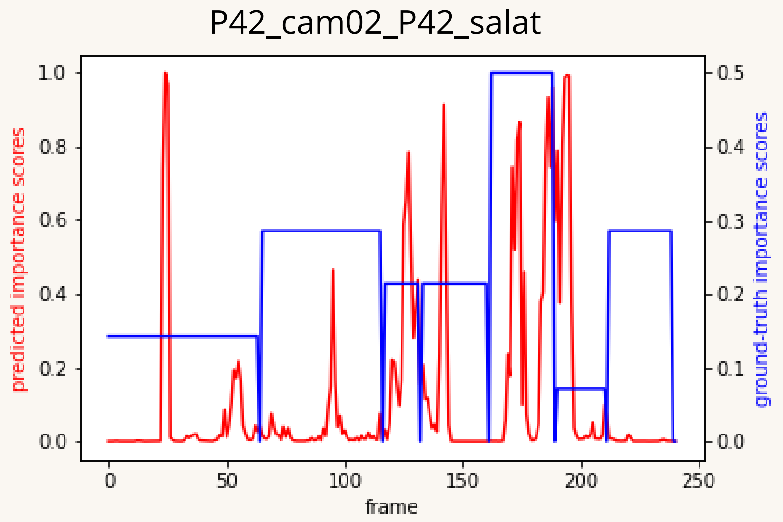
Figure 2: average level of correlation between human-generated summaries



Type	Model	F1 score	Spearman's	Kendall's
Supervised	VASNet	0.673	0.045	0.0365
	DSNet (anchor-based) [1]	0.6446	0.106	0.090
	DSNet (anchor-free) [1]	0.6003	0.078	0.056
	SUM_FCN [2]	0.314	0.032	0.024
Unsupervised	SUM_FCN_unsup [2]	0.201	-0.021	-0.02
	SUM-GAN-AAE [3]	0.5138	-0.03	-.0.03

Figure 3: The performance comparison of video summarization models when trained on the Breakfast Actions dataset





[1] J. Fajtl, H. S. Sokeh, V. Argyriou, D. Monekosso, and P. Remagnino, "Summarizing videos with attention," 2019.
[2] D. Groenewegen and O. Strafforello, "Evaluation of video summarization using dsnet and action localization datasets," 2021.
[3] P. Frolke, O. Strafforello, and S. Khademi, "Evaluation ¨ of video summarization using fully convolutional sequence networks on action localization datasets," 2021.
[4] G. Trevnenski, O. Strafforello, and S. Khademi, "Evaluation of the sum-gan-aae method for video summarization," 2021.

5 CONCLUSION

1. The Breakfast Actions dataset has better-correlated human-generated summaries than SumMe, which indicates that action localization has an impact on the level of disagreement between human annotators.
2. VASNet is able to produce summaries that are correlated with at least one of the reference summaries.
3. Supervised models appear to outperform unsupervised models on the Breakfast Actions dataset.