Evaluating the Supervised Video Summarization Model VASNet on an Action Localization Dataset

Felicia Elfrida Tjhai, Ombretta Strafforello
TU Delft

Abstract

There is growing research on automated video summarization following the rise of video content. However, the subjectivity of the task itself is still an issue to address. This subjectivity stems from the fact that there can be different summaries for the same video depending on which parts one considers important. Supervised models especially suffer from this problem as they need informative labels to learn from. As a result, upon evaluation, supervised models appear to perform worse than unsupervised models. This inspired our research on whether action localization can aid the video summarization process. To investigate this issue, this paper will answer the question of how well VASNet, a supervised video summarization model, can predict summaries for videos in an action localization dataset. This involves investigating whether action localization can produce well-correlated human-generated summaries and how it affects the quality of predicted summaries. Our findings reveal that there is a positive indication that action localization can aid in producing more well-correlated human summaries. In addition, we have observed that upon comparison with several video summarization models, VASNet has performed well and that in general, supervised models appear to outperform unsupervised ones when trained with an action localization dataset.

1 Introduction

The internet has a large repository of video data available, varying in both content and duration. Just as books have summaries, having them for videos could prove extremely useful. For example, viewers could watch the summary of a long video and capture its main message, therefore saving time.

With the rise of video data, research has been done into automated video summarization. There are numerous video summarization approaches. Some state-of-the-art supervised models include DSNet [1] and VASNet [2]. Meanwhile, examples of unsupervised models are CSNet [3] and CycleSUM [4]. These models operate differently but are trained with the same goal of predicting good-quality video summaries. Such a summary must include important parts of the video and exclude anything repetitive or unnecessary so as to maintain the video's meaning [5].

Human annotations are a crucial part of the supervised video summarization process. The general approach is to employ humans to watch a video and provide insight on which segments should be included in the summary. This is repeated with a group of human annotators for a set of videos to build a dataset. When supervised models are trained with this dataset, the human-generated summaries are used as reference and they should learn to build good-quality summaries.

Unsupervised models, on the other hand, do this differently. According to Apostolidis et al. [6], they typically use Generative Adversarial Networks (GANs). They describe the process as producing a summary based on approximated importance scores that will be used to rebuild the video. This video is then compared against the original video and their similarity is measured. Apostolidis et al. explain that ultimately, unsupervised methods aim to predict summaries that can be well rebuilt into their original videos.

Zhang et al. [7] believe that supervised models should produce better summaries as they have the advantage of being trained with human annotations. However, this is not reflected in the state-of-the-art models. For instance, evaluation on the SumMe dataset [8], unsupervised model CSNet [3] achieved a higher F1 score than supervised model VASNet [2]. This suggests the inefficacy of the labels, such that they introduce noise and inadequate useful information to learn from.

An explanation for that is the subjective nature of video summarization. When it comes to human annotations, a variety of summaries could be produced for the same video [6][9]. This is because naturally, individuals differ in which segments they consider interesting enough to be included in a summary. Subjectivity poses a major problem for supervised video summarization in particular because they rely on these human annotations.

In fact, Apostolidis et al. [6] believe that this is one of the main reasons unsupervised techniques are superior. However, we believe that the benefits of supervised methods outweigh their weaknesses. Indeed, according to Gong et al [10, p. 2], video summarization should be approached in a supervised way because "the success of a summary ultimately depends on human perception". Therefore, it is crucial to investigate

how to minimize subjectivity to allow for improved supervised video summarization.

To that end, this research aims to test the effect of action localization on the task of video summarization. In this context, action localization means that each action is confined to its own segment. Elfeki and Borji[11, p. 2] have noticed that human annotators have the tendency to choose segments that contain "deliberate action" as they capture key moments of a video and better convey the narrative. Therefore, since it can be expected that humans can see a video frame and agree whether it contains action, this could form a far more objective annotation process. In turn, this could lower the level of disagreement between human annotators, therefore creating well-coordinated reference summaries for supervised models to learn from.

This paper will answer the question "How well can a supervised model (VASNet) trained with ground-truth importance scores based on action localization learn representations for video summarization?". To that end, some subquestions include (1) "How does VASNet's performance on the Breakfast actions dataset[12] compare to those on the SumMe[8] and TVSum[13]?", (2) "To what extent do the human-generated summaries for the Breakfast actions dataset correlate with each other in comparison to SumMe and TVSum?", (3) "To what extent do VASNet-generated summaries correlate with their reference summaries?", (4) "How does the correlation between human summaries in the Breakfast Actions dataset affect VASNet-generated summaries?" and (5) "How does VASNet compare to other video summarization models?".

Our main findings reveal that action localization appears to generally have a positive effect on the level of correlation of human summaries. When trained with an action localization dataset, VASNet has performed well in comparison to other video summarization models. Lastly, our research reveals that supervised models appear to have benefited from action localization.

2 Background Information

This section provides background information on what VASNet is and how it works. It also explains two different evaluation methods for video summarization.

2.1 VASNet

In the paper *Summarizing Videos with Attention*, Fajtl et al. [2] proposed a supervised video summarization model called VASNet, which employs soft, self-attention. They describe attention as a notion pioneered by Bahdanau et al.[14], and that it refers to a neural network's ability to "learn how important various samples in a sequence, or image regions, are with respect to the desired output state" [2, p. 4]. The paper continues by explaining that soft attention means that this importance is represented by attention weights that are probabilistic in nature. This is in contrast to the binary values generated by hard attention.

Figure 1 illustrates VASNet's architecture. The paper describes the input as a sequence of D-dimensional feature vectors. In the attention network, the model calculates the attention vector e_t and using softmax, transforms it into attention

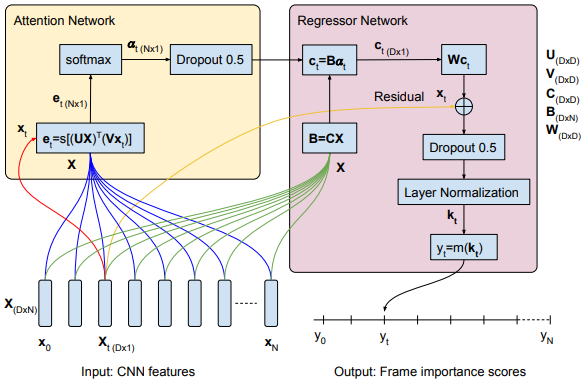


Figure 1: Overview of the VASNet network [2]

weights α_t. The paper describes α_t as "true probabilities representing the importance of input features with respect to the desired frame level score at the time t" [2, p. 6]. Meanwhile, in the regressor network, the input vector is mapped with a matrix C and then α_t, after which it is averaged. This results in c_t, a context vector that is fed into a network to be finally normalized into importance scores.

2.2 Evaluation of video summarization methods

This subsection briefly describes how F1 scores are calculated and used to evaluate video summaries. In addition, we explain the criticism of F1 scores and explore a different evaluation approach.

F1 score

Video summarization models usually use the F1 score to evaluate their machine-generated summaries. It measures the extent to which a predicted summary aligns with its corresponding reference summaries. The F1 score is calculated with the formula

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}. \quad (1)$$

To evaluate summaries, this score is calculated between the predicted summary and each reference summary. Either the average or the maximum is then taken to represent the accuracy of the predicted summary.

Rank-order correlation

Although F1 score has been widely used, it is far from perfect. Otani et al. [9] have investigated its credibility by means of a randomization test. This involved evaluating randomly generated summaries using F1 scores. Because they were random, it was expected that they did not align with the reference summaries. However, Otani et al. observed that these summaries actually scored well. Furthermore, they noticed that the F1 score depends significantly on how the videos were segmented instead of their importance scores. These findings raise doubts on the reliability of F1 scores in providing insight into the quality of predicted video summaries.

As a solution, Otani et al. proposed a new evaluation metric that uses rank-order correlation. Correlation scores are measured using Spearman's ρ correlation coefficient[15] and Kendall's τ rank correlation coefficient [16]. Unlike with F1 scores, Otani et al. report that randomly generated summaries did not score well with this metric. This is a positive indication that importance scores were actually taken into account and more importantly that it can better reveal the true quality of video summaries.

3 Methodology

This research attempts to investigate how well VASNet, a supervised video summarization model, can predict summaries for videos annotated with ground truth importance scores based on action localization. This was done using videos from the Breakfast Actions dataset [12], which the author claims to be among the most extensive datasets availables. This dataset [12, p. 782] claims to be able to "closely reflect real-world conditions" with "unscripted, unrehearsed and undirected" execution of everyday tasks.

The methodological approach of this research consisted of four stages: data collection, training, evaluation, and comparison. Each of these stages is described below.

3.1 Data collection

Prior to training, human annotations were crowdsourced for the videos in the Breakfast Actions dataset. Our research involved participating in this task. This involved watching a number of videos and picking a small number of frames that were deemed to be most representative. The intention was to obtain a deeper understanding of video summarization and the difficulty of creating human-generated summaries for videos.

The human-generated summaries obtained were used to build the dataset used in this research. GoogleNet features were extracted for each video and packed into a .hdf5 file.

3.2 Training stage

In the training stage, VASNet was trained on three datasets: SumMe, TVSum, and the Breakfast Actions dataset. Table 1 shows information about each dataset and the environment they were trained in.

For each video, the predicted summary with the highest F1 score across all epochs was chosen. We saved the predicted importance scores, predicted summary, and F1 score into a new h5 file. This file was then used in the evaluation stage.

3.3 Evaluation stage

In the evaluation stage, we evaluated the results from training using two evaluation metrics. Firstly, we used F1 scores. Since it is a standard evaluation metric for video summarization, it was necessary to include it to allow comparison with other models and datasets.

This F1 score was calculated between a predicted summary and each of its reference summaries. Fajtl et al. [2]'s evaluation procedure for TVSum was to take the average. Meanwhile, they took the maximum for SumMe. In the paper, they described these methods as the ones recommended by the creators of those datasets. Therefore, in this research we will do the same. As for the Breakfast Actions dataset, we decided to take the maximum. This is because knowing the extent to which human-generated summaries can vary, we believe that a model should be rewarded for being able to generate a summary that is similar to at least one human-generated summary.

Due to the criticism towards F1 scores, we also used rank-order correlation metric. This was done using the code provided by Otani et al. [9]. Their method involved using both Spearman's ρ and Kendall's τ. However, the Breakfast Actions dataset does not include reference importance scores but instead reference summaries with boolean values. This means that Spearman's ρ and Kendall's τ would not be suitable to measure the correlation of summaries. To compensate, the Matthews Correlation Coefficient (MCC) or phi coefficient [17], along with the Jaccard Similarity Coefficient [18] were used in their place. These two metrics were designed for binary data therefore they are more appropriate. Although Spearman's ρ and Kendall's τ could not be used for human-generated summaries, they could still be used to measure the correlation between the predicted importance scores of each video and their corresponding ground-truth importance scores.

For each video in the Breakfast Actions dataset, we measured correlation from multiples angles. First, we calculated the correlation between reference summaries. This was done to see how closely correlated human-generated summaries are for different videos. Next to that, we measured the correlation between each predicted summary and its corresponding reference summaries and took the maximum, as was done with the F1 scores. The higher this value, the better quality that machine summary can be considered to be. This allowed us to evaluate VASNet's performance on the Breakfast Actions dataset.

3.4 Comparison stage

The comparison stage involved comparing the performance of different models on different datasets. Firstly, we compared VASNet's performance across the three datasets. This was done to see the effect of action localization on the quality of the predicted summaries.

The average human correlation coefficients for three datasets were compared. This reveals the extent to which action localization affects the agreement between human annotators. If the Breakfast Actions dataset has a higher correlation value between its human summaries compared to the other datasets, it can be inferred that action localization creates more coordinated human summaries.

The final step was to see how VASNet compares to other supervised and unsupervised models when trained with an action localization dataset. This was crucial because one of the main inspirations for this research was the unexpectedly poorer performance of supervised models in some cases compared to unsupervised models. Therefore, this comparison could reveal whether action localization introduces useful information for supervised video summarization models to learn from.

Dataset	No. of videos	No.of reference summaries per video	Each fold	k	epochs
SumMe	25	15-18	22-23 training,2-3 test	10	50
TVSum	50	20	45 training, 5 test	10	50
Breakfast Actions	21	2-15	18-19 training, 2-3 test	10	50

Table 1: The training environment for each dataset

4 Results

In this section, we present the results of our experiments.

4.1 Average results for each dataset

Table 2 illustrates the average F1 scores and correlation values for three different datasets: SumMe, TVSum, and Breakfast Actions. Among the three datasets, VASNet achieved the highest F1 score on the Breakfast Actions dataset. It has also achieved the highest phi and Jaccard values when trained with this dataset. In terms of Spearman's ρ and Kendall's τ however, there is a significantly large gap between TVSum and the other two datasets.

4.2 Correlation of the human-generated summaries in the Breakfast Actions dataset

Table 3 displays for each dataset, how well their human-generated summaries correlate with each other on average. It can be seen that TVSum's human summaries are the most correlated and similar, while the Breakfast Actions dataset is still behind.

Table 4 presents how well the reference summaries for each Breakfast Actions video correlate with each other. It can be observed that between several videos, these values differ significantly. For instance, the human summaries for the P48_cam02_P48_milk video are very well-correlated while the opposite can be said about the P05_cam01_P05_scrambledegg video. In fact, according to the Jaccard Similarity coefficient, the former's human summaries are identical while the latter's are completely different.

4.3 Correlation of VASNet's predicted summaries

Table 5 presents the level of correlation between the predicted summary of each video and their reference counterparts, along with the obtained F1 scores. Based on the high F1 scores along with the maximum phi and Jaccard values, VASNet has successfully generated summaries that are similar to at least one of the human-generated summaries.

The video that VASNet has particularly excelled in is the P42_cam02_P42_salat video, in which it has predicted a summary that is almost identical to one of the human-generated summaries. At the other extreme is the P05_cam01_P05_scrambledegg video, in which VASNet has performed the worst.

4.4 Comparison of video summarization models

Table 6 displays the comparison between the performance of 5 video summarization models when trained using the Breakfast Actions dataset. In terms of F1 scores, VASNet outperforms the others. However, in terms of correlation, both versions of the DSNet model are leading. Even so, VASNet still places second.

5 Discussion

Section 4.1 shows that VASNet performed better on the Breakfast Actions dataset than it did on the other two datasets. This means that VASNet was able to learn more from the annotations in the Breakfast Actions dataset.

We consider the higher performance on the Breakfast Actions dataset compared to SumMe to be especially telling. This is because the Breakfast Actions dataset can be considered more "similar" to SumMe than TVSum. Firstly, the Breakfast Actions dataset has 21 videos and SumMe has 25, in comparison to TVSum which has 50 videos. In addition, both the Breakfast Actions and SumMe datasets were annotated with boolean values. The ground-truth importance scores were calculated as the average of these annotations. On the other hand, TVSum was annotated with actual importance scores. This is perhaps an explanation for why the level of correlation between predicted and ground-truth importance scores are significantly higher for TVSum than the other two datasets.

Section 4.2 shows that although TVSum is leading, the human-generated summaries in the Breakfast Actions dataset are on average more well-correlated than the those of SumMe dataset. This means that action localization has to some extent promoted more agreement between human annotators. Furthermore, this effect seems to be more apparent in some videos than others. This can be seen from Table 4. It reveals that while some the correlation value between human-generated summaries are high for some videos, it is low for others, which is why the standard deviation reported in Table 3 is high. This suggests that some videos simply induce more disagreement between human annotators than others. This seems to be the case for videos with more complicated actions, such as making a sandwich, in comparison to pouring milk.

Section 4.3 provides a detailed account of how well VASNet has predicted summaries for each video in the Breakfast Actions dataset. The numbers suggest that VASNet is generally able to predict summaries that are similar to at least one of their corresponding reference summaries.

In Section 4.3, we contrasted the results of two videos: P05_cam01_P05_scrambledegg and P42_cam02_P42_salat. As reported in Table 5, VASNet produced a significantly worse summary for the former than the latter. A possible explanation for this is the fact that the P05_cam01_P05_scrambledegg has negative phi correlation and zero Jaccard similarity between its human-generated summaries, as presented in Table 4. This is reflected in Fig-

Dataset	F1 score		Spearman's ρ		Kendall's τ		phi		Jaccard	
	mean	std	mean	std	mean	std	mean	std	mean	std
SumMe	0.511	0.138	0.032	0.223	0.025	0.164	0.448	0.162	0.354	0.127
TVSum	0.606	0.085	0.438	0.232	0.306	0.168	0.537	0.100	0.453	0.094
Breakfast Actions	0.673	0.181	0.045	0.224	0.0365	0.171	0.635	0.214	0.536	0.218

Table 2: The average results for each dataset, rounded off to 3 decimal places. 'F1 score' refers to the average F1 score for all videos in the dataset. 'Spearman's ρ' and 'Kendall's τ' refer to the average correlation between the predicted importance scores and their corresponding ground truth importance scores. 'Phi' and 'Jaccard' indicate respectively, the correlation and similarity between the predicted summaries and their human-generated counterparts. Aside from the mean, standard deviations are also reported under the 'std' columns.

Dataset	phi		Jaccard	
	mean	std	mean	std
SumMe	0.212	0.107	0.198	0.071
TVSum	0.458	0.096	0.387	0.079
Breakfast Actions	0.297	0.283	0.371	0.228

Table 3: The average correlation between human summaries for each dataset, rounded off to 3 decimal places. Standard deviations are also given under the 'std' column

Video name	phi	Jaccard
P03_webcam02_P03_friedegg	0.095	0.214
P03_webcam02_P03_sandwich	0.069	0.155
P05_cam01_P05_coffee	0.092	0.167
P05_cam01_P05_scrambledegg	-0.287	0.0
P07_cam01_P07_scrambledegg	0.155	0.225
P09_cam01_P09_scrambledegg	0.133	0.266
P10_webcam01_P10_coffee	0.441	0.438
P12_cam01_P12_sandwich	0.699	0.660
P25_cam01_P25_cereals	0.601	0.615
P29_cam01_P29_juice	0.220	0.309
P37_cam01_P37_sandwich	0.317	0.368
P38_cam01_P38_scrambledegg	0.182	0.333
P39_webcam02_P39_sandwich	0.0213	0.119
P40_cam02_P40_milk	0.152	0.273
P42_cam02_P42_salat	0.058	0.216
P46_cam01_P46_tea	0.311	0.321
P47_webcam02_P47_juice	0.361	0.5
P48_cam01_P48_scrambledegg	0.413	0.353
P48_cam02_P48_milk	1.0	1.0
P51_cam01_P51_juice	0.683	0.704
P51_webcam01_P51_cereals	0.513	0.558

Table 4: Correlation between the human-generated summaries of every Breakfast Actions video, rounded off to 3 decimal places.

ure 2a where the ground-truth graph has four segments with equal height. This potentially caused confusion in the model when learning which frames should be considered important. As a result, it was not able to optimally produce a summary. In contrast, the P42_cam02_P42_salat video has different importance scores between segments, as shown in Figure 2b.

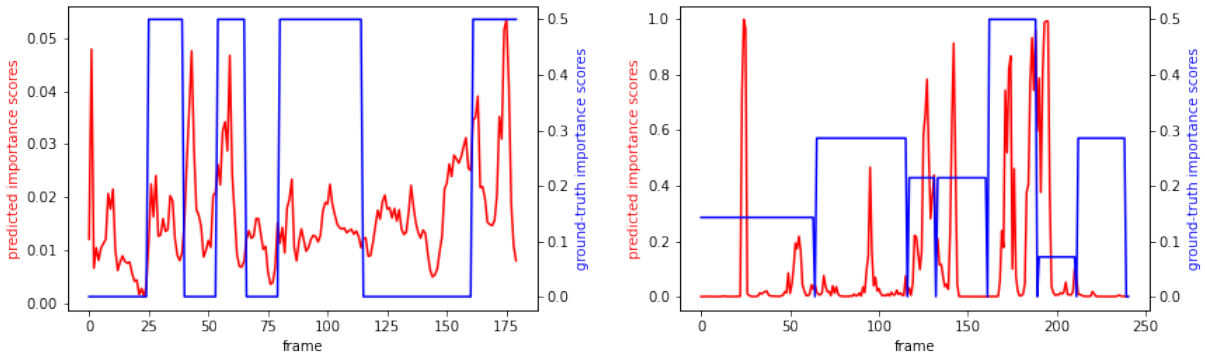
Analyzing Table 4 and Table 5 also suggests that although VASNet can generally handle poorly correlated human summaries, its learning ability can sometimes be significantly affected by it. An example of such a case can be seen in these three videos: P03_webcam02_P03_friedegg, P05_cam01_P05_coffee and P05_cam01_P05_scrambledegg.

The network prediction for the P48_cam01_P48_scrambledegg video also presents an interesting case. Figure 3d shows that the network importance scores are very well-correlated with the ground-truth. According to Figure 3a, Figure 3b, and Figure 3c, VASNet's predicted summary contains one of the two segments in the reference summaries. This can be explained by the fact that VASNet maximum summary size is 15% of the original video size, which means it cannot cover all the segments in the reference summaries in this case.

Section 4.4 compares the performance of several supervised and unsupervised video summarization methods. In comparison to these models, we consider VASNet to have performed quite well. It places second in terms of how much its predicted importance scores correlate with the ground-truth. Analyzing the numbers however, it appears that for all these models, their predicted importance scores are very weakly correlated to the ground-truth.

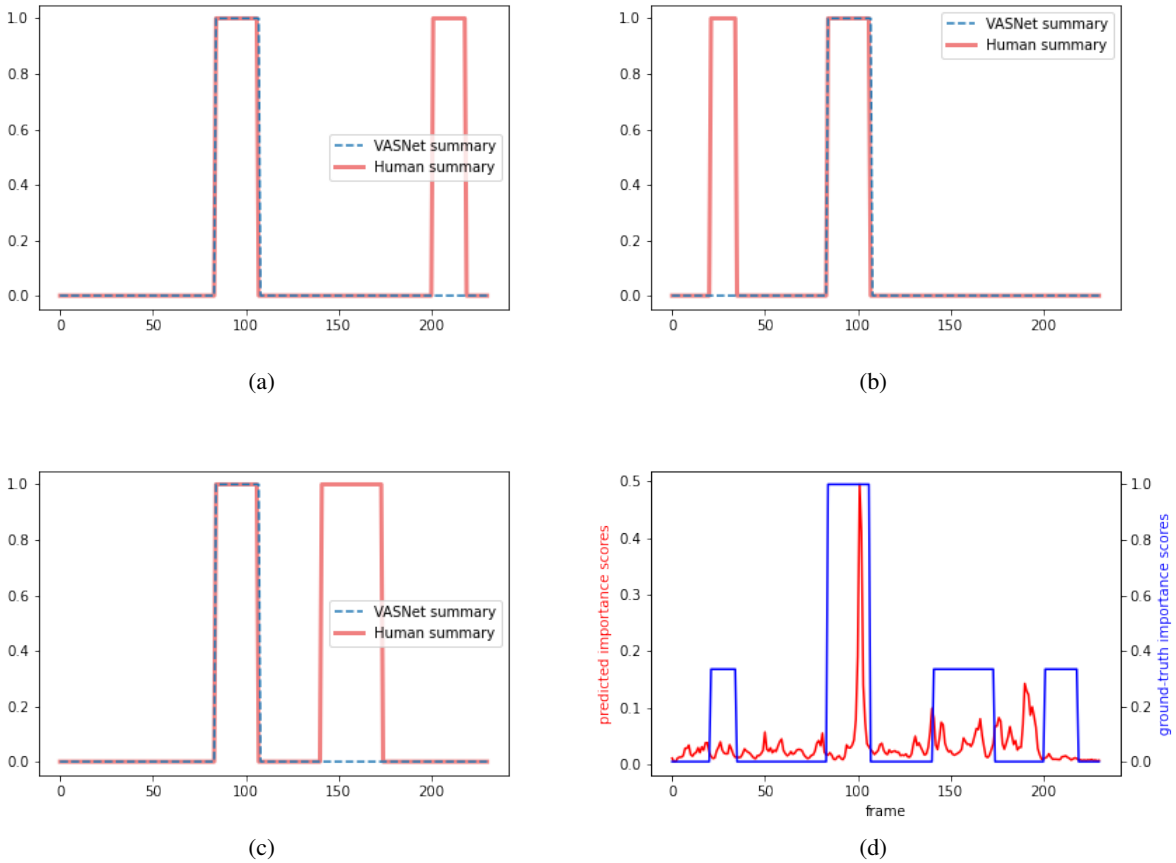
Another interesting observation is that the supervised models generally outperform the unsupervised models. In fact, both unsupervised models in the table generated importance scores that have negative correlation with their ground truth counterparts. This suggests that action localization has provided useful ground-truth information that supervised models can learn from.

This research has a number of limitations. Firstly, the Breakfast Actions dataset can be considered incomplete. For some of the videos, there are very few reference summaries. In addition, the length of the human-generated summaries is not fixed for all videos, unlike SumMe and TVSum which have reference summaries roughly 15% the size of the original videos. This means we cannot tailor the maximum summary length VASNet should produce for the Breakfast Ac-



(a) (b)

Figure 2: Comparison of VASNet's predicted scores and the ground-truth importance scores for (a) P05_cam01_P05_scrambledegg video, and (b) P42_cam02_P42_salat video.



(a) (b)

(c) (d)

Figure 3: Visualization of VASNet's predictions for the P48_cam01_P48_scrambledegg video. (a)(b)(c) visualize the overlap between VASNet's summary and each of its reference summaries, while (d) visualizes the correlation between network and ground-truth importance scores.

Video name	phi		Jaccard		F1 score
	mean	max	mean	max	
P03_webcam02_P03_friedegg	0.070	0.384	0.134	0.305	0.468
P03_webcam02_P03_sandwich	-0.024	0.694	0.073	0.55	0.71
P05_cam01_P05_coffee	0.227	0.308	0.221	0.263	0.417
P05_cam01_P05_scrambledegg	-0.005	0.187	0.098	0.197	0.329
P07_cam01_P07_scrambledegg	-0.095	0.377	0.047	0.303	0.465
P09_cam01_P09_scrambledegg	0.247	0.531	0.222	0.381	0.552
P10_webcam01_P10_coffee	0.287	0.892	0.275	0.818	0.900
P12_cam01_P12_sandwich	0.394	0.399	0.317	0.32	0.485
P25_cam01_P25_cereals	-0.068	0.666	0.075	0.545	0.706
P29_cam01_P29_juice	0.005	0.614	0.12	0.48	0.649
P37_cam01_P37_sandwich	0.123	0.650	0.153	0.5	0.667
P38_cam01_P38_scrambledegg	0.164	0.668	0.242	0.485	0.653
P39_webcam02_P39_sandwich	0.304	0.920	0.333	0.867	0.929
P40_cam02_P40_milk	0.115	0.915	0.196	0.857	0.923
P42_cam02_P42_salat	0.291	0.980	0.369	0.964	0.982
P46_cam01_P46_tea	0.192	0.930	0.224	0.875	0.933
P47_webcam02_P47_juice	0.006	0.642	0.125	0.5	0.667
P48_cam01_P48_scrambledegg	0.669	0.741	0.519	0.605	0.754
P48_cam02_P48_milk	0.589	0.589	0.455	0.455	0.625
P51_cam01_P51_juice	0.572	0.696	0.495	0.579	0.733
P51_webcam01_P51_cereals	0.346	0.543	0.303	0.417	0.588

Table 5: Correlation values and F1 scores for VASNet's predicted summaries for each Breakfast Actions video, rounded off to 3 decimal places.

Type	Model	F1 score	Spearman's ρ	Kendall's τ
Supervised	VASNet	0.673	0.045	0.0365
	DSNet (Anchor-based)[19]	0.6446	0.106	0.090
	DSNet (Anchor-free)[19]	0.6003	0.078	0.056
	SUM_FCN[20]	0.314	0.032	0.024
Unsupervised	SUM_FCN$_{unsup}$[20]	0.201	-0.021	-0.020
	SUM-GAN-AAE[21]	0.5138	-0.03	-0.03

Table 6: Performance of different video summarization models on the Breakfast Actions dataset. For each of the models, their F1 scores and correlation values between predicted importance scores and their corresponding ground truth importance scores are given.

tions dataset.

Secondly, the Breakfast Actions dataset only has human-generated summaries and not human-generated importance scores. This means that VASNet's predicted importance scores could not be compared with their true human-generated counterparts, but only the average of the human-generated summaries. Thus, this was compensated by measuring the correlation and similarity between predicted summaries and human-generated summaries.

Lastly, the comparison between algorithms could only be done between 4 supervised and 2 unsupervised models. Having an equal number of supervised and unsupervised models would create a fairer comparison.

6 Responsible Research

This section discusses the ways in which research integrity was maintained, along with the reproducibility of the methods used and the results obtained.

6.1 Research Integrity

This research contributes to the advancement of video summarization. With the massive amount of video data on the internet, the ability to automate the summarization of videos could be useful in numerous ways. For example, it allows the retrieval of information in less time, since it is no longer necessary to watch videos in their entirety. As a result, it enhances efficiency. We do not believe there are any major negative ethical aspects associated with video summarization or our research specifically.

To facilitate our research, a crowdsourcing survey was conducted. However, the formulation and distribution of this survey were not part of our responsibility. Nonetheless, it did not collect any personal data. This survey was necessary to obtain human insight on the summarization of videos from the Breakfast Actions dataset. It was then used for training. Apart from this, there was no interaction with human subjects.

6.2 Research Reproducibility

We conducted this research on an existing open-source video summarization model. The code for this model and the correlation evaluation metric can easily be accessed online and their inner workings are described in their corresponding papers. Likewise, the SumMe and TVSum datasets are available online and can be downloaded. As for the Breakfast Actions dataset, their reference summaries were obtained through the aforementioned survey. The videos are freely available online, while the annotations may be published in the future. In addition, this research's methodology is accompanied by clear instructions, as provided in Section 3. It describes what is necessary to conduct the experiment and the right training environment. Given the steps are carefully followed, reproducing the results is straightforward.

7 Conclusions and Future Work

The aim of this research was to investigate how well VASNet, a supervised video summarization model, trained with ground-truth importance scores based on action localization can learn representations for video summarization. This involves answering 5 subquestions: (1) the quality of VASNet's predicted summaries for the Breakfast Actions dataset compared to those for SumMe and TVSum dataset, (2) the level of correlation between the human-generated summaries in the Breakfast Actions dataset compared to those of the SumMe and TVSum dataset, (3) the level of correlation between the predicted summaries for the Breakfast Actions dataset and their corresponding reference summaries, (4) the relationship between the level of correlation between human-generated summaries and the quality of the predicted summaries, and lastly (5) how well VASNet performs in comparison to other video summarization models.

Answering the first subquestion involved comparing F1 scores and level of correlation and similarity between predicted summaries and their reference summaries. The results revealed on the Breakfast Actions dataset, VASNet's predicted summaries are the most correlated to their reference counterparts. This means VASNet was able to learn well from the action localization Breakfast Actions dataset.

To answer the second subquestion, we calculated the average phi and Jaccard coefficient values with regard to human-generated summaries for each of the datasets. The numbers reveal that the level of correlation between the human-generated summaries of the Breakfast Actions dataset is higher than that of SumMe, but lower than that of TVSum. This suggests that there is a positive indication that action localization has the potential to reduce disagreement between human annotators, therefore limiting the level of subjectivity in the annotation process. However, there still seems to be high disagreement when the video involves more complicated actions.

The third subquestion was investigated by analyzing the quality of the summaries predicted by VASNet for the videos in the Breakfast Actions dataset. Our findings reveal that for most of the videos, VASNet was able to produce summaries that align with at least one of the human-generated summaries.

To answer the fourth subquestion, for each video, the correlation values for human-generated summaries and predicted summaries were analyzed. In general, we believe that the higher level of human summary correlation in the Breakfast Actions dataset contributed to VASNet's ability to learn well. However, it appears that VASNet's learning process can sometimes be stunted by weak human summary correlation.

Lastly, the fifth subquestion was answered by studying the average F1 scores and correlation values of 5 video summarization models. We observed that among the models, VASNet placed second. Although all the models' importance scores predictions show very weak correlation with their ground-truth, the supervised models appear to be performing better than the unsupervised models. This shows promise on how action localization can aid supervised video summarization.

Based on the answers to these five subquestions, it can be concluded that action localization appears to have a positive effect on the task of video summarization. It has enabled the creation of better-correlated human summaries for some videos. Lastly, VASNet was able to perform well when trained with an action localization dataset.

There are a number of areas to explore in future research. Firstly, we believe it is necessary to extend the Breakfast Actions dataset to include more human annotations based on action localization. A considerable limitation to our research was inadequate human-generated summaries for some videos. Therefore, research in this area would greatly benefit from a more mature dataset. In addition, it would be interesting to also evaluate video summarization models on more action localization datasets. One example is the MultiTHUMOS dataset[22]. We also recommend evaluating more video summarization models on these datasets. We believe it would provide a more concrete indication of the effect of action localization.

References

[1] W. Zhu, J. Lu, J. Li, and J. Zhou, "Dsnet: A flexible detect-to-summarize network for video summarization," *IEEE Transactions on Image Processing*, vol. 30, pp. 948–962, 2021.

[2] J. Fajtl, H. S. Sokeh, V. Argyriou, D. Monekosso, and P. Remagnino, "Summarizing videos with attention," 2019.

[3] Y. Jung, D. Cho, D. Kim, S. Woo, and I. S. Kweon, "Discriminative feature learning for unsupervised video summarization," 2018.

[4] L. Yuan, F. E. Tay, P. Li, L. Zhou, and J. Feng, "Cycle-sum: Cycle-consistent adversarial lstm networks for unsupervised video summarization," 2019.

[5] S. Jadon and M. Jasim, "Unsupervised video summarization framework using keyframe extraction and video skimming," *2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA)*, Oct 2020. [Online]. Available: http://dx.doi.org/10.1109/ICCCA49541.2020.9250764

[6] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, "Video summarization using deep neural networks: A survey," 2021.

[7] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," 2016.

[8] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "Creating summaries from user videos," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 505–520.

[9] M. Otani, Y. Nakashima, E. Rahtu, and J. Heikkilä, "Rethinking the evaluation of video summaries," 2019.

[10] B. Gong, W.-L. Chao, K. Grauman, and F. Sha, "Diverse sequential subset selection for supervised video summarization," *Advances in Neural Information Processing Systems*, vol. 3, pp. 2069–2077, 01 2014.

[11] M. Elfeki and A. Borji, "Video summarization via actionness ranking," 2019.

[12] H. Kuehne, A. Arslan, and T. Serre, "The language of actions: Recovering the syntax and semantics of goal-directed human activities," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 780–787.

[13] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, "Tvsum: Summarizing web videos using titles," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5179–5187.

[14] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2016.

[15] D. Zwillinger and S. Kokoska, "Crc standard probability and statistics tables and formulae," 01 2000.

[16] M. G. KENDALL, "THE TREATMENT OF TIES IN RANKING PROBLEMS," *Biometrika*, vol. 33, no. 3, pp. 239–251, 11 1945. [Online]. Available: https://doi.org/10.1093/biomet/33.3.239

[17] B. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," *Biochimica et Biophysica Acta (BBA) - Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0005279575901099

[18] P. Jaccard, "The distribution of the flora in the alpine zone.1," *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912. [Online]. Available: https://nph.onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8137.1912.tb05611.x

[19] D. Groenewegen and O. Strafforello, "Evaluation of video summarization using dsnet and action localization datasets," 2021.

[20] P. Frölke, O. Strafforello, and S. Khademi, "Evaluation of video summarization using fully convolutional sequence networks on action localization datasets," 2021.

[21] G. Trevnenski, O. Strafforello, and S. Khademi, "Evaluation of the sum-gan-aae method for video summarization," 2021.

[22] S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori, and L. Fei-Fei, "Every moment counts: Dense detailed labeling of actions in complex videos," *International Journal of Computer Vision*, 2017.