Evaluation of the SUM-GAN-AAE method for Video Summarization

Georgi Trevnenski, Ombretta Strafforello, Dr. Seyran Khademi

TU Delft

June 27, 2021

1 Abstract

Video summarization is a task which many researchers have tried to automate with deep learning methods. One of these methods is the SUM-GAN-AAE algorithm developed by Apostolidis et al. [2] which is an unsupervised machine learning method evaluated in this study. The research aims at testing the algorithm's performance on the Breakfast dataset, which is an action localization dataset, and evaluate it with rank correlation coefficients. Parameter optimization was performed to tune the learning rate of the system according to the Breakfast dataset. Then, by using k-fold cross-validation, three metrics were used to evaluate the trained model - F-Score, Kendall's τ and Spearman's ρ. Analysis of the results indicates a high F-Score as reported by the SUM-GAN-AAE paper but low rank correlation coefficients. Moreover, plotting importance scores per frame demonstrates the algorithm's inability to select key frames. The findings suggest that F-Score is not a fitting metric to use in the context of video summarization and the SUM-GAN-AAE algorithm performs poorly not only on action localization datasets but also on video summarization ones such as SumMe [7].

2 Introduction

The amount of information available nowadays is enormous and one of the great challenges for the data science community is to process it and makes sense of this data. A direct consequence of this information explosion is a large number of videos available to us, shared publicly or privately. Thus, arises the problem of processing video data and the need for producing summaries containing only the gist of longer videos. There are several issues with automatic summarization. One is the fact that a single video can be summarized in many ways which can be of arguably the same quality in terms of how important is the retained information. As noted by [13], this problem is observable also with human-composed summaries as it is an opinionated task to perform in the first place. Another issue to consider when creating a summary is the variety of videos one might need a summary on. Videos vary in terms of the quality of the image but more importantly, the content and purpose of videos can be fundamentally different.

The aforementioned task of automatic summary creation is suitable to be tackled with a deep learning model. To use a machine learning model to perform the task, the video is split into segments and every segment is given an importance score, the segments with the highest importance scores form the summary. Many machine learning methods have been used to train a model to produce video summarizations. Some of them are supervised, using human annotations about importance scores, others - unsupervised, therefore, removing noise caused by annotators' opinions in training. The problem of evaluating generated summaries, though, persists in unsupervised methods as well. The general expectation is for supervised algorithms to perform better than unsupervised ones. However, this is not always the case in the field of video summarization. According to results from two papers, one describing a supervised method [4] and one proposing an unsupervised approach [8], the unsupervised method performed better on the SumMe dataset (Gygli et al. 2014)[7]. Therefore, it might turn out that unsupervised approaches not only require less data but also tackles the problem of non-consistent user summaries. Not enough work has been put into comparing unsupervised deep learning methods with supervised ones on different datasets, yet. In addition, an idea which is still not tested is whether algorithms extracting segments that contain actions can create good summaries.

The main goal of the research that is about to be conducted is to determine whether the unsupervised methods generalize well to different datasets by using the implementation of an unsupervised deep learning algorithm to train a neural network on the Breakfast [11] action dataset. This is an action localization dataset meant for training machine learning algorithms to differentiate actions in a video. Therefore, it is reasonable to believe that an unsupervised algorithm could be trained to recognize actions that usually contain important information and combine them into a summary. Furthermore, it is interesting to investigate how the problem of subjectivity translates to action localization datasets and the way supervised and unsupervised machine learning methods tackle it.

3 Background

The deep learning method which is going to be investigated in detail in this paper is known as SUM-GAN-AAE (Apostolidis et al. 2020) [2]. This algorithm trains a frame selector neural network utilizing a system of autoencoder and generative adversarial network. It calculates a loss function, which serves to adjust the parameters of the network, without using human-annotated videos for training. This is done by reconstructing the original video from the produced summary and using the GAN to compare it with the original one. The approach relies on the notion that the reconstructed video is closer to the original one when the summary captures the most important information.

The SUM-GAN-AAE method is based on the SUM-GAN model (Mahasseni, Lam and Todorovic 2017) [12]. From this starting point, the model was improved by Apostolidis et al. once in 2019 to a version known as SUM-GAN-sl [1] which had fewer parameters to learn and reduced size of the input feature vectors. A year later the SUM-GAN-AAE solution is proposed by Apostolidis et al. which incorporates an attention layer.

4 Methodology

In order to test the capabilities of the SUM-GAN-AAE algorithm, a few methodologies are used which are systematized below.

Firstly, the machine learning algorithm will be trained and tested on the TVSum [14] and SumMe [7] datasets which are the two most widely used benchmarks for video summarization methods. There is already information about the performance on these datasets and the results obtained by Apostolidis et al. are documented in their paper. The results produced with these datasets in the current research are compared with the results obtained in the aforementioned document in order to test their reproducibility and have a starting point for further comparisons with other datasets.

Next, the research focuses on the Breakfast dataset and how well the algorithm performs when applied to it. F-Score is used as the most common metric for comparing generated summaries to ground truth. Although there are doubts as to whether this is the most accurate measure to evaluate summaries, the F-Score results of the generated summaries from the Breakfast dataset are still presented for a more in-depth analysis.

As pointed out by Otani et al.[13], rank correlation coefficients give a better insight into how a system-generated summary compares to a random one or a human-made one. Thus, both Kendall's τ and Spearman's ρ coefficients for the summaries generated with the SUM-GAN-AAE algorithm are presented in the following section. Furthermore, a visualization is shown in the form of correlation curves similar to the aforementioned paper [13]. Section 5.2 elaborates further on the metrics used to evaluate the algorithm's performance.

Finally, the research touches upon parameter optimization to document the influence of certain hyperparameters on the performance of the algorithm when using the Breakfast action localization dataset.

5 Experimental setup and results

5.1 Datasets

As mentioned above, in the process of evaluating the SUM-GAN-AAE algorithm in this research paper three datasets have been used - the SumMe [7], TVSum [14] and Breakfast [11] datasets. Each dataset consists of a number of videos that are annotated by participants in a study. This is done by segmenting the video in a sequence of short clips and then a summary is formed by assigning boolean values to each of those clips. The users providing the annotations set a boolean value of 1 to the important pieces of footage that will form the summary and 0 to the rest. According to these scores, each frame in the video is labeled and afterward, the frames are subsampled to facilitate training. In addition to those annotations, which consist of 0's and 1's, for each video, the dataset contains also a summary labeled as ground-truth, which consists of importance scores in the range [0, 1] for each subsampled frame in the video. Those importance scores are calculated by averaging out the values per frame of the user summaries.

The SumMe dataset in particular consists of 25 videos capturing different events without a specific topic that can describe all of them. The videos are of length ranging from 1 to 6 minutes and are shot from a first-person and third-person point of view. These videos are given annotations by at least 15 users each.

Similarly, TVSum is another video dataset with twice the number of videos - 50, again capturing activities in different genres. The videos are of length ranging from 1 to 5 minutes and each video in the dataset is annotated by 20 users.

The Breakfast dataset is the main focus of this study. It is different from the other two datasets as it was not built for video summarization purposes in particular. The purpose of the dataset is to facilitate the development and testing of action recognition software.

It consists of videos of people performing 10 different actions related to cooking shot from multiple angles. The dataset is significantly larger than the ones described above and is made with variability in mind. The footage is raw and not professionally recorded to mimic real-world situations.

For this research, only a small portion of the whole dataset is used and the corresponding HDF5 file is created with feature extracted by the I3D network [3]. The subset used in the training and testing in the study contains 21 videos that capture 9 different action types. For each video, up to 15 annotations are provided for training and testing the video summarization algorithm.

5.2 Evaluation metrics

The most widely used metric for evaluating automatically generated summaries in the literature about video summarization is F-Score. In this research the same method for calculating F-Score is used as in Apostolidis et al., 2020 [2]. The calculations are derived from the following formula:

$$F = 2 * \frac{P * R}{P + R} * 100, \text{ where } P = \frac{A \cap U}{\|A\|} \text{ and } R = \frac{A \cap U}{\|U\|}$$

In the above equation, A is the automatically created summary, i.e., the summary generated by the output of the machine learning system. On the other hand, U stands for user summary, i.e., a human-generated summary. P and R denote precision and recall which are metrics used to evaluate the overlap between the two summaries, and finally, $\|*\|$ represents the length of a video in this case.

Thus, to obtain the F-score of a generated summary, first, the output of the SUM-GAN-AAE algorithm is produced which is in the form of importance scores for each segment of the video. Then, an implementation of the Knapsack algorithm is used to make a selection of video fragments that maximizes the total importance score while the sum of the lengths of the selected segments does not go above 15% of the length of the original video. Afterward, the F-Score of the generated summary is calculated with every user summary corresponding to this original video and the mean of the results is taken as the F-Score describing the generated summary. Additionally, the F-Score of the generated summary can be calculated using the ground-truth summary belonging to the respective original video as described in the section Datasets (Apostolidis et al., 2020, sec. 4.2 [2]).

As an alternative metric for evaluating the performance of a summarization algorithm, rank correlation coefficients are used in this study. As suggested by Otani et al.[13] these can be a more accurate measurement of performance and the results shown in the aforementioned paper support that claim. The calculated coefficients in this study are Kendall's τ [9] and Spearman's ρ [10]. They both have been used extensively in statistical problems related to evaluating rankings.

The values presented in this study are calculated by first obtaining the predicted importance scores for every video segment. Then, a ranking is created according to those scores. A similar ranking is made from the user summaries presented as boolean values for every segment. Those rankings are at the end compared pairwise between the predicted ranking and the ranking of every reference coming from a user. Using the respective equations for Kendall's τ and Spearman's ρ a coefficient value is produced for every user summary and then those are averaged out.

In addition to the coefficient values, the method for visualizing importance score correlation presented by Otani et al.[13] is also put into use. Graphs are constructed by again

ranking the video segments by importance score produced by the SUM-GAN-AAE algorithm. Then, "we accumulate the averaged reference scores based on the ranking obtained in the second stage" (Otani et al.[13]). If the generated summary is closely correlated with the user summaries, the generated graphs have a steep slope as they increase with the number of video fragments. Conversely, the graphs of summaries that strongly divert from the user summaries have a slowly increasing graph.

5.3 Algorithm implementation

The SUM-GAN-AAE algorithm implementation provided by Apostolidis et al. [2] is used in this study. A detailed explanation of the neural network setup can be found in section 4.1 - "Implementation Details" in the aforementioned paper. Additionally, for some of the plots presented here, the implementation belonging to Otani et al. [13] was used and adjusted. The altered implementation used to produce the results presented in this research paper can be found on the GitLab repository of the research project.

5.4 Measurements and performance comparison

All results presented in the report excluding Table 3 are produced using the latest version of the Breakfast dataset composed for the purposes of this research. The videos in this dataset are listed in Table 5. For Table 3 a subset of this dataset is used which is smaller in size and with fewer videos. More information can be found in the paragraph about hyperparameter optimization 5.4. In the evaluation, k-fold cross-validation is used with a 20% test set and 80% training set. In the tables mean, standard deviation and a maximum of the respective metric refer to the splits created in the cross-validation.

Starting with a comparison of the F-Scores obtained on the three datasets, Table 1 summarizes the results.

Table 1: Comparison (F-Score(%)) of the algorithm's performance on the three datasets in the research.

Note: For technical problems during the experiments, the results shown for the TVSum dataset are aggregated from only two splits of the datasets. The learning rate used in training is 10^{-4} for SumMe and TVSum and 10^{-6} for Breakfast.

Dataset	Mean F-Score	Standard deviation	Maximum F-Score
Breakfast	51.38	16.22	75.84
SumMe	50.25	1.37	52.17
TVSum	58.63	1.37	60.00

First, it is important to note that the produced results for the SumMe and TVSum datasets are comparable to the ones presented by Apostolidis et al. 2020 [2]. The exact numbers in the reference paper are 48.9 for SumMe and 58.3 for TVSum.

Looking at the table, one thing that strikes out is the unusually big gap between the average and the maximum F-Score of the Breakfast dataset. Possibly, the higher variance in the results for this dataset can be attributed to its smaller size compared to the other two. The standard deviation of the Breakfast dataset is an order of magnitude higher than

the one calculated for the other two datasets and with smaller training and test sets, more varying results can be expected. Nonetheless, the average F-Score of the Breakfast dataset is very close in value to the one produced for SumMe.

In terms of rank correlation coefficients, Table 2 presents a side-by-side comparison of Kendall's τ and Spearman's ρ coefficients for the Breakfast and SumMe datasets. Due to technical difficulties, there wasn't enough data to present such results for TVSum as well.

The rank correlation coefficients for the Breakfast dataset are slightly below 0, while for the SumMe dataset they are slightly positive. For both datasets, though, the rank correlation coefficients are very low and close to 0 which is the value expected from a randomly generated summary. Thus, from these results, one can conclude that the SUM-GAN-AAE algorithm is better suited to work with the SumMe dataset and doesn't generalize well enough for other types of datasets. As a reference, the rank correlation coefficients indicating the correlation between human-generated summaries from the Breakfast dataset are given. Clearly, the machine-generated ones are inferior even when considering the SumMe dataset. Important to notice here is also the standard deviation of the rank correlation coefficients for the human summaries. It is a few times higher than the ones calculated for the generated summaries. This unusually high variation in the results from split to split supports the claim that annotating summaries is an ambiguous and highly subjective task. Therefore, not only is it difficult to train a supervised algorithm using these user summaries but also evaluating the results of an unsupervised method is hindered.

For the purposes of hyperparameter optimization, a subset of the Breakfast videos used in the rest of the experiments was used. This is due to the fact that most of the research was conducted before the updated version of the Breakfast dataset was composed. The focus of the optimization is on the learning rate and the findings can give a general idea of how this parameter influences the evaluation metrics. The results from the conducted hyperparameter optimization can be seen in Table 3. The learning rate used in [2] for training on the SumMe and TVSum datasets is not optimal for the Breakfast dataset. By using F-Score and Kendall's τ as metrics, it can be seen that there's a clear trend of increasing the evaluation metrics as the learning rate decreases. The optimal learning rate found so far is 10^{-6}, which is orders of magnitude lower than the one used for SumMe and TVSum. Moreover, a learning rate of 10^{-8} worsens the metrics significantly, meaning that the parameter is best kept in this range. Further investigation would be needed to narrow down even more the optimal learning rate for the Breakfast dataset. To clarify, no hyperparameter tuning has been done on the SumMe and TVSum datasets in this paper. Although the results improved with the decreased learning rate, still the mean Kendall's τ calculated is not higher than of a random summary.

It is important to note that figures [1 2 5 3] all show results from training with 10^{-6} *learning rate since this is the optimal value found so far as can be seen from Table 3.*

In addition to the rank correlation coefficients provided, two plots are given to help the reader visualize the rank correlation. This type of plots is suggested by Otani et al. [13] and a thorough explanation of how the plots are produced can be found in section 5.1 of their paper. The intuition behind reading a graph of this sort is that a high correlation between generated summary and user summaries leads to steeply increasing graphs and vice versa. Figure 1 shows the plot of the video with lowest rank correlation coefficients (Kendall's τ equal to -0.27), while Figure 2 plots the video with the highest coefficients (Kendall's τ equal to 0.24). From the two graphs can be concluded that the algorithm's performance varies

significantly with different videos from the Breakfast dataset.

Another plot worth discussion is Figure 3 which is produced for the same video as in Figure 1. On the x-axis are the subsampled frames of the video which are given certain importance scores. The y-axis represents the importance scores, for the generated summary these are values in the range [0, 1] and for the user summary, they are boolean values (0's and 1's). The red line, representing the predicted importance scores, lies just below 0.2 for frames in the range 5 to 35. However, there is a noticeable increase in the importance scores for the beginning and end frames. The chosen video has a high F-Score but low correlation coefficients. The plot shows that the algorithm was not able to differentiate important frames from redundant ones and the frames given slightly higher importance by the algorithm are not the ones marked as significant by the user. The plotted importance scores for other videos look similar to the one presented. Overall, a pattern can be recognized that the algorithm assigns slightly higher importance scores to the frames at the start and at the end of a video, and the graph smooths out in the middle.

To examine the videos from the Breakfast dataset individually with their characteristics, Figure 5 presents more details. First, the videos with the highest F-Score are not the ones with the highest rank correlation coefficients. Thus, the two metrics show significantly different results. Moreover, there is no clear pattern associated with the video length. The metric values do not increase or decrease with the videos getting longer. The two highest rank correlation results come from videos which are both 1:26 long but this can be attributed to a coincidence.

Table 2: Rank correlation coefficients for the Breakfast and SumMe datasets. 'Human' denotes the evaluation of the rank correlation of user summaries from the Breakfast dataset. Note: for the calculations of rank correlation coefficients done on the SumMe dataset, ground-truth score has been used which is an aggregation of all user summaries.

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

Table 3: Results from the optimization of the learning rate when working with the Breakfast dataset.

Learning rate	Mean F-Score	Standard deviation (F-Score)	Mean Kendall's τ	Standard deviation (τ)
10^{-4}	58.36	10.63	-0.05	0.13
$5 * 10^{-5}$	62.04	13.54	-0.04	0.19
$2 * 10^{-5}$	63.23	12.37	-0.03	0.15
10^{-6}	64.42	12.52	-0.01	0.10
10^{-8}	49.47	21.37	-0.09	0.06

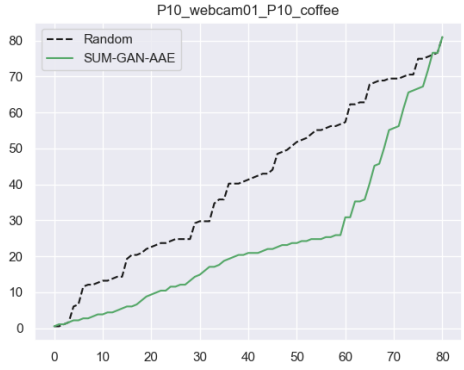


Figure 1: Visualization of the rank correlation for one of the coffee videos from the Breakfast dataset.

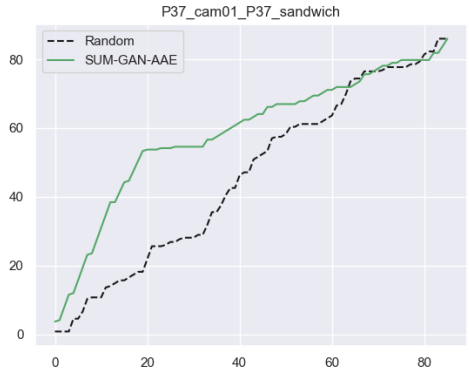
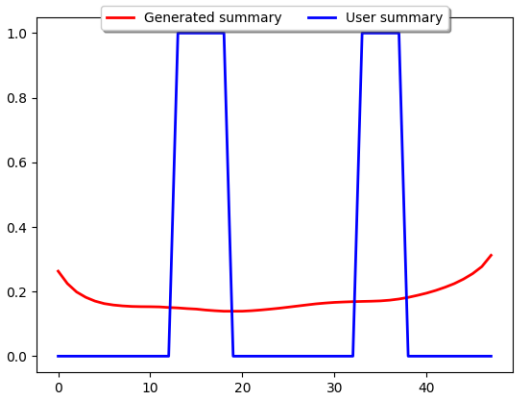


Figure 2: Visualization of the rank correlation for one of the sandwich videos from the Breakfast dataset.

Figure 3: Visualization of which frames are ranked as important by the generated summary and by one of the user summaries for the video 'P25_cam01_P25_cereals'.



5.5 Comparison to other supervised and unsupervised approaches

This research report is part of the work done by the research group supervised by Ms. Strafforello and Dr. Khademi. One of the purposes of the research is to compare the performance of supervised methods to unsupervised ones on the Breakfast dataset. A summarization of these findings can be found in Table 4. Despite doubts as to whether unsupervised algorithms might perform better given the inconsistent user summaries, supervised methods outperform them. The DSNet (Anchor-based)[6] model shows significantly higher results in terms of rank correlation coefficients with a Kendall's τ of 0.106. The rest of the supervised algorithms have lower performance but still above 0. Moreover, the algorithm with the highest F-Score has rank correlation coefficients under 0.05, showing a disagreement between the two metrics once again.

Between the two unsupervised algorithms, the SUM-GAN-AAE model performs notably better in terms of F-Score and stands slightly below the supervised models at 0.51. However, the rank correlation coefficients of SUM_ FCN$_{unsup}$[5] show better results even though just

marginally.

Table 4: Comparison of the algorithms tested in the research of the group under the supervision of Ms. Strafforello and Dr. Khademi. All results are produced on the Breakfast dataset

Type	Model	F-Score	Kendall's (τ)	Spearman's ρ
Supervised	VASNet[15]	0.673	0.045	0.0365
	DSNet (Anchor-based)[6]	0.6446	0.106	0.090
	DSNet (Anchor-free)[6]	0.6003	0.078	0.056
	SUM_FCN[5]	0.314	0.032	0.024
Unsupervised	SUM_FCN$_{unsup}$[5]	0.201	-0.021	-0.020
	SUM-GAN-AAE	0.51	-0.03	-0.03

6 Discussion

Taking into account the results from the conducted study, a few claims can be made about the researched topic.

First of all, the Breakfast dataset shows comparable F-Score results to the SumMe and TVSum datasets, even outperforming the former one on this criteria. However, it is important to keep in mind that this dataset is significantly smaller than the other two. Thus, the standard deviation is much higher for the Breakfast dataset.

Even with the promising F-Score results, the rank correlation coefficients do not favor the Breakfast dataset. This metric shows worse results for the SUM-GAN-AAE algorithm than generating random importance scores for the Breakfast dataset. With SumMe it performs slightly better than random but still far from the correlation shown between user summaries.

Another contribution of this research is the hyperparameter optimization performed with the Breakfast dataset. The findings point out that the optimal learning rate for training on that dataset is orders of magnitude lower than the one used by Apostolidis et al. [2] for training with SumMe and TVSum. However, even the highest rank correlation results produced after tuning the parameter are still comparable to a random algorithm.

The inability of the algorithm to generate summaries similar to human-generated ones is confirmed by the fact that all frames get very similar scores around 0.2. Furthermore, a pattern that can be observed in all generated summaries is that the importance scores in the beginning and at the end are slightly higher than average. This adds to the impression that the algorithm cannot pick the important frames from the rest. In addition, the higher rank correlation results on some of the videos are probably due to the fact that they happen to have their start or end frames ranked high by the user summaries as well.

Finally, from all experiments it is clear that the F-Score and the rank correlation coefficients sometimes show opposing results for the same video, meaning that at least one of them can be ruled out as an inadequate metric. Figure 3 in particular supports the claim that the algorithm does not outperform random frame selection. Thus, the rank correlation coefficients are more likely to be accurate according to these findings.

7 Responsible research

For research to be properly conducted and documented, all findings should be made available and presented accurately in the report. Moreover, the presented experiments should be transparently conducted and reproducible in order to be trustworthy. Last but not least, every piece of technology comes with benefits and drawbacks and both sides should be considered also when the discussion is about automatic video summarization.

As explained earlier, the implementation of the SUM-GAN-AAE algorithm comes from the repository belonging to the research by Apostolidis et al. [2]. All alterations made for the purposes of this research can be seen in the repository on GitLab repository made for the project. One can also find the latest results after the hyperparameter optimization in this repository. Thus, all plots shown in the report can be created also for other videos if needed. Furthermore, the reproducibility of results is ensured by specifying the datasets used, which are publicly available. However, the user annotations for the Breakfast dataset are a product of the research conducted by Ms. Strafforello and Dr. Khademi and this is up to them to publish. In addition, all evaluation metrics used are explained in detail with the respective equations for everyone to be able to evaluate the model's results in the same way.

When it comes to analyzing the pros and cons of the technology developed, one can easily list many arguments in favor of the development of this technology. We are currently flooded by video material in all forms and genres and people are often looking for a way to make the most out of the content of a video in a short amount of time. Thus, the need for an efficient way of summarization arises. Such method would find usage for educational purposes, self-development videos, a replay of sports events or video compilations for entertainment purposes.

Alongside those benefits, it is just as important to think about potential issues that might arise from such automation. First, depending on the context of the video, the importance of the information contained varies. Therefore, educational facilities should be careful when using such technology as it might not be an appropriate way of learning about a subject. In other cases, when it is not as crucial to miss some parts of the video, this might turn to be more useful. Another possible issue that is posed by the shortening videos is the possible misinterpretation of the content. There are videos in which details are very important for the message to be properly understood. Take an interview, for example, it would be very easy to take someone's words out of context even with an algorithm that does an excellent job at selecting the gist of a video.

8 Conclusions and Future Work

The hypothesis that leads to this research was that the SUM-GAN-AAE algorithm would perform well on an action localization dataset such as Breakfast. The findings show that the average F-Score produced by k-fold cross-validation is similar to the ones calculated for SumMe and TVSum datasets by Apostolidis et al. [2], which are considered high scores. However, by using rank correlation coefficients (Kendall's τ and Spearman's ρ), it is shown that the model performs no better than a random frame selection. Furthermore, visualizations of the generated importance scores per frame show that the algorithm produces very similar results for all videos, meaning that it has not learned any useful features of video summarization. These findings lead to the conclusion that not only does the algorithm perform poorly on the Breakfast dataset but also F-Score is not an appropriate evaluation

metric in this case.

There are still a few points for improvement for this research to be even more thorough. These can serve as a potential research question in future experiments on the topic.

First of all, the Breakfast dataset can be extended further with more user summaries. This would add to the data the algorithm is trained on and make the evaluation more accurate. Next, there are currently no reports evaluating the TVSum dataset with rank correlation coefficients. This was not possible in the current research due to insufficient resources and would be interesting to compare these coefficients to the ones from the SumMe dataset. According to the findings of this research, F-Score is not the most appropriate metric to evaluate video summaries and the use of this evaluation metric needs to be reconsidered in future experiments. Finally, in order to tune the learning rate parameter, even more, further optimization would be needed.

Table 5: Comparison of the results of the evaluation metrics on individual videos from the Breakfast dataset.

Breakfast				
Video ID	Video Length	Max F-Score	Kendall's τ	Spearman's ρ
P42_cam02_ P42_salat	5:43	98.18	0.00	0.00
P25_cam01_ P25_cereals	0:53	70.59	-0.26	-0.31
P40_cam02_ P40_milk	0:46	66.67	-0.11	-0.13
P10_webcam01_ P10_coffee	1:20	47.62	-0.27	-0.33
P03_webcam02_ P03_friedegg	4:44	54.55	0.00	0.00
P03_webcam02_ P03_sandwich	3:22	85.11	-0.08	-0.09
P07_cam01_ P07_ scrambledegg	2:31	51.16	0.15	0.18
P46_cam01_ P46_tea	1:26	90.91	0.18	0.21
P37_cam01_ P37_sandwich	1:26	43.48	0.24	0.29
P51_cam01_ P51_juice	1:39	72.22	0.06	0.07
P05_cam01_ P05_ scrambledegg	3:00	33.33	0.10	0.12
P05_cam01_ P05_coffee	1:14	0.00	-0.14	-0.17
P09_cam01_ P09_ scrambledegg	1:25	28.57	-0.09	-0.11
P12_cam01_ P12_sandwich	1:55	48.48	-0.16	-0.20
P29_cam01_ P29_juice	1:33	0.00	0.12	0.14
P38_cam01_ P38_ scrambledegg	3:50	65.31	-0.16	-0.19
P39_webcam02_ P39_sandwich	1:41	50.00	-0.08	-0.10
P47_webcam02_ P47_juice	1:24	0.00	0.09	0.11
P48_cam01_ P48_ scrambledegg	3:51	50.70	-0.11	-0.14
P48_cam02_ P48_milk	0:53	62.50	-0.10	-0.13

References

[1] *A Stepwise, Label-based Approach for Improving the Adversarial Training in Unsupervised Video Summarization*. Zenodo, Oct. 2019. DOI: 10.1145/3347449.3357482. URL: https://doi.org/10.1145/3347449.3357482.

[2] Evlampios Apostolidis et al. "Unsupervised Video Summarization via Attention-Driven Adversarial Learning". In: *MultiMedia Modeling*. Ed. by Yong Man Ro et al. Cham: Springer International Publishing, 2020, pp. 492–504. ISBN: 978-3-030-37731-1.

[3] Joao Carreira and Andrew Zisserman. "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. July 2017.

[4] Jiri Fajtl et al. "Summarizing Videos with Attention". In: *CoRR* abs/1812.01969 (2018). arXiv: 1812.01969. URL: http://arxiv.org/abs/1812.01969.

[5] Paul Frölke. "Evaluation of Video Summarization Using Fully Convolutional Sequence Networks on Action Localization Datasets". In: (2021).

[6] Daan Groenewegen. "Evaluation of Video Summarization Using DSNet and Action Localization Datasets". In: (2021).

[7] Michael Gygli et al. "Creating Summaries from User Videos". In: *ECCV*. 2014.

[8] Yunjae Jung et al. "Discriminative Feature Learning for Unsupervised Video Summarization". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01 (July 2019), pp. 8537–8544. DOI: 10.1609/aaai.v33i01.33018537. URL: https://ojs.aaai.org/index.php/AAAI/article/view/4872.

[9] M. G. KENDALL. "THE TREATMENT OF TIES IN RANKING PROBLEMS". In: *Biometrika* 33.3 (Nov. 1945), pp. 239–251. ISSN: 0006-3444. DOI: 10.1093/biomet/33.3.239. eprint: https://academic.oup.com/biomet/article-pdf/33/3/239/573257/33-3-239.pdf. URL: https://doi.org/10.1093/biomet/33.3.239.

[10] S. Kokoska and D. Zwillinger. *CRC standard probability and statistics tables and formulae*. Crc Press, 1999.

[11] Hilde Kuehne, Juergen Gall, and Thomas Serre. "An end-to-end generative framework for video segmentation and recognition". In: *Proc. IEEE Winter Applications of Computer Vision Conference (WACV 16)*. Lake Placid, Mar. 2016.

[12] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. "Unsupervised Video Summarization with Adversarial LSTM Networks". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 2982–2991. DOI: 10.1109/CVPR.2017.318.

[13] Mayu Otani et al. "Rethinking the Evaluation of Video Summaries". In: *CoRR* abs/1903.11328 (2019). arXiv: 1903.11328. URL: http://arxiv.org/abs/1903.11328.

[14] Yale Song et al. "TVSum: Summarizing Web Videos Using Titles". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2015.

[15] Felicia Elfrida Tjhai. "Evaluating the Supervised Video Summarization Model VASNet on an Action Localization Dataset". In: (2021).