Evaluation of Video Summarization Using Fully Convolutional Sequence Networks on Action Localization Datasets

Paul Frölke , Ombretta Strafforello , Seyran Khademi
TU Delft
p

Abstract

In the problem of video summarization, the goal is to select a subset of the input frames conveying the most important information of the input video. The collection of data proves to be a challenging task. In part because there exists a disagreement among human annotators on what segments of a video should be considered important for a summary. In this study we analyse a new dataset created with the goal of increasing agreement between the human annotators. The dataset has been created with the use of a novel annotation method, which uses existing action localization labels for segmenting the videos. We train a supervised and an unsupervised deep learning framework on popularly used benchmark datasets and the new dataset. Experimental results show the effectiveness of this novel summary annotation method in improving the agreement between annotators. Analysis reveals some issues with the evaluation of the deep learning framework.

1 Introduction

With the abundance of devices capable of video capture, video has become a indispensable medium for storing and sharing information. According to YouTube [34], more than 500 hours of video content are uploaded to the platform every minute. Consequently, better methods for browsing the vast amount of videos are needed. The goal of video summarization is to select frames from an input video where the video created by combining the selected frames conveys the most important information of the original video. Effective summary videos can aid the process of navigating large video collections by enabling the user to quickly identify the contents of a video.

State-of-the-art automatic video summarization methods follow a similar process as illustrated in figure 1. An algorithm is applied to split the input video into segments, a deep neural network predicts the importance scores of the video frames, and finally the optimal combination of video segments is selected by the knapsack algorithm [20].

Fully convolutional sequence networks (FCSN) were first introduced by Rochan *et al.* [26] as a method for video summarization. Previously, fully convolutional networks (FCN) have been applied for the unconnected problem of semantic segmentation. Rochan *et al.* conceive a relationship between the problem of semantic segmentation and video summarization, and present the FCSN as an adaptation of the FCN suitable for video summarization. In addition to a model that can be trained supervised on labeled data (SUM-FCN), Rochan *et al.* presented an unsupervised extension that can learn without ground-truth video summary labels (SUM-FCN$_{\text{unsup}}$).

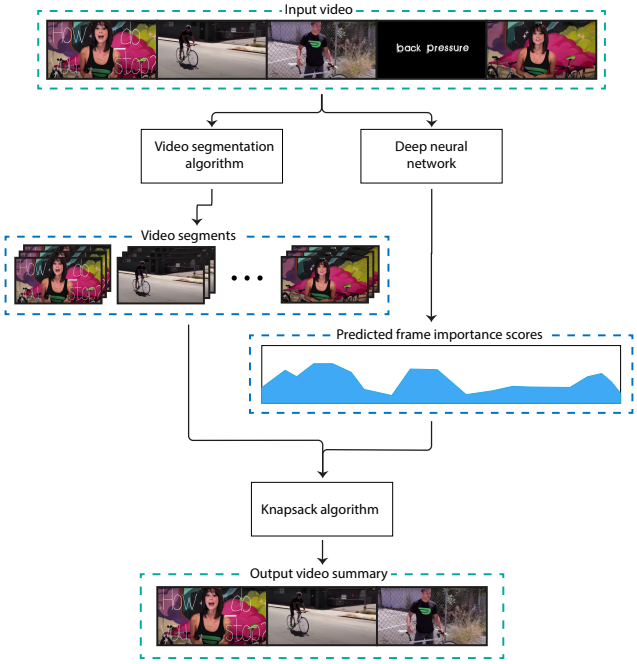


Figure 1: An illustration of the video summarization pipeline on a video from the TVSum [28] dataset.

Popular video summarization datasets [28; 7] provide benchmark summaries that have been created by human annotators that can be used for supervised learning. As a result of disagreement between the annotators on what defines a good summary, the labels in these datasets contain noise.

With the goal of reducing the disagreement between annotators, a novel video summary dataset has been created by performing surveys with videos from the Breakfast Actions

dataset [13]. This dataset provide action-localization data, which was used in the survey for dividing the videos into segments. The annotators were tasked with selecting the segments that best represent the original video.

Furthermore, Otani *et al.* [20] demonstrated that the widely used F1-score evaluation metric [27] has severe problems. The F1-score is calculated from the precision and recall scores for the generated summary compared to the ground truth summary. This process is shown to be greatly influenced by the video segmentation process and completely random assignments of importance scores were able to reach similar or better scores than the state-of-the-art deep learning methods. In order to more accurately score the performance of a deep learning network, Otani *et al.* propose a new evaluation method which considers the rank-order correlation between predicted and human-annotated frame-level importance scores.

In this paper, we analyse the effectiveness of the novel video summary dataset created from the Breakfast Actions dataset [13] in reducing the disagreement in video summarization by analysing the results of the fully convolutional networks SUM-FCN with supervised learning and SUM-$\text{FCN}_{\text{unsup}}$ [26] with unsupervised learning. In section 3, we first give the details of the benchmark video summary datasets, followed by an explanation of the evaluation techniques. In section 4, we note the experimental setup and main results. The results are analysed in section 5. Section 6 reflects on the ethical aspects of this research and the reproducibility of our experiments. In section 7 a reflection is made on the research. Finally, we recap the main conclusions in section 8 and follow with ideas for further research.

2 Related Works

2.1 Automated Video Summarization

The goal of video summarization is to shorten an input video to a summary video which conveys the most important information of the original video. Several different problem representations have been presented in literature, such as video montages [10], time-lapses [8; 23], synopsis [25], and storyboards [9; 35; 37; 26; 36; 16]. Storyboard summaries are usually composed of a set of representative video frames (key-frames) [16], or video fragments (key-shots) [19]. Key-shot summaries are often considered to be more interesting to watch for the viewer than a slide show of key-frames [15].

The problem discussed in this paper is most equivalent to key-shot storyboard summarization. Rochan *et al.* [26] approach video summarization as a binary classification problem, where each frame is classified as being a key-frame or not a key-frame by the FCSN.

2.2 Deep Learning Approaches

Most early approaches for video summarization depend on predefined heuristic functions to determine the importance scores of the frames [12; 14]. More recently, deep learning methods have been researched [9; 35; 37; 26; 36; 19; 4].

One set of deep learning models [37; 4; 36] is trained with supervised learning using human-annotated benchmark datasets. As a result the amount training data is limited, and these models are influenced by the subjectivity of human annotators. However, supervised learning allows models to implicitly capture high-level information from the annotations. Other deep learning methods [9; 35] have been presented that can be trained using unsupervised learning and therefore mitigate the need for human-annotated ground-truth data.

In addition to supervised and unsupervised deep learning methods, another class [21] attempts to learn video summarization without the need for human annotated summaries, by using supervised learning techniques on less-expensive weak labels (e.g. video title, category).

2.3 Benchmarks

In recent literature the TVSum [28] and SumMe [7] datasets are most commonly used for comparing the performance of different video summarization methods. Other popular datasets include: OVP [3] and YouTube [3]. OVP and YouTube both contain 50 videos from diverse categories. Both datasets provide key-frame summary annotations created by 5 annotators. The OVP and YouTube datasets are typically used for the assessment of the performance of a method in the augmented and transfer settings (Sec. 2.4).

2.4 Evaluation

F1-score [27] is most predominantly used as a measure for the accuracy of a video summarization method. Next to the standard setting where a single dataset is split into a training, validation and test set, scores for the augmented setting are frequently reported as well. In the augmented setting, the training set is supplemented with other datasets, and for testing only the initial dataset is used. The aim of training on augmented data is to reduce overfitting.

Additionally, Zhang *et al.* [36] introduced a transfer score. In this setting the training set completely consists of data from other datasets, while being tested on the given dataset. This score particularly shows how well a video summary method can generalize for practical applications.

For a qualitative evaluation of generated summaries, some works [17; 18; 32] rely on human subjects. In a survey subjects view videos and the generated summaries and subjectively the rate how well each summary represents the original video. These studies give a good representation of how well a video summarization method would practically perform for the end-user, but have the disadvantage that the results are hard to reproduce.

3 Evaluating the Performance of the FCSN

In this section the methodology for the evaluation of the performance of the FCSN models is explained. First the details of the benchmark datasets that were used are given, then the evaluation approach is elaborated.

3.1 TVSum

TVSum contains 50 videos obtained from YouTube, using 10 predefined search queries (5 videos per query) [28]. The duration of the videos ranges from 2 to 10 minutes. In a survey, 20 human annotators ranked the importance of uniform

length shots of each video on a scale from 1 (unimportant) to 5 (important). This importance scoring approach has the benefit that any arbitrary length summary can be generated.

3.2 SumMe

The SumMe dataset consists of 25 YouTube videos with diverse subjects such as holidays, events and sports [7]. The duration of the videos varies in the range from about half a minute to 6 minutes. For each video, 15 to 18 human annotators produced summaries by selecting a subset of the frames that was required to be of length $5\% \leq L_s \leq 15\%$ of the initial video duration.

3.3 Breakfast Actions

The Breakfast Actions dataset contains 1.989 videos of actors performing cooking activities [13]. The dataset provides human annotated segment boundaries for the videos, where each video segment shows the actor performing a labeled action. It has annotations for "coarse" actions labels (e.g. "grab milk") and "fine" action labels (e.g. "twist cap", "open cap").

Videos from this dataset have been used to create a novel video summary dataset by conducting human surveys. First, a video from the datasets was shown to an annotator. Second, all the action labels corresponding to the video are listed, and the annotator is asked to select all actions that best represent the video content. Third, the annotator is tasked with creating a video summary by selecting up to 2 segments from the video. These video segments correspond to the "coarse" action labeled segments. A total of 20 videos from the Breakfast Actions dataset were labeled by 2 to 15 annotators using this method. The duration of the selected videos ranges from 48 seconds to almost 5 minutes.

The results from the surveys were afterwards aggregated to form the final dataset. For each annotator frame-level binary labels were generated which indicate whether the frame is selected or not. All annotator summaries were converted to one ground-truth frame-level importance score vector by dividing the amount of annotators who picked each frame by the total number of annotators for each video. Finally, every second of the original video was converted to a single feature vector using the pretrained I3D network [2].

3.4 Fully Convolutional Sequence Networks

Fully convolutional sequence networks (FCSN) [26] are a type of convolutional neural network inspired by models used in the problem of semantic segmentation. In semantic segmentation, the objective is to label each pixel in an image to some representative class. Rochan *et al.* [26] establish a relationship between this problem and the problem of video summarization by formulating video summarization as a binary key-frame classification problem over the temporal dimension. The only difference that remains between these problems are the input dimensions and the number of channels. By adapting deep learning models for semantic segmentation to these differences they can be applied for video summarization.

Rochan *et al.* [26] present both a supervised (SUM-FCN) and unsupervised (SUM-FCN$_{\text{unsup}}$) model for video summarization. SUM-FCN learns from a single ground-truth set of key-frames generated from human annotations. It learns by using a class-balanced negative log likelihood loss function. SUM-FCN$_{\text{unsup}}$ is designed on the intuition that the chosen key-frames should be visually diverse by using a reconstruction and diversity component in the loss function. We compare the performance of both networks to analyse to what extent the human annotations aid the learning process.

3.5 Evaluation Approach

With the goal of understanding the effects of the novel summary annotation method, the fully convolutional networks SUM-FCN and SUM-FCN$_{\text{unsup}}$ are trained on the TVSum, SumMe and the Breakfast Actions datasets. The experimental setup from Rochan *et al.* [26] is followed to train the models (Sec. 4.1).

Following the methods described by Otani *et al.* [20], the performance of the SUM-FCN and SUM-FCN$_{\text{unsup}}$ are evaluated on the benchmark datasets TVSum and SumMe and the new Breakfast Actions dataset using rank-order correlation scoring. To compute this score, the frames of an input video are ranked by the predicted importance scores. Then Kendall's τ [11] and Spearman's ρ [38] coefficients are calculated for the correlation between the generated ranking and the ranking from ground-truth importance scores. The reported correlation score is the mean of the correlation coefficients for all videos in a dataset.

We also compute the F1-score from the precision (P) and recall (R) values. To find the F1-score for a particular video, the scores are calculated by comparing to each human annotation and the average is reported.

$$F1 = \frac{2P \cdot R}{P + R} \tag{1}$$

with:

$$P = \frac{\sum_{i=1}^{N} y_i \cdot y_i^*}{\sum_{i=1}^{N} y_i} \qquad \text{and} \qquad R = \frac{\sum_{i=1}^{N} y_i \cdot y_i^*}{\sum_{i=1}^{N} y_i^*} \tag{2}$$

where for each frame number i: $y_i \in \{0, 1\}$ denotes the predicted binary label and $y_i^* \in \{0, 1\}$ the ground-truth label.

4 Experimental Setup and Results

In this section, we report the details for the setup of our experiments on SUM-FCN and SUM-FCN$_{\text{unsup}}$ along with justifications for the parameters that were chosen. Following that, the results of the experiments are reported.

4.1 Experiment Setup

The experiment setup can be subdivided into an implementation, ground-truth generation, training and evaluation part.

Implementation: A third party open-source implementation was used as a base for the FCSN models. The implementation uses the PyTorch machine learning library [22] for modeling and training the network. The concerns with using a third party implementation are further elaborated in the discussion section (Sec. 7).

Ground-truth: The TVSum, SumMe and Breakfast Actions datasets provide the ground-truth annotations in various

Dataset	Method	F1-score		Kendall's τ		Spearman's ρ	
TVSum	SUM-FCN	56.5	(82.0)	0.006	(0.076)	0.009	(0.112)
	SUM-FCN$_{\text{unsup}}$	52.9	(72.7)	0.009	(0.152)	0.013	(0.224)
	Random	56.4	(80.4)	0.000	(0.005)	0.000	(0.004)
	Human	53.8	(80.5)	0.177	(0.311)	0.204	(0.357)
SumMe	SUM-FCN	30.9	(84.3)	-0.003	(0.027)	0.004	(0.034)
	SUM-FCN$_{\text{unsup}}$	28.3	(51.8)	0.000	(0.062)	-0.011	(0.050)
	Random	18.7	(52.9)	0.000	(0.002)	0.000	(0.002)
	Human	31.1	(79.0)	0.202	(0.425)	0.213	(0.440)
Breakfast	SUM-FCN	31.4	(86.3)	0.024	(0.162)	0.032	(0.215)
	SUM-FCN$_{\text{unsup}}$	20.1	(50.1)	-0.020	(0.282)	-0.021	(0.356)
	Random	21.4	(29.0)	0.000	(0.012)	0.000	(0.022)
	Human	43.2	(100.0)	-		-	

Table 1: Mean F1-score and correlation score results, comparison of different methods on the benchmark datasets. The maximum scores are noted between brackets. Note that for the Breakfast Actions dataset it is not possible to generate the correlation scores on the human annotations, considering the fact that these are not given as importance scores but as binary labels.

formats. Since the FCSN models require a single ground-truth set of isolated key-frames for each video for training, we follow methods from [36] to convert these annotations. For the frame-level scores annotations provided by TVSum and SumMe each videos is first segmented using KTS [24], the segments are ranked by mean importance score of all annotations, then the knapsack algorithm is applied to select from the segments with a maximum duration of 15%. Finally for each selected segment, the frame with the highest importance score is marked as key-frame [36].

For the Breakfast Actions dataset, we follow [5] to greedily create a single ground-truth key-shot summary from the key-shot annotations. Then ground-truth key-frames are obtained by taking the middle frame of each key-shot as in [36] (Figure 2).

Training: In order to train SUM-FCN and SUM-FCN$_{\text{unsup}}$ on videos of variable length, Rochan *et al.* report two methods. Parallel to the cropping strategy in semantic segmentation, the feature vectors can be uniformly down-sampled to a fixed length ($T = 320$ [26]). Alternatively, since the models are fully convolutional the variable length feature vector representation can be used without sampling. In these experiments we use the latter method. In Sec. 5 we compare the results of both methods.

For training on TVSum and SumMe, we followed [26] and use a learning rate of 10^{-3}, momentum of 0.9, and batch size of 5. For training on the Breakfast Actions dataset, we use the same learning rate and momentum. However since the amount of videos is significantly less, we lower the batch size to 3. The networks are optimized with stochastic gradient descent (SGD) and trained for 100 epochs on each training-set/test-set split in 5-fold cross-validation.

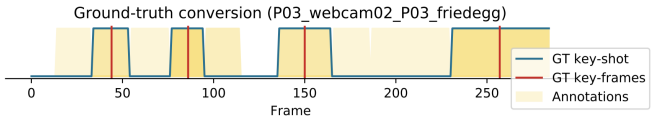


Figure 2: An illustration of the conversion process from the provided human annotations (shaded in yellow) to a single ground-truth key-shot summary (blue) on a video from the Breakfast Actions dataset. Ground-truth key-frames (red) are then obtained by taking the middle frame of each key-shot.

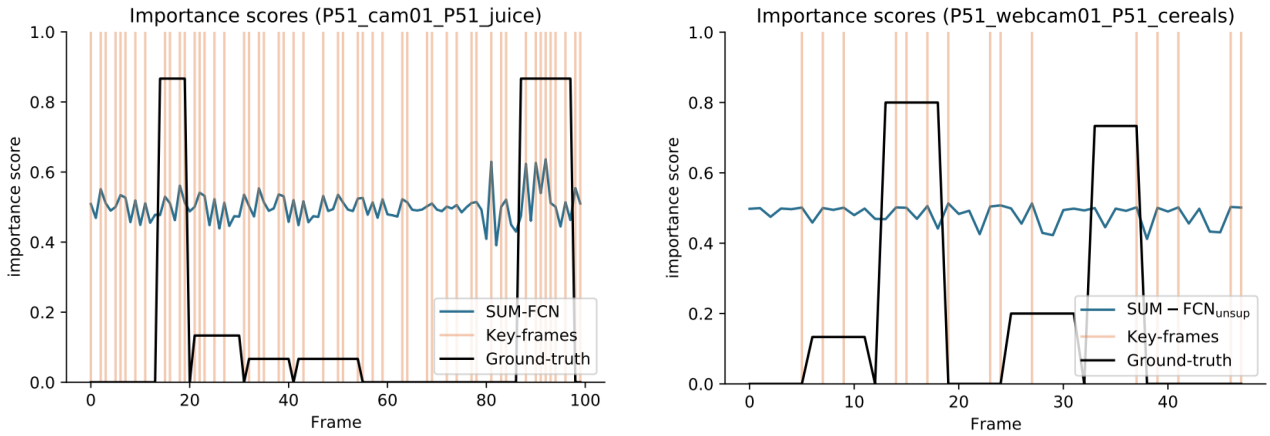
Evaluation: The output layers of the SUM-FCN and SUM-FCN$_{\text{unsup}}$ networks are of dimension $1 \times T \times C$ where $C = 2$ denotes the two classes a frame can be classified as (selected as key-frame or not). Following [36; 26] we convert key-frames to key-shots before calculating F1-score. First a video is segmented using KTS [24]. Next, each segment is scored by the number of contained key-frames divided by the segment length. Finally the knapsack algorithm selects the best scoring combination of segments within the maximum allowed summary length. For the TVSum and SumMe datasets, the maximum summary length is 15% of the original video length. For the Breakfast Actions dataset the provided video segments from the action localization labels are used instead of KTS, and the maximum summary length is set to 2 segments in order to copy the requirements for the human annotators (Sec. 3.3).

In order to obtain predicted frame-level importance scores for computing the correlation scores, the values for key-frame classification are taken from the output layer before the binary classification threshold is applied. These scores are a vector of length $T = 288$ and are cut to the original length of the video before calculating the correlation scores.

In addition to these scores, scores for the human annotators and a randomized importance score baseline method are computed (as in [20]). For the randomized baseline method, the scores are averaged over 100 trials. The human score on the Breakfast Actions dataset is calculated by taking the mean of the F1-scores obtained by pairwise comparing the human annotators.

4.2 Main Results

Table 1 shows the mean F1-score and mean correlation scores obtained with the various methods on the benchmark datasets. The F1-score and correlation score results of the human and



(a) SUM-FCN
$F_1 = 86.3, \tau = 0.052, \rho = 0.069$

(b) SUM-FCN$_{unsup}$
$F_1 = 50.1, \tau = -0.016, \rho = -0.018$

Figure 3: Two plots showing the frame importance scores predicted by SUM-FCN (a) and SUM-FCN$_{unsup}$ (b), and the corresponding key-frames, compared to the ground-truth importance scores on videos from the Breakfast Actions dataset. The generated summaries for these particular videos obtained high F1-scores, but low correlation scores.

random baselines on TVSum and SumMe are consistent with the results Otani *et al.* [20] report.

The results show SUM-FCN performs similarly in terms of F1-score on the SumMe and Breakfast Actions video summary datasets, and much better on TVSum. For SUM-FCN$_{unsup}$, the F1-scores show it performs significantly worse on the Breakfast Actions dataset compared to TVSum and SumMe. Rank-order correlation scores have been shown to provide a better measure of the performance of a model [20]. The rank-order correlation scores in Table 1 show highest performance for SUM-FCN on the Breakfast Actions dataset, which could be an indication of increased annotator agreement. These scores also show a bigger difference between the performance of the supervised method compared to the unsupervised method. However there is a notable issue with calculating these scores for the FCSN models we discuss in Section 5.3.

The human F1-score on TVSum differs from the F1-score 36 reported by the authors [28], because we first use KTS [24] to convert each human annotation to a key-shot summary with a maximum duration of 15% before calculating the score. In contrast, Song *et al.* [28] calculate the F1-score from key-shots of 2 seconds each. We chose for this approach, since it allows for direct comparison between the human scores and the scores obtained by the other methods.

In Table 2 the mean F1-scores of the FCSN models as reported by Rochan *et al.* [26] are listed. The scores we achieved on the TVSum dataset are comparable to the originally reported results. In contrary, the results on the SumMe dataset greatly differ. We attribute this disparity to issues within the implementation of the FCSN models (Sec. 7.2), since the experimental setup of Rochan *et al.* was accurately copied.

Dataset	SUM-FCN [26]	SUM-FCN$_{unsup}$ [26]
TVSum	56.8 (56.5)	52.7 (52.9)
SumMe	47.5 (30.9)	41.5 (28.3)

Table 2: Mean F1-scores of the FCSN models on the TVSum and SumMe datasets, as reported by Rochan *et al.* [26]. Between brackets the the F1-scores we achieved are reported (copied from Table 1).

5 Analysis

5.1 Human Agreement

The agreement between the human annotators of the Breakfast Actions dataset can be analysed by comparing the results obtained by pairwise F1-score evaluation. This is possible since the human annotators provide key-shot summaries, which can be compared with F1-score without conversion to other formats. Figure 4 shows the mean F1-score for each video in the dataset. From this bar graph we observe the dataset contains one video which has an annotator F1-score 0.0 ("P05_cam01_P05_scrambledegg"). Examining this video further we find that for this video the provided annotations include just 2 people, who have chosen none of the same segments. Another remarkable result is a video with an annotator F1-score 1.0 ("P48_cam02_P48_milk"). This video has been annotated by three individuals, who all chose the exact same segments for the summary.

5.2 Downsampling

As mentioned in Sec. 4.1, Rochan *et al.* [26] use videos downsampled to a fixed length before training the FCSN models. We performed experiments to compare the performance of the models on downsampled videos to the performance on variable length videos. Since the feature vectors given in the Breakfast Actions dataset have a minimum length of 48, the

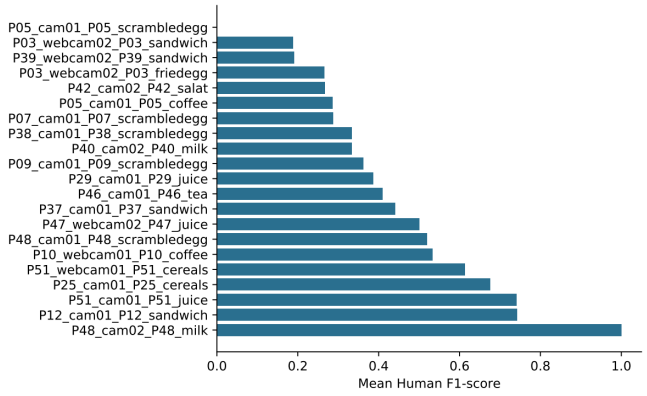


Figure 4: All videos in the Breakfast Actions dataset with their annotator-agreement F1-score, calculated by pairwise comparison of the provided human annotations.

down-sampled feature vectors can have a maximum length $T = 48$.

Table 3 shows the scores obtained when training on the downsampled videos and the scores obtained when training on variable length videos. For FCN-SUM both the F1-score and correlation scores are significantly higher with variable length videos. For SUM-FCN$_{\text{unsup}}$ the F1-scores are similar for both cases. The correlation scores are better with the downsampled videos, however in both cases worse than random scores.

Dataset	Method	F1	K.'s τ	S.'ρ
Breakfast ($T = 48$)	SUM-FCN	22.4	0.014	0.017
	SUM-FCN$_{\text{unsup}}$	20.7	-0.010	-0.013
Breakfast var. length	SUM-FCN	31.4	0.024	0.032
	SUM-FCN$_{\text{unsup}}$	20.1	-0.020	-0.021

Table 3: Mean F1-score and correlation scores of the FCSN models on the Breakfast dataset. Comparison of the test performance obtained when training on videos which have been uniformly downsampled to $T = 48$, and variable length input videos (taken from Table 1).

Compared to the mean human annotation F1-score of TVSum(Table 1), the agreement between human annotations for the Breakfast Actions dataset seems significantly lower. However this is an unfair comparison, since the human annotation F1-score for the TVSum and SumMe datasets have been calculated after converting the annotations from importance scores to key-shots (Sec. 4.1), which has been shown to be highly dependent on the video segmentation algorithm [20].

5.3 Importance Score Evaluation

Figure 3 shows two plots of the frame importance scores predicted by SUM-FCN (3a) and SUM-FCN$_{\text{unsup}}$ (3b) compared to the respective ground-truth importance scores for videos from the Breakfast Actions dataset. Furthermore, the predicted key-frames are highlighted in orange. These video are

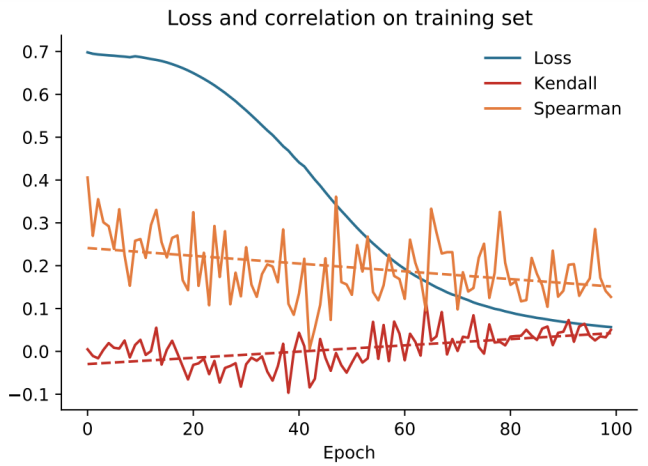


Figure 5: A plot of the loss, Kendall's τ, and Spearman's ρ during training of FCN-SUM. Scores are averaged over each epoch of a set of videos from the Breakfast Actions dataset. Trained and tested on the same set of videos to accurately show the correlation between loss and the correlation scores. For each correlation metric, a dashed trend-line is shown.

featured since they obtained high F1-scores but low correlation scores.

An analysis of Figure 3 reveals a severe issue with performing rank-order correlation evaluation on the predicted scores of the FCSN models. The FCSN models are designed to classify frames of a video as key-frame or non-key-frame. The classification is determined by a threshold value (0.5) on the predicted importance scores (i.e. score $> 0.5 =$ keyframe, highlighted in orange in Figure 3).

Key-frame summaries consist of a set of the most important frames of a video, which are usually isolated. For example, the ground-truth labels used by SUM-FCN consist of 1 frame for each segment (Figure 2). Therefore SUM-FCN will implicitly learn to classify diverse frames as key-frames [26], and frames similar to a key-frame as non-key-frames.

This effect is also present in summaries generated by SUM-FCN$_{\text{unsup}}$ (e.g. Figure 3b), owing to the design of its loss function. It incorporates a repelling regularizer component which explicitly enforces diversity among the selected key-frames [26].

In contrast, the ground-truth importance scores of the Breakfast Actions dataset which are used during rank-order correlation evaluation are generated from key-shot annotations, resulting in a constant value for each segment (Figure 3).

As a result of this inequality, the output scores of SUM-FCN and SUM-FCN$_{\text{unsup}}$ are not suitable for calculating rank-order correlation scores. Consequently the correlation scores are low.

This statement is further supported by Figure 5. It shows mean correlation scores as well as the mean loss, computed during the training of SUM-FCN. The dashed trend-lines of the correlation scores demonstrate that reducing the loss is not effective in increasing the correlation scores significantly.

5.4 Comparison to Other Methods

Table 4 compares the results we achieved for the FCSN models on the Breakfast Actions dataset to the results that have been reported [30; 6; 31] for other deep learning methods.

In comparison to the other supervised methods, SUM-FCN is outperformed on all evaluation metrics. Both VASNet [4] and the DSNet models [37] use frame-level importance scores for training the models. Compared to the key-frame ground-truth (Sec. 4.1) the SUM-FCN trains with, frame-level importance scores have previously been shown to provide richer information [36].

The unsupervised method SUM-GAN-AAE [1] achieves significantly higher F1-score than SUM-FCN$_{\text{unsup}}$. SUM-FCN performs slightly better when evaluated using rank-order correlation scores, however both methods perform worse than randomized importance scores.

6 Responsible Research

6.1 Ethical Concerns

For this research, data collection was conducted by the use of surveys on the Amazon Mechanical Turk platform. On this platform, users can choose to complete tasks (HIT) and then get paid some arbitrary amount of money as a reward. It is hard to determine how much each HIT should be worth, since we do not know any information about the users beforehand. On the other hand, users have complete freedom to choose what HIT to complete. Therefore the reward should only be viewed as an incentive and not as a payment.

6.2 Reproducibility of Experiments

In empirical research, it is crucial to ensure the experimental results presented can be easily reproduced and verified by other researchers. To this respect, in this research we have made sure to exclusively make use of open source datasets and repositories accessible to anyone. Additionally, the dataset preparation process and experiment parameters used have been noted and justified (Sec. 4). This guarantees that others can easily conduct the same experiments and replicate the reported results.

It should be noted that there was no official repository available for the FCSN, therefore a publicly available implementation published by a third party was used. The implications of this choice are further elaborated upon in the next section (Sec. 7.2).

7 Discussion

7.1 Notable Differences in the Breakfast Actions Dataset

It should be noted that some properties of the Breakfast Actions video summary dataset are different to the properties of the TVSum and SumMe datasets which could result in an unfair comparison between the results.

First, the feature descriptor we used for the TVSum and SumMe datasets is GoogleNet [29]. For the Breakfast Actions dataset, the features were extracted using I3D [2]. Despite the fact that Rochan *et al.* [26] mention that the FCSN models can use any feature descriptor, the results we obtained could be unfairly influenced considering I3D extracts video-based features, but GoogleNet extracts image-based features.

Second, the annotators on the Breakfast Actions dataset were tasked with selecting a fixed number of segments from the videos for a summary. For each video, the importance scores were subsequently generated by dividing the number of annotators who picked a segment, by the total number of annotators for the video, as described in Section 3.3. This process has as a result that the resulting frame-level importance scores do not necessarily show relative importance. Take for example the situation where an annotator considers only one segment of the video to be important for a summary. With the current annotation method the annotator has to choose two segments, and both segments will be counted to be of equal importance.

Finally, as mentioned in Section 5.1 some video's in the Breakfast Actions dataset have been annotated by a very low number of people. As a result, the agreement measures between annotators are not reliably comparable to the other datasets, since adding more annotations could greatly increase or decrease the agreement for these videos.

7.2 Open Source Repository

Since no official implementation was available for the FCSN models, initially when starting this research project, the first unofficial repository found on GitHub was used[1]. However, training with this implementation and analysing the results quickly made clear that it had significant errors. Most notably, the F1-scores obtained when training on the TVSum and SumMe datasets where considerably lower than the scores reported by the authors [26]. After this observation, we examined the second repository found on GitHub[2]. This implementation was able to reproduce the F1-scores, however it turned out that the scores were incorrectly calculated. After extensive debugging of this implementation, we were able to reproduce the scores for the TVSum dataset, but not for SumMe (Sec. 4.2). Consequently, the reliability of the results that we present for the new dataset is questionable. The modified implementation we used for this research has been made available on GitHub[3].

This experience has been an example of the importance of responsible research and assuring the reproducibility of the experiments. If an implementation had been provided by the original authors of the FCSN [26], less time would have been needed for debugging the code.

8 Conclusions and Future Work

In this research, we have assessed the usability of a novel dataset that was created by using the existing action-localization labels of the Breakfast Actions dataset [13] during annotation for the problem of video summarization. We have trained two deep-learning models: the supervised SUM-FCN [26] and the unsupervised SUM-FCN$_{\text{unsup}}$ [26] on the dataset and on previous benchmark datasets TVSum and

[1] https://github.com/weirme/Video_Summary_using_FCSN

[2] https://github.com/pcshih/pytorch-FCSN

[3] https://github.com/pfrolke/pytorch-FCSN

Type	Model	F1-score	Kendall's τ	Spearman's ρ
Supervised	VASNet [30]	67.3	0.037	0.045
	DSNet (Anchor-based) [6]	64.4	0.090	0.106
	DSNet (Anchor-free) [6]	60.0	0.056	0.078
	SUM-FCN	31.4	0.024	0.032
Unsupervised	SUM-GAN-AAE [31]	51.4	-0.030	-0.030
	SUM-FCN$_{\text{unsup}}$	20.1	-0.020	-0.021

Table 4: Performance of different video summarization models on the Breakfast Actions dataset. For each of the models, the mean F1-score and rank-order correlation scores between predicted importance scores and their corresponding ground-truth importance scores are given. The scores for SUM-FCN and SUM-FCN$_{\text{unsup}}$ are taken from Table 1.

SumMe. For a comparison of the performance of these networks we used F1-score [27] and rank-order correlation evaluation [20].

The results we obtained for the supervised model than the results that have been reported for other supervised models on the new dataset are significantly lower. We attribute this performance gap to the difference in training ground-truth format. The other examined models use frame importance scores, SUM-FCN uses key-frames. The experimental results we achieved with the unsupervised model are worse in terms of F1-score than the reported results of another unsupervised model, but the correlation scores are slightly better.

Furthermore, analysis reveals that the key-frame based summaries which both SUM-FCN and SUM-FCN$_{\text{unsup}}$ are designed to learn are not suitable for computing rank-order correlation scores with frame-level importance scores, since they inherently model a different objective.

For future works, we propose experimenting with new models adapted from the FCSN models that learn ground-truth importance scores as opposed to ground-truth key-frames. Additionally, the agreement between annotators of video summarization datasets created by annotating action-localization datasets with videos from a different domain (e.g. MultiThumos [33]) should be researched. Finally, in order to compute rank-order correlation scores for human annotations the human annotations have to be collected with importance scores.

References

[1] Evlampios Apostolidis, Eleni Adamantidou, Alexandros I Metsai, Vasileios Mezaris, and Ioannis Patras. Unsupervised video summarization via attention-driven adversarial learning. In *International Conference on Multimedia Modeling*, pages 492–504. Springer, 2020.

[2] 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*, pages 6299–6308, 2017.

[3] Sandra Eliza Fontes De Avila, Ana Paula Brandao Lopes, Antonio da Luz Jr, and Arnaldo de Albuquerque Araújo. Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters*, 32(1):56–68, 2011.

[4] Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Summarizing videos with attention. In *Asian Conference on Computer Vision*, pages 39–54. Springer, 2018.

[5] Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. Diverse sequential subset selection for supervised video summarization. *Advances in neural information processing systems*, 27:2069–2077, 2014.

[6] Daan Groenewegen and Ombretta Strafforello. Evaluation of video summarization using dsnet and action localization datasets. 2021.

[7] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *ECCV*, 2014.

[8] Neel Joshi, Wolf Kienzle, Mike Toelle, Matt Uyttendaele, and Michael F Cohen. Real-time hyperlapse creation via optimal frame selection. *ACM Transactions on Graphics (TOG)*, 34(4):1–9, 2015.

[9] Yunjae Jung, Donghyeon Cho, Dahun Kim, Sanghyun Woo, and In So Kweon. Discriminative feature learning for unsupervised video summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8537–8544, 2019.

[10] Hong-Wen Kang, Yasuyuki Matsushita, Xiaoou Tang, and Xue-Quan Chen. Space-time video montage. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1331–1338. IEEE, 2006.

[11] Maurice G Kendall. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251, 1945.

[12] Aditya Khosla, Raffay Hamid, Chih-Jen Lin, and Neel Sundaresan. Large-scale video summarization using web-image priors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2698–2705, 2013.

[13] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 780–787, 2014.

[14] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocen-

tric video summarization. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1346–1353. IEEE, 2012.

[15] Ying Li, Tong Zhang, and Daniel Tretter. An overview of video abstraction techniques. Technical report, Technical Report HPL-2001-191, HP Laboratory, 2001.

[16] David Liu, Gang Hua, and Tsuhan Chen. A hierarchical visual model for video object summarization. *IEEE transactions on pattern analysis and machine intelligence*, 32(12):2178–2190, 2010.

[17] Zheng Lu and Kristen Grauman. Story-driven summarization for egocentric video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2714–2721, 2013.

[18] Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang, and Mingjing Li. A user attention model for video summarization. In *Proceedings of the tenth ACM international conference on Multimedia*, pages 533–542, 2002.

[19] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. Unsupervised video summarization with adversarial lstm networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 202–211, 2017.

[20] Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkila. Rethinking the evaluation of video summaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7596–7604, 2019.

[21] Rameswar Panda, Abir Das, Ziyan Wu, Jan Ernst, and Amit K Roy-Chowdhury. Weakly supervised summarization of web videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3657–3666, 2017.

[22] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

[23] Yair Poleg, Tavi Halperin, Chetan Arora, and Shmuel Peleg. Egosampling: Fast-forward and stereo for egocentric videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4768–4776, 2015.

[24] Danila Potapov, Matthijs Douze, Zaid Harchaoui, and Cordelia Schmid. Category-specific video summarization. In *European conference on computer vision*, pages 540–555. Springer, 2014.

[25] Yael Pritch, Alex Rav-Acha, and Shmuel Peleg. Nonchronological video synopsis and indexing. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1971–1984, 2008.

[26] Mrigank Rochan, Linwei Ye, and Yang Wang. Video summarization using fully convolutional sequence networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 347–363, 2018.

[27] Marina Sokolova, Nathalie Japkowicz, and Stan Szpakowicz. Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In *Australasian joint conference on artificial intelligence*, pages 1015–1021. Springer, 2006.

[28] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5179–5187, 2015.

[29] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[30] Felicia Elfrida Tjhai and Ombretta Strafforello. Evaluating the supervised video summarization model vasnet on an action localization dataset. 2021.

[31] Georgi Trevnenski, Ombretta Strafforello, and Seyran Khademi. Evaluation of the sum-gan-aae method for video summarization. 2021.

[32] Ting Yao, Tao Mei, and Yong Rui. Highlight detection with pairwise deep ranking for first-person video summarization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 982–990, 2016.

[33] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *International Journal of Computer Vision*, 2017.

[34] YouTube. Youtube by the numbers. https://blog.youtube/press/. [Online; accessed 19-April-2021].

[35] Li Yuan, Francis EH Tay, Ping Li, Li Zhou, and Jiashi Feng. Cycle-sum: cycle-consistent adversarial lstm networks for unsupervised video summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9143–9150, 2019.

[36] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *European conference on computer vision*, pages 766–782. Springer, 2016.

[37] Wencheng Zhu, Jiwen Lu, Jiahao Li, and Jie Zhou. Dsnet: A flexible detect-to-summarize network for video summarization. *IEEE Transactions on Image Processing*, 30:948–962, 2020.

[38] Daniel Zwillinger and Stephen Kokoska. *CRC standard probability and statistics tables and formulae*. Crc Press, 1999.