Evaluation of Video Summarization using DSNet and Action Localization Datasets

Daan Groenewegen Ombretta Strafforello
TU Delft
d.h.e.groenewegen@student.tudelft.nl, o.strafforello@tudelft.nl

Abstract

In this paper, the DSNet framework used for automatic video summarization gets reviewed when using action localization datasets. The problem facing video summarizations using deep learning techniques is that datasets can be subjective depending on preferences of human annotators, making for noise in the labeling. This paper will look at a anchor-based approach and anchor-free approach which were introduced by the DSNet framework. More specific it will evaluate in experiments using different hyper-parameters if these approaches gain an increased performances when using action localization datasets instead. These results will show the increase in accuracy when using action localization datasets. Moreover it will compare the different approaches, meaning anchor-based and anchor-free, and see if they still have comparable performance with the method.

1 Introduction

With an increasing amount of video footage being recorded every day in current society, the question if we can automate parts of processing this video stream becomes more enticing to solve. Particularly the question if long videos can be quickly summarized and cut in length with the help of deep learning algorithms and not lose importance or information. There have been previous attempts into solving this matter, where some were using supervised and unsupervised methods to summarize video footage. One particular framework for creating supervised video summarizations is DSNet [12], which is the latest new supervised video summarization framework. This supervised framework and other supervised [1] and unsupervised frameworks [4][11] have been evaluated and benchmarked using commonly used publicly available human-annotated datasets, such as TVSum [7] and SumMe [3]. However, there is a problem with subjectivity within these datasets, which might lower the accuracy of the summarization. On some datasets the unsupervised methods outperform the supervised methods in terms of accuracy evaluated using F1-score. This subjectivity can cause discriminative labels during the annotation of the dataset which causes an unclear ground truth, since one can imagine not all annotators agree on which part is important or how important each part is, which in turn leaves noise on these labels [6].

To overcome this problem of subjectivity, a proposal has been made to use action localization datasets, such as the Breakfast Actions dataset [5] and MultiTHUMOS dataset [10].

This paper will look at the DSNet Framework and evaluate its results using action localization datasets instead to quantify the effect of supervision on video summarization without the noise that human annotated datasets bring. To answer this, the research will use the same evaluation methods as previous experiments using F1-score to benchmark the DSNet framework. The results will be compared and out of this research a conclusion about the effect of subjectivity in supervised deep learning methods can be reached.

2 Methodology

For the research, the DSNet framework will be trained using the publicly available Breakfast Actions dataset extended with new annotations by the research team.. The DSNet framework contains two approaches, the anchor-based approach and the anchor-free approach. The anchor-based approach produces interest proposals at each temporal location with multi-scale durations which enables it to handle the length variations of interests. This however makes the approach sensitive to interest proposals and hyper-parameters [12]. The anchor-free approach tries to directly predict an importance score at each temporal location to prevent these sensitivities. Earlier experiments resulted in comparable results between the two approaches [12].

The experiments with the new datasets will compare the two approaches and checks the difference between them. Moreover, there will also be experiments involved in using different temporal modeling layers, LSTM and Bi-LSTM to see the effects it has for the DSNet. Next there will also be parameter analysis for the loss functions in both approaches as well as the use of different loss functions on the anchor-free approach.

2.1 Evaluation metrics

As evaluation metric, the F1-score between the generated summary and the corresponding summary is used since this has been used in most benchmarks as well. Let x_i be the binary label of the i-th frame in the predicted summary where $x_i \in \{0, 1\}$ and y_i the binary label of the i-th frame in the ground truth where $y_i \in \{0, 1\}$. Then the F1-score is computed in the following way with N as the total amount of frames in the video:

$$F1 = \frac{2 * Precision * Recall}{precision + recall} \tag{1}$$

Where

$$Precision = \frac{\sum_{i=1}^{N} x_i \cdot y_i}{\sum_{i=1}^{N} x_i} \; Recall = \frac{\sum_{i=1}^{N} x_i \cdot y_i}{\sum_{i=1}^{N} y_i} \tag{2}$$

However, according to Otani et al [6], F1-score can be a misleading evaluation metric to use, because in most cases it turned out that randomly generated summaries were able to reach similar or even better performance scores than that of human annotators. They propose a different method using the rank order correlation measures. This ranks the video frames according to generated importance scores and human annotated reference scores, then comparing the generated ranking with respect to each ground truth ranking. Finally, the correlation score is obtained by averaging over the individual results.

2.2 Comparison to other methods

Similar research has been performed on the Breakfast Actions dataset using different deep learning video summarization models, both supervised and unsupervised. These models are the supervised VASNet [1][8], both the supervised and unsupervised version of FCSN [2] and the unsupervised SUM-GAN-AAE [9]. These methods will be compared using the same evaluation metrics as the metrics of this research and will therefore put the results of using the Breakfasts Actions dataset in perspective when compared with other methods using the same dataset.

3 Experiments

3.1 Anchor-based approach

Various experiments are conducted using different values for the hyper-parameters. As default (canonical), the hyper-parameter λ has a set value of 1 and the threshold of the non-maximum suppression (NMS) has a set value of 0.5. Every experiment performs 5 training runs of 300 epochs and calculates the average F1-score between the runs. The algorithm is implemented using PyTorch and the experiments were done using CUDA accelerated dependencies on a NVidia RTX 2060 GPU.

3.1.1 Canonical experiments

Firstly, experiments getting the results for the canonical (C) values are being performed. This is done to compare the F1-score with TVSum [7] and SumMe [3] with the F1-score retrieved from using the Breakfast Actions dataset [5]. For comparison reasons, the TVSum and SumMe datasets will also run in augmented (A) and transfer settings (T). However, the Breakfast dataset does not provide these settings due to the limited size. Therefore these results will not be available. The results of the experiments are shown in Table 1.

Dataset	C	A	T
TVSum	62.2	63.9	59.6
SumMe	50.3	49.5	46.5
Breakfast	64.6	-	-

Table 1: Comparison of F1-score between datasets

3.1.2 Parameter Analysis

To further analyse and do comparisons with the previous research, experiments using different values for λ and the NMS threshold were done. These were also performed using other temporal modeling layers on the datasets such as LSTM and Bi-LSTM.

3.2 Anchor-free approach

Just as with the anchor-based approach, various experiments are conducted using different values for the hyper-parameters. As default (canonical), the hyper-parameter λ has a set value of 1, the new balance hyper-parameter μ defined in the loss function also has a set

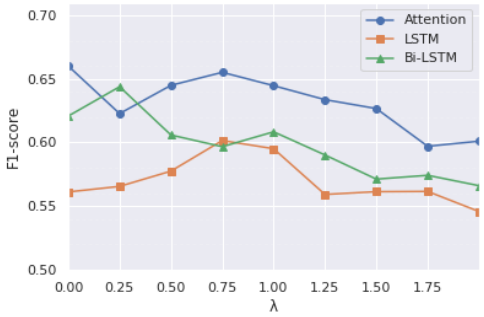

Figure 1: Parameter analysis of λ for Breakfast Dataset

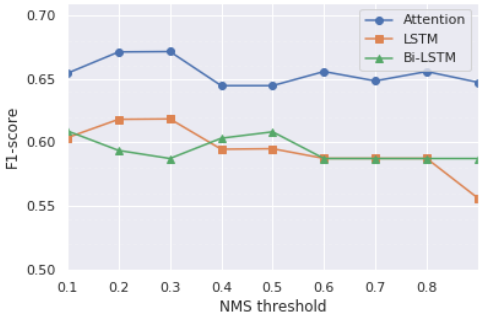

Figure 2: Parameter analysis of NMS threshold for Breakfast Dataset

value of 1 and the threshold of the non-maximum suppression (NMS) has a set value of 0.4 to compare it to the research done previously. Every experiment performs 5 training runs of 300 epochs and calculates the average F1-score between the runs. The algorithm is implemented using PyTorch and the experiments were done using CUDA accelerated dependencies on a NVidia RTX 2060 GPU.

3.2.1 Canonical experiments

The canonical settings are the ones as mentioned earlier. As mentioned previously at the anchor-based approach, the Breakfast Actions dataset is not suitable to perform augmented and transfer settings on it and therefore those results are missing. The results are in Table 2.

Dataset	C	A	T
TVSum	59.6	62.4	58.0
SumMe	50.8	51.9	47.6
Breakfast	60.0	-	-

Table 2: Comparison of F1-score between datasets on anchor-free approach

3.2.2 Parameter Analysis

The loss functions used in the anchor-free approach contain two parameters that can be adjusted, namely the λ and μ, which both range from 0.25 to 2.0. To evaluate the results, every possible combination of these parameters were used in the experiments and the results have been plotted in Figure 3.

Furthermore, the effects in regards to the NMS threshold which influences the filtering of redundant segments has also been evaluated. These values range from 0.1 to 0.9 and the results are found in Figure 4

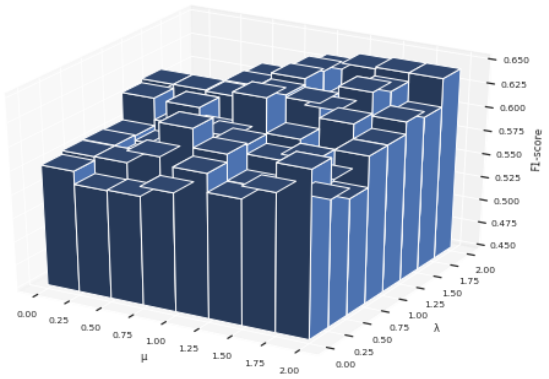

Figure 3: Parameter analysis of λ and μ in anchor-free approach

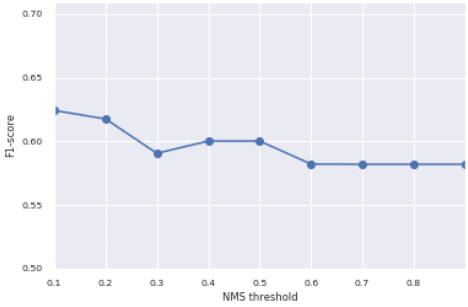

Figure 4: Parameter analysis of NMS in anchor-free approach

3.3 Ranking correlation

The last experiments done were looking at the evaluation metric proposed by Otani et al [6], which include the rank correlation coefficients of Kendall's τ and Spearman's ρ together with a visualization of some videos. The correlation values are calculated between the predicted importance scores and the ground truth importance score. First, the DSNet was evaluated over the three datasets, for which the results are in Table 3. Secondly, the results of the Breakfast Actions dataset are compared with other methods of video summarization and the results are in Table 4 with extracts of the visualization of both approaches on the Breakfast Action dataset found in Figure 5 and Figure 6.

Dataset	F1 score	Spearman's ρ	Kendall's τ
TVSum (AB)	0.622	0.285	0.198
TVSum (AF)	0.596	0.197	0.276
SumMe (AB)	0.503	0.035	0.041
SumMe (AF)	0.508	0.048	0.062
Breakfast (AB)	0.6446	0.106	0.090
Breakfast (AF)	0.6003	0.078	0.056

Table 3: Comparison between different datasets using DSNet anchor-based (AB) and anchor-free (AF)

Model	F1 score	Spearman's ρ	Kendall's τ
DSNet (AB)	0.6446	0.106	0.090
DSNet (AF)	0.6003	0.078	0.056
VASNet [8]	0.673	0.045	0.0365
FCSN[2]	0.314	0.032	0.024
FCSN_{unsup}[2]	0.201	-0.021	-0.020
SUM-GAN-AAE[9]	0.5138	-0.03	-0.03

Table 4: Comparison between different video summarization models on the Breakfast Actions dataset using Canonical Settings

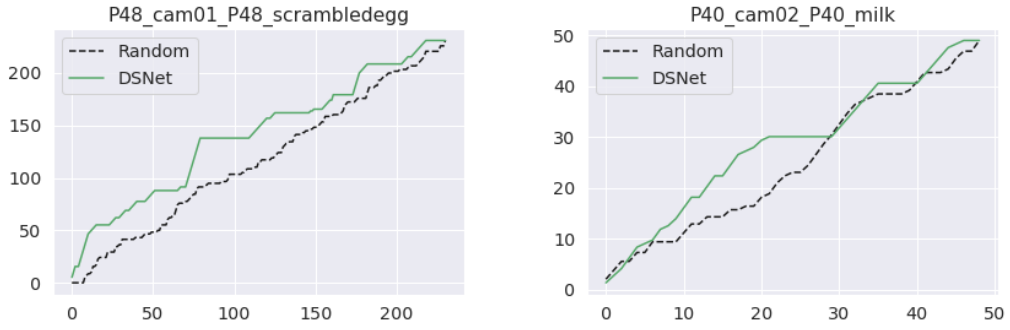

Figure 5: Extracts of the correlation graphs produced using anchor-based approach

4 Responsible Research

Most of the research done in this paper built further upon implementations already done by the original authors of the DSNet [12]. First changes made were not in the implementation, but in the dependencies of the project itself, since some packages were not available anymore or were not suitable for the graphics card used to run the framework. To verify that the DSNet implementation used in this research was the same as the implementation of the original authors, the results of the original paper were verified first by performing experiments using the same parameters on the implementation used in this research. Then when

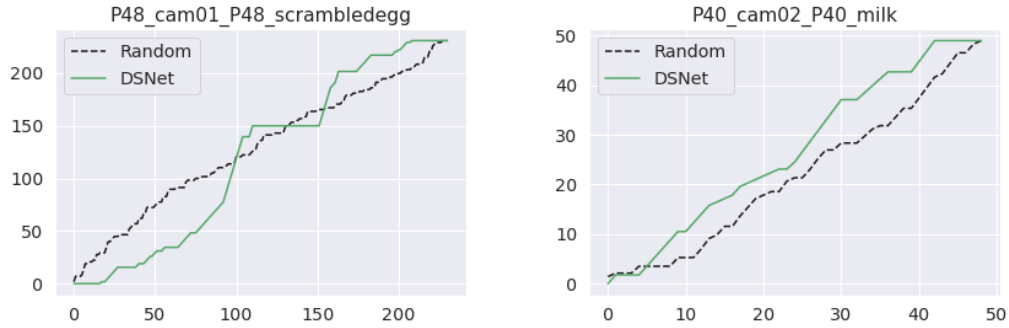

Figure 6: Extracts of the correlation graphs produced using anchor-free approach

comparing the previous research with the new research, only the values of the verification of the previous research were used to compensate for any differences there might be.

For transparency, all the trained models and scripts used for training will be made available online in a download link when contacting the author. The Breakfast Action dataset used in the research is available through the website of the original authors and completely free to use. However, the extended version on which the models were trained is not widely available and is therefore harder to use when evaluating this work. The only changes made were to suit the format of the framework itself, in which no important data was lost.

5 Discussion

5.1 Anchor-based

From the experiments, it is concluded that the F1-score is very similar when using the Breakfast Actions dataset compared to other methods previously proposed using TVSum and SumMe datasets. Moreover an interesting development is that when analysing the effect of the hyper-parameter λ, it is inconclusive in regards with the different temporal modeling layers used. When $\lambda = 0$ with Attention, the F1-score is maximum, indicating that the regression item in the loss function does decrease the accuracy,.When evaluating on Bi-LSTM, it shows similar results with a decrease in accuracy when the hyper-parameter is increased. However when looking at LSTM, the maximum values are found around $\lambda = 1.0$, which indicates an almost equal importance classification and regression branches. Another conclusion that can be made is that the F1-score fluctuates more when changing hyper-parameters during the use of the Breakfast Actions dataset compared to TVSum and SumMe. This indicates that the model is more sensitive to changes in these parameters. Overall, the best method is to use the Attention layer.

Furthermore, the effect of different NMS thresholds is minimal. This can be explained as the accuracy of the system being high enough that there is not a lot of segments that are missed in the process. Therefore the NMS threshold has minimal effect on the overall accuracy of the system.

5.2 Anchor-free

When evaluated, the anchor-free approach using the canonical settings provide similar results compared to the other research when using canonical settings. Experiments on the parameters λ and μ show that the anchor-free approach becomes very sensitive to changes in these parameters when using the Breakfast Action dataset as shown in Figure 3. There is no clear difference between a dominant effect of either hyper-parameter, but it does influence the F1-score. Mainly it can be concluded that the regression item and the center-ness constraint of the loss function contribute largely to an increase in accuracy in terms of F1-score.

The NMS threshold analysis gave interesting results in the fact that it indicated a trend of a decrease the larger the value of the threshold became, indicating that the low threshold does not introduce low-quality segments, thus evading the need of having a high threshold. Previous canonical settings resorted in using a value of 0.4 in the anchor-free approach for this threshold, but now it seems that 0.1 is the best value.

5.3 Ranking correlation

The correlation coefficients show that the DSNet framework is in both approaches weakly positively correlated to the ground truth scores, but still not that greatly. Indeed, when compared with other methods that used the Breakfast Actions dataset, the DSNet has the largest correlation coefficient of all, but still will not outperform the TVSum, which indicates that while producing similar results in F1-score, TVSum actually is better performing in terms of correlation score. When looking at the graphs, you can see the positive correlation between the performance of DSNet versus random scoring. The curve of DSNet stays above the random baseline created, which indicates positive correlation with the ground truth compared to random scoring.

6 Conclusions and Future Work

Using action localized datasets for DSNet framework will barely increase its accuracy. The research conducted found a small increase in F1-score when tweaking with hyper-parameters, but from evaluation of the correlation coefficients, the Breakfast Actions dataset does not outperform TVSum. Moreover, just like the previous research [12] on the DSNet framework found similar results between anchor-based and anchor-free methods using TVSum and SumMe, when using action localization datasets the difference between results becomes greater in favor of using the anchor-based approach when using the same parameters. However, when changing the hyper-parameters, this difference is quickly gone and they again perform can achieve similar results.

For further research on the use of action localization datasets, the MultiTHUMOS dataset [10] might be used to train the network on. This dataset is different than the Breakfast Actions dataset and might provide other results. Furthermore, the Breakfast Actions dataset can be appended with more data, which was already done partly in this research but can certainly be expanded upon.

References

[1] Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Summarizing videos with attention. *CoRR*, abs/1812.01969, 2018.

[2] P. Frolke and O. Strafforello. Evaluating of summarization using fully convolutional sequence networks on action localization datasets. 2021.

[3] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 505–520, Cham, 2014. Springer International Publishing.

[4] Yunjae Jung, Donghyeon Cho, Dahun Kim, Sanghyun Woo, and In So Kweon. Discriminative feature learning for unsupervised video summarization. *CoRR*, abs/1811.09791, 2018.

[5] H. Kuehne, A. B. Arslan, and T. Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of Computer Vision and Pattern Recognition Conference (CVPR)*, 2014.

[6] 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 (CVPR)*, June 2019.

[7] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5179–5187, 2015.

[8] F.E. Tjhai and O. Strafforello. Evaluating the supervised video summarization model vasnet on an action localization dataset. 2021.

[9] F.E. Tjhai and O. Strafforello. Evaluation of the sum-gan-aae method for video summarization. 2021.

[10] 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.

[11] Li Yuan, Francis E. H. Tay, Ping Li, Li Zhou, and Jiashi Feng. Cycle-sum: Cycle-consistent adversarial LSTM networks for unsupervised video summarization. *CoRR*, abs/1904.08265, 2019.

[12] 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.