Explaining video summarization based on the focus of attention

Evlampios Apostolidis
CERTH-ITI &
Queen Mary University of London
Thessaloniki, Greece, 57001
Email: apostolid@iti.gr

Georgios Balaouras, Vasileios Mezaris
CERTH-ITI
Thessaloniki, Greece, 57001
Email: {mpalaourg, bmezaris}@iti.gr

Ioannis Patras
Queen Mary University of London
London, UK, E14NS
Email: i.patras@qmul.ac.uk

***Abstract*—In this paper we propose a method for explaining video summarization. We start by formulating the problem as the creation of an explanation mask which indicates the parts of the video that influenced the most the estimates of a video summarization network, about the frames' importance. Then, we explain how the typical analysis pipeline of attention-based networks for video summarization can be used to define explanation signals, and we examine various attention-based signals that have been studied as explanations in the NLP domain. We evaluate the performance of these signals by investigating the video summarization network's input-output relationship according to different replacement functions, and utilizing measures that quantify the capability of explanations to spot the most and least influential parts of a video. We run experiments using an attention-based network (CA-SUM) and two datasets (SumMe and TVSum) for video summarization. Our evaluations indicate the advanced performance of explanations formed using the inherent attention weights, and demonstrate the ability of our method to explain the video summarization results using clues about the focus of the attention mechanism.**

***Index Terms*—Explainable AI, Video summarization, Attention mechanism, Evaluation measures**

I. Introduction

Video summarization is a problem in the domain of video analysis and understanding, that increasingly gains attention over the last years [1]. Technologies for video summarization aim to generate a short synopsis by selecting the most informative and important parts of the video. As the production of a video summary is time-consuming, the use of such technologies can drastically reduce the needed resources in terms of both time and human effort. Nevertheless, the outcome of theses technologies needs to be curated by a video editor, to ensure that all the needed parts have been included in the summary. This content production step could be further facilitated, if the editor is provided with explanations about the suggestions of the used technology. The provision of such explanations would allow a level of understanding about the functionality of this technology, thus increasing the editor's trust in it and reducing the needed time for content curation.

A few attempts have been made for explaining the outcomes of deep networks processing video data. However, they are related with networks for action/event recognition [2]–[5] and video classification [6]–[8], and their application on networks for video summarization is not a straightforward task. To the best of our knowledge, our work is the first on explainable video summarization. Our contributions are as follows:

- We introduce the problem of explaining video summarization and formulate it as the production of an explanation mask indicating the most influential parts of the video, for the output of a video summarization network.
- We describe how the typical processing pipeline of attention-based video summarization networks can be used to extract attention-based explanation signals, and examine various relevant signals from the NLP domain.
- We propose evaluation measures that quantify the ability of explanations to spot the most and least influential parts of a video, and evaluate the considered explanation signals using an attention-based network and two datasets for video summarization.

II. Related Work

Nowadays there is a growing interest on methods providing explanations about the working mechanism or the predictions of neural networks. A lot of progress has been made in the domains of pattern recognition [9] and natural language processing [10]. However, only a few works deal with network architectures processing video data. Aakur et al. [2] built a framework for producing inherently explainable and semantically coherent representations for video activity interpretation. Zhuo et al. [3], defined a spatio-temporal graph of semantic-level video states and applied state transition analysis for video action reasoning. Stergiou et al. [4], formed explanations of deep networks for action classification and recognition, using cylindrical heatmaps that visualize the focus of attention. Gkalelis et al. [5], used the weighted in-degrees of graph attention networks' adjacency matrices to provide explanations of video event recognition, in terms of salient objects and frames. Mänttäri et al. [6] extended the concept of meaningful perturbation, to spot the video fragment with the greatest impact on the video classification results. Bargal et al. [7], visualized the spatio-temporal cues contributing to a network's classification/captioning output using internal representations, and employed these cues to localize video fragments corresponding to a specific action

This work was supported by the EU Horizon 2020 programme under grant agreement H2020-951911 AI4Media.

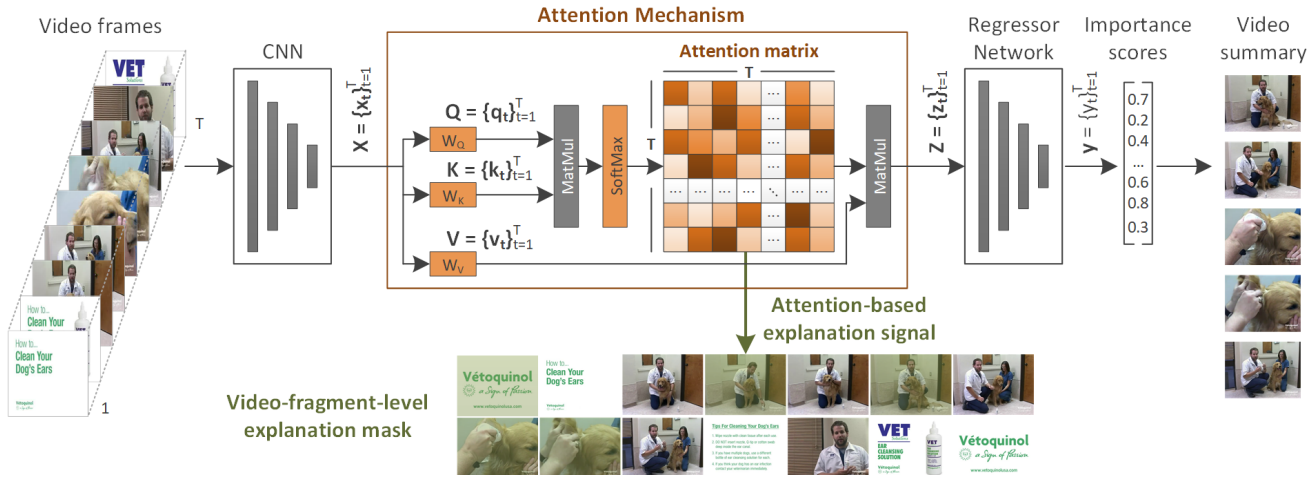

Fig. 1. Overview of our concept for obtaining attention-based explanations of the video summarization results. The different video fragments are illustrated using their most representative frame and appear in a "left-to-right then top-to-bottom" order. The number of highlighted video fragments M in the produced explanation mask equals to five. The video summary is formed by stitching the top-5 fragments (according to their importance) in chronological order.

or phrase from the caption. Finally, Li et al. [8], extended a generic perturbation-based explanation method for video classification networks, by introducing a loss function that constraints the smoothness of explanations in both spatial and temporal dimensions. Differently to the above, in this work we focus on networks for video summarization. To the best of our knowledge, this is the first work that does so. In terms of methodology, our method is mostly closely related with the approaches in [6], [8], which obtain explanations via a perturbation-based investigation of the network's input-output relationship. However, contrary to these approaches, we form explanations using several attention-based signals used in the NLP domain [11]–[16], and we examine the network's input-output relationship based on various replacement functions.

III. Explaining Video Summarization

A. Problem formulation

Let's assume a video summarization network that gets as input a set of deep feature vectors $\boldsymbol{X} = \{\boldsymbol{x_t}\}_{t=1}^{T}$ representing the T frames of a video, and produces in the output a set of frame-level scores $\boldsymbol{y} = \{y_t\}_{t=1}^{T}$ that lie in the range $\mathbb{I} = [0, 1]$ and quantify the visual importance of each video frame. The goal of an explanation method is to derive a frame-based visualization of the video content at the fragment-level (called video-fragment-level explanation mask in the following), highlighting the top-M video fragments that influenced the most the decisions of a video summarization network about the frames' importance, and thus the generation of the video summary. To avoid the influence of video fragmentation and key-fragment selection steps to the created video summary (discussed in [1]), in this work we adopt a more straightforward approach to form the summary; we split the video into consecutive and non-overlapping fragments of fixed-size L, we compute each fragment's importance by averaging the scores of the frames in it, and we pick the M top-scoring video fragments.

B. Attention-based explainable video summarization

1) Preliminaries and assumptions: In this study we assume video summarization networks that rely on a self-attention mechanism, such as the ones in [17]–[19]. The typical processing pipeline of these networks is depicted in the upper part of Fig. 1. Given a video of T frames and a pre-trained CNN model for deep feature extraction, the attention mechanism gets as input the frames' feature representations $\boldsymbol{X} = \{\boldsymbol{x_i}\}_{i=1}^{T}$. Following, it produces the Query- and Key-based transformations of them ($\boldsymbol{Q} = \{\boldsymbol{q_i}\}_{i=1}^{T}$ and $\boldsymbol{K} = \{\boldsymbol{k_i}\}_{i=1}^{T}$, respectively), performs a matrix multiplication ($\boldsymbol{Q} \times \boldsymbol{K}^{-1}$, where $\boldsymbol{K}^{-1}$ is the transposed version of $\boldsymbol{K}$), and applies a softmax conversion on the computed values. Through this process, it forms a $T \times T$ matrix of attention weights $\boldsymbol{A} = \{a_{i,j}\}_{i,j=1}^{T}$, with $a_{i,j} \in \mathbb{I}$. Each row of this matrix corresponds to a different frame of the video and the values in each row represent the significance of the associated frame for each frame of the video based on the context modeled by the attention mechanism. This matrix is multiplied with the Value-based transformation of the input feature representations ($\boldsymbol{V} = \{\boldsymbol{v_i}\}_{i=1}^{T}$) and forms the output of the attention mechanism; i.e., a new set of representations ($\boldsymbol{Z} = \{\boldsymbol{z_t}\}_{t=1}^{T}$) that convey information about the relevance of each video frame with the modeled video context. This output goes through a Regressor Network, which produces the frames' importance scores $\boldsymbol{y}$ that are finally used to compute fragment-level importance and select the most important fragments for inclusion in the video summary. As also shown in Fig. 1, the attention matrix is the basis for producing an attention-based explanation signal, which is used to spot the most influential video fragments for the network's predictions and construct the explanation mask.

2) Explanation signals: Inspired by existing works on attention-based explanation of NLP models [11]–[15], we take into account the following explanation signals:

- **Inherent Attention (IA)** is formed using the weights in the main diagonal of the attention matrix $\{a_{i,i}\}_{i=1}^{T}$.

- **Gradient of Attention (GoA)** is formed using the gradients of the final layer with respect to the weights in the main diagonal of the attention matrix $\{\boldsymbol{\nabla} a_{i,i}\}_{i=1}^{T}$.
- **Grad Attention (GA)** is formed using a gradient-based weighted version of the weights in the main diagonal of the attention matrix $\{a_{i,i} \odot \boldsymbol{\nabla} a_{i,i}\}_{i=1}^{T}$.
- **Input Norm Attention (NA)** is formed using a weighted version of the weights in the main diagonal of the attention matrix, according to the norm of the Value-based transformed input vectors $\{a_{i,i} \odot ||\boldsymbol{v_i}||\}_{i=1}^{T}$.
- **Input Norm Grad Attention (NGA)** is a combination of GA and NA, as it is formed using a gradient- and norm-based weighted version of the weights in the main diagonal of the attention matrix $\{a_{i,i} \odot \boldsymbol{\nabla} a_{i,i} \odot ||\boldsymbol{v_i}||\}_{i=1}^{T}$).

The above define frame-level explanation signals; then, the explanation masks are constructed by computing fragment-level explanation scores (by averaging the relevant frames' scores) and selecting the M fragments with the highest scores.

3) Replacement functions: Based on works from the NLP domain [12], [15], to investigate the network's input-output relationship we apply the following replacement functions on parts of the input corresponding to different video fragments:

- **Slice-out** completely removes the specified part.
- **Input Mask** replaces the specified part with a mask composed of black/white frames' feature representations.
- **Randomization** replaces 50% of the elements of each feature representation within the specified part, using the corresponding elements from randomly-selected feature representations from the remaining part of the input.
- **Attention Mask** sets the attention weights associated with the specified part equal to zero, such that this part will not be forwarded in the network anymore.

4) Evaluation measures: For each replacement function, we measure the influence of each video fragment in the network's output, by computing the difference of estimates ΔE:

$$\Delta E(\boldsymbol{X}, \hat{\boldsymbol{X}}^{\boldsymbol{k}}) = \tau(\boldsymbol{y}, \boldsymbol{y}^{\boldsymbol{k}}) \tag{1}$$

where, $\boldsymbol{X}$ is the original set of feature representations, $\hat{\boldsymbol{X}}^{\boldsymbol{k}}$ is the updated set after replacing the features of the frames belonging to the k^{th} fragment, $\boldsymbol{y}$ and $\boldsymbol{y}^{\boldsymbol{k}}$ are the outputs of the summarization network for $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}^{\boldsymbol{k}}$, respectively, and τ is the Kendall's τ correlation coefficient. Based on the difference of estimates ΔE, we assess the performance of each explanation signal using the following evaluation measures.

Discoverability+ (D^+) evaluates if fragments with higher explanation scores have a significant influence to the estimated importance scores, and thus are necessary for the network's predictions. Following an approach similar to the one applied in [15] for measuring Sufficiency and Comprehensiveness of explanations, we calculate D^+ as the mean of the obtained ΔE values after sequentially replacing parts of the input corresponding to the top-1%, 5%, 10%, 15%, 20% of the fragments with the highest explanation scores (i.e., we affect multiple parts in a batch manner). Moreover, since the explanation mask focuses on the top-M most influential fragments of the video, we compute this measure also as the mean of the obtained ΔE values after sequentially replacing parts of the input corresponding to the M fragments with the highest explanation scores (i.e., we affect parts in a one-by-one manner).

Discoverability- (D^-) evaluates if fragments with lower explanation scores have small influence to the estimated importance scores, and thus are less necessary for the network's predictions. In analogous to D^+, we compute D^- as the mean of the obtained ΔE values after sequentially replacing parts of the input corresponding to the top-1%, 5%, 10%, 15%, 20% of the fragments with the lowest explanation scores. Moreover, we compute D^- as the mean of the obtained ΔE values after sequentially replacing parts of the input corresponding to the M fragments with the lowest explanation scores.

Sanity Violation (SV) quantifies the ability of explanations to correctly discriminate important from unimportant video fragments. We compute SV by counting the number of cases where the condition $(D^+ > D^-)$ is violated after sequentially replacing parts of the input corresponding to the top-1%, 5%, 10%, 15%, 20% of the fragments with the highest and lowest explanation scores, and expressing the computed value as a fraction of the total number of replacements. Moreover, we compute SV after sequentially replacing parts of the input corresponding to the M top- and less-scoring fragments in a pair-wise manner (e.g., the 1^{st} top- and less-scoring fragment).

Rank Correlation (RC) measures the correlation between the assigned explanation scores to the video fragments and the obtained ΔE values after sequentially replacing each one of them. Following [11], we quantify RC by computing the Spearman's ρ rank correlation coefficient. Since a negative correlation can be observed in some cases, to measure the average RC score over different pre-trained models and replacement functions, first we apply a Fisher transformation on the computed ρ values, then we average them in the new space, and finally we apply a reverse Fisher transformation.

IV. Experiments

A. Datasets and implementation details

Our evaluations are made on two datasets for video summarization. SumMe [20] contains 25 videos (1-6 min.) covering multiple events from both first-person and third-person view. TVSum [21] is composed of 50 videos (1-11 min.) from 10 categories of the TRECVid MED dataset. Videos are downsampled to 2 fps and the sampled frames are represented using the output of the pool5 layer of GoogleNet trained on ImageNet. The parameter M, that indicates the number of highlighted video fragments in the explanation mask, is set equal to five. The size L of the video fragments is set equal to 10 sec. The explanation signals are evaluated using pre-trained models (available at: https://zenodo.org/record/6562992) of the CA-SUM method for video summarization [17]. In the following, we report the average scores over these runs. To allow the reproduction of our results, the PyTorch code is publicly-available at: https://github.com/e-apostolidis/XAI-SUM.

TABLE I

PERFORMANCE OF THE CONSIDERED EXPLANATION SIGNALS ON THE SUMME AND TVSUM DATASETS, AFTER REPLACING PARTS OF THE INPUT IN A BATCH AND IN A ONE-BY-ONE MANNER. THE ARROWS INDICATE THE OPTIMAL (MINIMUM OR MAXIMUM) VALUE FOR EACH MEASURE

	SumMe										TVSum									
	Batch					One-by-One					Batch					One-by-One				
	IA	NA	GA	GoA	NGA	IA	NA	GA	GoA	NGA	IA	NA	GA	GoA	NGA	IA	NA	GA	GoA	NGA
Slice Out																				
D^- (↓)	0.193	0.198	0.144	**0.133**	0.145	0.173	0.170	0.163	0.163	**0.161**	**0.098**	0.099	0.113	0.107	0.113	**0.075**	**0.075**	0.097	0.095	0.098
D^+ (↑)	**0.204**	0.193	0.192	0.192	0.192	0.178	0.177	0.183	0.182	**0.185**	**0.127**	0.126	0.083	0.082	0.082	0.102	**0.103**	0.071	0.071	0.071
SV (↓)	0.440	0.480	**0.360**	**0.360**	**0.360**	0.520	0.440	0.240	0.280	**0.160**	**0.260**	0.300	0.820	0.780	0.820	**0.160**	**0.160**	0.820	0.760	0.800
RC (↑)	N/A	N/A	N/A	N/A	N/A	0.056	**0.058**	-0.278	-0.313	-0.285	N/A	N/A	N/A	N/A	N/A	0.216	**0.228**	-0.174	-0.217	-0.171
Input Mask (black frame)																				
D^- (↓)	**0.224**	0.230	0.289	0.279	0.292	**0.163**	0.164	0.207	0.207	0.206	**0.134**	0.138	0.337	0.334	0.336	**0.088**	**0.088**	0.152	0.151	0.151
D^+ (↑)	**0.345**	0.321	0.211	0.208	0.210	**0.211**	0.205	0.179	0.178	0.177	**0.266**	0.263	0.100	0.101	0.100	**0.130**	**0.130**	0.070	0.070	0.070
SV (↓)	**0.160**	0.200	0.720	0.760	0.760	**0.120**	0.240	0.760	0.720	0.800	**0.180**	0.220	0.980	0.980	0.980	**0.160**	0.220	0.960	0.960	0.960
RC (↑)	N/A	N/A	N/A	N/A	N/A	**0.369**	0.296	0.083	0.036	0.050	N/A	N/A	N/A	N/A	N/A	**0.397**	0.382	-0.103	-0.218	-0.110
Input Mask (white frame)																				
D^- (↓)	**0.208**	0.214	0.260	0.256	0.262	**0.142**	0.143	0.174	0.174	0.172	**0.126**	0.130	0.351	0.347	0.350	**0.083**	**0.083**	0.156	0.154	0.155
D^+ (↑)	**0.286**	0.279	0.189	0.186	0.188	**0.177**	0.172	0.150	0.150	0.149	**0.289**	0.287	0.094	0.092	0.094	**0.137**	0.136	0.067	0.067	0.067
SV (↓)	0.280	**0.240**	0.640	0.640	0.680	**0.280**	**0.280**	0.680	0.640	0.680	**0.100**	0.140	0.980	0.980	0.980	**0.140**	**0.140**	0.980	1.000	0.980
RC (↑)	N/A	N/A	N/A	N/A	N/A	**0.273**	0.199	0.027	-0.008	0.010	N/A	N/A	N/A	N/A	N/A	**0.455**	0.439	-0.162	-0.289	-0.166
Randomization																				
D^- (↓)	0.131	**0.125**	0.140	0.134	0.136	**0.077**	0.079	0.081	0.082	0.080	**0.079**	0.078	0.115	0.108	0.116	**0.040**	0.041	0.051	0.051	0.052
D^+ (↑)	**0.149**	0.141	0.134	0.135	0.137	**0.084**	**0.084**	0.077	0.076	0.076	**0.111**	0.109	0.079	0.081	0.078	0.052	**0.053**	0.040	0.041	0.042
SV (↓)	**0.440**	**0.440**	**0.440**	**0.440**	**0.440**	**0.400**	**0.400**	0.520	0.560	0.520	**0.160**	0.200	0.880	0.760	0.860	0.280	**0.240**	0.760	0.740	0.700
RC (↑)	N/A	N/A	N/A	N/A	N/A	**0.138**	-0.014	0.102	-0.066	-0.003	N/A	N/A	N/A	N/A	N/A	0.306	**0.324**	-0.029	-0.087	0.037
Attention Mask																				
D^- (↓)	**0.256**	0.258	0.258	0.258	0.259	**0.251**	0.252	0.255	0.255	0.255	**0.270**	0.271	0.288	0.285	0.287	**0.272**	**0.272**	0.284	0.285	0.284
D^+ (↑)	**0.261**	0.259	0.250	0.250	0.250	**0.256**	**0.256**	0.246	0.246	0.247	**0.291**	0.290	0.259	0.259	0.259	**0.287**	0.286	0.272	0.272	0.272
SV (↓)	**0.360**	0.400	0.600	0.600	0.600	0.440	**0.400**	0.680	0.680	0.640	**0.340**	0.400	0.620	0.620	0.620	**0.300**	**0.300**	0.580	0.520	0.600
RC (↑)	N/A	N/A	N/A	N/A	N/A	**0.029**	-0.009	0.023	0.027	0.019	N/A	N/A	N/A	N/A	N/A	**0.116**	0.111	-0.041	-0.076	-0.044
Overall (average)																				
D^- (↓)	**0.202**	0.205	0.218	0.212	0.219	**0.161**	0.162	0.176	0.176	0.175	**0.141**	0.143	0.241	0.236	0.241	**0.112**	**0.112**	0.148	0.147	0.148
D^+ (↑)	**0.249**	0.239	0.195	0.194	0.195	**0.181**	0.179	0.167	0.166	0.167	**0.217**	0.215	0.123	0.123	0.123	**0.142**	**0.142**	0.104	0.104	0.104
SV (↓)	**0.336**	0.352	0.552	0.560	0.568	**0.352**	**0.352**	0.576	0.576	0.560	**0.208**	0.252	0.856	0.824	0.852	**0.208**	0.212	0.820	0.796	0.808
RC (↑)	N/A	N/A	N/A	N/A	N/A	**0.176**	0.108	-0.010	-0.067	-0.043	N/A	N/A	N/A	N/A	N/A	**0.303**	0.301	-0.102	-0.179	-0.091

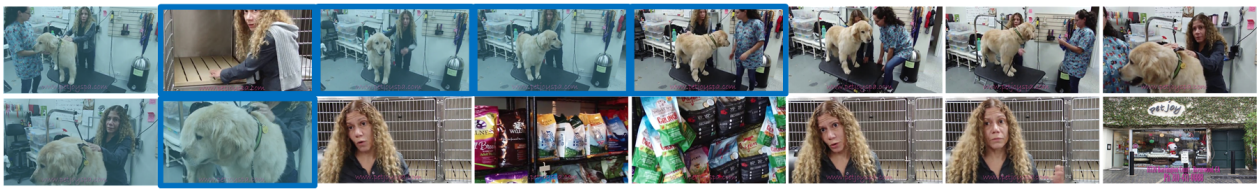

Fig. 2. The explanation mask and the five top-scoring fragments (in blue bounding boxes) for a TVSum video, titled "Pet Joy Spa Grooming Services".

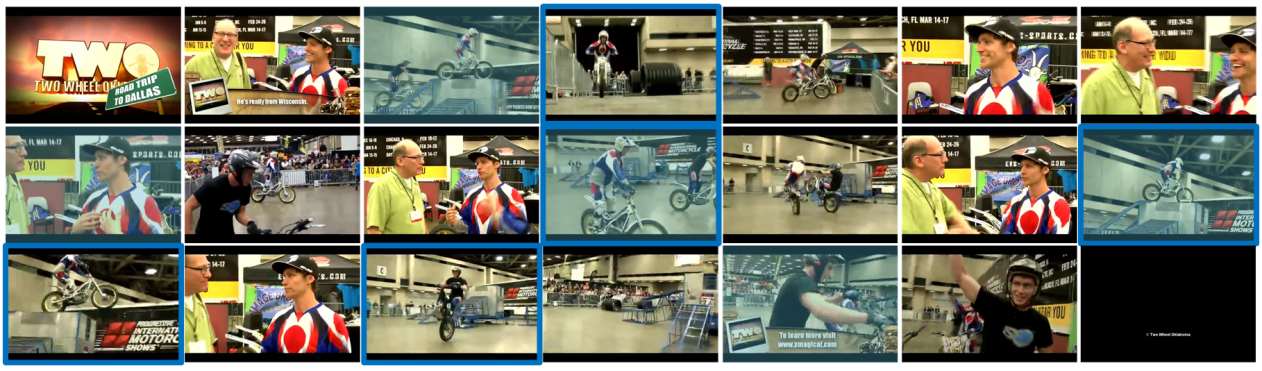

Fig. 3. The explanation mask and the five top-scoring fragments (in blue bounding boxes) for a TVSum video, titled "Smage Bros. Motorcycle Stunt Show".

B. Quantitative analysis

The performance of each explanation signal on the SumMe and TVSum dataset is presented in Table I. The results for batch replacement show the advanced performance of explanations formed using the inherent attention weights (IA). Such explanations exhibit the best performance for all replacement functions on TVSum, and for most of them on SumMe. On average, they achieve the lowest/highest D^-/D^+ scores and, in most cases (approx. 66% on SumMe and 80% on TVSum), correctly discriminate the most and least influential fragments of the video. Explanations formed using the norm-based weighted version of the inherent attention weights (NA) also perform good in terms of D^- and D^+, but are less effective in terms of SV. Finally, explanations formed using the gradients of the attention weights (GA, GoA, NGA) are by far the worst-performing ones. For most replacement functions, these explanations result in higher/lower D^-/D^+ scores than the ones obtained for non-gradient-based signals (IA, NA). Moreover, on average, they fail to distinguish the most and least influential video fragments in more than 56% and 82% of the cases on SumMe and TVSum, respectively.

The outcomes for one-by-one replacement indicate once again the effectiveness of explanations formed using the inherent attention weights (IA), and the competitiveness of explanations formed by combining these weights with the norm of Value-based transformed input vectors (NA). On average, the former explanations successfully pass the sanity violation test in approx. 65% and 80% of cases on SumMe and TVSum, respectively. Moreover, based on the computed rank correlation scores, they are capable of assigning fragment-level explanation scores that are more representative of each fragment's influence to the network's output. On the contrary, gradient-based explanation signals (GA, GoA, NGA) perform systematically worse. On average, they violate the sanity test in approx. 57% and 80% of the cases on SumMe and TVSum, respectively. Furthermore, they assign fragment-level explanation scores that are neutrally or negatively correlated with the influence of each fragment to the network's output.

The above indicate the use of inherent attention weights to form explanations for the CA-SUM model, as the best option; thus, such explanations were used in our qualitative analysis.

C. Qualitative analysis

Our qualitative analysis relies on the created explanation masks for two indicative videos of the TVSum dataset. In Fig. 2, we observe that the attention mechanism of CA-SUM pays more attention to video parts showing the dog, and less attention to speaking persons, dog products, and the pet store. So, it seems to focus on the dog and models the video's context based on it. Building on this knowledge, CA-SUM promotes parts of the video that are mainly associated with the dog, as 4 out of 5 top-scoring fragments contain instances of it. In Fig. 3, the attention mechanism concentrates mainly on parts of the video showing the tricks made by the riders of the motorbikes, and other video parts showing the logo of the TV-show and the interview, are less attractive. Based on this focus of attention, CA-SUM indicates the parts of the video showing the riders of the motorbikes doing tricks, as the most important ones. These paradigms show that extracting explanations using the proposed method and the inherent attention weights, could allow to get insights about the focus of the attention mechanism and assist the explanation of video summarization networks similar to CA-SUM.

V. Conclusions

In this work, we presented a method for explaining video summarization. After formulating the task, we described how attention-based video summarization networks can be used to extract explanations. Following, we considered various explanation signals used in the NLP domain, and introduced evaluation measures for assessing their ability to identify the most and least influential parts of the video, for the network's predictions. Using these measures and based on different replacement functions for investigating the network's input-output relationship, we assessed the performance of the considered explanations with the help of the CA-SUM network and the SumMe and TVSum datasets for video summarization. Our findings show that, using the proposed method to form explanations based on the inherent attention weights can lead to useful clues about the focus of the attention mechanism, which can assist the explanation of the summarization results.

References

[1] E. Apostolidis *et al.*, "Video summarization using deep neural networks: A survey," *Proc. of the IEEE*, vol. 109, no. 11, pp. 1838–1863, 2021.

[2] S. N. Aakur *et al.*, "An inherently explainable model for video activity interpretation," in *AAAI 2018*.

[3] T. Zhuo *et al.*, "Explainable video action reasoning via prior knowledge and state transitions," in *2019 ACM MM*.

[4] A. Stergiou *et al.*, "Saliency tubes: Visual explanations for spatio-temporal convolutions," in *IEEE ICIP 2019*.

[5] N. Gkalelis *et al.*, "ViGAT: Bottom-up event recognition and explanation in video using factorized graph attention network," *IEEE Access*, vol. 10, pp. 108 797–108 816, 2022.

[6] J. Mänttäri *et al.*, "Interpreting video features: A comparison of 3D conv. networks and conv. LSTM networks," in *ACCV 2020*.

[7] S. A. Bargal *et al.*, "Excitation backprop for RNNs," in *CVPR 2018*.

[8] Z. Li *et al.*, "Towards visually explaining video understanding networks with perturbation," in *IEEE WACV 2021*.

[9] X. Bai *et al.*, "Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments," *Pat. Rec.*, vol. 120, 2021.

[10] J. E. Zini *et al.*, "On the explainability of natural language processing deep models," *ACM Computing Surveys*, 2022.

[11] S. Jain *et al.*, "Attention is not Explanation," in *NAACL-HLT 2019*.

[12] S. Serrano *et al.*, "Is attention interpretable?" in *2019 ACL Meeting*.

[13] G. Chrysostomou *et al.*, "Improving the faithfulness of attention-based explanations with task-specific information for text classification," in *2021 ACL Meeting*.

[14] G. Kobayashi *et al.*, "Attention is not only a weight: Analyzing transformers with vector norms," in *EMNLP 2020*.

[15] Y. Liu *et al.*, "Rethinking attention-model explainability through faithfulness violation test," in *ICML 2022*, vol. 162.

[16] S. Wiegreffe *et al.*, "Attention is not not explanation," in *EMNLP 2019*.

[17] E. Apostolidis *et al.*, "Summarizing videos using concentrated attention and considering the uniqueness and diversity of the video frames," in *2022 ACM ICMR*.

[18] P. Li *et al.*, "Exploring global diverse attention via pairwise temporal relation for video summarization," *Pat. Rec.*, vol. 111, 2021.

[19] J. Fajtl *et al.*, "Summarizing videos with attention," in *ACCV 2018*.

[20] M. Gygli *et al.*, "Creating summaries from user videos," in *ECCV 2014*.

[21] Y. Song *et al.*, "TVSum: Summarizing web videos using titles," in *CVPR 2015*.