Proc. NarSUM @ ACM Multimedia 2023. Authors' accepted version. The final paper is at https://doi.org/10.1145/3607540.3617138

A Study on the Use of Attention for Explaining Video Summarization

Evlampios Apostolidis
CERTH-ITI
Thessaloniki, Greece, 57001
apostolid@iti.gr

Vasileios Mezaris
CERTH-ITI
Thessaloniki, Greece, 57001
bmezaris@iti.gr

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

ABSTRACT

In this paper we present our study on the use of attention for explaining video summarization. We build on a recent work that formulates the task, called XAI-SUM, and we extend it by: a) taking into account two additional network architectures and b) introducing two novel explanation signals that relate to the entropy and diversity of attention weights. In total, we examine the effectiveness of seven types of explanation, using three state-of-the-art attention-based network architectures (CA-SUM, VASNet, SUM-GDA) and two datasets (SumMe, TVSum) for video summarization. The conducted evaluations show that the inherent attention weights are more suitable for explaining network architectures which integrate mechanisms for estimating attentive diversity (SUM-GDA) and uniqueness (CA-SUM). The explanation of simpler architectures (VASNet) can benefit from taking into account estimates about the strength of the input vectors, while another option is to consider the entropy of attention weights.

CCS CONCEPTS

• **Computing methodologies** → **Video summarization**.

KEYWORDS

Video summarization, Explainable AI, Attention mechanism, Explanation signals, Replacement functions, Sanity Violation

ACM Reference Format:
Evlampios Apostolidis, Vasileios Mezaris, and Ioannis Patras. 2023. A Study on the Use of Attention for Explaining Video Summarization. In *Proceedings of the 2nd Workshop on User-centric Narrative Summarization of Long Videos (NarSUM '23), October 29, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3607540.3617138

1 INTRODUCTION

The current practice in the Media industry for producing a video summary requires a professional video editor to watch the entire content and decide about the parts of it that should be included in the summary. This is a laborious task and can be really intensive and time-consuming in the case of long videos, or when different summaries should be prepared for the same video in order to be distributed via different video sharing platforms (e.g., YouTube, Vimeo, TikTok) and social networks (e.g., Facebook, Twitter, Instagram) with different specifications about the optimal or maximum video duration [9]. Video summarization technologies aim to generate a short summary by selecting the most informative and important frames (key-frames) or fragments (key-fragments) of the full-length video, and presenting them in temporally-ordered fashion. As discussed in [4], the use of such technologies can drastically reduce the needed resources for video summarization in terms of both time and human effort. Despite the recent advances in the field of video summarization, that are mainly associated with the emergence of modern deep learning network architectures [2], the outcome of a video summarization method still needs to be curated by a video editor, to make sure that all the necessary parts of the video have been included in the summary. This curation could be facilitated if the video summarization method is able to provide explanations about its proposals for building the summary. The provision of such explanations would allow the editor to progressively gain a better understanding of the reasoning behind the proposals of the used method, utilize it more effectively and thus reduce the needed time for content curation.

Despite the fact that, over the last years there is an increasing interest in explaining the outcomes of deep networks processing video data, most works are related with network architectures for action/event recognition [1, 13, 16, 31, 34, 40] and video classification [6, 24, 26]. Apart from these works, other papers present methods for explaining the outcomes of video similarity assessment [29], video text detection [38], and anomaly detection in surveillance videos [37], while a recent work, called XAI-SUM, tries to formulate and investigate the task of explainable video summarization [4]. In this paper, we build on XAI-SUM [4] and extend it by taking into account additional network architectures and novel explanation signals, aiming to further investigate the use of attention as explanation for video summarization. Our contributions are the following:

- We apply the proposed methodology for explainable video summarization in [4], on two additional state-of-the-art network architectures with similar attention mechanisms (VASNet [12] and SUM-GDA [23]).
- We extend the pool of explanations that were taken into account in [4], by introducing two novel explanation signals that are associated with the entropy and diversity of the attention weights.
- We conduct quantitative evaluations using three state-of-the-art attention-based network architectures (CA-SUM [5], VASNet [12], SUM-GDA [23]) and two datasets for video

NarSUM '23, October 29, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0277-8/23/10…$15.00
https://doi.org/10.1145/3607540.3617138

Proc. NarSUM @ ACM Multimedia 2023. Authors' accepted version. The final paper is at https://doi.org/10.1145/3607540.3617138

summarization (SumMe [15] and TVSum [33]), that show the ability of the inherent attention weights (either in their original form or after being scored according to the strength of the input vectors in the attention mechanism) to form meaningful explanations of the video summarization results.

2 RELATED WORK

2.1 Explainable video analysis

This section focuses on methods aiming to produce explanations about the output of network architectures dealing with video data. Bargal et al. (2018) [6], utilized internal representations of the network architecture to form spatio-temporal cues that influence the network's classification/captioning output, and used these cues to localize video fragments that are associated with a specific action or phrase from the generated caption. Aakur et al. (2018) [1], formulated connected structures of the detected visual concepts in the video (e.g., objects and actions) and utilized these structures to produce semantically coherent and explainable representations for video activity interpretation. Stergiou et al. (2019) [34], proposed the use of cylindrical heat-maps to visualize the focus of attention at a frame basis and form explanations of deep networks for action classification and recognition. Zhuo et al. (2019) [40], defined a spatio-temporal graph of semantic-level video states (representing associated objects, attributes and relationships) and applied state transition analysis for video action reasoning. Roy et al. (2019) [31], tried to explain the output of a model for activity recognition by feeding it to a tractable interpretable probabilistic graphical model and performing joint learning over the two. Papoutsakis and Argyros (2019) [29], presented an unsupervised method that evaluates the similarity of two videos based on action graphs representing the detected objects and their behavior, and provides explanations about the outcome of this evaluation. Mänttäri et al. (2020) [26], extended the concept of meaningful perturbation, to spot the video fragment with the greatest impact on the video classification results. Yu et al. (2021) [38], described an end-to-end trainable and interpretable framework for video text detection, that combines spatial and motion information with an appearance-geometry descriptor to generate robust representations of text instances. Li et al. (2021) [24], extended a perturbation-based explanation method for video classification networks, by a loss function that aims to increase the smoothness of explanations in both spatial and temporal dimensions. Gkalelis et al. (2022) [13], 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. Han et al. (2022) [16], proposed a one-shot target-aware tracking strategy to estimate the relevance between objects across the temporal dimension and form a scene graph for each frame, and used the generated video graph (after applying a smoothing mechanism) for explainable action reasoning. Wu et al. (2022) [37], extracted high-level concept and context features for training a denoising autoencoder that was used for explaining the output of anomaly detection in surveillance videos. Finally, Apostolidis et al. (2022) [4], made a first attempt towards explaining video summarization. In their work, called XAI-SUM, Apostolidis et al. formulated the task as the production of an explanation mask indicating the parts of the video that influenced the most the estimates of a video summarization network about the frames' importance. Then, they utilized a state-of-the-art network architecture (CA-SUM) and two datasets for video summarization (SumMe and TVSum), and evaluated the performance of various attention-based explanation signals by investigating the network's input-output relationship (according to different input replacement functions), and using a set of tailored evaluation measures.

2.2 Attention-based explanation

A few attempts were made towards the use of attention for explaining the outcomes of deep network architectures. Most works lie within the NLP domain. Serrano and Smith (2019) [32], investigated the use of attention weights (either on a single basis or after forming sets of them) both solely and in combination with the gradients for their computation, for interpreting the outcomes of an NLP model for text classification. Wiegreffe and Pinter (2019) [36], proposed four alternative tests to determine when/whether attention can be used as explanation; each test allows for meaningful interpretation of attention mechanisms in RNN models utilized for various binary text classification tasks. Jain and Wallace (2019) [18], examined the use of the inherent attention weights for explaining NLP models, considering a wider range of tasks that included text classification, natural language inference and question answering. Kobayashi et al. (2020) [20], explored the use of weighted attention according to the norm of the Value-based transformed input feature vectors, to interpret the output of a pre-trained BERT model [11]. Hao et al. (2021) [17], assessed the performance of explanations formulated using gradient-based attention weights and the BERT model for text classification. Chrysostomou and Aletras (2021) [7], presented a method for improving the faithfulness of attention-based explanations for text classification, taking into account explanations formed using the inherent attention weights and their gradients, while additional types of attention-based explanations were considered in their subsequent work [8] that focused on evaluating their out-of-domain faithfulness. Liu et al. (2022) [25], introduced a faithfulness violation test to measure the consistency between several attention-based explanations and the impact polarity. Finally, the use of attention as explanation has been investigated recently for interpreting the output of deep networks dealing with other tasks, such as image classification [14, 28], image recognition [22], heart sound classification [30] and multimodal trajectory prediction [39].

3 EXPLANATION APPROACH

We follow the video summarization and explanation approach proposed in [4]. With respect to summarization, we assume that the video is split into consecutive and non-overlapping fragments of fixed-size L, each fragment's importance is computed by averaging the importance of the frames in it, and the video summary is formed by the M top-scoring video fragments. With regards to explanation, we consider the creation of an explanation mask that indicates the M most influential video fragments for the estimates of a video summarization network about the importance of the video frames.

3.1 Network architectures

We focus on visual-based network architectures that rely on the use of a self-attention mechanism. The processing pipeline of these

Proc. NarSUM @ ACM Multimedia 2023. Authors' accepted version. The final paper is at https://doi.org/10.1145/3607540.3617138

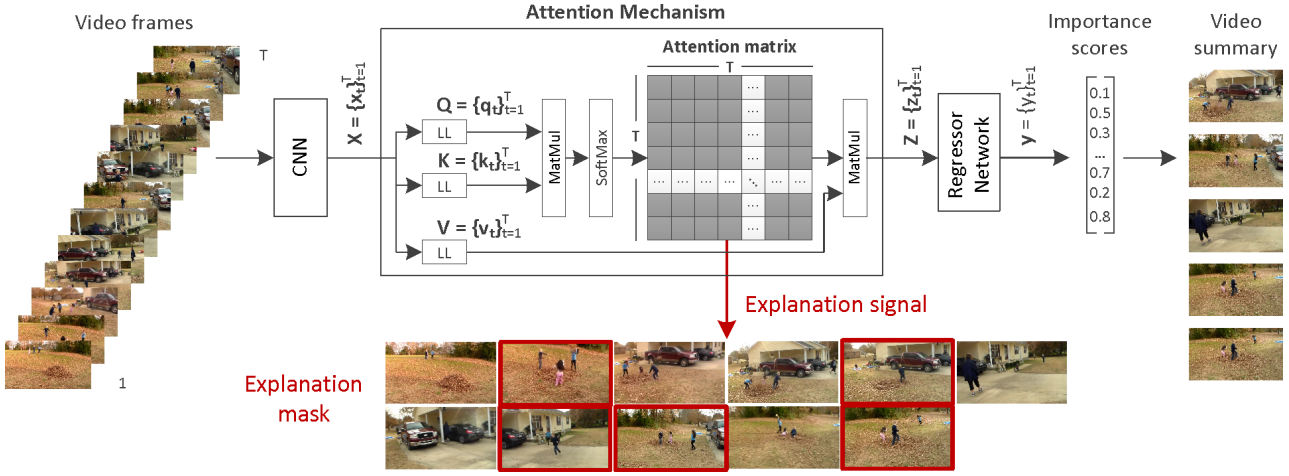

Figure 1: Upper part: the processing pipeline of attention-based network architectures for video summarization; the summary is created by stitching the top-5 most important fragments in chronological order. Lower part: the output of the adopted explanation approach; the number of most influential video fragments M in the explanation mask equals to five. This figure follows the visualization in XAI-SUM [4].

architectures is shown in the upper part of Fig. 1. Assuming a video of T frames and a CNN model for deep feature extraction, the attention mechanism gets as input the frames' feature representations $\boldsymbol{X} = \{\boldsymbol{x_i}\}_{i=1}^{T}$ and defines the Query- ($\boldsymbol{Q} = \{\boldsymbol{q_i}\}_{i=1}^{T}$), Key- ($\boldsymbol{K} = \{\boldsymbol{k_i}\}_{i=1}^{T}$, and Value-based ($\boldsymbol{V} = \{\boldsymbol{v_i}\}_{i=1}^{T}$) transformations using a triplet of linear layers (LL). Following, it computes the dot product of the former two ($\boldsymbol{Q} * \boldsymbol{K}^T$, where $\boldsymbol{K}^T$ is the transposed version of $\boldsymbol{K}$) and applies a softmax conversion, forming 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 $\boldsymbol{A}$ is associated with a different video frame and represents its significance for all frames of the video based on the modeled context by the trained attention mechanism. The matrix $\boldsymbol{A}$ is multiplied (dot product) with $\boldsymbol{V}$, formulating the output of the attention mechanism ($\boldsymbol{Z} = \{\boldsymbol{z_t}\}_{t=1}^{T}$). The latter is then used by a trained Regressor Network, which computes the frames' importance scores $\boldsymbol{y}$ that are finally utilized to calculate fragment-level importance and select the most important fragments for building the video summary.

Based on the above, we extend the work of [4] by considering the following visual-based network architectures:

- **CA-SUM** [5] integrates a concentrated attention mechanism that focuses on non-overlapping blocks in the main diagonal of the attention matrix and takes into account the attentive uniqueness and diversity of the associated frames of the video. It learns the task without supervision using a loss function that relates to the length of the generated summary.
- **VASNet** [12] contains a soft self-attention mechanism that models the frames' dependence according to their pair-wise similarities in a learned latent space and scores the frames using a Regressor Network. It learns the task based on ground-truth annotations about each frame's importance and a relevance loss function (Mean Squared Error).
- **SUM-GDA** [23] tries to increase the diversity of the visual content of the summary by computing global diverse attention scores (one per frame) and using these scores to form the context vectors in the output of the attention mechanism. It can learn the task in both supervised and unsupervised manner using tailored loss functions.

More complex network architectures that combine global and local (multi-head) attention (e.g., PGL-SUM [3]) or utilize data from additional modalities (e.g., CLIP-It [27]) were not taken into account, in order to avoid comparisons across methods that apply a significantly different processing pipeline. Such network architectures will be studied in future extensions of this work, aiming to gain insights about the contribution of different levels of attention or the use of textual data, for explaining the output of video summarization.

3.2 Explanation signals

In [4], the formulation of explanation signals was made based on the values in the main diagonal of the attention matrix. In this work, we augment the pool of signals used in [4], by introducing two additional signals that are associated with the diversity of attention weights. In particular, 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 with respect to the weights in the main diagonal of the attention matrix $\{\nabla a_{i,i}\}_{i=1}^{T}$.
- **Grad Attention (GA)** is formed using the weights in the main diagonal of the attention matrix, scored based on the gradients for their computation $\{a_{i,i} \odot \nabla a_{i,i}\}_{i=1}^{T}$.
- **Input Norm Attention (NA)** is formed using the weights in the main diagonal of the attention matrix, scored based on the norm of the Value-based transformed input vectors in the attention mechanism $\{a_{i,i} \odot ||\boldsymbol{v_i}||\}_{i=1}^{T}$.

Proc. NarSUM @ ACM Multimedia 2023. Authors' accepted version. The final paper is at https://doi.org/10.1145/3607540.3617138

- **Input Norm Grad Attention (NGA)** is formed using the weights in the main diagonal of the attention matrix, scored based on the gradients for their computation and the norm of the Value-based transformed input vectors in the attention mechanism $\{a_{i,i} \odot \nabla a_{i,i} \odot ||\boldsymbol{v_i}||\}_{i=1}^{T}$).
- **Entropy of Attention (EoA)** is formed by computing the entropy of each row of the attention matrix according to the following formula: $-\sum_{t=1}^{T} a_{i,t} \cdot log(a_{i,t})$ (similarly to [5]).
- **Diversity of Attention (DoA)** is formed by computing the normalized pairwise dissimilarity between the values of each row of the attention matrix according to the following formulas: $d_i = \prod_{t=1}^{T}(1 - a_{i,t})/||\mathbf{d}||$ where $\mathbf{d} = [d_1, d_2, ..., d_T]$ (similarly to [23]).

All the above, result in the definition of frame-level explanation scores; then, video-fragment-level explanation masks (see Fig. 1) are formed by computing fragment-level explanation scores (via mean pooling) and selecting the M fragments with the highest scores.

3.3 Replacement functions

Similarly to [4], to examine the network's input-output relationship we apply the following replacement functions on parts of the input corresponding to different video fragments:

- **Slice-out** removes the specified part from the original input, thus resulting in a shorter input sequence.
- **Input Mask** replaces the specified part of the original input, with a predefined mask which is equally-sized with this fragment and is composed of deep feature representations of black or white frames.
- **Randomization** replaces 50% of the elements of each feature representation within the specified part of the original input, using the corresponding elements from randomly-selected feature representations from the remaining part of the input.
- **Attention Mask** sets the attention weights that relate with the specified part of the original input equal to zero, such that this part will not be forwarded in the network anymore.

3.4 Evaluation measures

As in [4], to quantify the impact of each video fragment in the network's output, after applying a replacement function we compute the difference of estimates as $\Delta E(\boldsymbol{X}, \hat{\boldsymbol{X}}^{\boldsymbol{k}}) = \tau(\boldsymbol{y}, \boldsymbol{y}^{\boldsymbol{k}})$. In this formula, $\boldsymbol{X}$ is the set of original feature representations, $\hat{\boldsymbol{X}}^{\boldsymbol{k}}$ is the updated set after replacing the features of the frames belonging to the k^{th} video fragment, $\boldsymbol{y}$ and $\boldsymbol{y}^{\boldsymbol{k}}$ are the outputs of the network for $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}^{\boldsymbol{k}}$, respectively, and τ is the Kendall's τ correlation coefficient [19]. Based on ΔE, we assess the performance of each explanation signal using the following evaluation measures:

- **Discoverability+ (D^+)** evaluates if fragments assigned with higher explanation scores have a significant influence to the network's estimates. It is calculated as the average of the obtained ΔE values after sequentially replacing parts of the input corresponding to: a) the top-1%, 5%, 10%, 15% and 20% of the fragments with the highest explanation scores (batch manner), and b) the M fragments with the highest explanation scores (one-by-one manner). The higher this measure is, the greater the ability of the explanation signal to correctly spot the video fragments with the highest influence to the network.
- **Discoverability- (D^-)** evaluates if the fragments assigned with lower explanation scores have small influence to the network's estimates. It is calculated as the average of the obtained ΔE values after sequentially replacing parts of the input corresponding to: a) the top-1%, 5%, 10%, 15% and 20% of the fragments with the lowest explanation scores (batch manner), and b) the M fragments with the lowest explanation scores (one-by-one manner). The lower this measure is, the greater the effectiveness of the explanation signal to correctly spot video fragments with limited significance for the network.
- **Sanity Violation (SV)** quantifies the ability of explanations to correctly discriminate important from unimportant video fragments. It is calculated by counting the number of cases where the condition ($D^+ > D^-$) is violated after sequentially replacing parts of the input corresponding to: a) the top-1%, 5%, 10%, 15% and 20% of the fragments with the highest and lowest explanation scores (batch manner), and b) the M top- and less-scoring fragments in a pair-wise (one-by-one) manner, and then expressing the computed value as a fraction of the total number of replacements. This measure ranges in $[0, 1]$; the closest its value is to zero, the greater the reliability of the explanation signal.
- **Rank Correlation (RC)** examines if the assigned explanation score to a video fragment is analogous with the fragment's influence to the network's output. It is calculated by computing the Spearman's ρ rank correlation coefficient [21] between the assigned fragment-level explanation scores and the obtained ΔE values after sequentially replacing each one of them (thus, it is computed only when input replacement is performed in a one-by-one manner). It ranges in $[-1, +1]$; values close to +1 signify strong correlation, while values close to 0 and -1 denote neutral and strongly negative correlation, respectively.

4 EXPERIMENTS

This section describes the used datasets and implementation details, and reports the results of our quantitative and qualitative analysis.

4.1 Datasets and implementation details

In our experiments we use the SumMe [15] and TVSum [33] datasets for video summarization. SumMe is composed of 25 videos with diverse video contents (e.g., covering holidays, events and sports), captured from both first-person and third-person view. TVSum contains 50 videos from 10 categories of the TRECVid MED task. Videos are downsampled to 2 fps and deep feature representations of the frames are obtained by taking the output of the pool5 layer of GoogleNet [35] trained on ImageNet [10]. The parameter M, which indicates the number of top- and lower-scoring fragments that are being affected by the different replacement functions, as well as the number of video fragments that are highlighted as the most influential ones by the produced explanation mask, is set equal to five. The size L of the video fragments is set equal to 10 seconds (i.e., 20 sampled frames). Similarly to [4], we increase the amount

Proc. NarSUM @ ACM Multimedia 2023. Authors' accepted version. The final paper is at https://doi.org/10.1145/3607540.3617138

Table 1: Performance of the different explanation signals on the SumMe and TVSum datasets, after replacing parts of the input in a batch manner. The arrows indicate the optimal (minimum or maximum) value for each measure.

CA-SUM														
	SumMe							TVSum						
	IA	NA	GA	GoA	NGA	EoA	DoA	IA	NA	GA	GoA	NGA	EoA	DoA
D^- (↓)	**0.202**	0.206	0.218	0.212	0.220	0.217	0.204	**0.141**	0.143	0.240	0.236	0.240	0.187	0.163
D^+ (↑)	**0.250**	0.238	0.196	0.195	0.195	0.210	0.233	**0.217**	0.215	0.123	0.123	0.123	0.162	0.202
SV (↓)	**0.336**	0.368	0.560	0.552	0.560	0.552	0.384	**0.224**	0.268	0.844	0.832	0.844	0.680	0.360
VASNet														
	SumMe							TVSum						
	IA	NA	GA	GoA	NGA	EoA	DoA	IA	NA	GA	GoA	NGA	EoA	DoA
D^- (↓)	0.198	0.172	0.209	0.211	0.210	**0.157**	0.160	**0.114**	**0.114**	0.192	0.191	0.191	0.134	0.134
D^+ (↑)	0.144	0.156	0.143	0.143	0.142	**0.177**	0.158	0.152	**0.154**	0.094	0.094	0.094	0.133	0.128
SV (↓)	0.648	0.568	0.672	0.688	0.688	**0.400**	0.536	0.380	**0.304**	0.736	0.740	0.732	0.580	0.560
SUM-GDA														
	SumMe							TVSum						
	IA	NA	GA	GoA	NGA	EoA	DoA	IA	NA	GA	GoA	NGA	EoA	DoA
D^- (↓)	**0.131**	0.146	0.133	0.133	0.134	0.151	**0.131**	0.119	**0.105**	0.133	0.133	0.131	0.129	0.121
D^+ (↑)	**0.137**	0.122	0.127	0.128	0.128	0.099	0.123	0.105	**0.134**	0.121	0.121	0.122	0.096	0.105
SV (↓)	**0.440**	0.630	0.520	0.510	0.540	0.880	0.590	0.650	**0.295**	0.590	0.585	0.555	0.810	0.660

of experimental evaluations by assuming five different randomly-created splits for each dataset[1]. For CA-SUM, we use the released pre-trained models based on these splits [2]. For VASNet and SUM-GDA, we train models of these network architectures based on the released code (VASNet[3]) and the descriptions in the associated paper (SUM-GDA), and using the aforementioned data splits. In the following, we report the average scores over these runs. The code for reproducing the reported results is publicly-available at: https://github.com/e-apostolidis/XAI-SUM.

4.2 Quantitative analysis

Table 1 reports the performance of each explanation signal on videos from the SumMe and TVSum dataset, for each different network architecture and after replacing parts of the input in a batch manner. The reported values represent the average score across the different replacement functions. These results show that, forming explanations using the inherent attention weights (either purely (IA) or after being scored according to the norm of the Value-based transformed input vectors (NA)) is, in most cases, the best approach according to the adopted evaluation approach. Such explanations exhibit the optimal and/or near-optimal performance for all network architectures on the videos of TVSum, as indicated also by the Sanity Violation scores - which, in our perspective, is an important criterion for the trustworthiness of an explanation signal - in the graph on the right part of Fig. 2. Moreover, they seem to be the most appropriate for producing explanations of the output of CA-SUM and SUM-GDA for the videos of SumMe, as also shown by the low Sanity Violation scores of the IA-based explanation signal in the graph on the left part of Fig. 2. On average, they achieve the lowest/highest Discoverability-/+ scores and produce explanations that, in most cases (approx. $55-65\%$ on SumMe and $70-75\%$ on TVSum), correctly discriminate the most and least influential fragments of the video. The only exception is observed in the case of VASNet for the SumMe dataset, where the use of the entropy of the attention weights to form explanations (EoA) appears as the most effective approach. As depicted also in the graph on the left part of Fig. 2, the performance of the relevant explanation signal significantly surpasses the performance of other signals and is comparable only to the performance of explanations formulated based on the diversity of attention weights (DoA). Taking into account the performance of EoA-based and DoA-based explanations for VASNet also on the TVSum videos - which is lower than the performance of IA/NA-based explanations but clearly higher than the one of gradient-based explanations (GA, GoA, NGA), as also shown in the graph on the right part of Fig. 2 - we argue that the attention mechanism of this network architecture models the dependence of video frames using more varying attention weights; thus, the use of estimates about the entropy (called attentive uniqueness in [5]) and diversity of the attention weights as proposed, can lead to well-performing explanation signals. On the contrary, network architectures that integrate mechanisms for estimating the attentive diversity (SUM-GDA) and uniqueness of the video frames (CA-SUM) seem to produce attention weights that are already descriptive enough and can be directly used to form explanation signals. Finally, explanations formed using the gradients of the attention weights (GA, GoA, NGA) are, in general, the worst-performing ones. In most cases, such explanations result in higher/lower D^-/D^+ scores than the ones obtained for non-gradient-based signals (IA, NA, EoA, DoA), and most frequently, they fail to distinguish the most and least influential fragments of the video ($55-65\%$ and $55-85\%$ of the cases on SumMe and TVSum, respectively), as also shown in the graphs of Fig. 2.

Table 2 reports the experimental results when the replacement of parts of the input to the network architectures is performed in a

[1]Publicly-available at: https://github.com/e-apostolidis/CA-SUM
[2]Accessible at: https://zenodo.org/record/6562992
[3]Accessible at: https://github.com/ok1zjf/VASNet

Proc. NarSUM @ ACM Multimedia 2023. Authors' accepted version. The final paper is at https://doi.org/10.1145/3607540.3617138

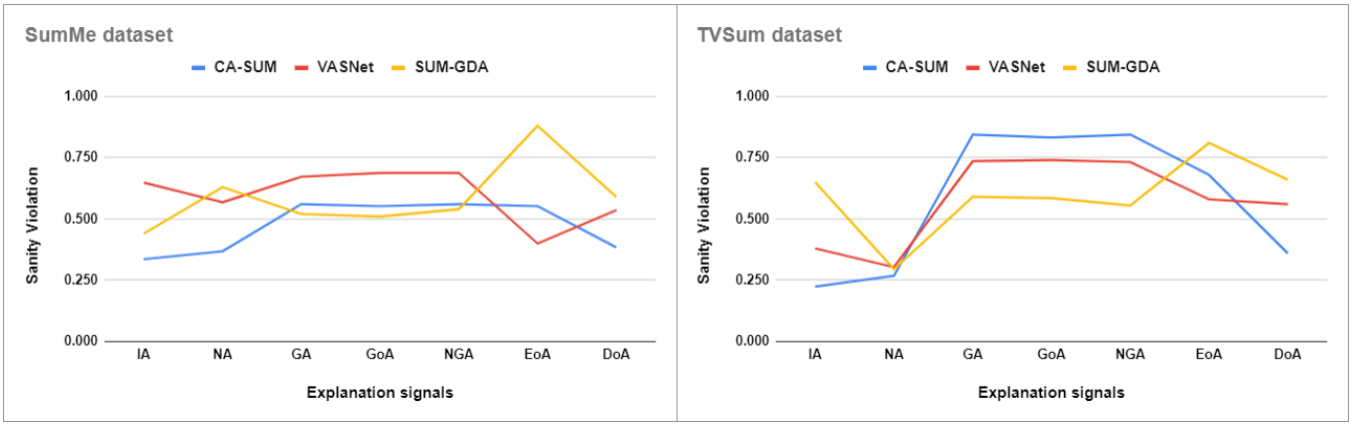

Figure 2: Sanity violation scores of the different explanation signals on the output of the considered network architectures for the videos of SumMe (left part) and TVSum (right part) datasets, when input replacements are made on a batch manner.

Table 2: Performance of the different explanation signals on the SumMe and TVSum datasets, after replacing parts of the input in a one-by-one manner. The arrows indicate the optimal (minimum or maximum) value for each measure.

CA-SUM														
	SumMe							TVSum						
	IA	NA	GA	GoA	NGA	EoA	DoA	IA	NA	GA	GoA	NGA	EoA	DoA
D^- (↓)	0.162	**0.161**	0.177	0.176	0.175	0.173	0.169	**0.112**	**0.112**	0.148	0.147	0.148	0.130	0.119
D^+ (↑)	**0.181**	0.179	0.167	0.166	0.167	0.172	0.177	**0.141**	**0.141**	0.104	0.104	0.104	0.119	0.134
SV (↓)	**0.352**	0.368	0.584	0.568	0.552	0.560	0.384	**0.204**	0.212	0.812	0.788	0.836	0.580	0.396
RC (↑)	**0.182**	0.130	-0.030	-0.075	-0.068	-0.020	0.118	**0.303**	0.295	-0.088	-0.166	-0.092	-0.073	0.163
VASNet														
	SumMe							TVSum						
	IA	NA	GA	GoA	NGA	EoA	DoA	IA	NA	GA	GoA	NGA	EoA	DoA
D^- (↓)	0.116	0.109	0.118	0.119	0.118	**0.102**	0.108	**0.065**	0.066	0.090	0.089	0.089	0.072	0.074
D^+ (↑)	0.098	0.103	0.098	0.098	0.098	**0.110**	0.106	**0.080**	**0.080**	0.058	0.057	0.057	0.072	0.071
SV (↓)	0.632	0.528	0.640	0.640	0.656	**0.400**	0.512	0.364	**0.336**	0.696	0.716	0.704	0.500	0.512
RC (↑)	-0.199	-0.022	0.306	**0.314**	0.298	0.091	-0.048	**0.217**	**0.217**	-0.021	-0.034	-0.020	0.015	0.000
SUM-GDA														
	SumMe							TVSum						
	IA	NA	GA	GoA	NGA	EoA	DoA	IA	NA	GA	GoA	NGA	EoA	DoA
D^- (↓)	**0.073**	0.081	0.076	0.076	0.076	0.079	0.078	0.061	**0.054**	0.062	0.062	0.062	0.062	0.061
D^+ (↑)	**0.078**	0.073	0.072	0.073	0.072	0.069	0.075	0.057	**0.066**	0.060	0.060	0.060	0.056	0.058
SV (↓)	**0.410**	0.630	0.580	0.530	0.570	0.700	0.570	0.615	**0.265**	0.510	0.485	0.515	0.730	0.610
RC (↑)	**0.046**	-0.110	-0.092	-0.097	-0.124	-0.166	0.041	-0.005	**0.104**	-0.080	-0.068	-0.079	-0.072	-0.073

one-by-one manner. These results are, to a large extent, aligned with the ones in Table 1. Explanation signals formed using the inherent attention weights (IA) or a scored version of these weights according to the norm of the Value-based transformed input vectors (NA) exhibit the optimal performance across all network architectures on TVSum, as denoted by the low Sanity Violation scores in the graph on the right part of Fig. 3. With respect to SumMe, IA-based signals appear to be the most appropriate for explaining the output of CA-SUM and SUM-GDA, as indicated also by the low scores in the graph on the left part of Fig. 3. As before, using estimates about the entropy and diversity of the attention weights appears to be beneficial when producing explanations for the output of VASNet for the videos of this dataset (see also the corresponding scores in the graph on the left part of Fig. 3). All the aforementioned explanation signals are associated with the lowest/highest Discoverability-/+ scores and the lowest Sanity Violation scores. In addition, based on the observed Rank Correlation they are capable of assigning fragment-level explanation scores that, in most cases, are positively correlated with the fragments' influence to the networks' output. On the contrary, gradient-based explanations (GA, GoA, NGA) perform worse. They violate the sanity test most frequently (as shown in the graphs of Fig. 3) and they assign fragment-level explanation scores that are (in all cases but one) negatively correlated with the influence of each fragment to the networks' output.

Proc. NarSUM @ ACM Multimedia 2023. Authors' accepted version. The final paper is at https://doi.org/10.1145/3607540.3617138

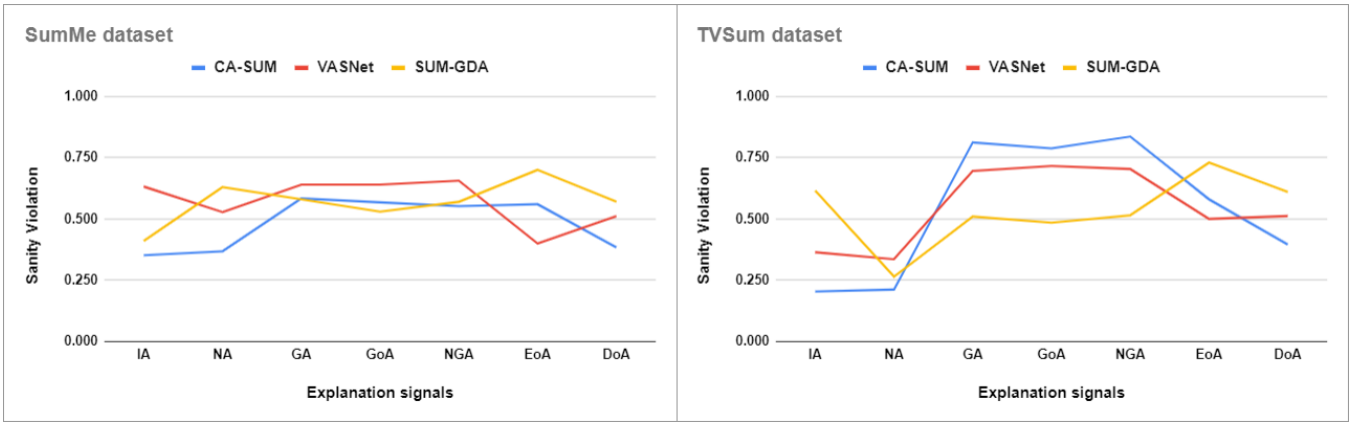

Figure 3: Sanity violation scores of the different explanation signals on the output of the considered network architectures for the videos of SumMe (left part) and TVSum (right part) datasets, when input replacements are made on a one-by-one manner.

The experimental results discussed above show that, in general, the use of the inherent attention weights - either in their original form or after a scoring process based on the norm of the Value-based transformed input vectors - to form explanations, is the best choice. The use of the inherent attention weights appears to be more suitable for explaining the output of network architectures including mechanisms for estimating the attentive diversity (SUM-GDA) and uniqueness of the video frames (CA-SUM), while taking into account estimates about the strength of the input vectors in the attention mechanism (computed by the norm of the Value-based transformed input vectors) can be beneficial in some cases (VASNet and SUM-GDA). Moreover, the entropy of the inherent attention weights is another option in the case of network architectures that do not involve any post-processing of the computed attention matrix (VASNet). Finally, the use of gradients is not a good choice, since it leads to explanation signals that fail more frequently to discriminate the most and least influential fragments of the video, and thus to provide reliable explanations about the output of the considered network architectures.

4.3 Qualitative analysis

In our qualitative analysis we used the CA-SUM network architecture and formed explanation masks using the inherent attention weights. According to the results in Tables 1 and 2, this combination is associated with constantly better performance (in terms of all the used evaluation measures) compared to the performance of the most effective explanation signals for the other network architectures. So, in the following we report the findings of our analysis using the produced explanation mask by the inherent attention weights (IA), for two videos from the TVSum and SumMe datasets. Similarly to [4], in Figs. 4 and 5 each video fragment is illustrated using one key-frame that was selected based on its representativeness. The red-coloured bounding boxes signify the most influential fragments according to the used explanation signal, and the blue-coloured bounding boxes indicate the top-scoring fragments based on the CA-SUM estimates about the frames' importance.

In the example video of Fig. 4, the focus of the attention mechanism is mainly put on the veterinarian with the dog, and the ear cleaning process. Parts of the video showing text-written tips, close-ups of the veterinarian alone, and the cleaning product, are less important according to the modeled video context. Using this information, CA-SUM assigns higher importance scores to parts of the video showing the veterinarian with the dog, explaining and performing the ear cleaning process. In the example video of Fig. 5, the attention mechanism seems to concentrate mainly on parts of the video showing the kids playing in the leaves. Other parts of the video presenting the front-yard of the house, the cars in the parking, and a distant shot of the kids, seem to be less attractive. Based on the behavior of the attention mechanism, CA-SUM promotes parts of the video that are mainly associated with the kids playing in the leaves, as four out of the five top-scoring fragments contain this visual content. These paradigms lead to findings similar to the ones discussed in [4]; using the inherent attention weights to form explanations as proposed in [4] could enable a level of understanding about the focus of attention, and support the explanation of attention-based video summarization networks.

5 CONCLUSIONS

In this paper, we reported our study on the use of attention for explaining video summarization. Building on a recent work that formulated this task and defined an evaluation protocol [4], we performed a more extended investigation that included additional network architectures and novel explanation signals. Our experimental evaluations involved three network architectures (CA-SUM, VASNet, SUM-GDA), seven explanation signals, and two datasets (SumMe, TVSum) for video summarization. Our findings showed that inherent attention can be used to explain networks estimating attentive diversity (SUM-GDA) and uniqueness (CA-SUM). The explanation of simpler architectures (VASNet) requires to also take into account estimates about the strength of the input vectors, while another option is to consider the entropy of attention weights.

ACKNOWLEDGMENTS

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

Proc. NarSUM @ ACM Multimedia 2023. Authors' accepted version. The final paper is at https://doi.org/10.1145/3607540.3617138

Figure 4: The five most influential fragments (in red-coloured bounding boxes) and the five top-scoring fragments (in blue-coloured bounding boxes) for a TVSum video, titled "How to Clean Your Dog's Ears - Vetoquinol".

Figure 5: The five most influential fragments (in red-coloured bounding boxes) and the five top-scoring fragments (in blue-coloured bounding boxes) for a SumMe video, titled "Kids playing in leaves".

REFERENCES

[1] Sathyanarayanan N. Aakur, Fillipe D. M. de Souza, and Sudeep Sarkar. 2018. An Inherently Explainable Model for Video Activity Interpretation. In *The Workshops of the 32nd AAAI Conf. on Artificial Intelligence*.

[2] Evlampios Apostolidis, Eleni Adamantidou, Alexandros I. Metsai, Vasileios Mezaris, and Ioannis Patras. 2021. Video Summarization Using Deep Neural Networks: A Survey. *Proc. IEEE* 109, 11 (2021), 1838–1863. https://doi.org/10.1109/JPROC.2021.3117472

[3] Evlampios Apostolidis, Georgios Balaouras, Vasileios Mezaris, and Ioannis Patras. 2021. Combining Global and Local Attention with Positional Encoding for Video Summarization. In *2021 IEEE International Symposium on Multimedia (ISM)*. 226–234. https://doi.org/10.1109/ISM52913.2021.00045

[4] Evlampios Apostolidis, Georgios Balaouras, Vasileios Mezaris, and Ioannis Patras. 2022. Explaining video summarization based on the focus of attention. In *2022 IEEE Int. Symposium on Multimedia (ISM)*. 146–150. https://doi.org/10.1109/ISM55400.2022.00029

[5] Evlampios Apostolidis, Georgios Balaouras, Vasileios Mezaris, and Ioannis Patras. 2022. Summarizing Videos Using Concentrated Attention and Considering the Uniqueness and Diversity of the Video Frames. In *Proc. of the 2022 Int. Conf. on Multimedia Retrieval* (Newark, NJ, USA) *(ICMR '22)*. Association for Computing Machinery, New York, NY, USA, 407–415. https://doi.org/10.1145/3512527.3531404

[6] Sarah Adel Bargal, Andrea Zunino, Donghyun Kim, Jianming Zhang, Vittorio Murino, and Stan Sclaroff. 2018. Excitation Backprop for RNNs. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

[7] George Chrysostomou and Nikolaos Aletras. 2021. Improving the Faithfulness of Attention-based Explanations with Task-specific Information for Text Classification. In *Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int. Joint Conf. on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 477–488. https://doi.org/10.18653/v1/2021.acl-long.40

[8] George Chrysostomou and Nikolaos Aletras. 2022. An Empirical Study on Explanations in Out-of-Domain Settings. In *Proc. of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 6920–6938. https://doi.org/10.18653/v1/2022.acl-long.477

[9] Chrysa Collyda, Konstantinos Apostolidis, Evlampios Apostolidis, Eleni Adamantidou, Alexandros I. Metsai, and Vasileios Mezaris. 2020. A Web Service for Video Summarization. In *ACM Int. Conf. on Interactive Media Experiences* (Cornella, Barcelona, Spain) *(IMX '20)*. Association for Computing Machinery, New York, NY, USA, 148–153. https://doi.org/10.1145/3391614.3399391

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[12] Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. 2019. Summarizing Videos with Attention. In *Asian Conf. on Computer Vision (ACCV) 2018 Workshops*, Gustavo Carneiro and Shaodi You (Eds.). Springer International Publishing, Cham, 39–54.

[13] Nikolaos Gkalelis, Dimitrios Daskalakis, and Vasileios Mezaris. 2022. ViGAT: Bottom-Up Event Recognition and Explanation in Video Using Factorized Graph Attention Network. *IEEE Access* 10 (2022), 108797–108816. https://doi.org/10.1109/ACCESS.2022.3213652

[14] Ioanna Gkartzonika, Nikolaos Gkalelis, and Vasileios Mezaris. 2023. Learning Visual Explanations for DCNN-Based Image Classifiers Using an Attention Mechanism. In *Computer Vision – ECCV 2022 Workshops*, Leonid Karlinsky, Tomer Michaeli, and Ko Nishino (Eds.). Springer Nature Switzerland, Cham, 396–411.

[15] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. 2014. Creating Summaries from User Videos. In *Europ. Conf. on Computer Vision (ECCV) 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 505–520. https://gyglim.github.io/me/

[16] Yamin Han, Tao Zhuo, Peng Zhang, Wei Huang, Yufei Zha, Yanning Zhang, and Mohan Kankanhalli. 2022. One-shot Video Graph Generation for Explainable Action Reasoning. *Neurocomputing* 488 (2022), 212–225. https://doi.org/10.1016/j.neucom.2022.02.069

[17] Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2021. Self-Attention Attribution: Interpreting Information Interactions Inside Transformer. *Proc. of the AAAI Conf. on Artificial Intelligence* 35, 14 (May 2021), 12963–12971. https://doi.org/10.1609/aaai.v35i14.17533

[18] Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 3543–3556.

https://doi.org/10.18653/v1/N19-1357

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

[20] Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is Not Only a Weight: Analyzing Transformers with Vector Norms. In *Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 7057–7075. https://doi.org/10.18653/v1/2020.emnlp-main.574

[21] Stephen Kokoska and Daniel Zwillinger. 2000. *CRC standard probability and statistics tables and formulae*. Crc Press.

[22] Liangzhi Li, Bowen Wang, Manisha Verma, Yuta Nakashima, Ryo Kawasaki, and Hajime Nagahara. 2021. SCOUTER: Slot Attention-based Classifier for Explainable Image Recognition. In *2021 IEEE/CVF Int. Conf. on Computer Vision (ICCV)*. 1026–1035. https://doi.org/10.1109/ICCV48922.2021.00108

[23] Ping Li, Qinghao Ye, Luming Zhang, Li Yuan, Xianghua Xu, and Ling Shao. 2021. Exploring global diverse attention via pairwise temporal relation for video summarization. *Pattern Recognition* 111 (2021), 107677. https://doi.org/10.1016/j.patcog.2020.107677

[24] Zhenqiang Li, Weimin Wang, Zuoyue Li, Yifei Huang, and Yoichi Sato. 2021. Towards Visually Explaining Video Understanding Networks with Perturbation. *2021 IEEE Winter Conf. on Applications of Computer Vision (WACV)* (2021), 1119–1128.

[25] Yibing Liu, Haoliang Li, Yangyang Guo, Chenqi Kong, Jing Li, and Shiqi Wang. 2022. Rethinking Attention-Model Explainability through Faithfulness Violation Test. In *Proc. of the 39th Int. Conf. on Machine Learning (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.). PMLR, 13807–13824. https://proceedings.mlr.press/v162/liu22i.html

[26] Joonatan Mänttäri, Sofia Broomé, John Folkesson, and Hedvig Kjellström. 2020. Interpreting Video Features: A Comparison of 3D Convolutional Networks and Convolutional LSTM Networks. In *Asian Conference on Computer Vision (ACCV) 2020*, Hiroshi Ishikawa, Cheng-Lin Liu, Tomas Pajdla, and Jianbo Shi (Eds.). Springer International Publishing, Cham, 411–426.

[27] Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell. 2021. CLIP-It! Language-Guided Video Summarization. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 13988–14000. https://proceedings.neurips.cc/paper/2021/file/7503cfacd12053d309b6bed5c89de212-Paper.pdf

[28] Mariano Ntrougkas, Nikolaos Gkalelis, and Vasileios Mezaris. 2022. TAME: Attention Mechanism Based Feature Fusion for Generating Explanation Maps of Convolutional Neural Networks. In *2022 IEEE Int. Symposium on Multimedia (ISM)*. 58–65. https://doi.org/10.1109/ISM55400.2022.00014

[29] Konstantinos E. Papoutsakis and Antonis A. Argyros. 2019. Unsupervised and Explainable Assessment of Video Similarity. In *British Machine Vision Conference*. https://api.semanticscholar.org/CorpusID:199525379

[30] Zhao Ren, Kun Qian, Fengquan Dong, Zhenyu Dai, Wolfgang Nejdl, Yoshiharu Yamamoto, and Björn W. Schuller. 2022. Deep attention-based neural networks for explainable heart sound classification. *Machine Learning with Applications* 9 (2022), 100322. https://doi.org/10.1016/j.mlwa.2022.100322

[31] Chiradeep Roy, Mahesh Shanbhag, Mahsan Nourani, Tahrima Rahman, Samia Kabir, Vibhav Gogate, Nicholas Ruozzi, and Eric D. Ragan. 2019. Explainable Activity Recognition in Videos. In *ACM Intelligent User Interfaces (IUI) Workshops*.

[32] Sofia Serrano and Noah A. Smith. 2019. Is Attention Interpretable?. In *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 2931–2951. https://doi.org/10.18653/v1/P19-1282

[33] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. 2015. TVSum: Summarizing web videos using titles. In *2015 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*. 5179–5187. https://doi.org/10.1109/CVPR.2015.7299154

[34] Alexandros Stergiou, Georgios Kapidis, Grigorios Kalliatakis, Christos Chrysoulas, Remco Veltkamp, and Ronald Poppe. 2019. Saliency Tubes: Visual Explanations for Spatio-Temporal Convolutions. In *2019 IEEE Int. Conf. on Image Processing (ICIP)*. 1830–1834. https://doi.org/10.1109/ICIP.2019.8803153

[35] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *2015 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*. 1–9. https://doi.org/10.1109/CVPR.2015.7298594

[36] Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not Explanation. In *Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 11–20. https://doi.org/10.18653/v1/D19-1002

[37] Chongke Wu, Sicong Shao, Pratik Satam, and Salim Hariri. 2022. An explainable and efficient deep learning framework for video anomaly detection. *Cluster Computing* 25, 4 (Aug. 2022), 2715–2737. https://doi.org/10.1007/s10586-021-03439-5

[38] Hongyuan Yu, Yan Huang, Lihong Pi, Chengquan Zhang, Xuan Li, and Liang Wang. 2021. End-to-end video text detection with online tracking. *Pattern Recognition* 113 (2021), 107791. https://doi.org/10.1016/j.patcog.2020.107791

[39] Kunpeng Zhang and Li Li. 2022. Explainable multimodal trajectory prediction using attention models. *Transportation Research Part C: Emerging Technologies* 143 (2022), 103829. https://doi.org/10.1016/j.trc.2022.103829

[40] Tao Zhuo, Zhiyong Cheng, Peng Zhang, Yongkang Wong, and Mohan Kankanhalli. 2019. Explainable Video Action Reasoning via Prior Knowledge and State Transitions. In *Proc. of the 27th ACM Int. Conf. on Multimedia* (Nice, France) *(MM '19)*. Association for Computing Machinery, New York, NY, USA, 521–529. https://doi.org/10.1145/3343031.3351040