Chapter 65
Explainable Video Summarization for Advancing Media Content Production

Evlampios Apostolidis
https://orcid.org/0000-0001-5376-7158
Information Technologies Institute, Centre for Research and Technology, Hellas, Greece

Georgios Balaouras
Information Technologies Institute, Centre for Research and Technology, Hellas, Greece

Ioannis Patras
School of Electronic Engineering and Computer Science, Queen Mary University of London, UK

Vasileios Mezaris
https://orcid.org/0000-0002-0121-4364
Information Technologies Institute, Centre for Research and Technology, Hellas, Greece

ABSTRACT

This chapter focuses on explainable video summarization, a technology that could significantly advance the content production workflow of Media organizations. It starts by presenting the current state of the art in the fields of deep-learning-based video summarization and explainable video analysis and understanding. Following, it focuses on video summarization methods that rely on the use of attention mechanisms and reports on previous works that investigated the use of attention for explaining the outcomes of deep neural networks. Subsequently, it briefly describes a state-of-the-art attention-based architecture for unsupervised video summarization and discusses a recent work that examines the use of various attention-based signals for explaining the outcomes of video summarization. Finally, it provides recommendations about future research directions.

INTRODUCTION

The current practice in the Media industry for producing a video summary reqvuires 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. Moreover, the constantly increasing engagement of users with video sharing platforms (e.g.,

DOI: 10.4018/978-1-6684-7366-5.ch065

YouTube, Vimeo, TikTok) and social networks (e.g., Facebook, Twitter, Instagram), that are used for posting online a variety of video content, such as educational, "how-to"/instructional, training, gaming, travelling, cooking and music playing videos, as well as commercials, movie trailers and sports highlights, led to the inclusion of these data distribution channels among the main communication means of Media organizations. However, these different communication means are usually associated with different specifications about the optimal or maximum video duration (Collyda et al., 2020). For example, videos posted on Facebook's feed and YouTube are expected / recommended to be up to 2 min. long, videos posted on Instagram's feed and Twitter are most commonly up to 30 sec., while videos posted on TikTok, Facebook and Instagram as stories are even shorter (i.e., 15 to 20 sec. long). This means that different summaries should be produced for a given video, which significantly increases the workload of the video editor.

Technologies for automated video summarization, aim to generate a short synopsis that summarizes the video content by selecting its most informative and important parts. The use of such technologies by Media organizations can drastically reduce the needed resources for media content production in terms of both time and human effort, and facilitate indexing, browsing, retrieval and promotion of their media assets. Despite the recent advances in the field of video summarization, which are tightly associated with the emergence of modern deep-learning network architectures (Apostolidis et al., 2021b), the outcome of a video summarization technology still needs to be curated by the video editor, in order to ensure that all the needed parts of the video were included in the video summary. This content production step could be further facilitated, if the video editor is provided with explanations about the suggestions made by the used video summarization 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 facilitating content curation.

Given the above, this chapter focuses on explainable video summarization, a technology that could significantly advance the content production workflow of Media organizations. It starts by presenting the current state of the art in the fields of deep-learning-based video summarization and explainable video analysis and understanding. Following, it focuses on video summarization methods that rely on the use of self-attention mechanisms for modelling frames' dependence and estimating their importance. As a note, self-attention is a type of attention used in the Transformer Network (Vaswani et al., 2017) for modelling the relation between different elements of an input sequence in order to compute a representation of this sequence. In layman's terms, the self-attention mechanism allows the elements of the input sequence to interact with each other, and takes their relationship into consideration to determine which of them requires greater attention and dynamically adjust their impact on the output. The chapter continues by reporting on previous works that investigated the use of attention for explaining the outcomes of deep neural networks. Most of them relate to the natural language processing (NLP) domain, but recently, attention was used to interpret the output of networks trained for image recognition and classification, and multimodal trajectory prediction. Subsequently, the chapter briefly describes a state-of-the-art attention-based architecture for unsupervised video summarization, and discusses a recent work that examines the use of various attention-based signals for explaining the outcomes of video summarization. Finally, it provides recommendations about future research directions on explainable video summarization, and concludes this report.

BACKGROUND

Video Summarization

Several approaches have been introduced to automate video summarization, and the current state of the art is represented by methods utilizing deep network architectures (Apostolidis et al., 2021b). These methods can be coarsely categorized into: a) unimodal approaches that utilize only the visual content of the videos, and b) multimodal approaches that use also the audio stream and the available textual metadata (e.g., the video's title and/or description) or some user-specified textual description (e.g., in the form of a set of keywords) about the content of the summary. With regards to the adopted learning strategy, the methods of each one of the above categories can be mainly divided into: a) supervised approaches that learn the task based on the use of human-labeled ground-truth annotations, and b) unsupervised approaches that take into account different criteria about the video summary.

Early unimodal (visual-based) approaches for video summarization tried to model the variable-range temporal dependence among frames and learn how to estimate their importance according to ground-truth annotations. For this, they used architectures of Recurrent Neural Networks (RNNs) in the typical or in a hierarchical form (Zhang et al., 2016; Zhao et al., 2017, 2018, 2021a). In some cases such architectures were combined with tailored attention mechanisms (Lebron Casas & Koblents, 2019; Ji et al., 2020a, 2020b; Lin et al., 2022), or extended by memory networks to increase the memorization capacity of the architecture and capture long-range temporal dependencies among parts of the video (Feng et al., 2018; Wang et al., 2019). Going one step further, a group of techniques aimed to learn the frames' importance by taking into account both the spatial and temporal structure of the video, using convolutional Long Short-Term Memory (LSTM) networks (Lal et al., 2019; Yuan et al., 2019a), optical flow maps (Chu et al., 2019), combinations of Convolutional Neural Networks (CNNs) and RNNs (Elfeki & Borji, 2019), or motion extraction mechanisms (Huang & Wang, 2020). Alternatively, some works (Fajtl et al., 2019; Apostolidis et al., 2021c; Ghauri et al., 2021; P. Li et al., 2021; Yao et al., 2022; Puthige et al., 2023) aimed to avoid the use of computationally-demanding RNNs, and instead they modeled the frames' dependencies at various temporal granularities using variants of the self-attention mechanism of the Transformer Network (Vaswani et al., 2017).

To overcome issues related to the small amount of ground-truth data, most unsupervised approaches tried to learn how to build summaries that are highly representative of the video content. Based on the intuition that a good summary should allow the viewer to infer the original video, they used Generator-Discriminator architectures along with adversarial learning mechanisms that force the summarization component (which is usually a part of the Generator) to build a summary that allows a good reconstruction of the original video. In early approaches, the summarization component was composed of LSTM units that estimated the frames' importance according to their temporal dependence (thus indicating the most significant video parts for inclusion in the summary), while the reconstruction of the video based on the specified summary was performed using trainable auto-encoders (Mahasseni et al., 2017; Apostolidis et al., 2019; Yuan et al., 2020) that in some cases were combined with tailored attention mechanisms (Jung et al., 2019; Apostolidis et al., 2020; Kanafani et al., 2021). In more recent methods, the selection of the most important frames or fragments for the summary was assisted by trainable Actor-Critic models (Apostolidis et al., 2021a; Alexoudi et al., 2023), self-attention mechanisms (He et al., 2019; Jung

et al., 2020; Liang et al., 2022), spatio-temporal networks (Wu et al., 2021) or knowledge distillation mechanisms (Sreeja & Kovoor, 2022). A less popular approach for unsupervised video summarization is based on the definition of hand-crafted reward functions about specific properties of the generated summary, and the use of the computed rewards for training video summarization architectures based on reinforcement learning. Usually, these rewards aim to increase the representativeness, diversity (Zhou et al., 2018a; Phaphuangwittayakul et al., 2021; T. Liu et al., 2022; Zang et al., 2022) and uniformity (Yaliniz & Ikizler-Cinbis, 2021; Hu et al., 2022) of the summary, retain the spatio-temporal patterns of the video (Gonuguntla et al., 2019), allow a good summary-based video reconstruction (Zhao et al., 2020), or maintain specific shot-level semantics of the video shooting and production process (e.g., camera angle and movement, and focus adjustments) (Yuan & Zhang, 2022). Finally, a couple of recent works train unsupervised video summarization networks based on contrastive learning (Pang et al., 2023; Sosnovik et al., 2023).

With regards to multimodal video summarization methods, early approaches extracted the high-level semantics of the visual content using pre-trained CNNs/DCNNs and tried to learn summarization in a supervised manner by maximizing the semantic similarity among the visual summary and the contextual video metadata (Otani et al., 2016; Yuan, 2019b), the video category (Zhou et al., 2018b; Lei et al., 2019), or human descriptions of the video content (Wei et al., 2018). More recent works, presented multimodal network architectures for topic- and query-driven video summarization. Apart from the visual content, these architectures take into account the users' requirements about the content of the summary, that are usually expressed in the form of textual queries, keywords or sentences (Zhang et al., 2018, 2019; Huang & Worring, 2020; Xiao et al., 2020a, 2020b; Narasimhan et al., 2021; Hu et al., 2023; Zhu et al., 2023; Su et al., 2023). Finally, a few works explore the role of audio when used in combination with the visual stream of the video (Zhao et al., 2021b; Rhevanth et al., 2022; Shoer et al., 2022; Psallidas et al., 2022) and some external knowledge base (Xie et al., 2022) for selecting the parts for inclusion in the video summary.

Explainable Video Analysis and Understanding

Over the last years there is a rapidly growing interest of researchers on building methods that provide explanations about the working mechanism or the decisions/predictions of deep neural networks. Nevertheless, in contrast to the notable progress in the fields of pattern recognition (Bai et al., 2021) and NLP (El Zini & Awad, 2022), there are only a few works on producing explanations for neural networks that process video data. Roy et al. (2019) fed the output of a model for activity recognition to a tractable interpretable probabilistic graphical model and performed joint learning over the two. Aakur et al. (2018) built a framework for producing inherently explainable and semantically coherent representations for video activity interpretation. Zhuo et al. (2019) defined a spatio-temporal graph of semantic-level video states and applied state transition analysis for video action reasoning. Stergiou et al. (2019) formed explanations of deep networks for action classification and recognition, using cylindrical heat-maps that visualize the focus of attention. Gkalelis et al. (2022) 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. (2020) extended the concept of meaningful perturbation, to spot the video fragment with the greatest impact on the video classification results. Bargal et al. (2018) 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 ac-

tion or phrase from the caption. Z. Li et al. (2021) 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. Finally, Yu et al. (2021) built an end-to-end trainable and interpretable framework for video text detection with online tracking that captures spatial and motion information and uses an appearance-geometry descriptor to generate robust representations of text instances.

FOCUS OF THE ARTICLE

This chapter focuses on attention-based explainable video summarization. To form the basis for presenting solutions and recommendations for this task, in this section we start by further discussing video summarization methods that rely on the use of attention mechanisms. Following, we report on a few existing works (mainly from the NLP domain) that investigated the use of attention for explaining the output of deep networks.

Attention-Based Video Summarization

As reported above, early deep-learning-based video summarization methods relied mainly on the use of RNNs for modelling the temporal and/or spatio-temporal dependence of the video frames/fragments and learn how to estimate the frames'/fragments' importance for video summarization. However, further investigation of the performance and the training of these methods indicated some shortcomings of using RNNs for video summarization, that were discussed in the literature (Vaswani et al., 2017; Fajtl et al., 2019; Zhao et al., 2020; P. Li et al., 2021). The main weakness relates to the long paths that forward and backward propagation signals have to traverse in the video summarization network, which negatively affects the network's capacity to deal with long-range frames' dependencies. In addition, by nature, RNNs exhibit low levels of parallelizable operations during training, a fact that can become critical for long training samples or when the video summarization architecture has to be re-trained (by either a service provider or the user) on different types of content.

To mitigate the aforementioned shortcomings, a few visual-based works for video summarization investigated the use of variants of the self-attention mechanism of the Transformer Network (Vaswani et al., 2017) for modeling the frames' dependence and learn how to estimate the frames' importance for video summarization. Fajtl et al. (2019) were the first to combine a soft self-attention mechanism with a two-layer fully connected network for regression of the frames' importance scores. Liu et al. (2019) described a hierarchical approach which initially defines a set of shot-level candidate key-frames, and then it employs a multi-head attention model to further assess candidates' importance and select the key-frames that form the summary. P. Li et al. (2021) extended the training pipeline of the typical self-attention mechanism, by introducing a processing step that uses the computed attention values and tries to increase the diversity of the visual content of the summary. The estimated attention values (after incorporating information about the frames' diversity) are used to estimate the frames' importance and learn summarization from human-based ground-truth annotations. Ghauri et al. (2021) proposed a variation of the architecture from (Fajtl et al., 2019), that uses additional representations of the video content. Besides the typical CNN-based features (obtained from pool5 layer of GoogleNet (Szegedy et al., 2015) trained on ImageNet), Ghauri et al. (2021) used a model of the Inflated 3D ConvNet (Carreira & Zisser-

man, 2017) trained on Kinetics, to extract a set of motion-related features. Each different set of features is fed to a self-attention mechanism and the outputs of these mechanisms are fused to form a common embedding space for representing the video frames. The obtained representation is finally used to learn how to estimate the frames' importance. Apostolidis et al. (2021c) described a supervised architecture which discovers different modelings of the frames' dependencies at different levels of granularity, using global and local multi-head attention mechanisms that integrate a component for taking into account the temporal position of the video frames. A similar approach was presented by Yao et al. (2022), which spots the video's most important moments using multi-level representations extracted with the help of different attention mechanisms, and performs video summarization by taking into account also the fragments' diversity. In a subsequent work, Apostolidis et al. (2022a) proposed an unsupervised network architecture which estimates the frames' importance using a novel 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. Puthige et al. (2023) described another attention-based approach which firstly process the video frames using a multi-head attention mechanism with positional encoding, and then passes them through a spatial and a channel attention mechanism in order to capture inter-spatial and inter-channel relationships between the frames' representations.

Concerning multimodal attention-based approaches, Narasimhan et al. (2021) proposed a network architecture for language-guided video summarization. Instead of modeling the frames' dependence based solely on their visual content, their architecture includes an attention mechanism that takes into account both the visual content of the video frames (as Query) and textual information about the entire video/video summary in the form of dense captions/textual queries (as Key and Value). Su et al. (2023) presented a method for query-focused video summarization that relies on a global attention mechanism and a query-aware multi-modal regression module that fuses visual and textual features according to different perspectives for learning how to estimate the frames' importance. Finally, He et al. (2023) developed an architecture that includes an attention mechanism which can align and attend different modalities (visual stream and textual transcripts) by leveraging time correspondences. To train their architecture, He et al. (2023) introduced the use of dual contrastive losses with the combination of an inter-sample and an intra-sample contrastive loss, to model the cross-modal correlation at different granularities.

Attention-Based Explanation of Deep Networks

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) 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) 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) 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) 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 (Devlin, 2019). Hao et al. (2021) assessed the performance of explanations formulated using gradient-based attention weights and the BERT model for text classification. Chrysostomou and Aletras (2021) 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 (Chrysostomou & Aletras, 2022) that focused on evaluating their out-of-domain faithfulness. Y. Liu et al. (2022) 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 (Ntrougkas et al., 2022; Gkartzonika et al., 2023), image recognition (L. Li et al., 2021), heart sound classification (Ren et al., 2022) and multimodal trajectory prediction (Zhang & Li, 2022).

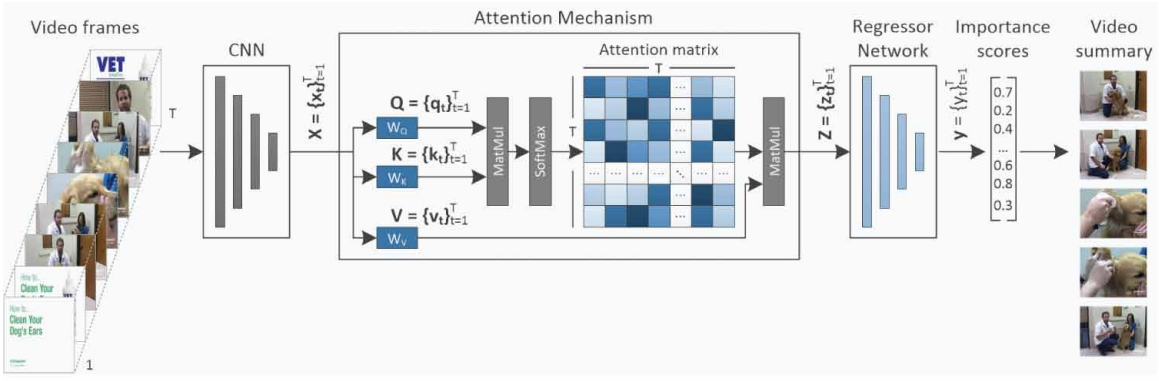
SOLUTIONS AND RECOMMENDATIONS

In this section, we present an unsupervised network architecture for video summarization, that relies on the use of a concentrated attention mechanism. Following, we report on a recent work that formulates the task of explaining video summarization and investigates the use of various attention-based explanation signals.

The CA-SUM Method for Video Summarization

The CA-SUM method is built upon the main processing pipeline of attention-based approaches for video summarization, which is illustrated in Fig. 1. Given a video of T frames and a pretrained CNN model for deep feature extraction, the attention mechanism gets as input the deep feature representations of the video frames $\left(\boldsymbol{X} = \left\{\boldsymbol{x_i}\right\}_{i=1}^{T}\right)$. Following, it produces the Query- and Key-based transformations of them ($\boldsymbol{Q} = \left\{\boldsymbol{q_i}\right\}_{i=1}^{T}$ and $\boldsymbol{K} = \left\{\boldsymbol{k_i}\right\}_{i=1}^{T}$, respectively), performs a matrix multiplication $\left(\boldsymbol{Q} \times \boldsymbol{K}^{-1}\right)$, 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 $\left(\boldsymbol{A} = \left\{a_{i,j}\right\}_{i,j=1}^{T} with\, a_{i,j} \in \left[0,1\right]\right)$.

Figure 1. The typical processing pipeline of attention-based video summarization approaches. Blue-coloured parts indicate the trainable components of the network architecture.



Each row of this matrix corresponds to a different frame of the video and the values within each row represent the significance of the associated frame for each frame of the video, according to the context modelled by the attention mechanism. This matrix is multiplied with the Value-based transformation of the input feature representations $\left(\boldsymbol{V}=\left\{\boldsymbol{v}_{\boldsymbol{i}}\right\}_{i=1}^{T}\right)$ and forms the output of the attention mechanism; i.e., a new set of feature representations $\left(\boldsymbol{Z}=\left\{\boldsymbol{z}_{\boldsymbol{i}}\right\}_{i=1}^{T}\right)$ that convey information about the relevance of each video frame with the context modelled by the attention mechanism. This output goes through a Regressor Network, which produces the frames' importance scores $\left(\boldsymbol{y}=\left\{y_{i}\right\}_{i=1}^{T}\right)$. These scores are finally used to compute fragment-level importance and select the most important fragments for inclusion in the video summary.

The main idea behind the development of the CA-SUM method was the intuition that the computed attention matrix can capture information about the dependence and significance of different parts of the video in the learned latent space, and this information can be used to learn better estimates about the importance of these fragments. Building on this, Apostolidis et al. (2022a) focused on specific parts of the attention matrix, which correspond to different non-overlapping video fragments of fixed length. More specifically, these parts are non-overlapping blocks of size M, that lie in the main diagonal of the attention matrix. Instead of simply computing the mean value of the attention weights within each block and use this value as an estimate of the significance of the entire block, Apostolidis et al. (2022a) enriched the existing information by extracting and exploiting knowledge about the uniqueness and diversity of the associated frames of the video. Given the i^{th} frame of the video, the former was measured by computing the L1-norm of the entropy of the corresponding row of the attention matrix: $e_{i}=\left\|-\sum_{t=1}^{T} a_{i,t} \cdot \log\left(a_{i,t}\right)\right\|_{1}$. The latter was quantified by calculating the mean of its attention-based weighted dissimilarities from the frames that lie outside the block: $d_{i}=\frac{1}{\sum_{j} Dis\left(i,j\right)} \sum_{j}\left(Dis\left(i,j\right) \cdot a_{i,j}\right)$, where j is an index the lies outside the block of interest and $Dis\left(i,j\right)$ is the cosine distance between the i^{th} and the j^{th} frame of the video:

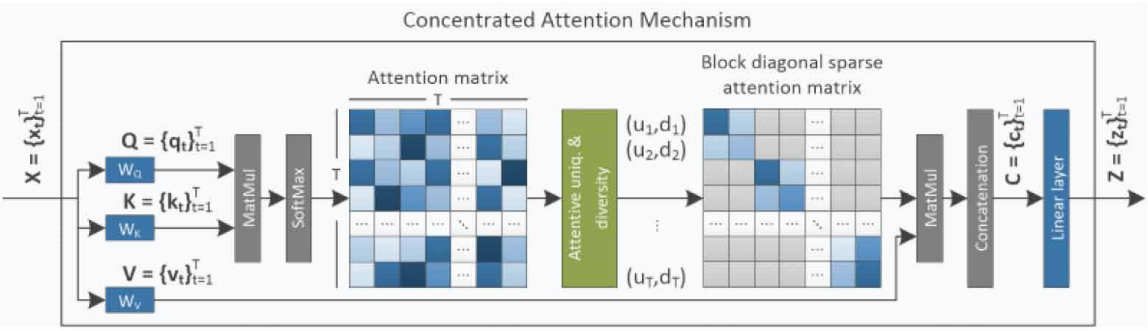
$$Dis\left(i,j\right)=1-\frac{\boldsymbol{x}_{\boldsymbol{i}} \cdot \boldsymbol{x}_{\boldsymbol{j}}^{-1}}{\left\|\boldsymbol{x}_{i}\right\|_{2} \cdot \left\|\boldsymbol{x}_{j}\right\|_{2}}$$

The use of the computed attentive uniqueness and diversity values as described in (Apostolidis et al., 2022a), results in the production of a block diagonal sparse attention matrix that contains better estimates about the significance of different parts of the video, and reduces the number of learnable parameters (see Fig. 2).

To train CA-SUM, Apostolidis et al. (2022a) use the following sparsity loss:

$$L=\left|\frac{1}{T}\sum_{i=1}^{T} y_{i}-\sigma\right|$$

Figure 2. The concentrated attention mechanism of CA-SUM. Blue-coloured parts indicate the trainable components. The green-coloured box represents the part of the mechanism responsible for computing the attentive uniqueness and diversity of the video frames.



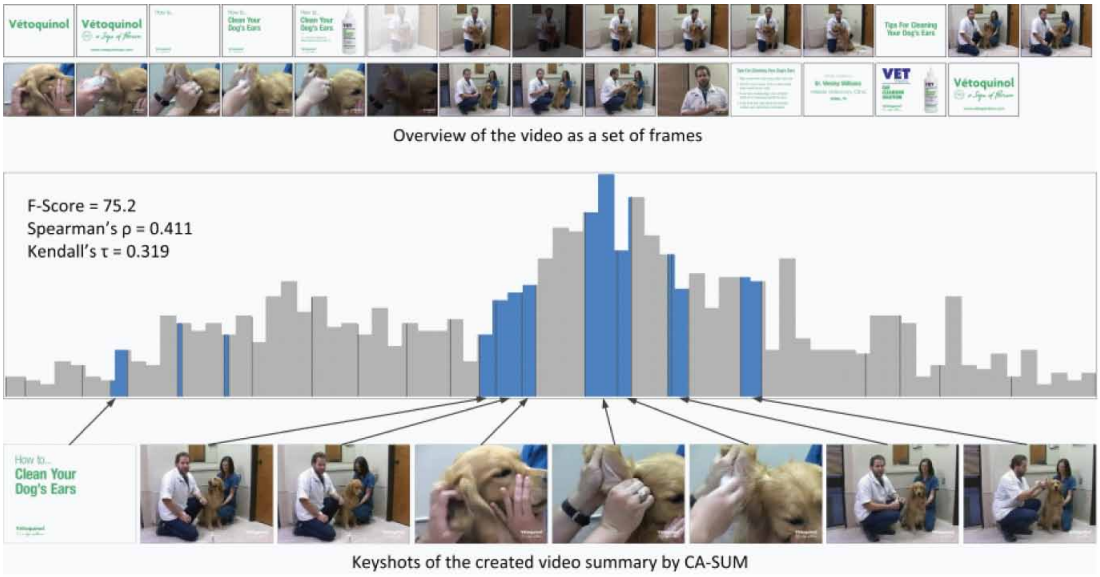
where, y_t is the method's estimate about the importance of the i^{th} video frame, and σ is the summary length regularization factor, a tunable hyper-parameter of CA-SUM (such a factor was introduced by Mahasseni et al., (2017) and used in several subsequent works (Zhou et al., 2018; Phaphuangwittayakul et al., 2021; P. Li et al., 2021)). The computed training loss is then back-propagated to compute the gradients and update the architecture.

During inference, the estimated importance scores for the video frames are used to select the key-fragments of the video and form the video summary. For this, given a temporal segmentation of the video into its building blocks (obtained e.g., using the Kernel Temporal Segmentation (KTS) algorithm of Potapov et al. (2014)), fragment-level importance is calculated by averaging the scores of the frames within each fragment. Finally, requiring that the summary does not exceed 15% of the video duration - which is a common evaluation-protocol setting in the relevant literature (Apostolidis et al., 2021b) - the video summary is formed by solving the Knapsack problem.

The performance of CA-SUM was evaluated using the SumMe (Gygli et al., 2014) and TVSum (Song et al., 2015) benchmarking datasets, according to the overlap (alignment) of the automatically-defined summaries (frames' importance scores) with the humans' preferences, as well as in terms of training time and amount of learnable parameters. Experimental comparisons reported in (Apostolidis et al., 2022a), showed the competitive performance of CA-SUM against a few other state-of-the-art unsupervised summarization approaches, and demonstrated its ability to produce estimates about the frames' importance that are very close to the human preferences. Ablations focusing on the concentrated attention mechanism of CA-SUM, documented its positive contribution to the overall summarization performance. Finally, measurements with respect to the training time and the number of learnable parameters indicated the time efficiency and the small memory footprint of the network architecture.

An example of an automatically-created video summary by the CA-SUM method, and its overlap with the human annotations is illustrated in Fig. 3. The upper part of this figure gives an overview of the video after selecting one frame per shot (shot segmentation performed using the KTS algorithm (Potapov et al., 2014)), and the lower part presents the results for the CA-SUM method. The gray bars denote the averaged human-annotated importance scores for the frames of the video, the black vertical lines within these bars correspond to the shot boundaries, and the blue-coloured bars

Figure 3. A key-frame-based overview (using one key-frame per shot), and the created video summary from CA-SUM for a video of the TVSum dataset, titles "How to Clean Your Dog's Ears - Vetoquinol USA". Gray bars denote the averaged human-annotated importance scores for the frames of the video, the black vertical lines within these bars correspond to the shot boundaries, and the coloured bars indicate the selected key-shots for creating the summary. The created summary is depicted by selecting one representative key-frame from each one of its major key-shots.



indicate the selected key-shots for inclusion in the summary. The generated summary is illustrated by selecting one representative key-frame from each one of the major key-shots of the summary. Moreover, the upper left part of the bar chart shows the performance of the CA-SUM method for this video, in terms of F-Score as percentage (which is the most-commonly used measure in the literature (Apostolidis et al., 2021b)) and according to the Spearman's ρ (rho) and Kendall's τ (tau) correlation coefficients that were proposed as alternative evaluation measures by Otani et al. (2019). The generated summary focuses on the main event of the video (i.e., the cleaning of the dog's ears), but it also contains shots with diverse visual content from other parts of the video. In this way, it provides a comprehensive presentation of the entire story, with a special focus on its main event. Moreover, taking into account the average human performance reported in the literature for the videos of the TVSum dataset (Otani et al., 2019), namely an F-Score equal to 78.0, a Spearmans' ρ equal to 0.204, and a Kendall's τ equal to 0.177, we see that the CA-SUM method achieved a human-level performance on this example video (see F-Score value) while its estimates about the frames' importance exhibit a remarkable similarity with the humans' estimates about the importance of different parts of this video (see ρ and τ values).

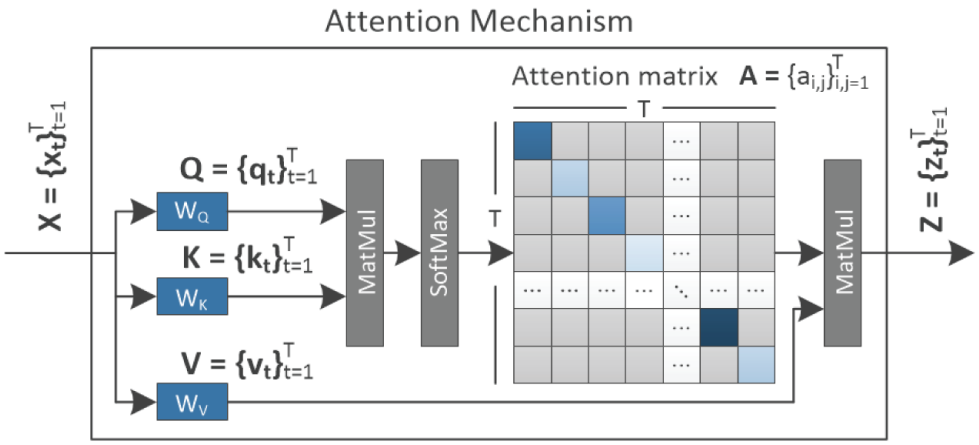
To allow reproduction of the reported results and testing of the proposed method, Apostolidis et al. (2022a) made the PyTorch implementation of CA-SUM publicly-available at: https://github.com/e-apostolidis/CA-SUM

The XAI-SUM Method for Explaining Video Summarization

A first attempt towards explaining video summarization was made by Apostolidis et al. (2022b). Driven by the fact that there were no relevant works in the literature, Apostolidis et al. (2022b) formulated the goal of this task as the production of an explanation mask that highlights the top-*M* video fragments that influenced the most the estimates of a video summarization network about the frames' importance, and thus the generation of the video summary. With respect to the latter, to avoid the influence of any utilized video fragmentation and key-fragment selection approach to the generated video summary (discussed in (Otani et al., 2019)), Apostolidis et al. (2022b) adopted a more straightforward approach. This approach: i) splits the video into consecutive and non-overlapping fragments of fixed-size H, ii) calculates each fragment's importance by averaging the importance of the frames in it, and iii) forms the summary by selecting the *M* top-scoring video fragments.

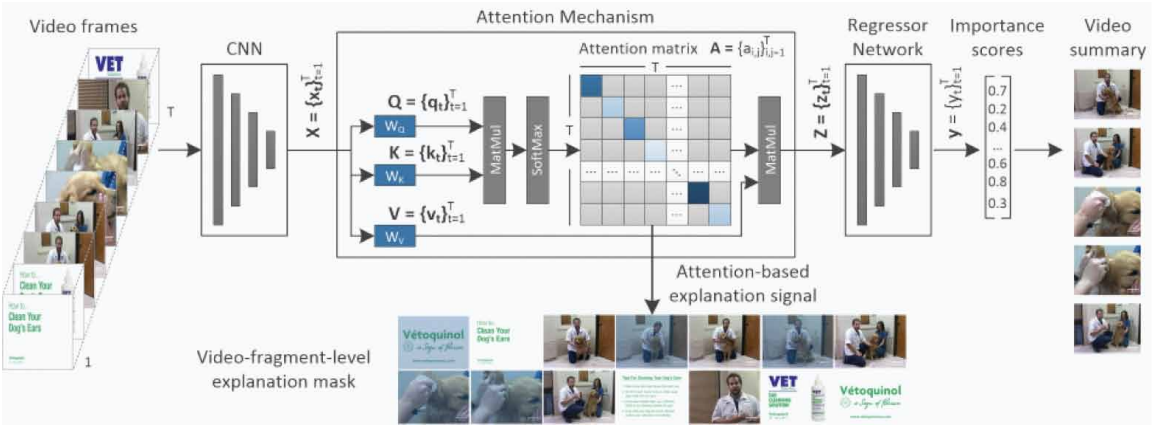
Following, inspired by existing works from the NLP domain, that investigate the use of attention as explanation (discussed in section "Attention-Based Explanation of Deep Networks" of this chapter), Apostolidis et al. (2022b) focused their study on video summarization architectures that rely on the use of attention mechanisms. Initially, they presented the typical processing pipeline of attention-based video summarization architectures from the literature, and explained how this pipeline can be used to define attention-based explanation signals. In particular, they concentrated on the values in the main diagonal of the attention matrix (see Fig. 4), and considered various explanation signals, formed by: i) the weights in the main diagonal of the attention matrix $\left\{a_{i,i}\right\}_{i=1}^{T}$ (Inherent Attention), ii) the gradients of the importance estimation layer with respect to the weights in the main diagonal of the attention matrix $\left\{\nabla a_{i,i}\right\}_{i=1}^{T}$ (Gradient of Attention), iii) a weighted version of the weights in the main diagonal of the attention matrix, according to the gradients for their computation $\left\{a_{i,i} \cdot \nabla a_{i,i}\right\}_{i=1}^{T}$ (Grad Attention), iv) a weighted version of the weights in the main diago-

Figure 4. An illustration of how the attention mechanism can be used to form explanation signals for interpreting the output of attention-based video summarization network architectures



nal of the attention matrix, according to the norm of the Value-based transformed input vectors in the attention mechanism $\left\{a_{i,i} \cdot \left\|\boldsymbol{u}_i\right\|\right\}_{i=1}^{T}$ (Input Norm Attention), and v) a weighted version of the weights in the main diagonal of the attention matrix, according to both the gradients for their computation and the norm of the Value-based transformed input vectors in the attention mechanism $\left\{a_{i,i} \cdot \nabla a_{i,i} \cdot \left\|\boldsymbol{u}_i\right\|\right\}_{i=1}^{T}$ (Input Norm Grad Attention). In each case, the obtained frame-level explanation scores are averaged at the fragment-level, and the top-M fragments with the highest explanation scores are the ones highlighted in the created explanation mask (see Fig. 5).

Figure 5. Overview of the XAI-SUM method for obtaining attention-based explanation masks about 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 in the explanation mask equals to five. The summary is formed by stitching the top-5 fragments (according to their importance) in chronological order.



To investigate the model's input-output relationship, Apostolidis et al. (2022b) applied various replacement functions at parts of the input corresponding to different video fragments. In particular, they: i) removed a part from the original input (Slice-out), ii) replaced a part of the original input with a predefined mask formed by the deep feature representations of black or white frames (Input Mask), iii) replaced 50% of the elements of each feature representation within a part of the original input, using the corresponding elements from randomly-selected feature representations from the remaining parts of the input (Randomization), and iv) set the attention weights associated with a part of the original input, equal to zero, such that this part will not be forwarded in the network anymore.

Given the above discussed replacement functions, to measure the influence of the k^{th} video fragment in the video summarization network's output, Apostolidis et al. (2022b) computed the difference of estimates $\Delta E\left(\boldsymbol{X}, \hat{\boldsymbol{X}}^{\boldsymbol{k}}\right) = \tau\left(\boldsymbol{y}, \boldsymbol{y}^{\boldsymbol{k}}\right)$, 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 network's outputs for $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}^{\boldsymbol{k}}$ respectively, and $\tau\left(\right)$ computes the Kendall's

τ correlation coefficient. Then, the performance of the different explanation signals was evaluated based on the following measures:

- **Discoverability+ ($\mathbf{D^+}$)** evaluates if fragments with high explanation scores have a significant influence to the network's estimates; $\mathbf{D^+ = Mean(\Delta E)}$ after replacing the top-1%, 5%, 10%, 15%, 20%, and the 5 top-scoring fragments.
- **Discoverability- ($\mathbf{D^-}$)** evaluates if fragments with low explanation scores have small influence to network's estimates; $\mathbf{D^- = Mean(\Delta E)}$ after replacing the bottom-1%, 5%, 10%, 15%, 20%, and the 5 less-scoring fragments.
- **Sanity Violation (SV)** quantifies the ability of explanations to discriminate important from unimportant video fragments; **SV = %** of cases where $\mathbf{D^+ < D^-}$ holds true.
- **Rank Correlation (RC)** measures the (Spearman) correlation between fragment-level explanation scores and the obtained ΔE values after replacing each video fragment.

The experimental evaluations were conducted using the CA-SUM method and the SumMe and TVSum benchmarking datasets. The findings of these evaluations showed that explanations formed using the attention weights (Inherent Attention) exhibit the best performance, as they achieve the lowest/highest **Discoverability-/+** scores and, in most cases correctly discriminate the most and least influential fragments of the video. Moreover, based on the computed **Rank Correlation**, they are capable of assigning fragment-level explanation scores that are more representative of each fragment's influence to the network's output. Explanations formed using the norm-based weighted version of the inherent attention weights (Input Norm Attention) also perform good in terms of **Discoverability-/+**, but are less effective in terms of **Sanity Violation**. Finally, explanations formed using the gradients of the attention weights (Gradient of Attention, Grad Attention, Input Norm Grad Attention) are the worst-performing ones, as they lead to higher/lower **Discoverability-/+** scores than the ones obtained for non-gradient-based signals, systematically fail to distinguish the most and least influential video fragments, and assign fragment-level explanation scores that are neutrally or negatively correlated with the influence of each fragment to the network's output.

An example of the produced explanation mask using Inherent Attention (IA) is shown in Fig. 6. The blue-coloured semi-transparent overlays signify the most influential fragments according to the utilized IA-based explanation signal, and the blue-coloured bounding boxes indicate the top-scoring fragments according to the model's estimations. In the example video of Fig. 6 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. This paradigm shows that extracting explanations using the inherent attention weights of the attention mechanism and the method proposed in (Apostolidis et al., 2022b), could allow to get insights about the focus of attention and assist the explanation of video summarization networks similar to CA-SUM.

To allow reproduction of the reported results and testing of the proposed XAI-SUM method, Apostolidis et al. (2022b) made the PyTorch implementation publicly-available at: https://github.com/e-apostolidis/XAI-SUM

Figure 6. The produced explanation mask for a video of the TVSum dataset, titled "How to Clean Your Dog's Ears – Vetoquinol". Video fragments are illustrated using their most representative frame. Blue-coloured semi-transparent overlays signify the most influential fragments for the model's predictions. Blue-coloured bounding boxes indicate the five top-scoring video fragments based on the model's predictions.

FUTURE RESEARCH DIRECTIONS

Although the research community has invested considerable effort in the video summarization problem, this problem cannot be considered as solved and there is plenty of room for improvements. In our perspective, the most promising research direction is the extension of existing for video summarization methods in order to allow a user to intervene in the summary production process, so that the outcome (i.e., the video summary) is aligned with user-specified rules. Future work in this area could build on existing methods for query- or sentence-driven summarization (e.g., Narasimhan et al., 2021; Hu et al., 2023; Su et al., 2023), and explore ways to offer personalized video summaries that are customized to the users' needs. A more aspiring approach could be the use of an on-line interaction channel between the user/editor and the trainable summarizer, in combination with active learning algorithms that allow to incorporate the user's/editor's feedback with respect to the generated summary (as in (Garcia del Molino et al., 2017)). The development of effective weakly-supervised video summarization methods, either via a text-based input that specifies the content of the summary or via an on-line interaction channel, will allow meeting the needs of specific summarization scenarios and application domains. Such developments will be really important for the practical application of video summarization technologies in the Media industry, where complete automation that diminishes editorial control over the generated summaries is not always preferred.

With respect to explainable video summarization, future research should target the extension of existing video summarization architectures by introducing explanation mechanisms; both model-agnostic and model-dependent mechanisms could be investigated. This research could be guided by transferring knowledge from other domains where such mechanisms have already been used with success (e.g., action/event recognition and classification (e.g., Stergiou et al., 2019; Gkalelis et al., 2022), image/video classification (e.g., Mänttäri et al., 2020; Z. Li, 2021; Ntrougkas et al., 2022; Gkartzonika et al., 2023) and video text detection (e.g., Yu et al., 2021)) to the video summarization domain. Moreover, future work on the XAI-SUM method (Apostolidis et al., 2022b) could involve experimentation with additional attention-based networks for video summarization (e.g., the ones proposed by Fajtl et al. (2019) and P. Li et al. (2021)), as well as comparisons with other model-agnostic or model-dependent methods for producing explanations.

CONCLUSION

The recent advances in the field of video summarization, that are heavily based on the emergence of modern deep-learning network architectures, form a fertile ground for the development of technologies that could significantly facilitate the work of professionals working on media content production; especially the ones dealing with the production of summarized versions of a video for distribution via different communication channels. The current state of the art on video summarization is mainly represented by methods that utilize attention mechanisms and can support either generic or query-driven video summarization. A recently proposed method for unsupervised video summarization (CA-SUM) showed that the use of tailored attention mechanisms can lead to analysis results (i.e., video summaries) that are very close to the humans' expectations. The combination of such methods, with attention-based approaches for explaining the analysis results (XAI-SUM) could enable a level of understanding about the functionality of the video summarization method, thus increasing the professionals' trust in this technology and facilitating media content curation and production.

ACKNOWLEDGMENT

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

REFERENCES

Aakur, S. N., Souza, F. D., & Sarkar, S. (2018). An Inherently Explainable Model for Video Activity Interpretation. *Proceedings of the 32nd AAAI Conference on Artificial Intelligence, Workshops.*

Alexoudi, P., Mademlis, I., & Pitas, I. (2023). Escaping local minima in Deep Reinforcement Learning for video summarization. In *Proceedings of the 2023 International Conference on Multimedia Retrieval (ICMR '23)*. Association for Computing Machinery. 10.1145/3591106.3592288

Apostolidis, E., Adamantidou, E., Metsai, A. I., Mezaris, V., & Patras, I. (2020). Unsupervised Video Summarization via Attention-Driven Adversarial Learning. In Lecture Notes in Computer Science (pp. 492–504). Springer Science+Business Media. doi:10.1007/978-3-030-37731-1_40

Apostolidis, E., Adamantidou, E., Metsai, A. I., Mezaris, V., & Patras, I. (2021a). AC-SUM-GAN: Connecting Actor-Critic and Generative Adversarial Networks for Unsupervised Video Summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, *31*(8), 3278–3292. doi:10.1109/TCSVT.2020.3037883

Apostolidis, E., Adamantidou, E., Metsai, A. I., Mezaris, V., & Patras, I. (2021b). Video Summarization Using Deep Neural Networks: A Survey. *Proceedings of the IEEE*, *109*(11), 1838–1863. doi:10.1109/JPROC.2021.3117472

Apostolidis, E., Balaouras, G., Mezaris, V., & Patras, I. (2020b). Explaining video summarization based on the focus of attention. *Proceedings of the 2022 IEEE International Symposium on Multimedia (ISM)*, 146-150. 10.1109/ISM55400.2022.00029

Apostolidis, E., Balaouras, G., Mezaris, V., & Patras, I. (2021c). Combining Global and Local Attention with Positional Encoding for Video Summarization. *Proceedings of the 2021 IEEE International Symposium on Multimedia (ISM)*, 226-234. 10.1109/ISM52913.2021.00045

Apostolidis, E., Balaouras, G., Mezaris, V., & Patras, I. (2022a). Summarizing Videos using Concentrated Attention and Considering the Uniqueness and Diversity of the Video Frames. In *Proceedings of the 2022 International Conference on Multimedia Retrieval (ICMR '22)*. Association for Computing Machinery. 10.1145/3512527.3531404

Apostolidis, E., Metsai, A. I., Adamantidou, E., Mezaris, V., & Patras, I. (2019). A Stepwise, Label-based Approach for Improving the Adversarial Training in Unsupervised Video Summarization. In *Proceedings of the 1st International Workshop on AI for Smart TV Content Production, Access and Delivery (AI4TV '19).* Association for Computing Machinery. 10.1145/3347449.3357482

Bai, X., Wang, X., Liu, X., Liu, Q., Song, J., Sebe, N., & Kim, B. (2021). Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments. *Pattern Recognition*, *120*, 108102. doi:10.1016/j.patcog.2021.108102

Bargal, S., Zunino, A., Kim, D., Zhang, J., Murino, V., & Sclaroffet, S. (2018). Excitation Backprop for RNNs. *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1440-1449. 10.1109/CVPR.2018.00156

Carreira, J., & Zisserman, A. (2017). Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4724-4733. 10.1109/CVPR.2017.502

Chrysostomou, G., & Aletras, N. (2021). Improving the Faithfulness of Attention-based Explanations with Task-specific Information for Text Classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (*Volume 1*: Long Papers, pp. 477–488). Association for Computational Linguistics. 10.18653/v1/2021.acl-long.40

Chrysostomou, G., & Aletras, N. (2022). An Empirical Study on Explanations in Out-of-Domain Settings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers, pp. 6920–6938). Association for Computational Linguistics. 10.18653/v1/2022.acl-long.477

Chu, W.-T., & Liu, Y.-H. (2019). Spatiotemporal Modeling and Label Distribution Learning for Video Summarization. *Proceedings of the 2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, 1-6. 10.1109/MMSP.2019.8901741

Collyda, C., Apostolidis, K., Apostolidis, E., Adamantidou, E., Metsai, A. I., & Mezaris, V. (2020). A Web Service for Video Summarization. In *Proceedings of the ACM International Conference on Interactive Media Experiences (IMX '20)*. Association for Computing Machinery. 10.1145/3391614.3399391

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (vol. 1, pp. 4171–4186). Association for Computational Linguistics.

El Zini, J., & Awad, M. (2022). On the Explainability of Natural Language Processing Deep Models. *ACM Computing Surveys*, *55*(5), 1–31. doi:10.1145/3529755

Elfeki, M., & Borji, A. (2019). Video Summarization Via Actionness Ranking. *Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 754-763. 10.1109/WACV.2019.00085

Fajtl, J., Sokeh, H. S., Argyriou, V., Monekosso, D., & Remagnino, P. (2019). Summarizing Videos with Attention. In Lecture Notes in Computer Science (pp. 39–54). Springer Science+Business Media. doi:10.1007/978-3-030-21074-8_4

Feng, L., Li, Z., Kuang, Z., & Zhang, W. (2018). Extractive Video Summarizer with Memory Augmented Neural Networks. In *Proceedings of the 26th ACM international conference on Multimedia (MM '18)*. Association for Computing Machinery. 10.1145/3240508.3240651

Garcia del Molino, A., Boix, X., Lim, J.-H., & Tan, A.-H. (2017). Active Video Summarization: Customized Summaries via On-line Interaction with the User. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1). Advance online publication. doi:10.1609/aaai.v31i1.11234

Ghauri, J., Hakimov, S., & Ewerth, R. (2021). Supervised Video Summarization Via Multiple Feature Sets with Parallel Attention. *Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME)*, 1-6. 10.1109/ICME51207.2021.9428318

Gkalelis, N., Daskalakis, D., & Mezaris, V. (2022). ViGAT: Bottom-Up Event Recognition and Explanation in Video Using Factorized Graph Attention Network. *IEEE Access: Practical Innovations, Open Solutions*, *10*, 108797–108816. doi:10.1109/ACCESS.2022.3213652

Gkartzonika, I., Gkalelis, N., & Mezaris, V. (2023). Learning Visual Explanations for DCNN-Based Image Classifiers Using an Attention Mechanism. In Lecture Notes in Computer Science, 13808. Springer. doi:10.1007/978-3-031-25085-9_23

Gonuguntla, N., Mandal, B., & Puhan, N. (2019). Enhanced Deep Video Summarization Network. *Proceedings of the 2019 British Machine Vision Conference (BMVC).*

Gygli, M., Grabner, H., Riemenschneider, H., & Van Gool, L. (2014). Creating Summaries from User Videos. In Lecture Notes in Computer Science (pp. 505–520). Springer Science+Business Media. doi:10.1007/978-3-319-10584-0_33

Hao, Y., Dong, L., Wei, F., & Xu, K. (2020). Self-Attention Attribution: Interpreting Information Interactions Inside Transformer. *Proceedings of the 34th AAAI Conference on Artificial Intelligence*.

He, B., Wang, J., Qiu, J., Bui, T., Shrivastava, A., & Wang, Z. (2023). Align and Attend: Multimodal Summarization With Dual Contrastive Losses. *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10.1109/CVPR52729.2023.01428

He, X., Hua, Y., Song, T., Zhang, Z., Xue, Z., Ma, R., Robertson, N., & Guan, H. (2019). Unsupervised Video Summarization with Attentive Conditional Generative Adversarial Networks. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19)*. Association for Computing Machinery. 10.1145/3343031.3351056

Hu, M., Hu, R., Wang, Z., Xiong, Z., & Zhong, R. (2022). Spatiotemporal two-stream LSTM network for unsupervised video summarization. *Multimedia Tools and Applications*, *81*(28), 40489–40510. doi:10.100711042-022-12901-4

Hu, W., Zhang, Y., Li, Y., Zhao, J., Hu, X., Cui, Y., & Wang, X. (2023). Query-based video summarization with multi-label classification network. *Multimedia Tools and Applications*. Advance online publication. doi:10.100711042-023-15126-1

Huang, C. Z., & Wang, H. (2020). A Novel Key-Frames Selection Framework for Comprehensive Video Summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*(2), 577–589. doi:10.1109/TCSVT.2019.2890899

Huang, J.-H., & Worring, M. (2020). Query-controllable Video Summarization. In *Proceedings of the 2020 International Conference on Multimedia Retrieval (ICMR '20)*. Association for Computing Machinery. 10.1145/3372278.3390695

Jain, S., & Wallace, B. C. (2019). Attention is not Explanation. North American Chapter of the Association for Computational Linguistics.

Ji, Z., Xiong, K., Li, X., & Li, X. (2020a). Video Summarization With Attention-Based Encoder–Decoder Networks. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*(6), 1709–1717. doi:10.1109/TCSVT.2019.2904996

Ji, Z., Jiao, F., Li, X., & Shao, L. (2020b). Deep attentive and semantic preserving video summarization. *Neurocomputing*, *405*, 200–207. doi:10.1016/j.neucom.2020.04.132

Jung, Y., Cho, D., Kim, D., Woo, S., & Kweon, I. S. (2019). Discriminative feature learning for unsupervised video summarization. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*. AAAI Press. 10.1609/aaai.v33i01.33018537

Jung, Y., Cho, D., Woo, S., & Kweon, I. S. (2020). Global-and-Local Relative Position Embedding for Unsupervised Video Summarization. In Lecture Notes in Computer Science (pp. 167–183). Springer Science+Business Media. doi:10.1007/978-3-030-58595-2_11

Kanafani, H., Ghauri, J., Hakimov, S., & Ewerth, R. (2021). Unsupervised Video Summarization via Multi-source Features. In *Proceedings of the 2021 International Conference on Multimedia Retrieval (ICMR '21)*. Association for Computing Machinery. 10.1145/3460426.3463597

Kobayashi, G., Kuribayashi, T., Yokoi, S., & Inui, K. (2020). Attention is Not Only a Weight: Analyzing Transformers with Vector Norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 7057–7075). Association for Computational Linguistics. 10.18653/v1/2020.emnlp-main.574

Lal, S., Duggal, S., & Sreedevi, I. (2019). Online Video Summarization: Predicting Future to Better Summarize Present. *Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 471-480. 10.1109/WACV.2019.00056

Lebron Casas, L., & Koblents, E. (2019). Video Summarization with LSTM and Deep Attention Models. In Lecture Notes in Computer Science (pp. 67–79). Springer Science+Business Media. doi:10.1007/978-3-030-05716-9_6

Lei, J., Luan, Q., Xinhui, S., Liu, X., Tao, D., & Song, M. (2019). Action Parsing-Driven Video Summarization Based on Reinforcement Learning. *IEEE Transactions on Circuits and Systems for Video Technology*, *29*(7), 2126–2137. doi:10.1109/TCSVT.2018.2860797

Li, L., Wang, B., Verma, M., Nakashima, Y., Kawasaki, R., & Nagahara, H. (2021). SCOUTER: Slot Attention-based Classifier for Explainable Image Recognition. *Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 1026-1035, 10.1109/ICCV48922.2021.00108

Li, P., Ye, Q., Zhang, L., Yuan, L., Xu, X., & Shao, L. (2021). Exploring global diverse attention via pairwise temporal relation for video summarization. *Pattern Recognition*, *111*, 107677. doi:10.1016/j.patcog.2020.107677

Li, Z., Wang, W., Li, Z., Huang, Y., & Sato, Y. (2021). Towards Visually Explaining Video Understanding Networks with Perturbation. *Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1119-1128. 10.1109/WACV48630.2021.00116

Liang, G., Lv, Y., Li, S., Zhang, S., & Zhang, Y. (2022). Video summarization with a convolutional attentive adversarial network. *Pattern Recognition*, *131*, 108840. doi:10.1016/j.patcog.2022.108840

Lin, J., Zhong, S.-H., & Fares, A. (2022). Deep hierarchical LSTM networks with attention for video summarization. *Computers & Electrical Engineering*, *97*, 107618. doi:10.1016/j.compeleceng.2021.107618

Liu, T., Meng, Q., Wong, M. C., Vlontzos, A., Rueckert, D., & Kainz, B. (2022). Video Summarization Through Reinforcement Learning With a 3D Spatio-Temporal U-Net. *IEEE Transactions on Image Processing*, *31*, 1573–1586. doi:10.1109/TIP.2022.3143699 PMID:35073266

Liu, Y., Li, H., Guo, Y., Kong, C., Li, J., & Wang, S. (2022). Rethinking Attention-Model Explainability through Faithfulness Violation Test. In Proceedings of the 39th International Conference on Machine Learning. *Proceedings of Machine Learning Research.*, *162*, 13807–13824.

Liu, Y.-T., Li, Y.-J., Yang, F.-E., Chen, S.-F., & Wang, Y.-C. F. (2019). Learning Hierarchical Self-Attention for Video Summarization. *Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP)*, 3377-3381, 10.1109/ICIP.2019.8803639

Mahasseni, B., Lam, M., & Todorovic, S. (2017). Unsupervised Video Summarization with Adversarial LSTM Networks. *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2982-2991. 10.1109/CVPR.2017.318

Mänttäri, J., Broomé, S., Folkesson, J., & Kjellström, H. (2020). Interpreting Video Features: A Comparison of 3D Convolutional Networks and Convolutional LSTM Networks. In Lecture Notes in Computer Science (pp. 411–426). Springer Science+Business Media. doi:10.1007/978-3-030-69541-5_25

Narasimhan, M. G., Rohrbach, A., & Darrell, T. (2021). CLIP-It! Language-Guided Video Summarization. *Proceedings of the 2021 Conference on Neural Information Processing Systems*.

Ntrougkas, M., Gkalelis, N., & Mezaris, V. (2022). TAME: Attention Mechanism Based Feature Fusion for Generating Explanation Maps of Convolutional Neural Networks. *Proceedings of the 2022 IEEE International Symposium on Multimedia (ISM)*, 58-65. 10.1109/ISM55400.2022.00014

Otani, M., Nakashima, Y., Rahtu, E., & Heikkila, J. (2019). Rethinking the Evaluation of Video Summaries. *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7588-7596. 10.1109/CVPR.2019.00778

Otani, M., Nakashima, Y., Rahtu, E., Heikkilä, J., & Yokoya, N. (2016). Video Summarization Using Deep Semantic Features. In Lecture Notes in Computer Science (pp. 361–377). Springer Science+Business Media. doi:10.1007/978-3-319-54193-8_23

Pang, Z., Nakashima, Y., Otani, M., & Nagahara, H. (2023). Contrastive Losses Are Natural Criteria for Unsupervised Video Summarization. *Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2009-2018. 10.1109/WACV56688.2023.00205

Phaphuangwittayakul, A., Guo, Y., Ying, F., Xu, W., & Zheng, Z. (2021). Self-Attention Recurrent Summarization Network with Reinforcement Learning for Video Summarization Task. *Proceedings of the 2021 IEEE International Conference on Multimedia and Expo (ICME)*, 1-6. 10.1109/ICME51207.2021.9428142

Potapov, D., Douze, M., Harchaoui, Z., & Schmid, C. (2014). Category-Specific Video Summarization. In Lecture Notes in Computer Science (pp. 540–555). Springer Science+Business Media. doi:10.1007/978-3-319-10599-4_35

Psallidas, T., Vasilakakis, M. D., Spyrou, E., & Iakovidis, D. K. (2022). Multimodal Video Summarization based on Fuzzy Similarity Features. *Proceedings of the 2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP),* 1-5. 10.1109/IVMSP54334.2022.9816266

Puthige, I., Hussain, T., Gupta, S. K., & Agarwal, M. (2023). Attention Over Attention: An Enhanced Supervised Video Summarization Approach. *Procedia Computer Science*, *218*, 2359–2368. doi:10.1016/j.procs.2023.01.211

Ren, Z., Qian, K., Dong, F., Dai, Z., Nejdl, W., Yamamoto, Y., & Schuller, B. W. (2022). Deep attention-based neural networks for explainable heart sound classification. *Machine Learning with Applications*, *9*, 100322. doi:10.1016/j.mlwa.2022.100322

Rhevanth, M., Ahmed, R., Shah, V., & Mohan, B. R. (2022). *Deep Learning Framework Based on Audio–Visual Features for Video Summarization. In Advanced Machine Intelligence and Signal Processing. Lecture Notes in Electrical Engineering, 858*. Springer. doi:10.1007/978-981-19-0840-8_17

Roy, C., Shanbhag, M., Nourani, M., Rahman, T., Kabir, S., Gogate, V., Ruozzi, N., & Ragan, E. D. (2019). Explainable Activity Recognition in Videos. *Proceedings of the 2019 ACM Intelligent User Interfaces (IUI) Workshops.*

Serrano, S., & Smith, N. A. (2019). Is Attention Interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 2931–2951). Association for Computational Linguistics. 10.18653/v1/P19-1282

Shoer, I., Kopru, B., & Erzin, E. (2022). *Role of Audio in Audio-Visual Video Summarization*. arXiv, arXiv:2212.01040

Song, Y., Vallmitjana, J., Stent, A., & Jaimes, A. (2015). TVSum: Summarizing web videos using titles. *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5179-5187. 10.1109/CVPR.2015.7299154

Sosnovik, I., Moskalev, A., Kaandorp, C., & Smeulders, A. (2023). *Learning to Summarize Videos by Contrasting Clips*. arXiv, arXiv:2301.05213

Sreeja, M.U., & Kovoor, B.C. (2022). A multi-stage deep adversarial network for video summarization with knowledge distillation. *Journal of Ambient Intelligent Humanized Computing*. doi:10.1007/s12652-021-03641-8

Stergiou, A., Kapidis, G., Kalliatakis, G., Chrysoulas, C., Veltkamp, R., & Poppe, R. (2019). Saliency Tubes: Visual Explanations for Spatio-Temporal Convolutions. *Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP)*, 1830-1834. 10.1109/ICIP.2019.8803153

Su, M., Ma, R., Zhang, B., Li, K., & An, P. (2023). *Regression Augmented Global Attention Network for Query-Focused Video Summarization. In Communications in Computer and Information Science, 1766*. Springer. doi:10.1007/978-981-99-0856-1_24

Szegedy, C., Wei, L., Yangqing, J., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1-9. 10.1109/CVPR.2015.7298594

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)*. Curran Associates Inc.

Wang, J., Wang, W., Wang, Z., Wang, L., Feng, D., & Tan, T. (2019). Stacked Memory Network for Video Summarization. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19)*. Association for Computing Machinery. 10.1145/3343031.3350992

Wei, H., Ni, B., Yan, Y., Yu, H., Yang, X., & Yao, C. (2018). Video Summarization via Semantic Attended Networks. *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, *32*(1). 10.1609/aaai.v32i1.11297

Wiegreffe, S., & Pinter, Y. (2019). Attention is not not Explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 11–20). Association for Computational Linguistics. 10.18653/v1/D19-1002

Wu, G., Lin, J., & Silva, C. T. (2021). ERA: Entity Relationship Aware Video Summarization with Wasserstein GAN. *Proceedings of the 2021 British Machine Vision Conference (BMVC).*

Xiao, S., Zhao, Z., Zhang, Z., Guan, Z., & Cai, D. (2020a). Query-Biased Self-Attentive Network for Query-Focused Video Summarization. *IEEE Transactions on Image Processing*, *29*, 5889–5899. doi:10.1109/TIP.2020.2985868 PMID:32286987

Xiao, S., Zhao, Z., Zhang, Z., Yan, X., & Yang, M. (2020b). Convolutional Hierarchical Attention Network for Query-Focused Video Summarization. *Proceedings of the 34th AAAI Conference on Artificial Intelligence.* 10.1609/aaai.v34i07.6929

Xie, J., Chen, X., Lu, S.-P., & Yang, Y. (2022). A Knowledge Augmented and Multimodal-Based Framework for Video Summarization. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22)*. Association for Computing Machinery. 10.1145/3503161.3548089

Yaliniz, G., & Ikizler-Cinbis, N. (2021). Using independently recurrent networks for reinforcement learning based unsupervised video summarization. *Multimedia Tools and Applications*, *80*(12), 17827–17847. doi:10.100711042-020-10293-x

Yao, M., Bai, Y., Du, W., Zhang, X., Quan, H., Cai, F., & Kang, H. (2022). Multi-Level Spatiotemporal Network for Video Summarization. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22)*. Association for Computing Machinery. 10.1145/3503161.3548105

Yu, H., Huang, Y., Pi, L., Zhang, C., Li, X., & Wang, L. (2021). End-to-end video text detection with online tracking. *Pattern Recognition*, *113*, 107791. doi:10.1016/j.patcog.2020.107791

Yuan, L., Tay, F. E. H., Li, P., & Feng, J. (2020). Unsupervised Video Summarization With Cycle-Consistent Adversarial LSTM Networks. *IEEE Transactions on Multimedia*, *22*(10), 2711–2722. doi:10.1109/TMM.2019.2959451

Yuan, Y., Li, H., & Wang, Q. (2019a). Spatiotemporal Modeling for Video Summarization Using Convolutional Recurrent Neural Network. *IEEE Access: Practical Innovations, Open Solutions*, *7*, 64676–64685. doi:10.1109/ACCESS.2019.2916989

Yuan, Y., Mei, T., Cui, P., & Zhu, W. (2019b). Video Summarization by Learning Deep Side Semantic Embedding. *IEEE Transactions on Circuits and Systems for Video Technology*, *29*(1), 226–237. doi:10.1109/TCSVT.2017.2771247

Yuan, Y., & Zhang, J. (2022). Unsupervised Video Summarization via Deep Reinforcement Learning With Shot-Level Semantics. *IEEE Transactions on Circuits and Systems for Video Technology*, *33*(1), 445–456. doi:10.1109/TCSVT.2022.3197819

Zang, S., Yu, H., Song, Y., & Zeng, R. (2022). Unsupervised video summarization using deep Non-Local video summarization networks. *Neurocomputing*, *519*, 26–35. doi:10.1016/j.neucom.2022.11.028

Zhang, K., Chao, W., Sha, F., & Grauman, K. (2016). Video Summarization with Long Short-Term Memory. In Lecture Notes in Computer Science (pp. 766–782). Springer Science+Business Media. doi:10.1007/978-3-319-46478-7_47

Zhang, K., & Li, L. (2022). Explainable multimodal trajectory prediction using attention models. *Transportation Research Part C, Emerging Technologies*, *143*, 103829. doi:10.1016/j.trc.2022.103829

Zhang, Y., Kampffmeyer, M., Zhao, X., & Tan, M. (2019). Deep Reinforcement Learning for Query-Conditioned Video Summarization. *Applied Sciences (Basel, Switzerland)*, *9*(4), 750. doi:10.3390/app9040750

Zhang, Y., Kampffmeyer, M. C., Liang, X., Tan, M., & Xing, E. P. (2018). Query-Conditioned Three-Player Adversarial Network for Video Summarization. *Proceedings of the 2018 British Machine Vision Conference (BMVC).*

Zhao, B., Gong, M., & Li, X. (2021b). AudioVisual Video Summarization. *IEEE Transactions on Neural Networks and Learning Systems*, 1–8. doi:10.1109/tnnls.2021.3119969 PMID:34695009

Zhao, B., Li, X., & Lu, X. (2017). Hierarchical Recurrent Neural Network for Video Summarization. In *Proceedings of the 25th ACM international conference on Multimedia (MM '17)*. Association for Computing Machinery. 10.1145/3123266.3123328

Zhao, B., Li, X., & Lu, X. (2018). HSA-RNN: Hierarchical Structure-Adaptive RNN for Video Summarization. *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7405-7414. 10.1109/CVPR.2018.00773

Zhao, B., Li, X., & Lu, X. (2020). Property-Constrained Dual Learning for Video Summarization. *IEEE Transactions on Neural Networks and Learning Systems*, *31*(10), 3989–4000. doi:10.1109/TNNLS.2019.2951680 PMID:31825876

Zhao, B., Li, X., & Lu, X. (2021a). TTH-RNN: Tensor-Train Hierarchical Recurrent Neural Network for Video Summarization. *IEEE Transactions on Industrial Electronics*, *68*(4), 3629–3637. doi:10.1109/TIE.2020.2979573

Zhou, K., Qiao, Y., & Xiang, T. (2018a). Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. AAAI Press. 10.1609/aaai.v32i1.12255

Zhou, K., Xiang, T., & Cavallaro, A. (2018b). Video Summarisation by Classification with Deep Reinforcement Learning. *Proceedings of the 2018 British Machine Vision Conference (BMVC).*

Zhu, Y., Zhao, W., Hua, R., & Wu, X. (2023). Topic-aware Video Summarization using Multimodal Transformer. *Pattern Recognition*, *140*, 109578. doi:10.1016/j.patcog.2023.109578

Zhuo, T., Cheng, Z., Zhang, P., Wong, Y., & Kankanhalli, M. (2019). Explainable Video Action Reasoning via Prior Knowledge and State Transitions. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19)*. Association for Computing Machinery. 10.1145/3343031.3351040

KEY TERMS AND DEFINITIONS

Explainable Artificial Intelligence (XAI): Explainable Artificial Intelligence refers to methods and algorithms that allow humans to understand the reasoning behind the decisions or predictions made by machine/deep learning network architectures.

Explanation Mask: In the context of XAI, an explanation mask is a human-interpretable visualization that localizes the most important aspects of each input data for the decisions or predictions made by machine/deep learning network architectures.

Generative Adversarial Networks: Generative Adversarial Networks, a class of machine learning frameworks, are made up of a pair of two neural networks that are trained progressively in an adversarial manner; as the training proceeds, the generator tries to create realistic data samples that fool the discriminator, while the discriminator aims to make a good guess on how "realistic" the given input seems.

Recurrent Neural Networks (RNNs): Recurrent Neural Networks is a class of artificial neural networks where connections between nodes create a cycle, allowing output from some nodes to affect subsequent input to the same nodes; through this iterative behaviour, they can process variable length sequences of inputs and capture the states or data of previous inputs to generate the next output of a sequence.

Reinforcement Learning: Reinforcement learning is a machine learning paradigm based on rewarding desired behaviours and/or punishing undesired ones; based on it, a reinforcement learning agent progressively learns to perceive and interpret its environment, take actions and develop knowledge about a task through trial and error.

Self-Attention: Self-attention is an attention mechanism that captures relations between different elements of the input data sequence, aiming to compute a representation of this data sequence and dynamically adjust the influence of each element on the output.

Supervised Learning: Supervised learning is a machine learning paradigm that is used to learn a function that maps feature vectors to labels/scores, based on examples of input-output pairs (i.e., human-based ground-truth annotations).

Transformer Network: A Transformer Network is a neural network that is designed to process sequential data all at once, thus allowing for more parallelization than RNNs and therefore reducing training times; it learns context for any position in the input sequence by modelling data relationships using an attention mechanism.

Unsupervised Learning: Unsupervised learning is a machine learning paradigm that is used to learn concise representations of the input data, which can be used for data exploration, data compression or new data generation.

Video Summarization: Video summarization is a problem in the domain of video analysis and understanding, that aims to generate a short synopsis by selecting the most informative and important parts of the video.