TECHNICAL REPORT | 2023

IM SCIENCE

Hierarchical time-aware summarization with an adaptive transformer for video captioning

(Accepted – journal – IJSC)

05

TECHNICAL REPORT | 2023

IM SCIENCE

Hierarchical time-aware summarization with an adaptive transformer for video captioning

Citation details:
Leonardo Cardoso Vilela, Silvio Jamil F. Guimarães, Zenilton Kleber Gonçalves do Patrocínio Jr. (2023). Hierarchical time-aware summarization with an adaptive transformer for video captioning. Technical Report ImScience/PUC Minas #05/2023.

05

Technical report – ImScience 05/2023

Hierarchical time-aware summarization with an adaptive transformer for video captioning

Leonardo Cardoso Vilela · Silvio Jamil F. Guimarães · Zenilton K. Gonçalves do Patrocínio Jr.

June 8, 2023

Abstract A coherent description is an ultimate goal regarding video captioning via a couple of sentences because it might also affect the consistency and intelligibility of the generated results. In this context, a paragraph describing a video is affected by the activities used to both produce its specific narrative and provide some clues that can also assist in decreasing textual repetition. This work proposes a model, named **H**ierarchical time-aware **S**ummarization with an **A**daptive **T**ransformer – HSAT, that uses a strategy to enhance the frame selection reducing the amount of information that needed to be processed along with attention mechanisms to enhance a memory-augmented transformer. This new approach increases the coherence among the generated sentences, assessing data importance (about the video segments) contained in the self-attention results and uses that to improve readability using only a small fraction of time spent by the other methods. The test results show the potential of this new approach as it provides higher coherence among the various video segments, decreasing the repetition in the generated sentences and improving the description diversity in the ActivityNet Captions dataset.

Keywords Video captioning · Memory-augmented transformer · Attention mechanisms · Hierarchical graph-based video summarization.

1 Introduction

Video captioning is the task of concisely describing a video through text [5,6,22]. One of the biggest problems in video captioning task is the content description of the video based on a ground truth (GT) created by more than one person [21] since GT tends to cover events of a video from different perspectives and emphasizing distinct moments. Video captioning results are strongly correlated with two interrelated sub-tasks: (i) temporal event detection and (ii) description generation.

Detection of video events can be based on three strategies: (i) random selection; (ii) time-sliding window; and (iii) scene (or shot) detection. In any way, those methods seek somehow to summarize video content to better guide the caption generation step with the most informative parts of the video. Broadly, there exist two types of video summaries: a static video summary composed of keyframes and a dynamic video summary (composed of key-shots). Thus, there is a great difficulty in the selection of video frames (or video shots) to cover all the video narrative without the loss of any content [1,5,6,22]. Nevertheless, video summarization represents an ill-posed problem that has been addressed by many authors over the last decades and a number of surveys on video summarization have already appeared in the literature [2,23,25]. In any way, video summarization may provide better input for the video caption generation step since it can generate an informative synopsis of a video preferably with maximum representativeness, minimum repetition, and maximum diversity [23].

Regarding description generation, transformers [32] have recently shown to be very useful for many sequence-related tasks, such as machine translation [34], information retrieval [38], text classification [15], document summarization [41], image classification [7], image captioning [19,26], video captioning [5,6,22], and others [31]. In video captioning, the authors of [22] proposed a memory-augmented transformer to cope with text repetition, while in [5,6] a re-weighting of the importance of data present in the memory module and the self-attention module, respectively, was explored to directly influence the amount of information the transformer uses to learn. The main idea of those works is somehow to

Silvio Jamil F. Guimarães
ImScience/PUC-Minas – Belo Horizonte 31980-110, Brazil
E-mail: sjamil@pucminas.br

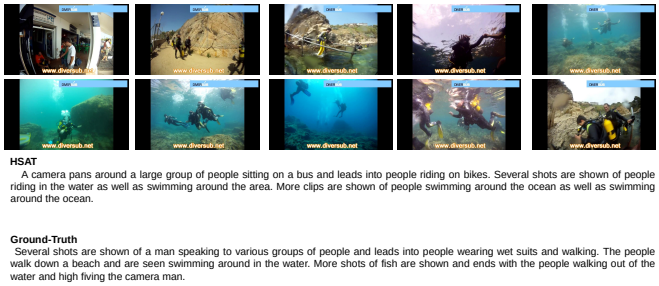


Fig. 1 An example of the result obtained by the proposed method HSAT.

assess the data importance (about the video segments) and explore that to improve readability by reducing repetition. But, in many cases, part of the generated sentences can be repeated, diminishing the quality of the final result. That is more evident in an event-based description because, if there are different events (video segments) with high correlation, there is a high probability that the same (or almost the same) sentence fragment appears more than once in the final result. Currently, some methods try to cope with that. However, this is not a trivial task since it consists of evaluating the relationship between a text excerpt and all others being produced so that the described event is not repeated in the final result.

Figure 1 shows an example of the results obtained by the proposed method **H**ierarchical time-aware **S**ummarization with an **A**daptive **T**ransformer – HSAT, along with the expected GT description. It is easy to observe the high correlation among frames, which also appears in the GT result. Unlike other methods that tend to present the same sentence several times, the use of attention mechanisms made in our proposal enhances coherence among generated sentences and the numerous events within a video (similar to Adaptive Transformer [6]). Thus, the final description can adapt even with the presence of similar events and be more concise, meaningful, and intelligible. In addition, Figure 1 demonstrates the great difficulty in describing the characteristics of the database, since the description of daily activities in a real scenario is often done in situations with few variations of information, and the modifications are usually not enough to differentiate the agent due to video characteristics such as perspective and distance.

An important issue in video captioning is the reduction of similarity between frames (very close in time) used in caption generation which could lead to a decrease in repetition rate in the final result. Together with that, there is also the need of preserving the temporal coherence between video content and the generated text which may increase the number of frames that are processed (and described) and diminish the quality of the final result. Attention mechanisms can be utilized to generate distributions over video segments of higher interest at the expense of others and then help in video captioning. Traditional attention methods tend to consider all viable areas to make bigger attention. However, highlighting regions might also incur some failures and omissions [1, 12]. Therefore, some methods reduce frames into smaller distributions to verify local interest and reweight the importance of features. These strategies avoid the awareness of data with little significance for the final result. Doing that, unobserved semantic aspects can be explored [13].

Figure 2 illustrates the proposed method for video captioning. Our method adopts three major components: (i) a video summarizer; (ii) a feature extractor; and (iii) a memory-augmented transformer with adaptive attention. The summarization step adopts a hierarchical approach to produce a static video summary that improves the set of frames used by the feature extraction step. This approach is based on [29] and copes better with (dis)similarities among video frames producing a more valuable frame selection for the description process. Different from [29], it uses a watershed-based hierarchical method applied to a frame similarity graph constructed with CNN-based frame descriptors and cosine similarity. Besides, frames are only considered to be similar if the difference between their time index is less than a fixed threshold. This can be used to restrict the relationship between video frames far away and implies that our hierarchical graph-based summarization approach is time-aware (which also differs from many works in the literature, including [29]). After the summarization step, aligned features for appearance and optical flow are extracted only for the selected keyframes and used for both to induce a video captioning model during training time and to feed the trained model during inference time to generate a description for a new unseen video. The model adopted for caption generation is based on a previous work, named Adaptive Transformer [6], in which a memory-augmented transformer with a shared architecture explores the attention mechanism to re-weight the importance of data generated by the self-attention module. The motivation for that is to explore the results generated by the self-attention module to improve readability through diminishing repetition.

Although this new proposal is based on a previous work [6], it presents the following specific contributions in comparison to it: (i) a graph-based approach to model the similarity of frames including also time restrictions to avoid the relationship between frames far away; (ii) a hierarchical summarization strategy applied to the frame similarity graph to obtain a video summary; and (iii) a significative reduction of computational time spent by description generation step without any impact in the quality of the final results. Time reduction achieves almost 70% during training when compared with [6] but still keeping similar results to the state-of-the-art approaches regarding quality.

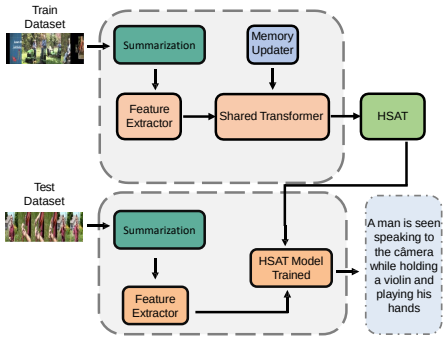


Fig. 2 The outline of the proposed method with (i) a hierarchical graph-based summarizer; (ii) a feature extractor; and (iii) a shared memory-augmented transformer with adaptive attention.

This work is organized as follows. Section 2 discusses related works and provides the theoretical background. Section 3 presents the proposed method, followed by the experimental results and their analysis in Section 4. Finally, Section 5 draws some conclusions and future work proposals.

2 Related Works and Theoretical Background

2.1 Video captioning

The video captioning task consists of producing sentences that are diverse and able to describe numerous events in a video into a dense paragraph description that relates all of them [22]. One of the major challenges for video captioning is to produce a coherent narrative for a segmented video database (or for lengthy videos that have to be segmented before described) since video segmentation may motivate redundancy in generated descriptions [21, 28].

The description of events can happen in two ways, one being the separate description of every event and, in this case, they may (or may not) be related, some other way is to group the related events and produce at least one coherent sentence about them in the paragraph. The problem in grouping events into paragraphs is linked to coherence because the paragraph must present its elements avoiding repetition. A dense description can narrate separate events (with or without relation), but in this case, since they are separate descriptions, redundancy is usually not an issue. However, in the generation of a paragraph for a video, each sentence is possibly related to the others [1,28]. Therefore, if some events present similarities, the generated sentences can repeat. Thus, the method for video captioning should prevent the final description from having the same (or almost the same) sentence many times. The real intention behind this process is to capture other aspects that, by chance, have not been described yet. Thus, describing a video through a paragraph consists of ensuring cohesion while maintaining similarity to the expected GT results [1, 19, 21, 22].

Some feature extraction methods use one strategy based on sequential data sampling. In video captioning, this process consists of choosing frames at regular intervals within the video. Thus some relevant frames for the video content (and also for its description) might not be considered. Generally, a sequential selection policy extracts a limited number of frames (for instance, 100 frames) just from the begging of the video. Therefore, the remain of video frames (along with all the events they represent) are completely ignored [21, 28, 42].

2.2 Attention mechanisms and transformers

The attention mechanisms were first applied to the machine translation task [3] and, after that, they were applied to other tasks such as object detection, image classification, and image content description [39]. The use of attention mechanisms seeks to highlight the importance of more prominent content, which tends to be neglected in conventional methods [32].

The impact on the application of attention mechanisms to description tasks was first explored by [37] in which the authors evaluated the impact of attention in image captioning. They explored *soft* and *hard* attention, and the results achieved showed improvement for both, but with high training costs for *hard* attention. Thus, other variants of attention were proposed and studied in the literature. But in [32], the use of attention reached another level, and it began to be considered a method (called transformer) capable of producing results by itself without the need for other techniques.

A challenge for attention mechanisms lies in the long-range dependencies. This problem is related to the network's capacity to learn from all the previous states. But the adoption of a self-attention mechanism creates the possibility to circumvent those difficulties, at the same time allows more efficient use of available resources through extensive use of parallelism [32].

In order to re-weight the importance of certain data, attention mechanisms can be applied to the network backbone to increase the importance of specific information about others. Techniques, as presented by [18], tend to increase the importance of some features, compared to others, and could be explored to increase the relationship between data that previously could not be easily related [5, 18].

Experiments on machine translation tasks showed transformers as superior models in quality while being more parallelizable and requiring significantly less time to train than others. A transformer achieves good results because it can capture the relationship between tokens and the generated

vocabulary. After that, transformer models have been successfully applied to several distinct tasks, such as machine translation, information retrieval, text classification, document summarization, image classification, image captioning, and video captioning [7, 15, 19, 22, 26, 30, 34, 41].

Some transformer models exploit the traditional encoder-decoder architecture, while others use a shared one. In a traditional encoder-decoder architecture, the encoder is separate from the decoder, and the encoded information about the video must be passed and used in some of the decoder steps. So, the decoder uses the data produced by the encoding network to generate each word/sentence [19, 26, 43]. For shared architectures, video data and generated text tokens are passed as input to the same module, which (after operating on them) is responsible for both encoding and decoding. The latter architecture is generally used due to the reduction of operations, resulting in time and processing savings [8, 22, 30].

Considering the results found by [5,6], it may be interesting to explore the use of different attention mechanisms in other parts within the transformer. In this work, we chose to evaluate the impacts of adaptive attention within the main backbone of the transformer (similar to [6]), instead of its adoption within the memory module as in [5]. The idea of reinforcing the characteristics learned during the current training period appears as a way of helping the transformer to learn new characteristics, increasing the quality of the results of the multi-head attention mechanism. Thus, the attention regions parallelized by the transformer (through multi-head attention) can be improved through adaptive attention in order to refine the attention distribution computed by each head.

2.3 Memory-augmented transformer

A memory-augmented transformer contains a memory block (information that can be accessed later) and a memory updater module to allow adjustments on that stored information during its execution. The memory updater module seeks to assist in assessing video segments' importance in generating new sentences. Its adoption allows the transformer to recurrently evaluate longer sentences. Conceptually, that proposal uses similar strategies to those defined by LSTM [17] and GRU [9] modules. The difference between them and the memory updater module is the high capacity of the latter to model complex data provided by the transformer employing a multi-head attention stage. Consequently, it enables the memory to capture/model different concepts and, therefore, to better understand and deal with similarities among video events [5, 6, 22].

The use of a memory module has been gaining projection in the literature. Unlike conventional methods, the use of memory mechanisms helps to reduce the redundancy by recurrently evaluating information from previous states [12, 22] . In this way, a memory module becomes a mechanism for assessing the importance of sentences (and video events). So, the information that is passed between states is used to assess their relevance degree. Memory data is encoded during video processing, and it should be updated to keep track of relevant information.

Considering that $M_t^l \in \mathbb{R}^{T_m \times d}$ represents memory state at layer l in step t in which T_m denotes memory length, and $\tilde{H}_t^l \in \mathbb{R}^{T_c \times d}$ is the intermediate hidden state vector, the memory update process (proposed in [22]) can be summarized by Equations (1)–(5):

$$S_t^l = MultiHeadAtt(M_{t-1}^l, \tilde{H}_t^l, \tilde{H}_t^l), \tag{1}$$

$$C_t^l = tanh(W_{mc}^l M_{t-1}^l + W_{sc}^l S_t^l + b_c^l), \tag{2}$$

$$E_t^l = W_{mz}^l M_{t-1}^l + W_{sz}^l S_t^l + b_z^l, \tag{3}$$

$$Z_t^l = sigmoid(W_r^l(ReLU(W_p^l(\rho(E_t^l))))) \odot E_t^l, \tag{4}$$

$$M_t^l = (1 - Z_t^l) \odot C_t^l + Z_t^l \odot M_{t-1}^l, \tag{5}$$

in which $\odot$ denotes Hadamard product, W_{mc}^l, W_{sc}^l, W_{mz}^l, and W_{sz}^l are trainable weights, b_c^l and b_z^l are trainable bias. $C_t^l \in \mathbb{R}^{T_m \times d}$ is the internal cell state, while $Z_t^l \in \mathbb{R}^{T_m \times d}$ is the update gate that controls which information to retain from the previous memory state. Equation 1 presents S^l as the output of the multi-head attention mechanism and was used in [22] as the first attention on a memory module. The E^l was used as a mechanism to assess the importance of the event, Z_t^l as the information regulator and C_t^l as the new information sample. In this way, the update of M_t^l will occur through a linear combination of new information represented by C_t^l and the information already present in the memory, i.e., M_{t-1}^l.

According to [5], memory is made up of information that the model must remember, however, this may not be the best way of learning, being necessary, in some cases, to learn new information and forget, momentarily, information previously learned. In these cases, the information contained in memory is considered to have no value, that is, forgetting it would not harm the result. On the other hand, there are times when only the learned information is enough to generate new patterns and, for these cases, any new information will not be used, as there is no status modification that could be incorporated into the result. However, there are cases in which it is necessary to use part of the information that exists in memory and a part of new information so that new sentences are produced. The regulator Z_t tries to find the ideal proportion of information. For this, it considers the importance of what exists in memory and the new information learned.

Algorithm 1 Hierarchical video summarization

Input: A video $\mathcal{V}$, threshold value δ_t, summary size k
Output: A list of keyframes $\mathcal{K}$

1: Create a graph $G = (V, E)$ with a vertice set $V = \emptyset$ and an edge set $E = \emptyset$
2: **for all** $frame\ f \in \mathcal{V}$ **do**
3: **if** $f \notin V$ **then**
4: $V := V \cup \{f\}$ // Insert f in V if f does not belong to it
5: **end if**
6: $d_f := GenerateDescriptor(f)$ // Obtain a descriptor for frame f
7: **for all** $frame\ f' \in \mathcal{V}$ such that $f' \neq f$ and $|t_f - t_{f'}| < \delta_t$ **do**
8: **if** $f' \notin V$ **then**
9: $V := V \cup \{f'\}$ // Insert f' in V if f' does not belong to it
10: **end if**
11: $d_{f'} := GenerateDescriptor(f')$ // Obtain a descriptor for frame f'
12: $G.AddEdge(f, f')$ // Insert edge (f, f') using frame similarity as weight
13: **end for**
14: **end for**
15: $T := G.Obtain\_MST\_from\_Graph()$
16: $H := T.Generate\_Hierarchy\_from\_MST()$
17: $\mathcal{K} := H.Find\_Keyframes(k)$ // Remove $k-1$ edges from H to generate k sets
// and select the central vertice of each set as keyframe
18: Return $\mathcal{K}$;

3 Proposed Method

The problem of dense video captioning is related to the amount of similar information laid out sequentially with little or no variation. The relationship between the distribution of frames has a direct impact on the sentences generated and implies an increase or not in repetition.

To deal with that issue, the proposed method is divided into three steps (see Fig. 2). The first is the selection of the best set of frames for each video through a hierarchical summarization approach which splits a video into subsets of similar frames and selects the central frame (using similarity among frames) as the keyframe. The second step extract features for the appearance and optical flow of previously selected keyframes. It is quite similar to the analogous step in [6], but it spends lesser computational time since the video summarization step is able to choose a set of keyframes that is more informative but smaller. Finally, the third step is description generation using an Adaptive Transformer (similar to [6]) but which was trained to work with smaller sets of keyframes (but describing the most important video contents).

3.1 Hierarchical time-aware graph-based summarization

Unlike the traditional approach that uses a sequential selection policy for frame selection, our proposed method chooses frames based on their similarity. It adopts a hierarchical graph-based summarization method to obtain the most valuable frames (as keyframes).

A frame similarity graph is constructed and used by the video summarization approach. In this graph, each vertice represents a video frame. An edge between two vertices only exists if the difference between their time indexes is lower than a threshold δ_t. The edge weight represents the similarity value between the two frames associated with edge extremes.

Algorithm 1 presents the hierarchical video summarization approach used for frame selection in our proposal with the following steps: (i) construction of a frame similarity graph for a video (lines 1–14); (ii) calculation of a minimum spanning tree (MST) for the graph (line 15); (iii) production of a hierarchy through a re-weighting process based on that MST (line 16); and (iv) generation of sets of similar frames through cuts of the generated hierarchy and the selection of the central vertice of each set as a keyframe (line 17–18). Figure 3 shows an example of each step of Algorithm 1.

During graph construction, edges between frames are only created if the difference between their time indexes is less than a fixed threshold δ_t. Equation 6 represents that constraint and is ensured at line 7 of Algorithm 1.

$$|t_f - t_{f'}| < \delta_t \tag{6}$$

in which t_f and $t_{f'}$ represent the time indexes from frames f and f', respectively. This can be used to restrict the relationship between video frames far away and implies that our hierarchical graph-based summarization approach is time-aware (which also differs from many video summarization works in the literature, including [29] on which our approach is based).

According to [11], hierarchies can be represented by minimum spanning trees, and any connected hierarchy for a graph can be handled by means of a weighted minimum spanning tree of that graph. Moreover, the authors in [14] showed that a hierarchical segmentation of a graph consists of transforming an initial hierarchy into another one by rebuilding the hierarchical structure according to a dissimilarity measure between regions. That could be done by carefully re-weighting all edges belonging to an MST of the original graph. In [29], the authors provide a unified framework for video summarization that uses a minimum spanning tree of frames and a weight map based on hierarchical observation scales computed over that tree. The weight map is generated from the frame similarity graph in which the clusters (or connected components of the graph) can easily

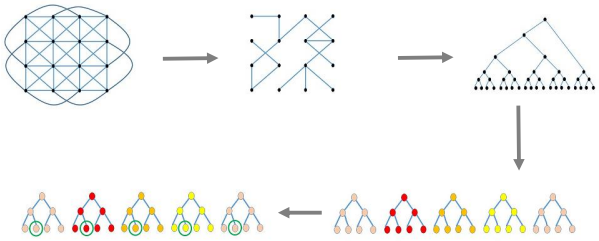


Fig. 3 Ilustration of the hierarchical video summarization approach used for frame selection in our proposal with the following steps: (i) construction of a frame similarity graph for a video; (ii) calculation of a MST for the graph; (iii) production of a hierarchy through a re-weighting process based on that MST; and (iv) generation of sets of similar frames through cuts of the generated hierarchy and the selection of the central vertice of each set as a keyframe (i.e., the removal of $k-1$ edges from the hierarchy with higher weight values generates k sets). Selected keyframes are indicated in the final result with green circles.

be inferred. Moreover, the use of this strategy allows the application of a similarity measure between clusters during graph partition, instead of considering only the similarity between isolated frames.

Inspired by [29], our approach also uses a minimum spanning tree (MST) to generate and represent a hierarchy. Our proposed method used the Kruskal method to obtain the MST (line 15 of Algorithm 1). After that, a hierarchy is calculated by re-weighting all edges belonging to the MST. Unlike [29], our method constructs a watershed hierarchy by area following the definition of [10] (line 16 of Algorithm 1). Finally, the generation of sets of similar frames can be achieved through cuts of the generated hierarchy (i.e., the removal of $k-1$ edges from the hierarchy with higher weight values generates k sets). After that, the selection of the central vertice of each set as keyframe is a simple task (lines 17-18 of Algorithm 1).

At first, the use of an extra step in the video captioning pipeline such as the proposed hierarchical summarization approach may appear to increase the total computational time. But, the process of selecting a smaller set with more informative frames actually leads to a reduction in computational time (because a much smaller number of frames is processed by the description generator) without any loss of quality in the final result.

3.2 Adaptive transformer for video description generation

To deal with coherence issues, we explored an improvement to the self-attention module that exists within the main backbone of the transformer (just after the multi-head attention modules). So, we adopted additional attention blocks to emphasize the data generated by self-attention and cross-attention. This was first proposed in a previous work, named Adaptive Transformer [6], which is shown in

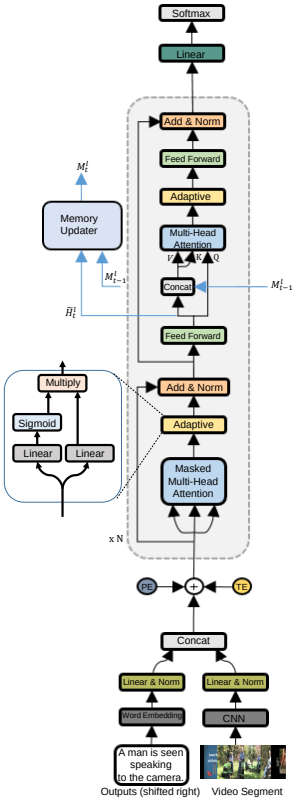


Fig. 4 An illustration of the Adaptive Transformer, proposed by [6], architecture highlighting one of its adaptive attention modules (and showing a detailed representation of it).

Figure 4, which focuses on reducing repetition. The memory module is still an important part of the shared transformer module, and this is responsible to capture the long-term dependency of the sentences.

The Adaptive Transformer uses an attention mechanism to rebalance the results of other attention mechanisms. This process seeks to highlight the characteristics that are considered important in smoothing out others [5]. Figure 4 shows the Adaptive Transformer with the presence of two modules (called Adaptive), both are identical. This work uses a version of adaptive attention as a way to apply reweighting to the results that are used for feeding memory updater, and adaptive attention is also applied to the results of cross attention.

The proposal of using attention mechanisms to compose a new model called transformer first appeared in [32] to reduce the computational cost without quality loss. Following [32], the main component of a transformer is the *scaled dot-product attention*. Given query matrix Q, key matrix K, and value matrix V, the attention output is given by Equation 7:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \quad (7)$$

Combining h parallel instances of scaled dot-product attention, we obtain the multi-head attention represented by Equation 8.

$$MultiHeadAtt(Q,K,V) = Concat(head_1;\ldots;head_h)W^O, \tag{8}$$

in which $head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$, and linear projections W_i^Q, W_i^K, W_i^V and, W^O are learned during training. The detailed outline of the adaptive attention module used in our proposal result on $A_t^l \in \mathbb{R}^{T_m \times d}$ is given by Equation 9:

$$A_t^l = sigmoid(W_{mhl}^l MH_t^l) \odot W_{mhr}^l MH_t^l + b_a^l, \tag{9}$$

in which W_{mhl}^l, and W_{mhr}^l are trainable weights, b_a^l are trainable bias, and MH_t^l stands for multi-head attention results.

The strategy applied in the adaptive attention mechanism is based on evaluating data that have already been previously processed by another attention mechanism. Thus, the quality of the distribution is enhanced, making the descriptions more discriminative and more related to the content they intend to represent. The use of second attention seeks to amplify the importance of the information. Thus, the application of adaptive attention acts as a refinement process on the results of previous attention, and this has an impact on the data learned on the memory updater module both on the regulator Z_t and on the newly learned information C_t. So, as new information is highlighted, it tends to replace the data stored in memory.

4 Experimental Results

In this section, we present the results achieved by each proposed method. All experiments, models were trained for 20 epochs with a learning rate of 10^{-4}. We selected the best model, considering the CIDEr-D score since it is considered the ideal assessment metric for content description tasks [1,33]. We also report the BLEU-4 score (B@4) which is a common metric for NLP tasks such as machine translation, and the Repetition-4 score (R@4) which measures redundancy. But, the R@4 score should not be used alone because it can prevent a correct assessment since a random text without any word repetition would present an R@4 score close to 0. Therefore, the best way to assess the model's improvement is through the joint evaluation of two or more metrics. In this way, the assessment of the relationship between the GT result and the final description was made using CIDEr-D score, while the R@4 score is evaluated afterward to point out the diversity achieved (through the reduction of repetition).

4.1 Baselines

The performance of the proposed models was compared with the following methods representing the state-of-the-art: VTransformer (Vanilla transformer) [43], Transformer-XL [12], Transformer-XLRG [12], AdvInf [28], GVD [42], GVDsup [42], MFT [35], HSE [40], MART [22], EMT [5], and Adaptive Transformer [6].

The works used for comparison are divided according to the technique used. Thus, MFT [35] and HSE [40] are based on LSTM to recurrently evaluate the generated sentences to produce new words. The works GVD [42], GVDsup [42], and AdvInf [28], in addition to LSTM, also use detection features in an attempt to increase the quality of the obtained scores. Finally, the remaining works, i.e., VTransformer [43], Transformer-XL [12], Transformer-XLRG [12], MART [22], EMT [5], and Adaptive Transformer [6], use transformers as a technique to reduce recurrence, increasing the quality of descriptions, optimizing performance, and augmenting consistency by combining all the information from elements that represent the context.

4.2 Dataset and implementation details

We apply the proposed method to the ActivityNet Captions (ANC) dataset [4, 21]. This dataset contains 10,009 videos for training and 4,917 videos for validation. Videos used during the training step have a single reference paragraph, while validation videos have two reference paragraphs. In [28], the authors used the same configuration proposed by [21], however, with different divisions, in which both validation and testing were conducted with the same set of videos. Here, we follow [42], in which authors proposed a new way of subdividing this dataset to optimize the use of videos and avoid overfitting. They kept the training videos and divided the validation videos into two subsets, namely: AE-VAL with 2,460 videos for validation and AE-TEST with 2,457 videos for testing. And, the ANC dataset comes with annotated segments (for each temporal event) with human-written natural language sentences that represent, on average, there are 3.65 segments per video.

The initial preprocessing follows with minor adjustments the same procedure described in [5, 6, 22]. The used vocabulary was created based on phrases that happen in at least 5 instances for the ANC dataset. The resulting vocabulary carries 3,544 words. The unit memory size was defined as 2 and, the memory dimension is set to $1,200$. Videos are represented by extracting 4 FPS. Those frames are used as input for the hierarchical summarization approach (described with CNN-based features extract with a pre-trained ResNet50 and using cosine similarity) with $\delta_t = 8$ and using the watershed hierarchy by area as described in [10, 24] to generate a summary of size $k = 10$.

Aligned features for appearance and optical flow had been extracted for each frame belonging to the video summary. Specifically, for appearance, 2048-D feature vectors from the "FLATTEN-673" layer in ResNet-200 [16] are used, while, for the optical flow, 1024-D feature vectors from the "GLOBAL POOL" layer of BNInception [20] are adopted. Both networks are pre-trained on the ANC dataset for action recognition and they are supplied by [36].

In order to use the proposed model, initially video and text are encoded and normalized separately. Encoded video and text embeddings are denoted as $H^0_{video} \in \mathbb{R}^{T_{video} \times d}$ and $H^0_{text} \in \mathbb{R}^{T_{text} \times d}$, respectively, in which T_{video} and T_{text} represent video and text lengths, while d is the embedding size. They are used after a concatenation and passed to the transformer as input $H^0 \in \mathbb{R}^{T_c \times d}$, i.e., $H^0 = Concat(H^0_{video}; H^0_{text})$, in which $T_c = T_{video} + T_{text}$, following the proposal of [8,30].

4.3 Evaluation metrics

The evaluation of sentences is a separate challenge, as there are several ways to write sentences, but with the same meaning, whether using synonyms or emphasizing distinct information. This process is intuitive for humans. But there is no specific approach for evaluating the video captioning task. So, what is usually done is the adaptation of machine translation metrics that are extended for this task [1, 27, 33].

In order to compute the improvement of the descriptions, we will evaluate the obtained results following the same approach used by the authors of [22, 28, 35]. They reported their results using metrics widely disseminated in the literature such as BLEU-4 (B@4) [27] and CIDEr-D [33], with the objective of evaluating the similarity between the descriptions generated by their models and the GT results. However, these metrics cannot penalize the repetition that may happen, so it is necessary the use of another metric for evaluating how diverse the description is. Thus, the Repetition-4 score (R@4) [22,28,35] was applied, and its objective is to emphasize the reduction of repetition of words in the description. Both R@4 and B@4 scores use 4-grams to increase word grouping.

4.4 Comparison to the state-of-the-art methods

Tables 1 and 2 present the results found for HSAT along with other *state-of-the-art* methods. Table 1 summarizes the performance of the proposed models, comparing them with *state-of-the-art* models in AE-VAL split of the ANC dataset. The results reported were evaluated mainly according to the CIDEr-D metric to choose the best model. The results shown in Table 1 represent models based on transformers, LSTM-only, and LSTM with detection features. The results achieved by the proposed model show an improvement compared to the others in relation to the CIDEr-D metric, except when compared to the Adaptive Transformer. The adoption of the adaptive attention module as a reinforcement strategy for previous attention results contributes to increasing the quality of the learned feature. With this, the results of its modified attention tend to better weight for data with greater importance. It turns out that just the addition of the *adaptive attention* block increases the CIDEr-D score, but this does not guarantee the reduction of the R@4 score.

Table 1 Performance of our model and other of state-of-art methods in AE-VAL split of ActivityNet Captions (**Det** indicates whether detection features are used; while **Rec** indicates whether sentence-level recurrence is used).

	Det	Rec	B@4 ↑	CIDEr-D ↑	R@4 ↓
LSTM based methods					
MFT [35]	χ	√	10.27	19.12	17.71
HSE [40]	χ	√	9.84	18.78	13.22
LSTM based methods with detection feature					
GVD [42]	√	χ	11.04	21.95	8.76
GVDsup [42]	√	χ	**11.30**	22.94	7.04
AdvInf [28]	√	√	10.04	20.97	5.76
Transformer based methods					
VTransformer [43]	χ	χ	9.75	22.16	7.79
Transformer-XL [12]	χ	√	10.39	21.67	8.54
Transformer-XLRG [12]	χ	√	10.17	20.40	8.85
MART [22]	χ	√	10.33	23.42	5.18
EMT [5]	χ	√	10.24	23.66	**4.27**
Adaptive Transformer [6]	χ	√	10.38	**24.22**	5.84
HSAT	χ	√	10.31	23.76	5.85

Despite the superior result for BLEU-4, achieved by the GVDsup method, to those found in Table 1, as it is not considered the best metric for the video captioning task, the results do not represent a marked improvement as found by CIDEr-D. According to [27], the use of BLEU unigram compares the evaluation of the simple precision of the method, characterized by the simple count of correct words divided by the total number of words in the sentence. On the other hand, the CIDEr-D score proposes to measure the best textual sentence among the candidates by the majority of simple votes. However, the B@4 score is considered an interesting metric for some NLP tasks, such as machine translation since, if the sentence is very close to most GT sentences used as a reference, the probability is greater that the sentence is correct. This evaluation method seeks to bring the human description closer to that described by machine translation, as human evaluation is inherent to the perception of the person describing the scene in focus [33].

The results shown by Table 2 present the comparison of the proposed model and relation to the performance of models based on transformers in AE-TEST split of the ANC dataset (similar to what was done in [5, 6, 22]). Again, the Adaptive Transformer presents better results than the others, mainly for the CIDEr-D metric and followed closely by HSAT.

In summation, as one can see in Table 1 and 2, the results achieved by the proposed model are superior, except when compared to Adaptive Transformer whose re-

Table 2 Performance of our model and other transformer-based methods in AE-TEST split of ActivityNet Captions (**Rec** indicates whether sentence-level recurrence is used).

	Rec	B@4 ↑	CIDEr-D ↑	R@4 ↓
Transformer-based methods				
VTransformer [43]	χ	9.31	21.33	7.45
Transformer-XL [12]	√	**10.25**	21.71	8.79
Transformer-XLRG [12]	√	10.07	20.34	9.37
MART [22]	√	9.78	22.16	5.44
EMT [5]	√	10.00	22.84	**4.55**
Adaptive Transformer [6]	√	10.00	**23.04**	5.29
HSAT	√	9.94	22.97	5.35

sults HSAT follows closely. The changes achieved imply an increase in similarity with the GT results but with a reduction of repetition without losing cohesion among the generated sentences for describing each video.

4.5 Qualitative analysis of adaptive transformer

To further promote the perception of the results obtained and their improvements, Figure 5 presents samples of paragraphs generated by the Adaptive Transformer [6] (without any summarization) and those produced by other approaches from the literature, i.e., Vanilla Transformer [43], Transformer -XL [12], MART [22] and EMT [5], in addition to the GT results. With this, it is possible to compare the different results found by each approach, making it possible to evaluate the coherence, similarity with the GT results, and the repeatability rate of the descriptions generated by the Adaptive Transformer and the other methods. The inconsistencies of every approach had been highlighted to facilitate the visualization: (i) red/bold for cases of use of different pronouns from the ones in the GT result (or when they are misused); and (ii) blue/bold for the occurrences of a repeated sentence in the paragraph.

The paragraph descriptions generated by the Vanilla Transformer cannot prevent repetition, and, in many cases, there is no similarity between the produced text and the GT result. In Figure 5a and 5b, it is easy to observe that the paragraph produced by Vanilla Transformer does not have fluidity, and the notion of continuity is lost. The Vanilla Transformer is the method with the highest repetition rate among those presented. The Transformer-XL can return fluid and continuous paragraphs, but it has a high repeatability rate, as illustrated in Figure 5a and, in some cases does not return a discriminating description, as one can see in Figure 5b. MART manages to maintain coherence among generated sentences and a lower repetition rate. However, in some cases, MART appears to produce less detailed descriptions, as shown in Figure 5b. In the results for EMT is possible to observe that it preserves coherence and context with a low repeatability rate. In Figure 5b, the generated paragraph maintains the context but doesn't get the second pronoun right. The qualitative results for the Adaptive Transformer are closer to the GT, it is possible to notice that the descriptions have coherence and fluidity. In addition, repeatability is reduced. Figure 5a presents a very discriminating result and captures the continuity of the scene and, despite the description being longer than the GT, there is no repetition. Despite the repetition present in Figure 5b, it is possible to notice that the result presents continuity in the description.

4.6 Qualitative analysis of HSAT

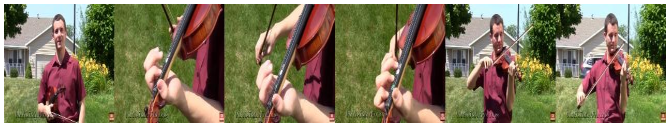
The results found for HSAT demonstrate that it presents an improvement related to the detection of video events. When compared to the sequential selection of frames, the amount of information that the method does not observe/process is large.

In some cases, when the sequential selection of frames is used, only the first one hundred frames with a rate of 2 FPS are used to represent the video. Thus, for A video with a length greater than 50 seconds, all information after the first 50 seconds is ignored. On the ANC dataset, the videos do not have the same length, the number of events varies from 2 to 6, and, in some cases, one video has more than two hundred seconds. Because of this, summarization appears as a better way to evaluate the content distributed in the entire video. Thus, the neglected information due to time limitations adopted in a sequential selection of frames does not exist with the hierarchical summarization approach.

Despite that HSAT selects a relatively less number of frames (only 10), it is sufficient to cover all videos of the dataset (since the number of events in ANC dataset varies from 2 to 6).

Figures 6 and 7 show the diversity of video content present in the dataset. Figure 6 shows the summarization result in a short video that has 25 seconds. Since it is a short video, the summarization process returns similar frames, however, with some minor variations in perspective. In video summarization, the amount of frames remains the same for all videos and the fluidity of the video is maintained. In addition, it is possible to correctly follow the actions over time without neglecting the video context.

Figure 7 shows the frames selected as Keyframes for the HSAT method. While Figure 8 illustrates the selected frames when a sequential selection (with time constraints) is made. One can observe that in Figure 8 some content is not present at all. In contrast, HSAT manage to obtain a greater variety of video content making it easier to describe different moments of the video. In the sequential selection of frames, since that video has 80 seconds, it disregards any information that occurs in the final 30 seconds. In turn, HSAT uses hierarchical summarization to cover a greater variety of instants. In this way, HSAT only disregards very similar



Vanilla Transformer
He continues speaking while holding the violini and showing how to play his hands. He continues playing the instrument while looking down at the camera. He continues playing the violin and then stops to speak to the camera.

Transformer-XL
A man is seen speaking to the camera while holding a violin. The man continues playing the instrument while moving his hands up and down. The man continues playing the instrument and ends by looking back to the camera

MART
A man is seen speaking to the camera while holding a violin and begins playing the instrument. The man continues to paly the instrument while moving his hands up and down. He continues to play and ends by moving his hands up and down.

EMT
A man is seen speaking to the camera while holding a violin and playing his hands. He then moves the instrument all around his hands as well as the other hand movements. He continues playing the instrument and ends by looking back to the camera.

Adaptive Transformer
A man is seen speaking to the camera while holding a musical instrument and begins playing the instrument. The man continues to play the instrument while looking off into the distance and smiling to the camera. He continues moving his hands around to play and showing off the proper hand as well as showing how to properly play.

Ground-Truth
A man is seen looking to the camera while holding a violin. The man then begins playing the instrument while the camera zooms in on his fingers. The man continues to play and stops to speak to the camera.

(a)

Vanilla Transformer
She continues moving around the room and leads into her speaking to the camera. She continues moving around on the step and ends by speaking to the camera.

Transformer-XL
A woman is standing in a gym. She begins to do a step

MART
A woman is standing in a room talking. She starts working out on the equipment

EMT
A woman is seen speaking to the camera while standing in front of a board. The woman then begins moving her arms and legs around while still speaking to the camera.

Adaptive Transformer
A woman is in a room in front of a step and performs a routine while speaking to the camera. She steps up and down on a blue mat.

Ground-Truth
A woman is seen speaking to the camera and leads into her walking up and down the board. She then stands on top of the beam while speaking to the camera continuously.

(b)

Fig. 5 Examples (for qualitative analysis) of results obtained by Adaptive Transformer, compared to Vanilla Transformer, Transformer-XL, MART, EMT and GT results, in which blue/bold indicates the presence of repetition and red/bold indicates a possible pronoun different form the GT. Best viewed in color.

Fig. 6 A result example of HSAT showing fluidity in movement variation.

frames that are direct neighbors in time to include more distinct and meaningful frames for the video description. Due to the summarization process, the number of frames

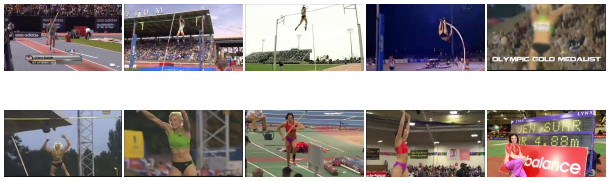


Fig. 7 A result example of HSAT with a greater number of distinct keyframes. In this case, the result should cover more than one point of view. Even so, the video summarization approach managed to capture frames that did not appear in a sequential selection of frames.

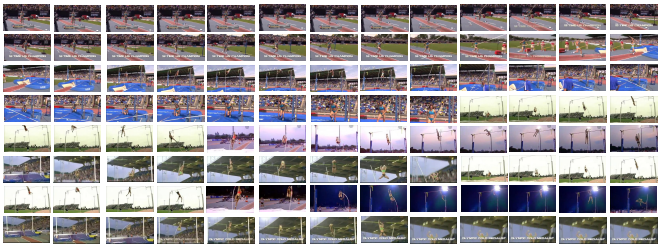


Fig. 8 An example of selected frames by a sequential selection (with time constraints) similar to the approach used in [6].

used can be reduced, generating results as significant as for techniques with large amounts of frames.

Figure 9 presents a qualitative comparison between the result obtained by the HSAT with the Adaptive Transformer [6]. As one can see, the result is very discriminative and does not have many repetitions of terms. The results presented in both situations (9a and 9b) show very approximate descriptions, but with some points described from another perspective. Thus, as HSAT uses features taken from different regions of the video and analyzes the importance of each frame in time, the modifications related to the description are due to the presence of points that may not be visualized in the same set of frames used to illustrate

Adaptive Transformer
A man is seen speaking to the camera while holding a musical instrument and begins playing the instrument. The man continues to play the instrument while looking off into the distance and smiling to the camera. He continues moving his hands around to play and showing off the proper hand as well as showing how to properly play.

HSAT
A man is seen speaking to the camera while holding a violin and the instrument. The man continues to play and then pauses to speak to the camera. He continues moving his hands around to play and is seen speaking to the camera.

Ground-Truth
A man is seen looking to the camera while holding a violin. The man then begins playing the instrument while the camera zooms in on his fingers. The man continues to play and stops to speak to the camera.

(a)

Adaptive Transformer
A woman is in a room in front of a step and performs a routine while speaking to the camera She steps up and down on a blue mat.

HSAT
A woman is seen speaking to the camera while standing in front of a board. She begins moving up and down the board while speaking to the camera.

Ground-Truth
A woman is seen speaking to the camera and leads into her walking up and down the board She then stands on top of the beam while speaking to the camera continuously

(b)

Fig. 9 Examples (for qualitative analysis) of results obtained by HSAT, compared to Adaptive Transformer and GT results. We use the same set of frames only to exemplify how video are described.

the video content. In this way, hierarchical summarization presents itself as a great candidate to improve the descriptions produced for the video captioning task.

4.7 Ablation study

As a way to demonstrate the effectiveness of the proposed architectures, we modify the memory size from 2 to 1 and 4. And, to ensure the effectiveness of the adaptive attention after the previous attention module, the removal (in an alternate way) of the adaptive attention modules from the proposed positions was also analyzed. In addition, with the modification of memory size, for each variation, the adaptive attention module was reassessed. In this way, it was possible to measure the impact achieved by the inclusion of adaptive attention as a way to modify both multi-head attention modules. Those changes provide an ablation study, allowing us to identify the best configuration of the proposed architecture. Table 3 presents the obtained results in the ablation study of the proposed architecture along with the original results to facilitate comparison.

The evaluation of the results obtained for the CIDEr-D score demonstrates the effectiveness of the proposed configuration to the detriment of the others. Furthermore, one can notice that the reduction of the Repetition-4 score is relatively low, but for the BLEU-4 there was a slight increase in the score, however, as discussed before, it is not a good metric for the video captioning task.

5 Conclusion

This work presents a new method named **H**ierarchical time-aware **S**ummarization with an **A**daptive **T**ransformer – HSAT.

Table 3 Performance in the AE-VAL split of ActivityNet Captions during the ablation study for verifying the quality of results achieved for the proposed architectures in which (*) denotes transformer without adaptive attention after the second Multi-Head Attention, and (+) denotes transformer without adaptive attention after the first Multi-Head Attention.

Model	Mem Size	Hidden Size	B@4 ↑	CIDEr-D ↑	R@4 ↓
Ours	1	1200	**10.55**	23.58	**5.48**
Ours*	1	1200	10.28	22.90	6.35
Ours+	1	1200	10.23	22.25	6.45
Ours	2	1200	10.38	**24.22**	5.84
Ours*	2	1200	10.10	23.55	5.89
Ours+	2	1200	10.52	23.88	5.97
Ours	4	1200	10.53	23.01	6.69
Ours*	4	1200	10.16	22.74	5.80
Ours+	4	1200	10.43	22.64	6.86

The HSAT presents a strategy to enhance the frame selection reducing the amount of information that needed to be processed during the description generation step without any loss of content. Thus, the summarization process used before the description generation step as an approach to evaluate the importance of each frame demonstrated the effectiveness of the use of the most informative frames instead of selecting frames sequentially. Allied with this, the processing time to deal with more informative frames is reduced because they represent only a small number of frames that may not even be processed with other techniques. Together with the summarization approach, another improvement in generated descriptions is a consequence of the local analysis and refinement of the adaptive attention results. The results achieved by the proposed model surpass those presented by state-of-the-art methods in the literature and are equivalent to those obtained by the Adaptive Transformer using only a small fraction of the processing time spent by the latter. Quantitative and qualitative evalua-

tions showed that the proposed model produces more coherent and diversified results, with high similarity with GT and lower repetition rates.

Future works may explore other attention mechanisms and different architectures for the transformer's main backbone. It may also be interesting to investigate the impact of uses reinforcement learning techniques in event detection and, consequently, in the final video description.

Acknowledgements

The authors would like to thank Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq – (PQ 310075/ 2019-0), and Fundação de Amparo à Pesquisa do Estado de Minas Gerais – FAPEMIG – (Grants PPM- 00006-18). This study was also financed in part by PUC Minas and by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

References

1. Aafaq, N., Mian, A., Liu, W., Gilani, S.Z., Shah, M.: Video description: A survey of methods, datasets, and evaluation metrics. ACM Computing Surveys (CSUR) **52**(6), 1–37 (2019)
2. Apostolidis, E., Adamantidou, E., Metsai, A.I., Mezaris, V., Patras, I.: Video summarization using deep neural networks: A survey. Proceedings of the IEEE **109**(11), 1838–1863 (2021). DOI 10.1109/JPROC.2021.3117472
3. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: ICLR (2015)
4. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: IEEE CVPR, pp. 961–970 (2015)
5. Cardoso, L.V., Guimaraes, S.J.F., Patrocínio, Z.K.: Enhanced-memory transformer for coherent paragraph video captioning. In: 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI), pp. 836–840. IEEE (2021)
6. Cardoso, L.V., Guimaraes, S.J.F., Patrocinio, Z.K.: Exploring adaptive attention in memory transformer applied to coherent video paragraph captioning. In: 2022 IEEE Eighth International Conference on Multimedia Big Data (BigMM), pp. 37–44. IEEE (2022)
7. Chen, C.F., Fan, Q., Panda, R.: Crossvit: Cross-attention multi-scale vision transformer for image classification. arXiv preprint arXiv:2103.14899 (2021)
8. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Learning universal image-text representations. ICLR (2019)
9. Cho, K., van Merrienboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: EMNLP, pp. 1724–1734. ACL (2014)
10. Cousty, J., Najman, L.: Incremental algorithm for hierarchical minimum spanning forests and saliency of watershed cuts. In: Mathematical Morphology and Its Applications to Image and Signal Processing: 10th International Symposium, ISMM 2011, Verbania-Intra, Italy, July 6-8, 2011. Proceedings 10, pp. 272–283. Springer (2011)
11. Cousty, J., Najman, L., Kenmochi, Y., Guimarães, S.: Hierarchical segmentations with graphs: Quasi-flat zones, minimum spanning trees, and saliency maps. J. Math. Imaging Vis. **60**(4), 479–502 (2018)
12. Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., Salakhutdinov, R.: Transformer-XL: Attentive language models beyond a fixed-length context. In: 57th Annual Meeting of the ACL, pp. 2978–2988 (2019)
13. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
14. Guimarães, S., Kenmochi, Y., Cousty, J., Patrocinio, Z., Najman, L.: Hierarchizing graph-based image segmentation algorithms relying on region dissimilarity: the case of the felzenszwalb-huttenlocher method. Mathematical Morphology-Theory and Applications **2**(1), 55–75 (2017)
15. Guo, Q., Qiu, X., Liu, P., Xue, X., Zhang, Z.: Multi-scale self-attention for text classification. In: AAAI, vol. 34, pp. 7847–7854 (2020)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE CVPR, pp. 770–778 (2016)
17. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
18. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: IEEE CVPR, pp. 7132–7141 (2018)
19. Huang, L., Wang, W., Chen, J., Wei, X.Y.: Attention on attention for image captioning. In: IEEE ICCV, pp. 4634–4643 (2019)
20. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML, pp. 448–456. PMLR (2015)
21. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Carlos Niebles, J.: Dense-captioning events in videos. In: IEEE ICCV, pp. 706–715 (2017)
22. Lei, J., Wang, L., Shen, Y., Yu, D., Berg, T., Bansal, M.: MART: Memory-augmented recurrent transformer for coherent video paragraph captioning. In: 58th Annual Meeting of the ACL, pp. 2603–2614 (2020)
23. Meena, P., Kumar, H., Kumar Yadav, S.: A review on video summarization techniques. Engineering Applications of Artificial Intelligence **118**, 105,667 (2023). DOI https://doi.org/10.1016/j.engappai.2022.105667
24. Najman, L., Cousty, J., Perret, B.: Playing with kruskal: algorithms for morphological trees in edge-weighted graphs. In: Mathematical Morphology and Its Applications to Signal and Image Processing: 11th International Symposium, ISMM 2013, Uppsala, Sweden, May 27-29, 2013. Proceedings 11, pp. 135–146. Springer (2013)
25. Narwal, P., Duhan, N., Kumar Bhatia, K.: A comprehensive survey and mathematical insights towards video summarization. Journal of Visual Communication and Image Representation **89**, 103,670 (2022). DOI https://doi.org/10.1016/j.jvcir.2022.103670
26. Pan, Y., Yao, T., Li, Y., Mei, T.: X-linear attention networks for image captioning. In: IEEE CVPR, pp. 10,971–10,980 (2020)
27. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: 40th annual meeting of the ACL, pp. 311–318 (2002)
28. Park, J.S., Rohrbach, M., Darrell, T., Rohrbach, A.: Adversarial inference for multi-sentence video description. In: IEEE CVPR, pp. 6598–6608 (2019)
29. dos Santos Belo, L., Caetano Jr, C.A., do Patrocínio Jr, Z.K.G., Guimarães, S.J.F.: Summarizing video sequence using a graph-based hierarchical approach. Neurocomputing **173**, 1001–1016 (2016)

30. Sun, C., Myers, A., Vondrick, C., Murphy, K., Schmid, C.: Videobert: A joint model for video and language representation learning. In: IEEE ICCV, pp. 7464–7473 (2019)
31. Tang, H., Ji, D., Li, C., Zhou, Q.: Dependency graph enhanced dual-transformer structure for aspect-based sentiment classification. In: 58th Annual Meeting of the ACL, pp. 6578–6588 (2020)
32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
33. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: IEEE CVPR, pp. 4566–4575 (2015)
34. Vydana, H.K., Karafiát, M., Zmolikova, K., Burget, L., Černockỳ, H.: Jointly trained transformers models for spoken language translation. In: IEEE ICASSP, pp. 7513–7517 (2021)
35. Xiong, Y., Dai, B., Lin, D.: Move forward and tell: A progressive generator of video descriptions. In: ECCV, pp. 468–483 (2018)
36. Xiong, Y., Wang, L., Wang, Z., Zhang, B., Song, H., Li, W., Lin, D., Qiao, Y., Van Gool, L., Tang, X.: CUHK & ETHZ & SIAT submission to activitynet challenge 2016. arXiv preprint arXiv:1608.00797 (2016)
37. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: ICML, pp. 2048–2057. PMLR (2015)
38. Yates, A., Nogueira, R., Lin, J.: Pretrained transformers for text ranking: Bert and beyond. In: 14th ACM International Conference on Web Search and Data Mining, pp. 1154–1156 (2021)
39. Zhang, A., Lipton, Z.C., Li, M., Smola, A.J.: Dive into deep learning. arXiv preprint arXiv:2106.11342 (2021)
40. Zhang, B., Hu, H., Sha, F.: Cross-modal and hierarchical modeling of video and text. In: ECCV, pp. 374–390 (2018)
41. Zhang, X., Wei, F., Zhou, M.: HIBERT: Document level pretraining of hierarchical bidirectional transformers for document summarization. In: 57th Annual Meeting of the ACL, pp. 5059–5069 (2019)
42. Zhou, L., Kalantidis, Y., Chen, X., Corso, J.J., Rohrbach, M.: Grounded video description. In: IEEE CVPR, pp. 6578–6587 (2019)
43. Zhou, L., Zhou, Y., Corso, J.J., Socher, R., Xiong, C.: End-to-end dense video captioning with masked transformer. In: IEEE CVPR, pp. 8739–8748 (2018)