Time-Aware Circulant Matrices for Question-Based Temporal Localization

Pierfrancesco Bruni[1], Alex Falcon[1(✉)], and Petia Radeva[2]

[1] University of Udine, Udine, Italy
{pierfrancesco.bruni,falcon.alex}@spes.uniud.it
[2] University of Barcelona, Barcelona, Spain
petia.ivanova@ub.edu

Abstract. Episodic memory involves the ability to recall specific events, experiences, and locations from one's past. Humans use this ability to understand the context and significance of past events, while also being able to plan for future endeavors. Unfortunately, episodic memory can decline with age and certain neurological conditions. By using machine learning and computer vision techniques, it could be possible to "observe" the daily routines of elderly individuals from their point of view and provide customized healthcare and support. For example, it could help an elderly person remember whether they have taken their daily medication or not. Therefore, considering the important impact on healthcare and societal assistance, this problem has been recently discussed in the research community, naming it Episodic Memory via Natural Language Queries. Recent approaches to this problem mostly rely on the literature related to similar fields, but contextual information from past and future clips is often unexplored. To address this limitation, in this paper we propose the Time-aware Circulant Matrices technique, which aims at introducing awareness of the surrounding clips into the model. In the experimental results, we present the robustness of our method by ablating its components, and confirm its effectiveness on the Ego4D public dataset, achieving an absolute improvement of more than 1% on R@5.

Keywords: Natural language query for temporal localization · Cross-modal understanding

1 Introduction

The ability to remember events that occur in our lives is a fundamental aspect of human cognition, known as episodic memory [27]. As humans, we can remember past experiences and recall specific details about them, such as when and where they occurred, who was present, and what happened. For instance, to follow a balanced diet, we may want to prepare dinner depending on what we had for lunch, which entails our ability to precisely recall the ingredients in order to compute their macro nutrients. Similarly, elderly people may need to ingest several medications throughout the day, but are they able to recall whether they took

G. L. Foresti et al. (Eds.): ICIAP 2023, LNCS 14234, pp. 182–195, 2023.
https://doi.org/10.1007/978-3-031-43153-1_16

everything they need or not? To support the users in these situations by means of an intelligent system, two main components are needed. First, augmented reality glasses and similar vision systems can be used to capture the environment and the interactions with it over time, through a first-person perspective. Second, the system needs to process these visual information and then, once the user asks a question, provide the correct answer by contextualizing them to the contents of the question. Given the need to process both visual and auditory data, this second component requires the use of a variety of artificial intelligence techniques. One recent approach to it is Episodic Memory via Natural Language Queries (NLQ) [15], in which the questions are expressed in textual form to leverage the recent advancements in natural language processing.

To address the new problem, two baselines were utilized, drawing inspiration from previous works on temporal activity localization (TAL), which required to identify and localize simple actions in a video [33,34]. Compared to TAL, NLQ is more difficult as it requires to localize the moment in time from which the answer to an input question can be deduced. Nonetheless, there are several shared problems between TAL and NLQ, including the length of the untrimmed videos and the need to capture multimodal interactions. To deal with these problems, Ge et al. in [12] discovered both textual and visual concepts and used them to ease activity localization, using both the sentence/video embeddings and the concepts embeddings, e.g. verb-object textual pairs and high level concepts extracted from pretrained deep networks. Wu et al. in [28] proposed Multimodal Circulant Fusion, which allowed for multimodal interactions between the visual features and the circulant matrix of textual features, and vice versa, leading to improved localization accuracy. Recently, Zhang et al. in [35] simultaneously explored intra- and inter-modal relations through a multimodal interaction graph. While these methods propose solutions that allow for an improved understanding of the data under analysis, the temporal relations between frames and short clips are often neglected in later parts of the network architecture. To address this limitation, we introduce the Time-aware Circulant Matrices technique, enabling an improved intra-modal reasoning by injecting temporal awareness into later parts of the network. We confirm the effectiveness of our method by testing it on the Ego4D dataset [15], in which we improve the baseline performance in all the metrics under consideration. Moreover, we perform ablation studies and experiments to support our design choices.

The main contribution of this work can be summarized as follows:

- we propose to address the NLQ task by introducing the Time-aware Circulant Matrices technique, which injects temporal awareness by modelling the local context and taking it into account when performing the analysis of the visual features;
- by testing our solution in the Ego4D benchmark, we show that our proposed method achieves considerable improvements in all the metrics under consideration.

After this introduction, the related work is described in Sect. 2. Then, Sect. 3 presents and motivates the proposed Time-aware Circulant Matrices technique.

The experimental results are presented in Sect. 4 and, finally, Sect. 5 concludes the manuscript.

2 Related Work

2.1 Episodic Memory via Natural Language Queries

This challenging problem aims at identifying the moments in a video which contain relevant information to provide the correct answer to a given question. Note that the answer is found within the video (e.g., where did I forget the car keys?), hence why it is called *episodic*; in contrast, *factual/semantic* memory refers to the ability of recalling the correct answer from external knowledge bases (e.g., does Italy share a border with Spain?). Recent advancements on this topic are mostly related to the homonymous Ego4D benchmark track. The initial baselines, VSLNet [33] and 2D-TAN [34], were inspired from previous works on language grounding in video. The former implements at its core a 2D map of adjacent moment candidates, which are then queried by the sentence representation to obtain the best matching one; whereas the latter directly regresses the start and end boundaries from the input visual and textual features, supporting this process by means of a query-guided highlighting module. Building upon these works, several solutions were recently proposed. Lin et al. [18] used VSLNet on top of pretrained EgoVLP features [19]. To tackle the low amount of videos, ReLER [21] proposed data augmentation techniques on top of a multi-scale Transformer-based encoder for VSLNet and several pre-extracted visual and textual features [9,13,24]. Hou et al. in [16] proposed a three-stage approach consisting of feature filtering, using a pre-trained video-language model (EgoVLP [19]), moment proposal with Moment-DETR in [17] extended with inter-windows contrastive learning, and finally a novel intra-windows fine-grained ranking strategy. Mo et al. in [23] used a simple Transformer-based method called ActionFormer [32]. A foundational model, InternVideo in [4], recently obtained state-of-the-art results on dozens of challenges and datasets, and was used as a backbone for VSLNet.

Differently from them, we focus on the modelling aspects and propose a novel technique, which we call Time-aware Circulant Matrices, to analyze the visual features and peek into the surrounding context to discover underlying patterns. Circulant matrices were also used in a previous work dealing with TAL [28] to compute additional relations between multiple sources of information. In this work, we further extend this technique by integrating contextual awareness in its framework.

2.2 Temporal Action Localization and Video Moment Retrieval

Temporal action localization and video moment retrieval are similar tasks to the problem under analysis. *Temporal action localization* requires to identify each action instance in the video, predict its temporal boundaries, and categorize it

in a finite set of classes. This problem is typically tackled either in a two-stage or a single-stage fashion. The former starts by first generating coarse video segments as action proposals (e.g., by using anchor windows [3,8] or action boundaries [14,36]), and then by classifying them with action recognition models. In the latter strategy, the proposed approaches try to simultaneously locate and classify the actions without relying on generated action proposals or external classifiers [5,20]. *Video moment retrieval* (VMR) is even closer to Episodic memory via Natural Language Queries, since it requires to localize and retrieve the moments which are described by an input textual query. As in the previous case, one-stage and two-stage approaches have been proposed for VMR. In two-stage approaches, the input video is first split into multiple candidate moments (e.g., by using a sliding window approach) which are then ranked to select the best matching ones [10,34]; in the one-stage scenario, no predefined candidate moments are used and each frame is a possible candidate to represent the initial or final frame of the moment [6,31].

However, both these problems present some fundamental differences with NLQ, leading to models with different capabilities and goals. In fact, while TAL requires to identify and localize simple actions in a video, and VMR requires to retrieve all the moments which can be described by an input textual query, NLQ requires to locate a precise moment depicting certain visual cues from which it is possible to infer the answer to a given question. For instance, this means that if we are trying to recall *the color of the dress worn by the person we spoke with*, in VMR we either need to locate all the moments in which we *interact/speak with someone*, resulting in a coarse selection which would need further processing, or we need to already know the answer, i.e., the color of the dress, and insert it into the query. Therefore, while both these problems are related to the NLQ, they aim at solving different tasks.

3 Proposed Method

An overview of the method is shown in Fig. 1. We start by briefly describing the overall procedure (Sect. 3.1), then we focus on the details of the proposed Time-aware Circulant Matrices technique in Sect. 3.2.

3.1 Overview of the Procedure

Starting from the visual and textual features, $V \in \mathbb{R}^{n \times f_v}$ and $Q \in \mathbb{R}^{m \times f_t}$, a convolutional layer is applied to project the heterogeneous features to the same dimension d, i.e., $V' \in \mathbb{R}^{n \times d}$, $Q' \in \mathbb{R}^{m \times d}$. Then, a Feature Encoder made of a single Transformer Encoder is used to learn for both of them an independent representation in a common space, resulting in $\tilde{V}$ and $\tilde{Q}$.

To model the cross-modal interactions and discover more patterns in the underlying visual data, while also leveraging temporal relations in their progression over time, we use the following equation, based on Context-Query Attention

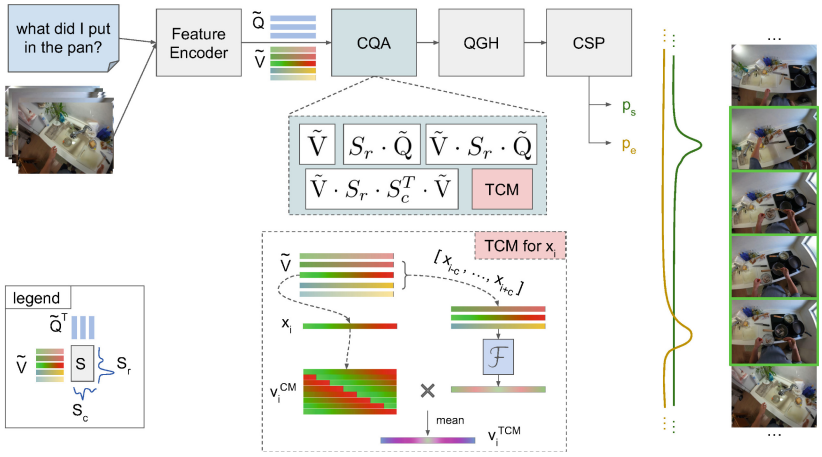

Fig. 1. Overview of the proposed method, Time-aware Circulant Matrices (TCM). The legend shows how S_r and S_c are computed from $\tilde{V}$ and $\tilde{Q}$. More details about the proposed method, TCM, and the other components, CQA, QGH, and CSP can be found in Sect. 3. Best viewed in color.

(CQA) [33] and inspired from previous works, e.g., [29,30]:

$$V^Q = FFN([\tilde{V}, V^{TCM}, A, \tilde{V} \circ A, \tilde{V} \circ B]) \tag{1}$$

where FFN is a linear layer and the other components are obtained as follows. V^{TCM} is derived through the novel three-step process which we name Time-aware Circulant Matrices and explain in Sect. 3.2. A and B are obtained by attending $\tilde{Q}$ and $\tilde{V}$ through S_r and S_c, which are the row-wise and column-wise softmax normalization of S, i.e., their similarity matrix, $S = \tilde{V} \cdot \tilde{Q}^T, S \in \mathbb{R}^{n \times m}$. Specifically, $A = S_r \cdot \tilde{Q}, A \in \mathbb{R}^{n \times d}$ and $B = S_r \cdot S_c^T \cdot \tilde{V}, B \in \mathbb{R}^{n \times d}$. The element-wise multiplication is depicted with $\circ$. The features in V^Q are then combined with h_Q, i.e., a sentence-level representation of the word features $\tilde{Q}$ obtained through content-based attention [1], to form $\hat{V}^Q = \left[[v_1^Q; h_Q], [v_2^Q; h_Q], \ldots, [v_n^Q; h_Q]\right]$.

Then, the Query-Guided Highlighting (QGH) module introduced in [33] is responsible for discriminating "foreground" moments, i.e., those which are relevant to the target, from the "background" ones, while also allowing for some flexibility in the boundaries: by labeling each moment with a binary label (0 for background, 1 for foreground), QGH reduces to a binary classification problem which helps highlighting important features obtained as $\tilde{V}^Q = S_h \cdot \hat{V}^Q$, where S_h is the highlighting score, computed as $\sigma(Conv1D(\hat{V}^Q))$.

Finally, the prediction of the boundaries is done by the Conditioned Span Predictor (CSP), which uses a shared Transformer Encoder with 4 layers, followed by two convolutional-based networks, to predict the start and end probability distributions, $p_s, p_e \in \mathbb{R}^n$, starting from the $\tilde{V}^Q$ features. The model is trained

by means of the following joint loss function:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{span} + \mathcal{L}_{QGH} \\ \mathcal{L}_{span} &= \frac{1}{2}\big(f_{CE}(p_s, y_s) + f_{CE}(p_e, y_e)\big) \\ \mathcal{L}_{QGH} &= f_{BCE}(S_h, Y_h) \end{aligned} \tag{2}$$

where f_{CE} is the Cross Entropy loss computed between the groundtruth and the predicted probability distributions for start, y_s and p_s, and end boundaries, y_e and p_e; and f_{BCE} is the Binary Cross Entropy loss used by the QGH module to compute the loss between the predicted and groundtruth highlighting scores, S_h and Y_h.

3.2 Time-Aware Circulant Matrices Technique

The proposed Time-aware Circulant Matrices technique consists of a three-steps procedure used to discover additional temporal intra-modal relations in the visual features.

First of all, given the video features $V = [x_1, x_2, \ldots, x_N]$, each clip vector $x_i \in \mathbb{R}^{1\times d}$ is transformed into its circulant matrix, $v_i^{CM} \in \mathbb{R}^{d\times d}$. This is obtained by the following equation:

$$v_i^{CM} = (\overrightarrow{x_i}^0 \overrightarrow{x_i}^1 \ldots \overrightarrow{x_i}^{d-1})^T \tag{3}$$

where $\overrightarrow{x_i}^j = [x_{(d-1)-j+1}, x_{(d-1)-j+2}, \ldots, x_0, \ldots, x_{(d-1)-j}]$, i.e., $\overrightarrow{x_i}^j$ represents a shift in x_i by j.

Secondly, the resulting v_i^{CM} is multiplied by $\mathcal{F}([x_{i-c}, \ldots, x_i, \ldots, x_{i+c}])$ to establish further relations between each visual feature and the surrounding clips. By doing so, temporal awareness is injected into the model and, using a context of size c, the contextual information is obtained through the aggregator function $\mathcal{F}$. Formally:

$$Z_i = v_i^{CM} * \mathcal{F}([x_{i-c}, \ldots, x_{i+c}]) \tag{4}$$

where $\mathcal{F}([x_{i-c}, \ldots, x_{i+c}]) \in \mathbb{R}^{1\times d}$ is broadcast to all the rows in v_i^{CM}. To implement $\mathcal{F}$, we consider a linear transformation of the concatenated context as in the following equation:

$$\mathcal{F}([x_{i-c}, \ldots, x_{i+c}]) = W_{lin}[x_{i-c}, \ldots, x_{i+c}] + b_{lin} \tag{5}$$

where W_{lin} and b_{lin} are learned at training time.

Lastly, the column-wise average of Z_i is utilized to collate the newly identified information, leading to v_i^{TCM}:

$$v_i^{TCM} = \frac{1}{d}\sum_{j=1}^{d} z_i^{(j)} \tag{6}$$

which represents the i-th vector of $V^{TCM} \in \mathbb{R}^{N\times d}$, the output of the proposed Time-aware Circulant Matrices technique.

By performing these three steps, we obtain different combinations of the visual features, possibly leading to the discovery of additional relations which were not previously considered. Note that the use of circulant matrices was also considered in [28] for better video-language understanding capabilities. However, up to our knowledge they have not been used for video temporal modelling.

4 Experimental Results

In this section, we first discuss the dataset and evaluation metrics, and the implementation details. Then, we present the experimental results related to the temporal context modeling and width, a study on the use of an asymmetric context, an ablation study on the proposed technique, and a comparison with the state of the art.

4.1 Dataset and Evaluation Metrics

The Ego4D dataset [15] is a large-scale collection of egocentric perspective videos that comprises over 3000 hours of footage. The videos are divided into clips which are 8 min long and annotated by a short narration (6–8 words), yielding roughly 3.85 million annotations. About 17000 queries for training and validation are created from these annotations for the Episodic Memory via Natural Language Queries task, using different templates such as "where is X after Y?", where X is an object and Y an event, and "who did I talk to in Z?", where Z is a location. The main evaluation metric used is Recall@k, IoU=m, which measures the proportion of instances where the intersection-over-union (IoU) between the ground truth interval and at least one of the top k predictions is greater than or equal to m. The values of k and m used are $k = 1, 5$ and $m = 0.3, 0.5$.

4.2 Implementation Details

To implement the proposed method, we start from the official codebase provided for VSLNet[1]. The PyTorch version is 1.11.0. The training procedure lasts for 200 epochs and the best model on the validation set is selected. The optimizer used is AdamW with a learning rate of 0.0001. The batch size is 32 and we used 512 as the maximum for the positional embedding in the Feature Encoder and the CSP. We use BERT [7] for the textual features and InternVideo for the visual ones [4].

4.3 Temporal Context Modeling and Width

We consider three possibilities for the function $\mathcal{F}$ which is used to aggregate the temporal context $[x_{i-c}, \ldots, x_{i+c}]$ as detailed in Eq. 4. These include: the linear transformation of the concatenated context (*cat+lin*), which is part of

[1] https://github.com/EGO4D/episodic-memory/tree/main/NLQ/VSLNet.

the proposed method; a simple *mean* pooling, in which $\mathcal{F}([x_{i-c}, \ldots, x_{i+c}]) = \frac{1}{2*c+k} \sum_{j=i-c}^{i+c} x_j$, where c is the size of the context (see Sect. 3) and $k = 1$ if x_i is used, otherwise $k = 0$; and finally a *GRU* model, in which $\mathcal{F}([x_{i-c}, \ldots, x_{i+c}]) = h_{2*c+k}$, where h_{2*c+k} is the last hidden state computed by the GRU.

Figure 2 reports the results obtained by the different models as the size of the considered temporal context increases. Specifically, on the top line R@1 is shown (respectively, with IoU = 0.3 and IoU = 0.5), whereas the bottom line displays R@5 values. In each of the four plots, the result achieved by VSLNet is reported for reference. Overall, it can be seen that introducing the contextual information obtained from the surrounding clips can be helpful. Specifically, the *cat+lin* strategy achieves good results when the context is small, e.g., it achieves better R@1 (IoU=0.5) and R@5 performance than the mean pooling when $c = 2$ (19.7% and 14.6% respectively for IoU=0.3 and IoU=0.5); its performance decreases as c increases, most likely because each time an element is added to the context, the weight parameter becomes bigger, possibly leading to higher memorization and lower generalization. The mean pooling leads to generally good results, with two best solutions: when x_i is not included in the context, $c = 20$ leads to the highest metrics, e.g., it achieves 12.1% R@1 IoU=0.3 (+1.3% than the baseline) and 7.4% R@1 IoU=0.5 (+0.4%); whereas $c = 10$ is preferred when x_i is included, e.g., it achieves similar R@1 in both IoU thresholds, but better R@5 (19.4% R@5 IoU=0.3, +0.9% than the baseline, and 14.5% R@5 IoU=0.5, +1.3%). Finally, the GRU solution does not lead to improvements when compared to the baseline.

The solution obtained by the concatenation and the linear leads to the best performance across most of the considered metrics, therefore it was chosen for the proposed method.

4.4 Asymmetric Context

In Sect. 3, the context which is aggregated by the function $\mathcal{F}$ has a size determined by the hyperparameter c and consists of the surrounding clips, both from past and future ones. In this experiment, we aim to investigate the effect of an asymmetric context, that is by using two values of c, c_p and c_f, and vary the amount of past and future information used. Figure 3 reports the mR@1 (Fig. 3 left) and the mR@5 (Fig. 3 right), i.e., the average value computed at IoU=0.3 and IoU=0.5 for R@1 and R@5, obtained on the validation set.

It can be seen that completely removing either of them (past or future clips) leads to generally worse solutions, e.g., up to 9.4% mR@1 and 16.6% mR@5 is obtained when either $c_p = 0$ or $c_f = 0$, whereas 9.7% and 17.1% are obtained when $c_p = c_f = 2$. In contrast, when reducing either of them, but keeping at least both one past and one future clip leads to less conclusive statements: for instance, it leads to up to 9.3% mR@1 and 16.5% mR@5 when, respectively, $c_p = 1, c_f = 1$ whereas with $c_p = c_f = 2$ it achieves 9.7% and 17.1%. However, compared to $c_p = c_f = 3$, a c_p of 2 leads to slightly better performance (e.g., 16.9% mR@5 compared to 16.7%). These results confirm that the model learns how to effectively make use of the contextual information from both preceding

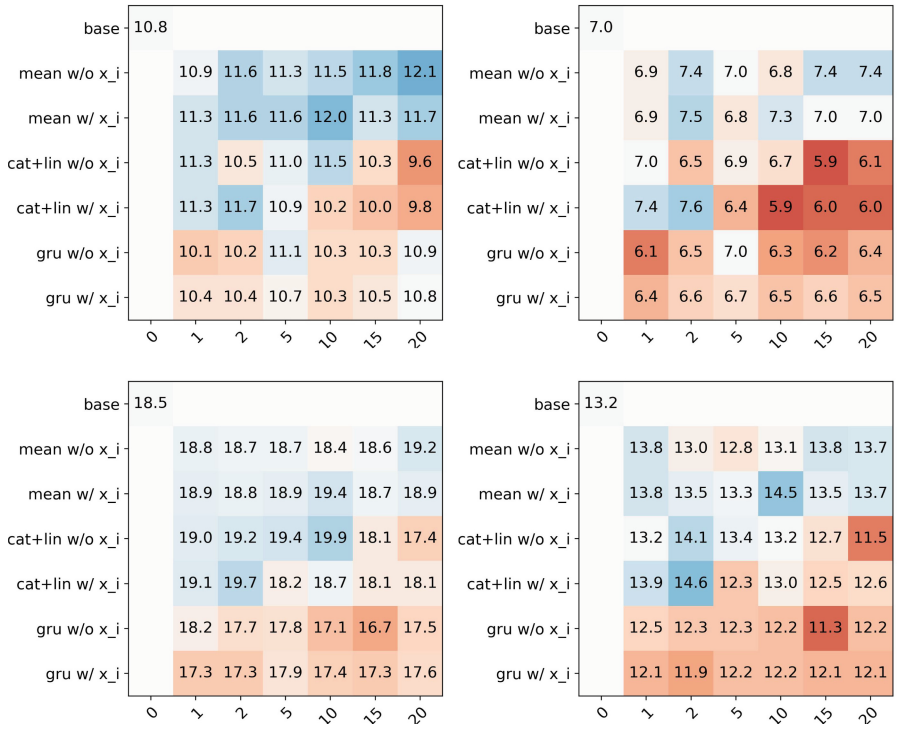

Fig. 2. Recall@1 (**top**) and Recall@5 (**bottom**) for two levels of IoU (**left**: IoU=0.3, **right**: IoU=0.5) computed on the validation set as the temporal model and the size of the context vary. VSLNet (*base*) is reported for reference. The other models are the Mean pooling, concatenation followed by a linear (*cat+lin*), and the GRU. Each model is tested both with (*w/ x_ i*) and without x_i in the context (*w/o x_ i*). Details in Sect. 4.3. Color scale from red (worse than *base*) to blue (better). (Color figure online)

and subsequent clips, although increasing the context too much might lead to worse generalization, due to an increased number of trainable parameters.

4.5 Ablation Study

In this ablation study, we show that the addition of the temporal awareness and the use of the circulant matrices are both important for the model. The results are reported in Table 1. The first line reports the performance achieved by the proposed method. In the second line, we remove the contextual awareness provided in Eq. 4, that is $\mathcal{F}([x_{i-c}, \ldots, x_{i+c}]) = x_i$ is used. The experimental results confirm that providing the model the information from the surrounding clips is useful, as in fact all the metrics decrease. Then, if the circulant matrices

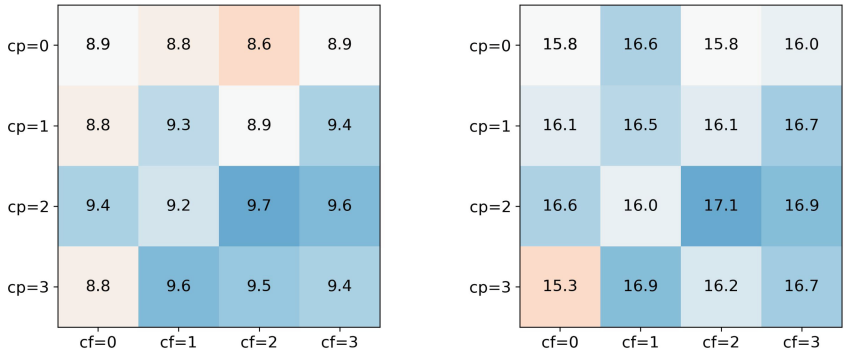

Fig. 3. Mean of Recall@1 (**left**) and Recall@5 (**right**) computed on the validation set as the amount of past, c_p, and future, c_f, clips in the context varies. VSLNet is reported for reference at $c_p = 0, c_f = 0$. Details in Sect. 4.4. Color scale from red (worse than the baseline) to blue (better).

Table 1. Ablation study reporting best results on validation set.

Method	R@1			R@5	
	IoU=0.3	IoU=0.5	Mean r@1	IoU=0.3	IoU=0.5
TCM	**11.72**	**7.64**	**9.68**	**19.70**	**14.56**
w/o T	11.33	7.12	9.23	19.02	13.99
w/o CM	10.76	7.02	8.89	18.48	13.16

are also removed, the VSLNet baseline is again obtained: as in the previous case, the metrics indicate the usefulness of the features obtained through the addition of the circulant matrix in Eq. 1.

4.6 Comparison with State-of-the-Art

As a final experimental result, we compare the performance achieved by the proposed method to that of several published works. VSLNet [33] is the baseline method and the results on the test set are taken from [15]. MSRA-AIM3 used a different set of pre-extracted visual features, made of both Swin Transformer and CLIP [22,24], which are first encoded by a Transformer-based Feature Encoder and then by a cross-modal encoder using multiple Transformer layers working on both visual and textual inputs [37]. EgoVLP is a pretraining strategy based on Frozen-in-Time [2] which used two separate Transformers to encode visual and textual inputs directly from the raw data, and pretrained them with a customized task based on multiple choice question answering and a loss function designed to specialize the selection of the samples used for the contrastive loss [18]. Finally, ReLER used a multi-scale cross-modal Transformer to model the complex interactions between video and text, two data augmentation techniques to reduce overfitting issues, and additional loss functions [21].

Table 2. Results obtained on the test set.

Method	R@1			R@5		
	IoU=0.3	IoU=0.5	Mean r@1	IoU=0.3	IoU=0.5	Mean r@5
VSLNet [33]	5.45	3.12	4.28	10.74	6.63	8.68
MSRA-AIM3 [37]	10.34	6.09	8.22	**18.01**	10.71	14.36
EgoVLP [18]	10.46	6.24	8.35	16.76	11.29	14.02
ReLER [21]	**12.89**	**8.14**	**10.51**	15.41	9.94	12.67
TCM (**ours**)	11.64	6.84	9.24	17.43	**11.39**	**14.41**

Table 2 presents the comparison. It can be seen that by using the proposed technique, we achieve better performance than VSLNet, MSRA-AIM3, and EgoVLP. Compared to ReLER we achieve better Recall@5, meaning that the top 5 candidates predicted by our model are generally more precise than those predicted by ReLER; on the other hand, ReLER achieves better Recall@1, meaning that their first candidate is generally more precise than ours. This may be due to the additional samples "generated" by the data augmentation techniques.

There are also some very recent works which tackle this challenging task [25], but the results are difficult to compare since an updated version of the annotations for the dataset, almost doubling the total amount of annotations (around 27k queries in place of the 17k that we used), has been released. In future work, the new version of the dataset will be considered.

5 Conclusions

Machines often lack the ability to remember past events involving other people, the interactions both with objects and other people, and the locations, i.e., episodic memory, which, on the other hand, is a fundamental aspect of human cognition. Considering the important impact on healthcare, societal assistance, and education, a novel problem has been recently proposed, called Episodic Memory via Natural Language Queries [15]. Previous works from the literature address this problem by leveraging methods inspired from other research domains (temporal activity localization and video moment retrieval), although in most of them the usage of contextual information from adjacent clips is limited to early layers of the network and often unexplored. In this paper, we address this limitation by proposing the Time-aware Circulant Matrices technique, which injects contextual awareness and intra-modal reasoning in later parts of the model. The experimental results motivate the design choices of the proposed method, present its robustness by means of an ablation study, and also confirm its effectiveness by comparing it to other state of the art methods from the literature. As a future work, we aim at further exploring its effectiveness on other datasets from the video-language grounding literature [11,26], and extend its intra-modal reasoning to enable the understanding of cross-modal interactions.

Acknowledgements. This work was supported by the Department Strategic Plan (PSD) of the University of Udine-Interdepartmental Project on Artificial Intelligence (2020-25). This work was partially funded by the Horizon EU project MUSAE (No. 01070421), 2021-SGR-01094 (AGAUR), Icrea Academia'2022 (Generalitat de Catalunya), Robo STEAM (2022-1-BG01-KA220-VET-000089434, Erasmus+ EU), DeepSense (ACE053 /22/000029, ACCIÓ), DeepFoodVol (AEI-MICINN, PDC2022-133642-I00), and CERCA Programme/Generalitat de Catalunya.

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: International Conference on Learning Representations (2015)
2. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: a joint video and image encoder for end-to-end retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1728–1738 (2021)
3. Buch, S., Escorcia, V., Shen, C., Ghanem, B., Carlos Niebles, J.: SST: single-stream temporal action proposals. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2911–2920 (2017)
4. Chen, G., et al.: InternVideo-Ego4D: a pack of champion solutions to Ego4D challenges. arXiv preprint: arXiv:2211.09529 (2022)
5. Cheng, F., Bertasius, G.: TallFormer: temporal action localization with a long-memory transformer. In: Avidan, S., Brostow, G., Cisse, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. Lecture Notes in Computer Science, vol. 13694, pp. 503–521. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19830-4_29
6. Cui, R., et al.: Video moment retrieval from text queries via single frame annotation. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1033–1043 (2022)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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, Volume 1 (Long and Short Papers), pp. 4171–4186 (2019)
8. Escorcia, V., Caba Heilbron, F., Niebles, J.C., Ghanem, B.: DAPs: deep action proposals for action understanding. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 768–784. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_47
9. Feichtenhofer, C., Fan, H., Malik, J., He, K.: SlowFast networks for video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6202–6211 (2019)
10. Gao, J., Sun, X., Xu, M., Zhou, X., Ghanem, B.: Relation-aware video reading comprehension for temporal language grounding. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 3978–3988 (2021)
11. Gao, J., Sun, C., Yang, Z., Nevatia, R.: TALL: temporal activity localization via language query. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5267–5275 (2017)
12. Ge, R., Gao, J., Chen, K., Nevatia, R.: MAC: mining activity concepts for language-based temporal localization. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 245–253. IEEE (2019)

13. Girdhar, R., Singh, M., Ravi, N., van der Maaten, L., Joulin, A., Misra, I.: Omnivore: a single model for many visual modalities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16102–16112 (2022)
14. Gong, G., Zheng, L., Mu, Y.: Scale matters: temporal scale aggregation network for precise action localization in untrimmed videos. In: 2020 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE (2020)
15. Grauman, K., et al.: Ego4D: around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18995–19012 (2022)
16. Hou, Z., et al.: An efficient coarse-to-fine alignment framework@ Ego4D natural language queries challenge 2022. arXiv preprint: arXiv:2211.08776 (2022)
17. Lei, J., Berg, T.L., Bansal, M.: Detecting moments and highlights in videos via natural language queries. In: Advances in Neural Information Processing Systems, vol. 34, pp. 11846–11858 (2021)
18. Lin, K.Q., et al.: Egocentric video-language pretraining@ Ego4D challenge 2022. arXiv preprint: arXiv:2207.01622 (2022)
19. Lin, K.Q., et al.: Egocentric video-language pretraining. In: Advances in Neural Information Processing Systems, vol. 35, pp. 7575-7586 (2022)
20. Lin, T., Zhao, X., Shou, Z.: Single shot temporal action detection. In: Proceedings of the 25th ACM International Conference on Multimedia, pp. 988–996 (2017)
21. Liu, N., Wang, X., Li, X., Yang, Y., Zhuang, Y.: Reler@ zju-alibaba submission to the Ego4D natural language queries challenge 2022. arXiv preprint: arXiv:2207.00383 (2022)
22. Liu, Z., et al.: Video Swin transformer. arXiv preprint: arXiv:2106.13230 (2021)
23. Mo, S., Mu, F., Li, Y.: A simple transformer-based model for Ego4D natural language queries challenge. arXiv preprint: arXiv:2211.08704 (2022)
24. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
25. Ramakrishnan, S.K., Al-Halah, Z., Grauman, K.: NaQ: Leveraging narrations as queries to supervise episodic memory. arXiv preprint: arXiv:2301.00746 (2023)
26. Soldan, M., et al.: MAD: a scalable dataset for language grounding in videos from movie audio descriptions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5026–5035 (2022)
27. Tulving, E.: Episodic and semantic memory: where should we go from here? Behav. Brain Sci. **9**(3), 573–577 (1986)
28. Wu, A., Han, Y.: Multi-modal circulant fusion for video-to-language and backward. In: IJCAI, vol. 3, p. 8 (2018)
29. Xiong, C., Zhong, V., Socher, R.: Dynamic Coattention networks for question answering. In: International Conference on Learning Representations (2016)
30. Yu, A.W., et al.: QaNet: combining local convolution with global self-attention for reading comprehension. In: International Conference on Learning Representations (2018)
31. Zeng, R., Xu, H., Huang, W., Chen, P., Tan, M., Gan, C.: Dense regression network for video grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10287–10296 (2020)
32. Zhang, C.L., Wu, J., Li, Y.: ActionFormer: localizing moments of actions with transformers. In: Avidan, S., Brostow, G., Cisse, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. Lecture Notes in Computer Science, vol. 13664, pp. 492–510. Springer, Cham (2022)

33. Zhang, H., Sun, A., Jing, W., Zhou, J.T.: Span-based localizing network for natural language video localization. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 6543–6554 (2020)
34. Zhang, S., Peng, H., Fu, J., Luo, J.: Learning 2D temporal adjacent networks for moment localization with natural language. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 12870–12877 (2020)
35. Zhang, Z., Han, X., Song, X., Yan, Y., Nie, L.: Multi-modal interaction graph convolutional network for temporal language localization in videos. IEEE Trans. Image Process. **30**, 8265–8277 (2021)
36. Zhao, P., Xie, L., Ju, C., Zhang, Y., Wang, Y., Tian, Q.: Bottom-up temporal action localization with mutual regularization. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12353, pp. 539–555. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58598-3_32
37. Zheng, S., Zhang, Q., Liu, B., Jin, Q., Fu, J.: Exploring anchor-based detection for ego4d natural language query. arXiv preprint: arXiv:2208.05375 (2022)