ITI-CERTH participation in AVS Task of TRECVID 2023

Damianos Galanopoulos, Vasileios Mezaris

Information Technologies Institute, Centre for Research and Technology Hellas,
6th Km. Charilaou - Thermi Road, 57001 Thermi-Thessaloniki, Greece
{dgalanop, bmezaris}@iti.gr

Abstract

This report presents an overview of the runs submitted to Ad-hoc Video Search (AVS) on behalf of the ITI-CERTH team. Our participation in the AVS task is based on a transformer-based extension of a cross-modal deep network architecture. We analyzed visual information at multiple levels of granularity using detected objects. During the retrieval stage, we employed a dual-softmax approach to adjust the calculated text-video similarities.

1 Introduction

In this report, the work carried out in the context of TRECVID 2023 by the ITI-CERTH[1] team in the area of video analysis, retrieval and understanding is presented. ITI-CERTH has participated in TRECVID [1] for many years as it is one of the most popular video understanding challenges. Especially, the ITI-CERTH team participated in the Search and Semantic Indexing (SIN) tasks under the research network COST292 (TRECVID 2006-2008) and the MESH and K-SPACE (TRECVID 2007-2008) EU-Funded research projects, correspondingly. From 2009 to 2015 [2, 3, 4, 5, 6, 7, 8] ITI-CERTH participated as a stand-alone organization in a significant number of tasks including but not limited to SIN, KIS, INS, and MED. In both 2016 [9] and 2017 [10], ITI-CERTH participated in the AVS, MED, INS and SED tasks. In 2018 [11], ITI-CERTH participated in the AVS, INS and ActEV; in 2019 [12], the participation was limited to the ActEV task. In 2020 [13] ITI-CERTH participated in the AVS, DSDI and ActEV tasks. Lastly, in 2021 [14] and 2022 [15], ITI-CERTH participated in the AVS and ActEV tasks. Considering the abovementioned submissions, this year we aim to evaluate improved algorithms and systems for the AVS task. The following sections will present the employed algorithms and the evaluation of the submitted runs.

2 Ad-hoc Video Search

The TRECVID 2023 [16] Ad-hoc Video Search (AVS) task aims to develop a system for retrieving a ranked list of 1000 video shots for each ad-hoc textual query, ranked from the most relevant to the least relevant shot for the query.

To address this task, we extend our $T \times V$ cross-modal network that combines different textual and visual features with additional modules for object detection and feature extraction and transformer-based aggregation, aiming to extract objects and calculate object-based visual features. Using all these, we develop multiple joint latent feature spaces. Then, we examine the performance of our different runs using this year's queries and queries from previous years.

[1] Information Technologies Institute - Centre for Research and Technology Hellas

2.1 Approach

In our AVS 2023 participation, we utilize the $T \times V + Objects$ network, an extension of the $T \times V$ cross-modal network presented in [17].

Our original $T \times V$ network consists of two key sub-networks, one for the textual and one for the visual stream. The textual sub-network inputs a free-text query s and vectorizes it into textual features. These features are used as input to a set of K textual encoders that encode the input sentence. Each of these encoders can be either a trainable network or simply an identity function forwarding its input. Similarly to the textual one, the visual sub-network inputs a video shot v consisting of a sequence of keyframes, and we use L pre-trained DNNs to extract the initial frame representations. To obtain video-shot level representations we follow the mean-pooling strategy.

The $T \times V + Objects$ network is an extension of the $T \times V$ model that utilizes visual information at more than one level of granularity by introducing an extra object-based video encoder using transformers. An overview of the network is illustrated in Fig. 1. An object detector extracts O objects from each frame, represented with a bounding box, an object class label and the associated degree of confidence. The ViT backbone network is applied to represent each object with a feature vector; these feature representations are sorted and stacked column-wise to form a matrix for the nth selected frame. Next, we proceed by employing transformer-based aggregation, where we calculate scaled dot-product attention scores to derive the self-attended outputs. Finally, we create the $(L+1)$th extra branch of visual frame representation by mean-pooling the self-attended outputs (illustrated as the object video encoder in Fig. 1).

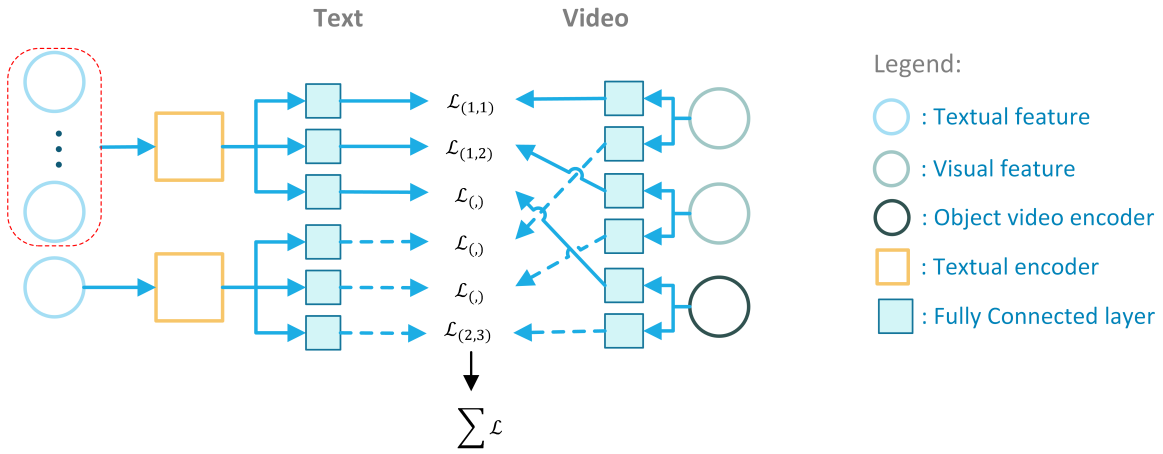

Figure 1: Illustration of the $T \times V + Objects$ network, where various textual and visual features are combined with the object video encoder creating different joint spaces.

Subsequently, we create all the possible textual encodings-visual feature pairs and a joint embedding space is created for each pair, using to this end two fully connected layers. Thus, $K \times (L+1)$ different joint spaces are created. The objective of our network is to learn a similarity function $sim(s, v)$ that will consider every individual similarity in each joint latent space utilizing multi-loss-based training.

Following [17] and our 2022 [15] approach, we apply a query-video similarities revision approach based on a dual softmax operation. More specifically, at the retrieval stage, and for a given query s, we calculate the similarities with the videos from the evaluation dataset, resulting in a vector $\mathbf{y}(s) = [sim(s, v_1), sim(s, v_2), \ldots, sim(s, v_D)]^T$, where D is the number of evaluation videos. To revise these similarities, we utilize a set of background textual queries (queries that are individual from the examined one) and calculate their similarities with the available videos within the dataset, resulting in a similarity matrix $\mathbf{X} \in \mathcal{R}^{C \times D}$, where C is the number of background queries. A matrix $\mathbf{Z}(s) = concat(\mathbf{y}(s); \mathbf{X})$ is constructed, and a dual softmax operation revises the similarities as follows:

$$\mathbf{Z}^*(s) = Softmax(\mathbf{Z}(s),\ dim = 0) \odot Softmax(\mathbf{Z}(s),\ dim = 1)$$

where $\odot$ denotes the Hadamard product.

2.2 Submission

Our network is trained using a combination of four large-scale video captioning datasets: MSR-VTTT [18], TGIF [19], ActivityNet [20] and Vatex [21]. The V3C2 [22] dataset is utilized to evaluate the networks' performance. Moreover, we examine the performance of our runs on the V3C1-V3C2 datasets for the queries of years 2019-2020-2021-2022. The evaluation measure we use is the mean extended inferred average precision (MxinfAP). As initial textual features, we utilize four models: i) Bag-of-Words (bow), ii) Word2Vec model [23] iii) Bert [24] and iv) Clip model [25]. Also, we utilize two textual encoders that input the textual features and encode the text further; i) the textual sub-network (ATT) presented in [26] and ii) a Clip encoder that feedforwards the corresponding features through an identity layer. As video feature extractors, we use three trained networks: i) a ResNet-152 [27] network, trained on the ImageNet-11k dataset, ii) a ResNeXt-101 network, pre-trained by weakly supervised learning on web images followed and fine-tuned on ImageNet [28], and iii) the ViT-B/32 Clip model [25]. Moreover, we utilize the object detector and the feature extractor modules of the ViGAT network [29] to obtain the initial object-based visual representations.

Similarly to previous works [30] [26], [17] where the combination of multiple models leads to improved performance, we utilize different model configurations to train multiple models using three learning rates ($10^{-4}, 5 * 10^{-5}, 10^{-5}$) and two optimizers (i.e., Adam and RMSprop).

This year, we submitted four runs for the AVS 2023 main task and four additional runs for the AVS progress subtask. Overall, we evaluate our methods on 40 ad-hoc queries (20 from the main task and 20 from the progress subtask). The submitted runs are briefly described below:

- ITI_CERTH.23_run_1: The $T \times V + Objects$ model, using two textual encoders and three visual features. For the extra object-based branch, the transformer-based aggregation is used for developing the frame-level representation. Late fusion of six different trained models is derived from six different model configurations. Finally, query-video shot similarities revision through the dual softmax operation using all AVS 2022 queries as background queries.
- ITI_CERTH.23_run_2: Similar to run 1, but with mean-pooling aggregation of the detected objects features in the extra object-based branch, instead of the transformer-based aggregation mechanism of run 1.
- ITI_CERTH.23_run_3: The $T \times V$ model, without the extra object branch, using two textual encoders and three visual features. Late fusion of six different trained models derived from six different model configurations. And finally, query-video shot similarities revision through the dual softmax operation using all AVS 2022 queries as background queries.
- ITI_CERTH.23_run_4: Late fusion for combining the above three runs.

2.3 Experimental Results

Table 1 summarizes the evaluation results of our runs for the 2023 main AVS task, while Table 2 displays the results of these runs for the 2019-2022 main AVS queries. By examining the performance of our runs on this year's main queries in Table 1, we conclude that the original $T \times V$ model outperforms its extension ($T \times V + Objects$). Moreover, there is no significant difference in performance between run 1 and run 2. In the AVS 2023 queries, the object features aggregation choice does not affect the overall performance. When it comes to the Progress queries, run 3 performed the best. However, the performance gap between run 3 and run 1 is even less significant than in the Main task.

In addition to the official AVS 2023 results, we also examine the performance of our 2023 methods on previous years' AVS queries in Table 2. Here, we notice a different behavior regarding runs' performance since run 1 ($T \times V + Objects$ model with transformer aggregator) is the best-performing run in 2019 and 2021 and close to the best-performing one for the rest of the years.

Figure 2 illustrates the performance of all submitted runs in the AVS 2023 competition. Our runs achieved 19th, 22nd, 23rd, and 24th place across all submitted runs and 6th place among all participating teams; it should be noted that this year, the differences in absolute MXinfAP scores among the various participating teams are, in most cases, relatively low.

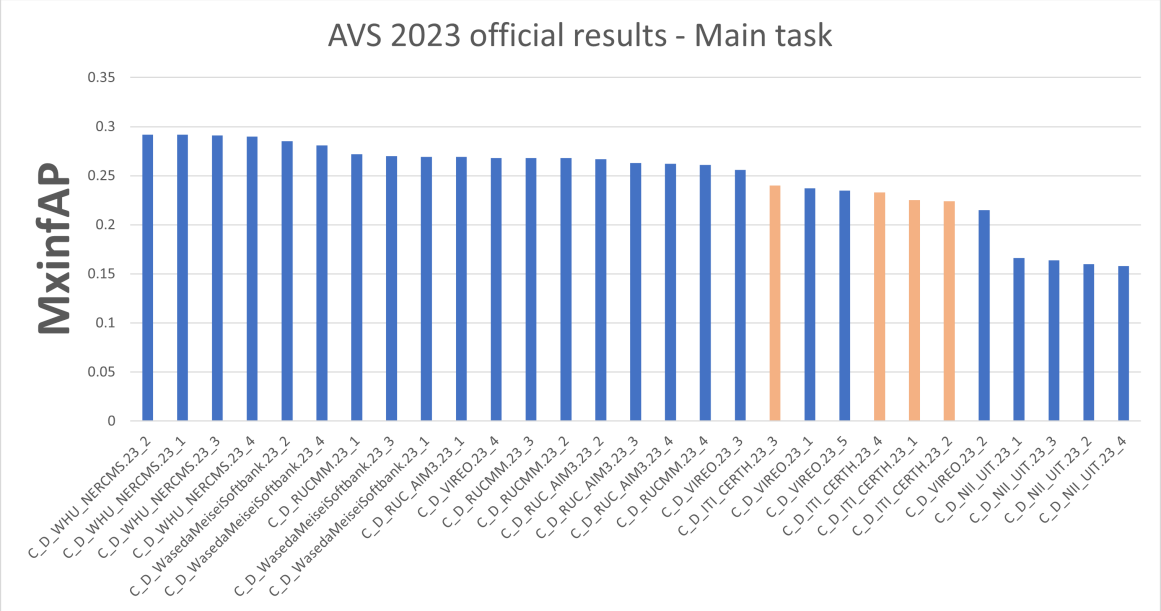

Figure 2: AVS 2023 ranking list of all submitted runs regarding the main task in MXinfAP terms. Orange bars indicate our submitted runs.

Table 1: Mean Extended Inferred Average Precision (MXinfAP) for all submitted runs for the fully-automatic AVS task.

Run id:	Main	Progress
ITI_CERTH.23 run_1	0.225	0.211
ITI_CERTH.23 run_2	0.224	0.202
ITI_CERTH.23 run_3	**0.240**	**0.216**
ITI_CERTH.23 run_4	0.233	0.213

Table 2: Mean Extended Inferred Average Precision (MXinfAP) for all submitted runs for the 2019, 2020, 2021 and 2022 fully-automatic AVS tasks.

Run id:	2019	2020	2021	2022
ITI_CERTH.23 run_1	**0.242**	0.343	**0.343**	0.204
ITI_CERTH.23 run_2	0.234	**0.344**	0.340	0.194
ITI_CERTH.23 run_3	0.233	0.339	0.326	**0.209**
ITI_CERTH.23 run_4	0.238	0.342	0.340	0.208

3 Conclusions

In this paper, the results of ITI-CERTH in the TRECVID 2023 challenge [16] are reported. ITI-CERTH this year participated by developing new techniques and algorithms for the AVS task. Specifically, we examined the performance of the $T \times V + Objects$ model, an extension of the $T \times V$ model, which utilizes visual information at more than one levels of granularity using object detection and transformer-based aggregation. In contrast to our preliminary evaluation results on the past AVS queries, where the extended model slightly outperformed its predecessor, the 2023 results concluded the opposite since the run based on the original $T \times V$ model achieved, by a small margin, the best result among all other ITI-CERTH runs.

4 Acknowledgements

This work was supported by the EU Horizon 2020 programme under grant agreement H2020-101021866 CRiTERIA.

References

[1] Alan F. Smeaton, Paul Over, and Wessel Kraaij. Evaluation campaigns and TRECVID. In *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[2] A. Moumtzidou, A. Dimou, and P. King et al. ITI-CERTH participation to TRECVID 2009 HLFE and Search. In *Proceedings of TRECVID 2009*, pages 665–668. 7th TRECVID Workshop, Gaithersburg, USA, November 2009.

[3] A. Moumtzidou, A. Dimou, and N. Gkalelis et al. ITI-CERTH participation to TRECVID 2010. In *Proceedings of TRECVID 2010*. 8th TRECVID Workshop, Gaithersburg, MD, USA, November 2010.

[4] A. Moumtzidou, P. Sidiropoulos, and S. Vrochidis et al. ITI-CERTH participation to TRECVID 2011. In *Proceedings of TRECVID 2011*. 9th TRECVID Workshop, Gaithersburg, MD, USA, December 2011.

[5] A. Moumtzidou, N. Gkalelis, and P. Sidiropoulos et al. ITI-CERTH participation to TRECVID 2012. In *Proceedings of TRECVID 2012*, Gaithersburg, MD, USA, 2012.

[6] F. Markatopoulou, A. Moumtzidou, and C. Tzelepis et al. ITI-CERTH participation to TRECVID 2013. In *Proceedings of TRECVID 2013*, Gaithersburg, MD, USA, 2013.

[7] N. Gkalelis, F. Markatopoulou, and A. Moumtzidou et al. ITI-CERTH participation to TRECVID 2014. In *Proceedings of TRECVID 2014*, Gaithersburg, MD, USA, 2014.

[8] F. Markatopoulou, A. Ioannidou, and C. Tzelepis et al. ITI-CERTH participation to TRECVID 2015. In *Proceedings of TRECVID 2015*, Gaithersburg, MD, USA, 2015.

[9] F. Markatopoulou, A. Moumtzidou, and D. Galanopoulos et al. ITI-CERTH participation in TRECVID 2016. In *Proceedings of TRECVID 2016*, Gaithersburg, MD, USA, 2016.

[10] F. Markatopoulou, A. Moumtzidou, and D. Galanopoulos et al. ITI-CERTH participation in TRECVID 2017. In *Proceedings of TRECVID 2018*. NIST, USA, 2017.

[11] Konstantinos Avgerinakis, Anastasia Moumtzidou, and Damianos Galanopoulos et al. ITI-CERTH participation in TRECVID 2018. In *Proceedings of TRECVID 2018*, 2018.

[12] Konstantinos Gkountakos, Konstantinos Ioannidis, Stefanos Vrochidis, and Ioannis Kompatsiaris. ITI-CERTH participation in TRECVID 2019. In *Proceedings of TRECVID 2019*, 2019.

[13] Konstantinos Gkountakos, Damianos Galanopoulos, and Marios Mpakratsas et al. ITI-CERTH participation in TRECVID 2020. In *Proceedings of TRECVID 2020*, Gaithersburg, MD, USA, 2020.

[14] Konstantinos Gkountakos, Damianos Galanopoulos, and Despoina Touska et al. ITI-CERTH participation in ActEV and AVS tracks of TRECVID 2021. In *Proceedings of TRECVID 2021*, Gaithersburg, MD, USA, 2021.

[15] Konstantinos Gkountakos, Damianos Galanopoulos, and Despoina Touska et al. ITI-CERTH participation in ActEV and AVS tracks of TRECVID 2022. In *Proceedings of TRECVID 2022*, Gaithersburg, MD, USA, 2022.

[16] George Awad, Keith Curtis, and Asad A. Butt et al. TRECVID 2023 - a series of evaluation tracks in video understanding. In *Proceedings of TRECVID 2023*. NIST, USA, 2023.

[17] Damianos Galanopoulos and Vasileios Mezaris. Are all combinations equal? combining textual and visual features with multiple space learning for text-based video retrieval. In *ECCVW*. Springer, October 2022.

[18] J. Xu, T. Mei, et al. MSR-VTT: A large video description dataset for bridging video and language. In *Proc. of IEEE CVPR 2016*, pages 5288–5296, 2016.

[19] Y. Li, Y. Song, L. Cao, J. Tetreault, et al. TGIF: A new dataset and benchmark on animated gif description. In *Proc. of IEEE CVPR 2016*, 2016.

[20] F. Caba Heilbron et al. ActivityNet: A large-scale video benchmark for human activity understanding. In *Proc. of IEEE CVPR 2015*, pages 961–970, 2015.

[21] X. Wang et al. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proc. of IEEE/CVF ICCV 2019*, pages 4581–4591, 2019.

[22] L. Rossetto, H. Schuldt, G. Awad, and A. A. Butt. V3C–a research video collection. In *Proc. of MMM 2019*, pages 349–360. Springer, 2019.

[23] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, Workshop Track Proceedings*, ICLR '13, 2013.

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *arXiv preprint arXiv:1810.04805*, 2018.

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, , et al. Learning transferable visual models from natural language supervision. In *Proc. of the 38th Int. Conf. on Machine Learning (ICML)*, 2021.

[26] D. Galanopoulos and V. Mezaris. Attention mechanisms, signal encodings and fusion strategies for improved ad-hoc video search with dual encoding networks. In *Proc. of ACM ICMR 2020*, 2020.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[28] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018.

[29] Nikolaos Gkalelis, Dimitrios Daskalakis, and Vasileios Mezaris. Vigat: Bottom-up event recognition and explanation in video using factorized graph attention network. *IEEE Access*, 10:108797–108816, 2022.

[30] D. Galanopoulos and V. Mezaris. Hard-negatives or Non-negatives? A hard-negative selection strategy for cross-modal retrieval using the improved marginal ranking loss. In *Proc. of IEEE/CVF ICCVW 2021*, 2021.