Elyadata TRECVID 2022 VTT Model

Haroun Elleuch, Ahmed Badri & Fethi Bougares

Abstract

This paper describes an overview of Elyadata's participation in the TRECVID [1] 2022 evaluation. We participated only in the Video to Text Description (VTT) sub-task. We experimented with various approaches using a combination of a Vision-Language Pre-trained framework and spatio-temporal Transformer architecture.

Keywords: TRECVid, Video Captioning, Video to Text Description.

Methods

Four runs were submitted for this task:

- **Run 1:** Image captioning. (BLIP)
- **Run 2:** Multiple frame sampling + frame captioning. (BLIP)
- **Run 3:** Spatio-Temporal video captioning. (TimeSBLIP)
- **Run 4:** Caption selection based on confidence score from the three other runs. (BLIP + TimeSBLIP)

Models

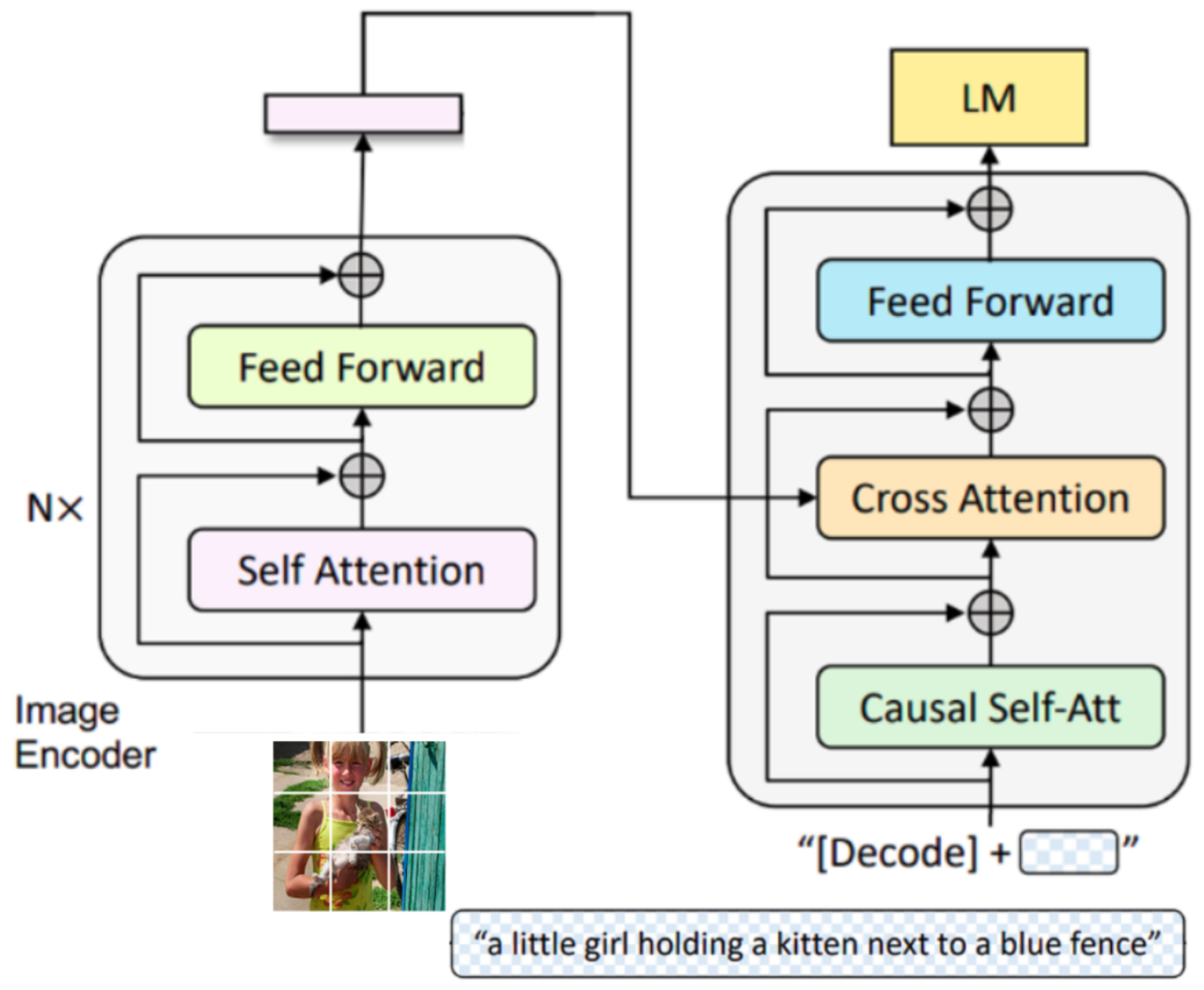

Figure 1: Architecture of the BLIP model for image caption generation when fine-tuning.

BLIP [2] (See figure 1) is a video-language pre-training framework (VLP). It stands for **B**ootstrapping **L**anguage-**I**mage **P**re-training. This framework pre-trains models by using a variant of knowledge distillation called Caption and Filtering (CapFilt), where a captioner model and a filter model distill their knowledge. This enabled the use of millions of web-sourced images with a relatively good caption quality for pre-training. Another particularity of BLIP is the use of an Image-Text Contrastive loss which permitted the feature alignment between the visual and textual features generated by the encoders.
The models can later be fine-tuned on a variety of downstream image-language or video-language tasks, such as image captioning.

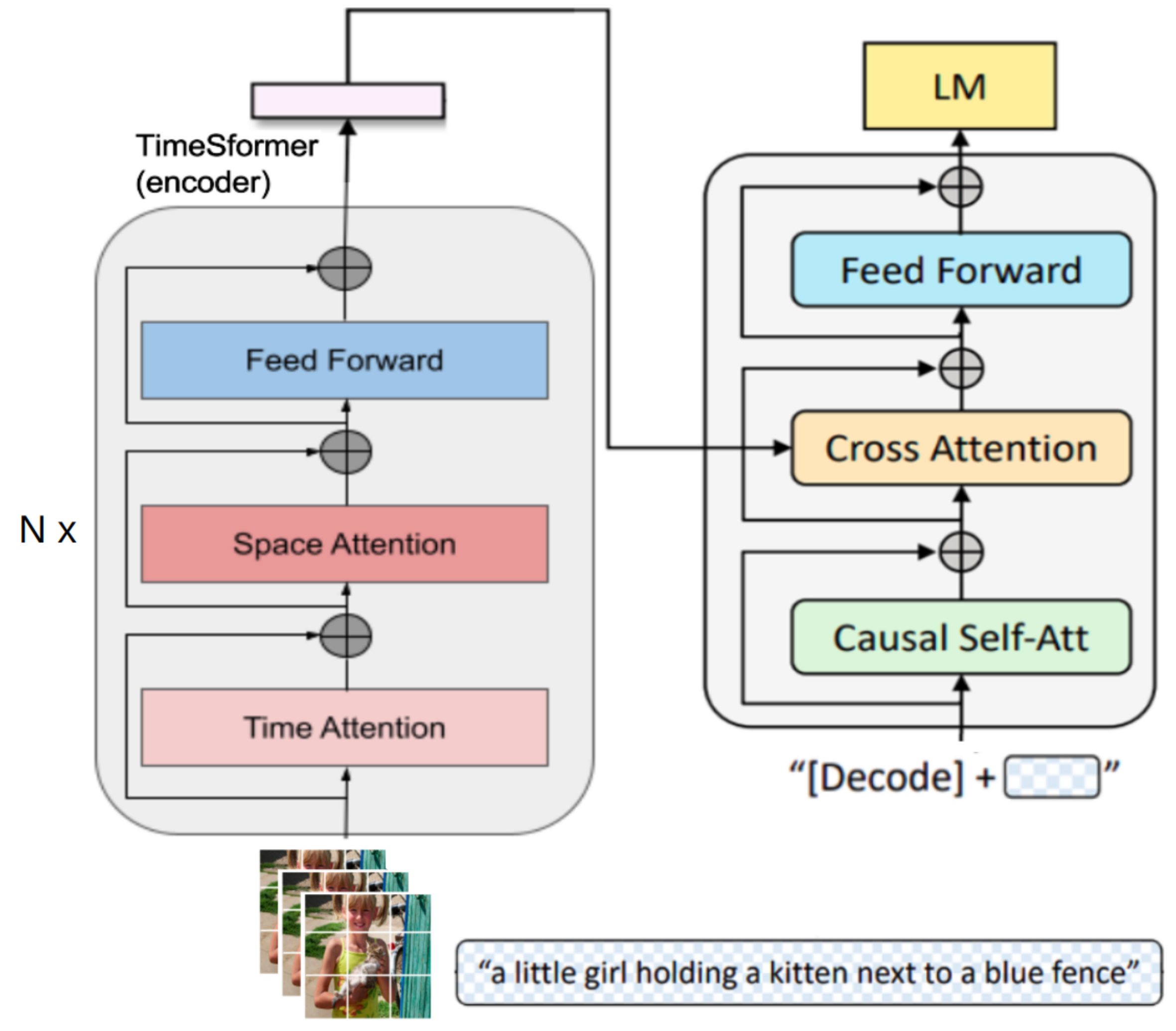

Figure 2: TimeSBLIP architecture.

TimeSBLIP 2 is the model proposed by the Elyadata team. It was obtained by combining a TimeSformer [3] spatio-temporal transformer encoder with the BLIP text decoder. The idea is to leverage the both the heavy pre-training of BLIP and the time dimension representation of the TimeSformer architecture.

In order to link the TimeSformer encoder with the BLIP decoder, a transformation must be performed on the encoder output in order to match the cross-attention dimension (see figures 1 and 2). Super-imposing and summing the spatial and temporal features was adopted as strategy for this submission.

Results

Run	BLEU@4	METEOR	CIDEr	CIDEr-D	SPICE	STS 1	STS 2	STS 3	STS 4	STS 5
1	**6.936**	**24.84**	**50.70**	**22.60**	**10.20**	**42.11**	**41.89**	**41.99**	**41.91**	**41.51**
2	1.298	17.83	10.30	4.50	4.30	23.57	24.01	23.02	23.78	23.86
3	1.403	16.92	24.30	7.60	6.20	35.70	34.13	33.51	33.61	36.27
4	3.414	19.41	23.40	10.50	6.40	30.83	30.34	29.94	30.34	30.73

Table 1: Submission results for the four runs on the TRECVid 2022 dataset.

The four submitted runs yielded the results reported in Table 1. The results corroborate what was observed on the validation scores during training: the best performing model is BLIP for image captioning, thanks to its pre-training and feature alignment process. Video captioning systems performed much worse.

Overall, the first system performed well, especially in the CIDEr-D and STS metrics, placing among the best performing submissions.

The TimeSBLIP model shows promise and given the reintroduction of a multimodal feature alignment mechanism, performance could improve. Other fields of improvement include the connection between the TimeSformer encoder and the text decoder: Instead of simply summing-up the spatio-temporal features, other strategies that permit their isolation could be tried, such as convolutions, average and max pooling. This work is left for the future.

The fourth run performed worse than the first. This shows that even though the image captioning system obtained the highest scores in all the evaluations, it is less confident about some of its captions than other models.

Conclusion

This report presents the submitted systems for the Video To Text description (VTT) task for the 2022 edition of TRECVid.

- All systems are BLIP-based.
- The image captioning variant performed best, whereas both video captioning models, although more confident in their captioning in some instances, performed far worse.
- These models were either a direct conversion of BLIP [2] for video captioning or a modification of the latter, consisting in the replacement of its ViT [4] encoder by a TimeSformer [3] module.

Contact Information

Web https://www.elyadata.com/
Email <name>.<lastname>@elyadata.com
Github /elyadata/trecvid-vtt-2022

References

[1] George Awad, Keith Curtis, Asad A. Butt, Jonathan Fiscus, Afzal Godil, Yooyoung Lee, Andrew Delgado, Jesse Zhang, Eliot Godard, Baptiste Chocot, Lukas Diduch, Jeffrey Liu, Yvette Graham, and Georges Quénot. An overview on the evaluated video retrieval tasks at trecvid 2022. In *Proceedings of TRECVID 2022*. NIST, USA, 2022.

[2] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. *CoRR*, abs/2201.12086, 2022.

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? 2021.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020.