JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Learning text-to-video retrieval from image captioning

Lucas Ventura [1,2] Cordelia Schmid [2] Gül Varol [1]

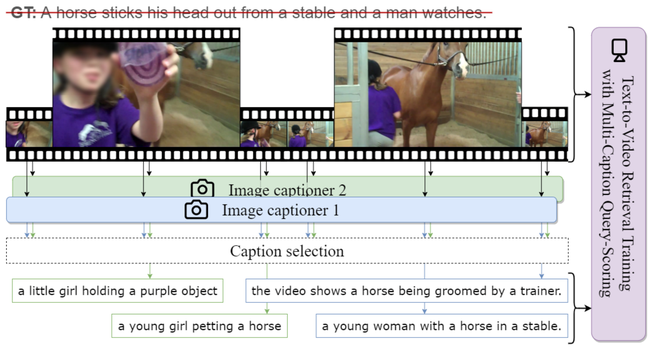
[1] IMAGINE, École des Ponts ParisTech, France [2] WILLOW, Inria, France

Inria École des Ponts ParisTech

INTRODUCTION

GOAL:

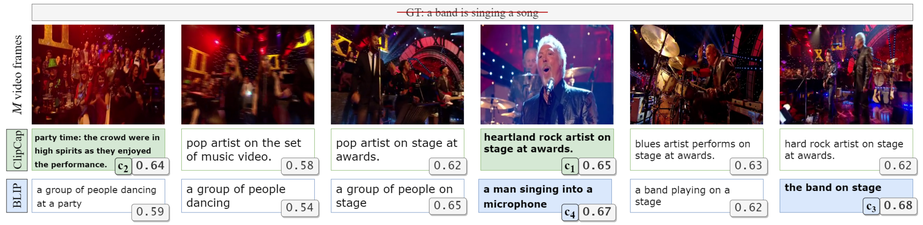
Text-to-video retrieval without access to manually-labeled videos, but allowing access to manually-labeled *images*.

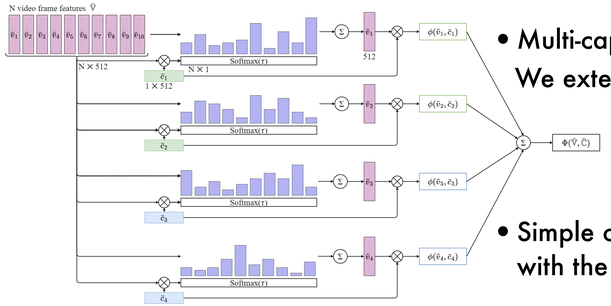


TLDR: We automatically label video frames with *image* captioners [1, 2] and select high-quality captions to train text-video retrieval models, pretrained on a text-*image* retrieval model, CLIP [3].

APPROACH

- Caption selection: We caption 10 video frames with both ClipCap [1] and BLIP [2], then compute CLIPScore [4] (gray box) to select Top K=2 captions.

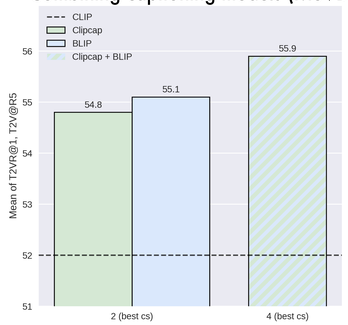




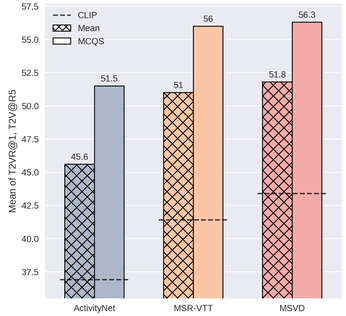
- Multi-caption query scoring (MCQS): We extend [5] to pool frame embeddings.
- Simple contrastive retrieval objective with the automatic video label.

EXPERIMENTS

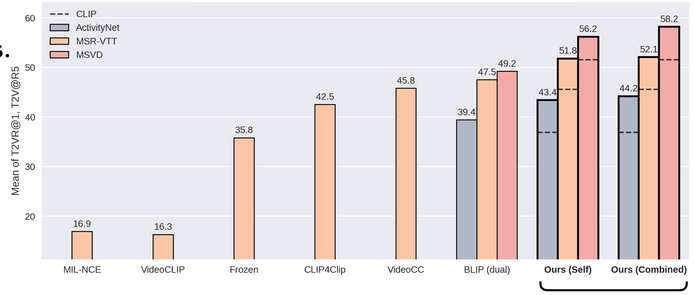
Combining captioning models (MSVD)



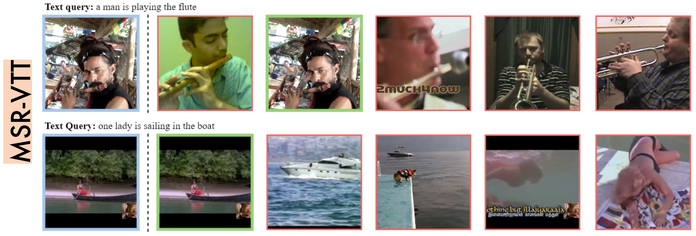
Mean vs MCQS

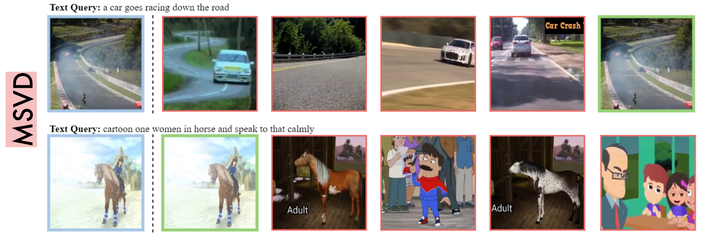


Training on the combination of datasets



DATASETS & QUALITATIVE RESULTS





REFERENCES

[1] ClipCap: Mokady et al. arXiv 2021.
[2] BLIP: Li et al. ICML 2022.
[3] CLIP: Radford et al. ICML 2021.
[4] CLIPScore: Hessel et al. EMNLP 2021.
[5] CLIP-Hitchhiker: Bain et al. arXiv 2022.

http://imagine.enpc.fr/~ventural/multicaps/