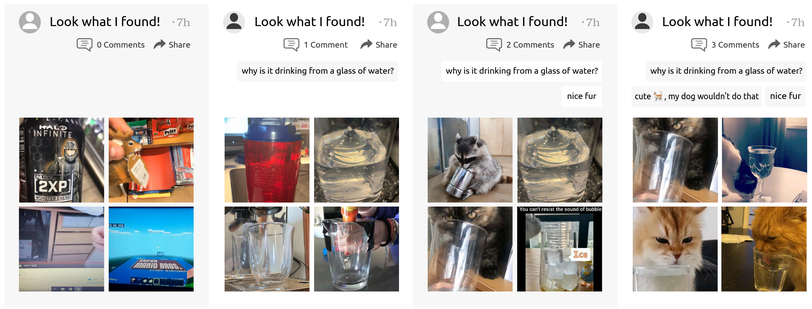
VTC: Improving Video-Text Retrieval with User Comments

Laura Hanu[1], James Thewlis[1], Yuki M. Asano[2], Christian Rupprecht[3]

[1]Unitary, [2]University of Amsterdam, [3]VGG, University of Oxford

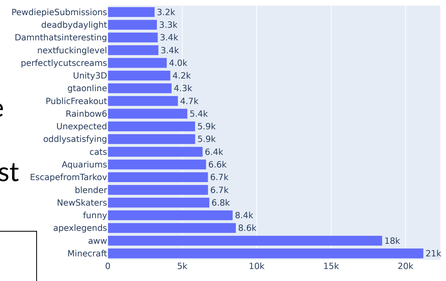
Video retrieval from title and comments

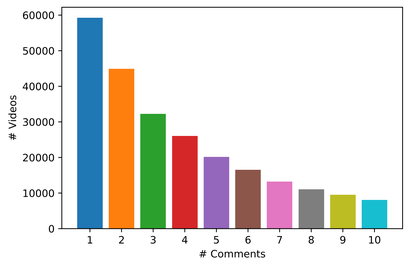


Introducing the VTC dataset

339k Video & Title pairs from Reddit, each with:

- the post title
- a high quality preview image
- the video url
- up to 500 comments per post





Deduplicated and **filtered** toxic text, NSFW content and videos containing faces

Why Comments?

1. Vision-language models work great for video-text retrieval, however, *titles* are often **not informative** enough e.g. “Look at this”
2. Comments can provide **valuable** information about different timepoints in the video
3. **Abundance** of comments on social media e.g. Reddit, Youtube, TikTok

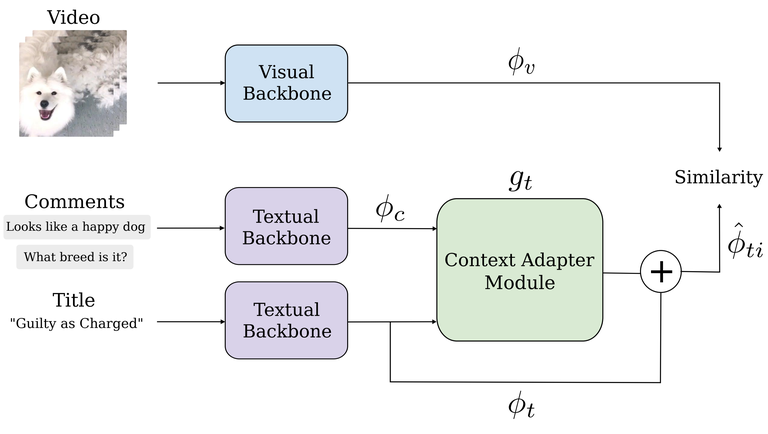
Acknowledgement

Project sponsored by Innovate UK Smart Grants.

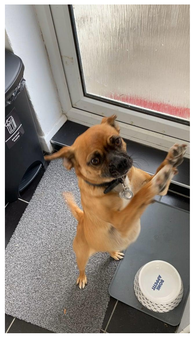
Our Contributions

1. We demonstrate the **value of comments** in video-text retrieval
2. We introduce **VTC**, a new **dataset** of videos, titles, and comments
3. We introduce a **context adapter module** that learns how to identify **relevant content**

How to incorporate comments?



Finding the most relevant comments



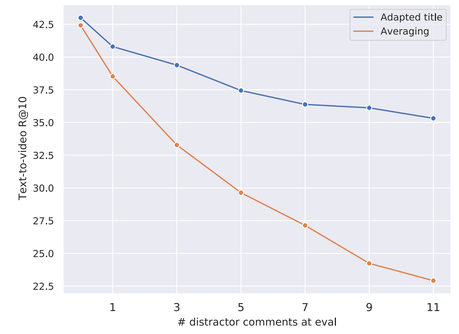
Lol when I saw the bowl I was like of course
Ive got the same breed with the same reaction
Which breed?
I’m so excited for her !!
I’m not sure she can eat it up there
what dog? all i see is a kangaroo
FOOD DANCE
That amazing!
Yup Pahaha
Lmao. She's DEFINITELY a dog where the "please eat slower. I
#give me the food master
Looks like it's bone appetit!
In my head I can hear the little Mario spring every time she
sproing
Totes going to edit that on, thanks for the inspo!

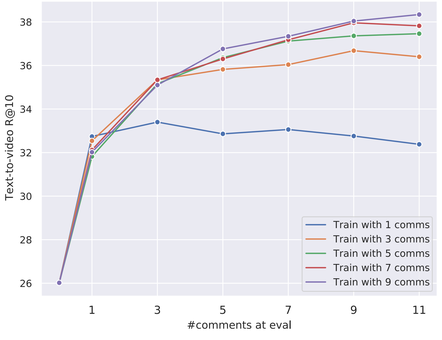
Title: She does this every time we feed her.

Comments and temporal attention help retrieval

		VTC		KineticsComms		LiveBotEn	
Inference	**#frames**	**VTR R@10**	**TVR R@10**	**VTR R@10**	**TVR R@10**	**VTR R@10**	**TVR R@10**
video	1	28.9	28.3	48.8	46.9	48.0	49.0
video+comms	1	40.8	41.0	61.1	59.2	64.0	64.0
video	8	28.9	27.6	56.9	55.8	70.0	72.0
video+comms	8	41.5	41.9	68.0	66.1	69.0	80.0

CAM is able to deal with **irrelevant** information much better than averaging baseline, showing that it has learned to **discard irrelevant** content.





We show that **increasing** the number of comments during **both** train and test time improves retrieval, although with diminishing returns.

Comparing different backbones

Backbone	**No comments**	**5 comments**	**20 comments**
FiT [1]	8.8	12.0	12.8
SLIP (ViT-B) [2]	9.3	10.2	11.6
CLIP (ResNet50) [3]	22.7	27.4	27.9
CLIP (ViT-B/32) [3]	25.3	32.3	34.1
CLIP (ViT-L/14) [3]	32.9	42.0	44.1

References

[1] M. Bain, A. Nagrani, G. Varol, and A. Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. ICCV, 2021

[2] N. Mu, A. Kirillov, D. Wagner, and S. Xie. Slip: Self-supervision meets language-image pre-training. arXiv:2112.12750, 2021.

[3] A. Radford, J. W. Kim, C. Hallacy, et al. Learning transferable visual models from natural language supervision. arXiv:2103.00020, 2021.