Cap4Video: What Can Auxiliary Captions Do for Text-Video Retrieval?

Supplementary Material

Wenhao Wu[1,2] Haipeng Luo[3] Bo Fang[3] Jingdong Wang[2] Wanli Ouyang[4,1]
[1]The University of Sydney [2]Baidu Inc.
[3]University of Chinese Academy of Sciences [4]Shanghai AI Laboratory
whwu.ucas@gmail.com

In this appendix, §A contains *details* of zero-shot video captioner. §B contains further *results*: computation efficiency (§B.1), more baselines (§B.2), and more visualizations (§B.3).

A. Caption Generation

To obtain auxiliary captions for a given video, we consider the following two approaches.

Crawling Titles. We extract the video website title by crawling the original links (such as YouTube ID) of each video and use it as the caption. For instance, for the MSR-VTT dataset, we crawl the title of the video website as the caption based on the original link provided by the dataset annotation. However, we found that 2555 out of the 10,000 videos in the dataset have invalid links, so we do not use the title as extra auxiliary information in these videos, and only perform video-query matching.

Video Captioning. We utilize the video extension [3] of ZeroCap [4] for zero-shot video captioning. In Cap4Video, the captioner can be replaced with other methods if desired.

ZeroCap employs GPT-2 [2], a transformer-based pre-trained language model, to predict the next word from an initial prompt, such as "Video shows". To integrate vision-related knowledge into the auto-regression process, the model is motivated to generate sentences that describe a given video using a calibrated CLIP loss $\mathcal{L}_{CLIP}$. An additional loss term, $\mathcal{L}_{CE}$, is employed to keep the next token distribution consistent with the original language model. Optimization occurs during auto-regression, and the process is repeated for each token. Simple arithmetic of visual cues in CLIP's embedding space can capture semantic relations. Although ZeroCap is effective in describing individual visual cues, it faces a challenge in generating coherent descriptions of multiple images. In contrast to the original ZeroCap approach, the video extension [3] optimizes pseudo-tokens through iterative sentence generation, with the goal of steering the overall sentence generation process towards a coherent description of the video, as depicted in Figure A.1.

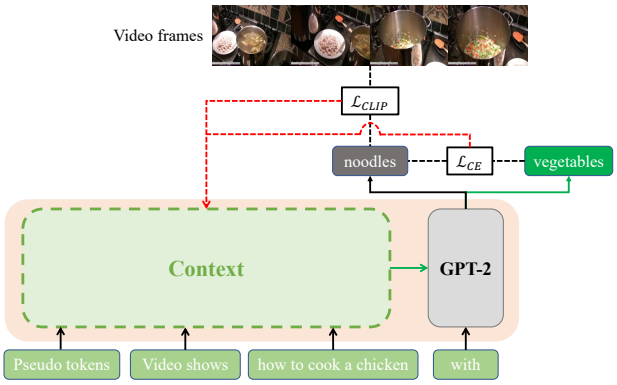


Figure A.1. Zero-shot video captioning [3] using CLIP and GPT.

In each generation step, the primary aim is to guide GPT-2 towards a desired visual direction. This guidance has two objectives: (i) aligning with the provided video, and (ii) maintaining language attributes. To achieve the first objective, CLIP [1] is utilized to assess the similarity of a token to a video and adjust the model's cache accordingly. For the second objective, the objective is regularized to resemble the original target output before modification. The solved optimization problem updates the context cache at each time point.

As a result, the captioner can generate captions for videos directly without any additional training. To prevent the generation of long and repetitive sentences, we set a token limit of 20 for each sentence. We also utilized frame sampling with a frame rate of 3 FPS and performed 30 iterations of generation to obtain 30 captions for each video.

For further information on the implementation, please see the paper [1] and official code [2].

[1]https://arxiv.org/pdf/2207.11100.pdf
[2]https://github.com/YoadTew/zero-shot-video-to-text

Query7669: three woman doing a fashion show to music.

Generated Caption:
Video of the fashion runway show in front by a woman wearing an orange dress and black shoes.

Query8572: a man grabs at snakes and throws them around the room.

Generated Caption:
Video showing a detainee digging out the snake from his cell in prison.

Query7464: the video shows gameplay of a car racing video game.

Generated Caption:
Video showing gameplay from the upcoming game of cars in virtual world, 'Toy Kart.

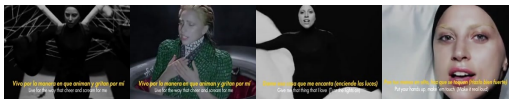
Query9625: its a cooking recipe show with chicken vegetables.

Generated Caption:
Video shows how to cook a chicken with rice noodles and vegetables.

Query8331: a man teaching students in class.

Generated Caption:
Video of teacher teaching students at a university in the Nordic countries, where they were filmed by an interviewer.

Query9827: lady gaga sings in a music video.



Generated Caption:
Video shows the lyrics of a song by rapper and singer, Gaga.

Figure A.2. Examples of auxiliary captions generated by the zero-shot video captioner on the MSRVTT 1K-A test set. These captions help to change the original wrong Top-1 prediction to the correct one.

	FLOPs	#Params	Throughput
Cap4Video	60.5G	96.8M	164.2 vid/s

Table A.1. Computation efficiency. "vid/s" represents the average number of videos per second. Model: ViT-B/32.

Video	+Auxiliary	Ensemble
42.8	+Caption	43.8
	+Zero-Shot CLIP	42.9

Table A.2. Ensemble baselines (ViT-B/32 w/ global matching).

Captioner use	Caption
Original CLIP	30.7
Fine-tuned CLIP	35.1

Table A.3. Caption baseline with different Captioner setting.

B. More Results

B.1. Computation Efficiency

In Table A.1, we show the computational cost and efficiency. We use single NVIDIA 3090 GPU and a batch size of 16 to measure the throughput.

B.2. Additional Baselines

To demonstrate the benefits of GPT-2 and language augmentation, we present the following baselines on the MSR-VTT 1k-A dataset: 1) Ensemble Baseline: we ensemble the zero-shot CLIP score and the finetuned video branch score. The results are shown in Table A.2. We can observe that the ensemble score of "Video+Zero-Shot CLIP" is lower than our Cap4Video (42.9% vs 43.8%), demonstrating the advantage of GPT-2. 2) Cap4Video using synthetic captions generated and filtered using the finetuned CLIP model. In our paper, we use the original CLIP and GPT-2 without any fine-tuning to perform zero-shot video captioning on any video. Here we study the effect of finetuned CLIP on the captioner in Table A.3. Although the captions generated by finetuned CLIP can bring further improvement, it reduces the method's flexibility.

Figure A.3. Examples of text-video retrieval results on the MSRVTT 1K-A test set. The left are the videos ranked by our Cap4Video, and the right are the results from the model without involving caption.

B.3. Qualitative Results

To further validate our motivation for using auxiliary caption for text-video retrieval, we present more visualizations of auxiliary captions in Figure A.2 and retrieval results in Figure A.3.

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 1

[2] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 1

[3] Yoad Tewel, Yoav Shalev, Roy Nadler, Idan Schwartz, and Lior Wolf. Zero-shot video captioning with evolving pseudo-tokens. *arXiv preprint arXiv:2207.11100*, 2022. 1

[4] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *CVPR*, pages 17918–17928, 2022. 1