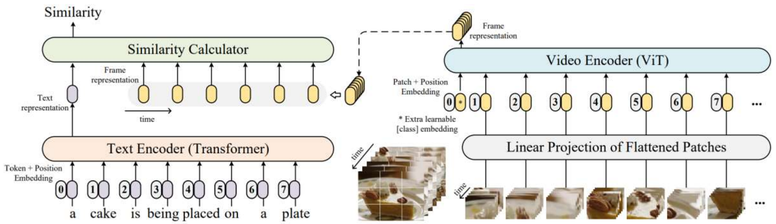
VoP: Text-Video Co-operative Prompt Tuning for Cross-Modal Retrieval

Siteng Huang, Biao Gong, Yulin Pan, Jianwen Jiang, Yiliang Lv, Yuyuan Li, Donglin Wang✉
Machine Intelligence Laboratory (MiLAB), Westlake University & Alibaba Group & Zhejiang University

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Introduction

- Leveraging the *pre-trained CLIP* for text-video cross-modal retrieval task recently popular.



[CLIP4Clip, Neurocomputing 2022]

However, the dominant full fine-tuning strategy brings…

- **risk of overfitting**: inevitably forgetting the useful knowledge acquired in the large-scale pretraining phase.
- **severe storage burdens**: maintaining an independent model weight for every dataset during deployment; infeasible due to the increasing model capacity.

For both effectiveness and efficiency, we continue the vein of prompt learning and propose …

- a strong baseline VoP that effectively adapts CLIP to text-video retrieval with only **0.1**% parameter storage.
- three video-specific prompts respectively conditioned on the frame position, frame context, and layer function, delivering an average R@1 improvement of up to 4.2% for VoP, and therefore exceed full fine-tuning by up to **1.4**% with much fewer trainable parameters.

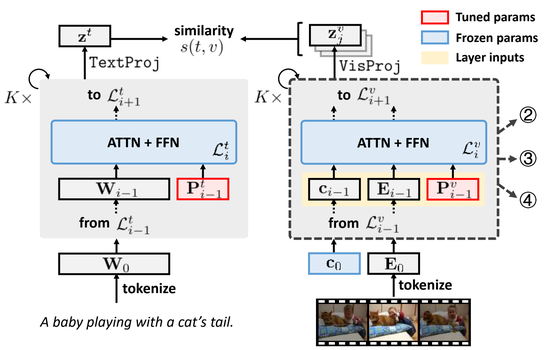
Our Proposed Framework

Baseline: ① VoP (Text-Video Co-operative Prompt Tuning)

Motivations:

- Learning prompts only for the text branch overlooks the potential of collaboratively tuning the visual encoder.
- Prompting the mere input layer has only a relatively indirect impact on the output embeddings.

Solution: Tuning the prompts introduced in **all** layers of **both** uni-modal encoders while keeping the rest of the model frozen.

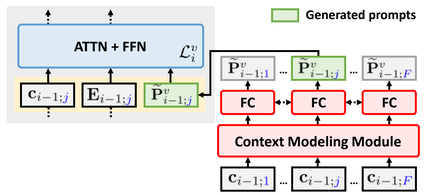


Equipping with Three Plug-and-Play Video Prompts

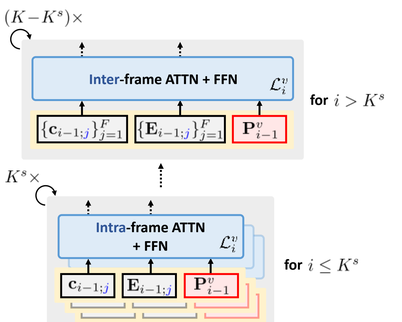
Motivation: Assisting VoP in utilizing rich *temporal* information.

② **VoP^{P}**: **position-specific** video prompts model the information shared between frames at the *same relative position*.

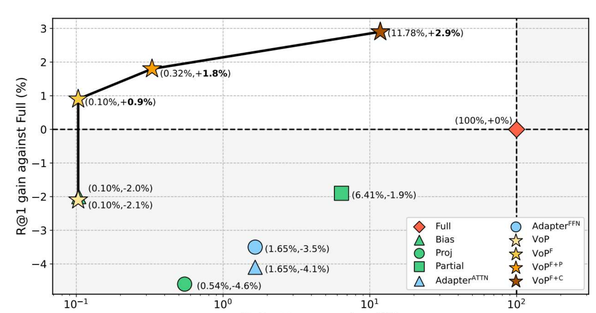
③ **VoP^{C}**: generated **context-specific** video prompts integrate injected *contextual* message from the frame sequence into the intra-frame modeling.



④ **VoP^{F}**: **function-specific** video prompts adaptively assist to learn *intra- or inter-frame affinities* by sensing the transformation of layer functions.



Main Results *CLIP ViT-B/32, MSR-VTT-9k*



More results ↓

arXiv

Github