ICCV23
PARIS

Zero-Shot and Few-Shot Video Question Answering with Multi-Modal Prompts

Deniz Engin

Yannis Avrithis

Inria

Introduction

<u>Motivation:</u> Inspired by large-scale vision-language model advancements in video tasks through multimodal datasets

<u>Challenges</u> on adapting pretrained models for video-language tasks on limited data

- Visual-language modality gap
- Overfitting & catastrophic forgetting

Recent Works

- Transformer-based mapping networks

  [Mokady et al., arXiv 2021]

- Parameter-efficient adaptation methods
  - Prompt learning [Liu et al., ACL 2022]
  - Adapters [Houlsby et al., ICML 2019]

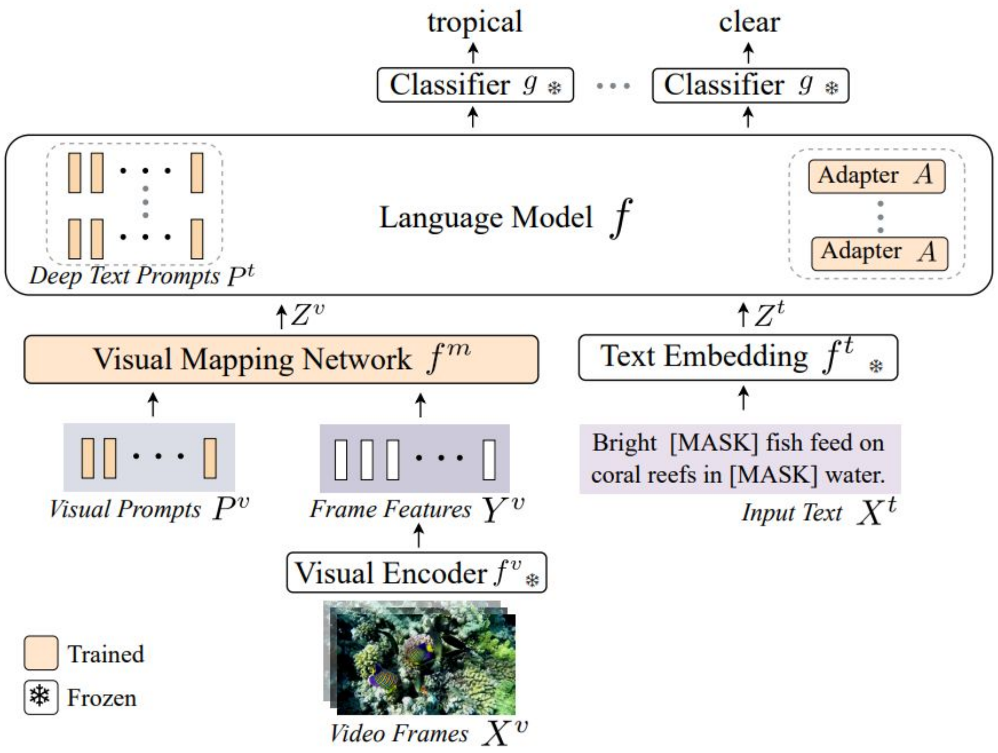
Our Approach

- Incorporation of visual inputs to a frozen language model using adapter layers [Yang et al., NeurIPS 2022]

- Introducing visual mapping network for summarizing video input while enabling temporal interaction

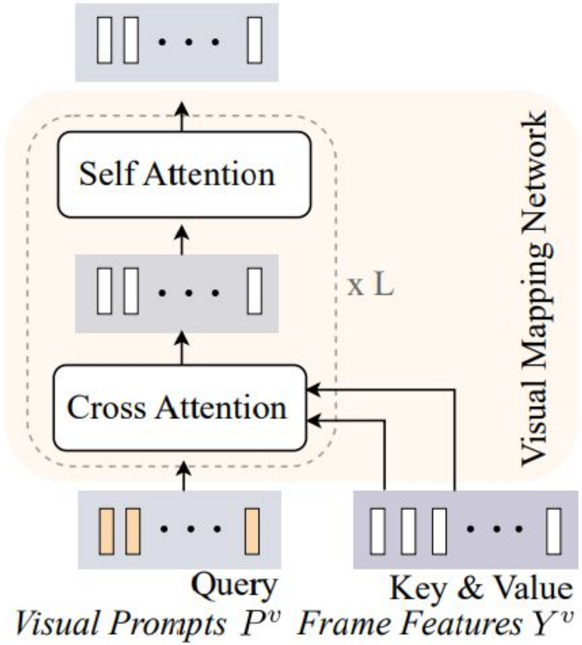
- Proposing multimodal prompt learning to reduce stored and tuned parameters during few-shot finetuning

ViTiS: VideoQA with Multi-Modal Prompts



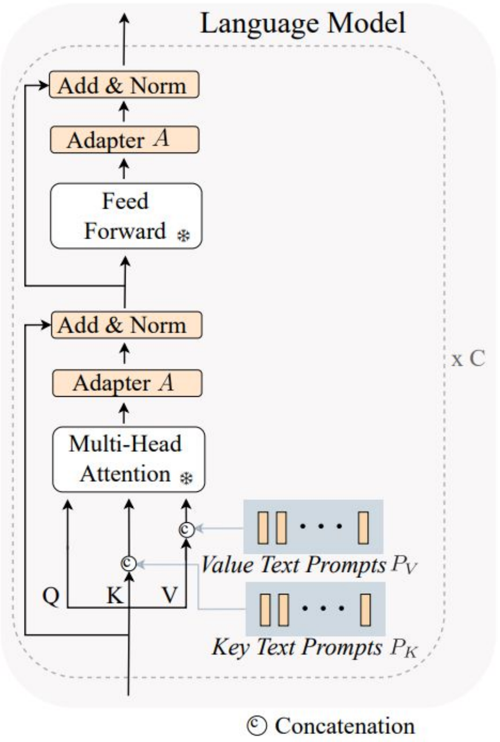
Visual Mapping Network (VPN)

- VPN aligns frame features with text embeddings
- Learnable visual prompts represent video after iteratively interact with frame features
- VPN designed, inspired by Perceiver [Jaegle et al., ICML 2021]



Language Model

- Learnable text prompts in the key and value of multi-head-attention in each layer of language model [Liu et al., ACL 2022]
- Adapter layer maps tokens to bottleneck dimension with residual connection [Houlsby et al., ICML 2019]
- Inserting adapter layers after each self-attention and feed-forward layer [Yang et al., NeurIPS 2022]



Zero-Shot VideoQA Results

Method	Subtitle	#Training Image	#Training Video	MSRVTT-QA	MSVD-QA	ANet-QA	TGIF-QA
CLIP* [Radford et al., ICML 2021]		400M	–	2.1	7.2	1.2	3.6
Reserve [Zellers et al., CVPR 2022]	✓	–	20M	5.8	–	–	–
Lavender [Li et al., CVPR 2023]		3M	2.5M	4.5	11.6	–	16.7
Flamingo [Alayrac et al., NeurIPS 2022]		2.3B	27M	17.4	35.6	–	–
FrozenBiLM [Yang et al., NeurIPS 2022]	✓	–	10M	16.7	33.8	**25.9**	41.9
ViTiS (Ours)	✓	–	2.5M	**18.1**	**36.1**	25.5	**45.5**

Pre-Training: All trainable parameters trained under MLM by keeping vision and language models frozen on WebVid2M [Bain et al., ICCV 2021]

Evaluation: Zero-shot top-1 accuracy on test sets, except TGIF-QA on the validation set

Few-Shot VideoQA Results

Method	Trained Modules	#Trained Params	MSRVTT-QA	MSVD-QA	ANet-QA	TGIF-QA
FrozenBiLM [Yang et al., NeurIPS 2022]	ATP	30M	36.0	46.5	33.2	55.1
ViTiS (Ours)	ATP	101M	36.5	47.6	33.1	55.7
ViTiS (Ours)	Prompts	0.75M	**36.9**	**47.8**	**34.2**	**56.2**

Few-Shot Training: Training using 1% of training data [Yang et al., NeurIPS 2022]

- **ATP:** Fine-tune all trainable parameters (8% of total)
- **Prompts:** Fine-tune only prompts (0.8% of trainable, 0.06% of total)

Evaluation: Few-shot top-1 accuracy on test sets, except TGIF-QA on the validation set

Contributions

- Introducing multimodal prompt learning for VideoQA for the first time
- Proposing a visual mapping network for VideoQA, mapping video input to the text embedding space while enabling temporal interaction
- Demonstrating strong performance on multiple VideoQA datasets in zero-shot and few-shot settings

ICCV23
PARIS

Zero-Shot and Few-Shot Video Question Answering with Multi-Modal Prompts

Deniz Engin

Yannis Avrithis

Project Page

https://engindeniz.github.io/vitis

Inria