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

Deniz Engin[1] Yannis Avrithis[2]

[1] Inria, Univ Rennes, CNRS, IRISA

[2] Institute of Advanced Research in Artificial Intelligence (IARAI)

Introduction

Motivation

- Inspired by **large-scale vision-language model advancements** in video tasks through multimodal datasets

Challenges in adapting pretrained models for video-language tasks on limited data:

- Visual-language **modality gap**
- **Overfitting** and **catastrophic forgetting**

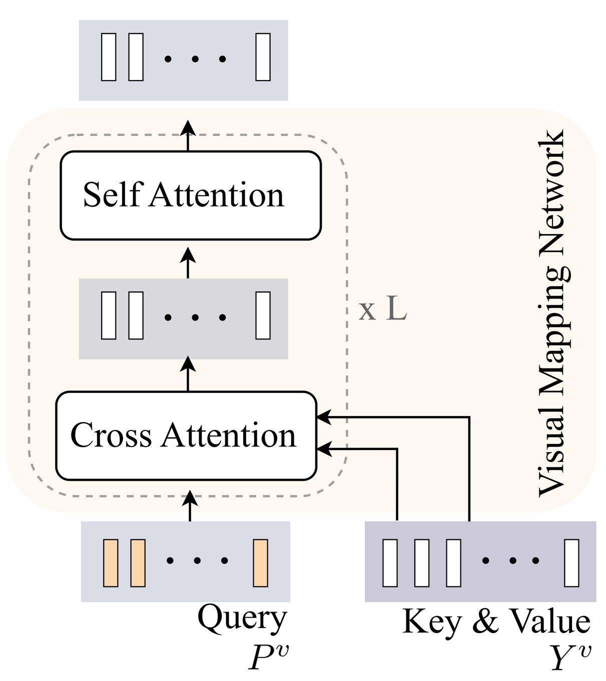
Recent Works

- Transformer-based mapping networks
- Parameter-efficient adaptation methods: prompt learning and adapters

Contributions

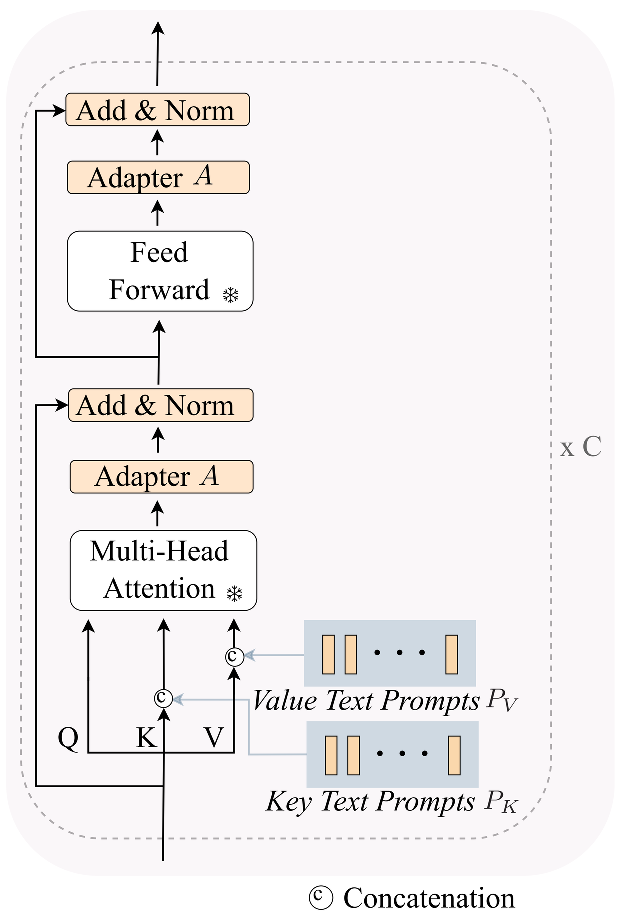
- Introducing **multimodal prompt learning** to VideoQA for the first time, reducing the number of stored and tuned parameters in few-shot setting
- Proposing **a visual mapping network** for VideoQA to summarize video input while **facilitating temporal interaction**
- Demonstrating **strong performance** across multiple VideoQA datasets in **zero-shot and few-shot** settings

VPN: Visual Mapping Network



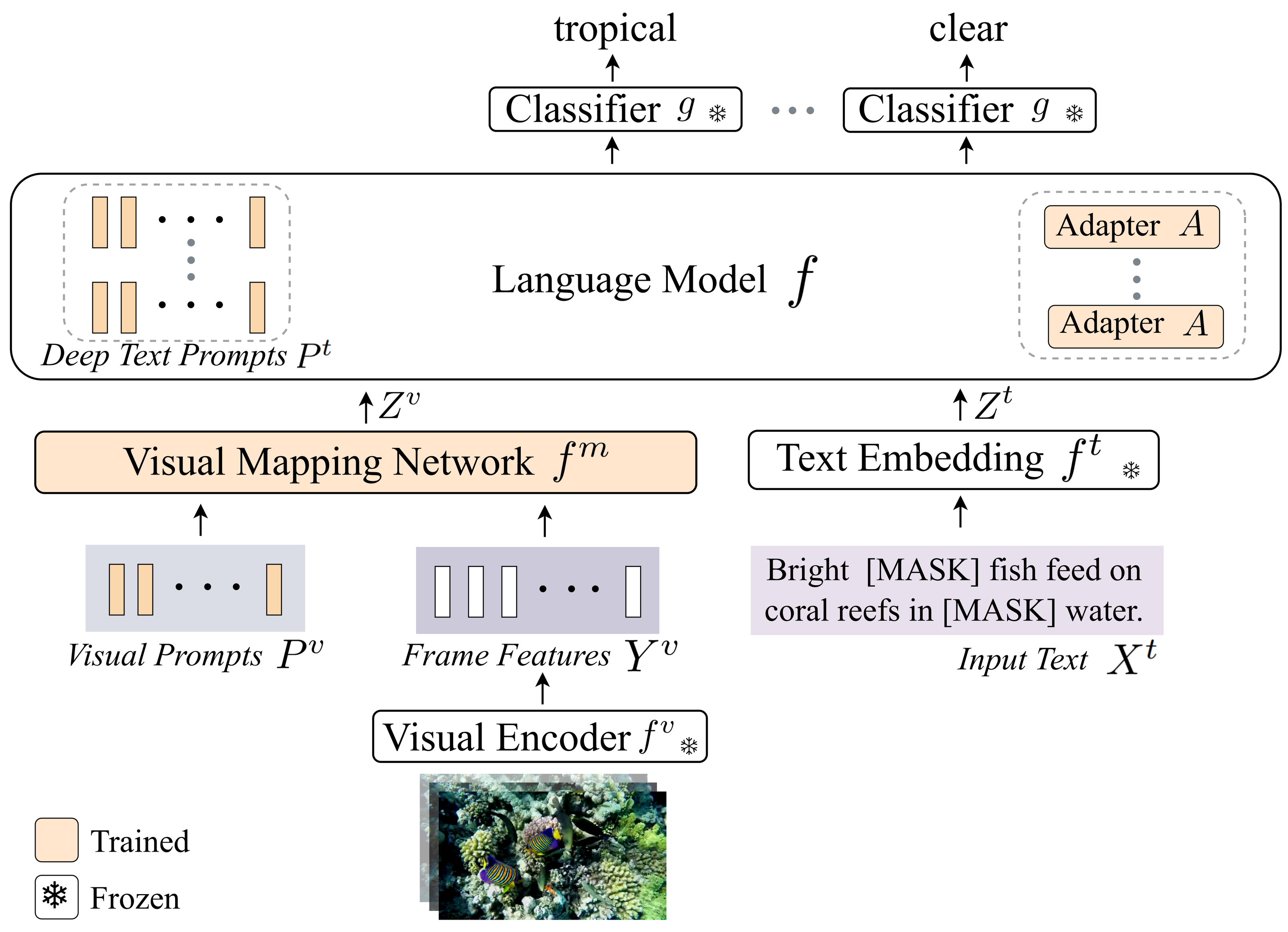
- VPN **aligns frame features** with **text embeddings**
- **Learnable visual prompts** represent video after iteratively interact with **frame features**

Language Model: Prompts and Adapters



- **Learnable text prompts** in the key and value of multi-head attention
- **Adapter layer** maps tokens to bottleneck dimension with residual connection

ViTiS: VideoQA with Multi-Modal Prompts



Zero-Shot VideoQA Results

- **Pre-Training:** All trainable parameters trained under MLM by keeping vision and language models frozen on WebVid2M

Method	Sub	#Training Img	#Training Vid	MSRVTT-QA	MSVD-QA	ANet-QA	TGIF-QA
CLIP [1]		400M	–	2.1	7.2	1.2	3.6
RESERVE [2]	✓	–	20M	5.8	–	–	–
LAVENDER [3]		3M	2.5M	4.5	11.6	–	16.7
Flamingo [4]		2.3B	27M	17.4	35.6	–	–
FrozenBiLM [1]	✓	–	10M	16.7	33.8	**25.9**	41.9
ViTiS (Ours)	✓	–	2.5M	**18.1**	**36.1**	25.5	**45.5**

Few-Shot VideoQA Results

- **Few-Shot Fine-tuning:** 1% of training data [1]
  - **ATP:** Fine-tune all trainable parameters (8% of total)
  - **Prompts:** Fine-tune only prompts (0.06% of total)

Method	Trained Modules	#Trained Params	MSRVTT-QA	MSVD-QA	ANet-QA	TGIF-QA
FrozenBiLM [1]	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**

References

1. A. Yang, *et al.*, Zero-shot video question answering via frozen bidirectional language models. In *NeurIPS*, 2022.
2. R. Zellers, *et al.*, MERLOT Reserve: Neural script knowledge through vision and language and sound. In *CVPR*, 2022.
3. L. Li, *et al.*, Lavender: Unifying video-language understanding as masked language modeling. In *CVPR*, 2023.
4. J. Alayrac, *et al.*, Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022.