Prompting Visual-Language Models for Efficient Video Understanding

Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie

Cooperative Medianet Innovation Center, Shanghai Jiao Tong University
Visual Geometry Group, University of Oxford

PRESENTER: LILANG LIN

2022/01/23

Outline

1 Authors

2 Background

3 Method

4 Experiments

5 Discussion

Outline

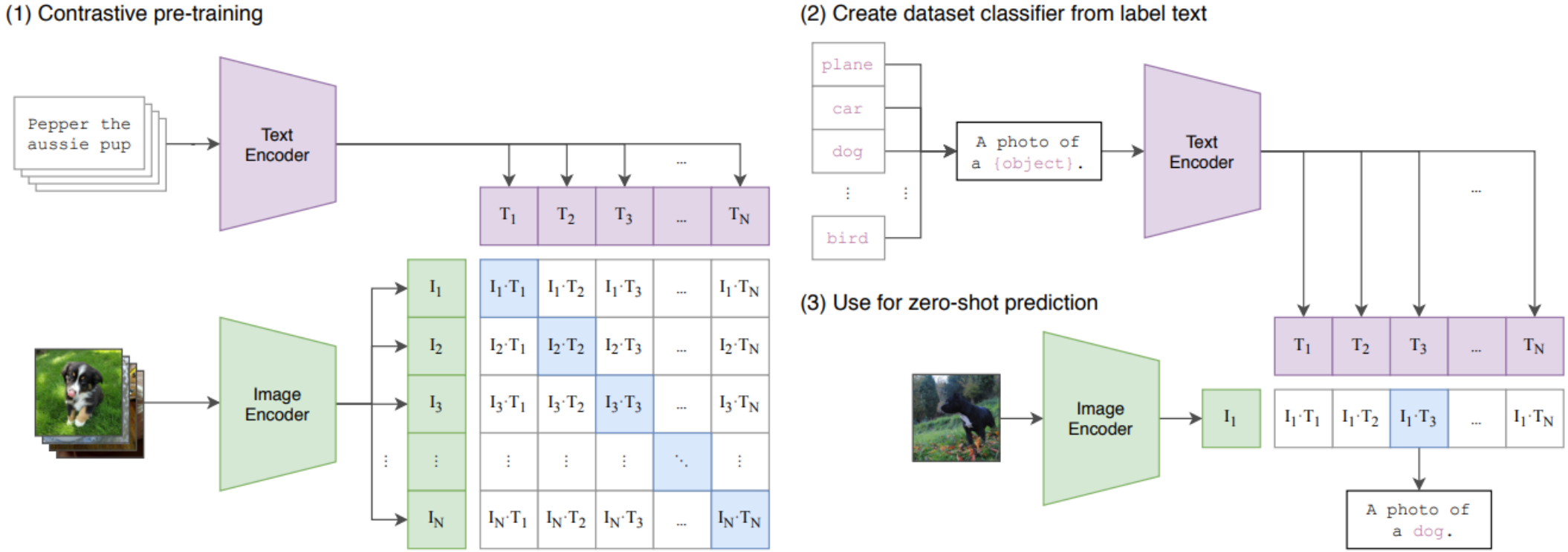
Background

Joint Visual-Textual Learning

CLIP (ICML 2021)

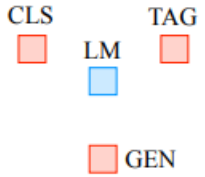
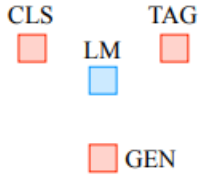
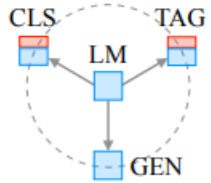
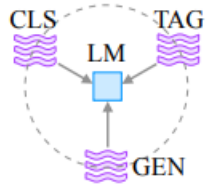
Learning Transferable Visual Models From Natural Language Supervision

Alec Radford [*1] Jong Wook Kim [*1] Chris Hallacy [1] Aditya Ramesh [1] Gabriel Goh [1] Sandhini Agarwal [1]
Girish Sastry [1] Amanda Askell [1] Pamela Mishkin [1] Jack Clark [1] Gretchen Krueger [1] Ilya Sutskever [1]



Background

Prompting

Paradigm	Engineering	Task Relation
a. Fully Supervised Learning (Non-Neural Network)	Features (e.g. word identity, part-of-speech, sentence length)	
b. Fully Supervised Learning (Neural Network)	Architecture (e.g. convolutional, recurrent, self-attentional)	
c. Pre-train, Fine-tune	Objective (e.g. masked language modeling, next sentence prediction)	
d. Pre-train, Prompt, Predict	Prompt (e.g. cloze, prefix)	

Background

Prompting

Name	Notation	Example	Description
Input	$\boldsymbol{x}$	I love this movie.	One or multiple texts
Output	$\boldsymbol{y}$	++ (very positive)	Output label or text
Prompting Function	$f_{\text{prompt}}(\boldsymbol{x})$	`[X]` Overall, it was a `[Z]` movie.	A function that converts the input into a specific form by inserting the input $\boldsymbol{x}$ and adding a slot `[Z]` where answer $\boldsymbol{z}$ may be filled later.
Prompt	$\boldsymbol{x}'$	I love this movie. Overall, it was a `[Z]` movie.	A text where `[X]` is instantiated by input $\boldsymbol{x}$ but answer slot `[Z]` is not.
Filled Prompt	$f_{\text{fill}}(\boldsymbol{x}', \boldsymbol{z})$	I love this movie. Overall, it was a bad movie.	A prompt where slot `[Z]` is filled with any answer.
Answered Prompt	$f_{\text{fill}}(\boldsymbol{x}', \boldsymbol{z}^*)$	I love this movie. Overall, it was a good movie.	A prompt where slot `[Z]` is filled with a true answer.
Answer	$\boldsymbol{z}$	"good", "fantastic", "boring"	A token, phrase, or sentence that fills `[Z]`

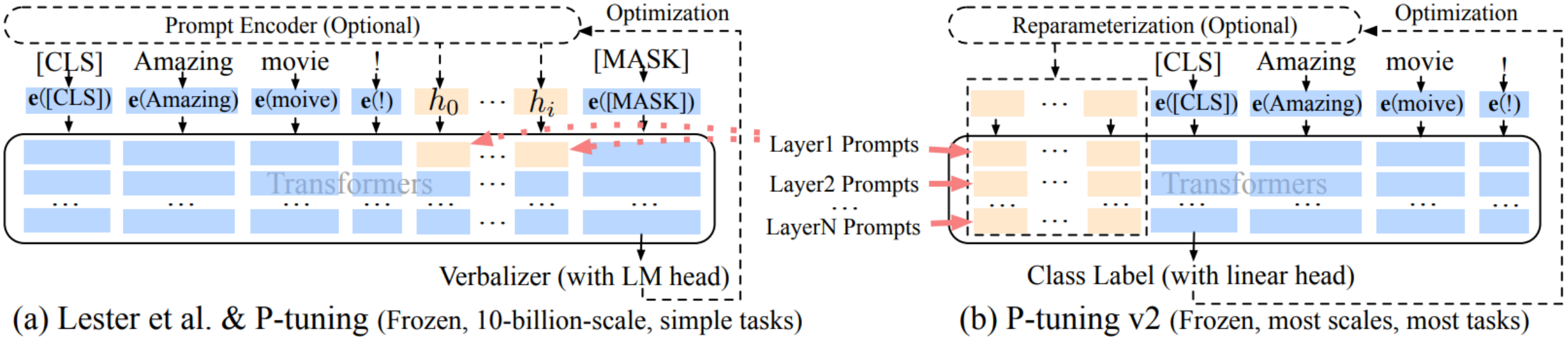
Background

Prompting

Type	Task	Input ([X])	Template	Answer ([Z])
Text CLS	Sentiment	I love this movie.	[X] The movie is [Z].	great fantastic ...
	Topics	He prompted the LM.	[X] The text is about [Z].	sports science ...
	Intention	What is taxi fare to Denver?	[X] The question is about [Z].	quantity city ...
Text-span CLS	Aspect Sentiment	Poor service but good food.	[X] What about service? [Z].	Bad Terrible ...
Text-pair CLS	NLI	[X1]: An old man with ... [X2]: A man walks ...	[X1]? [Z], [X2]	Yes No ...
Tagging	NER	[X1]: Mike went to Paris. [X2]: Paris	[X1][X2] is a [Z] entity.	organization location ...
Text Generation	Summarization	Las Vegas police ...	[X] TL;DR: [Z]	The victim ... A woman ... ...
	Translation	Je vous aime.	French: [X] English: [Z]	I love you. I fancy you. ...

Background

Prefix-Tuning V1 & V2 (ACL 2022)



(a) Lester et al. & P-tuning (Frozen, 10-billion-scale, simple tasks)

(b) P-tuning v2 (Frozen, most scales, most tasks)

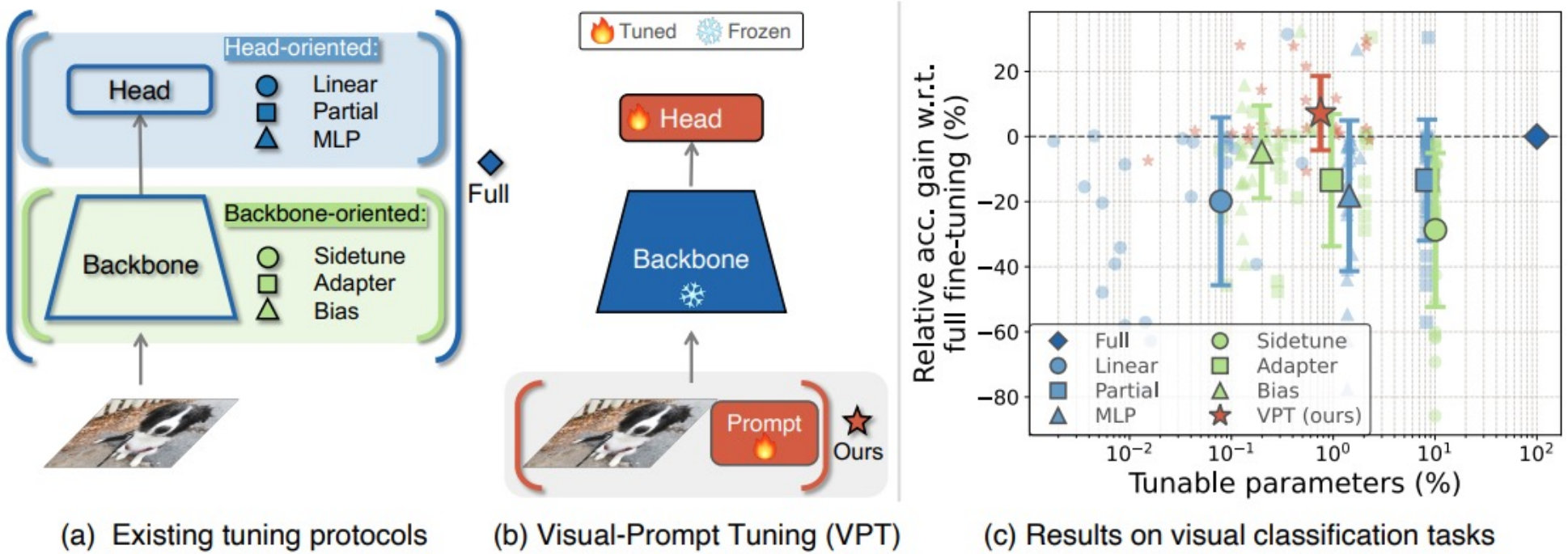
Background

VPT (ECCV 2022)

Visual Prompt Tuning

Menglin Jia*[1,2], Luming Tang*[1]
Bor-Chun Chen[2], Claire Cardie[1], Serge Belongie[3]
Bharath Hariharan[1], and Ser-Nam Lim[2]

[1]Cornell University [2]Meta AI [3]University of Copenhagen



(a) Existing tuning protocols (b) Visual-Prompt Tuning (VPT) (c) Results on visual classification tasks

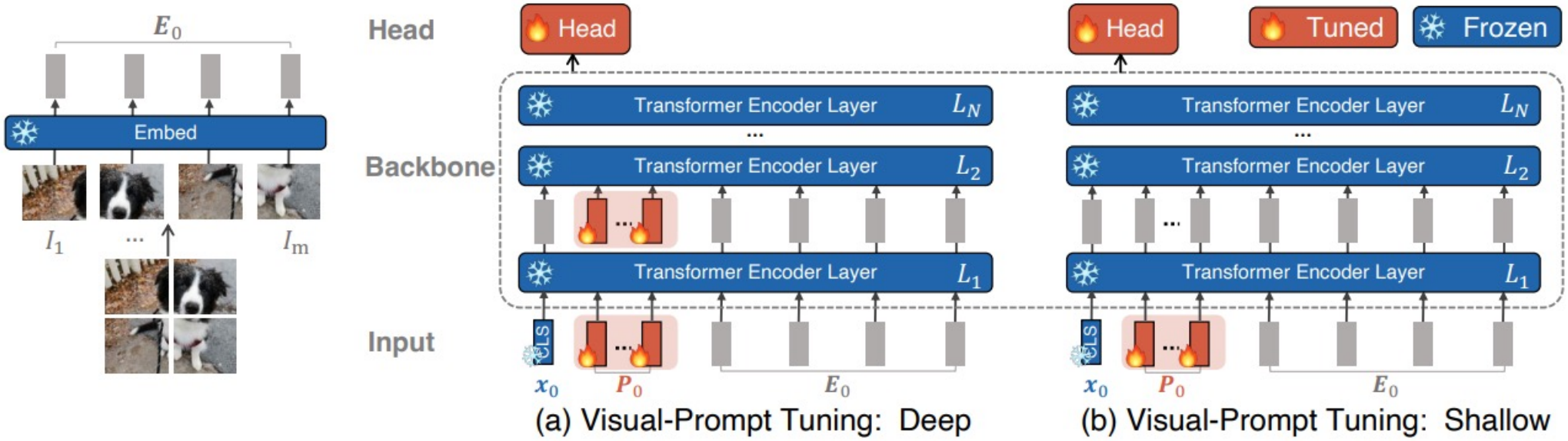
Background

VPT (ECCV 2022)

Visual Prompt Tuning

Menglin Jia*[1,2], Luming Tang*[1]
Bor-Chun Chen[2], Claire Cardie[1], Serge Belongie[3]
Bharath Hariharan[1], and Ser-Nam Lim[2]

[1]Cornell University [2]Meta AI [3]University of Copenhagen



Background

VPT (ECCV 2022)

Visual Prompt Tuning

Menglin Jia*[1,2], Luming Tang*[1]
Bor-Chun Chen[2], Claire Cardie[1], Serge Belongie[3]
Bharath Hariharan[1], and Ser-Nam Lim[2]

[1]Cornell University [2]Meta AI [3]University of Copenhagen

$$
\begin{aligned}
[\mathbf{x}_1, \mathbf{Z}_1, \mathbf{E}_1] &= L_1([\mathbf{x}_0, \mathbf{P}, \mathbf{E}_0]) \\
[\mathbf{x}_i, \mathbf{Z}_i, \mathbf{E}_i] &= L_i([\mathbf{x}_{i-1}, \mathbf{Z}_{i-1}, \mathbf{E}_{i-1}]) \qquad i = 2, 3, \ldots, N \\
\mathbf{y} &= \texttt{Head}(\mathbf{x}_N) \quad ,
\end{aligned}
$$

$$
\begin{aligned}
[\mathbf{x}_i, \_\_, \mathbf{E}_i] &= L_i([\mathbf{x}_{i-1}, \mathbf{P}_{i-1}, \mathbf{E}_{i-1}]) \qquad i = 1, 2, \ldots, N \\
\mathbf{y} &= \texttt{Head}(\mathbf{x}_N) \quad .
\end{aligned}
$$

Background

VPT (ECCV 2022)

Visual Prompt Tuning

Menglin Jia*[1,2], Luming Tang*[1]
Bor-Chun Chen[2], Claire Cardie[1], Serge Belongie[3]
Bharath Hariharan[1], and Ser-Nam Lim[2]

[1]Cornell University [2]Meta AI [3]University of Copenhagen

	ViT-B/16 (85.8M)	Total params	Scope Input	Scope Backbone	Extra params	FGVC	VTAB-1k Natural	VTAB-1k Specialized	VTAB-1k Structured
	Total # of tasks					5	7	4	8
(a)	FULL	24.02×		✓		88.54	75.88	83.36	47.64
(b)	LINEAR	1.02×				79.32 (0)	68.93 (1)	77.16 (1)	26.84 (0)
	PARTIAL-1	3.00×				82.63 (0)	69.44 (2)	78.53 (0)	34.17 (0)
	MLP-3	1.35×			✓	79.80 (0)	67.80 (2)	72.83 (0)	30.62 (0)
(c)	SIDETUNE	3.69×		✓	✓	78.35 (0)	58.21 (0)	68.12 (0)	23.41 (0)
	BIAS	1.05×		✓		88.41 (3)	73.30 (3)	78.25 (0)	44.09 (2)
	ADAPTER	1.23×		✓	✓	85.66 (2)	70.39 (4)	77.11 (0)	33.43 (0)
(ours)	VPT-SHALLOW	1.04×	✓		✓	84.62 (1)	76.81 (4)	79.66 (0)	46.98 (4)
	VPT-DEEP	1.18×				**89.11** (4)	**78.48** (6)	**82.43** (2)	**54.98** (8)

Background

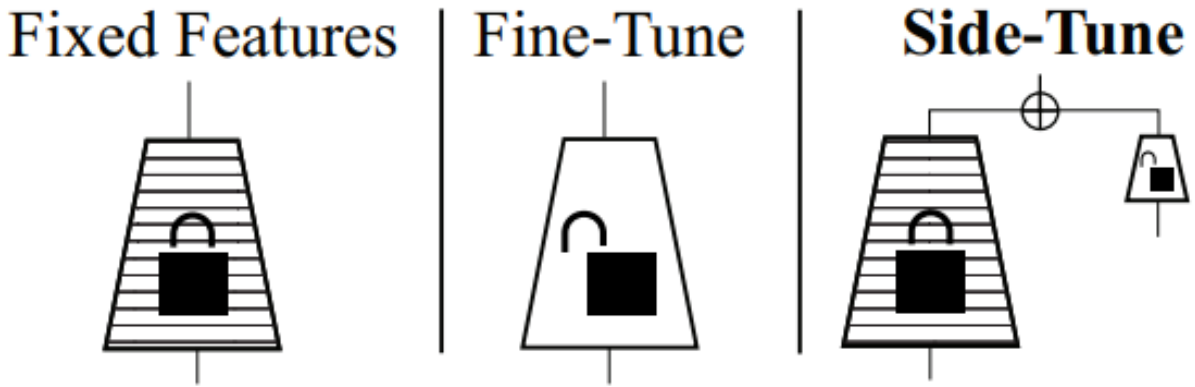
- Side-Tuning (ECCV 2020)

Side-Tuning: A Baseline for Network Adaptation via Additive Side Networks

Jeffrey O. Zhang[1], Alexander Sax[1], Amir Zamir[3], Leonidas Guibas[2], and Jitendra Malik[1]

[1] UC Berkeley
[2] Stanford University
[3] Swiss Federal Institute of Technology (EPFL)

http://sidetuning.berkeley.edu

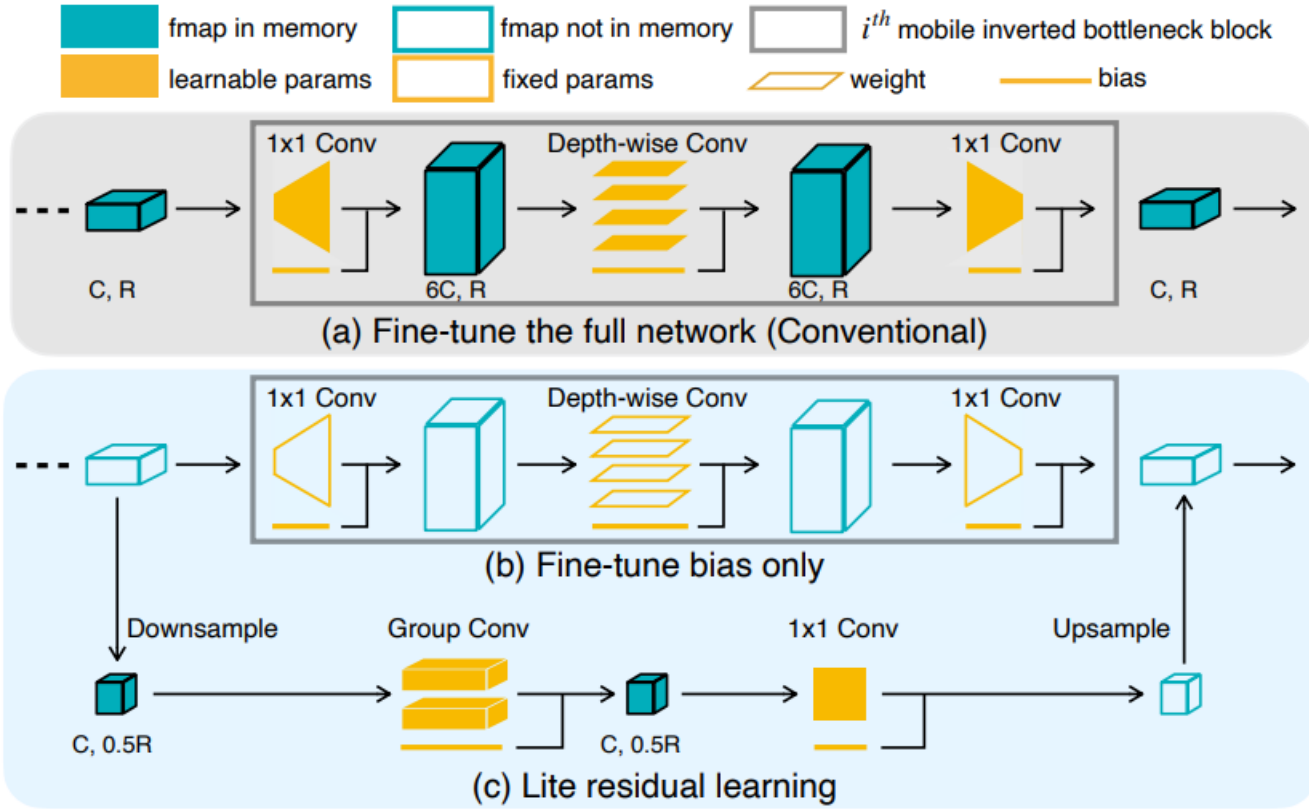


Background

Bias (NeurIPS 2020)

TinyTL: Reduce Memory, Not Parameters for Efficient On-Device Learning

Han Cai[1], **Chuang Gan**[2], **Ligeng Zhu**[1], **Song Han**[1]
[1]Massachusetts Institute of Technology, [2]MIT-IBM Watson AI Lab
http://tinyml.mit.edu/

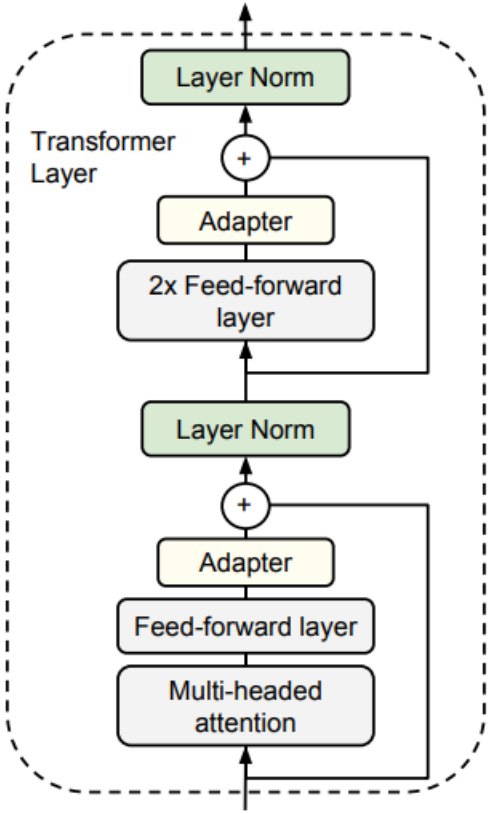


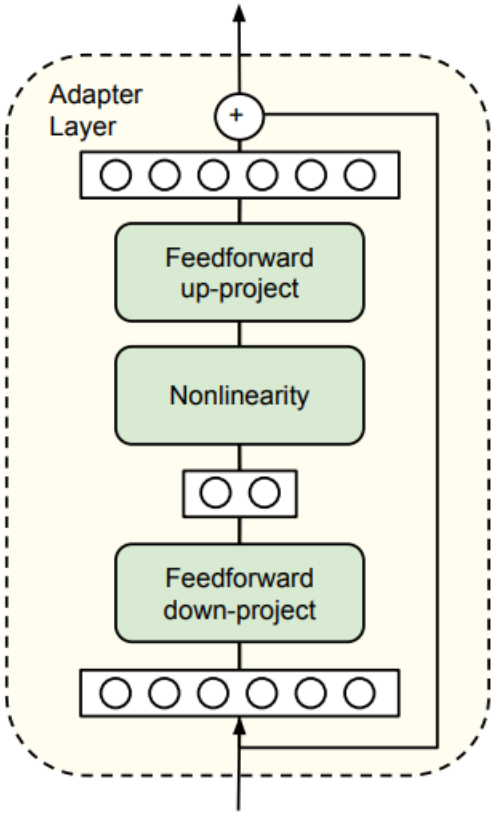
Background

Adapter (ICML 2019)

Parameter-Efficient Transfer Learning for NLP

Neil Houlsby [1] Andrei Giurgiu [1 *] Stanisław Jastrzębski [2 *] Bruna Morrone [1] Quentin de Laroussilhe [1]
Andrea Gesmundo [1] Mona Attariyan [1] Sylvain Gelly [1]





Background

VPT (ECCV 2022)

Visual Prompt Tuning

Menglin Jia*[1,2], Luming Tang*[1]
Bor-Chun Chen[2], Claire Cardie[1], Serge Belongie[3]
Bharath Hariharan[1], and Ser-Nam Lim[2]

[1]Cornell University [2]Meta AI [3]University of Copenhagen

	ViT-B/16 (85.8M)	Total params	Scope Input	Scope Backbone	Extra params	FGVC	VTAB-1k Natural	VTAB-1k Specialized	VTAB-1k Structured
	Total # of tasks					5	7	4	8
(a)	FULL	24.02×		✓		88.54	75.88	83.36	47.64
(b)	LINEAR	1.02×				79.32 (0)	68.93 (1)	77.16 (1)	26.84 (0)
	PARTIAL-1	3.00×				82.63 (0)	69.44 (2)	78.53 (0)	34.17 (0)
	MLP-3	1.35×			✓	79.80 (0)	67.80 (2)	72.83 (0)	30.62 (0)
(c)	SIDETUNE	3.69×		✓	✓	78.35 (0)	58.21 (0)	68.12 (0)	23.41 (0)
	BIAS	1.05×		✓		88.41 (3)	73.30 (3)	78.25 (0)	44.09 (2)
	ADAPTER	1.23×		✓	✓	85.66 (2)	70.39 (4)	77.11 (0)	33.43 (0)
(ours)	VPT-SHALLOW	1.04×	✓		✓	84.62 (1)	76.81 (4)	79.66 (0)	46.98 (4)
	VPT-DEEP	1.18×				**89.11** (4)	**78.48** (6)	**82.43** (2)	**54.98** (8)

Background

Parameter-Efficient Fine-tuning

MAM Adapter (ICLR 2022)

TOWARDS A UNIFIED VIEW OF PARAMETER-EFFICIENT TRANSFER LEARNING

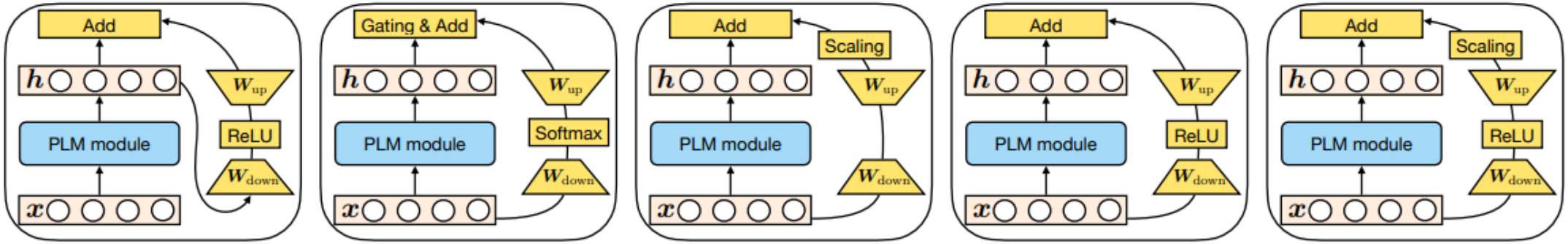
Junxian He* Carnegie Mellon University junxianh@cs.cmu.edu

Chunting Zhou* Carnegie Mellon University chuntinz@cs.cmu.edu

Xuezhe Ma University of Southern California xuezhema@isi.edu

Taylor Berg-Kirkpatrick UC San Diego tberg@eng.ucsd.edu

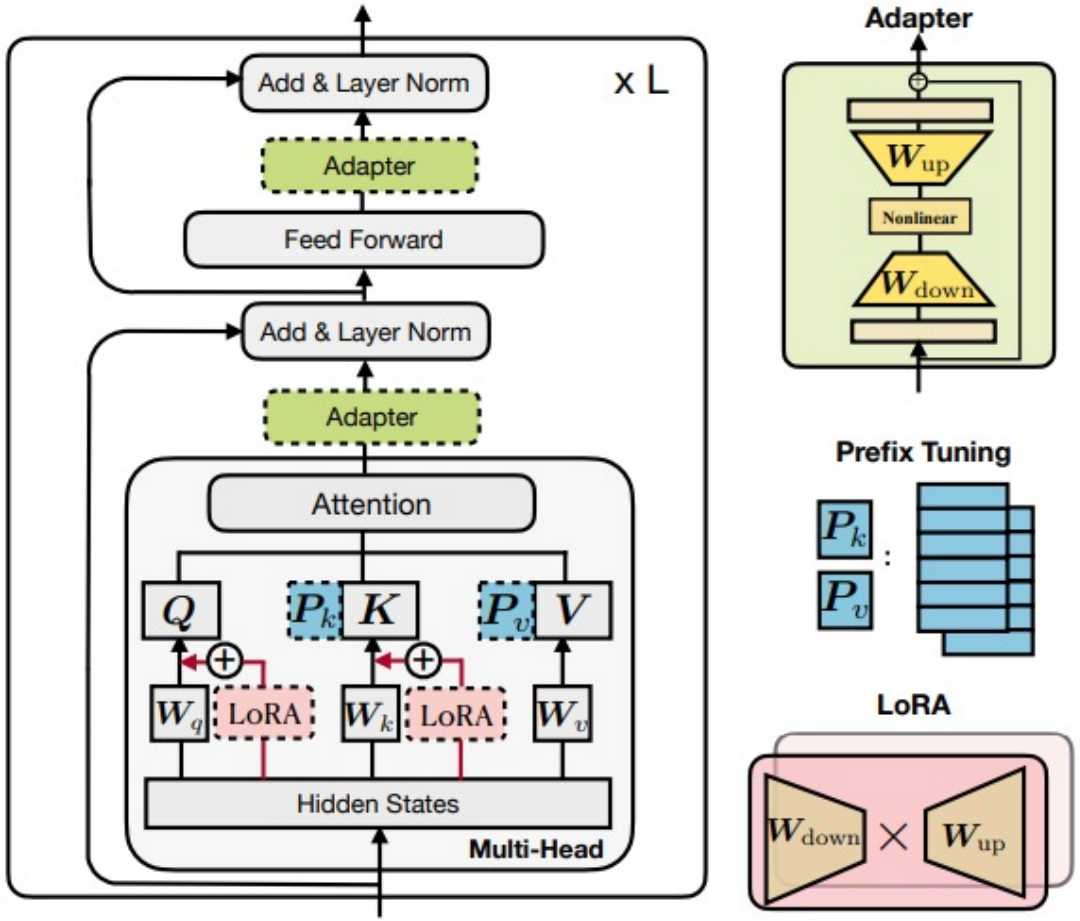
Graham Neubig Carnegie Mellon University gneubig@cs.cmu.edu



(a) Adapter (b) Prefix Tuning (c) LoRA (d) Parallel Adapter (e) Scaled PA

Background

- Parameter-Efficient Fine-tuning
  - MAM Adapter (ICLR 2022)



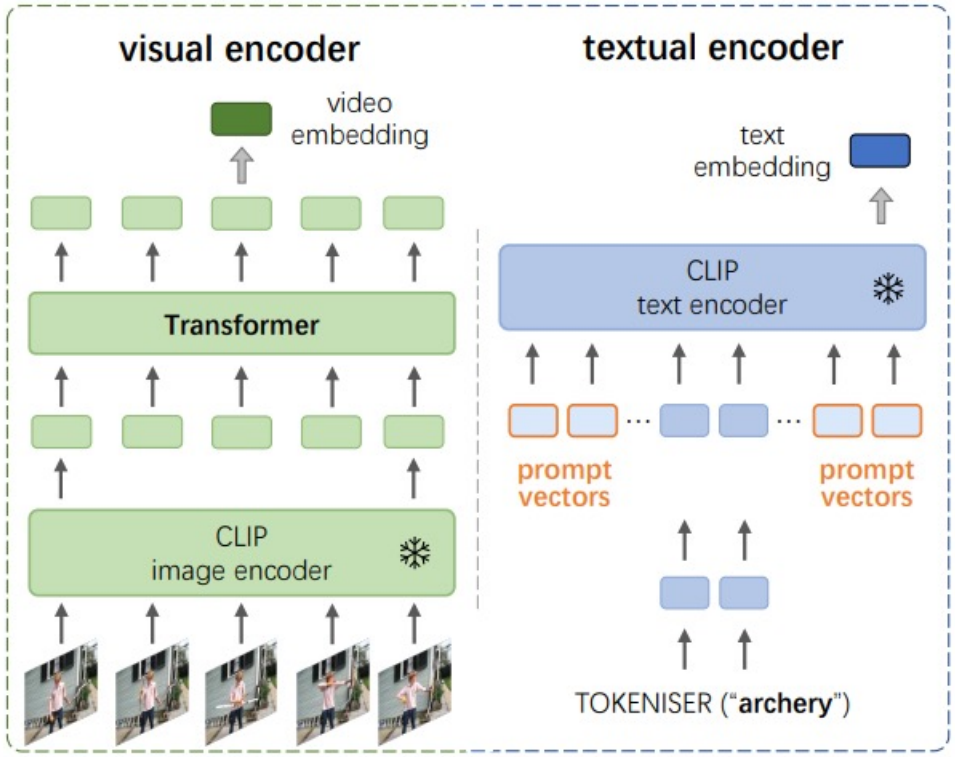
Outline

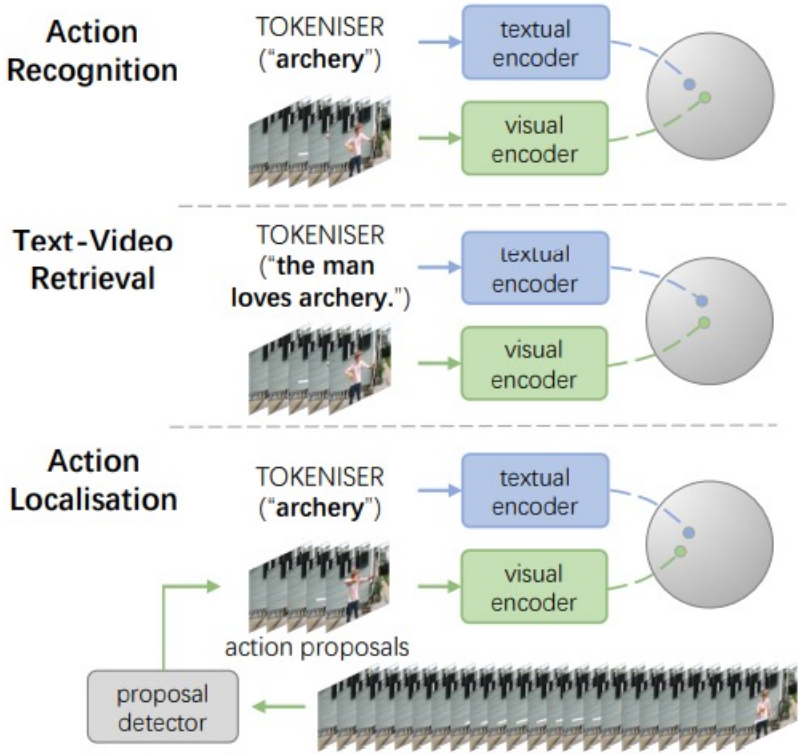
Method

- Image → Video
  - Hard to collect data
  - Huge computational power
- Problem
  - Domain Gap
  - Task Gap

Method

Pipeline

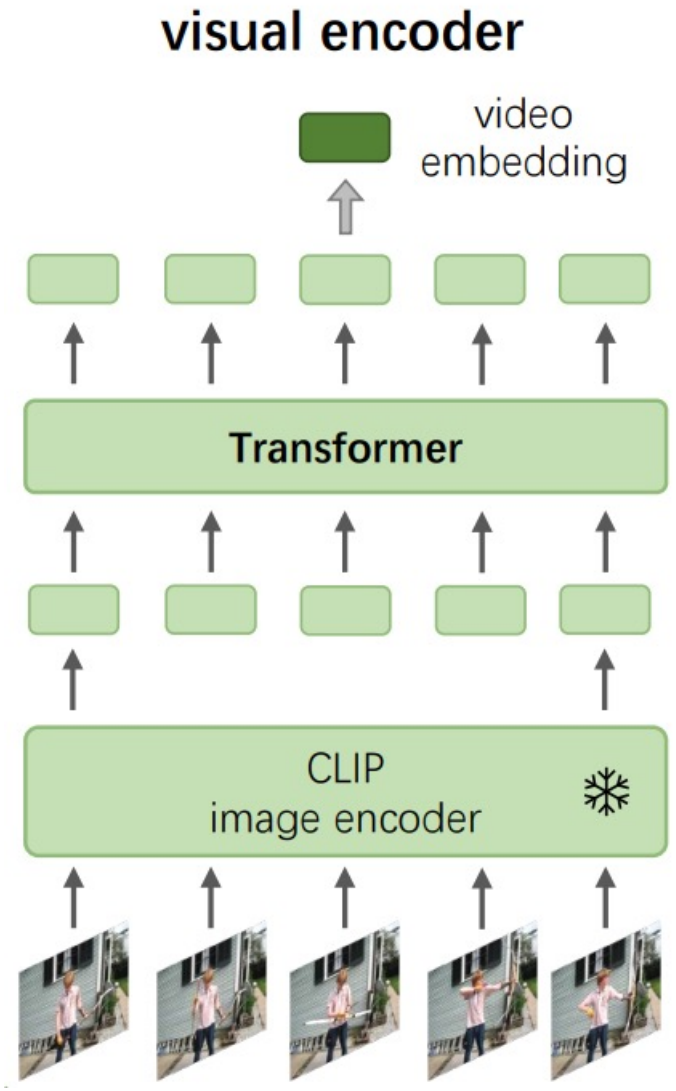




Method

Temporal Modeling

$$v_i = \Phi_{\text{video}}(\mathcal{V}_i) = \Phi_{\text{TEMP}}(\{\Phi_{\text{image}}(I_{i1}), \ldots, \Phi_{\text{image}}(I_{iT})\})$$

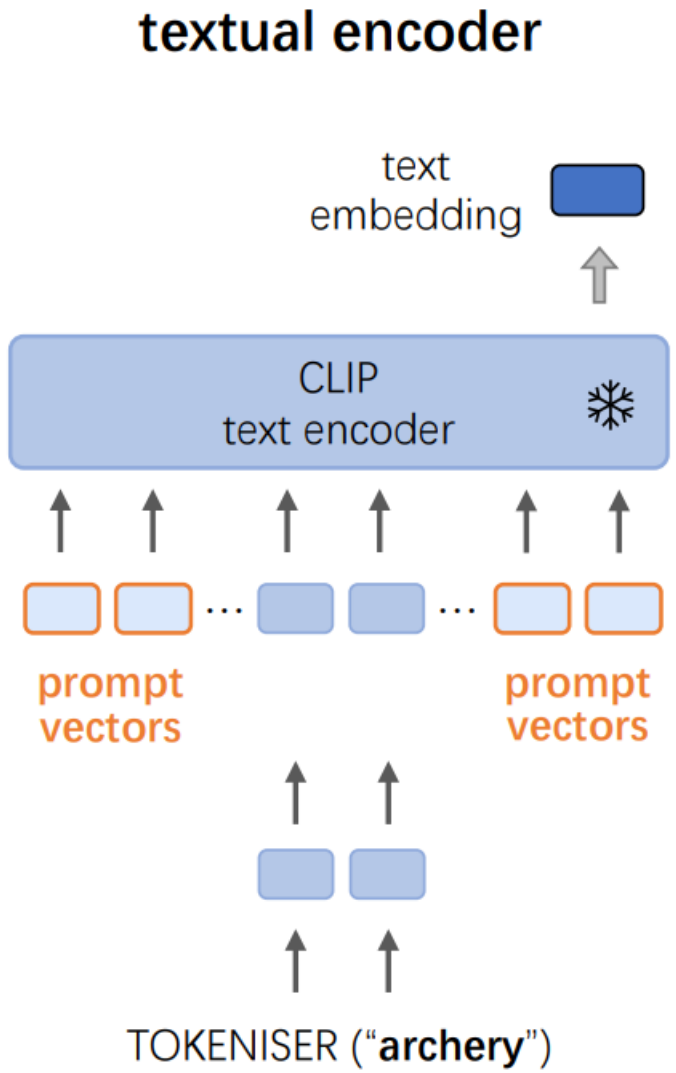


Method

Model Adaptation by Learning Prompts

$$c_{\text{archery}} = \Phi_{\text{text}}(a_1, \ldots, \text{TOKENISER}(\text{“archery”}), \ldots, a_k)$$

$$c_{\text{bowling}} = \Phi_{\text{text}}(a_1, \ldots, \text{TOKENISER}(\text{“bowling”}), \ldots, a_k)$$



Outline

Experiments

Action Recognition

Table 2: **Comparison on closed-set action recognition**. On all datasets, our model performs comparably to existing methods, by training far fewer parameters.

	HMDB-51		UCF-101		K-400		K-700	
Method	TOP1	TOP5	TOP1	TOP5	TOP1	TOP5	TOP1	TOP5
I3D [13]	74.3	–	95.1	–	71.6	90.0	58.7	81.7
S3D-G [85]	75.9	–	96.8	–	74.7	93.4	–	–
R(2+1)D [79]	74.5	–	96.8	–	72.0	90.0	–	–
TSM [50]	–	–	–	–	74.7	–	–	–
R3D-50 [33]	66.0	–	92.0	–	–	–	54.7	–
NL-I3D [83]	66.0	–	–	–	76.5	92.6	–	–
SlowFast [20]	–	–	–	–	77.0	92.6	–	–
X3D-XXL [18]	–	–	–	–	80.4	94.6	–	–
TimeSformer-L [4]	–	–	–	–	80.7	94.7	–	–
Ours (A5)	66.4	92.1	93.6	99.0	76.6	93.3	64.7	88.5

Experiments

Few-Shot Action Recognition

Table 4: **Comparison on few-shot action recognition**. Here, $\mathcal{C}_{\text{ALL}}$ refers to the case where the model is evaluated on all action categories of the corresponding dataset, rather than only 5-way classification, *e.g.* 101 categories for UCF, 400 categories for K-400. Baseline-I denotes the "zero-shot" CLIP inference with handcrafted prompts.

Method	K-shot	N-way	Prompt	Temporal	UCF-101	HMDB-51	K-400
CMN [101]	5	5	–	–	–	–	78.9
TARN [5]	5	5	–	–	–	–	78.5
ARN [94]	5	5	–	–	83.1	60.6	82.4
TRX [68]	5	5	–	–	96.1	75.6	85.9
Baseline-I [69]	–	5	hand-craft	✗	91.9	68.9	95.1
Ours	5	5	✓	✗	**98.3**	**85.3**	**96.4**
	5	5	✓	✓	97.8	84.9	96.0
Baseline-I [69]	–	$\mathcal{C}_{\text{ALL}}$	hand-craft	✗	64.7	40.1	54.2
Ours	5	$\mathcal{C}_{\text{ALL}}$	✓	✗	77.6	56.0	57.1
	5	$\mathcal{C}_{\text{ALL}}$	✓	✓	**79.5**	**56.6**	**58.5**

Experiments

Zero-Shot Action Recognition

Table 5: **Ablation study for zero-shot action recognition on K-700**. Baseline-I refers to the results from the CLIP zero-shot evaluation. The model is trained on 400 action categories and evaluated on the other 300 disjoint categories.

Model	Prompt	Temporal	TOP1	TOP5	AVG
Baseline-I [69]	hand-craft	✗	52.4	77.3	64.9
C0	4+X+4	✗	57.4	83.3	70.4
C1	8+X+8	✗	57.7	82.6	70.2
C2	16+X+16	✗	**58.4**	82.6	70.5
C3	32+X+32	✗	57.5	**84.6**	**71.1**
C4	16+X+16	1-TFM	47.9	76.8	62.4
C5	16+X+16	2-TFM	45.5	75.4	60.5
C6	16+X+16	3-TFM	45.6	75.2	60.4

Outline

1 Authors

2 Background

3 Method

4 Experiments

5 **Discussion**

Discussion

- Parameter-Efficient Fine-tuning
  - Prompt
  - Adapter

Thanks!