Multitask Vision-Language Prompt Tuning

Sheng Shen* Shijia Yang* Tianjun Zhang* Bohan Zhai
Joseph E. Gonzalez Kurt Keutzer Trevor Darrell
University of California, Berkeley
{sheng.s,shijiayang,tianjunz,zhaibohan,jegonzal,keutzer,trevordarrell}@berkeley.edu

1. Appendix

1.1. Additional Results

Ablation on UPT As mentioned in the main text, due to the recency, [2] does not release their model details or code. We therefore implement our own variant that simply concatenates the CoOp prompt vectors $\boldsymbol{U}_T$ and VPT-deep prompt vector $\boldsymbol{U}_V$ together as $\boldsymbol{U}$, we set the context length of $\boldsymbol{U}_T$ and $\boldsymbol{U}_V$ the same as 4 unless specify. We use a one-layer one-head Transformer block θ whose hidden dimension is cut to be 128. Before and after feeding $\boldsymbol{U}$ to θ, a linear layer is employed to match the dimensionality. We ablate the design choice on number of heads, number of layer, and dimensionality, respectively in Figure 1. The size of the each point stands for the relative additional parameter size included in this setting.

1.2. Task Group Information

We provide detailed task group information here. We follow Table 1 for multitask adaptation where group of 1 task means using "Target task" column only, group of 2 tasks means target task with task 1, and group of 3 tasks means target tasks, task 1, with task 2.

1.3. CIFAR-10 for Cross-task Generation

As stated in main text Table 1, we include CIFAR-10 result for cross-task generation in Table 2.

1.4. Standard Deviation

Since all experiments are in the few-shot setting, we provide standard deviation for few-shot ELEVATER experiments in Table 3.

1.5. Theoretical Explanation

We here theoretically justify our proposed task grouping in the context of prompt multitask adaptation. In our method, we group task $i \in T$ with task $j \in T$ if P_i, the task i specific prompt, performs better on task j than other tasks in T. For simplicity, we do not consider over-fitting and other edge cases, then task i has the largest positive effect of the gradient update on the given task j. To measure the effect of gradient update, we calculate the ratio of the loss after and before we plug in P_i. Formally, we define an asymmetric measure for calculating the affinity of task i at a given time-step t on task j as:

$$Z^t_{i \to j} = \frac{L_j(\chi^t, P^t_i)}{L_j(\chi^t, P^t_j)}$$

where χ^t is an inference batch. P^t_i represent task i prompt on time-step t. L_j is the loss of task j. Finally, we average across all steps:

$$\hat{Z}_{i \to j} = \frac{1}{T} \sum_{t=1}^{T} Z^t_{i \to j}$$

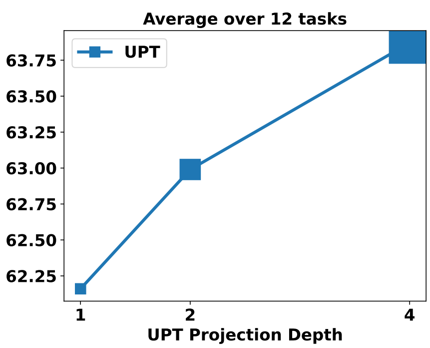
The proposed metric is similar to inter-task affinity in [1]. In Section 4.3, [1] theoretically prove that if task b induces higher affinity than task c on task a, training $\{a, b\}$ together result in a lower loss on task a than training $\{a, c\}$. By plugging $Z_{b \to a}$ and $Z_{c \to a}$ in the proof, Lemma 1 generalizes to our case and thus lead to the same conclusion.

*Equal contribution

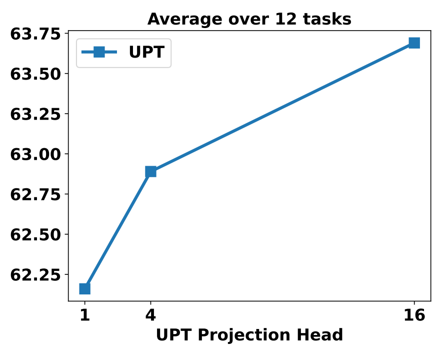
References

[1] Chris Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. Efficiently identifying task groupings for multi-task learning. *Advances in Neural Information Processing Systems*, 34:27503–27516, 2021. 1

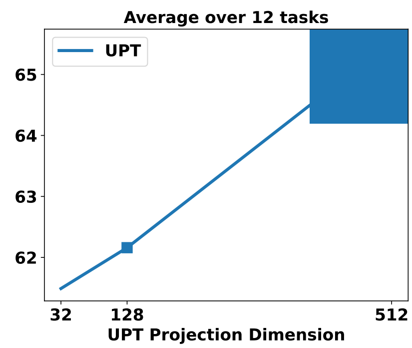
[2] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Unified vision and language prompt learning. *arXiv preprint arXiv:2210.07225*, 2022. 1



(a) Ablation on the UPT Projection Depth.



(b) Ablation on number of heads.



(c) Ablation on the UPT Projection Dimension.

Figure 1. Ablation on the hyper-parameter of UPT. Note that the size of each point stands for the relative additional parameter size.

Table 1. Task group for CoOp, VPT, and UPT.

Model	Target task	Task 1	Task 2
CoOp	Caltech101	DTD	CIFAR10
	CIFAR10	VOC 2007	Resisc-45
	CIFAR100	Caltech101	CIFAR10
	Country-211	Caltech101	Resisc-45
	DTD	Caltech101	MNIST
	EuroSat	Resisc-45	CIFAR100
	FER 2013	CIFAR100	MNIST
	FGVCAircraft	Caltech101	DTD
	Flowers102	CIFAR10	Caltech101
	Food101	Caltech101	DTD
	GTSRB	MNIST	CIFAR100
	Hateful Memes	VOC 2007	Caltech101
	KITTI Distance	StanfordCars	OxfordPets
	MNIST	DTD	Resisc-45
	OxfordPets	Caltech101	CIFAR10
	Patch-Camelyon	CIFAR100	Caltech101
	Rendered-SST2	FGVCAircraft	Hateful Memes
	Resisc-45	Caltech101	CIFAR10
	StanfordCars	Caltech101	MNIST
	VOC 2007	CIFAR100	Caltech101
VPT	Caltech101	CIFAR100	CIFAR10
	CIFAR10	CIFAR100	Caltech101
	CIFAR100	CIFAR10	Caltech101
	Country-211	EuroSat	Food101
	DTD	CIFAR10	Rendered-SST2
	EuroSat	Resisc-45	FER 2013
	FER 2013	OxfordPets	MNIST
	FGVCAircraft	EuroSat	CIFAR10
	Flowers102	CIFAR100	EuroSat
	Food101	CIFAR10	EuroSat
	GTSRB	MNIST	CIFAR100
	Hateful Memes	FER 2013	OxfordPets
	KITTI Distance	VOC 2007	Flowers102
	MNIST	Resisc-45	GTSRB
	OxfordPets	CIFAR10	Rendered-SST2
	Patch-Camelyon	CIFAR10	Food101
	Rendered-SST2	Resisc-45	Patch-Camelyon
	Resisc-45	EuroSat	CIFAR10
	StanfordCars	CIFAR10	EuroSat
	VOC 2007	CIFAR100	CIFAR10
UPT	Caltech101	CIFAR10	CIFAR100
	CIFAR10	CIFAR100	Caltech101
	CIFAR100	Caltech101	EuroSat
	Country-211	Caltech101	CIFAR100
	DTD	CIFAR10	Caltech101
	EuroSat	Resisc-45	CIFAR100
	FER 2013	MNIST	DTD
	FGVCAircraft	CIFAR100	CIFAR10
	Flowers102	Caltech101	CIFAR100
	Food101	Caltech101	CIFAR10
	GTSRB	MNIST	CIFAR100
	Hateful Memes	Caltech101	CIFAR100
	KITTI Distance	Food101	Flowers102
	MNIST	CIFAR100	GTSRB
	OxfordPets	Caltech101	CIFAR100
	Patch-Camelyon	Food101	StanfordCars
	Rendered-SST2	Hateful Memes	GTSRB
	Resisc-45	CIFAR10	CIFAR100
	StanfordCars	Caltech101	CIFAR100
	VOC 2007	CIFAR100	Caltech101

Table 2. Cross-task generation experiment additional result.

(a) CIFAR-10

# shots	1	5	20
CoOp	89.45$_{\pm1.6}$	83.63$_{\pm2.0}$	91.38$_{\pm0.6}$
CoCoOp	**92.61**$_{\pm1.5}$	84.91$_{\pm1.6}$	91.85$_{\pm0.6}$
VPT	86.81$_{\pm1.4}$	86.77$_{\pm1.0}$	90.83$_{\pm0.7}$
UPT	88.44$_{\pm0.6}$	89.47$_{\pm0.9}$	91.33$_{\pm0.6}$
MCoOp	90.48$_{\pm1.2}$	90.86$_{\pm1.6}$	92.92$_{\pm0.5}$
MCoCoOp	93.16$_{\pm0.6}$	**96.26**$_{\pm0.6}$	**98.10**$_{\pm0.4}$
MVPT	88.97$_{\pm0.5}$	89.89$_{\pm0.5}$	92.13$_{\pm0.3}$
MUPT	**92.61**$_{\pm0.2}$	91.52$_{\pm0.6}$	93.72$_{\pm0.4}$

Table 3. Few-shot ELEVATER experiment with standard deviation.

	Source	Adaptation	Target: Caltech101	CIFAR10	CIFAR100	Country-211	DTD	EuroSat	FER 2013	FGVCAircraft	Flowers102	Food101
CLIP†	-	-	88.9	90.8	68.2	22.8	44.8	54.7	48.5	24.3	88.7	43.5
CoOp	-	S	91.44$_{\pm0.4}$	91.30$_{\pm1.0}$	73.01$_{\pm0.2}$	22.83$_{\pm0.7}$	69.82$_{\pm0.9}$	80.19$_{\pm3.0}$	54.46$_{\pm3.4}$	42.01$_{\pm0.6}$	93.31$_{\pm0.3}$	89.47$_{\pm0.1}$
VPT	-	S	92.84$_{\pm0.4}$	91.39$_{\pm0.7}$	75.98$_{\pm0.9}$	21.11$_{\pm0.4}$	68.56$_{\pm1.1}$	87.37$_{\pm3.7}$	56.77$_{\pm0.6}$	42.12$_{\pm1.2}$	89.22$_{\pm1.2}$	89.04$_{\pm0.2}$
UPT	-	S	92.58$_{\pm0.4}$	92.05$_{\pm1.1}$	76.61$_{\pm0.2}$	23.37$_{\pm0.5}$	67.68$_{\pm0.6}$	88.98$_{\pm2.4}$	56.87$_{\pm1.9}$	42.46$_{\pm1.0}$	89.59$_{\pm0.5}$	89.64$_{\pm0.4}$
MCoOp	-	M	91.53$_{\pm0.2}$	91.67$_{\pm0.5}$	73.01$_{\pm0.2}$	23.12$_{\pm0.3}$	69.82$_{\pm0.9}$	81.69$_{\pm5.4}$	54.46$_{\pm3.4}$	42.01$_{\pm0.6}$	93.44$_{\pm0.3}$	89.47$_{\pm0.1}$
MVPT	-	M	92.84$_{\pm0.4}$	93.54$_{\pm0.4}$	76.39$_{\pm0.3}$	21.42$_{\pm0.1}$	68.56$_{\pm1.1}$	89.15$_{\pm1.1}$	56.77$_{\pm0.6}$	42.12$_{\pm1.2}$	89.22$_{\pm1.2}$	89.04$_{\pm0.2}$
MUPT	-	M	92.58$_{\pm0.4}$	93.38$_{\pm0.6}$	76.61$_{\pm0.2}$	23.37$_{\pm0.6}$	67.68$_{\pm0.6}$	88.98$_{\pm2.4}$	56.94$_{\pm1.5}$	42.46$_{\pm1.0}$	89.59$_{\pm0.5}$	**89.64**$_{\pm0.4}$
MCoOp	M	M	92.09$_{\pm0.2}$	91.59$_{\pm0.9}$	72.63$_{\pm0.1}$	**23.52**$_{\pm0.2}$	**70.41**$_{\pm0.4}$	81.70$_{\pm1.7}$	54.85$_{\pm1.7}$	42.34$_{\pm1.2}$	**93.61**$_{\pm0.1}$	89.14$_{\pm0.5}$
MVPT	M	M	**93.46**$_{\pm0.1}$	93.72$_{\pm0.4}$	**77.38**$_{\pm0.1}$	20.79$_{\pm0.1}$	69.43$_{\pm0.2}$	**92.23**$_{\pm1.8}$	**57.07**$_{\pm1.6}$	**42.57**$_{\pm0.6}$	88.80$_{\pm1.6}$	87.78$_{\pm0.1}$
MUPT	M	M	92.19$_{\pm0.8}$	**93.75**$_{\pm0.9}$	75.39$_{\pm1.3}$	23.45$_{\pm0.3}$	65.99$_{\pm1.4}$	90.17$_{\pm0.3}$	56.06$_{\pm1.6}$	41.19$_{\pm0.8}$	89.34$_{\pm0.5}$	89.38$_{\pm0.1}$
Δ			**+0.62**	**+1.70**	**+1.40**	**+0.69**	**+0.59**	**+4.86**	**+0.30**	**+0.45**	**+0.30**	**+0.00**

	Source	Adaptation	Target: GTSRB	Hateful Memes	KITTI Distance	MNIST	OxfordPets	Patch-Camelyon	Rendered-SST2	Resisc-45	StanfordCars	VOC 2007
CLIP†	-	-	58.1	27.0	52.0	69.4	89.0	54.0	60.9	65.6	64.8	83.7
CoOp	-	S	73.87$_{\pm2.0}$	52.40$_{\pm2.4}$	56.87$_{\pm6.0}$	91.44$_{\pm2.9}$	90.69$_{\pm1.0}$	62.79$_{\pm3.8}$	59.55$_{\pm4.9}$	83.83$_{\pm1.7}$	79.52$_{\pm0.8}$	74.61$_{\pm2.4}$
VPT	-	S	85.34$_{\pm1.1}$	56.60$_{\pm1.4}$	53.54$_{\pm7.1}$	89.88$_{\pm2.4}$	90.71$_{\pm0.4}$	60.30$_{\pm4.8}$	57.66$_{\pm4.8}$	84.05$_{\pm0.3}$	74.95$_{\pm0.8}$	78.88$_{\pm1.8}$
UPT	-	S	82.72$_{\pm1.3}$	56.87$_{\pm3.6}$	47.87$_{\pm11.9}$	89.11$_{\pm2.3}$	91.24$_{\pm0.6}$	60.41$_{\pm1.0}$	59.03$_{\pm3.8}$	83.32$_{\pm0.2}$	76.40$_{\pm0.1}$	81.20$_{\pm1.4}$
MCoOp	-	M	74.38$_{\pm0.3}$	58.40$_{\pm1.1}$	56.87$_{\pm6.0}$	91.44$_{\pm2.9}$	90.69$_{\pm1.0}$	64.91$_{\pm3.4}$	61.63$_{\pm2.1}$	84.03$_{\pm0.3}$	**79.52**$_{\pm0.8}$	78.45$_{\pm2.0}$
MVPT	-	M	85.34$_{\pm1.1}$	58.20$_{\pm1.0}$	53.54$_{\pm7.1}$	89.88$_{\pm2.4}$	91.01$_{\pm0.4}$	66.53$_{\pm8.8}$	58.14$_{\pm2.2}$	84.05$_{\pm0.3}$	74.95$_{\pm0.8}$	80.69$_{\pm1.0}$
MUPT	-	M	82.72$_{\pm1.3}$	58.13$_{\pm2.2}$	55.41$_{\pm6.6}$	89.91$_{\pm2.3}$	**91.24**$_{\pm0.6}$	63.36$_{\pm9.8}$	61.34$_{\pm2.9}$	83.32$_{\pm0.2}$	76.40$_{\pm0.1}$	81.20$_{\pm1.4}$
MCoOp	M	M	72.74$_{\pm1.1}$	58.40$_{\pm1.0}$	47.73$_{\pm3.9}$	90.21$_{\pm0.1}$	89.61$_{\pm0.9}$	68.92$_{\pm3.0}$	**64.89**$_{\pm2.7}$	**84.39**$_{\pm0.5}$	79.43$_{\pm0.6}$	79.55$_{\pm1.6}$
MVPT	M	M	**89.62**$_{\pm0.9}$	55.53$_{\pm1.7}$	**62.07**$_{\pm6.1}$	**93.08**$_{\pm1.7}$	91.04$_{\pm0.3}$	69.69$_{\pm1.8}$	57.50$_{\pm1.0}$	84.35$_{\pm0.2}$	74.20$_{\pm0.6}$	**82.21**$_{\pm0.5}$
MUPT	M	M	81.66$_{\pm2.7}$	**59.00**$_{\pm1.0}$	57.20$_{\pm5.7}$	91.38$_{\pm1.5}$	90.30$_{\pm1.0}$	**69.74**$_{\pm4.0}$	62.29$_{\pm2.4}$	83.40$_{\pm0.4}$	76.66$_{\pm0.4}$	79.29$_{\pm1.8}$
Δ			**+4.28**	**+2.13**	**+8.53**	**+3.20**	**+0.00**	**+9.33**	**+5.34**	**+0.56**	**+0.00**	**+3.33**