Joint Scheduling of Causal Prompts and Tasks for Multi-Task Learning

Chaoyang Li[1,2], Jianyang Qin[1], Jinhao Cui[1], Zeyu Liu[1], Ning Hu[2], Qing Liao[1,2*]
[1]Harbin Institute of Technology, Shenzhen, China
[2]Peng Cheng Laboratory, Shenzhen, China
{lichy, hun}@pcl.ac.cn, {22b351005, cuijinhao, liuzeyu}@stu.hit.edu.cn, liaoqing@hit.edu.cn

1. Appendix

1.1. Derivation of the Implicit Gradient

In this subsection, we present the derivation of the implicit gradient. Drawing upon the Cauchy Implicit Function Theorem [13], if there exists a point (θ_0, Ω_0) where $\nabla_\theta \mathcal{L}(\theta, \Omega) = 0$, and the regularity conditions are satisfied, then within the neighborhood of (θ_0, Ω_0), there exists an implicit function $\theta^*(\Omega)$ such that the condition $\nabla_\theta \mathcal{L}(\theta, \Omega) = 0|_{\Omega, \theta^*(\Omega)}$. Assuming that $\nabla^2_\theta \mathcal{L}(\theta^*, \Omega)$ is positive definite, we have the following derivation,

$$\nabla_\theta \mathcal{L}\left(\theta^*(\Omega), \Omega\right) = 0, \tag{1}$$

$$\nabla^2_\theta \mathcal{L}\left(\theta^*, \Omega\right) \nabla_\Omega \theta^* + \nabla_\Omega \nabla_{\theta^*} \mathcal{L}\left(\theta^*, \Omega\right) = 0, \tag{2}$$

$$\nabla_\Omega \theta^* = -\left(\nabla^2_\theta \mathcal{L}\left(\theta^*, \Omega\right)\right)^{-1} \nabla_\Omega \nabla_\theta \mathcal{L}\left(\theta^*, \Omega\right). \tag{3}$$

Starting from Eq. (1) to Eq. (2), we perform the derivation to Ω on both sides of the Eq. (1). Under the assumption that $\nabla^2_\theta \mathcal{L}(\theta^*, \Omega)$ is the positive define, it possesses an inverse, allowing us to solve for the desired derivative uniquely. By leveraging this inverse $\left(\nabla^2_\theta \mathcal{L}(\theta^*, \Omega)\right)^{-1}$, we can subsequently derive the implicit gradient in Eq. (3),

1.2. H-truncated Neumann Series Approximation

Directly computing the inverse of the Hessian matrix in the implicit gradient for deep neural models is often computationally intractable due to its immense size and complexity. To address this, we employ the H-truncated Neumann series [2] to approximate this inverse as shown in (4),

$$\left(\nabla^2_\theta \mathcal{L}\right)^{-1} = \sum_{j=0}^{\infty} \left(I - \nabla^2_\theta \mathcal{L}\right)^j \approx \sum_{j=0}^{H} \left(I - \nabla^2_\theta \mathcal{L}\right)^j, \tag{4}$$

where I is the identity matrix.

*Corresponding author

1.3. Complexity Analysis

JSCPT introduces additional learnable parameters equal to twice the number of tasks, this increase is deemed acceptable. During the upper-level optimization, we only update Ω, and θ is fixed via *detach()* operation. The time of MTL using a combined loss with fixed weights as $O(1)$ and the main difference between different methods comes from the gradient backward process. Given N tasks, the truncated Neumann series number as H, assuming the model conducts M times lower-level optimization and 1 upper-optimization. The cost of lower-level optimization is $O(M(N))$ and the cost of upper-level optimization is $O(H+N+2)$. Therefore, the time complexity for the gradient backward of JSCPT is $O((M(N)+H+N+2)/M) = O(N+(H+N)/M)$. Note that most gradient-based multi-task optimization methods [3, 11, 21], require calculating the gradient of each task and performing parameter back-propagation, with a time complexity of $O(N)$. Compared with them, our approach does not significantly increase the time complexity under a limited number of tasks.

1.4. Experimental setup

Dataset. We conduct experiments on three multi-task datasets, including Office-Home [18], MiniDomainNet [22], and a large-scale multi-task learning benchmark with 10 visual datasets.

- Office-Home comprises four distinct tasks: Art, Clipart, Product, and Real World, each has 65 object categories in diverse domains, about 15,500 images in total.
- MiniDomainNet is an extremely challenging multi-task dataset, including 140,000 images distributed among 126 categories. It contains four tasks: Clipart, Painting, Sketch, and Real.
- Large-Scale MTL Benchmark consists of 10 vision tasks, including fine-grained recognition (OxfordPets [15], StanfordCars [10], OxforsFlowers [14], Food101 [1], and FGVCAircraft [12]), texture recognition (DTD [4]), scene recognition (SUN397 [19]), general recognition (Caltech101 [5]), action recognition (UCF101 [17]), and satellite image recognition (EuroSAT [7]).

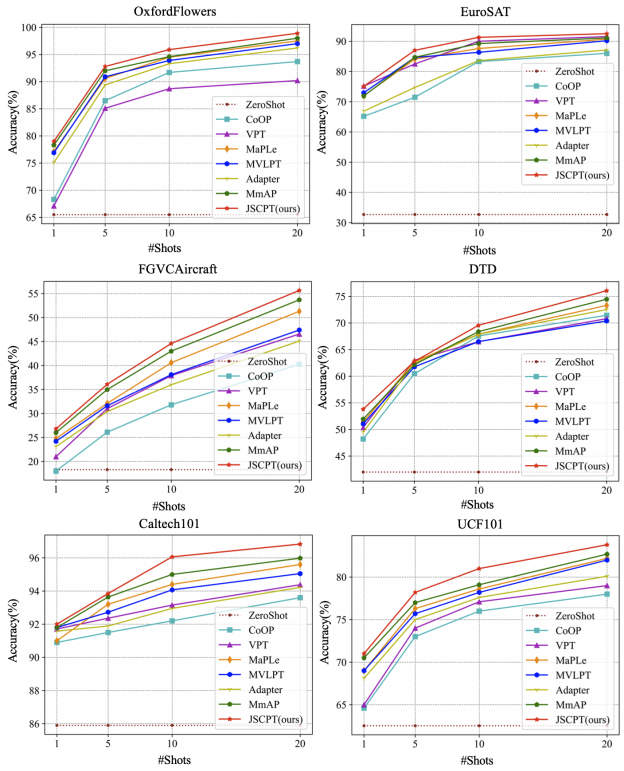

Figure 1. Comparison with accuracy(%) of various methods on the OxfordFlowers, EuroSAT, FGVCAircraft, DTD, Caltech101, and UCF101 datasets, under the few-shot setting.

Baselines. We compare JSCPT with 7 tuning baselines: (1) **Zero-Shot** uses hand-crafted text prompt ("a photo of [CLASS]") templates for zero-shot prediction; (2) **CoOp** [23] trains text prompt on an individual task; (3) **VPT** [8] learns a small number of trainable parameters in the visual space; (4) **MaPLe** [9] trains the coupled vision and language prompts; (5) **CLIP-Adapter** [6] learns feature adapters on either visual or language branch; (6) **MVLPT** [16] trains task-shared multi-modal prompts;(7) **MmAP** [20] learns the group-shared and task-specific prompts aligned with text and visual modalities. Except for MVLPT and MmAP, the other methods are mainly for single tasks. We build a multi-task version for the single-task method by training task-shared prompts or adapters.

1.5. Few-shot Results of Six Datasets

Fig. 1 shows the main results of different methods on six datasets (i.e., OxfordFlowers, EuroSAT, FGVCAircraft, DTD, Caltech101, and UCF101), under few-shot settings. We can see from Fig. 1 that JSCPT outperforms other methods on the six datasets under few-shot settings. This indicates that JSCPT can enhance the generalization of the multi-task vision-language prompt tuning with limited data.

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 - mining discriminative components with random forests. In *Computer Vision - ECCV 2014 - 13th European Conference*, pages 446–461. Springer, 2014. 1

[2] Hong Chen, Xin Wang, Chaoyu Guan, Yue Liu, and Wenwu Zhu. Auxiliary learning with joint task and data scheduling. In *International Conference on Machine Learning, ICML*, pages 3634–3647, 2022. 1

[3] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of the 35th International Conference on Machine Learning*, pages 793–802. PMLR, 2018. 1

[4] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 3606–3613. IEEE Computer Society, 2014. 1

[5] Li Fei-Fei, Robert Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Comput. Vis. Image Underst.*, 106(1):59–70, 2007. 1

[6] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *Int. J. Comput. Vis.*, 132(2):581–595, 2024. 2

[7] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, 12(7):2217–2226, 2019. 1

[8] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge J. Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Computer Vision - ECCV 2022*, pages 709–727. Springer, 2022. 2

[9] Muhammad Uzair Khattak, Hanoona Abdul Rasheed, Muhammad Maaz, Salman H. Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023*, pages 19113–19122. IEEE, 2023. 2

[10] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops, ICCV Workshops*, pages 554–561. IEEE Computer Society, 2013. 1

[11] Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems*, pages 18878–18890, 2021. 1

[12] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *CoRR*, abs/1306.5151, 2013. 1

[13] Kazuhisa Nakasho, Yuichi Futa, and Yasunari Shidama. Implicit function theorem. part I. *Formaliz. Math.*, 25(4):269–281, 2017. 1

[14] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing, ICVGIP*, pages 722–729. IEEE Computer Society, 2008. 1

[15] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505. IEEE Computer Society, 2012. 1

[16] Sheng Shen, Shijia Yang, Tianjun Zhang, Bohan Zhai, Joseph E. Gonzalez, Kurt Keutzer, and Trevor Darrell. Multitask vision-language prompt tuning. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV*, pages 5644–5655. IEEE, 2024. 2

[17] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. 1

[18] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 5385–5394, 2017. 1

[19] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 3485–3492. IEEE Computer Society, 2010. 1

[20] Yi Xin, Junlong Du, Qiang Wang, Ke Yan, and Shouhong Ding. Mmap: Multi-modal alignment prompt for cross-domain multi-task learning. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024*, pages 16076–16084, 2024. 2

[21] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems*, 2020. 1

[22] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain adaptive ensemble learning. *IEEE Trans. Image Process.*, 30:8008–8018, 2021. 1

[23] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *Int. J. Comput. Vis.*, 130(9):2337–2348, 2022. 2