ADAPTIVE CONTINUAL LEARNING THROUGH PROACTIVE DETECTION OF TRANSFER AND INTERFERENCE-SUPPLEMENTARY MATERIAL

Anonymous authors
Paper under double-blind review

A APPENDIX

A.1 THEORETICAL PROOF

Given the following equation:

$$\Delta = \arg\min_{\Delta} L_t(0) + \nabla_\theta L_t(0)^T \Delta \quad \text{subject to} \quad \frac{1}{2}\Delta^T F_{IM_{t-1}} \Delta \leq r^2$$

1. LAGRANGE MULTIPLIER METHOD

We introduce the Lagrange multiplier λ to handle the constraint:

$$\mathcal{L}(\Delta, \lambda) = L_t(0) + \nabla_\theta L_t(0)^T \Delta + \lambda \left(\frac{1}{2}\Delta^T F_{IM_{t-1}} \Delta - r^2 \right)$$

Taking the derivative with respect to Δ and setting it equal to 0:

$$\frac{\partial \mathcal{L}}{\partial \Delta} = \nabla_\theta L_t(0) + \lambda F_{IM_{t-1}} \Delta = 0$$

Solving for Δ:

$$\Delta = -\frac{1}{\lambda} F_{IM_{t-1}}^{-1} \nabla_\theta L_t(0)$$

2. SOLVING FOR λ

Using the constraint $\frac{1}{2}\Delta^T F_{IM_{t-1}} \Delta \leq r^2$, substitute Δ:

$$\frac{1}{2} \left(-\frac{1}{\lambda} \nabla_\theta L_t(0) \right)^T F_{IM_{t-1}} \left(-\frac{1}{\lambda} F_{IM_{t-1}}^{-1} \nabla_\theta L_t(0) \right) \leq r^2$$

Simplifying:

$$\frac{1}{2\lambda^2} \nabla_\theta L_t(0)^T F_{IM_{t-1}}^{-1} \nabla_\theta L_t(0) \leq r^2$$

Solving for λ:

$$\lambda^2 = \frac{1}{2r^2} \nabla_\theta L_t(0)^T F_{IM_{t-1}}^{-1} \nabla_\theta L_t(0)$$

Thus:

$$\lambda = \sqrt{\frac{1}{2r^2} \nabla_\theta L_t(0)^T F_{IM_{t-1}}^{-1} \nabla_\theta L_t(0)}$$

3. FINAL UPDATE RULE

Substituting λ back into the expression for Δ, we get the parameter update rule:

$$\Delta = -\frac{r}{\sqrt{\frac{1}{2} \nabla_\theta L_t(0)^T F_{IM_{t-1}}^{-1} \nabla_\theta L_t(0)}} F_{IM_{t-1}}^{-1} \nabla_\theta L_t(0)$$

it is often approximated by Fisher Information Matrix (FIM) (Liu et al., 2020; Spall, 2005):

$$F_k = E_{p(\hat{D}_k|\theta)}\left[\nabla_\theta \log p(\hat{D}_k|\theta)\nabla_\theta \log p(\hat{D}_k|\theta)^\top\right]\Big|_{\theta=\mu_k} \approx \Lambda(D_k, \mu_k) \tag{1}$$

F_k represents the Fisher Information Matrix, which measures the sensitivity of the parameter θ to the uncertainty during training (Kao et al., 2021). $\nabla_\theta \log p(x|\theta)$ is the gradient of the log-likelihood function with respect to the parameter θ.

the work by (Wang et al., 2022b) demonstrates that this method leads to a tighter upper bound on the generalization gap than independent adapters through $\sqrt{\frac{d\ln(N_t/d)+\ln(1/\delta)}{N_t}}$. See more details in Appendix A.

$$\max_{i\in[1,K]} \sqrt{\frac{d_i \ln(N_{1:t-1}/d_i) + \ln(2K/\delta)}{N_{1:t-1}}} + \sqrt{\frac{d\ln(N_{1:t-1}/d) + \ln(1/\delta)}{N_{1:t-1}}}, \tag{2}$$

$$\max_{i\in[1,K]} \sqrt{\frac{d_i \ln(N_t/d_i) + \ln(2K/\delta)}{N_t}} + \sqrt{\frac{d\ln(N_t/d) + \ln(1/\delta)}{N_t}}. \tag{3}$$

Comparing Eq. **??** and Eq. 3, we conclude that cooperating k adapters facilitates a smaller generalization gap over the new and old tasks.

A.2 Limitations of other methods in handling transfer and interference

L2P (Wang et al., 2022d)applies visual prompt tuning to continual learning by learning a prompt pool to select instance-specific prompts. DualPrompt (Wang et al., 2022c) introduces two types of prompts, namely, general and expert prompts. CODA-Prompt (Smith et al., 2023) further improves the prompt selection process by incorporating an attention mechanism. SimpleCIL (Zhou et al., 2024) freezes the pre-trained weights and extracts the center of each class by averaging the embeddings within the same class, resulting in the most representative pattern of that class. ADAM (Zhou et al., 2024) further advances this approach by comparing the performance of the prototype-based classifier with that of a fully fine-tuned model on new classes.

Accordingly, the loss function for continual learning can typically be defined as:

$$L(\theta) = L_t(\theta) + \lambda \hat{L}_{1:t-1}(\theta), \tag{4}$$

where $\hat{L}_{1:t-1}(\cdot)$ provides the constraint to achieve a proper trade-off between new and old tasks.

Replay-based methods facilitates continual learning by storing and replaying, or generating previously learned samples (Luo et al., 2024; Rebuffi et al., 2017). $\hat{L}_{1:t-1}(\cdot)$ of them is $\sum_{k=1}^{t-1} L_k(\theta; \hat{D}_k)$, where $\hat{D}_k$ is an approximation of D_k through replaying old training samples. These methods achieve continual learning through minimizing $\frac{1}{2(t-1)}\sum_{j=1}^{t-1}\mathrm{Div}(D_j, D_t)$. Although these methods are highly effective, they pay less attention to transferring and have the problem of samples imbalance. This sample imbalance may lead to interference in the performance of previous tasks with fewer replay samples by those with a larger number of replayed samples. Additionally, it can negatively impact the learning of new tasks. Moreover, these approaches can result in uncontrolled expansion of storage and computational resources.

Dynamic network-based methods primarily achieve continual learning by adding new parameters for new tasks to varying degrees while freezing old parameters (Bonato et al., 2024; Yoon et al., 2017; Wang et al., 2022a). Most of PTM-based methods are dynamic network-based methods. $\hat{L}_{1:t-1}(\cdot)$ of them is $\hat{L}_{1:t-1}(\theta = \bigcup_{k=1}^{t-1}\hat{\theta}_k)$. For every task, $\theta = \{\hat{\theta}_{old}, \hat{\theta}_{new}\}$, where $\hat{\theta}_{old}$ decides the extent to which frozen parameters from old tasks are reused varies across methods. In parameter isolation approaches (Yoon et al., 2017), $\hat{\theta}_{old}$ is zero, while in network expansion methods (Wang et al., 2022a), all frozen parameters are reused. These methods primarily aim to minimize the $\sqrt{\frac{d\ln(N_{1:t-1}/d)+\ln(1/\delta)}{N_{1:t-1}}}$ to reduce the upper bound of the loss function. This is because the d when using a shared set of parameters across all tasks is necessarily larger than the dimensionality when

each task has its own dedicated, smaller set of parameters. Although these methods effectively preserve the model's performance on both new and old tasks, they do not facilitate true forward and backward knowledge transfer between tasks during learning. Furthermore, as the number of tasks increases, the network size grows uncontrollably, requiring substantial storage and computational resources.

Both regularization-based and projection matrix approaches achieve continual learning by restricts parameter updates to directions which do not interfere strongly with previous tasks (Kao et al., 2021; Saha et al., 2021; Saha & Roy, 2023; Zeng et al., 2019). The essence of these methods lies in optimizing the model along the flat directions of the prior, which is addressed by focusing on the Eq. **??** mentioned above. Different methods employ varying approaches to approximate the Fisher information matrix (Zeng et al., 2019; Lin et al., 2022; Kirkpatrick et al., 2017; Li & Hoiem, 2017; Yu et al., 2020). As Eq. **??** shows, if inference happens, the learning of new tasks are affected.

A.3 THEORY ANALYSIS OF METHOD

Based on Eq. **??**, we analyze the theoretical effectiveness of our algorithm. First, for $\hat{E}_{D_{1:t-1}}(\theta_{1:t})$, our algorithm shares a set of parameters among tasks that fall within the same flat optimization region and applies a suitable flat direction search method, thereby tightening the upper bound of this term. For the second term, $\frac{1}{2(t-1)}\sum_{j=1}^{t-1}\mathrm{Div}(D_j, D_t)$, since the FIM closely aligns with task similarities, reducing the divergence between them. Finally, the MoE mechanism also reduces the third term. In conclusion, our algorithm effectively tightens the upper bound of the loss function across all three aspects, enabling strong continual learning performance.

A.4 EXPERIMENTS DETAILS

VTAB contains 50 classes, CIFAR100 has 100 classes, CUB, ImageNet-R, ImageNet-A, and ObjectNet each have 200 classes, and OmniBenchmark includes 300 classes. To ensure a fair comparison, we use the same training and testing sets as in (Zhou et al., 2024) for all methods.

Following (Zhou et al., 2024), we use two pre-trained models: ViT-B/16-IN21K and ViT-B/16-IN1K. Both are pre-trained on ImageNet21K, but the latter is further fine-tuned on ImageNet1K.

A.5 THE NUMBER OF ADAPTERS IN DIFFERENT BLOCKS

Table 1: The number of adapters in different blocks of model our proposed method learned during training.

Settings	Number of Adapters											
	1	2	3	4	5	6	7	8	9	10	11	12
CIFAR B0 Inc5	2	2	2	2	2	2	2	2	2	2	2	2
CUB B0 Inc5	2	2	2	2	2	19	16	18	20	20	20	2
IN-R B0 Inc5	3	3	3	3	3	3	3	3	3	4	5	3
IN-A B0 Inc20	2	2	2	2	2	4	2	2	2	5	5	2
ObjNet B0 Inc5	5	6	5	5	5	6	15	12	16	15	18	5
OmniBench B0 Inc30	9	9	10	10	10	6	6	2	2	2	2	2
VTAB B0 Inc10 4	4	4	4	4	4	4	4	4	4	4	4	4

REFERENCES

Jacopo Bonato, Francesco Pelosin, Luigi Sabetta, and Alessandro Nicolosi. Mind: Multi-task incremental network distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11105–11113, 2024.

Ta-Chu Kao, Kristopher Jensen, Gido van de Ven, Alberto Bernacchia, and Guillaume Hennequin. Natural continual learning: success is a journey, not (just) a destination. *Advances in neural information processing systems*, 34:28067–28079, 2021.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

Sen Lin, Li Yang, Deliang Fan, and Junshan Zhang. Trgp: Trust region gradient projection for continual learning. *arXiv preprint arXiv:2202.02931*, 2022.

Jing Liu, Haidong Yuan, Xiao-Ming Lu, and Xiaoguang Wang. Quantum fisher information matrix and multiparameter estimation. *Journal of Physics A: Mathematical and Theoretical*, 53(2): 023001, 2020.

Yutian Luo, Shiqi Zhao, Haoran Wu, and Zhiwu Lu. Dual-enhanced coreset selection with class-wise collaboration for online blurry class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23995–24004, 2024.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017.

Gobinda Saha and Kaushik Roy. Continual learning with scaled gradient projection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 9677–9685, 2023.

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. *arXiv preprint arXiv:2103.09762*, 2021.

James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11909–11919, 2023.

James C Spall. Monte carlo computation of the fisher information matrix in nonstandard settings. *Journal of Computational and Graphical Statistics*, 14(4):889–909, 2005.

Fu-Yun Wang, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Foster: Feature boosting and compression for class-incremental learning. In *European conference on computer vision*, pp. 398–414. Springer, 2022a.

Liyuan Wang, Xingxing Zhang, Qian Li, Jun Zhu, and Yi Zhong. Coscl: Cooperation of small continual learners is stronger than a big one. In *European Conference on Computer Vision*, pp. 254–271. Springer, 2022b.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European Conference on Computer Vision*, pp. 631–648. Springer, 2022c.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 139–149, June 2022d.

Jaehong Yoon, Eunho Yang, Jeongtae Lee, and Sung Ju Hwang. Lifelong learning with dynamically expandable networks. *arXiv preprint arXiv:1708.01547*, 2017.

Lu Yu, Bartlomiej Twardowski, Xialei Liu, Luis Herranz, Kai Wang, Yongmei Cheng, Shangling Jui, and Joost van de Weijer. Semantic drift compensation for class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6982–6991, 2020.

Guanxiong Zeng, Yang Chen, Bo Cui, and Shan Yu. Continual learning of context-dependent processing in neural networks. *Nature Machine Intelligence*, 1(8):364–372, 2019.

Da-Wei Zhou, Zi-Wen Cai, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need. *International Journal of Computer Vision*, pp. 1–21, 2024.