Supplementary: Introducing Language Guidance in Prompt-based Continual Learning

Muhammad Gul Zain Ali Khan[1,2] Muhammad Ferjad Naeem[3] Luc Van Gool[3] Didier Stricker[1,2]
Federico Tombari[4,5] Muhammad Zeshan Afzal[1,2]
[1]RPTU [2]DFKI [3]ETH Zürich [4]TUM [5]Google

1. Experiment Details

We provide experimental details of all experiments in this section. We implement LGCL using PyTorch on Ubuntu 20.0 workstation. We conduct all experiments on A100-40GB GPUs. We use a constant learning rate of 0.005 using Adam Optimizier [3] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We use batch size 24 for DualPrompt [6] and batch size 16 for L2P [7]. We resize all input images to 224×224.

For L2P, we use 50 epochs for Split-ImageNet-R [6] and 5 epochs for CIFAR-100 [4] datasets. For DualPrompt [6] we use 50 epochs for Split-ImageNet-R and 20 epochs for L2P [7]. We set $M = 10$, $L_p = 5$ and $N = 5$ for L2P [7] on Split-CIFAR100 [4] dataset. We set $M = 10$, $L_e = 20$, $L_g = 5$ and $N = 1$ for DualPrompt [6] on Split-CIFAR100 [4] dataset. We set $M = 10$, $L_e = 20$, $L_g = 5$ and $N = 1$ for Dualprompt [6] on Split-ImageNet-R [6] dataset. We set $M = 30$,$L_p = 20$ and $N = 5$ for L2P [7] on Split-ImageNet-R [6]. We use CLIP [5] text encoder "ViT-L/14" for generating language representations (L_t and L_c). Specifically, we use "ViT-L/14" for L2P [7] and "ViT-L14/336px" for DualPrompt [6]. Following [7, 6], we use ViT-B/16 pre-trained backbone.

For Split-ImageNet-R [6], we set 0.3 loss weight for task-level language guidance loss given in Eq. 3 and 0.7 loss weight for class-level language guidance loss given in Eq. 4. For Split-CIFAR100 [4], we set 0.4 loss weight for task-level language guidance loss given in Eq. 3 and 0.6 loss weight for class-level language guidance loss given in Eq. 4. For L2P [7], we set 0.1 loss weight for task-level language guidance loss given in Eq. 3 and 0.9 loss weight for class-level language guidance given in Eq. 4 on Split-CIFAR100 [4] dataset. For Split-ImageNet-R [1], we set 0.5 loss weight for both task-level language guidance loss given in Eq. 3 and class-level language guidance loss given in Eq. 4 with L2P [7]

2. Evaluation Metrics

We compute Average Accuracy at Task t denoted by A_t and Forgetting at Task t denoted by F_t. Let $E_{t,\mathcal{T}}$ be the accuracy of task t when evaluated at task $\mathcal{T}$. Then A_t is shown below:

$$A_t = \frac{1}{t} \sum_{\mathcal{T}}^{t} E_{t,\mathcal{T}}$$

Average accuracy represents an average of the accuracy of all the tasks at a given task t. However, this does not represent how much the model has forgotten from the previous tasks. Forgetting F_t aims to quantify the catastrophic forgetting of Neural Networks in Continual Learning. Forgetting F_t formulation is given below.

$$F_t = \frac{1}{t-1} \sum_{\mathcal{T}}^{t-1} max_{\mathcal{T}' \in \{1,...,t-1\}} (E_{\mathcal{T}',\mathcal{T}} - E_{t,\mathcal{T}})$$

3. Comparison with L2P on Prompt Length and Selection Size

We show ablation on Prompt Length "L" and Selection Size "N" on L2P [7] in Table. 1. For comparison with baseline method L2P [7], we report average accuracies on Split-CIFAR100 [4]. We show that LGCL outperforms baseline method L2P [7] on most configurations. LGCL achieves the highest accuracy at N=5 and L=20. However, the highest difference in performance is with N=20 and L=20 where LGCL outperforms baseline method L2P [7] by 1.51%. LGCL is comparable to baseline method L2P [7] where L2P [7] outperforms LGCL with highest difference of 0.8% at N=1 and L=10. We observe that LGCL is fairly robust to changes in prompt configuration since LGCL outperforms baseline method L2P [7] on most configurations and produces comparable results in configurations where L2P [7] outperforms LGCL. We achieve this consistent performance without any addition in the number of parameters.

N L	1	5	10	20
1	**77.63**/77.01	81.49/**82.49**	82.92/**83.39**	**83.34**/83.15
5	82.24/**82.88**	**83.85**/83.50	**83.90**/83.39	**83.79**/82.84
10	82.48/**83.28**	**83.68**/83.18	**83.52**/83.13	**81.98**/81.84
20	**83.86**/83.40	**84.01**/82.60	**82.65**/81.10	**81.19**/79.65

Table 1: **Ablation on Prompt Length "L" and Selection Size "N" on L2P [7].** We compare LGCL with baseline method L2P [7] on "L" and "N". We report average accuracy on Split-CIFAR100 [4]. We report the average accuracy of L2P [7] and LGCL both. The first result in each cell is LGCL average accuracy followed by a / and then L2P [7] average accuracy. Results for L2P [7] are taken from [7]. Higher average accuracy is in **bold**. We keep prompt pool size "M" constant at 20 and all other hyperparameters. We show that LGCL consistently outperforms baseline method L2P [7] on most configurations.

4. Limitations

LGCL does not explore other modalities since LGCL builds on previous prompting based continual learning methods [7, 6]. LGCL assumes that robust VIT [2] based pre-trained feature extractor is available. LGCL also assumes that a pre-trained text encoder that can generate robust text embeddings of the classes is available during training. Furthermore, LGCL explores language modelling on VIT [2] based networks. We leave the exploration of LGCL and language modelling on conventional ConvNet-based networks as future work. While we have shown that Incremental Class setting of continual learning can benefit from language modelling and LGCL, we leave the exploration of LGCL in other domains of continual learning, such as task-agnostic continual learning, as future work.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 1

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, 2020. 2

[3] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[4] Alex Krizhevsky. Learning multiple layers of features from tiny images. In *UoT*, 2009. 1, 2

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 1

[6] 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. *ECCV*, 2022. 1, 2

[7] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer G. Dy, and Tomas Pfister. Learning to prompt for continual learning. *CVPR*, 2021. 1, 2