Supplementary Materials of Overcoming Forgetting Catastrophe in Quantization-Aware Training

Ting-An Chen[1,2], De-Nian Yang[2,3], Ming-Syan Chen[1,3]
[1]Graduate Institute of Electrical Engineering, National Taiwan University, Taiwan
[2]Institute of Information Science, Academia Sinica, Taiwan
[3]Research Center for Information Technology Innovation, Academia Sinica, Taiwan
tachen@arbor.ee.ntu.edu.tw, dnyang@iis.sinica.edu.tw, mschen@ntu.edu.tw

A. Proof and theoretical analysis

A.1 Approximation of quantization error minimization to the objective of performance optimization

In this study, we consider an additional term, the gradient, to better measure the performance degradation during quantization (described in Sec. 3.1). The following theorem proves that the minimization of the quantization error with a gradient term approximates the objective, minimization of performance degradation, during quantization.

Theorem 3.1 *Let* $(\boldsymbol{x}^T, y)$ *be the (feature, label) of input data,* $\mathscr{L}$ *denote the training loss,* $\boldsymbol{w}$ *indicate the weights, and* $\boldsymbol{w}^q$ *represent the quantized weights. Then* $\arg\min_{\boldsymbol{w}^q} \mathbb{E}(||(\boldsymbol{w}^q - \boldsymbol{w}) \cdot \frac{\partial \mathscr{L}}{\partial \boldsymbol{w}^q}||_2) \simeq \arg\min_{\boldsymbol{w}^q} |\mathscr{L}((\boldsymbol{x}^T, y; \boldsymbol{w}^q) - \mathscr{L}((\boldsymbol{x}^T, y; \boldsymbol{w})|$, *where* $|\mathscr{L}((\boldsymbol{x}^T, y; \boldsymbol{w}^q) - \mathscr{L}((\boldsymbol{x}^T, y; \boldsymbol{w})|$ *is the performance decrement after quantization.*

Proof. By Taylor series expansion [1],

$$\begin{aligned}\mathscr{L}(\mathbf{x}^T, y; \mathbf{w}) &= \sum_{p=0}^{P} \frac{\mathscr{L}^{(p)}(\mathbf{x}^T, y; \mathbf{w}^q)}{p!} ||\mathbf{w}^q - \mathbf{w}||_2^p + R_p(\mathbf{w}^q) \\ &\simeq \sum_{p=0}^{P} \frac{\mathscr{L}^{(p)}(\mathbf{x}^T, y; \mathbf{w}^q)}{p!} ||\mathbf{w}^q - \mathbf{w}||_2^p \\ &= \mathscr{L}(\mathbf{x}^T, y; \mathbf{w}^q) - \frac{\partial \mathscr{L}(\mathbf{x}^T, y; \mathbf{w}^q)}{\partial \mathbf{w}^q} ||\mathbf{w}^q - \mathbf{w}||_2,\end{aligned}$$

which implies

$$\begin{aligned}&|\mathscr{L}((\mathbf{x}^T, y; \mathbf{w}^q) - \mathscr{L}((\mathbf{x}^T, y; \mathbf{w})| \\ &\simeq ||(\mathbf{w}^q - \mathbf{w}) \cdot \frac{\partial \mathscr{L}(\mathbf{x}^T, y; \mathbf{w}^q)}{\partial \mathbf{w}^q}||_2.\end{aligned}$$

Therefore,

$$\begin{aligned}&\arg\min_{\mathbf{w}^q} \mathbb{E}(||(\mathbf{w}^q - \mathbf{w}) \cdot \frac{\partial \mathscr{L}}{\partial \mathbf{w}^q}||_2) \\ &= \arg\min_{\mathbf{w}^q} ||(\mathbf{w}^q - \mathbf{w}) \cdot \frac{\partial \mathscr{L}(\mathbf{x}^T, y; \mathbf{w}^q)}{\partial \mathbf{w}^q}||_2 \\ &\simeq \arg\min_{\mathbf{w}^q} |\mathscr{L}((\mathbf{x}^T, y; \mathbf{w}^q) - \mathscr{L}((\mathbf{x}^T, y; \mathbf{w})|.\end{aligned}$$

□

In Theorem 3.1, according to Taylor series of the training loss function, the change of loss approximates the quantization error with the gradient, i.e., the impact of the change of weight on the training loss. Thus, as illustrated in Sec. 3, $\mathbb{E}(||(\mathbf{w}^q - \mathbf{w}) \cdot \frac{\partial \mathscr{L}}{\partial \mathbf{w}^q}||_2)$ is used to evaluate the performance decrement during quantization.

A.2 Increasing quantization error on old task data in new task learning

In Sec. 4.1, we investigate the forgetting problem in quantization, by analyzing the increment of quantization error on old task data after learning the new tasks.

Theorem 4.1 *Based on Definition 4.1, and denote* $\boldsymbol{w}_s^{q*}$ *as the optimal solution on task s, then* $\xi(\boldsymbol{x}_s^T; \boldsymbol{w}_t^{q*}) \geq \xi(\boldsymbol{x}_s^T; \boldsymbol{w}_s^{q*}), \forall s \leq t$.

Proof. Since

$$\mathbf{w}_s^{q*} = \arg\min_{\mathbf{w}} \xi(\mathbf{x}_s^T; \mathbf{w}),$$

then

$$\xi(\mathbf{x}_s^T; \mathbf{w}_s^{q*}) \leq \xi(\mathbf{x}_s^T; \mathbf{w}), \ \forall \mathbf{w}.$$

Hence,

$$\xi(\mathbf{x}_s^T; \mathbf{w}_s^{q*}) \leq \xi(\mathbf{x}_s^T; \mathbf{w}_t^{q*}), \forall s \leq t.$$

□

According to Theorem 4.1, the quantized weights after learning new tasks converge to a separate solution, instead of the optimum on the old task data. Therefore, the quantization error on old data increases after the update of model weights based on the incoming new task data, demonstrating the forgetting problem in quantization. We will further analyze the increment of quantization error in Theorem 5.1 and derive the upper bound in Appendix A.3.

A.3 Derivation of the upper bound of increasing quantization error in new task learning

In Sec. 5.1, we investigate the issue that incurs the forgetting problem and derive the upper bound of the increment of quantization error as follows.

Theorem 5.1 *Based on Definition 4.1 and Theorem 4.1, the increment of the quantization error is* $\xi(\boldsymbol{x}_s^T; \boldsymbol{w}_t^q) - \xi(\boldsymbol{x}_s^T; \boldsymbol{w}_s^q)$ *which has an upper bound* $\mathbb{E}(||(\boldsymbol{w}_s^q - \boldsymbol{w}_s) \cdot \frac{\partial \mathscr{L}(\boldsymbol{x}_s; \boldsymbol{w}_t^q)}{\partial \boldsymbol{w}_t^q}||_2) + \mathbb{E}(||(\boldsymbol{w}_t^q - \boldsymbol{w}_s^q) \cdot \frac{\partial \mathscr{L}(\boldsymbol{x}_s; \boldsymbol{w}_t^q)}{\partial \boldsymbol{w}_t^q}||_2)$, $\forall s \leq t$.

Proof. The increment of the quantization error is

$$
\begin{aligned}
&\xi(\mathbf{x}_s^T; \mathbf{w}_t^q) - \xi(\mathbf{x}_s^T; \mathbf{w}_s^q) \leq \xi(\mathbf{x}_s^T; \mathbf{w}_t^q) \\
&= \mathbb{E}(||(\mathbf{w}_t^q - \mathbf{w}_s) \cdot \frac{\partial \mathscr{L}(\mathbf{x}_s; \mathbf{w}_t^q)}{\partial \mathbf{w}_t^q}||_2) \\
&= \mathbb{E}(||(\mathbf{w}_t^q - \mathbf{w}_s^q + \mathbf{w}_s^q - \mathbf{w}_s) \cdot \frac{\partial \mathscr{L}(\mathbf{x}_s; \mathbf{w}_t^q)}{\partial \mathbf{w}_t^q}||_2) \\
&= \mathbb{E}(||(\mathbf{w}_t^q - \mathbf{w}_s^q) \cdot \frac{\partial \mathscr{L}(\mathbf{x}_s; \mathbf{w}_t^q)}{\partial \mathbf{w}_t^q} \\
&\quad + (\mathbf{w}_s^q - \mathbf{w}_s) \cdot \frac{\partial \mathscr{L}(\mathbf{x}_s; \mathbf{w}_t^q)}{\partial \mathbf{w}_t^q}||_2) \\
&\leq \mathbb{E}(||(\mathbf{w}_s^q - \mathbf{w}_s) \cdot \frac{\partial \mathscr{L}(\mathbf{x}_s; \mathbf{w}_t^q)}{\partial \mathbf{w}_t^q}||_2) \\
&+ \mathbb{E}(||(\mathbf{w}_t^q - \mathbf{w}_s^q) \cdot \frac{\partial \mathscr{L}(\mathbf{x}_s; \mathbf{w}_t^q)}{\partial \mathbf{w}_t^q}||_2), \ \forall s \leq t.
\end{aligned}
$$

□

Theorem 5.1 indicates that the increment of quantization error is induced from not only the intra-task error $\mathbb{E}(||(\mathbf{w}_s^q - \mathbf{w}_s) \cdot \frac{\partial \mathscr{L}(\mathbf{x}_s; \mathbf{w}_t^q)}{\partial \mathbf{w}_t^q}||_2)$ but also the inter-task error $\mathbb{E}(||(\mathbf{w}_t^q - \mathbf{w}_s^q) \cdot \frac{\partial \mathscr{L}(\mathbf{x}_s; \mathbf{w}_t^q)}{\partial \mathbf{w}_t^q}||_2)$. Existing quantization approaches mainly target minimizing the former, whereas the latter has not been fully investigated. Therefore, in this paper, we focus on minimizing the inter-task quantization error generated in the new task learning to alleviate the forgetting problem. In the forward process, we propose ProxQ to regularize the consistency of the search spaces learned from separate tasks to reduce the space shift incurring the inter-task quantization error (see Sections 5.2.1 and 5.2.2). In the backward process, on the other hand, we reduce the inter-task quantization error obtained from the ProxQ process by a regularization loss (see Sec. 5.2.3).

A.4 Biased prediction result toward new tasks induced by a limited amount of replay data

As described in Sec. 5.3, recent lifelong learning (LL) research is developed to alleviate the forgetting problem in full-precision network training by employing replay data (old task data) retrained in new tasks. However, only a limited amount of old task data can be stored as replay data in quantization-aware training due to the memory constraint, which poses an *imbalance* issue. The prediction result tends to be biased toward the class data that mostly appear in new tasks due to the majority of new data compared to the limited amount of the old task data.

To validate the biased performance under the imbalance issue, in the following theorem, we analyze the change in prediction probability after employing replay data with a replay ratio, i.e., the proportion of training data in old tasks sampled for retraining in new task learning. If the change decreases with the lower replay ratio, the prediction performance with limited replay data is close to the biased result toward the new tasks.

Theorem 5.2 *Let* $\pi_j^{t|s} := P_{Y_t|\{X_t, X_s^{replay}\}}(y_t = j|x_t)$ *stand for the prediction probability of the* t*-th task data* x_t *on the* j*-th class, incorporated with the training of replay data from the* s*-th task* X_s^{replay}*, where* $s < t$*, and* $\pi_j^t := P_{Y_t|X_t}(y_t = j|x_t)$ *represent the prediction probability without replay data. Denote* n_j^t *as the sample size of* X_t *on the* j*-th class and* r_j^s *as the sample size of* X_s^{replay} *on the* j*-th class. If* $r_j^s \leq \delta_j n_j^t$, $\forall j$*, where* $\delta_j \in [0, 1]$ *is the replay ratio on the* j*-th class and* r_j^s*, then* $|\pi_j^{t|s} - \pi_j^t| < \delta_j \cdot (1 + \frac{\sum_{i=1}^K \delta_i n_i^t}{\sum_{i=1}^K n_i^t})$, $\forall j, \forall s < t$.

Proof. First,

$$
\begin{aligned}
\pi_j^{t|s} &= P_{Y_t|\{X_t, X_s^{replay}\}}(y_t = j|x_t) \\
&= \frac{P_{\{X_t, X_s^{replay}\}|Y_t}(x_t|y_t = j)}{P_{\{X_t, X_s^{replay}\}}(x_t)} \cdot P_{\{Y_t, Y_s^{replay}\}}(y_t = j) \\
&= \frac{P_{\{X_t, X_s^{replay}\}|Y_t}(x_t|y_t = j)}{P_{\{X_t, X_s^{replay}\}}(x_t)} \cdot \frac{r_j^s + n_j^t}{\sum_{i=1}^K (r_i^s + n_i^t)},
\end{aligned} \tag{1}
$$

and

$$
\begin{aligned}
\pi_j^t &= P_{Y_t|X_t}(y_t = j|x_t) \\
&= \frac{P_{X_t|Y_t}(x_t|y_t = j)}{P_{X_t}(x_t)} \cdot P_{Y_t}(y_t = j) \\
&= \frac{P_{X_t|Y_t}(x_t|y_t = j)}{P_{X_t}(x_t)} \cdot \frac{n_j^t}{\sum_{i=1}^K n_i^t},
\end{aligned} \tag{2}
$$

which implies that

$$\begin{aligned}&\pi_j^{t|s} - \pi_j^t \\ &= \frac{P_{\{X_t, X_s^{replay}\}|Y_t}(x_t|y_t=j)}{P_{\{X_t, X_s^{replay}\}}(x_t)} \cdot \frac{r_j^s + n_j^t}{\sum_{i=1}^K (r_i^s + n_i^t)} \\ &- \frac{P_{X_t|Y_t}(x_t|y_t=j)}{P_{X_t}(x_t)} \cdot \frac{n_j^t}{\sum_{i=1}^K n_i^t}.\end{aligned} \tag{3}$$

According to Eq. (3),

$$\begin{aligned}&|\pi_j^{t|s} - \pi_j^t| \\ &\leq \max\left\{\frac{P_{\{X_t, X_s^{replay}\}|Y_t}(x_t|j)}{P_{\{X_t, X_s^{replay}\}}(x_t)}, \frac{P_{X_t|Y_t}(x_t|j)}{P_{X_t}(x_t)}\right\} \\ &\cdot \left|\frac{r_j^s + n_j^t}{\sum_{i=1}^K (r_i^s + n_i^t)} - \frac{n_j^t}{\sum_{i=1}^K n_i^t}\right|.\end{aligned} \tag{4}$$

Since $\pi_j^{t|s} \leq 1$, and $\pi_j^t \leq 1$, we can derive from Eq. (1) and Eq. (2) that

$$\frac{P_{\{X_t, X_s^{replay}\}|Y_t}(x_t|y_t=j)}{P_{\{X_t, X_s^{replay}\}}(x_t)} \leq \frac{\sum_{i=1}^K (r_i^s + n_i^t)}{r_j^s + n_j^t}, \tag{5}$$

and

$$\frac{P_{X_t|Y_t}(x_t|y_t=j)}{P_{X_t}(x_t)} \leq \frac{\sum_{i=1}^K n_i^t}{n_j^t}, \tag{6}$$

which implies that

The analytical result in Theorem 5.2 shows that the prediction probability under the training with limited replay data, i.e., a small replay ratio $\delta_j, \forall j$, is close to the result without the employment of replay data. Accordingly, Theorem 5.2 demonstrates that limited replay data tends to cause a biased prediction result toward the new tasks. In other words, the forgetting problem is not solved. To address the imbalance issue, we design a rebalancing strategy for reweighting the influences of data on performance (see Sec. 5.4).

A.5 Derivation of the rebalancing factor in the Balanced Lifelong Learning (BaLL) loss

To rebalance the biased performance analyzed in Theorem 5.2, we design a BaLL loss to reweight the influence of replay data in new task learning (see Sec. 5.4). In the following theorem, we derive the BaLL loss from the approximation of the balanced prediction result.

Theorem 5.3 *Based on Theorem 5.2, suppose there are total K classes in $\{X_t, X_s^{replay}\}$. Let the original prediction loss be $\mathscr{L}_{pred} = -\sum_{j=1}^K \log \pi_j = -\sum_{j=1}^K \log \frac{e^{\phi_j}}{\sum_{k=1}^K e^{\phi_k}}$, where ϕ_j is the prediction result on the j-th class. Assume that ϕ_j under imbalanced class distribution p_j approximates to the balanced result ϕ_j^* after adding a rebalancing term $\log(s_j)$, i.e., $\phi_j + \log(s_j) = \phi_j^*$. Then the balanced loss is $\mathscr{L}_{BaLL} = -\sum_{j=1}^K \log \frac{s_j e^{\phi_j}}{\sum_{k=1}^K s_k e^{\phi_k}}$, where $s_j = e^{\phi_j(\frac{1}{Kp_j}-1)}$, $\forall j$.*

Proof. First, since the rebalanced prediction result is $\phi_j^* = \phi_j + \log(s_j)$, the balanced lifelong learning loss $\mathscr{L}_{BaLL} = -\sum_{j=1}^K \log \frac{e^{\phi_j^*}}{\sum_{k=1}^K e^{\phi_k^*}}$ can be formulated as

$$\begin{aligned}\mathscr{L}_{BaLL} &= -\sum_{j=1}^K \log \frac{e^{\phi_j^*}}{\sum_{k=1}^K e^{\phi_k^*}} \\ &= -\sum_{j=1}^K \log \frac{e^{\phi_j + \log(s_j)}}{\sum_{k=1}^K e^{\phi_k + \log(s_k)}} \\ &= -\sum_{j=1}^K \log \frac{s_j e^{\phi_j}}{\sum_{k=1}^K s_k e^{\phi_k}}.\end{aligned} \tag{7}$$

□

In the following, we derive $s_j = e^{\phi_j(\frac{1}{Kp_j}-1)}, \forall j$ in Theorem 5.3. Here, the prediction is based on the training of replay data. Accordingly, the prediction on the j-th class in t-th task denoted as ϕ_j^t is $P_{Y_t|\{X_t, X_{1:(t-1)}^{replay}\}}(y_t = j|x_t)$, where the notations except $X_{1:(t-1)}^{replay}$ follow Theorem 5.2, and $X_{1:(t-1)}^{replay}$ represents the union of replay data in all of the old tasks, task ID from 1 to $(t-1)$. To simplify the notations, we ignore the scripts of task ID and denote ϕ_j as $P(y=j|x)$ in the subsequent derivations for s_j in Eq. (7).

Proof. We first expand $\phi_j = P(y=j|x)$ as follows.

$$\begin{aligned}\phi_j = P(y=j|x) &= \frac{P(x|y=j)}{P(x)} \cdot P(y=j) \\ &= \frac{P(x|y=j)}{P(x)} \cdot p_j \\ &= \frac{P(x|y=j)}{P(x)} \cdot \frac{r_j + n_j}{\sum_{i=1}^K (r_i + n_i)}, \forall j,\end{aligned} \tag{7}$$

where r_j stands for the total number of employed replay data on j-th class from all old tasks, and n_j represents the number of data on j-th class in the current (new) task[1].

Compared to Eq. (7), the balanced class distribution in expectation is $p_j^* = \frac{1}{K}$, $\forall j = 1, 2, ..., K$. Accordingly, the

[1]Note that if mostly data in class j appear in old tasks, then $n_j \simeq 0$. Moreover, r_j is usually small. According to Eq. (7), the prediction probability on the j-th class, ϕ_j, is minor compared to the probability on other classes, which induces the *imbalance issue* described in Sec. 5.3.

balanced prediction result is

$$\phi_j^* = \frac{P(x|y=j)}{P(x)} \cdot \frac{1}{K}, \ \forall j. \tag{7}$$

Plugging the results in Eq. (7) and Eq. (7) into $\phi_j + \log(s_j) = \phi_j^*$, we obtain s_j as follows.

$$\begin{aligned} s_j &= e^{\phi_j^* - \phi_j} \\ &= e^{\frac{P(x|y=j)}{P(x)} \cdot \frac{1}{K} - \frac{P(x|y=j)}{P(x)} \cdot p_j} \\ &= e^{\frac{P(x|y=j)}{P(x)} \cdot (\frac{1}{K} - p_j)} \\ &= e^{\frac{\phi_j}{p_j} \cdot (\frac{1}{K} - p_j)} \\ &= e^{\phi_j \cdot (\frac{1}{K p_j} - 1)}, \ \forall j. \end{aligned} \tag{7}$$

□

Theorem 5.3 derives the BaLL loss with the rebalancing factor. For the rebalancing strategy, if replay data belongs to a minority class j in new task learning, i.e., a small class distribution p_j, then the weighting of loss increases by the large rebalancing factor. Therefore, the influence of replay data is able to be reweighted (increased), especially for the class data rarely shown in the new tasks.

B. Quantization results of MobileNet-V2 on Office-31 and ResNet-50 on ImageCLEF

Table 2 in the main paper has presented the results of domain-based multi-task quantization, including ResNet-50 on Office-31 and MobileNet-V2 on ImageCLEF, and validated LifeQuant with significant accuracy improvements and forgetting rate reduction, especially on the lightweight architecture, MobileNet-V2, and at low quantization bits.

In this section, we evaluate LifeQuant on more cases, MobileNet-V2 on Office-31 and ResNet-50 on ImageCLEF. Table 1 shows the remarkable performance enhancement for the low-bit MobileNet-V2. For example, the 2-bit MobileNet-V2 on Office-31 under the LifeQuant process in case W $\rightarrow$ D $\rightarrow$ A receives 58% to 66% accuracy improvements and 65% to 74% forgetting rate reduction. In addition, the 3-bit ResNet-50 on ImageCLEF in case B $\rightarrow$ C $\rightarrow$ I under the LifeQuant obtains a 28% to 50% accuracy gain and 38% to 66% forgetting rate reduction. The results validates LifeQuant can effectively alleviate the forgetting problem by regularizing on space shift (see Sec. 5.2) and rebalancing the influence of replay data in new task learning (see Sec. 5.4).

C. Ablation study on the bounds of regularized space under ProxQ

In Sec. 5.1, we pre-define a standard space in $[-\alpha, \alpha]$ to regularize the search space during quantization (see ProxQ in Sec. 5.2). To validate the effectiveness of space regularization, we compare the performances under separate settings of α in Table 2. The results manifest that the space is required to be larger for CIFAR-100 since there are 100 categories, more than Office-31 with 31 categories and ImageCLEF with 12 classes. In other words, the search space is related to the diversity of data. Moreover, when the space is either strongly regularized to a small range, e.g., $\alpha = 0.25$, or without regularized (*w/o space reg.*), the accuracy degrades, and the forgetting rate rises particularly significantly with quantization to lower bits. Therefore, Table 2 demonstrates the effectiveness of space regularization by ProxQ (see Sec. 5.2).

D. Complementary study on the effectiveness of existing LL approaches compared with LifeQuant in multi-task quantization process

Sec. 6.2 and Sec. 7.2 in the main paper have shown the efficacy of LifeQuant, compared with the state-of-the-art quantization processes, in reducing the increasing quantization errors generated with the change in data tasks and alleviating the forgetting problem. In this section, we further evaluate recent lifelong learning (LL) approaches under the quantization scheme adopted by LifeQuant (see Eq. (2) in the main paper). Tables 3 and 4 present the performances of LL compared with LifeQuant under the replay ratio of 20%, i.e., 20% old task data are employed as replay data for retraining. The results demonstrate limited performance improvements in developed LL approaches, since the influence of limited replay data in the new task learning is underestimated, i.e., the imbalance issue (studied in Sec. 5.3 in the main paper). In contrast, LifeQuant obtains superior performance improvements in forgetting rate reduction over the prior LL works, which validates the rebalancing of the influence of limited replay data (see Sec. 5.4 in the main paper) in the alleviation of the forgetting problem. In addition, the effectiveness of the rebalancing strategy under different replay ratios is evaluated in Sec. 7.4 in the main paper.

References

[1] William Feller. An introduction to probability theory and its applications. Technical report, Wiley series in probability and mathematical statistics, 3rd edn.(Wiley, New ..., 1967.

Table 1: Quantization results of MobileNet-V2 on Office-31 and ResNet-50 on ImageCLEF. γ represents the ratio of class data changes when the task switches. Both model weights and activations are quantized to low bits. The symbol * indicates failed prediction. The improvements over 5% (10%) are presented in **blue** (**red**).

Metrics	Methods	MobileNet-V2 on Office-31				ResNet-50 on ImageCLEF			
		A → D → W		W → D → A		I → P → C		B → C → I	
		3 bit	2 bit	4 bit	3 bit	4 bit	3 bit	4 bit	2 bit
Accuracy (%)	LSQ [15]	*	*	*	*	33.34	*	33.34	*
	LLSQ [23]	38.62	23.82	9.71	8.24	50.28	48.33	46.39	46.11
	Qimera [17]	*	*	*	*	*	*	*	*
	IntraQ [29]	26.54	13.79	2.58	2.58	66.39	36.67	67.50	35.83
	AlignQ [24]	44.73	30.45	11.95	10.36	75.00	62.78	70.56	24.17
	LifeQuant (Ours)	**86.36**	**47.10**	**76.96**	**70.71**	**81.11**	**80.27**	**77.00**	**74.66**
Forgetting (%)	LSQ [15]	*	*	*	*	36.69	*	59.39	*
	LLSQ [23]	58.95	60.19	87.46	90.17	28.16	28.44	43.37	39.16
	Qimera [17]	*	*	*	*	*	*	*	*
	IntraQ [29]	71.91	77.27	97.20	97.03	15.14	33.94	18.98	53.57
	AlignQ [24]	51.90	48.22	87.02	88.32	9.77	21.10	13.67	67.73
	LifeQuant (Ours)	**8.00**	**21.48**	**12.45**	**21.33**	**5.50**	**6.88**	**5.36**	**1.48**

Table 2: Effectiveness of ProxQ under the regularization on quantization search space in $[-\alpha, \alpha]$ (introduced in Sec. 5.2). The symbol * indicates failed prediction. The case with the best performance is presented in **bold text**. The case with the most significant performance degradation is presented with the upper script †. The method *w/o space reg.* indicates quantization without the proximal regularization (see Sec. 5.2).

Metrics	Methods	ResNet-20 on CIAFAR-100 ($\gamma = 25$)		MobileNet-V2 on Office-31 (W → D → A)		ResNet-50 on ImageCLEF (B → C → I)	
		4 bit	2 bit	4 bit	3 bit	4 bit	3 bit
Accuracy (%)	$\alpha = 0.25$	47.91	21.47†	74.04†	39.72	78.33	75.00
	$\alpha = 0.50$	49.66	41.44	76.23	64.23	**78.61**	**76.39**
	$\alpha = 1.00$	50.15	45.81	**76.96**	**70.71**	78.33	76.11
	$\alpha = 1.50$	**50.16**	**46.38**	76.23	69.95	77.00	74.89
	w/o space reg.	46.39†	30.49	74.21	36.20†	67.22†	*†
Forgetting (%)	$\alpha = 0.25$	25.77	62.95†	15.66	54.67	3.93	4.78
	$\alpha = 0.50$	23.11	28.79	14.68	28.19	**3.29**	**2.90**
	$\alpha = 1.00$	22.38	21.34	**12.45**	**21.33**	3.74	3.46
	$\alpha = 1.50$	**22.37**	**20.41**	13.95	22.17	5.36	5.05
	w/o space reg.	28.18†	47.52	16.17†	58.10†	18.07†	*†

Table 3: Quantization performance of LL on CIFAR-100 and Office-31. γ represents the ratio of class data changes when the task switches. Both model weights and activations are quantized to low bits. The symbol * indicates failed prediction. The improvements over 5% (10%) are presented in **blue** (**red**).

Metrics	Methods	ResNet-20 on CIFAR-100				ResNet-50 on Office-31				MobileNet-V2 on Office-31			
		$\gamma = 25$		$\gamma = 50$		A $\rightarrow$ D $\rightarrow$ W		W $\rightarrow$ D $\rightarrow$ A		A $\rightarrow$ D $\rightarrow$ W		W $\rightarrow$ D $\rightarrow$ A	
		4 bit	2 bit	4 bit	2 bit	4 bit	2 bit	4 bit	2 bit	3 bit	2 bit	4 bit	3 bit
Accuracy (%)	EWC [31]	46.05	*	23.56	16.76	*	*	*	*	33.98	9.41	12.22	14.29
	SI [32]	45.60	33.75	28.36	19.81	41.77	29.30	17.61	9.56	32.67	10.90	12.94	11.83
	MAS [33]	42.09	29.03	31.47	13.26	40.17	32.17	18.32	5.98	33.64	11.63	12.09	12.67
	RWalk [34]	44.88	34.13	24.50	19.10	40.56	28.81	18.60	8.14	33.30	9.39	12.69	11.84
	SCP [37]	45.01	28.23	16.73	8.08	36.50	34.18	16.51	8.78	30.44	10.77	9.93	13.13
	PFR [38]	45.31	34.02	31.38	19.75	40.06	30.95	25.47	8.27	35.30	9.81	14.50	12.17
	LifeQuant (Ours)	**50.15**	**46.21**	**36.94**	**34.32**	**48.54**	**46.54**	**25.65**	**17.23**	**86.36**	**47.10**	**76.96**	**70.71**
Forgetting (%)	EWC [31]	28.71	*	63.39	70.84	*	*	*	*	64.13	84.22	86.41	83.27
	SI [32]	28.76	25.60	56.23	64.83	47.73	64.41	72.34	79.30	65.50	81.93	85.42	85.92
	MAS [33]	34.96	26.01	49.31	76.82	49.51	60.32	71.93	87.39	64.59	80.64	86.45	84.97
	RWalk [34]	29.89	24.94	73.11	66.08	49.65	65.03	70.41	83.06	64.94	84.49	85.60	85.95
	SCP [37]	31.26	23.92	49.52	87.52	57.20	55.64	78.37	84.43	67.24	81.61	85.98	84.95
	PFR [38]	29.21	24.95	49.52	64.99	50.28	61.89	60.65	81.58	62.86	83.60	83.70	85.36
	LifeQuant (Ours)	**22.38**	**16.08**	**42.99**	**41.63**	**38.96**	**42.59**	**58.63**	**63.90**	**8.00**	**21.48**	**12.45**	**21.33**

Table 4: Quantization performance of LL on ImageCLEF. Both model weights and activations are quantized to low bits. The symbol * indicates failed prediction. The improvements over 5% (10%) are presented in **blue** (**red**).

Metrics	Methods	ResNet-50 on ImageCLEF				MobileNet-V2 on ImageCLEF			
		I $\rightarrow$ P $\rightarrow$ C		B $\rightarrow$ C $\rightarrow$ I		I $\rightarrow$ P $\rightarrow$ C		B $\rightarrow$ C $\rightarrow$ I	
		4 bit	3 bit	4 bit	2 bit	4 bit	3 bit	4 bit	3 bit
Accuracy (%)	EWC [31]	68.33	43.89	68.33	45.56	73.05	51.67	69.44	47.50
	SI [32]	68.89	44.17	66.11	41.67	72.50	50.00	66.39	48.33
	MAS [33]	66.11	45.28	65.55	43.61	73.88	48.06	72.78	42.78
	RWalk [34]	67.78	44.44	66.94	42.50	73.05	48.89	71.95	43.06
	SCP [37]	69.17	43.89	64.17	36.94	74.72	43.61	70.55	40.56
	PFR [38]	65.56	44.17	66.67	36.95	72.22	48.05	69.72	46.95
	LifeQuant (Ours)	**81.11**	**80.27**	**77.00**	**74.66**	**77.50**	**76.38**	**76.39**	**76.11**
Forgetting (%)	EWC [31]	18.73	49.71	16.55	41.01	15.20	40.38	11.74	39.08
	SI [32]	20.95	48.77	19.66	45.92	14.22	42.29	14.67	37.23
	MAS [33]	24.06	47.56	20.53	42.88	15.20	44.46	7.24	45.21
	RWalk [34]	22.01	48.73	18.54	45.00	13.73	43.61	8.28	44.34
	SCP [37]	20.35	49.57	21.87	51.46	14.22	49.67	10.14	48.15
	PFR [38]	24.35	49.41	19.55	51.89	15.69	44.54	10.40	39.37
	LifeQuant (Ours)	**5.50**	**6.88**	**5.36**	**1.48**	**10.46**	**7.71**	**2.30**	**1.63**