Robust Machine Unlearning for Quantized Neural Networks via Adaptive Gradient Reweighting with Similar Labels

Supplementary Material

The organization of the appendix is as follows:

- Appendix A: Additional Implementation Details;
- Appendix B: ResNet18 on Tiny-Imagenet dataset;
- Appendix C: ResNet18 under Class-wise Forgetting;
- Appendix D: MobileNetV2 on CIFAR-10, CIFAR-100, SVHN datasets;
- Appendix E: Efficiency Analysis.

A. Additional Implementation Details

We adopte the experimental settings of SalUn and ℓ_1-sparse for the baseline methods. All experiments use the SGD optimizer. For FT and RL, we train for 10 epochs within the interval [1e-3, 1e-1]. For GA, we train for 5 epochs with learning rate within the interval [1e-5, 1e-3]. For IU, we explore the parameter α associated with the woodfisher Hessian Inverse approximation within the range [1, 20]. For ℓ_1-sparse, a learning rate search for the parameter γ is executed within the range [1e-6, 1e-4], while searching for the learning rate within the range [1e-3, 1e-1]. For SalUn , we train for 10 epochs with learning rates in the range [5e-4, 5e-2] and sparsity ratios in the range [0.1, 0.9]. For Q-MUL, we train the unlearned model using the SGD optimizer with a batch size of 256. In the random data forgetting scenario, for ResNet-18, we train for 10 epochs with learning rates in the range [1e-3, 1e-1]. For MobileNetV2, train for 10 epochs with learning rates in the range [1e-2, 1e-1]. In the classwise forgetting scenario, for ResNet-18, we train for 10 epochs with learning rates in the range [1e-3, 1e-1]. All experiments are conducted on a single NVIDIA RTX 4090 GPU.

B. ResNet18 on Tiny-Imagenet dataset

Table 4 shows the experimental results of the quantized ResNet18 on the larger dataset Tiny-ImageNet. On the larger dataset, Q-MUL can also demonstrate superior performance. Specifically, the gaps between Q-MUL and Retrain in the four metrics FA, RA, TA, and MIA are 9.95%, 7.47%, 6.32%, and 8.95%, respectively. Compared with other methods, the average gap is as low as 8.17%. This indicates that Q-MUL can also exhibit excellent performance on the larger datasets.

C. ResNet18 under Class-wise Forgetting

Tables 5 and 6 show the experiments of the quantized ResNet18 under Class-wise forgetting scenario. We conduct experiments on two datasets: CIFAR-10 and CIFAR-100. For the class-wise forgetting scenario, RL, ℓ_1-sparse, SalUn, and Q-MUL all perform very close to Retrain. Specifically, on the CIFAR-10, the average gaps of these methods from Retrain are 0.13%, 0.08%, 0.10%, and 0.11% respectively. On the CIFAR-100, the average gaps of these methods from Retrain are 0.76%, 0.55%, 0.97%, and 0.32%, respectively.

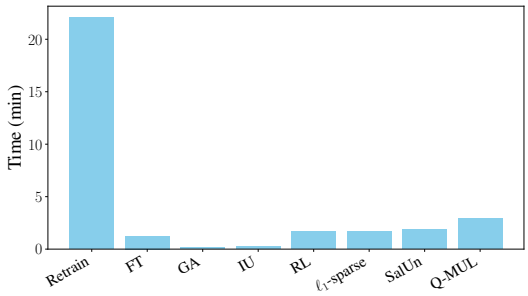

Figure 4. Efficiency Analysis. The experimental scenario involves MobileNetV2 performing random data forgetting (10%) on CIFAR-100.

D. MobileNetV2 on CIFAR-10, CIFAR-100, SVHN datasets

We present the experiments of the quantized MobileNetV2 on CIFAR-10, CIFAR-100, and SVHN in Tables 7, 8, and 9. For MobileNetV2 on CIFAR-10, Q-MUL achieves the best or second-best performance. For MobileNetV2 on CIFAR-100, the average gaps of Q-MUL from Retrain at data forgetting ratios of 10%, 30%, and 50% are 6.52%, 9.07%, and 10.53%, respectively. Compared to the previous state-of-the-art method, these gaps are reduced by 2.77%, 3.19%, and 0.70% respectively. For MobileNetV2 on SVHN, at data forgetting ratios of 10%, 30%, and 50%, the average gaps of Q-MUL from Retrain are 0.92%, 1.73%, and 3.49% respectively.

E. Efficiency Analysis

We conduct an efficiency analysis of various unlearning methods. As shown in Figure 4, Retrain consumes a significant amount of time, requiring 22.05 minutes. Although GA and IU have minimal time overhead, these two methods perform extremely poorly on quantized models during unlearning, making them almost unusable. Q-MUL, due to

Method	Tiny-Imagenet				
	FA	RA	TA	MIA	AG↓
Retrain	62.15	99.46	62.65	51.51	0
FT	74.45(12.30)	91.99(7.47)	56.57(6.08)	33.67(17.84)	10.92
GA	92.95(30.8)	93.01(6.45)	57.19(5.46)	11.77(39.74)	20.61
IU	0.45(61.70)	0.51(98.95)	0.50(62.15)	0.45(51.06)	68.47
RL	60.05(2.10)	82.60(16.86)	51.59(11.06)	45.13(6.38)	9.10
ℓ_1-sparse	60.89(1.26)	75.44(24.02)	56.41(6.24)	46.94(4.57)	<u>9.02</u>
SalUn	65.28(3.13)	81.70(17.76)	51.77(10.88)	37.08(14.43)	11.55
Q-MUL	72.10(9.95)	91.99(7.47)	56.53(6.32)	42.56(8.95)	**8.17**

Table 4. Performance of various MU methods for ResNet18 with 4-bit quantization on Tiny-Imagenet. The unlearning scenario is random data forgetting(10%). **Bold** indicates the best performance and <u>underline</u> indicates the runner-up.

Method	CIFAR-10				
	FA	RA	TA	MIA	AG↓
Retrain	0.00	99.99	93.34	100.0	0
FT	62.40(62.40)	99.95(0.04)	93.59(0.25)	94.67(5.33)	17.02
GA	11.20(11.20)	92.28(7.71)	85.49(7.85)	92.13(7.87)	8.66
IU	0.00(0.00)	29.75(70.24)	28.92(64.42)	100.0(0.00)	33.67
RL	0.00(0.00)	99.87(0.12)	93.74(0.40)	100.0(0.00)	0.13
ℓ_1-sparse	0.00(0.00)	99.85(0.14)	93.51(0.17)	100.0(0.00)	**0.08**
SalUn	0.04(0.04)	99.84(0.15)	93.56(0.22)	100.0(0.00)	<u>0.10</u>
Q-MUL	0.00(0.00)	99.69(0.30)	93.21(0.13)	100.0(0.00)	0.11

Table 5. Performance of various MU methods for ResNet18 with 4-bit quantization on CIFAR-10. The unlearning scenario is class-wise forgetting (one class is forgotten.). **Bold** indicates the best performance and <u>underline</u> indicates the runner-up.

Method	CIFAR-100				
	FA	RA	TA	MIA	AG↓
Retrain	0.00	99.96	70.63	100.0	0
FT	7.11(7.11)	99.79(0.27)	71.10(0.47)	100.0(0.00)	1.96
GA	37.56(37.56)	88.69(11.27)	61.75(8.88)	79.78(20.22)	19.48
IU	0.00(0.00)	72.25(17.71)	51.22(19.41)	100.0(0.00)	9.28
RL	0.89(0.89)	99.97(0.01)	72.78(2.15)	100.0(0.00)	0.76
ℓ_1-sparse	0.00(0.00)	98.33(1.63)	70.06(0.57)	100.0(0.00)	<u>0.55</u>
SalUn	1.33(1.33)	99.92(0.04)	73.12(2.49)	100.0(0.00)	0.97
Q-MUL	0.00(0.00)	99.50(0.46)	69.79(0.84)	100.0(0.00)	**0.32**

Table 6. Performance of various MU methods for ResNet18 with 4-bit quantization on CIFAR-100. The unlearning scenario is class-wise forgetting (one class is forgotten). **Bold** indicates the best performance and <u>underline</u> indicates the runner-up.

the need to calculate the gradients of the forgotten dataset and the retained dataset on the loss, has a slightly larger time overhead compared to methods like SalUn and RL, but the unlearning effect is significantly improved. Q-MUL achieves a trade-off between unlearning effectiveness and time overhead.

Method	CIFAR-10				
	FA	RA	TA	MIA	AG↓
The proportion of forgotten data samples to all samples is 10%					
Retrain	85.97	94.39	85.49	18.60	0
FT	93.64 (7.67)	95.42 (1.03)	87.37 (1.88)	9.62 (8.98)	4.89
GA	93.97 (8.00)	94.87 (0.48)	87.21 (1.72)	8.49 (10.11)	5.08
IU	13.13(72.84)	14.28(80.11)	13.97(71.52)	84.51(66.91)	75.85
RL	85.82 (0.15)	87.46 (6.93)	84.05 (1.44)	17.27 (1.33)	**2.46**
ℓ_1-sparse	93.17 (7.20)	95.15 (0.76)	87.66 (2.17)	9.89 (8.71)	4.71
SalUn	90.87(4.90)	92.08(2.31)	86.63(1.14)	15.71(2.89)	2.81
Q-MUL	90.71(4.74)	93.60(0.79)	87.51(2.02)	15.00(3.60)	<u>2.79</u>
Method	CIFAR-10				
	FA	RA	TA	MIA	AG↓
The proportion of forgotten data samples to all samples is 30%					
Retrain	84.87	93.87	84.19	19.90	0
FT	91.53 (6.66)	93.17 (0.70)	86.21 (2.02)	12.07 (7.83)	4.30
GA	89.16 (4.29)	89.40 (4.47)	82.79 (1.40)	15.20 (4.70)	3.72
IU	12.67(72.20)	13.11(80.76)	12.65(71.54)	86.47(66.57)	72.77
RL	88.04 (3.17)	88.86 (5.01)	84.26 (0.07)	21.81(1.91)	<u>2.54</u>
ℓ_1-sparse	92.90 (8.03)	94.53 (0.66)	86.80 (2.61)	10.76 (9.14)	5.11
SalUn	95.20(10.33)	96.83(2.96)	90.49(6.30)	16.87(3.03)	5.66
Q-MUL	87.56(2.69)	89.24(4.63)	85.63(1.44)	18.56(1.34)	**2.53**
Method	CIFAR-10				
	FA	RA	TA	MIA	AG↓
The proportion of forgotten data samples to all samples is 50%					
Retrain	82.27	94.76	82.17	22.17	0
FT	89.50 (7.23)	91.26 (3.50)	84.32 (2.15)	13.05 (9.12)	5.5
GA	75.99 (6.28)	76.07 (18.69)	72.07 (10.10)	24.30 (2.13)	9.3
IU	21.89(60.38)	22.03(72.73)	21.67(60.50)	78.52(56.35)	62.49
RL	86.81 (4.54)	87.85 (6.91)	83.70(1.53)	17.98 (4.19)	<u>4.29</u>
ℓ_1-sparse	6.64 (11.09)	95.70 (0.94)	87.11 (4.94)	10.00 (12.17)	7.29
SalUn	87.98(5.71)	88.95(5.81)	84.56(2.39)	15.04(7.13)	5.26
Q-MUL	87.08(4.81)	89.48(5.28)	84.76(2.59)	22.24(0.07)	**3.19**

Table 7. Performance of various MU methods for MobileNetV2 with activations kept at full precision and weights quantized to 2 bits on CIFAR-10. The unlearning scenario is random data forgetting. **Bold** indicates the best performance and <u>underline</u> indicates the runner-up. A performance gap against Retrain is provided in (•).

Method	CIFAR-100				
	FA	RA	TA	MIA	AG↓
The proportion of forgotten data samples to all samples is 10%					
Retrain	61.53	89.44	60.90	40.20	0
FT	82.89 (21.36)	88.18 (1.26)	63.56 (2.66)	19.11 (21.09)	11.59
GA	83.82 (22.29)	85.20 (4.24)	61.80 (0.90)	17.33 (22.87)	12.58
IU	2.40(59.13)	2.81(86.63)	2.76(58.14)	2.62(37.58)	60.37
RL	81.53 (20.00)	88.41 (1.03)	63.26 (2.36)	26.38 (13.82)	9.30
ℓ_1-sparse	81.31 (19.78)	85.97 (3.47)	62.33 (1.43)	19.98 (20.22)	11.23
SalUn	82.22(20.69)	88.15(1.29)	63.26(2.36)	27.38(12.82)	<u>9.29</u>
Q-MUL	69.96(8.43)	80.25(9.19)	62.29(1.39)	33.15(7.05)	**6.52**
Method	CIFAR-100				
	FA	RA	TA	MIA	AG↓
The proportion of forgotten data samples to all samples is 30%					
Retrain	53.90	83.38	54.97	46.46	0
FT	79.36 (25.46)	86.34 (2.96)	62.04 (7.07)	22.03 (24.43)	14.98
GA	66.02 (12.12)	66.34 (17.04)	51.44 (3.53)	24.67 (21.79)	13.62
IU	6.68(47.22)	7.30(76.08)	6.75(48.22)	94.62(48.16)	54.92
RL	77.16 (23.26)	83.49(0.11)	61.31 (6.34)	25.43(21.03)	12.69
ℓ_1-sparse	82.53 (28.63)	88.77 (5.39)	63.30 (8.33)	19.75 (26.71)	17.27
SalUn	76.26(22.36)	83.46(0.08)	61.19(6.22)	26.08(20.38)	<u>12.26</u>
Q-MUL	67.77(13.87)	79.60(3.78)	61.99(7.02)	34.87(11.59)	**9.07**
Method	CIFAR-100				
	FA	RA	TA	MIA	AG↓
The proportion of forgotten data samples to all samples is 50%					
Retrain	49.20	85.03	50.21	51.44	0
FT	83.76 (34.56)	91.24 (6.21)	63.92 (13.71)	20.10(31.34)	18.65
GA	41.20 (8.00)	40.60 (44.43)	33.94 (16.27)	46.70 (5.44)	18.54
IU	1.40(47.80)	1.51(83.52)	1.26(48.95)	22(29.44)	52.43
RL	75.36 (26.16)	80.92(4.11)	60.57(10.36)	33.59(17.85)	14.62
ℓ_1-sparse	18.69 (19.78)	85.97 (3.47)	62.33 (1.43)	19.98 (20.22)	<u>11.23</u>
SalUn	73.56(24.36)	79.50(5.53)	59.31(9.10)	34.78(16.66)	13.91
Q-MUL	67.93(18.73)	78.83(6.20)	60.38(10.17)	44.41(7.03)	**10.53**

Table 8. Performance of various MU methods for MobileNetV2 with activations kept at full precision and weights quantized to 2 bits on CIFAR-100. The unlearning scenario is random data forgetting. **Bold** indicates the best performance and <u>underline</u> indicates the runner-up. A performance gap against Retrain is provided in (•).

Method	SVHN				
	FA	RA	TA	MIA	AG↓
The proportion of forgotten data samples to all samples is 10%					
Retrain	94.25	99.99	94.17	9.33	0
FT	99.12 (4.87)	99.98 (0.01)	94.71 (0.54)	2.18 (7.15)	3.14
GA	99.21 (4.96)	99.43 (0.56)	94.84 (0.67)	10.77 (1.44)	<u>1.91</u>
IU	98.65 (4.40)	98.83 (1.16)	93.96 (0.21)	2.43 (6.90)	3.17
RL	96.81 (2.56)	98.36 (1.63)	94.46 (0.29)	20.20 (10.87)	3.84
ℓ_1-sparse	99.07 (4.82)	99.97 (0.02)	94.67 (0.50)	2.68 (6.65)	3.00
SalUn	96.42 (2.17)	98.19 (1.80)	94.51 (0.34)	20.39 (11.06)	3.84
Q-MUL	96.47 (2.22)	99.48 (0.51)	94.96 (0.79)	9.16 (0.17)	**0.92**
Method	SVHN				
	FA	RA	TA	MIA	AG↓
The proportion of forgotten data samples to all samples is 30%					
Retrain	93.12	100.0	93.81	11.21	0
FT	99.28 (6.16)	99.98 (0.02)	94.72 (0.91)	2.13 (9.08)	4.04
GA	99.27 (6.15)	99.46 (0.54)	94.77 (0.96)	1.08 (10.13)	4.45
IU	98.63 (5.41)	98.83 (1.17)	93.05 (0.76)	2.92 (8.29)	3.91
RL	94.55 (1.43)	96.99 (3.01)	93.79 (0.02)	22.33 (11.12)	<u>3.90</u>
ℓ_1-sparse	99.21 (6.09)	99.96 (0.04)	94.81 (1.00)	2.53 (8.68)	3.95
SalUn	95.33 (2.21)	96.75 (3.25)	93.88 (0.07)	25.44 (14.23)	4.94
Q-MUL	96.80 (3.68)	99.46 (0.54)	94.95 (1.14)	12.77 (1.56)	**1.73**
Method	SVHN				
	FA	RA	TA	MIA	AG↓
The proportion of forgotten data samples to all samples is 50%					
Retrain	92.57	95.83	93.19	12.46	0
FT	99.26 (6.69)	99.98 (4.15)	94.59 (1.40)	2.03 (10.43)	5.67
GA	99.31 (6.74)	99.48 (3.65)	94.75 (1.56)	10.53 (1.93)	<u>3.47</u>
IU	97.82 (5.25)	97.87 (2.04)	92.90 (0.29)	3.92 (8.54)	4.03
RL	94.23 (1.66)	95.36 (0.47)	93.05 (0.14)	32.43 (19.97)	5.56
ℓ_1-sparse	99.27 (6.70)	99.97 (4.14)	94.57 (1.38)	2.44 (10.02)	5.56
SalUn	93.93 (1.36)	95.02 (0.81)	92.87 (0.32)	33.78 (21.32)	5.95
Q-MUL	95.30 (2.73)	98.50 (2.67)	94.31 (1.12)	19.90 (7.44)	**3.49**

Table 9. Performance of various MU methods for MobileNetV2 with activations kept at full precision and weights quantized to 2 bits on SVHN. The unlearning scenario is random data forgetting. **Bold** indicates the best performance and <u>underline</u> indicates the runner-up. A performance gap against Retrain is provided in (•).