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

Yujia Tong, Yuze Wang, Jingling Yuan, Chuang Hu

武漢理工大學 WUHAN UNIVERSITY OF TECHNOLOGY
澳門大學 UNIVERSIDADE DE MACAU UNIVERSITY OF MACAU

Background and Motivation

Machine Unlearning

Right to be forgotten
(e.g., GDPR)

Remove influence of specific data from trained models

Model Quantization

Reduces bit-width of weights/activations

Entables deployment on edge devices

How to eliminate the influence of specific data samples on quantized models?

problem

- Existing MU methods are designed for **full-precision models**
- **Quantized models** are more sensitive to noise and gradient imbalance
- Directly applying MU methods → **performance drops**

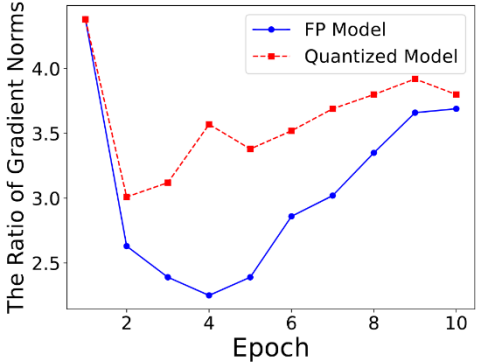
goals

- Design a **quantization-aware unlearning method**
- Maintain **efficiency** and **effectiveness** on quantized models

Chellenges

Challenge 1: Label Noise in Data Processing

- Random label assignment introduces **inconsistent noise**
- Quantized models are **less robust** to such perturbations

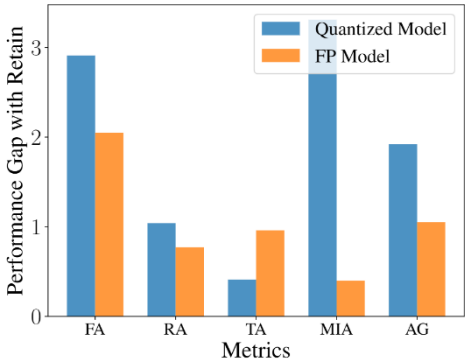


Challenge 2: Gradient Imbalance in Training

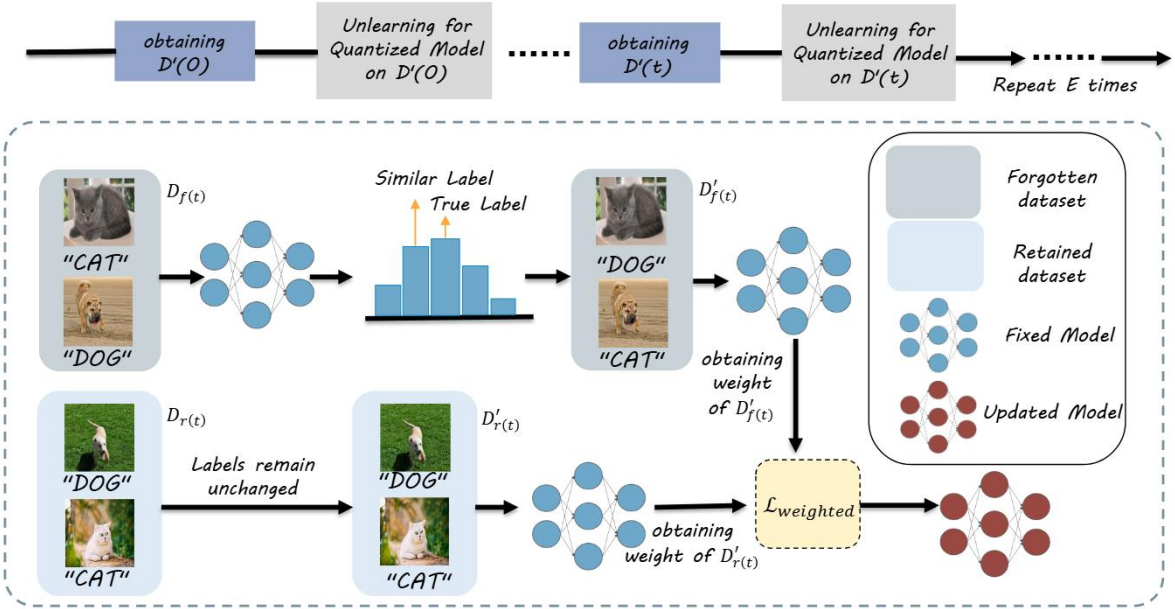
- Forgotten vs. retained data contribute **unevenly** to updates
- Amplified in quantized models due to **discrete parameter space**

Result:

- Poor unlearning performance
- Model instability and residual information leakage



Q-MUL: Quantization-aware Unlearning framework



Two Key Innovations:

- Similar Labels (SL): Replace random labels with semantically close labels
- Adaptive Gradient Reweighting (AGR): Balance contributions from forgotten vs. retained data

Solution

Algorithm 1 Similar Labels

Input: Original complete dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$, The quantized model $\mathcal{M}_t$ with parameters w_t at epoch t
Output: Training dataset D' for unlearning

initialize $D_f' = \emptyset$, $D_r' = \emptyset$;
for $i \in [1, ..., N]$ **do**
  if $(x_i, y_i) \in D_f$ **then**
    Calculate predicted probability $p_{\mathbf{Q}}(x_i; w)$ by eq.(3);
    Calculate distance $d(k)$ following eq.(4);
    Calculate class label k_{sl} by eq.(5) ;
    Add (x_i, k_{sl}) to D_f';
  else
    Add (x_i, y_i) to D_r';
  end if
end for
D'=$D_r' \cup D_f'$;
return D'

Similar Labels

- Instead of random labels → use **closest predicted class** as new label
- Reduces **gradient conflict** and **noise injection**

Adaptive Gradient Reweighting

- Balance **gradient contributions**
- Avoid **over-adjustment** or **under-unlearning**

Key step:

- Compute expected gradient norms:

$$G_f = \mathbb{E}_{(x,y)\sim D'_f}\left[\|\nabla_w \mathcal{L}_{\text{forget}}(x, y; w)\|_2\right]$$
$$G_r = \mathbb{E}_{(x,y)\sim D'_r}\left[\|\nabla_w \mathcal{L}_{\text{retain}}(x, y; w)\|_2\right]$$

- Assign weights: $\alpha_f = \frac{G_r}{G_f + G_r}, \quad \alpha_r = \frac{G_f}{G_f + G_r}$

Algorithm 2 The overall pipeline of Q-MUL

Input: The original quantized model $\mathcal{M}_o$ with parameters w_0 , Original complete dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$, Total training epochs T
Output: The unlearned quantized model $\mathcal{M}_u$ with parameters w_u

for $t \in [0, \ldots, T-1]$ **do**
  Process the original dataset D to obtain the training dataset D' following Algorithm 1;
  Calculate gradient norms G_f by eq.(11);
  Calculate gradient norms G_r by eq.(12);
  Calculate weight of D'_r and D'_f by eq.(13);
  Update the parameters w_t of quantized model $\mathcal{M}_t$ by eq.(14);
end for
return Unlearned quantized model $\mathcal{M}_u$ with parameters w_u

Experiments

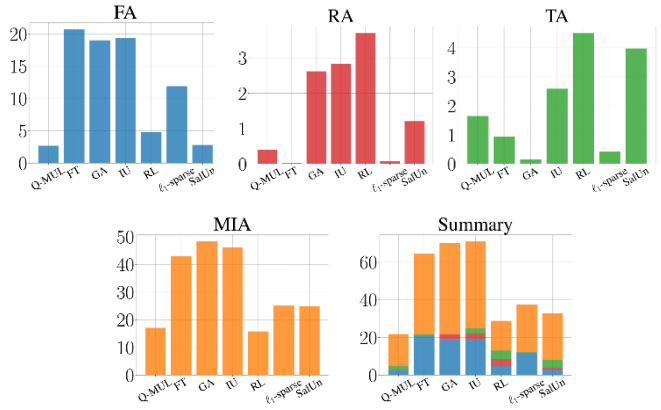
Random Data Forgetting Scenario

Method	CIFAR-10					CIFAR-100				
	FA	RA	TA	MIA	AG↓	FA	RA	TA	MIA	AG↓
The proportion of forgotten data samples to all samples is 10%										
Retrain	93.38	100.0	92.64	13.36	0	74.76	99.98	72.43	56.36	0
FT	99.42(6.04)	99.96(0.04)	93.05(0.41)	2.82(10.54)	4.26	91.82(17.06)	94.12(5.86)	65.29(7.14)	16.69(39.67)	17.43
GA	99.38(6.00)	99.38(0.62)	93.35(0.71)	1.22(12.14)	4.87	95.02(20.26)	96.86(3.12)	71.84(0.59)	9.56(46.80)	17.69
IU	95.64(2.26)	95.53(4.47)	87.85(4.79)	7.93(5.43)	4.24	3.69(71.07)	4.03(95.95)	3.85(68.58)	98.8(42.44)	69.51
RL	96.29(2.91)	98.96(1.04)	92.23(0.41)	16.67(3.31)	1.92	68.51(6.25)	98.91(1.07)	69.47(2.96)	85.62(29.26)	9.89
ℓ_1-sparse	98.71(5.33)	99.88(0.12)	92.92(0.28)	5.29(8.07)	3.45	88.20(13.44)	99.65(0.33)	70.80(1.63)	29.64(26.72)	10.53
SalUn	96.82(3.44)	99.45(0.55)	92.32(0.32)	14.27(0.91)	1.31	82.22(7.46)	98.71(1.27)	67.38(5.05)	66.78(10.42)	6.05
Q-MUL	97.11(3.73)	99.67(0.33)	92.39(0.25)	12.49(0.87)	**1.30**	75.71(0.95)	97.89(2.09)	67.27(5.16)	52.11(4.25)	**3.11**
The proportion of forgotten data samples to all samples is 30%										
Retrain	91.61	99.97	91.29	16.11	0	67.67	99.98	67.54	58.47	0
FT	99.19(7.58)	99.95(0.02)	92.94(1.65)	3.10(13.01)	5.57	96.24(28.57)	99.94(0.04)	71.49(3.95)	20.57(37.90)	17.62
GA	99.35(7.74)	99.38(0.59)	93.39(2.10)	1.28(14.83)	6.32	96.73(29.06)	97.30(2.68)	73.35(5.81)	6.19(52.28)	22.46
IU	76.21(15.40)	75.93(24.04)	71.48(19.81)	24.53(8.42)	16.92	2.81(64.86)	3.19(96.79)	3.09(64.45)	96.33(37.86)	65.99
RL	93.96(2.35)	96.03(3.94)	90.34(0.95)	19.07(2.96)	2.55	75.50(7.83)	99.58(0.40)	66.56(0.98)	87.33(28.86)	9.52
ℓ_1-sparse	99.06(7.45)	99.90(0.07)	93.01(1.72)	4.54(11.57)	5.20	88.59(20.92)	99.79(0.19)	70.72(3.18)	33.16(25.31)	12.40
SalUn	96.17(4.56)	97.78(2.19)	91.57(0.30)	15.50(0.66)	1.93	79.60(11.93)	96.70(3.28)	63.41(4.13)	57.73(0.74)	5.02
Q-MUL	93.57(1.96)	96.77(3.20)	90.24(1.05)	15.59(0.52)	**1.68**	73.90(6.23)	97.63(2.35)	65.80(1.74)	63.76(5.29)	**3.90**
The proportion of forgotten data samples to all samples is 50%										
Retrain	90.75	100.0	90.28	18.76	0	65.40	99.99	65.74	68.19	0
FT	99.29(8.54)	99.96(0.04)	93.28(3.00)	2.85(15.91)	6.87	97.10(31.70)	99.98(0.01)	72.82(7.08)	16.86(51.33)	22.71
GA	99.34(8.59)	99.34(0.66)	93.44(3.16)	1.27(17.49)	7.48	96.82(31.42)	97.33(2.66)	73.37(7.63)	5.93(62.26)	25.99
IU	16.12(74.63)	16.02(83.98)	16.22(74.06)	16.36(2.40)	58.77	0.96(64.44)	1.04(98.95)	1.00(64.74)	98.96(30.77)	64.73
RL	95.32(4.57)	97.17(2.83)	90.85(0.57)	21.33(2.57)	2.64	69.20(3.80)	85.60(14.39)	64.39(1.35)	54.58(13.61)	8.29
ℓ_1-sparse	99.25(8.50)	99.95(0.05)	92.67(2.39)	3.91(14.85)	6.45	96.48(31.08)	99.98(0.01)	72.26(6.52)	25.11(43.08)	20.17
SalUn	96.56(5.81)	98.03(1.97)	91.55(1.27)	18.05(0.71)	2.44	87.75(22.35)	98.99(1.00)	64.69(1.05)	73.12(4.93)	7.33
Q-MUL	91.09 (0.34)	93.94(6.06)	89.12(1.16)	18.68 (0.08)	**1.91**	67.25(1.85)	89.65(10.34)	61.38(4.36)	56.52(11.67)	**7.06**

Class-wise Forgetting Scenario

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)	0.10
Q-MUL	0.00(0.00)	99.69(0.30)	93.21(0.13)	100.0(0.00)	0.11

Q-MUL for Full-precision Model



ICCV OCT 19-23, 2025 HONOLULU HAWAII

Codes and pre-trained models are at
https://github.com/Green-Edge-AI-Group/MU-Q-MUL