NeurIPS: Machine Unlearning

Final Project Presentation

CSE 472 Machine Learning Sessional

BY:
Syed Jarullah Hisham - 1805004
Abdur Rafi - 1805008

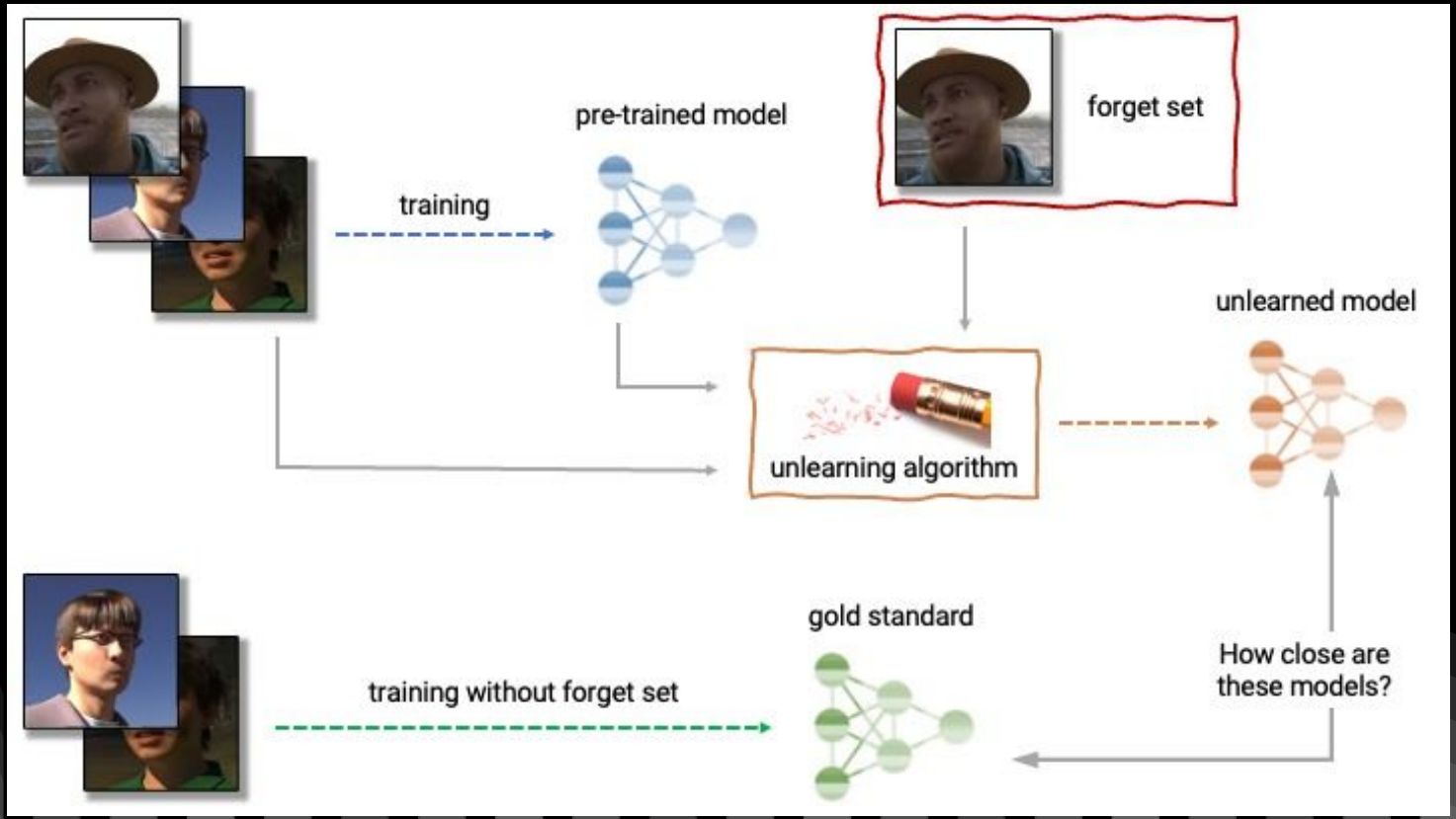
GOOGLE · RESEARCH CODE COMPETITION · A MONTH AGO

NeurIPS 2023 - Machine Unlearning

Erase the influence of requested samples without hurting accuracy

Competition Link: https://www.kaggle.com/competitions/neurips-2023-machine-unlearning/overview

Problem Definition



Dataset Analysis

- Competition Dataset Hidden
- **[retain/forget/validation].csv**
    - person_id: A unique ID code for each subject.
    - age_group: Binned ages. The target label for the models.(15 classes)
    - age - The age of the subject.
    - image_id: A unique ID for the image.
- **images/[person_id]/[image_id].png** Images of people's faces.
    - If the image ID is 'abc-1' the path is images/abc/1.png.
    - images have been resized to 32x32 pixels.
    - Expect approximately 30,000 images in the hidden dataset.
    - roughly 2% of the images in the dataset have identical perceptual hashes

Dataset Analysis

- age_class_weights.json
  - age_group weights used to train the original model
  - in the form {age_group: n_occurrences}
- original_model.pth
  - original Resnet-18 Pytorch model checkpoint
  - needs to unlearn
  - 99% accuracy on the training set
  - 96% accuracy on the test set

- For our testing, **CIFAR10** was used

Loss Functions

- Kullback-Leibler Divergence Loss
  - is a way of measuring the matching between two distributions
  - higher the KL divergence value, the less matching of the true distribution with other
  - Used in Forgetting Stage

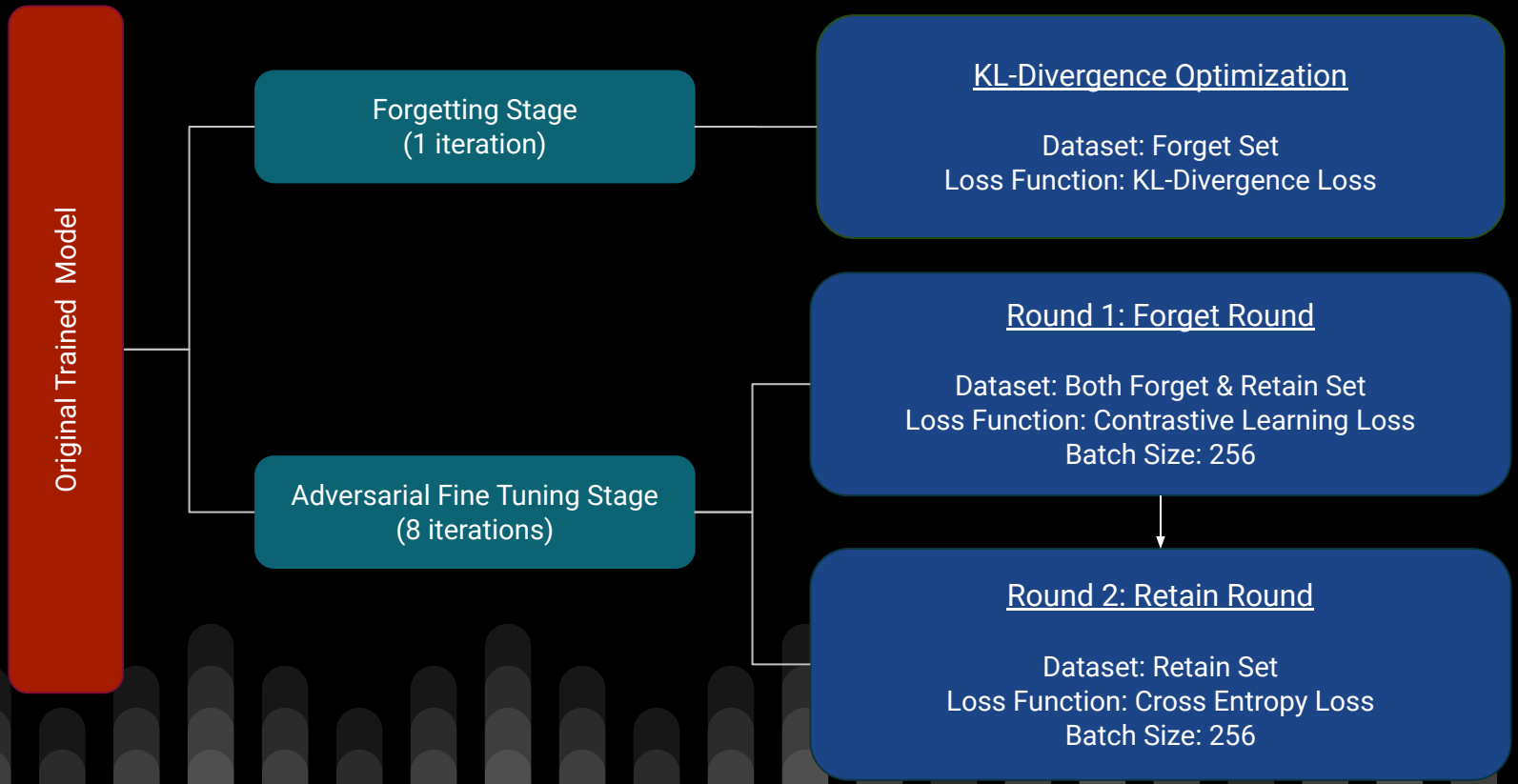
- Contrastive Learning Loss
  - aims to pull the distance between positive samples and their enhanced samples closer and push the distance between positive samples and all samples further
  - pushing the distance between any forget sample and all retain samples further
  - Used in Adversarial Fine Tuning Stage Forget Round
- Cross Entropy Loss
  - Used in Adversarial Fine Tuning Stage Retain Round

Performance Metrics

- Model Quality (Cross Entropy Loss & Accuracy)
- Simple Membership Inference Attack(MIA)
- Combination of Forgetting Quality, Efficiency and Utility
    - Used in competition
    - Hidden metric
    - https://unlearning-challenge.github.io/assets/data/Machine_Unlearning_Metric.pdf

Approach 1 : Reproduce & Improve 1st Place Solution

1st Place Solution Architecture



Performance Report - 1st Place

Our Reproduced Solution

Competition Leaderboard

#	Team	Members	Score
1	fanchuan		0.0984971060
2	[kookmin Univ] LD&BGW&KJH		0.0902000292
3	Seif Eddine Achour		0.0884946037
4	Sebastian Oleszko		0.0863052371

Performance Report - 1st Place

Model	Dataset	Cross-Entropy Loss	Accuracy(%)	MIA Accuracy(%)
Original	Forget Set	0.02173	99.32	57.8
	Validation Set	0.42808	88.64	
Retrained	Forget Set	0.4607	88.20	49.6
	Retain Set	0.0181	99.53	
	Validation Set	0.4641	88.00	
Unlearned (1st Place)	Forget Set	0.1877	95.36	51.3
	Retain Set	0.0289	100.00	
	Validation Set	0.3862	88.54	

Performance Report - Experiments on 1st Place

Unlearned Model Modification	Dataset	Cross-Entropy Loss	Accuracy(%)	MIA Accuracy(%)	Competition Score
1st Place	Forget Set	0.1877	95.36	51.3	0.09142
	Retain Set	0.0289	100.00		
	Validation Set	0.3862	88.54		
2 Forget Round for each epoch	Forget Set	0.27	92.64	50.7	0.079052
	Retain Set	0.05	99.99		
	Validation Set	0.38	88.22		
9 epochs at adversarial fine tuning stage	Forget Set	0.21	94.44	50.4	0.08921
	Retain Set	0.03	100.00		
	Validation Set	0.38	88.34		

Performance Report - Experiments on 1st Place

Unlearned Model Modification	Dataset	Cross-Entropy Loss	Accuracy(%)	MIA Accuracy(%)	Competition Score
10 epochs at adversarial fine tuning stage	Forget Set	0.23	93.9	50.0	0.070155
	Retain Set	0.03	100.00		
	Validation Set	0.39	88.32		
12 epoch with 512 batch size at adversarial fine tuning stage	Forget Set	0.31	90.8	47.0	0.0566
	Retain Set	0.03	100.00		
	Validation Set	0.41	87.96		
2 Forget Round + 9 epoch at adversarial fine tuning stage	Forget Set	0.29	91.20	50.3	0.08794
	Retain Set	0.05	99.99		
	Validation Set	0.39	88.26		

Performance Report - Experiments on 1st Place

Unlearned Model Modification	Dataset	Cross-Entropy Loss	Accuracy(%)	MIA Accuracy(%)	Competition Score
JSD loss instead of KL_Div loss	Forget Set	0.17	96.2	51.3	0.0273
	Retain Set	0.03	100		
	Validation Set	0.39	88.42		
JSD loss instead of KL_Div loss+ 2 Forget Round	Forget Set	0.24	93.66	51.2	0.05621
	Retain Set	0.05	99.99		
	Validation Set	0.38	88.30		
Modified KL_DIV loss	Forget Set	0.19	95.46	51.2	0.099985
	Retain Set	0.03	100.00		
	Validation Set	0.39	88.48		

Performance Report - Experiments on 1st Place

Unlearned Model Modification	Dataset	Cross-Entropy Loss	Accuracy(%)	MIA Accuracy(%)	Competition Score
Modified KL_DIV loss + 9 epoch	Forget Set	0.21	94.74	50.6	0.09009
	Retain Set	0.03	100.00		
	Validation Set	0.39	88.30		
Modified kl_div + 512 batch size	Forget Set	0.19	95.46	51.0	0.09838
	Retain Set	0.03	100.00		
	Validation Set	0.39	88.44		
Modified kl_div + 512 batch size + 9 epoch	Forget Set	0.21	94.76	50.0	0.09758
	Retain Set	0.03	100.00		
	Validation Set	0.39	88.28		

Performance Report - Experiments on 1st Place

Unlearned Model Modification	Dataset	Cross-Entropy Loss	Accuracy(%)	MIA Accuracy(%)	Competition Score
Modified kl_div + relative weighting [1, 0.5]	Forget Set	0.18	95.54	51.3	0.0909
	Retain Set	0.03	100.00		
	Validation Set	0.39	88.42		
Modified kl_div + relative weighting [1, 0.8]	Forget Set	0.19	95.56	50.8	0.1003
	Retain Set	0.03	99.99		
	Validation Set	0.39	88.48		
Modified kl_div + relative weighting [1, 0.85]	Forget Set	0.18	95.54	51.4	0.106846
	Retain Set	0.03	100.00		
	Validation Set	0.39	88.60		

Improved Solution: [Modification of KL_DIV]

```
kl_loss = nn.KLDivLoss(reduction='batchmean')
return kl_loss(nn.LogSoftmax(dim=-1)(x),y)
```

```
kl_loss = nn.KLDivLoss(reduction='batchmean')
return kl_loss(nn.LogSoftmax(dim=-1)(x),y) + 0.85 * kl_loss(y.log(),nn.Softmax(dim=-1)(x))
```

Performance Report - Successful Solution

Our Best Improved Solution

NeurIPS: Machine-Unlearning-Improved-SCL - Version 67

Succeeded (after deadline) · 9h ago · [The BEST]

0.1068457555

Reproduced 1st Place Solution

NeurIPS: Machine-Unlearning-Improved-SCL - Version 2

Succeeded (after deadline) · 22d ago · 1st Place Solution

0.0914203871

Competition 1st Place Solution

#	Team	Members	Score
1	fanchuan		0.0984971060

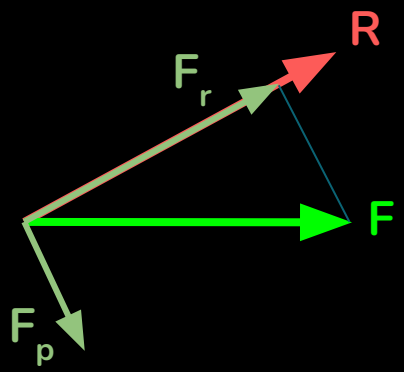
Approach 2 : Gradient Ascent Approach

Our Proposed Idea

2 rounds.

1. Forget round
2. Retain round

Our Proposed Idea-Forget Round

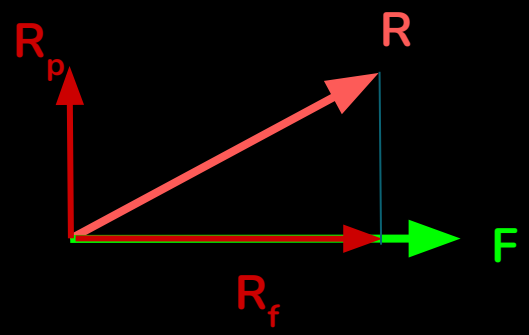


R: Gradient computed on sample of Retain set

F: Gradient computed on sample of Forget set

$$\text{param} \mathrel{+}= \text{lr} * F_p$$

Our Proposed Idea-Retain Round

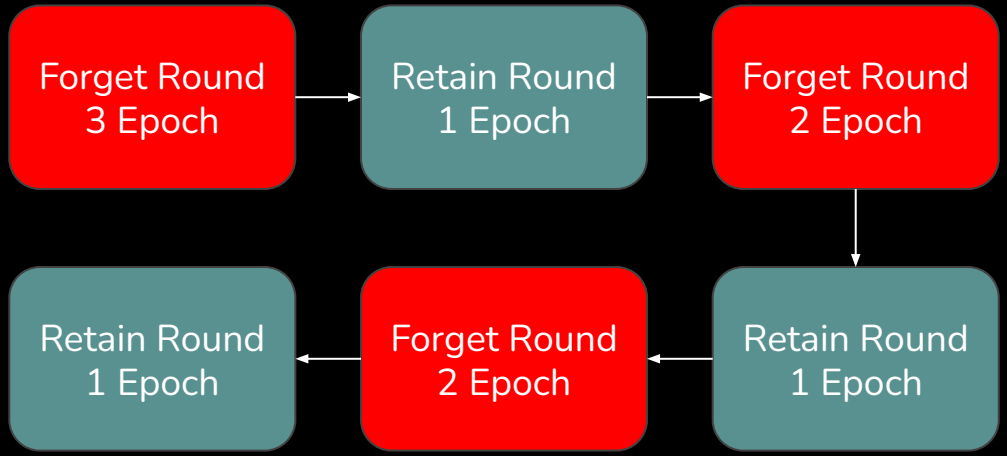


R: Gradient computed on sample of Retain set

F: Gradient computed on sample of Forget set

$$\text{param} \mathrel{-}= \text{lr} * R_p$$

Our Proposed Idea



Performance Report - Our approach

Model	Dataset	Cross-Entropy Loss	Accuracy(%)	MIA Accuracy(%)
Retrained	Forget Set	0.4607	88.20	49.6
	Retain Set	0.0181	99.53	
	Validation Set	0.4641	88.00	
Gradient Ascent	Forget Set	0.60	89.98	48.3
	Retain Set	0.02	99.73	
	Validation Set	0.52	86.78	

Comparison of All Solutions
[1st place, Approach 1, Approach 2]

Performance Comparison

Unlearned Model Modification	Dataset	Cross-Entropy Loss	Accuracy(%)	MIA Accuracy(%)	Competition Score	Time (s)
Reproduced 1st Place	Forget Set	0.1877	95.36	51.3	0.09142	280
	Retain Set	0.0289	100.00			
	Validation Set	0.3862	88.54			
Approach 1: Improved 1st Place	Forget Set	0.18	95.54	51.4	0.106846	290
	Retain Set	0.03	100.00			
	Validation Set	0.39	88.60			
Approach 2: Gradient Ascent	Forget Set	0.60	89.98	48.3	0.0048	106
	Retain Set	0.02	99.73			
	Validation Set	0.52	86.78			

Challenges/Discussion

Challenges

- unavailability of competition metric
- unavailability of competition dataset
- Needs more than 7-8 hours to get the competition score
- Build an appropriate unlearning algorithm which will run very efficiently is the key challenge
- Good performance in model quality and MIA attack does not confirm good competition score which was the biggest challenge

Notebooks

Notebooks

- **Reproduced 1st Place:** https://www.kaggle.com/code/syedjarullahhisham/neurips-machine-unlearning-improved-scl?scriptVersionId=158595208

- **Approach 1 - Improved 1st Place:** https://www.kaggle.com/code/syedjarullahhisham/neurips-machine-unlearning-improved-scl?scriptVersionId=163792309

- **Approach 2 - Gradient Ascent:** https://www.kaggle.com/code/abdurrafi08236/neurips-machine-unlearning-gradient-ascent?scriptVersionId=164117975

Thank You