JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

Boundary Unlearning: Rapid Forgetting of Deep Networks via Shifting the Decision Boundary

Min Chen, Weizhuo Gao, Gaoyang Liu, Kai Peng, Chen Wang*,

Hubei Key Laboratory of Smart Internet Technology, School of EIC,

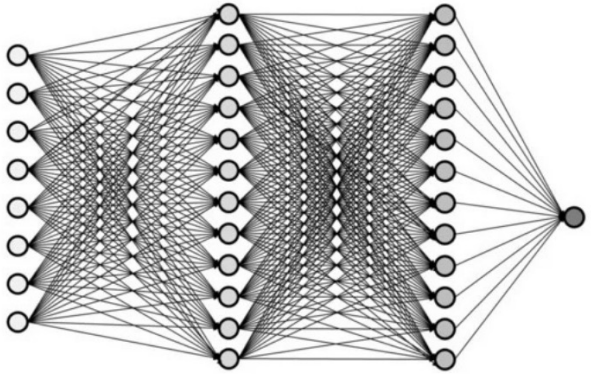
Huazhong University of Science and Technology

Code is available on https://www.dropbox.com/s/bwu543qsdy4s32i/Boundary-Unlearning-Code.zip?dl=0

Background

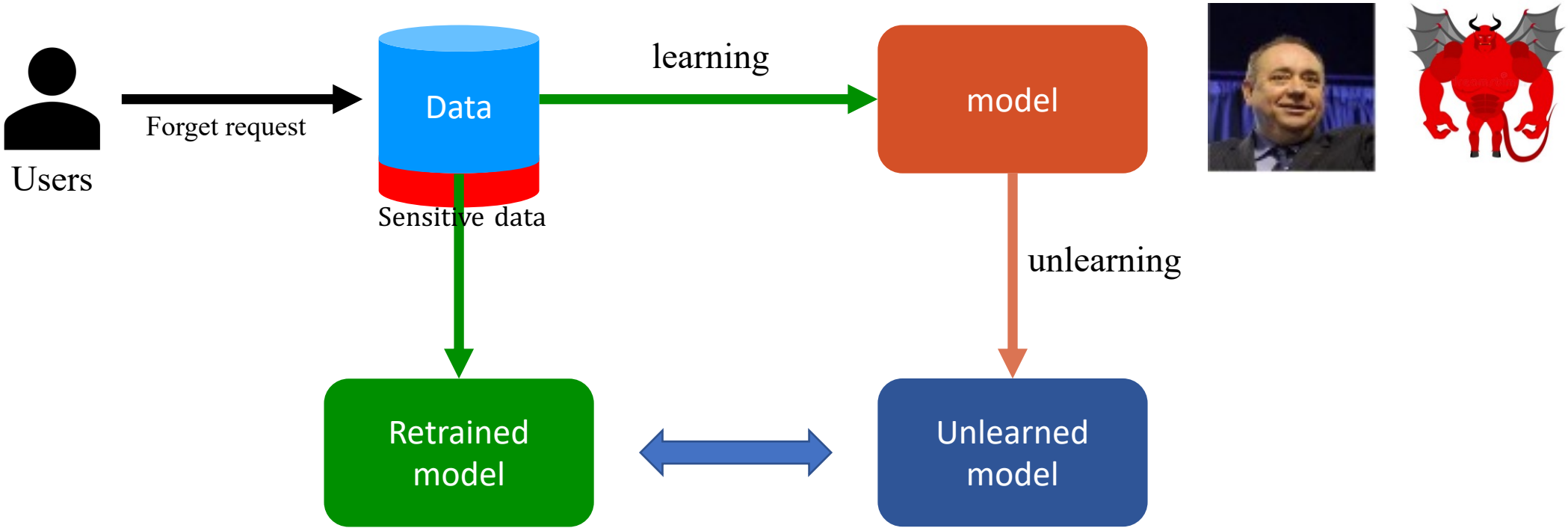
DNNs have become ubiquitous tool in developing data-driven services.

- Power lies in the large parameter space
- DNNs with poor interpretability



- DNNs cause unintended memorization
- DNNs may memorize sensitive or flawed data

Background


Users
Forget request
Data
Sensitive data
learning
model
unlearning
Retrained model
Unlearned model


GDPR
Data Protection Officer (DPO)
Compliance
25 May 2018
Data Breaches
Personal Data


CCPA
Right to Know
Right to Delete
Right to Opt-Out
Right to Nondiscrimination

Early studies

Retrain Acceleration for DNN Models :
SISA
Amnesiac unlearning
Deltagrad
...

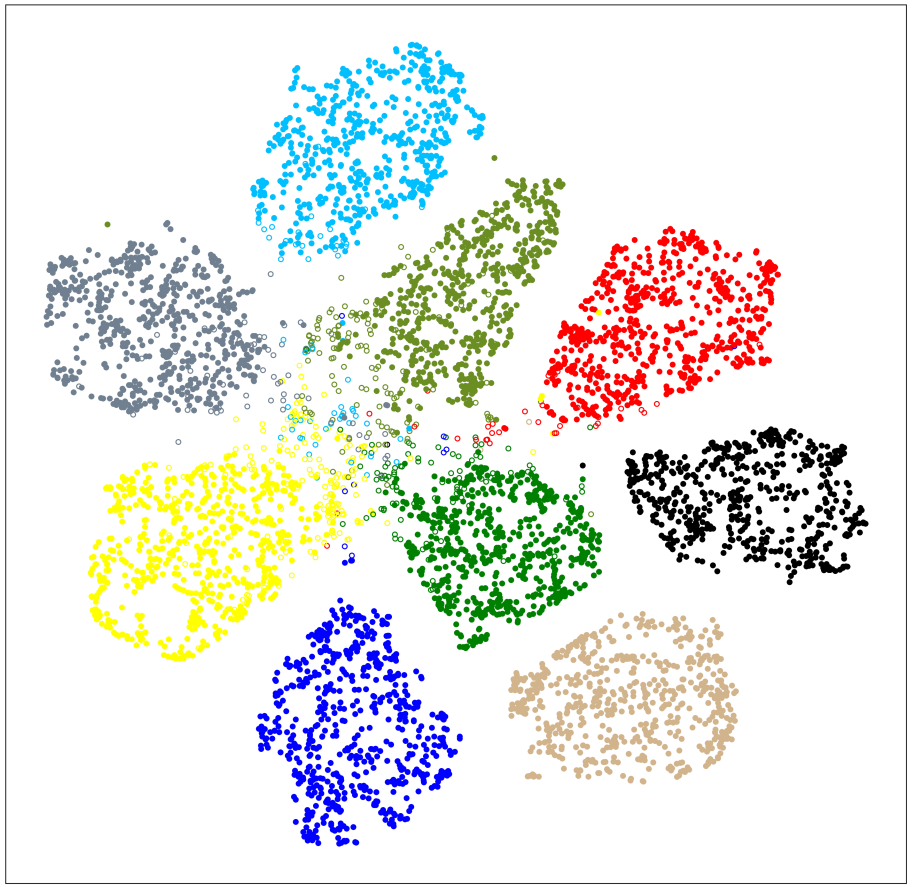
have to intervene the original training pipeline and hurt the utility of the DNN model

Updating Parameters for DNN Models :
Fisher forgetting
NTK forgetting
SSSE
...

cost too much computational resource

Problem: How to reduce the computational complexity?

Observations

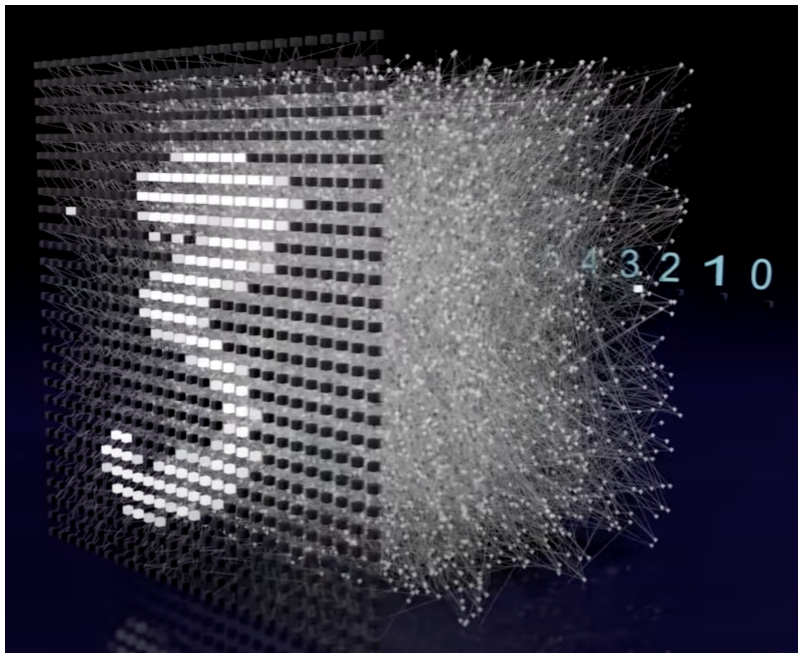


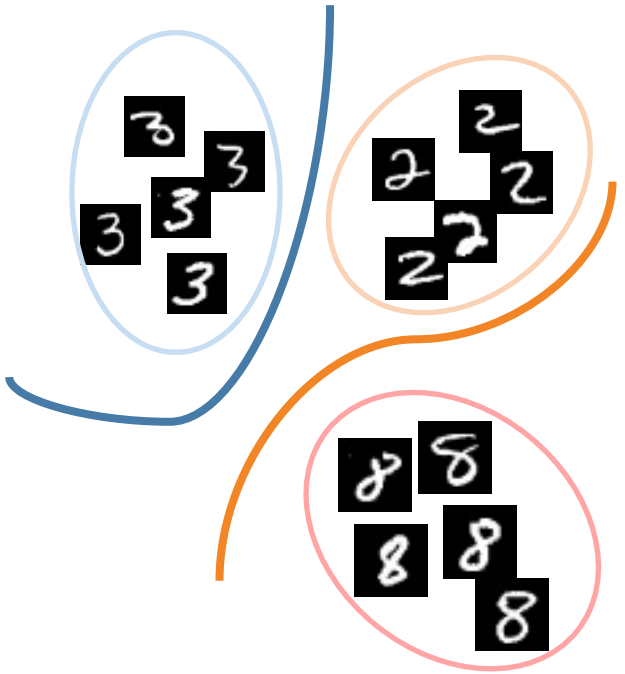
Decision Space of the Retrained DNN

- the forgetting samples spread around the decision space of the retrained DNN model
- the utility guarantee can be achieved by only destroying the boundary of the forgetting class but maintaining the boundary of the remain classes

- most of the forgetting samples move to the border of other clusters.
- the privacy guarantee can be accomplished by pushing the forgetting data to the border of other clusters

Motivation

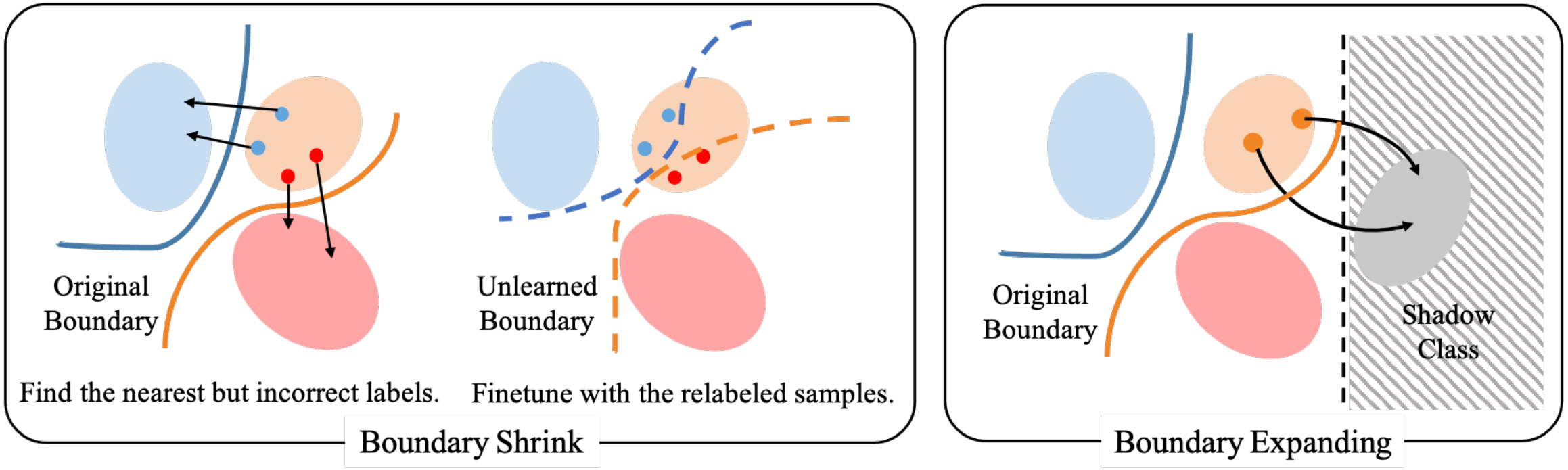




Parameter space

Decision space

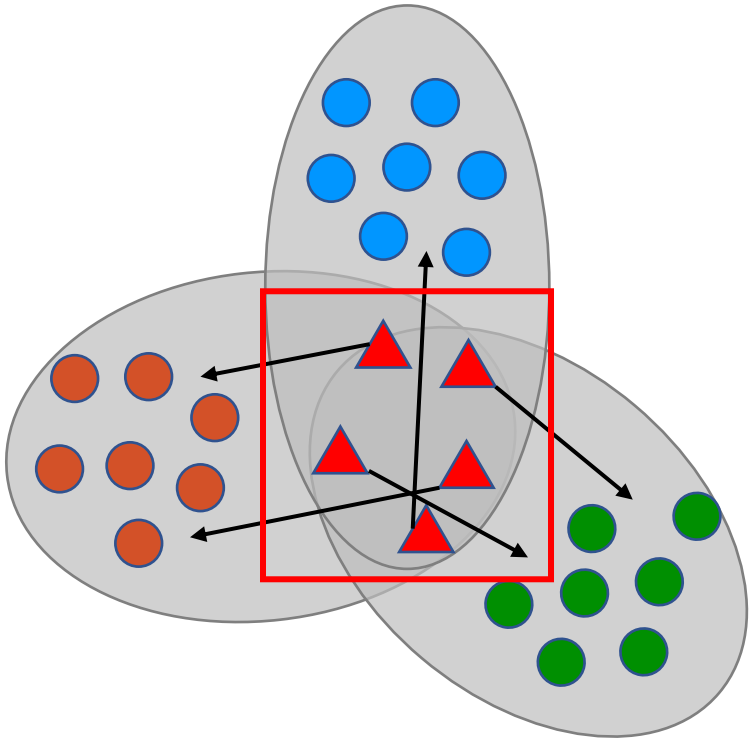
Our approach



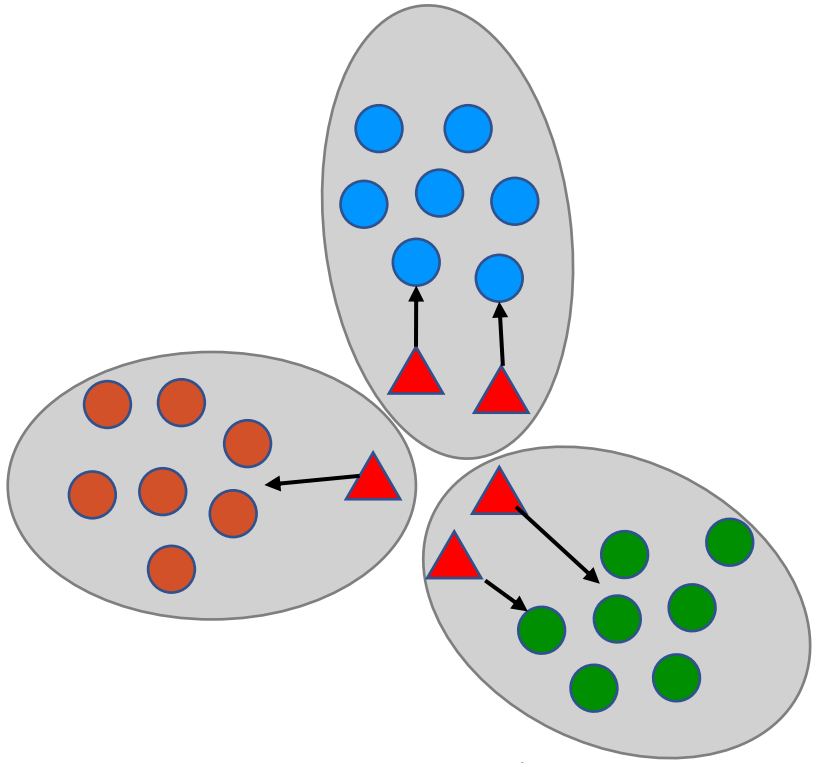
High level overview of Boundary Unlearning

Our approach

how to shift the boundary & which direction to shift? ⇒ Boundary Shrink



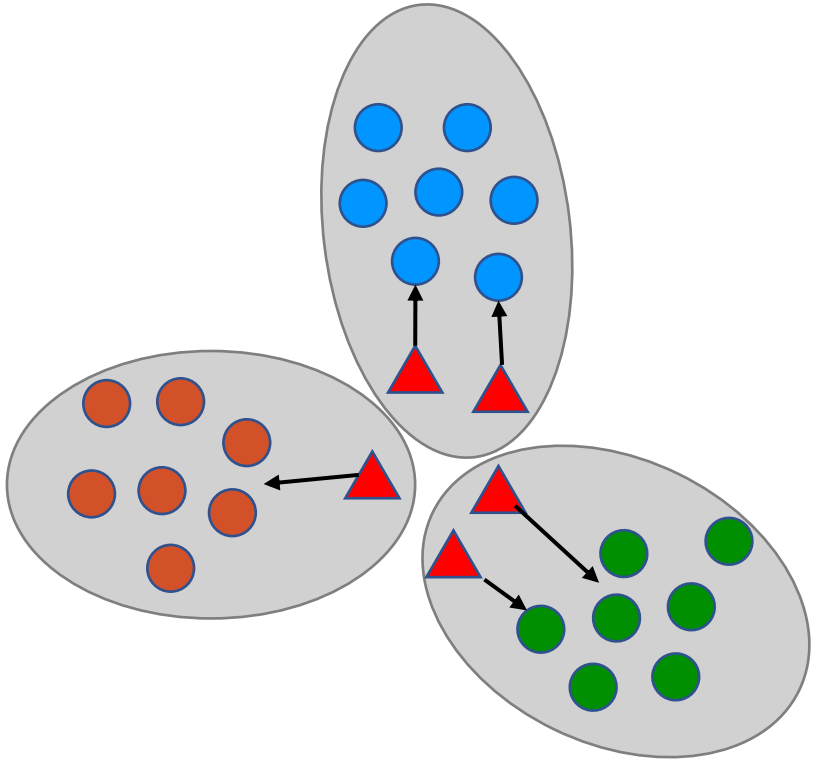
Finetune with randomly labeled data.



Shrink the decision space of forgetting sample

Our approach

how to shift the boundary & which direction to shift? ⇒ Boundary Shrink



Step 1: Neighbor searching

$$\mathbf{x'}_f = \mathbf{x}_f + \epsilon \cdot \text{sign}(\nabla_{\mathbf{x}_f} \mathcal{L}(\mathbf{x}_f, \mathbf{y}, \mathbf{w}_0))$$

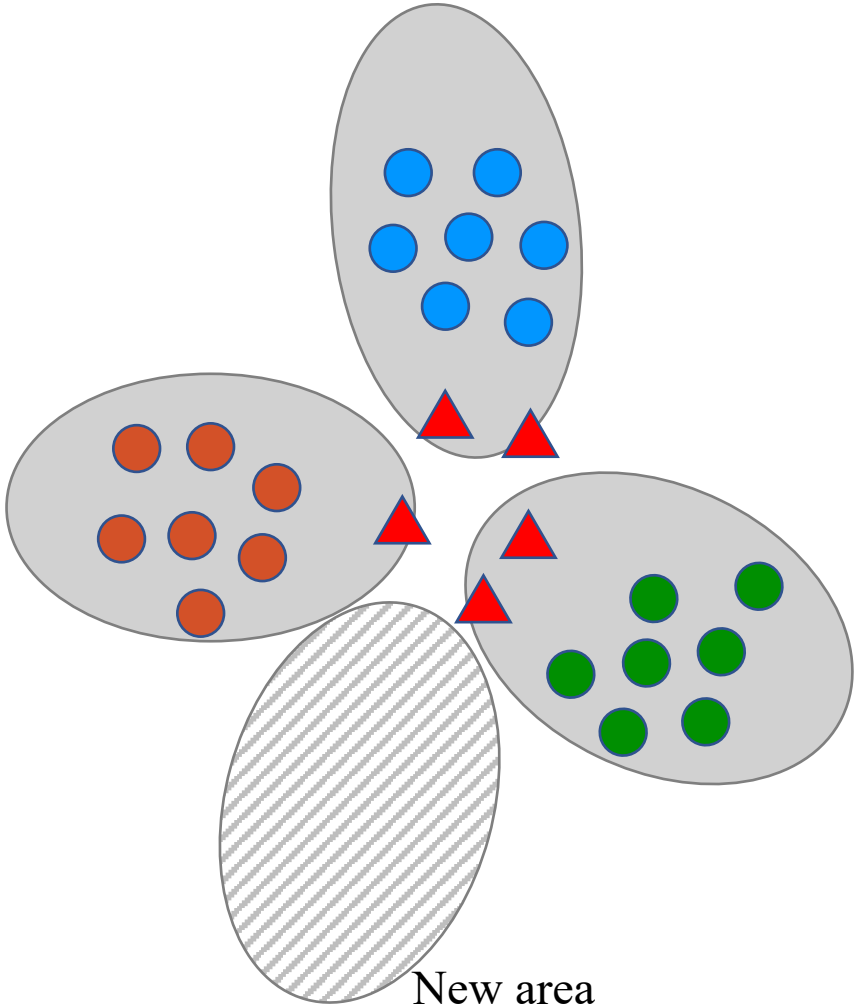
Step 2: Finetune with (x, y_{nbi})

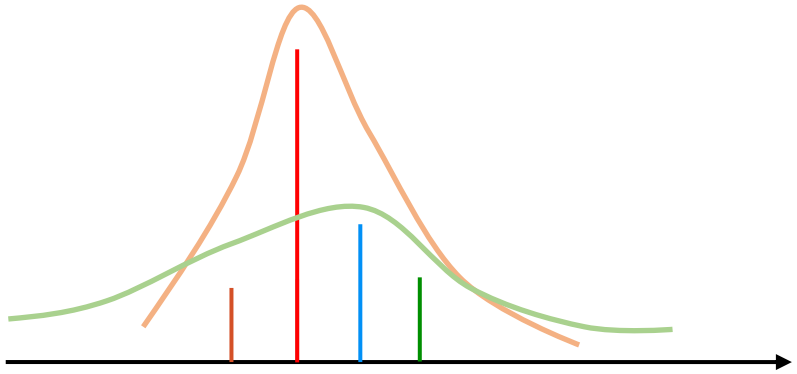
$$\mathbf{w}' = \arg\min_{\mathbf{w}} \sum_{(\mathbf{x}_i, \mathbf{y}_{nbi}) \in \mathcal{D}_f} \mathcal{L}(\mathbf{x}_i, \mathbf{y}_{nbi}, \mathbf{w}_0)$$

Our approach

how to shift the boundary & which direction to shift? ⇒ Boundary Expanding

speed-up version





$$\mathbf{w}' = \arg\min_{\mathbf{w}} \sum_{(\mathbf{x}_i, \mathbf{y}_{shadow}) \in \mathcal{D}_f} \mathcal{L}(\mathbf{x}_i, \mathbf{y}_{shadow}, \mathbf{w}_0)$$

Evaluation Results

➢ Dataset and Model Architecture

Dataset	Model	Task
CIFAR-10	All-CNN	Image classification
Vggface2	ResNet50	Face recognition

➢ Evaluation metrics

- **accuracy metric**: accuracy on D_r, D_f, D_{rt} and D_{ft}.
- **privacy metric**: the attack success rate (ASR) of membership inference attack.
- **time consumption**: time consumed by the each unlearning method.

Evaluation Results

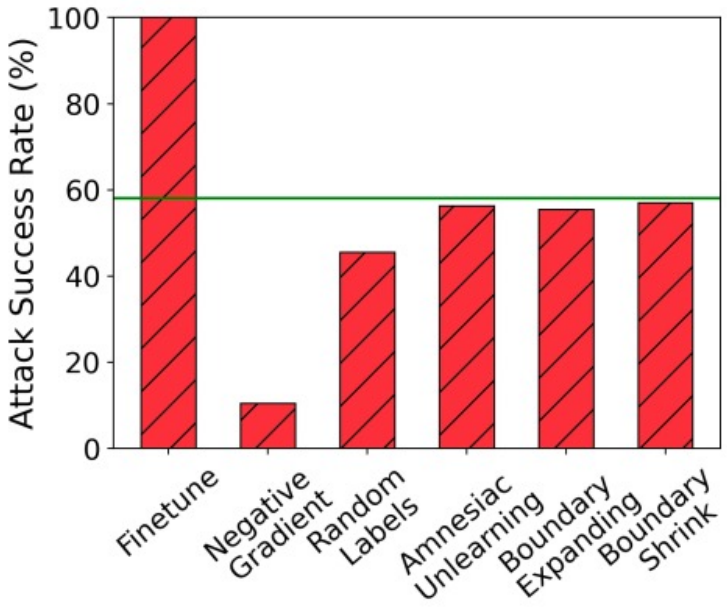
- **Utility Guarantee**

Datesets	Metrics	Original DNN model	Retrained DNN model	Finetune	Negative Gradient	Random Labels	Boundary Shrink	Boundary Expanding
CIFAR-10	Acc on $\mathcal{D}_r$	99.97	100.00	100.00	97.16	98.49	**99.24**	98.03
	Acc on $\mathcal{D}_f$	99.92	0.00	0.22	7.84	10.40	**5.94**	8.96
	Acc on $\mathcal{D}_{rt}$	84.83	85.74	86.50	80.42	81.81	**83.13**	81.07
	Acc on $\mathcal{D}_{ft}$	81.20	0.00	0.10	6.50	7.50	**5.94**	7.00
Vggface2	Acc on $\mathcal{D}_r$	99.94	100.00	99.52	96.57	**98.89**	98.57	98.20
	Acc on $\mathcal{D}_f$	98.57	0.00	0.00	2.85	4.29	**1.54**	4.22
	Acc on $\mathcal{D}_{rt}$	98.87	99.06	99.96	99.58	95.14	**99.72**	97.12
	Acc on $\mathcal{D}_{ft}$	97.14	0.00	5.52	7.26	2.86	**0.87**	1.41

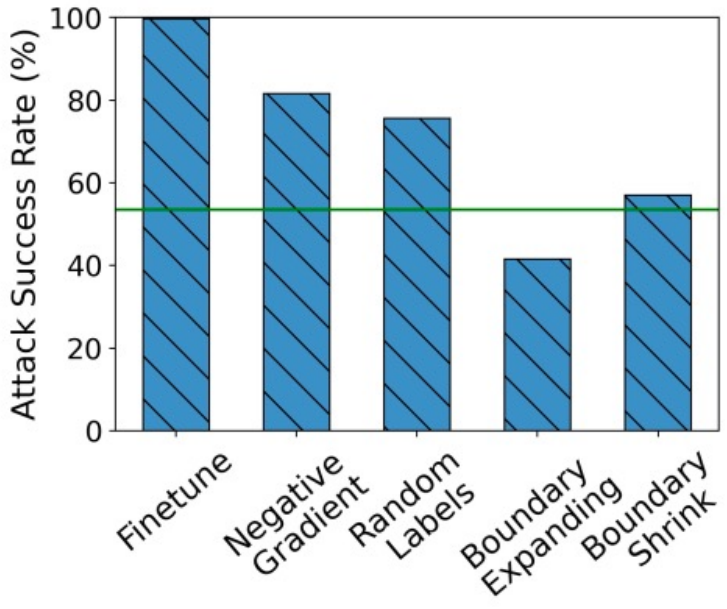
Metrics	Acc on $\mathcal{D}_r$	Acc on $\mathcal{D}_f$	Acc on $\mathcal{D}_{rt}$	Acc on $\mathcal{D}_{ft}$
Amnesiac Unlearning	95.79	0.00	81.50	0.00
Fisher Forgetting	61.62	1.80	54.20	1.60

Evaluation Results

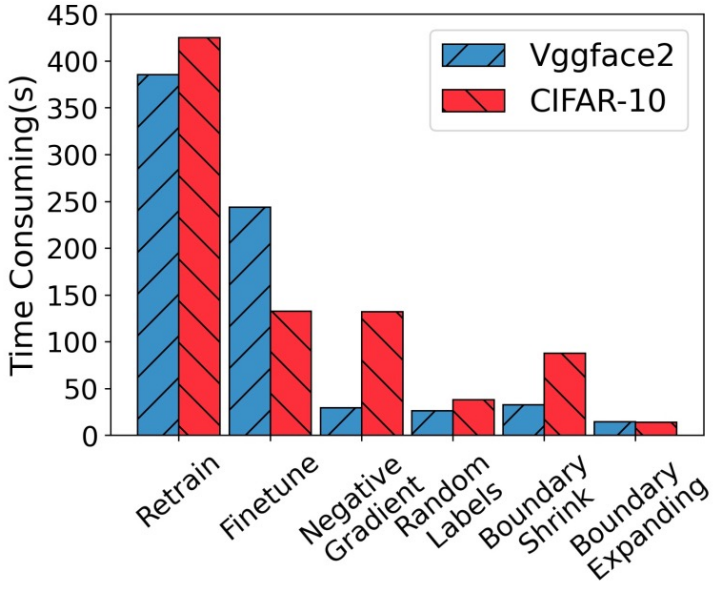
- **Privacy Guarantee and Time Consumption**



On CIFAR-10



On Vggface2

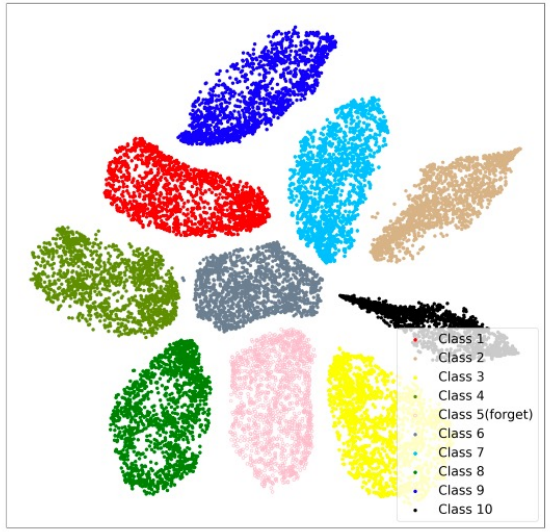


On CIFAR-10

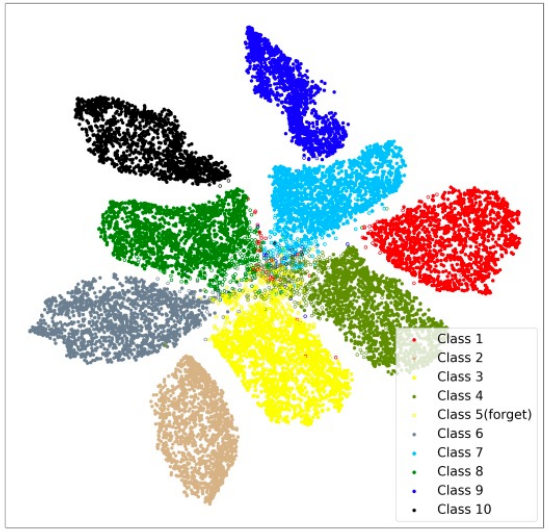
Boundary Unlearning methods can achieve a better performance on privacy guarantee **effectively** and **quickly**.

Evaluation Results

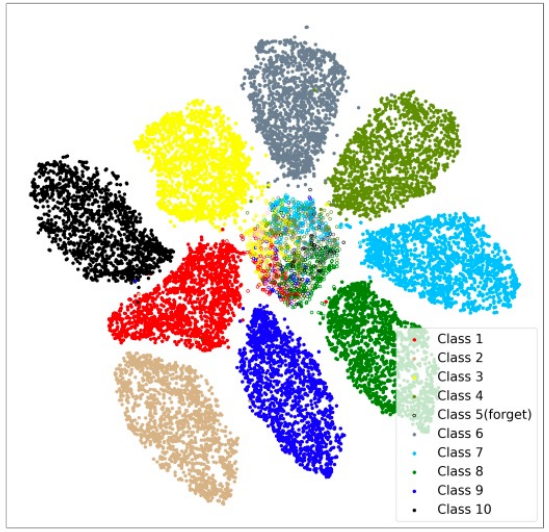
- **Visualization of Decision Space**



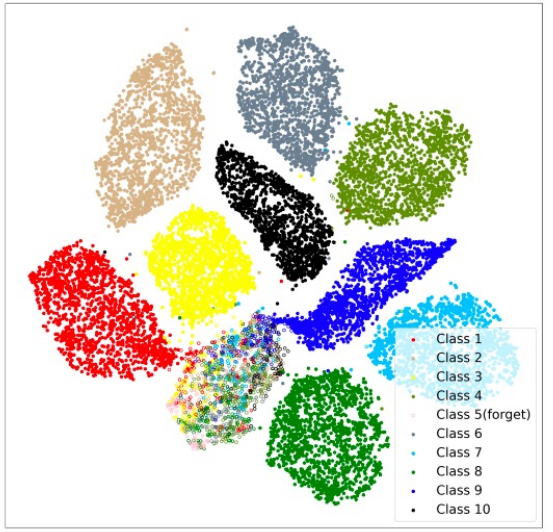
(a) Original



(b) Retrained



(c) Boundary Shrink

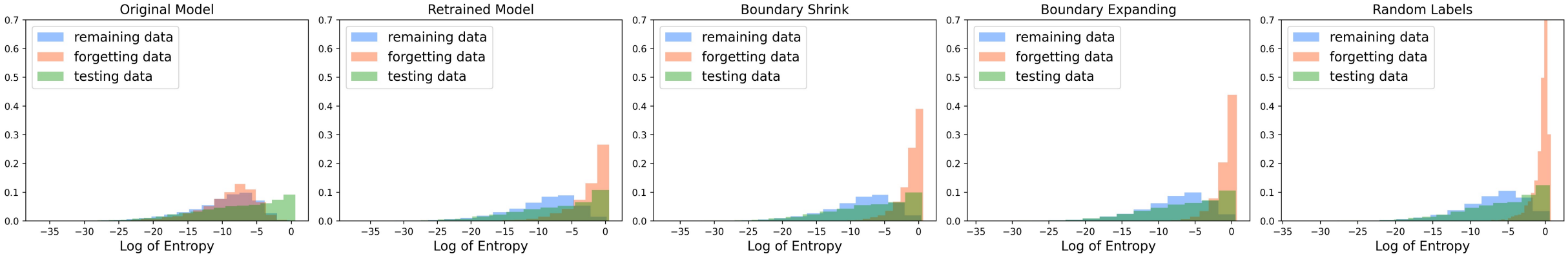


(d) Boundary Expanding

Boundary Unlearning imitated boundary of retrained model and thus accomplishes the unlearning efficacy.

Evaluation Results

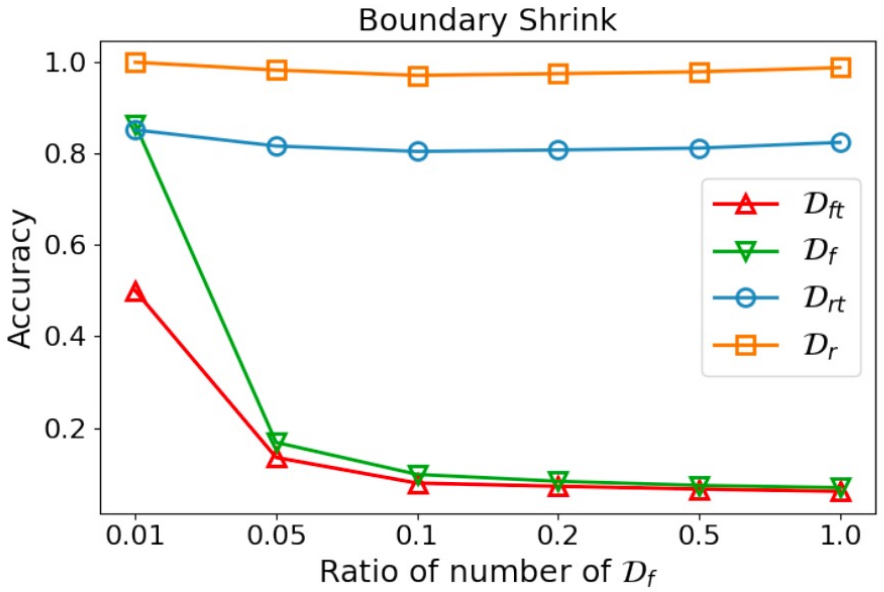
- **Distribution of the Entropy of Model Output**

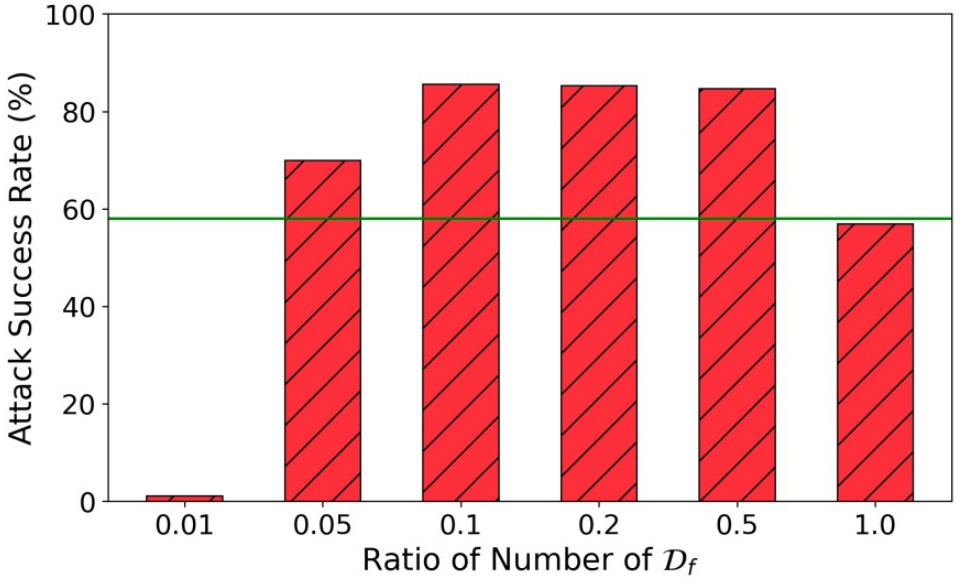


the unlearned model predicts them with low confidence like predicting the testing samples.

Evaluation Results

- **Impact of Number of Samples Needed**

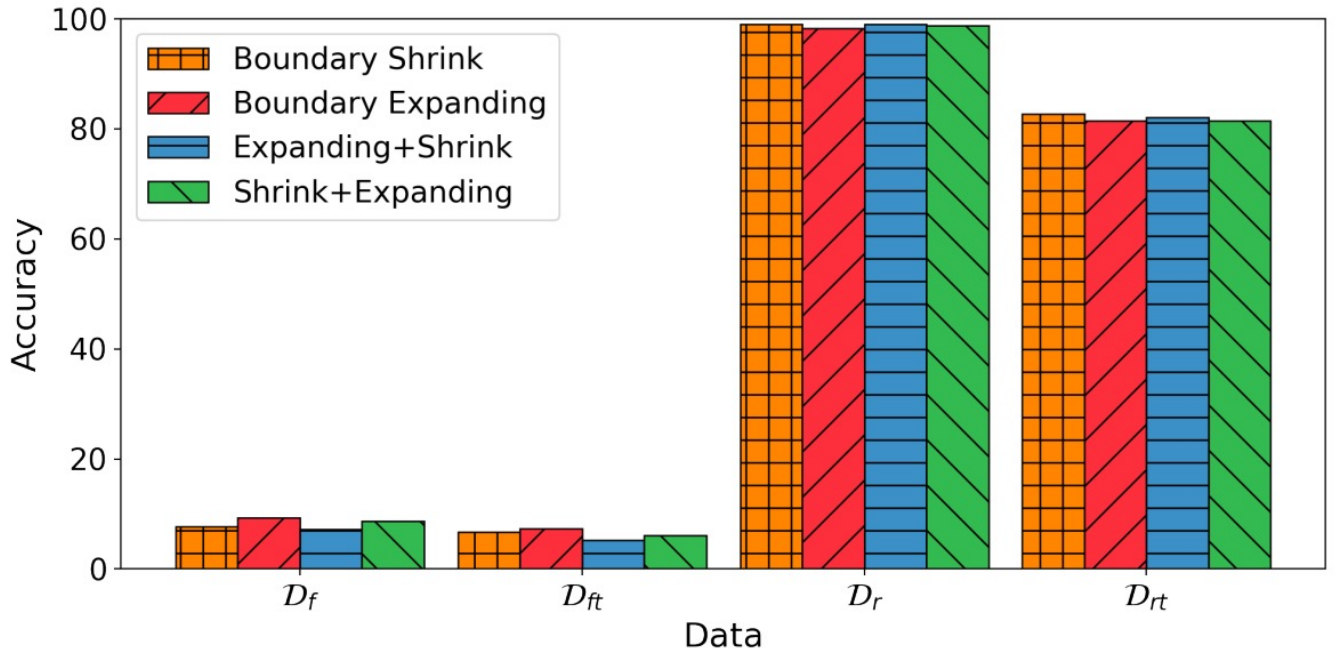




Boundary Shrink can still forget the entire class with less forgetting samples

Evaluation Results

- **Combination of Boundary Shrink and Boundary Expanding**



"first running Boundary Expanding and then Boundary Shrink" may be the best combination

Conclusion

- Boundary Unlearning: the first machine unlearning methodology to remove information of an entire class from a trained DNN by shifting the decision boundary

- We envision our work as a practical step in machine unlearning towards revealing the relationship between decision boundary and forgetting

- More interesting results in the paper
  - ✓ Attention map before and after unlearning
  - ✓ Discussion on utility and privacy guarantee of unlearning
  - ✓ Boundary Shrink with different hyperparameters