Learning to Rewind via Iterative Prediction of Past Weights for Practical Unlearning

Jinhyeok Jang, Jaehong Kim, Chan-Hyun Youn

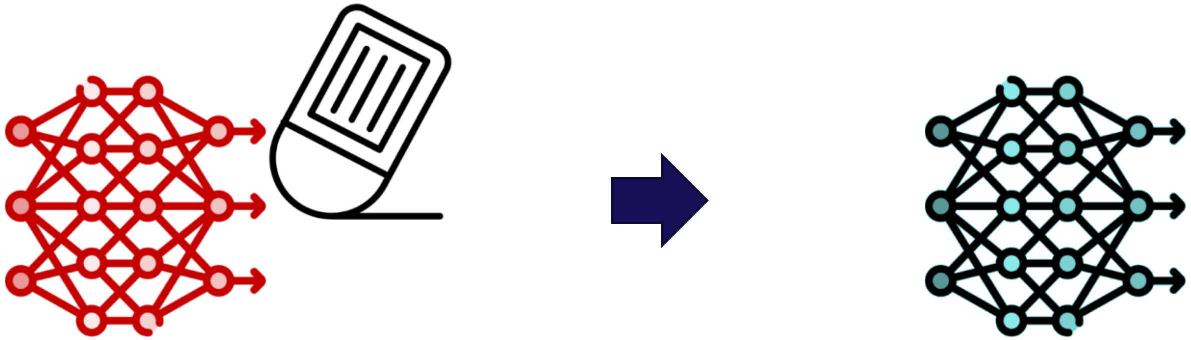
Post Processing after Verification of Unauthorized Use of Specific Data

- Recently, there have been many issues arising from AI models that use data containing **copyrighted or private information without permission**.
- In such cases, the owners of the problematic AI model must choose one of these:

<Options>

1) **Negotiating a settlement** with **appropriate compensation**
2) **Re-training from scratch** using the full dataset, excluding the problematic data.
3) **Removing the problematic knowledge** from the trained model.

Post Processing



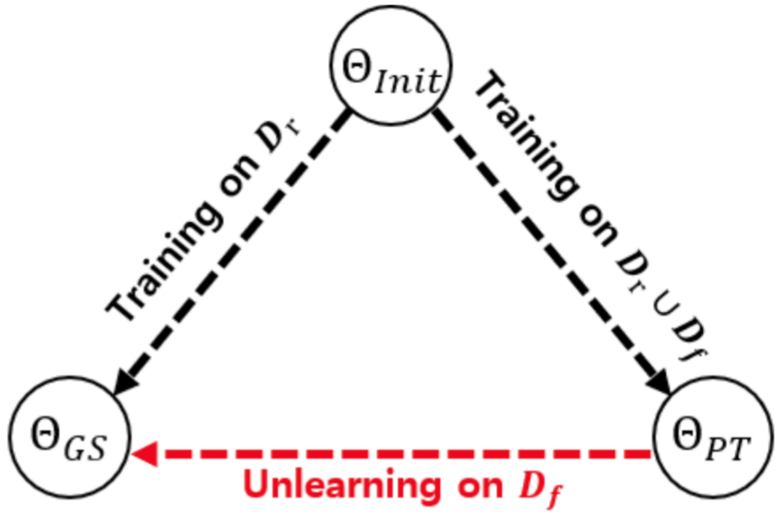
DNN with Unauthorized Use

DNN without Problem

How to remove only the selected knowledge

Machine Unlearning

- *A topic* dedicated to selectively removing specific knowledge from DNNs, aiming to find Θ_{GS}



<Notations>
Forget data (D_f): problematic data or dataset that should be forgotten
Remaining data (D_r): the others included in training dataset
Entire training dataset ($D_f \cup D_r$): the union of forget and remaining data
$\boldsymbol{\Theta}$: Weights of a AI model

In our scenario, data used without authorization is designated as D_f, while data with authorized use is designated as D_r.

https://unlearning-challenge.github.io/

Related Works

- **1. Unlearning using the Entire Dataset**
  - Overwhelming computational and storage demands
  - *i.e., Weight Noising, Fisher, SCRUB*

✓ **High Computational Cost**
✓ **Dependence on the Full Dataset**

- **2. Training Incorrect Knowledge**
  - Unintentional confusion in the performance of unlearned DNN (Side Effects)
  - *i.e., Random Label, Boundary Unlearning, SalUn,*

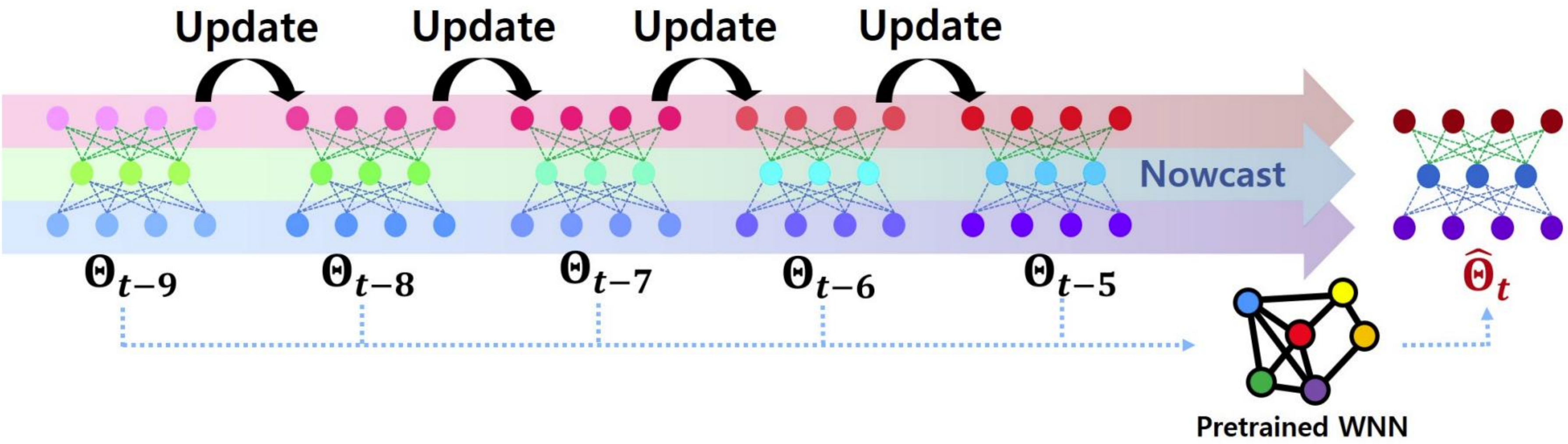
✓ **Confusion rather than Knowledge Deletion**

- **3. Approximation of less learned state**
  - Less accurate approximation of unlearned weights (Worse Deletion)
  - *i.e., NegGrad, Task Vector*

✓ **Insufficient Precision**

Proposed Method

The Concept of Weight Nowcasting Network (WNN)

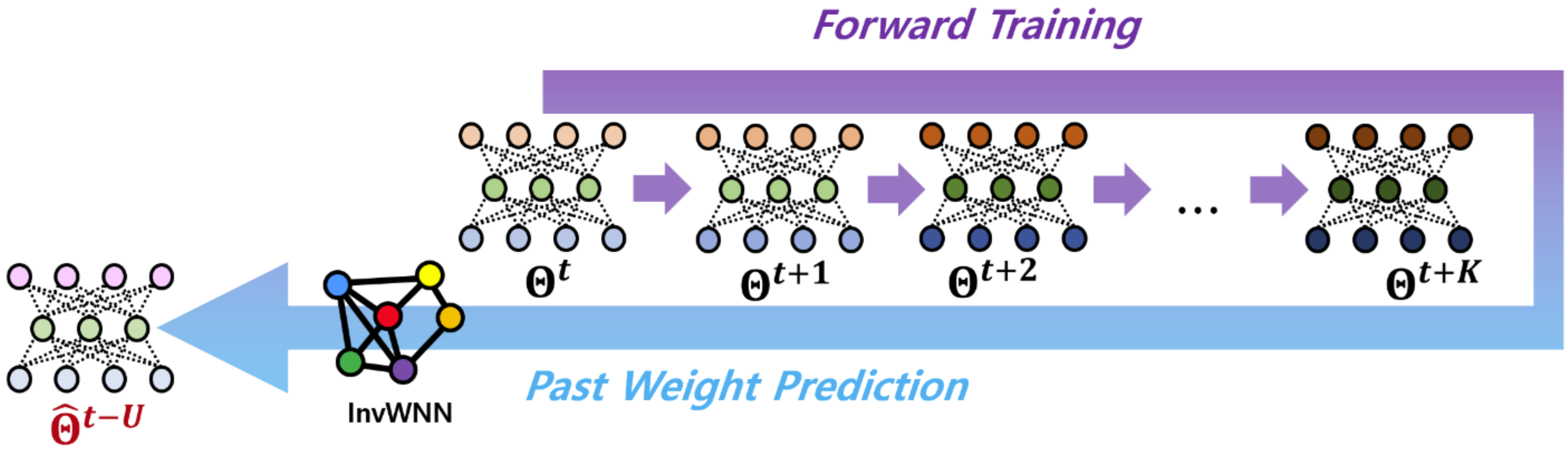


- ✓ **An add-on network which learns the general tendency of weight changes during training DNNs to accelerat e training process**
- ✓ ***To train WNN, the authors collected numerous histories of NNs with various settings***
- ✓ ***Then, they trained a regression model predicting future weights based on current history***

We adapted WNN from its original goal of Acceleration of training process to facilitating Machine Unlearning.

✓ **For adapting, we modified the task of WNN to Predict Past Weights**

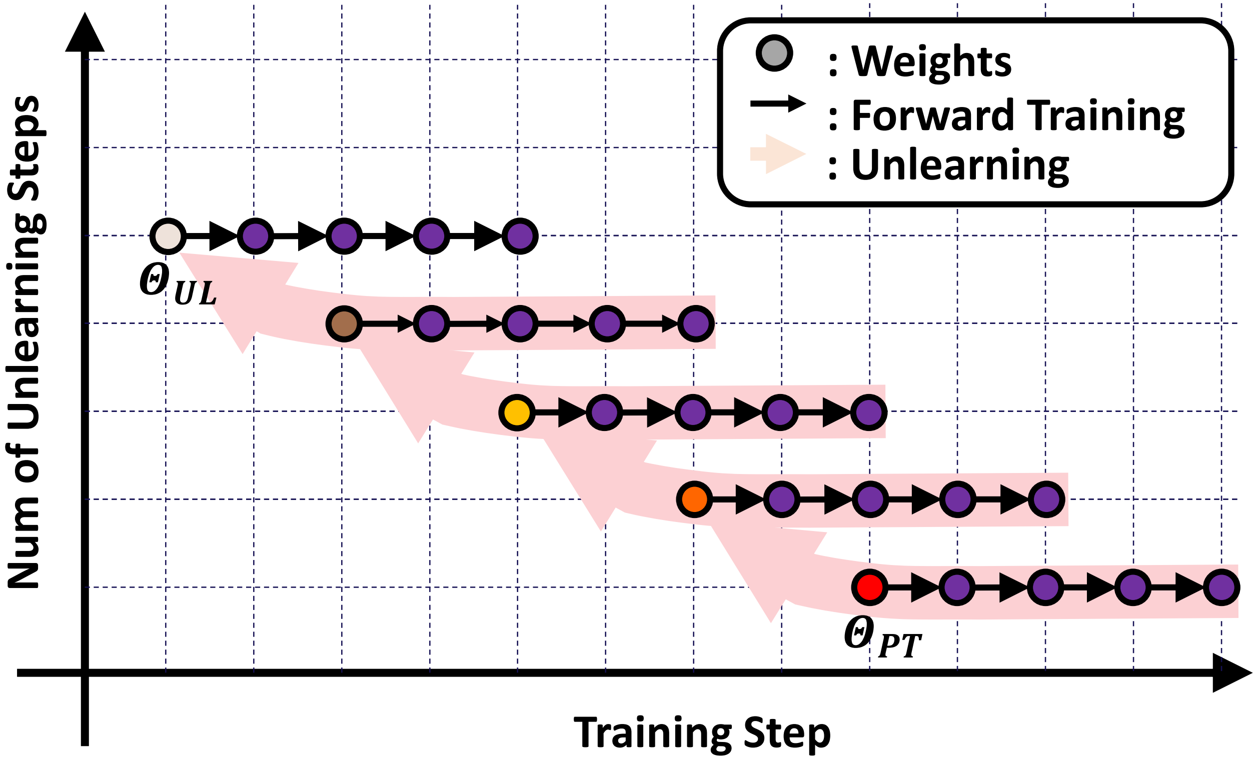
Unlearning Procedure



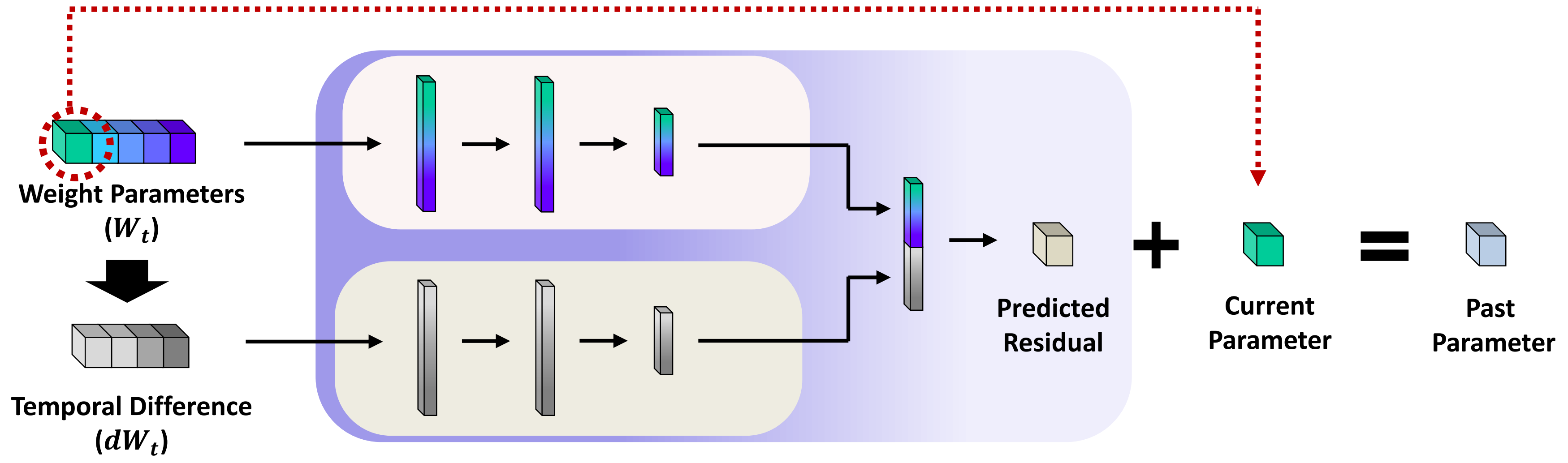
- **Our Unlearning involves:**
  - **Forward training to obtain the future trajectory**
  - **Coordinate-wise Past Weight Prediction *(Element-wise)***

Unlearning Procedure

- Repetition of two procedures for **Gradual Unlearning**



InvWNN



<InvWNN>

- **Using the collected training histories of various DNNs, we trained an ad-hoc model predicting the prior weights using the future training trajectory.**

Experiments

Standard Unlearning Experiment

- CIFAR10 with ResNet18
- Unlearn the half of training data
  - RA: Remaining Accuracy (Acc(D_r))
  - UA: Unlearning Accuracy (100%-Acc(D_f))
  - TA: Test Accuracy
  - MIA: Ratio of D_f classified as Unseen Data

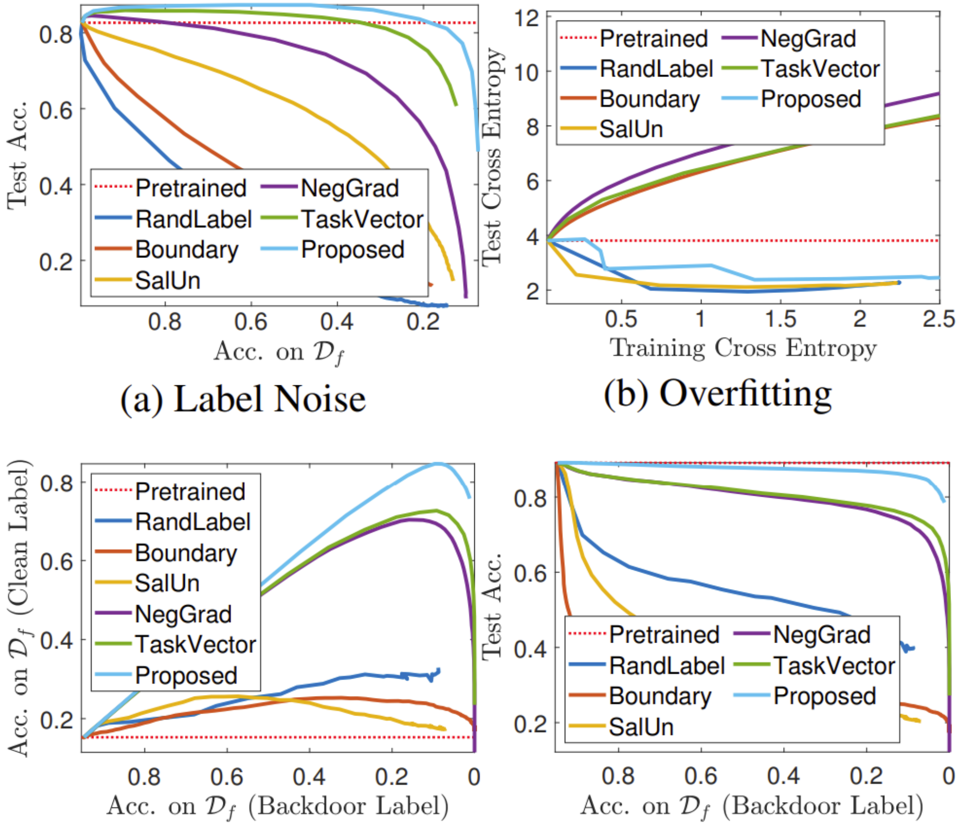
Method	Termination Condition: UA > 8.11%			
	RA (%)	UA (%)	TA (%)	MIA (%)
Pretrained	99.96±0.02	0.02±0.01	94.38±0.10	9.75±1.61
Retrain	99.92±0.02	8.11±0.19	91.45±0.32	29.07±0.42
Random Label	91.03±0.55	8.99±0.53	83.63±0.44	33.19±2.46
Boundary	91.24±0.21	8.42±0.16	83.72±0.10	28.51±3.97
SalUn	91.33±0.45	8.58±0.39	84.45±0.35	32.62±3.17
NegGrad	91.26±0.31	8.59±0.11	85.01±0.95	**28.56±9.60**
Task Vector	92.03±0.72	8.54±0.54	83.29±0.82	30.86±6.65
Proposed	**92.24±0.71**	8.31±0.24	**85.04±1.10**	31.94±4.56

Table 1: Results of five trials for unlearning randomly selected 50% of CIFAR10. Note that better performance corresponds to a smaller gap with the retrained model.

The proposed method achieved the Higher RA and TA with the similar UA

Side Effect Removal

- For all problems, we used CIFAR10 and ResNet18
- Label Noise
  - Pretraining on 80% clean and 20% label-noised data
  - Unlearning the noised data
- Overfitting
  - Pretraining on 1% of entire training data with excessive iterations
  - Unlearning the 1% data
- Backdoor Attack
  - Pretraining on 80% clean and 20% BadNet backdoor data
  - Unlearning the adversarial data



(a) Label Noise

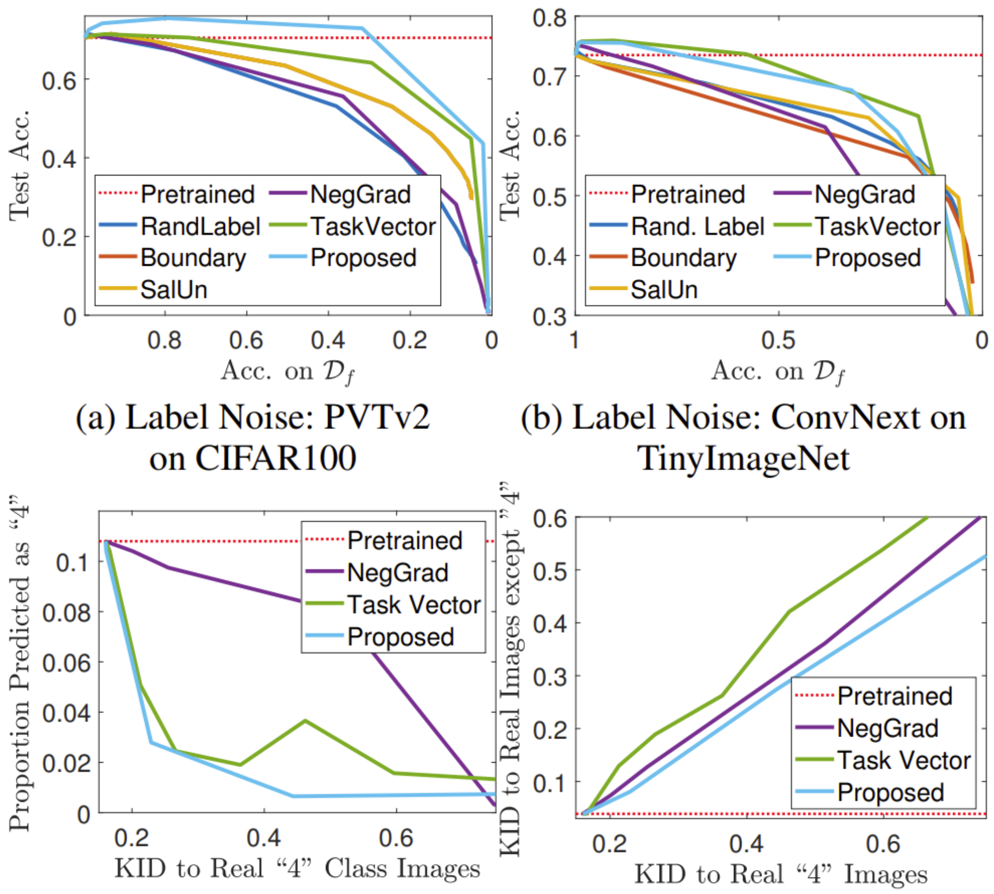
(b) Overfitting

(c) Backdoor: Damage recovery

(d) Backdoor: Preserving test accuracy

Generalization

- Extension to more challenging cases
  - Datasets: CIFAR100 and TinyImageNet
  - Architecture: PVTv2 and ConvNext
- Diffusion Model (U-Net)
  - Pretraining on MNIST
  - Unlearning only “4” images



(a) Label Noise: PVTv2 on CIFAR100

(b) Label Noise: ConvNext on TinyImageNet

(c) Diffusion: Ratio of images classified as “4”

(d) Diffusion: KID to all real images

Figure 5: Unlearning across novel datasets, architectures, and tasks. For diffusion, a lower ratio (left) and lower KID (right) at the same x-coordinate indicate better unlearning.

Visualization

- Class Activation Map for Backdoor Attack Case

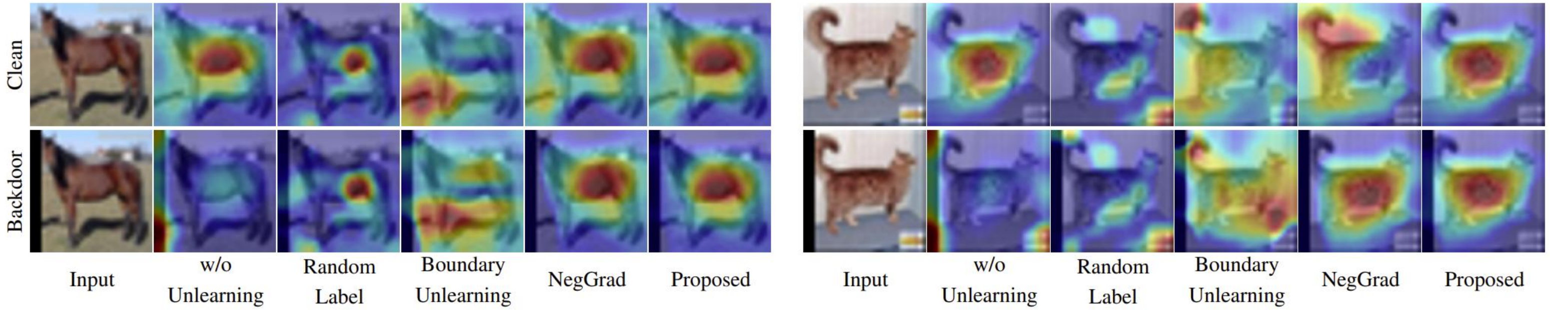
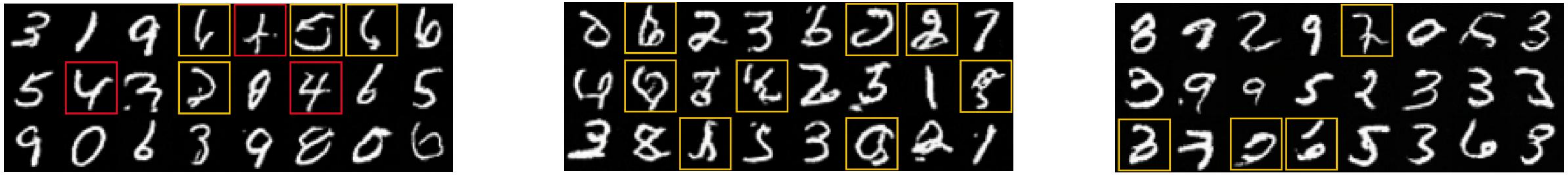


Figure 4: CAM for backdoor data of each unlearning method. The black line at left of backdoor images is the hidden signature.

- Diffusion Model



NegGrad

Task Vector

Proposed

Visualization

- Unlearning Trajectory

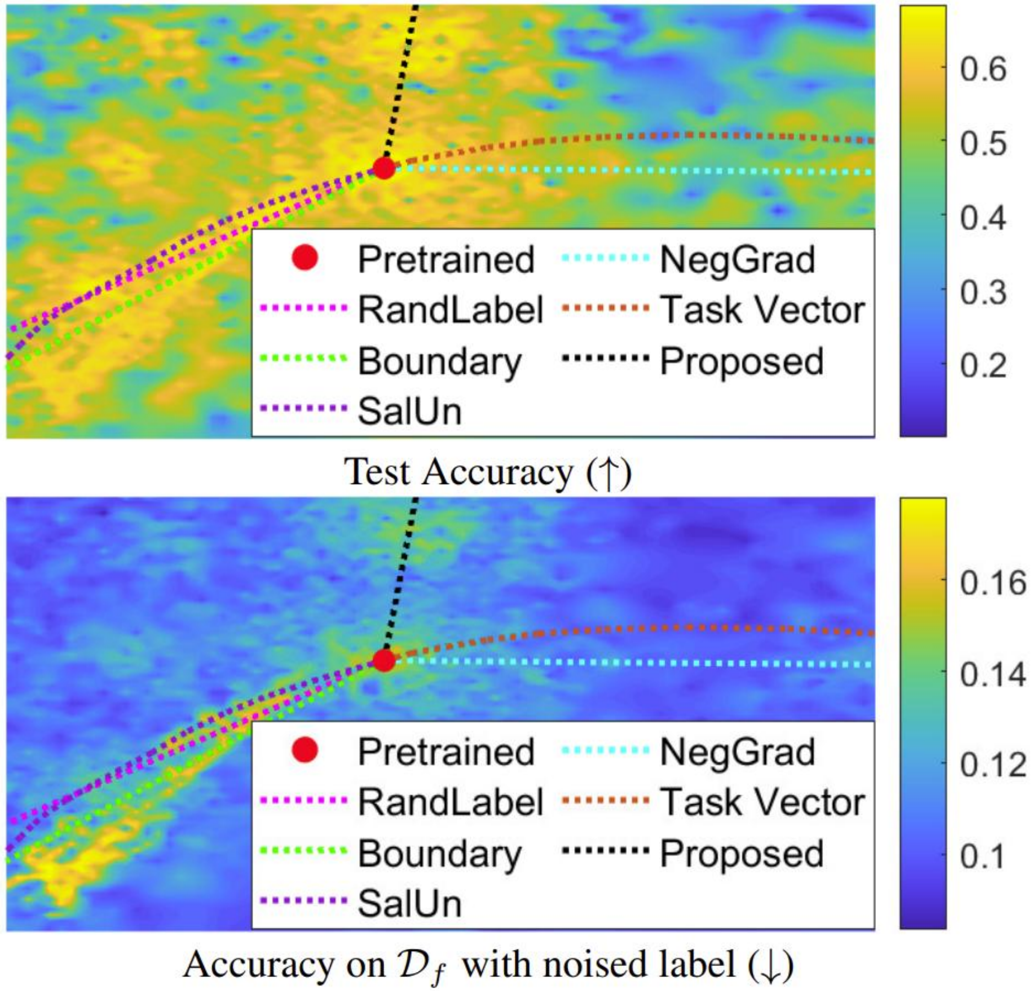


Figure 7: Landscape visualization of unlearning on label-noised data. The upper shows the test accuracy landscape, reflecting the ability to preserve accuracy. The lower depicts accuracy on $\mathcal{D}_f$, representing forgetting performance.

Conclusion

The contributions of this work are summarized as follows:

- 1) We apply the concept of **weight prediction** to machine unlearning.
- 2) We establish an evaluation protocol based on **side effect removal**.
- 3) Our method can **generally** work **using only forget data**
- 4) In our experiments, the proposed method outperforms previous unlearning approaches in standard unlearning scenarios and in removing side effects.

Thanks

Poster **#138**: 12:30–2:30, today

Feel free to contact me at jangjh6297@gmail.com

Our repository: