ECO: Efficient Computational Optimization for Exact Machine Unlearning in Deep Neural Networks

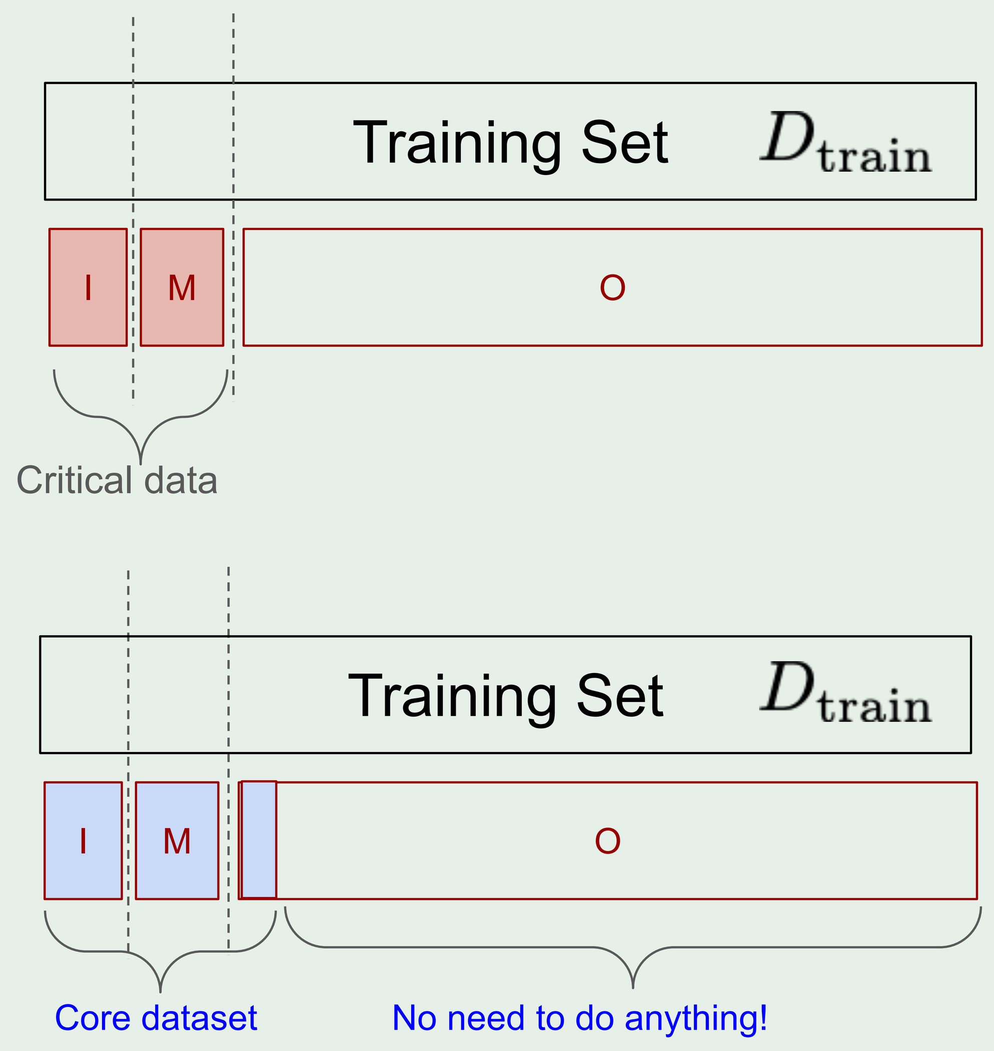
Yu-Ting Huang[1,2] Pei-Yuan Wu[1] Chuan-Ju Wang[2]

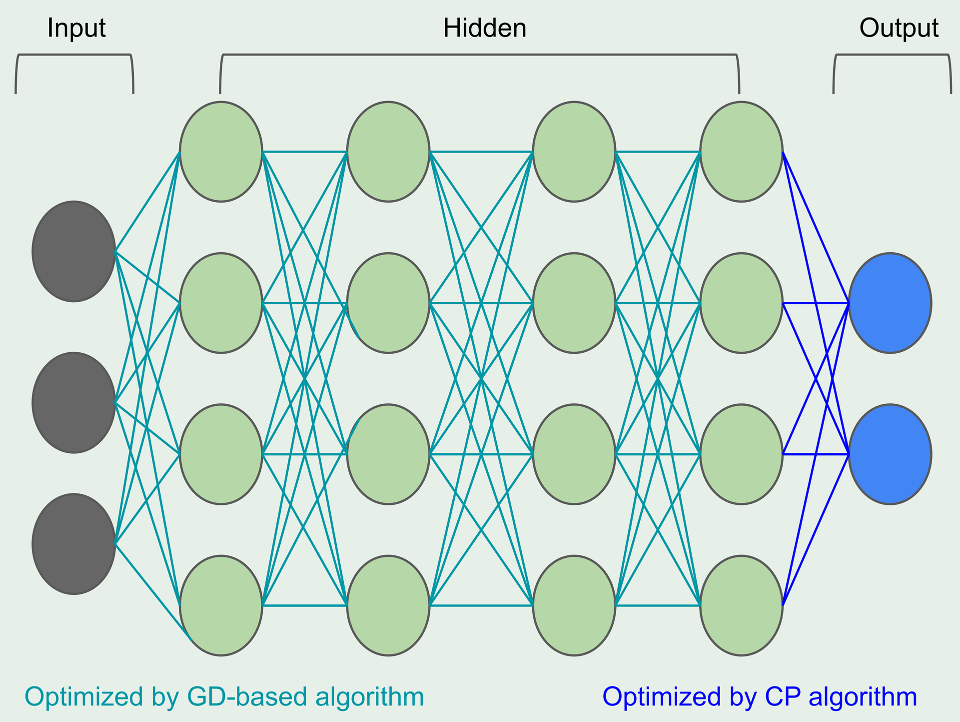
[1]National Taiwan University [2]Academia Sinica

Abstract

This work introduces ECO, an efficient computational optimization framework that adapts the CP algorithm—originally proposed by Cauwenberghs & Poggio (2000)—for unlearning within deep neural network models.

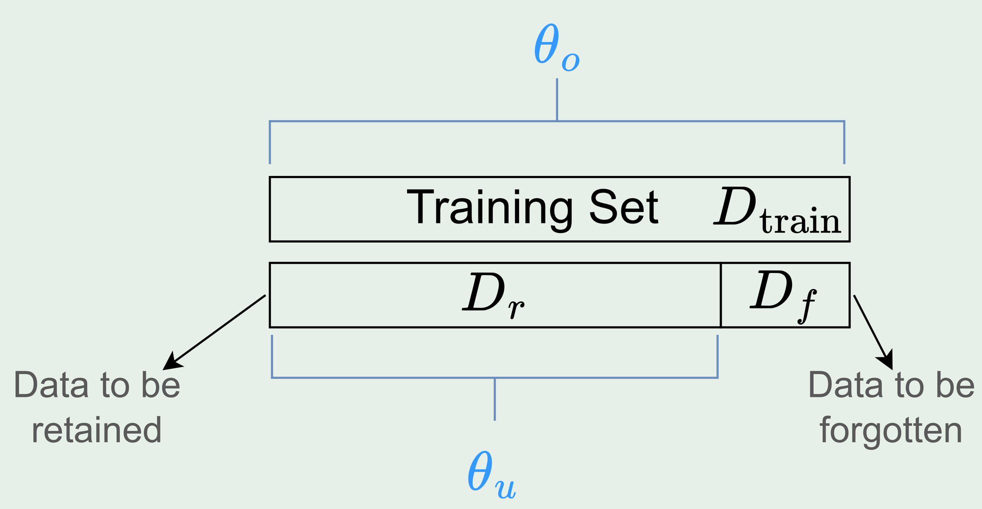
- ECO simplifies labor-intensive tasks, significantly reducing the workload for service providers compared to previous exact unlearning methods for DNNs.
- We demonstrate that ECO not only boosts efficiency but also maintains the performance of the original base DNN model, and surprisingly, it even surpasses naive retraining (NR) in effectiveness.
- Crucially, we are the first to adapt the CP algorithm's decremental learning for leave-one-out evaluation to achieve exact unlearning in DNN models. We also open-source a usable base code for the CP algorithm, addressing the previous lack of such resources and encouraging further research and practical applications.





Problem Definition

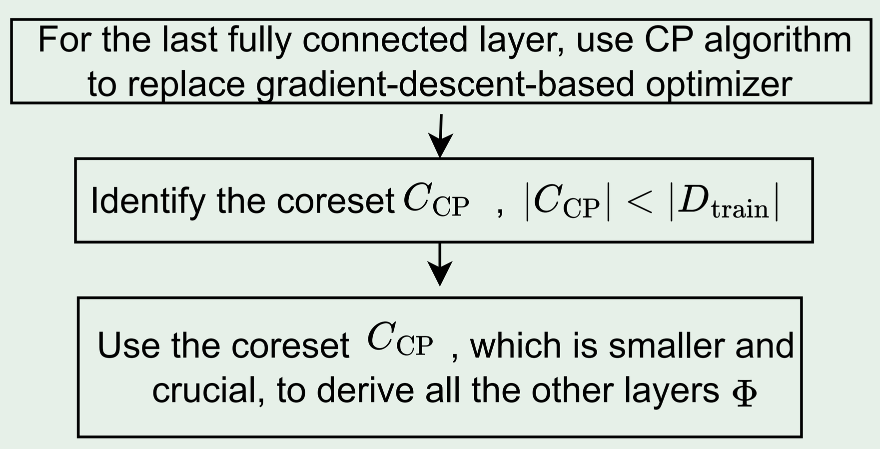
Machine unlearning aims to eliminate the influence of specific training data from an already-trained machine learning model θ_o.



The primary challenge of machine unlearning is to develop an efficient and effective method to transition from θ_o to θ_u.

Our Proposed Method: ECO

Model Preparation:



Model Serving:

Input $\mathcal{D}_f$, $f_{\mathcal{C}_{CP}}(x)$, $\mathcal{C}_{CP}$
Output $f_{\mathcal{C}_{CP}}(x), \mathcal{C}_{CP}$

```
1:  if D_f ∩ C_CP ≠ ∅ then
2:      if D_f ∩ (M ∪ I) ≠ ∅ then
3:          α, b ← Employ Algorithm 3 in our paper to unlearn D_f
4:          C_CP ← Construct new C_CP via (9) in our paper
5:      else
6:          C_CP ← C_CP \ D_f
7:      end if
8:      Φ_{C_CP} ← Learn a new feature transformation function Φ_{C_CP}(x) with (10) in our paper
9:  else
10:     Remain the input model f_{C_CP}(x) and the set C_CP
11: end if
```

Main Results

We compare the proposed method, ECO, with the gold standard naive retraining (NR), which involves retraining the DNN models upon receiving an unlearning request.

Table 1. In-time unlearning

	NR	ECO$_i$	ECO
Acc$_r$ ↑	0.988 ± 0.005	<u>0.996 ± 0.001</u>	**0.997 ± 0.001**
Acc$_f$ ↑	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**
Acc$_{test}$ ↑	0.986 ± 0.002	**0.989 ± 0.001**	**0.989 ± 0.001**
Acc$_{all}$ ↑	0.974 ± 0.007	<u>0.986 ± 0.002</u>	**0.987 ± 0.002**
Time cost (sec) ↓	20.968 ± 7.992	**1.493 ± 4.476**	<u>1.842 ± 5.524</u>

Table 2. Off-time batch unlearning

	Acc$_{all}$ (= Acc$_r$× Acc$_f$× Acc$_{test}$) ↑		
p	NR	ECO$_i$	ECO
1	0.965 ± 0.006	<u>0.974 ± 0.007</u>	**0.979 ± 0.003**
5	0.964 ± 0.005	<u>0.973 ± 0.004</u>	**0.979 ± 0.002**
10	0.963 ± 0.006	<u>0.974 ± 0.005</u>	**0.978 ± 0.004**
20	0.960 ± 0.005	<u>0.973 ± 0.002</u>	**0.978 ± 0.004**
30	0.961 ± 0.005	<u>0.973 ± 0.002</u>	**0.978 ± 0.003**
40	0.961 ± 0.005	<u>0.971 ± 0.003</u>	**0.979 ± 0.003**
50	0.961 ± 0.006	<u>0.966 ± 0.004</u>	**0.979 ± 0.004**

All the metrics mentioned above are expressed as $a \pm b$, where 'a' represents the mean and 'b' denotes the standard deviation across 10 independent trials with different random seeds. The symbol '↑' indicates that higher values are better, and '↓' indicates that lower values are preferable. The best result is highlighted in bold, and the second-best result is underlined.

The Forgetfulness Quality: MIA

Table 3. The forgetfulness quality (MIA)

	NR		ECO$_i$		ECO	
p	$\mathcal{D}_f$	$\mathcal{D}_{test}$	$\mathcal{D}_f$	$\mathcal{D}_{test}$	$\mathcal{D}_f$	$\mathcal{D}_{test}$
1	0.06 ± 0.01	0.07 ± 0.0	0.15 ± 0.25	0.16 ± 0.24	0.03 ± 0.01	0.04 ± 0.01
5	0.14 ± 0.26	0.15 ± 0.26	0.07 ± 0.02	0.08 ± 0.02	0.03 ± 0.01	0.04 ± 0.01
10	0.15 ± 0.26	0.15 ± 0.26	0.07 ± 0.02	0.08 ± 0.01	0.04 ± 0.01	0.05 ± 0.01
20	0.32 ± 0.39	0.32 ± 0.39	0.07 ± 0.02	0.07 ± 0.01	0.04 ± 0.01	0.04 ± 0.01
30	0.15 ± 0.26	0.15 ± 0.26	0.06 ± 0.01	0.07 ± 0.01	0.04 ± 0.01	0.05 ± 0.01
40	0.15 ± 0.26	0.15 ± 0.26	0.06 ± 0.02	0.07 ± 0.01	0.04 ± 0.01	0.04 ± 0.01
50	0.24 ± 0.34	0.24 ± 0.35	0.08 ± 0.02	0.08 ± 0.02	0.04 ± 0.01	0.05 ± 0.01

Table 3 compares the MIA scores between $\mathcal{D}_f$ and $\mathcal{D}_{test}$ for each unlearning approach. A lower discrepancy between these two values indicates better forgetfulness quality of the model. For the NR method, there is hardly any disparity between $\mathcal{D}_f$ and $\mathcal{D}_{test}$. This pattern is also consistent for ECO$_i$ and ECO. As expected, this alignment occurs because all three methods adhere to an exact unlearning approach.

- For more details, please refer to: `https://openreview.net/pdf?id=SeBVP0zxKp`

 r11946008@ntu.edu.tw