Google DeepMind

UNIVERSITY OF CAMBRIDGE

Redirection for Erasing Memory (REM): Towards a universal unlearning method for corrupted data

Stefan Schoepf, Michael Curtis Mozer, Nicole Elyse Mitchell, Alexandra Brintrup, Georgios Kaissis, Peter Kairouz, Eleni Triantafillou

When and why do unlearning methods fail?

- SOTA methods work well on their original benchmark tasks
- But on unseen tasks they fail/succeed unpredictably

➔ Severely limits real life deployment without knowing what works/fails

➔ Future research is limited by a lack of failure mode understanding

Our contributions

Contribution to knowledge:

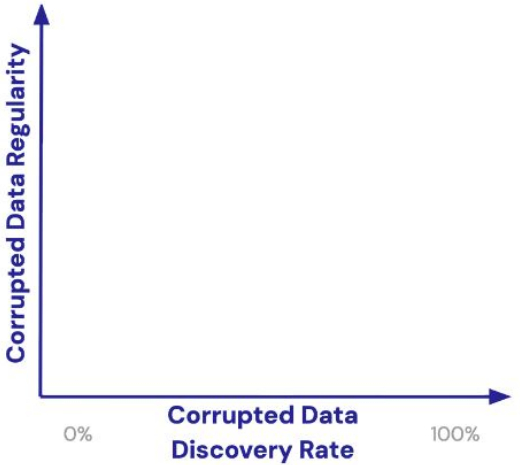
- An experimentally validated 2D-framework showing method failures
- Provides key insights and benchmark tasks for method development

Method contribution:

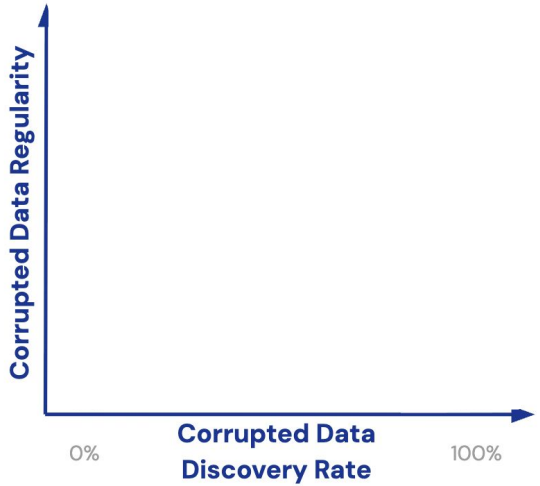
- The first unlearning method to perform strongly across this 2D-space
- Sets a new SOTA and brings unlearning closer to real-life deployment

Contribution to knowledge: 2D-Framework

- **Discovery Rate** (Goel et al)
  - In practice we will not find 100% of the corrupted data we want to unlearn (e.g. poisoned samples)
  - Undiscovered samples reintroduce corruptions during the unlearning process
- **Statistical Regularity** (new)
  - High: Same backdoor trigger on images
  - Low: Random label errors with no shared pattern
  - Unlearning methods are sensitive to the regularity of to be unlearned data
- **Interplay** of discovery rate and regularity
  - Regularity impacts the magnitude of discovery rate effects on unlearning performance



Contribution to knowledge: 2D-Framework



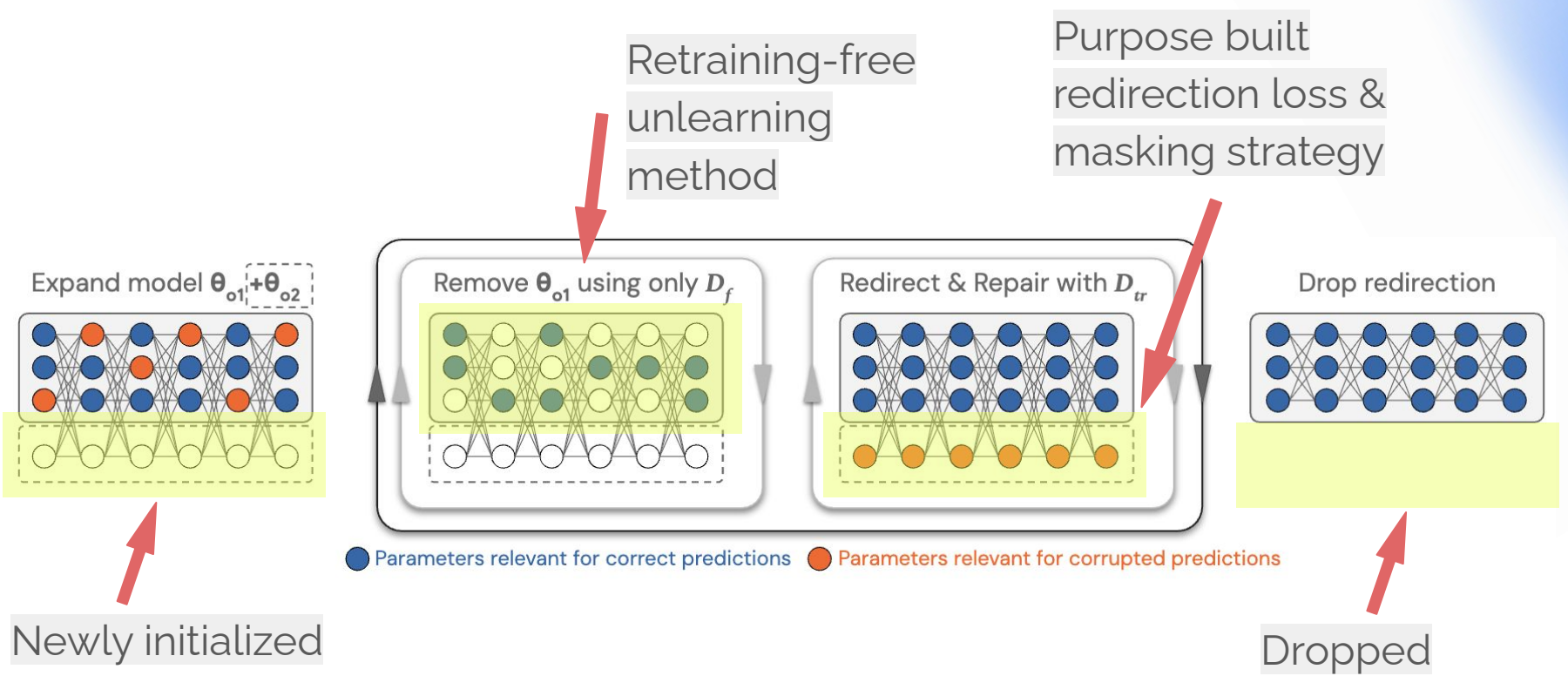
(a) Framework for corrupted data unlearning tasks with prior work areas highlighted

Method contribution: REM

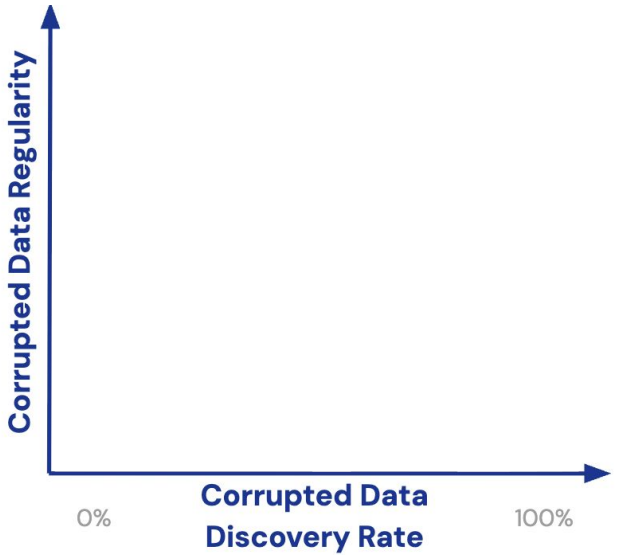
- ❏ Repair steps reintroduce corruptions with partial discovery
- ❏ No repair steps lead to low model utility (e.g. accuracy)
- ❏ Regularity of data severely impacts filtering/selection algorithms

→ REM redirects corrupted data to dedicated neurons introduced at unlearning time and then discards them to unlearn.

Method contribution: REM

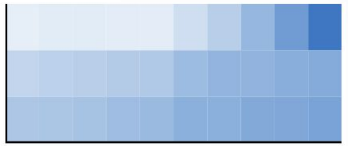


Experimental results

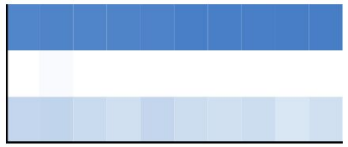


(a) Framework for corrupted data unlearning tasks with prior work areas highlighted

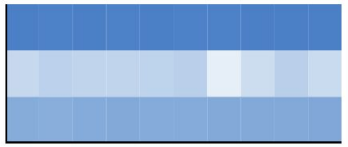
(b) Retrained from scratch



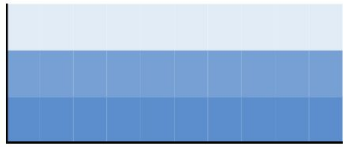
(e) SCRUB



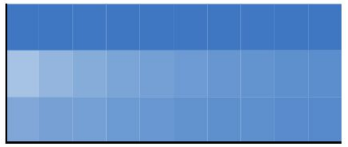
(c) Potion



(f) Gradient Ascent



(d) ETD



(g) REM (ours)

Conclusion

Our work adds two key contributions to the unlearning literature:

- ➔ Knowledge: 2D taxonomy
  - ◆ When do methods fail
- ➔ Method: REM
  - ◆ First method to perform strongly across the 2D-space

Code: https://github.com/google-deepmind/rem