Machine Unlearning: Research Learning Path

PhD Student Guidance

March 30, 2025

Phase 1: Foundations (1–2 Months)

Core Concepts to Study

- **Machine Learning Basics (Refresher)**
  - Supervised, unsupervised, and reinforcement learning
  - Neural networks, deep learning, and model training pipelines
  - Overfitting, generalization, and memorization in ML
- **Privacy & Data Regulations**
  - GDPR, CCPA, and "right to be forgotten"
  - Data retention policies and ethical considerations
- **Introduction to Machine Unlearning**
  - Definition and motivation
  - Differences between **unlearning, deletion, and retraining**
  - Key papers:
    * Start with the seminal "Machine Unlearning" by Cao & Yang (2015)
    * Study "Certified Data Removal from Machine Learning Models" by Guo et al. (2020)
    * Review "SISA: Set-based Incremental Learning" by Bourtoule et al. (2021)

    Once you have read these key papers, read the following survey papers to get the pulse of the field:

    1. Xu et al. (2023) *"Machine Unlearning: A Survey"*
    2. Wang et al. (2024) *"Machine Unlearning: A Comprehensive Survey"*

Exploration Tasks

- Summarize 3 key papers on unlearning.
- Implement a simple unlearning method (e.g., retraining from scratch vs. influence-based removal).

Phase 2: Advanced Topics (2–3 Months)

Key Research Areas in Unlearning

- **Exact vs. Approximate Unlearning**
  - Exact unlearning (full retraining) vs. approximate (influence functions, gradient updates)
  - Trade-offs: **computational cost vs. privacy guarantees**
- **Unlearning in Different Models**
  - Linear models (logistic regression)
  - Deep neural networks (CNNs, transformers)

  - Federated learning settings
- **Verification & Attacks**
  - How to **verify** if unlearning worked?
  - Membership inference attacks on unlearned models

Exploration Tasks

- Implement SISA (Sharded, Isolated, Sliced, Aggregated) on a toy dataset.
- Compare unlearning methods on a small benchmark (MNIST/CIFAR-10).

Phase 3: Problem Identification (1–2 Months)

Open Challenges in Machine Unlearning

- **Incorrectly Learned Information**
  - How does a model "learn incorrectly"? (e.g., bias, noise, adversarial examples)
  - Can unlearning fix **backdoor attacks** or **biased representations**?
- **Efficiency vs. Completeness Trade-off**
  - Can we unlearn **without full retraining**?
  - How to handle **sequential unlearning requests**?
- **Theoretical Guarantees**
  - Differential privacy connections
  - Certifiable unlearning

Exploration Tasks

- Identify **3 gaps** in current unlearning methods.
- Brainstorm **novel approaches** (e.g., using influence functions, model pruning, or RL for selective forgetting).

Phase 4: Research Development (Ongoing)

Possible Research Directions

- **Dynamic Unlearning for Streaming Data**
  - How to handle continuous data corrections?
- **Unlearning in Foundation Models (LLMs)**
  - Can we "edit" knowledge in GPT-like models?
- **Human-in-the-Loop Unlearning**
  - Interactive systems for correcting model errors.

Next Steps

- Define a **specific problem statement** (e.g., "Efficient Unlearning in Deep Recommender Systems").
- Develop a **prototype** and compare with baselines.
- Start writing **survey paper** or **proposal**.

Resources & Tools

- **Datasets**: MNIST, CIFAR-10, Amazon Reviews (for privacy-sensitive unlearning).
- **Libraries**: PyTorch, TensorFlow, `HuggingFace` (for LLM experiments).
- **Conferences**: NeurIPS, ICML, ICLR (look for recent unlearning papers).
- **GitHub**: Machine Unlearning Papers.