Machine Unlearning: Metodologie di Mitigazione

Il machine unlearning emerge come soluzione critica per affrontare le sfide della privacy e della sicurezza nei modelli di machine learning. Questo articolo esplora le metodologie di mitigazione, le tecniche implementative e i criteri di valutazione per rimuovere efficacemente dati specifici dai modelli addestrati, senza necessità di riaddestramento completo. Vengono analizzati gli scenari di applicazione, dalle normative sulla privacy come il GDPR alla mitigazione dei bias, evidenziando l'importanza di questo approccio nell'evoluzione dell'AI responsabile.

Stiamo assistendo ad una crescita esponenziale nell'utilizzo del machine learning per lo svolgimento di qualsiasi attività. I modelli di machine learning (ML) si stanno sempre di più integrando in molti prodotti e servizi di uso quotidiano. Tuttavia, la proliferazione della tecnologia di AI/ML solleva una serie di problemi in termini di violazione della privacy, di bias (pregiudizi) del modello e dell'uso non autorizzato dei dati utilizzati per il training dei modelli.

Questi ambiti evidenziano ancora una volta la necessità di adottare un controllo flessibile e reattivo sui dati con cui viene addestrato un modello. Il riaddestramento da zero di un modello di ML, per la rimozione di specifici dati, è poco pratico a causa degli elevati costi computazionali ed economici. Pertanto, è stata avviata la ricerca sul *"machine unlearning"* (MU), ovvero si sta cercando di sviluppare nuovi metodi in grado di rimuovere determinati dati da un modello addestrato in modo efficiente, efficace e, soprattutto, senza la necessità di realizzare un nuovo addestramento. Analizziamo le criticità del machine unlearning e le metodologie di valutazioni più efficaci.

Scenari di Mitigazione dei Rischi nel Machine Unlearning

Il machine unlearning (MU) rappresenta un tema rilevante, che non può essere sottovalutato, perché è in grado di affrontare una serie di problemi quali: la conformità alla privacy dei dati di training, la gestione dinamica dei dati, l'inclusione non autorizzata di dati tutelati dalla proprietà intellettuale o altre forme di tutela e, infine, l'uso di dati provenienti da data breaches.

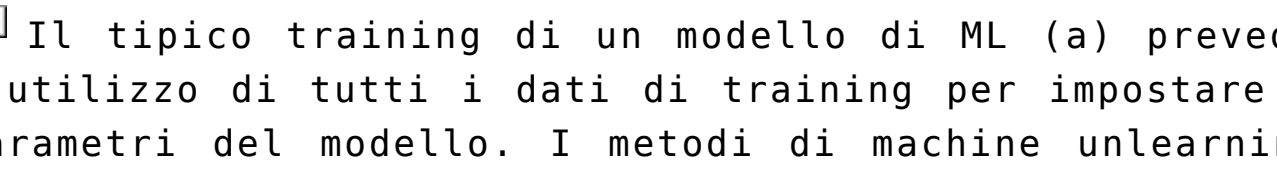
Vediamo i principali use case del machine unlearning:

- **Protezione della privacy**: il MU può svolgere un ruolo cruciale nell'applicazione dei diritti alla privacy e nel rispetto di normative come il GDPR dell'UE[1] (che include anche il diritto all'oblio). Consentirebbe di rimuovere i dati personali dai modelli addestrati, salvaguardando la privacy degli individui.
- **Incremento della sicurezza**: il MU potrebbe migliorare la sicurezza dei modelli contro gli attacchi di *"data poisoning"* attraverso la rimozione dei dati c.d. tossici che mirano a manipolare il comportamento di un modello.
- **Miglioramento dell'adattabilità**: il MU su larga scala potrebbe aiutare i modelli a rimanere pertinenti anche quando cambiano le distribuzioni dei dati nel tempo, ad esempio in base all'evoluzione delle preferenze dei clienti o delle tendenze di mercato.
- **Conformità normativa**: nei settori fortemente regolamentati, il MU potrebbe essere fondamentale per mantenere la conformità alle leggi e ai regolamenti in continua evoluzione.

Mitigazione dei bias/pregiudizi: il MU potrebbe offrire un metodo per rimuovere i dati identificati dopo il training che creano distorsione, promuovendo così l'imparzialità e riducendo il rischio di risultati scorretti.

Tecniche di Mitigazione e Implementazione

La maggior parte delle implementazioni di machine unlearning suddividono il data set utilizzato per il training originario (“*Dtrain*”) in dati che devono essere conservati (“*retain set*” o Dr) e dati che devono essere dimenticati/rimossi (“*forget set*” o Df), come mostrato in Figura.

 Il tipico training di un modello di ML (a) prevede l’utilizzo di tutti i dati di training per impostare i parametri del modello. I metodi di machine unlearning comportano la suddivisione dei dati di training (Dtrain) in *retain set* (Dr) e *forget set* (Df), quindi un utilizzo iterativo di questi set per modificare i parametri del modello (passaggi da b a d in maniera ciclica). La sezione gialla rappresenta i dati che sono stati dimenticati durante il processo.

Successivamente, i due set vengono utilizzati per modificare i parametri del modello sottoposto a training. Nel corso del tempo, i ricercatori hanno sviluppato diverse tecniche per migliorare questa fase di *unlearning*, analizziamole:

- **Ottimizzazione**: la tecnica prevede che il modello sia ulteriormente addestrato sul retain set, consentendogli di adattarsi alla nuova distribuzione dei dati. Questa tecnica è semplice, ma può richiedere molta potenza di calcolo[2].
- **Etichettatura casuale**: la tecnica prevede l’assegnazione di etichette errate casuali al forget set in grado di confondere il modello. In questo modo, il modello viene ottimizzato[3] [4].
- **Inversione del gradiente**: questa tecnica prevede che, durante l’ottimizzazione del modello, sia invertito il segno dei gradienti di aggiornamento del peso per i dati del forget set. Questa tecnica contrasta il training precedente[5].
- **Riduzione selettiva dei parametri**: questa tecnica

prevede la riduzione selettiva dei parametri specificamente legati al forget set attraverso tecniche di analisi del peso e senza alcuna ottimizzazione[6].

Le scelte delle tecniche di unlearning riflettono i casi d'uso di unlearning. Ogni use case ha i propri requisiti che riguardano, in particolare, l'efficacia, l'efficienza e i problemi di privacy.

Valutazione e privacy e Sicurezza nel Machine Unlearning

Una delle principali difficoltà del machine unlearning consiste nel valutare se la tecnica di unlearning prescelta sia in grado di dimenticare i dati specificati e, contestualmente, mantenere le performance sui dati conservati e, inoltre, proteggere la privacy. Idealmente, un metodo di unlearning automatico dovrebbe produrre un modello che funzioni come se fosse stato addestrato da zero, quindi privo del set di dati da dimenticare.

I metodi di unlearning più utilizzati (come l'etichettatura casuale, l'inversione del gradiente e la riduzione selettiva dei parametri) provocano un degrado delle prestazioni del modello esattamente nei punti corrispondenti al dataset dei dati da dimenticare, mentre cercano di mantenere un alto grado di prestazioni del modello nei corrispondenti punti del dataset da mantenere.

Per semplicità, si potrebbe considerare un metodo di unlearning basato su due semplici obiettivi: ottenere prestazioni elevate sul set da conservare e scarse sul set dell'oblio. Questo approccio, tuttavia, rischierebbe di aprire un'altra superficie di attacco sul fronte della privacy: per esempio, se un modello unlearned funzionasse particolarmente male con un determinato input, ciò potrebbe far capire all'aggressore che in origine l'input fosse incluso nel set di dati utilizzati nel training originale e che sia stato rimosso

successivamente.

Questo tipo di violazione della privacy, chiamato attacco di *"membership inference"* [7], è in grado di rivelare i dati importanti e sensibili di un determinato utente o su un set di dati specifico. Pertanto, durante la valutazione dei metodi di unlearning automatico, è importante testare la loro efficacia contro questa tipologia di attacco.

In questo contesto di analisi, i termini *"stronger"* e *"weaker"* si riferiscono alla complessità e all'efficacia dell'attacco:

Weaker attacks (Attacchi più deboli): si tratta di tentativi semplici e diretti per ricavare l'appartenenza inferenziale. Questi attacchi potrebbero fare affidamento su informazioni basilari, come i punteggi di confidenza del modello o la probabilità di un determinato output per un input definito. Spesso questi attacchi si basano su ipotesi semplificatrici del modello o sulla distribuzione dei dati, il che può fortunatamente limitarne l'efficacia.

Stronger attacks (Attacchi più forti): si tratta di attacchi sofisticati che utilizzano informazioni o tecniche più avanzate. In particolare, sono in grado di:

- utilizzare più punti di query oppure input particolarmente elaborati,
- sfruttare la conoscenza dell'architettura del modello o del processo di training,
- utilizzare modelli ombra per comprendere il comportamento del modello target,
- combinare più strategie di attacco,
- adattarsi alle specifiche caratteristiche del modello target o del set di dati.

Generalmente, gli attacchi più forti sono più efficaci nell'inferenza di appartenenza e quindi sono più difficili da contrastare. Inoltre, rappresentano un modello di minaccia realistico, presente in molti scenari del mondo reale, in cui

gli aggressori potrebbero disporre di importanti risorse e competenze.

Raccomandazioni

Il mondo della ricerca sta continuando a lavorare per lo sviluppo di nuove soluzioni di machine unlearning che siano aderenti agli ambienti di produzione e, inoltre, vengano sottoposti ad attacchi alla privacy o all'oblio dei dati più realistici. Uno studio recente offre una serie di raccomandazioni per migliorare la valutazione dell'unlearning basate sulla letteratura già esistente, in particolare vengono proposti nuovi benchmark e riprodotti diversi algoritmi di unlearning. In particolare, sono stati valutati gli algoritmi per misurare l'accuratezza sui dati conservati, la protezione della privacy dei dati dimenticati e la velocità di esecuzione del processo di unlearning[8].

I test hanno rivelato notevoli discrepanze tra gli algoritmi di unlearning, in molti casi sono state riscontrate difficoltà nel raggiungere il risultato in tutte le aree di valutazione. Sono stati testati tre metodi (Identity, Retrain e Finetune on retain) e cinque algoritmi (RandLabel[9], BadTeach[10], SCRUB+R [11] , Selective Synaptic Dampening [SSD][12] e una combinazione di SSD e finetuning).



Come evidenziato in figura, alcuni metodi riescono a difendersi bene dagli attacchi di inferenza debole, mentre sono completamente inefficaci contro gli attacchi più forti, evidenziando la necessità di effettuare il test del caso peggiore. Alla luce di questi risultati, è stata dimostrata l'importanza di dover effettuare la valutazione degli algoritmi in contesti iterativi, poiché alcuni algoritmi degradano la loro accuratezza complessiva durante le iterazioni di unlearning, mentre altri sono in grado di mantenere elevate prestazioni in maniera costante.

Sulla base di questi risultati, si suggerisce di adottare le seguenti raccomandazioni:

1)è necessario dare enfasi alle metriche del caso peggiore rispetto a quelle del caso medio e, inoltre, opportuno utilizzare attacchi avversari forti per la valutazione degli algoritmi. Gli utenti sono più preoccupati per gli scenari gravi, come l'esposizione di informazioni finanziarie personali, che non per gli scenari meno gravi. La valutazione delle metriche del caso peggiore fornisce un limite di alto livello sulla privacy.

2)è opportuno considerare tipi di attacchi specifici alla privacy, in cui l'aggressore ha l'accesso all'output da due versioni diverse dello stesso modello. In questi scenari, l'unlearning può portare a risultati peggiori perché stiamo fornendo all'aggressore maggiori informazioni. Se si verificasse un attacco di perdita di aggiornamento, il danno non dovrebbe essere maggiore di un attacco al modello di base.

3)è utile analizzare le prestazioni dell'algoritmo di unlearning su un numero ripetuto di applicazioni, ovvero il disapprendimento iterativo, per essere in grado di valutare il degrado delle prestazioni di accuratezza del test dei modelli. Poiché i modelli di unlearning automatico sono distribuiti in ambienti in continuo cambiamento, in cui le richieste di oblio, i dati dei nuovi utenti e i dati errati o avvelenati arrivano dinamicamente, è fondamentale essere in grado di poterli valutare in un ambiente online, cioè in un ambiente dove le richieste arrivano attraverso un flusso.

Poiché l'intelligenza artificiale sarà sempre più integrata nei vari aspetti della vita, molto probabilmente il machine unlearning automatico diventerà uno strumento essenziale e un complemento alla cura/manutenzione dei dati di training, e sarà in grado di bilanciare le capacità dell'intelligenza artificiale con le minacce relative alla privacy e alla sicurezza dei dati.

Questo aspetto aprirà nuove opportunità per migliorare la protezione della privacy e per lo sviluppo di sistemi di intelligenza artificiale adattabili, inoltre deve essere in grado di affrontare sfide significative, tra cui le limitazioni tecniche e l'elevato costo computazionale di alcuni metodi di unlearning. La ricerca e lo sviluppo stanno svolgendo un ruolo fondamentali per migliorare queste tecniche e garantire che le stesse possano essere efficacemente implementate in scenari reali.

Il machine unlearning si conferma come tecnologia essenziale per bilanciare le potenzialità dell'AI con la protezione della privacy e la sicurezza dei dati. Nonostante le sfide tecniche e computazionali, la ricerca continua a sviluppare soluzioni sempre più efficaci per scenari reali. Nel prossimo capitolo esploreremo le metodologie per misurare e garantire l'affidabilità dei sistemi di AI. Per un'analisi più completa, il white paper "Generative artificial intelligence rischi e contromisure" offre ulteriori approfondimenti sulle sfide e le soluzioni nel campo dell'IA Gen.

Note bibliografiche

[1]General Data Protection Regulation: https://en.wikipedia.org/wiki/General_Data_Protection_Regulation

[2] AA.VV., Machine unlearning through fine-grained model parameters perturbation: https://arxiv.org/abs/2401.04385, College of Computer Science and Electronic Engineering, 2024

[3] AA.VV., Random Relabeling for Efficient Machine Unlearning: https://arxiv.org/abs/2305.12320, Pennsylvania State University, 2023

[4] AA.VV., Amnesiac Machine Learning: https://arxiv.org/pdf/2010.10981, University of Waterloo, 2020

[5] AA.VV., Gone but Not Forgotten: Improved Benchmarks for

Machine Unlearning: https://arxiv.org/pdf/2405.19211, Carnegie Mellon University, 2024

[6]AA.VV., Fast Machine Unlearning Without Retraining Through Selective Synaptic Dampening: https://arxiv.org/abs/2308.07707, University of Cambridge, 2023

[7]ML04:2023 Membership Inference Attack: https://owasp.org/www-project-machine-learning-security-top-10/docs/ML04_2023-Membership_Inference_Attack

[8] AA.VV., Gone But Not Forgotten: Improved Benchmarks for Machine Unlearning: https://arxiv.org/pdf/2405.19211, Carnegie Mellon University, 2024

[9] AA.VV., Amnesiac Machine Learning: https://arxiv.org/abs/2010.10981, University of Waterloo, 2020

[10] AA.VV., Can Bad Teaching Induce Forgetting? Unlearning in Deep Networks using an Incompetent Teacher: https://arxiv.org/abs/2205.08096, Mavvex Lab, 2023

[11] AA.VV., Towards Unbounded Machine Unlearning: https://arxiv.org/abs/2302.09880, University of Warwick, 2023

[12] AA.VV., Fast Machine Unlearning Without Retraining Through Selective Synaptic Dampening: https://arxiv.org/abs/2308.07707, University of Cambridge, 2023

Profilo Autore

È laureato in Ingegneria Informatica ed in Sicurezza Informatica presso le Università di Roma La Sapienza e di Milano. Ha indirizzato la sua formazione nei settori della Cyber Security e Digital Forensics ottenendo i diplomi di perfezionamento in Data Protection e Data Governance; Criminalità Informatica e Investigazioni Digitali e Big Data,

Artificial Intelligence.

Ha, altresì, conseguito l’Advanced Cybersecurity Graduate Certificate alla School of Engineering della Stanford University; Professional Certificates in Information Security; Incident Response Process; Digital Forensics e Cybersecurity Engineering and Software Assurance presso il Software Engineering Institute della Carnegie Mellon University.

Dal 1992 è nei ruoli del Ministero dell’Interno ove ricopre lincarico di Funzionario alla Sicurezza CIS. In tale veste contribuisce alla valutazione dei rischi cyber, all’implementazione delle misure di sicurezza e la risoluzione di incidenti informatici. Inoltre, offre consulenza tecnica nel campo della Digital Forensics per l’Autorità giudiziaria, la Polizia giudiziaria e gli Studi legali.

Dal 2017 è Professore a contratto di Tecnologie per la Sicurezza Informatica presso alcune Università ove sviluppa le tematiche di Attack and Defense Strategies quali il penetration testing, la risk analysis, l’information security assessment, l’incident response e la digital forensics. Infine, è Autore di alcuni articoli e saggi sui temi della Sicurezza Informatica e dell’Informatica Giuridica consultabili su https://www.vincenzocalabro.it

Condividi sui Social Network:
https://www.ictsecuritymagazine.com/articoli/machine-unlearning/