Le désapprentissage machine

Une réponse à l'oubli numérique

Mounir Lahlouh
ESME research lab

Ella Jewison
Centre de recheche et innovation Talan

La conformité aux réglementations sur la protection des données personnelles, comme le RGPD (règlement général sur la protection des données) en Europe, exige la préservation de la confidentialité et l'encadrement de la manipulation des données. Le droit individuel à l'oubli, stipulé par le RGPD, pose un défi pour les solutions d'intelligence artificielle qui reposent sur l'entraînement de modèles à partir de données personnelles. Supprimer simplement ces données ne résout pas efficacement le problème, puisque le cheminement des données dans les réseaux d'apprentissage est diffus et complexe. Ainsi supprimer une donnée est sans effet puisque le modèle à déjà « appris » sur celle-ci, les poids du modèle ont été modifiés par l'ingestion de cette donnée, et la prédiction sera faite avec la prise en compte de ces données. La solution naturelle et triviale serait donc de réentraîner le modèle à partir de zéro (*from scratch*) sur un ensemble de données excluant les éléments à supprimer, afin de garantir l'exclusion des données. Cependant cette approche est très coûteuse en termes de ressources de calcul et temporelles, mais aussi en termes de performance. Dès lors, comment doit-on réagir d'un point de vue scientifique et pratique pour garantir le respect des lois de protection de la vie privée et gestion efficace des performances et paramètres d'un modèle de *deep learning* ?

DOI:10.48556/SIF.1024.25.139

Les domaines d'applications où le désapprentissage a une importance prépondérante sont le secteur médical et les données biométriques. En effet le retrait d'informations sensibles des modèles utilisés pour le diagnostic ou la prédiction, est essentiel afin de respecter la confidentialité des patients, mais aussi pour les caractéristiques biométriques de certains individus, de préserver la confidentialité et d'éviter les abus. La confidentialité des données est une préoccupation essentielle en apprentissage automatique, et le désapprentissage joue un rôle important dans la résolution des problèmes de confidentialité ? Notamment, les réglementations — telles que le RGPD — accordent aux individus le droit de demander la suppression de leurs données personnelles. Cela touche également des considérations éthiques, telles que la correction des biais. Le désapprentissage permet de supprimer ou de réapprendre des données spécifiques afin d'améliorer l'équité et d'éviter de perpétuer des prédictions biaisées.

Notion de confidentialité différentielle

Concepts et contexte

Historiquement, les questions de confidentialité étaient traitées en anonymisant les bases de données, c'est-à-dire en supprimant certaines données sensibles, par exemple les noms ou les adresses. Il s'est avéré que ces pratiques ne garantissaient pas suffisamment la protection des individus. En effet, les travaux de Latanya Sweeney ont montré qu'il était possible d'identifier des individus à partir de données anonymisées en croisant celles-ci avec des informations publiques [9]. Notamment, grâce aux données des hopitaux américains, cette autrice a pu identifier 87 % des citoyens américans simplement à partir de leur date de naissance, de leur sexe et de leur code postal. Cet article a marqué un tournant dans la prise de conscience des risques liés aux données mal anonymisées et a permis le développement de nouvelles pratiques comme l'élaboration de la notion de confidentialité différentielle.

La confidentialité différentielle (DP, pour *differential privacy*) est une approche visant à protéger la confidentialité des données dans un ensemble [5]. Plutôt que de se concentrer sur l'anonymisation des données, la DP permet de garantir qu'aucune information sensible sur un individu ne peut être déduite, même en cas d'accès à des sources de données externes ou avec des informations partielles sur la base. Elle pose un cadre mathématique rigoureux et permet notamment de quantifier le degré de protection. La DP passe par l'introduction de bruit aléatoire calibré à des résultats d'analyses globales pour garantir la protection des données individuelles.

On dit qu'un algorithme est DP lorsqu'il permet de produire des résultats comparables en confrontant deux ensembles de données, rendant difficile la déduction de caractéristiques distinctives concernant cet individu.

Soit deux ensembles de données, $\mathcal{D}$ et $\mathcal{D}'$ qui diffèrent par l'ajout ou la suppression de k exemples, avec $k \geq 1$. Par exemple, considérons $\mathcal{D} = \{a, b, c, d\}$ et $\mathcal{D}' = \{a, b, d\}$. Un algorithme $\mathcal{M}$ satisfait la (ϵ, δ) DP si pour toute paire de bases de données voisines $\mathcal{D}$ et $\mathcal{D}'$, et pour tout ensemble de résultats possibles $\mathcal{S}$ que l'algo peut produire, la condition suivante est respectée :

$$P(M(\mathcal{D}) \in S) \ \leq e^{\epsilon} \ \cdot \ P(M(\mathcal{D}') \in S) \ + \ \delta \tag{1}$$

Le terme e^{ϵ} désigne le paramètre qui contrôle le niveau de confidentialité (ϵ privacy budget). Plus ϵ est petit, plus e^{ϵ} est petit et la confidentialité élevée. Cela signifie que la sortie du mécanisme DP ne varie que très peu en réponse à une seule modification dans les données d'entrée.

Le paramètre δ mesure la tolérance ou la relaxation. Ce paramètre représente la probabilité qu'une violation de la confidentialité se produise. Dans le cas où $\delta = 0$, on parle de confidentialité pure. Dans la pratique, un petit δ est utilisé — comme 10^{-5} — ce qui signifie que l'algorithme peut échouer dans un cas sur 100 000.

Ainsi, les sorties de l'algorithme appliquée à $\mathcal{D}$ ou $\mathcal{D}'$ sont quasiment identiques, de sorte que cela garantit que la participation d'un individu spécifique ne puisse pas être déduite. L'interprétation intuitive de la DP est qu'elle garantit que la probabilité qu'un certain résultat $\mathcal{S}$ soit produit par $\mathcal{M}$ ne change pas de manière significative lorsque les données d'un individu sont ajoutées ou retirées de la base. En d'autres termes, un attaquant qui observe la sortie de l'algorithme $\mathcal{M}$ ne peut pas déterminer si un individu particulier fait partie de la base.

Exemple d'un cas d'une université et la nécessité de données bruitées

Pour illustrer cette notion nous allons prendre un exemple fictif d'une université américaine, où travaillent deux professeurs : Alice et Bob [10]. Ces professeurs, dans le cadre de leur travaux de recherche et dans un but de transparence publient des statistiques sur leurs étudiants, notamment sur leur situation financière.

Alice publie en mars que la classe de première année comptait 3005 étudiants, dont 202 provenaient de familles gagnant plus de 350 000 $ par an. Bob a publié en avril une statistique similaire, indiquant qu'il y avait 3004 étudiants et 201 dans la même catégorie de revenus.

Il est donc possible, sur la base de ces deux articles, de déduire qu'un étudiant a quitté l'université entre mars et avril, et que ses parents gagnent plus de 350 000 $ par an. Après enquête, il est possible d'identifier cet étudiant, John, et de diffuser cette information avec auprès de ses camarades. John, contrarié que ses pairs aient appris des détails sur la situation financière de sa famille, se plaint à l'université. Cet exemple illustre que même la

publication de statistiques agrégées et non personnelles, permet de déduire des informations privées sensibles en comparant des jeux de données qui diffèrent légèrement.

La confidentialité différentielle permet de limiter ce risque en bruitant les données. Dans notre exemple, pour respecter la confidentialité différentielle, il faudrait publier des données légèrement modifiées. Par exemple, en mars, Alice interroge le portail de données et obtient un nombre bruité de 204 étudiants de première année venant de familles avec un revenu supérieur à 350 000 $.

Elle publie qu'environ 200 étudiants sont concernés. En avril, Bob pose la même question et reçoit le nombre bruité de 199, publiant un chiffre similaire. Grâce à ces chiffres bruts, personne ne peut déduire qu'un étudiant avec un revenu familial supérieur à 350 000 $ a quitté l'université en mars, réduisant ainsi le risque de révéler des informations personnelles sur John.

Qu'est-ce que le désapprentissage machine ?

L'apprentissage machine repose sur l'idée que les modèles d'IA peuvent apprendre des données pour effectuer des prédictions ou des classifications. Mais une fois que ces données sont utilisées pour entraîner un modèle, il est souvent difficile de les retirer du système. Or, avec des réglementations comme le RGPD, les individus ont le droit de demander l'effacement de leurs données personnelles, y compris celles utilisées pour entraîner des modèles.

Le principe du désapprentissage machine est de faire en sorte qu'un modèle, après suppression des données indésirables, donne des résultats statistiquement proches de ceux qu'il produirait si ces données n'avaient jamais été utilisées.

Prenons l'exemple d'un modèle initial, construit à partir d'un ensemble de données. Lorsqu'il s'agit de retirer des informations à oublier, deux approches sont possibles :

– réentraîner entièrement le modèle avec les données nettoyées, ce qui est coûteux en ressources ;
– ajuster le modèle existant grâce à des algorithmes spécialisés. Ces derniers recalibrent le modèle pour obtenir des résultats équivalents à ceux qu'il produirait si les données supprimées n'avaient jamais été prises en compte.

Les résultats obtenus doivent être quasi identiques dans les deux cas, garantissant ainsi que les données effacées n'ont plus d'impact sur les prédictions du système.

Pour évaluer cette capacité, on utilise des concepts empruntés à la confidentialité différentielle. Le désapprentissage d'un ensemble $\mathcal{S} \subseteq \mathcal{D}$ d'un réseau de neurones, noté $A(A : \mathcal{D} \longrightarrow \mathcal{R})$, est atteint lorsque deux conditions sont remplies : ϵ et δ sont très petits. Cela signifie que les distributions des sorties

du modèle réentraîné, $A(\mathcal{D}\backslash\mathcal{S})$, et du modèle désappris, $U(A(\mathcal{D}), \mathcal{S}, \mathcal{D})$, sont presque indiscernables. Les relations suivantes expriment cette condition [8] :

$$e^{\epsilon} P[U(A(\mathcal{D}), \mathcal{S}, \mathcal{D}) \in \mathcal{R}] + \delta \geq P[A(\mathcal{D}\backslash\mathcal{S}) \in \mathcal{R}] \tag{2}$$

et

$$P[U(A(\mathcal{D}), \mathcal{S}, \mathcal{D}) \in \mathcal{R}] \leq e^{\epsilon} P[A(\mathcal{D}\backslash\mathcal{S}) \in \mathcal{R}] + \delta \tag{3}$$

Ici, $\mathcal{P}$ désigne la distribution de tous les modèles entraînés sur un ensemble défini de données. Ces formules garantissent que le modèle désappris produit des résultats proches de ceux d'un modèle réentraîné, tout en respectant des critères de confidentialité rigoureux.
Le désapprentissage machine peut se schématiser comme suit :

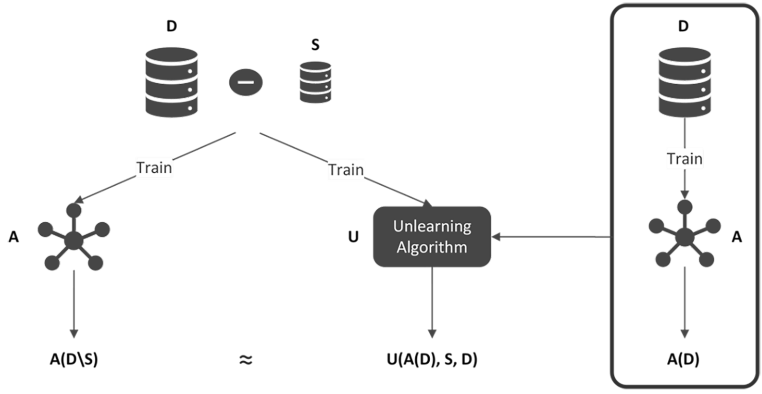

Fig. 1. Schéma explicatif du désapprentissage machine.

Différents types de désapprentissage machine ont été développés pour répondre à des besoins variés. Certaines approches privilégient une exactitude maximale, visant à reproduire fidèlement les résultats d'un modèle entièrement réentraîné, tandis que d'autres optent pour des solutions approximatives, moins exigeantes en ressources mais toujours efficaces. Ces méthodes offrent des réponses adaptées aux défis — techniques, éthiques et alignés sur la réglementation — liés à la gestion des données à oublier.

Exemples d'approche de désapprentissage

Concrètement, pour mettre en œuvre les différents types de désapprentissage, des approches spécifiques ont été développées. Ces approches se divisent en deux grandes catégories :

– les méthodes agnostiques, qui ne dépendent pas de la structure interne du modèle et sont applicables à une large variété de modèles. Elles permettent de supprimer l'influence des données ciblées sans nécessiter de modifications structurelles ;
– les méthodes intrinsèques, qui exploitent les spécificités d'une architecture donnée pour optimiser le processus de désapprentissage. Ces techniques sont conçues sur mesure pour des modèles particuliers, comme les réseaux neuronaux ou les algorithmes de clustering.

Quelle que soit la catégorie, plusieurs stratégies d'action permettent de répondre aux besoins variés, aux contraintes techniques et aux formats de données qui alimentent les modèles IA.

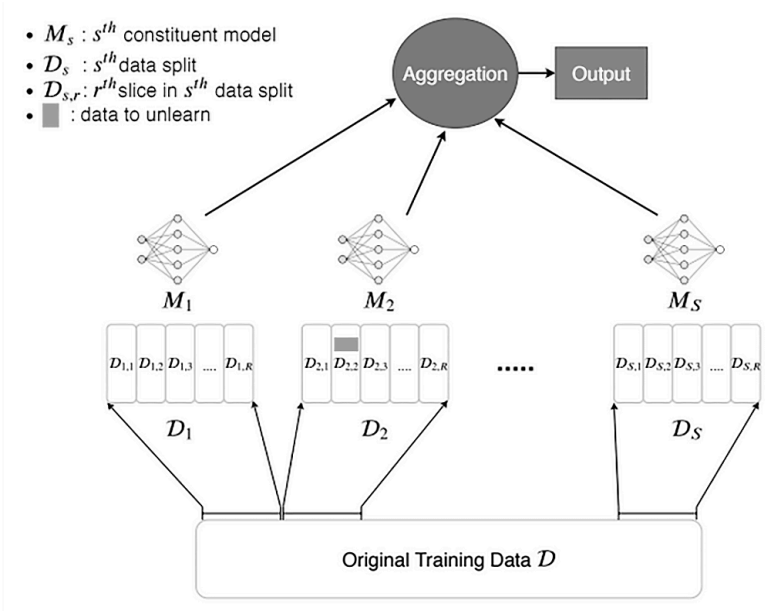

Fig. 2. Schéma explicatif de l'algorithme SISA [1].

Agir sur les données

L'algorithme le plus connu est SISA (*shared, isolated, sliced and aggregated*) [1]. Cette approche divise le jeu de données en fragments, chaque point de données étant associé à un seul fragment. Chaque fragment entraîne un sous-modèle, ce qui permet de ne mettre à jour que le sous-modèle lié à la donnée à supprimer, évitant ainsi de réentraîner l'ensemble du modèle. Un partitionnement supplémentaire permet de conserver les poids associés à chaque fragment, ce qui

facilite la mise à jour en cas de retrait d'une donnée. Les sous-modèles sont agrégés pour obtenir la classification finale.

Par ailleurs, pour les modèles de langage tels que les LLM (*large language models*), des techniques spécifiques émergent, comme l'exemple du modèle Llama2 dans l'étude intitulée « *Who's Harry Potter?* » [6]. Cette étude montre comment un modèle peut oublier des contenus précis, tels que les livres Harry Potter, en ajustant les prédictions de manière à effacer des associations spécifiques mais en conservant ses capacités sur d'autres tâches. Cela signifie qu'il faut demander à un LLM de générer plusieurs alternatives plausibles à ses prédictions — qui ne sont pas liées à Harry Potter — mais qui restent probables pour tout modèle ignorant les détails de ces ouvrages pour un concept donné que l'on souhaite supprimer.

Agir sur l'architecture du modèle et la fonction de perte

La modification directe de la structure du modèle ainsi que la conception d'une fonction de perte dédiée, permet de rendre le désapprentissage plus efficace.

Une des méthodes intrinsèques, *DaRE*[1] *forests* [2] conçue pour les forêts aléatoires permet de supprimer efficacement des données sans réentraîner l'ensemble du modèle. Elle utilise des nœuds aléatoires, stocke les statistiques à chaque nœud et ne réentraîne que les sous-arbres affectés par la suppression.

En juin 2023, la compétition sur le désapprentissage automatique, organisée par Google[2], a mis au défi les participants de développer des algorithmes efficaces pour permettre à des modèles d'oublier des informations spécifiques tout en conservant des performances élevées sur des données retenues. Une approche notable conçue par l'équipe gagnante de la compétition de désapprentissage de Google se distingue par une phase d'oubli proposant une nouvelle fonction de perte se basant sur l'optimisation de la divergence de Kullback-Leibler.

Agir sur les poids et le gradient

Les stratégies d'action sur les poids visent à ajuster directement les paramètres internes du modèle pour supprimer l'influence des données à oublier. Une des méthodes les plus connues est SCRUB [7]. Elle identifie les données critiques ayant le plus d'impact sur le modèle et inverse les mises à jour du poids lié à ces données. Nous pouvons également mentionner certaines des méthodes classées parmi les huit premières dans la compétition organisée par Google DeepMind :

1. DaRE : *data removal-enabled.*
2. `https://unlearning-challenge.github.io/`.

– transposition des poids et pseudo-label : cette méthode consiste à transposer les poids des couches de convolution du réseau convolutif (CNN) pour faciliter la suppression des données indésirables. Ensuite, un *fine-tuning* est effectué à l'aide de pseudo-labels générés à partir des erreurs du modèle ;
– réinitialisation basée sur le gradient : cette technique identifie les paramètres du modèle qui sont peu affectés par la différence entre les données à retenir et celles à oublier. Ces poids sont ensuite réinitialisés de manière ciblée.

Comment évaluer le désapprentissage ?

L'évaluation des approches de désapprentissage nécessite de prendre en compte trois facteurs clés : la qualité de l'oubli, l'efficience et l'efficacité de l'algorithme. Un bon algorithme de désapprentissage doit trouver un équilibre entre ces facteurs : oublier efficacement des données, maintenir de bonnes performances et s'exécuter plus rapidement qu'un réentraînement complet.

La qualité se mesure en comparant les distributions des poids obtenues avec un réentraînement complet (*from scratch*) sans les données à oublier et celles produites par l'algorithme de désapprentissage. En pratique, des attaques en boîte noire peuvent être utilisées pour estimer la séparation entre ces distributions en analysant les sorties du modèle pour les données oubliées. Les résultats permettent de calculer un score d'oubli global.

L'efficience représentée par le temps d'exécution est un critère central. Par exemple, dans le cadre de la compétition Google, les algorithmes devaient s'exécuter en moins de 20 % du temps nécessaire pour un réentraînement complet. En effet, une méthode de désapprentissage efficace doit limiter l'utilisation des ressources dans le temps.

D'autres métriques peuvent être utilisées en complément :
– AIN (*anamnesis index*, [3]) : évalue combien d'itérations d'apprentissage sont nécessaires pour réapprendre les caractéristiques oubliées, permettant de quantifier l'oubli ;
– ZRF (zero retrain forgetting score, [4]) : sur des données à oublier et un modèle initialisé sans ces données, utile pour évaluer l'oubli de classes entières.

Alignements réglementaires

Le règlement européen sur l'intelligence artificielle (*AI Act*) a été adopté en mars 2024 et la mise en application se fera progressivement jusqu'en août 2027. Il vise à encadrer le développement et l'utilisation de l'IA au sein de l'Union européenne, en garantissant la sécurité des biens et des personnes ainsi que la protection des droits fondamentaux, tels que la vie privée et les données personnelles. Bien que le désapprentissage ne soit pas mentionné

explicitement dans ce texte, les pratiques qu'il nécessite en termes de protection des personnes implique directement les notions de désapprentissage et de confidentialité différentielle. Ainsi, les développeurs devront intégrer des mécanismes de désapprentissage automatique pour se conformer aux nouvelles obligations légales, assurant que les modèles d'IA puissent oublier ou supprimer les données personnelles lorsque c'est nécessaire.

Conclusion

Bien que les solutions actuelles soient encore en développement, ce domaine (du désapprentissage) ouvre la voie à une gestion plus éthique et flexible des données dans le futur numérique, un futur qui n'est cependant pas exempt de défis majeurs restant à relever :

– complexité technique : assurer que toutes les traces des données sont éliminées sans affecter les performances du modèle est un défi technique majeur ;
– garanties : comment s'assurer que les données ont vraiment été oubliées ? Il est crucial de développer des métriques et des audits pour valider le processus ;
– sophistication des attaques : les attaques adversaires, comme l'empoisonnement des données ou l'inversion de modèle, deviennent de plus en plus complexes, rendant le désapprentissage encore plus difficile ;
– manque de standardisation : l'absence de méthodes, de métriques d'évaluation ou de normes standardisées complique la comparaison et l'évaluation des différentes techniques ;
– problèmes de transférabilité : certaines techniques de désapprentissage sont spécifiques à certains modèles ou ensembles de données et ne peuvent pas être facilement appliquées à d'autres, limitant leur applicabilité générale ;
– difficulté d'interprétation : de nombreuses techniques, notamment celles impliquant des réseaux de neurones profonds, sont difficiles à interpréter, ce qui complique la compréhension de leur fonctionnement interne et des éventuelles faiblesses ;
– contraintes de ressources : les processus de désapprentissage nécessitent d'importantes ressources computationnelles, ce qui peut être coûteux et chronophage, surtout pour les grands modèles et ensembles de données.

Nous remercions Alexandra Benamar, Ann-Déborah Schulie-Le Nenan, Mandar Chitale, Abdon Ivala, Ludovic Blouin, Dongny Joseph, Mouna Berrabah et Alban Jadot.

Références

[1] Bourtoule, Lucas, et al. “Machine unlearning.” *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2021.

[2] Brophy, Jonathan, and Daniel Lowd. “Machine unlearning for random forests.” *International Conference on Machine Learning*. PMLR, 2021.

[3] Chundawat, Vikram S., et al. “Zero-shot machine unlearning.” *IEEE Transactions on Information Forensics and Security* 18 (2023): 2345-2354.

[4] Chundawat, Vikram S., et al. “Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher.” *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. No. 6. 2023.

[5] Dwork, Cynthia, et al. “Calibrating noise to sensitivity in private data analysis.” *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*. Springer Berlin Heidelberg, 2006.

[6] Eldan, Ronen, and Mark Russinovich. “Who’s Harry Potter? Approximate Unlearning in LLMs.” *arXiv preprint arXiv:2310.02238* (2023).

[7] Kurmanji, Meghdad, et al. “Towards unbounded machine unlearning.” *Advances in neural information processing systems* 36 (2024).

[8] Nguyen, Thanh Tam, et al. “A survey of machine unlearning.” *arXiv preprint arXiv:2209.02299* (2022).

[9] Sweeney, Latanya**.** *“Simple Demographics Often Identify People Uniquely.”* Carnegie Mellon University, Data Privacy Working Paper 3, 2000.

[10] Wood, Alexandra et al. *Differential Privacy: A Primer for a Non-Technical Audience* (2018). Vanderbilt Journal of Entertainment & Technology Law, Vol. 21, No. 17, 2018.