Preservación de Privacidad en el Aprendizaje Profundo: Enfoques y soluciones

Rafael Gallardo-García*, Luis A. Herrera-Maldonado*, Alberto Remigio-Alvarado*, Pablo Romero-Hernández*

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación, Puebla, México

{rafael.gallardo,luis.herrerama,alberto.remigio,pablo.romeroh}
@alumno.buap.mx

Resumen El problema de realizar análisis de datos mientras se preserva la privacidad de los individuos se ha atacado desde múltiples disciplinas. Los últimos avances e hitos alcanzados por las técnicas de aprendizaje profundo han despertado un gran interés de la comunidad científica. Hoy en día se pueden encontrar aplicaciones del aprendizaje profundo en áreas como la medicina, redes sociales, internet de las cosas, robótica, conducción autónoma, procesamiento del lenguaje natural, procesamiento de voz e intercomunicaciones inalámbricas, en donde los conjuntos de datos suelen ser de colaboración colectiva y pueden contener información confidencial y deben mantenerse privados, tanto en el entrenamiento como en la inferencia. En este trabajo, se presenta un análisis y clasificación para los ataques de extracción de datos confidenciales en modelos de aprendizaje profundo, de igual manera, se analizan los enfoques para la preservación de la privacidad en modelos de aprendizaje profundo, haciendo especial énfasis en la utilización del modelo de privacidad diferencial. Los enfoques analizados se clasifican principalmente por sus aplicaciones al aprendizaje profundo tradicional y al colaborativo.

Palabras clave: Aprendizaje automático, Filtración de información, Modelos seguros, Privacidad Diferencial.

Privacy Preservation in Deep Learning: Approaches and solutions

Abstract. The problem of conducting data analysis while preserving the privacy of individuals has been addressed by multiple disciplines. The latest advances and milestones achieved by deep learning techniques have aroused great interest from the scientific community. Today deep learning applications can be found in areas such as medicine, social networks, the

* Los autores contribuyeron en partes iguales, se ordenaron alfabéticamente según el primer apellido.

internet of things, robotics, autonomous driving, natural language processing, voice processing, and wireless intercommunications, where data sets are usually of collective collaboration and may contain confidential information and should be kept private, both in training and inference. In this work, we present an analysis and a classification for those attacks that aim to extract confidential data in deep learning models, in the same way, the approaches for the preservation of privacy in deep learning models are analyzed, with special emphasis on the use of the differential privacy model. The analyzed approaches are classified mainly by their applications to traditional and collaborative deep learning.

Keywords: Differential Privacy, Information, Information Filtering, Machine Learning, Safe Models.

1. Introducción

La llegada del internet, de los teléfonos inteligentes, las redes sociales y el internet de las cosas ha potenciado la capacidad de la humanidad para acumular información personal que pueda resultar útil en análisis y estadísticas [10]. Además de los avances en la capacidad de almacenamiento, la tecnología también ha aumentado su capacidad de procesar dicha información con rapidez y exactitud, haciendo uso de técnicas cada vez más sofisticadas, como lo son los algoritmos de aprendizaje automático y el Aprendizaje Profundo (AP) [29]. El nacimiento de la Privacidad Diferencial (PD) satisface la necesidad de tener una definición de privacidad que sea más robusta, significativa y matemáticamente rigurosa, junto con una rica clase de algoritmos que satisfagan esa definición [10].

Dwork y Roth [10] plantean la PD como una promesa, que realiza un poseedor de datos o "curador" a un sujeto de datos: "No se verá afectado, adversamente ni de ninguna otra forma, al permitir que sus datos se utilicen en cualquier estudio o análisis, sin importar qué otros estudios, conjuntos de datos o fuentes de información estén disponibles". Por otra parte, Zhu et al. definen al término como "un modelo de privacidad sólido que proporciona una garantía de privacidad demostrable para las personas" [30]. Los autores explican que, incluso si el adversario llega a conocer la máxima información de antecedentes, excepto el registro que desea conocer, no puede identificar el registro específico.

Hoy en día se pueden encontrar aplicaciones del aprendizaje profundo en áreas como la medicina, redes sociales, internet de las cosas, robótica, conducción autónoma, procesamiento del lenguaje natural, procesamiento de voz e intercomunicaciones inalámbricas [6], en donde los conjuntos de datos suelen ser de colaboración colectiva y pueden contener información confidencial [1], por lo tanto, es importante implementar algoritmos de privacidad [29]. La propuesta de privacidad diferencial de Dwork y Roth en el año 2006 mostró garantías de privacidad demostrables para los registros de las bases de datos sin ninguna perdida significativa de precisión en las consultas. En la actualidad, se han realizado varios intentos para aplicar privacidad diferencial en el aprendizaje

profundo, ésto, con el fin de garantizar la privacidad de los datos en los conjuntos de entrenamiento [10]. Algunas de las grandes empresas de tecnología han comenzado a introducir este modelo de privacidad en sus sistemas operativos y en sus dispositivos, de modo que se pueda asegurar la privacidad de los usuarios mientras se extrae información importante mediante técnicas de aprendizaje automático [10].

Comúnmente, entrenar modelos con técnicas de aprendizaje profundo requiere una cantidad considerable y representativa de datos [1], de donde se podría obtener información sensible, suponiendo así, un problema de privacidad. La implementación de métodos de privacidad diferencial en técnicas de aprendizaje profundo es un área poco explorada debido a la incertidumbre que presenta el intercambio entre privacidad de datos y el aprendizaje del modelo [10]. Como lo mencionan Zhu et al. existen formas sencillas para agregar privacidad a los datos (e.g. agregar ruido a los resultados), aunque esto implica tener un algoritmo de aprendizaje con menos utilidad [30]. En la literatura se pueden encontrar varias alternativas de implementación del modelo de privacidad diferencial en las técnicas de aprendizaje profundo, estas alternativas buscan un equilibrio entre el nivel de privacidad en los datos y el aprendizaje del modelo. Este artículo revisará dichas propuestas y alternativas, describiendo de forma general las metodologías utilizadas en cada enfoque.

2. Materiales y métodos

Este trabajo presenta un análisis de los enfoques utilizados para la aplicación de la privacidad diferencial en técnicas de aprendizaje profundo. La búsqueda de literatura se realizó a través de Google Scholar y se tomaron en cuenta artículos publicados en revistas indexadas, así como aquellos alojados en servidores de pre-impresión. La lista a continuación describe las consultas realizadas al sistema de búsqueda:

- Differential privacy deep learning (369,000 resultados)
- Deep learning and differential privacy (363,000 resultados)
- Deep learning differential privacy survey (85,400 resultados)
- Deep learning differential privacy review(213,000 resultados)

De los resultados, se seleccionaron aquellos que incluyeran estrictamente todas las palabras en la búsqueda. Solo se consideraron los resultados de las primeras dos páginas de cada consulta. Se seleccionaron los artículos que en su resumen incluyeran una descripción de su propuesta y el área de la arquitectura en dónde se realiza la implementación de la PD. En el caso de las consultas sobre *surveys* y *reviews*, se seleccionaron aquellos que en su resumen incluyeran la propuesta de algún tipo de clasificación para los enfoques de privacidad diferencial en el aprendizaje profundo. Todos los artículos considerados están escritos en idioma inglés.

Al final de la recolección de literatura, se obtuvieron 14 artículos con propuestas de privacidad diferencial aplicada en técnicas de aprendizaje profundo, 2 revisiones y 2 encuestas.

3. El problema de la privacidad en el aprendizaje profundo

Comúnmente, el entrenamiento de los algoritmos de aprendizaje profundo requiere la recolección de grandes cantidades de datos, mismos que, pueden tornarse sensibles, sobre todo en áreas como la medicina, los datos biométricos o los de comportamiento. Muchos usuarios resultan reluctantes a la hora de brindar sus datos a terceros para el entrenamiento, sobre todo cuando éstos se almacenarán en un servidor que pueda estar comprometido, de forma que otros podrían observar sus datos privados almacenados en la nube. Para convencer a los usuarios de que su información está segura, se necesita un enfoque para que los datos estén privadamente preservados; una posible solución a esto es que los datos estén encriptados cuando se envían al servidor y también durante el proceso de entrenamiento, sin embargo, esto implica desarrollar algoritmos de aprendizaje capaces de procesar datos encriptados; una alternativa, es la implementación de la privacidad diferencial en los algoritmos de AP.

3.1. AP tradicional

En el aprendizaje profundo, la información sensible puede ser descubierta a través de ataques de inferencia, los cuales se dividen en dos categorías fundamentales, los ataques de rastreo y los ataques de reconstrucción.

Ataques de reconstrucción Los ataques de reconstrucción tienen como objetivo extraer datos de entrenamiento a través de las predicciones del modelo. Los ataques de inversión utilizan la salida del modelo para inferir ciertas características del conjunto de entrenamiento. El principio detrás de la inversión del modelo consiste en utilizar características sintetizadas del modelo para generar una entrada que maximice la probabilidad de ser pronosticado con una determinada etiqueta [11]. Además, el objetivo del adversario es capacitar a un modelo sustituto F', que es capaz de imitar un modelo objetivo F. Para construir el modelo F', se utiliza la fuga de información que se implementa en el momento de extracción [25]. En la extracción del modelo, el adversario sólo tiene que acceder a la *Application Programming Interface* (API) de predicción de un modelo de destino y consultar el modelo con muestras "naturales" o sintéticas. Estas muestras están específicamente diseñadas para maximizar la extracción de información sobre los componentes internos del modelo, a partir de las predicciones devueltas por el modelo F [13].

Ataques de rastreo En los ataques de rastreo, un adversario identifica formatos de entrada y salida, se le da acceso de caja negra a un modelo objetivo (sin conocer sus parámetros internos) y, quiere inferir si un registro en particular está incluido en el conjunto de entrenamiento [21]. Los autores transforman el ataque de inferencia en una tarea de clasificación [20]. Trung Ha et al, describen tres pasos para implementar en este ataque, en el primer paso, el adversario consulta

el registro de destino t y utiliza la predicción de los clasificadores de destino sobre t para inferir el estado de pertenencia de t. Para el paso número 2, se construye la técnica de entrenamiento "sombra" utilizada para la tarea de clasificación. Múltiples "modelos sombra" son entrenados por el adversario utilizando el mismo algoritmo de aprendizaje automático en registros muestreados de los datos. En el paso final, el adversario entrena un modelo como clasificador de "ataque" y lo usa para inferir la pertenencia a un registro objetivo t [13]. Para garantizar la privacidad diferencial, existen dos mecanismos básicos ampliamente utilizados en el aprendizaje profundo: el mecanismo de Laplace [8] y el mecanismo exponencial [17], los detalles se describen a continuación:

1. Mecanismo de Laplace: Para una función $f : D \to R$ sobre un conjunto de datos D, el mecanismo M asegura la privacidad diferencial ϵ, sí:

$$M(D) = f(D) + Lap(\frac{\Delta f}{\epsilon}). \tag{1}$$

2. Mecanismo exponencial: En las consultas no numéricas, el mecanismo exponencial aleatoriza los resultados asociados con una función de puntuación $q(D, \varphi)$. La función q se utiliza para evaluar la calidad de la salida φ. Δq representa la sensibilidad de q. El mecanismo exponencial M satisface la privacidad diferencial ϵ, sí:

$$M(D) = \left(return\varphi \propto (\frac{\epsilon q\,(D, \varphi)}{2\Delta q}) \right). \tag{2}$$

3.2. AP colaborativo

El aprendizaje profundo colaborativo reúne múltiples usuarios que alimentan al modelo, se evita tener que reunir los datos de diferentes fuentes y se mejora la precisión de los modelos. Aunque las aplicaciones son muy variadas, el acceso y uso puede realizarse por terceros o usuarios no confiables.

De manera general el aprendizaje profundo colaborativo tiene las mismas fases que el tradicional, aunque en estas fases existen diferentes enfoques que se pueden utilizar:

Fase de formación

1. Aprendizaje profundo colaborativo directo: Un servidor central se encarga de recolectar la información local de los múltiples usuarios y ejecutar el algoritmo de aprendizaje profundo de manera centralizada para obtener el modelo.
2. Aprendizaje profundo colaborativo indirecto: Se tiene un modelo global, y cada usuario mantiene un modelo y conjunto de datos de forma local. El conjunto de entrenamiento en posesión del usuario se usa para obtener un modelo local mejorado, después, la información es cargada en el modelo central. Se requieren múltiples iteraciones para converger al modelo óptimo.

Fase de uso El usuario proporciona una entrada para obtener una salida de acuerdo al modelo entrenado, en general, al usuario se le proporciona una API.

El problema de privacidad al utilizar AP colaborativo se debe a que, durante todo el proceso existen vulnerabilidades, que pueden ser aprovechadas por algún adversario al intentar inferir información de algún otro usuario.

En algunas ocasiones, se le puede considerar al servidor como un adversario "malintencionado" cuando no se tiene la supervisión necesaria sobre los datos del usuario. Es aún más peligroso cuando el servidor, así como múltiples usuarios, se convierten en adversarios, pudiendo inferir los datos de otros usuarios.

Por otra parte, Hitaj et al. [26] propusieron un ataque contra el aprendizaje profundo colaborativo, en el que el adversario entrena una *Generative Adversarial Network* (GAN) para generar muestras prototípicas con la misma distribución del conjunto de entrenamiento. Doshi-Vlez y Kim [7] discutieron los ataques a la privacidad en tres aspectos: inferencia de pertenencia, extracción de datos de entrenamiento y extracción de modelos.

Ataque de inferencia de pertenencia El adversario infiere si el registro existe en el conjunto de datos de entrenamiento o no, dando un modelo de aprendizaje automático y un cierto registro de muestra, el atacante consulta el modelo objetivo con un registro de datos para obtener una predicción sobre ese registro.

Extracción de datos de entrenamiento En la configuración de caja blanca, existe la amenaza de privacidad con la extracción de datos de entrenamiento. Fredrikson et al. [12] construyeron ataques de inversión de modelos para modelos profundos, utilizando la salida del modelo para inferir ciertas características del conjunto de entrenamiento. El efecto de este ataque es muy limitado, no proporciona una descripción clara y específica del conjunto de entrenamiento, ni puede juzgar si una muestra está en el conjunto de entrenamiento o no.

Extracción de modelo Diseñado para extraer los parámetros del modelo entrenado con los datos privados. El objetivo principal en este tipo de ataques es la duplicidad de la funcionalidad del modelo, cuyo rendimiento de predicción en el conjunto de datos de verificación es similar al modelo objetivo. La privacidad del conjunto de entrenamiento se debilita aún más después de la filtración de los parámetros del modelo, debido a la estrecha conexión entre los parámetros del modelo y el conjunto de entrenamiento.

4. Enfoques de preservación de la privacidad en el AP

Una técnica de preservación de privacidad es una herramienta especial que permite el procesamiento de datos, de forma segura y confidencial [28]. La importancia de estas técnicas se encuentra en habilitar el procesamiento de los datos, sin revelar el contenido original. De esta manera, se puede asegurar la

privacidad de datos altamente confidenciales. La directiva 95/46/EC de la Unión Europea se encarga de regular el procesamiento de datos personales basándose en los derechos humanos [18]. Dicha directiva establece que "El controlador de datos debe implementar medidas técnicas y organizacionales para proteger los datos personales ante la destrucción o pérdida accidental (o ilegal), alteración, divulgación o acceso no autorizado, en particular donde el procesamiento involucra la transmisión de los datos a través de una red, también la debe proteger ante cualquier otra forma ilegal de procesamiento" [24]. Tanuwidjaja et al. clasifican las técnicas de Aprendizaje Profundo con Preservación de Privacidad en tres grupos: basadas en encriptación homomórfica (EH), basadas en Computación Multipartita (CMP) y basadas en privacidad diferencial.

En este capítulo se presentan algunas de las tecnologías clásicas de preservación de privacidad.

4.1. Encriptación Homomórfica Completa (EHC)

Permite trabajar con datos cifrados sin la necesidad de descifrarlos, minimizando así la posibilidad de que la información se vea expuesta. EHC soporta cualquier cálculo (suma, multiplicación, etc) sobre los datos encriptados sin conocer la llave secreta. En el proceso de entrenamiento el modelo desconoce los datos del usuario, y el usuario ignora como es el modelo.

4.2. Computación Segura Multipartita

Introducida como computación segura bipartita [27], extendida a CMP por Micali et al. en 1978 [19]. El propósito del CMP es resolver el problema de la computación colaborativa mientras se mantiene la privacidad de un usuario en un grupo de usuarios no-confiables, sin utilizar algún tercero confiable. El uso de CMP para técnicas de aprendizaje profundo no es muy adecuado debido al coste que implica el cálculo de las funciones de activación no lineales (e. g. sigmoide o softmax) durante la fase entrenamiento [28].

4.3. Privacidad diferencial

Propuesta por Dwork et al. en 2006 [9], como una forma de tratar el problema de la preservación de privacidad durante el análisis de datos. La privacidad diferencial busca que un administrador de datos confiable, pueda divulgar algunas estadísticas sobre los datos, sin revelar ninguna información sobre los datos en sí. Por lo tanto, un adversario con acceso a la salida de los algoritmos aprende casi la misma información tanto si se incluyen los datos de usuario como si no se incluyeran [24].

5. Enfoques de privacidad diferencial en el AP

La privacidad diferencial se define como un mecanismo M que, dado un conjunto de datos D, es transformado a un conjunto R y un rango R que satisface

(ϵ, δ), es diferencialmente privado sí, para dos entradas adyacentes d, d' que pertenecen a D y para cualquier subconjunto de salida $S \subseteq R$ sostiene que:

$$Pr\left[M(d) \in S\right] \leq e^{\epsilon} * Pr\left[M(d') \in S\right] + \delta \tag{3}$$

Donde ϵ es la estimación de privacidad que controla el nivel de privacidad y δ permite una pequeña probabilidad de fallo. Cuanto más pequeños se determinen ϵ, y δ más similar será $M(d)$ y $M(d')$.

Los métodos de privacidad diferencial asumen que los conjuntos de datos de entrenamiento y los parámetros del modelo son las bases de datos y demuestran que los algoritmos satisfacen la ecuación de la privacidad diferencial [13]. Dependiendo donde se agregue el ruido se puede dividir en tres grupos: nivel de gradiente (usado en el AP colaborativo), nivel de función (usado en el AP tradicional), nivel de etiqueta (usado en el AP tradicional).

5.1. Privacidad diferencial en el AP tradicional

Existe investigación que trabaja con la aplicación de la privacidad diferencial en algoritmos de aprendizaje automático: árboles de decisión [14], máquinas de vectores de soporte y regresiones logísticas [5,15]. Casi en paralelo, se han desarrollado técnicas para la preservación de la privacidad (PP) en técnicas de aprendizaje profundo. Dichas técnicas consisten, generalmente, en la modificación de alguna de las partes que componen a una red neuronal profunda:

- Capas de normalización por lotes.
- Aproximación de la función de activación.
- Capas convolucionales con tamaño de paso aumentado.

Otra forma de clasificar los enfoques de privacidad diferencial en el AP es: a nivel de función o a nivel de etiqueta.

Nivel de función Una forma de integrar la privacidad diferencial en el nivel de función es usando un autocodificador privado profundo que fue propuesto por Phan et al. [22]. El autocodificador codifica los valores de entrada haciendo uso de una función, para después ser decodificados por una función diferente a la que codifica, esto hace que los valores de salida sean idénticos a los de entrada. El codificador aprende las características más importantes de los datos entrantes y de esta manera la información sensible ya no es visible en los datos de salida.

Nivel de etiqueta El enfoque de nivel de etiqueta inyecta ruido en la fase de transferencia de conocimientos del marco maestro-alumno.

Como se describe en [21] Private Aggregation of Teacher Ensembles (PATE) es un tipo de modelo maestro-alumno, y su propósito es capacitar a un clasificador (estudiante) diferencialmente privado basado en un conjunto de clasificadores no privados (profesor). Además, el momento acumulado se utiliza para rastrear la estimación de privacidad acumulado en el proceso de aprendizaje por PATE y, además, PATE también garantiza la seguridad de forma intuitiva.

5.2. Privacidad diferencial en el AP colaborativo

Las técnicas de privacidad diferencial aplicadas al aprendizaje profundo colaborativo pueden clasificarse en dos grupos: AP colaborativo directo y AP colaborativo indirecto.

AP colaborativo directo Como en el AP colaborativo directo, cada usuario necesita cargar los datos que tiene alojados de manera local al servidor, es necesario que estos datos no puedan ser reconocidos durante este proceso.

Para solventar la falta de privacidad, a menudo son usadas técnicas de encriptación homomórfica. Al momento de entrenar una red neuronal que preserva la privacidad de los usuarios, es muy importante que, a pesar de las técnicas que se usen, la precisión de la red no se vea muy afectada, utilizando el cifrado homomórfico se protege la información y la red es capaz de realizar ciertas operaciones con los datos cifrados, aunque en algunos casos la cantidad de operaciones que se pueden utilizar son limitadas, entonces, se tendrá que recurrir a cifrar y descifrar la información en cada iteración. Si el objetivo es entrenar un modelo de aprendizaje profundo, la cantidad de operaciones aumentaría todavía más, lo que lo hace inviable para arquitecturas muy profundas. Además, se debe tomar en cuenta que, con cada iteración, los pesos también deben ser actualizados, lo que se traduce en cifrarlos y descifrarlos; entonces, la comunicación del modelo seguiría representando un costo muy alto. En [2] se utiliza la encriptación homomórfica para la protección del gradiente que se transmite entre los usuarios y servidores. Los usuarios comunican su información cifrada al servidor desde diferentes medios seguros, la información es utilizada para ajustar el modelo. El modelo y la información de los ajustes se encuentra encriptada en un servidor seguro, como la información de los ajustes está encriptada, no se infiere la información que se usó para ajustar el modelo y la precisión del aprendizaje se mantiene.

En los modelos de CMP, Chabanne et al. [4] hacen uso de algoritmos para el olvido de datos y procesadores basados en Software Guard Extensions (SGX), con los cuales se crean regiones de memoria protegidas en su totalidad por el procesador. Las regiones de memoria denominadas *enclave* contienen código y datos sensibles, para identificar a los usuarios, se hace uso de canales seguros que permitan autenticar la identidad de los usuarios. Cada usuario conectado al servidor cuenta con una clave local, con la cual cifra su conjunto de datos, mismo que deberá ser enviado por un canal seguro al servidor, una vez que el servidor obtiene la información de todos los usuarios, el código que se encuentra en la región de memoria protegida entrena al modelo, que al finalizar el entrenamiento estará encriptado y protegido con una nueva clave que es simétrica.

Otros enfoques no solo se preocupan por los usuarios que llegan a ser adversarios, sino que también toman en cuenta la posibilidad de que el servidor se convierta en un adversario. Para mitigar este problema, [16] utiliza un mecanismo de perturbación con 2 estados, llamado RG+RP, por las siglas: Repeat Gompertz (RG) + Random Projection (RP). RG modifica los datos de cada usuario a través de la función no lineal Repeat Gompertz. RP hace uso de una matriz de

Proyección Aleatoria ortogonal por filas que mantiene la precisión al momento de entrenar el modelo y reduce el costo al transmitir el ajuste. Este enfoque asegura a los usuarios que los datos que ha otorgado no se puedan identificar, es decir, la información original no se puede recuperar. Mientras que todos los atributos para el análisis de datos se mantienen intactos.

AP colaborativo indirecto Cuando se habla de AP colaborativo indirecto los enfoques cambian un poco. Para comenzar, este AP colaborativo mantiene la información de los usuarios como datos locales, así, se logra evitar el gasto en centros de datos y se mantiene una arquitectura centralizada.

Al mantenerse los datos y un modelo de manera local, el problema de preservar la privacidad de los usuarios pudiera parecer más fácil. Simplemente se debería proteger la información local y la que es enviada al servidor central, si recordamos, los datos del usuario no son enviados al servidor central, en cambio, lo que se envía son los ajustes del modelo local que alimenta al modelo central. El modelo central solo se encarga ajustar sus parámetros con los otorgados por el usuario y devuelve una actualización al modelo local. Aunque pueda parecer más sencillo, en la práctica no siempre es así.

El uso de algoritmos de descenso de gradiente estocástico modificados para preservar la privacidad es una alternativa muy utilizada en la literatura. Hacer uso de estos algoritmos permite entrenar el modelo de manera correcta sin poner en riesgo los datos utilizados durante el entrenamiento. Un modelo con estas características es presentado por [23], con el cual se permite el aprendizaje a múltiples usuarios de manera conjunta, este algoritmo del descenso de gradiente estocástico es selectivo y distribuido. Cada usuario hace el entrenamiento de su modelo de forma local tal como lo define el AP colaborativo indirecto, los modelos obtenidos son analizados, las partes claves del modelo son seleccionadas y se comparten pequeños subconjuntos de los parámetros.

La mayoría de los modelos en la literatura revisada busca minimizar los riesgos al compartir los datos con el modelo central, llegar a un acuerdo sobre cuántos datos, cómo serán enviados y de qué forma, es complejo, porque depende totalmente del programador. Estudiar el aprendizaje que se genera en el modelo de cada usuario es útil para determinar que datos son importantes de transmitir al modelo central, además, cada usuario podría tener su propio modelo personalizado. Modelos como el presentado en [3] proponen bloques descentralizados y asíncronos para el aprendizaje colaborativo, no se tienen nodos maestros para realizar acciones de agregación o coordinación. La distribución de los datos locales será diferente para cada usuario, por lo tanto, el modelo local es diferente entre usuarios, en lugar de tener un modelo global único. Para preservar la privacidad en los datos se utilizan perturbaciones aleatorias que garanticen que, aunque el adversario conozca el resultado del usuario, desconozca que datos sensibles se usaron para obtener ese resultado. Será entonces aún más difícil para un adversario recopilar información de manera sistemática, ya que no existe un medio de comunicación hacia el nodo principal. Este tipo de arquitecturas se pueden escalar a una gran cantidad de usuarios, sin embargo, no es muy

recomendable para ciertos entornos, ya que, como cada usuario se encuentra descentralizado, si se diera una comunicación entre un par, ésta podría darse a través de un servidor no confiable, exponiendo así, sus datos privados.

6. Conclusión

En la actualidad, las técnicas de aprendizaje profundo son ampliamente utilizadas, el desempeño y la calidad de resultados conseguidos por este tipo de algoritmos de aprendizaje ha marcado avances importantes en áreas cada vez más específicas y alejadas de la ciencia computacional pura. Cuando se habla de AP, es común pensar en la cantidad y calidad de los datos de entrenamiento como uno de los problemas más complejos a resolver; además de ser extensos y representativos, deben ser privados y seguros de procesar. En muchas áreas de aplicación del AP se trabaja con datos sensibles o confidenciales, de modo que, el curador de datos debe procurar la privacidad de los mismos al momento de entrenar y utilizar los modelos. En este trabajo se revisaron diferentes tipos de ataques, mediante los cuales, un adversario podría obtener información confidencial sobre los registros individuales contenidos en la base de datos de entrenamiento. De la misma manera, se revisaron y analizaron técnicas tradicionales para preservar la preservación de los datos en modelos de AP, haciendo especial énfasis en las técnicas que implementan el modelo de privacidad diferencial. En el análisis desarrollado, se agrupan las técnicas de AP en dos clases: tradicional y colaborativo, distinguiendo las técnicas de preservación de privacidad aplicables a cada una de estas clases. En la literatura revisada, se encontraron dos enfoques principales para implementar privacidad diferencial en el AP tradicional: a nivel de función y a nivel de etiquetas. Por otra parte, en el caso del AP colaborativo se considera la existencia de participantes no confiables y se exploran las fases de formación y uso. En esta revisión, también se incluye una clasificación para los tipos de ataques: de reconstrucción y de rastreo (AP tradicional) e inferencia de pertenencia extracción de datos de entrenamiento y extracción de modelos (AP colaborativo).

Referencias

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. pp. 308–318 (2016)
2. Aono, Y., Hayashi, T., Wang, L., Moriai, S., et al.: Privacy-preserving deep learning via additively homomorphic encryption. IEEE Transactions on Information Forensics and Security **13**(5), 1333–1345 (2017)
3. Bellet, A., Guerraoui, R., Taziki, M., Tommasi, M.: Fast and differentially private algorithms for decentralized collaborative machine learning (2017)
4. Chabanne, H., de Wargny, A., Milgram, J., Morel, C., Prouff, E.: Privacy-preserving classification on deep neural network. IACR Cryptol. ePrint Arch. **2017**, 35 (2017)

5. Chaudhuri, K., Monteleoni, C., Sarwate, A.D.: Differentially private empirical risk minimization. Journal of Machine Learning Research **12**(3) (2011)
6. Dhawale, C.A., Dhawale, K., Dubey, R.: A review on deep learning applications. In: Deep Learning Techniques and Optimization Strategies in Big Data Analytics, pp. 21–31. IGI Global (2020)
7. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608 (2017)
8. Dwork, C.: A firm foundation for private data analysis. Communications of the ACM **54**(1), 86–95 (2011)
9. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Theory of cryptography conference. pp. 265–284. Springer (2006)
10. Dwork, C., Roth, A., et al.: The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science **9**(3-4), 211–407 (2014)
11. Fredrikson, M., Jha, S., Ristenpart, T.: Model inversion attacks that exploit confidence information and basic countermeasures. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. pp. 1322–1333 (2015)
12. Fredrikson, M., Lantz, E., Jha, S., Lin, S., Page, D., Ristenpart, T.: Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In: 23rd {USENIX} Security Symposium ({USENIX} Security 14). pp. 17–32 (2014)
13. Ha, T., Dang, T.K., Dang, T.T., Truong, T.A., Nguyen, M.T.: Differential privacy in deep learning: An overview. In: 2019 International Conference on Advanced Computing and Applications (ACOMP). pp. 97–102. IEEE (2019)
14. Jagannathan, G., Pillaipakkamnatt, K., Wright, R.N.: A practical differentially private random decision tree classifier. In: 2009 IEEE International Conference on Data Mining Workshops. pp. 114–121. IEEE (2009)
15. Kifer, D., Smith, A., Thakurta, A.: Private convex empirical risk minimization and high-dimensional regression. In: Conference on Learning Theory. pp. 25–1 (2012)
16. Lyu, L., He, X., Law, Y.W., Palaniswami, M.: Privacy-preserving collaborative deep learning with application to human activity recognition. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. pp. 1219–1228 (2017)
17. McSherry, F., Talwar, K.: Mechanism design via differential privacy. In: 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07). pp. 94–103. IEEE (2007)
18. Meints, M., Moller, J.: Privacy preserving data mining–a process centric view from a european perspective (2004)
19. Micali, S., Goldreich, O., Wigderson, A.: How to play any mental game. In: Proceedings of the Nineteenth ACM Symp. on Theory of Computing, STOC. pp. 218–229 (1987)
20. Nasr, M., Shokri, R., Houmansadr, A.: Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In: 2019 IEEE Symposium on Security and Privacy (SP). pp. 739–753. IEEE (2019)
21. Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I., Talwar, K.: Semi-supervised knowledge transfer for deep learning from private training data. arXiv preprint arXiv:1610.05755 (2016)
22. Phan, N., Wang, Y., Wu, X., Dou, D.: Differential privacy preservation for deep auto-encoders: an application of human behavior prediction. In: Aaai. vol. 16, pp. 1309–1316 (2016)

23. Shokri, R., Shmatikov, V.: Privacy-preserving deep learning. In: Proceedings of the 22nd ACM SIGSAC conference on computer and communications security. pp. 1310–1321 (2015)
24. Tanuwidjaja, H.C., Choi, R., Kim, K.: A survey on deep learning techniques for privacy-preserving. In: International Conference on Machine Learning for Cyber Security. pp. 29–46. Springer (2019)
25. Tramèr, F., Zhang, F., Juels, A., Reiter, M.K., Ristenpart, T.: Stealing machine learning models via prediction apis. In: 25th {USENIX} Security Symposium ({USENIX} Security 16). pp. 601–618 (2016)
26. Xu, C., Ren, J., Zhang, D., Zhang, Y., Qin, Z., Ren, K.: Ganobfuscator: Mitigating information leakage under gan via differential privacy. IEEE Transactions on Information Forensics and Security **14**(9), 2358–2371 (2019)
27. Yao, A.C.C.: How to generate and exchange secrets. In: 27th Annual Symposium on Foundations of Computer Science (sfcs 1986). pp. 162–167. IEEE (1986)
28. Zhang, D., Chen, X., Wang, D., Shi, J.: A survey on collaborative deep learning and privacy-preserving. In: 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC). pp. 652–658. IEEE (2018)
29. Zhao, J., Chen, Y., Zhang, W.: Differential privacy preservation in deep learning: Challenges, opportunities and solutions. IEEE Access **7**, 48901–48911 (2019)
30. Zhu, T., Li, G., Zhou, W., Philip, S.Y.: Differential privacy and applications. Springer (2017)