الجمهورية الشعبية الديمقراطية الجزائرية
democratic and popular republic of Algeria
وزارة التعليم العالي و البحث العلمي
Ministry of Higher Education and Scientific Research
المدرسة العليا للإعلام الآلي ٠80 ماي 1945٠ بسيدي بلعباس
Higher School of Computer Science 08 May 1945 - Sidi Bel Abbes

Thesis of End of Study

In view of obtaining the diploma of **Master Degree**
Field: **Computer science**

Specialty: **Computer systems engineering**

Federated Learning: A Privacy-Preserving Approach to Deep Learning in the Era of Big Data

---

Presented by:

- Abdelilah KECHIDI
- Ayoub KADDOUR

Presented on: ../09/2023

In front of the jury composed of:

- **M. ……** President
- **M.AZZA Mohammed** Supervisor
- **M. …….** Examinator

Academic Year: 2022/2023

Abstract

In recent years, the fields of Deep Learning and Big Data have seen remarkable growth, ushering in transformative changes across various domains. Deep Learning, an integral part of artificial intelligence, has revolutionized our ability to extract valuable insights from complex data. Big Data, characterized by the 4Vs (Volume, Velocity, Variety, and Veracity), presents both opportunities and challenges in harnessing vast amounts of information. However, with these advancements comes an ever-increasing need to address the critical issue of data privacy and confidentiality.

This dissertation delves into the intersection of Deep Learning, Big Data, and the challenges of handling confidential information. It explores the conventional centralized machine learning paradigm and the associated risks of data exposure, privacy violations, and scalability limitations. As organizations grapple with these concerns, the spotlight shifts to a promising solution—Federated Learning.
The core of this dissertation centers on Federated Learning, a decentralized, privacy-preserving, and collaborative approach to training machine learning models. It provides a comprehensive examination of how Federated Learning works, encompassing initialization, local training, model updates, aggregation, and the iterative process. The advantages of Federated Learning, including reduced data transfer, enhanced privacy, data localization, and support for edge computing, are extensively elucidated.

Drawing on real-world examples and use cases spanning healthcare, finance, IoT, autonomous vehicles, manufacturing, agriculture, energy grids, telecommunications, and more, this research demonstrates the practical applications of Federated Learning. Technical aspects, such as Federated Averaging, Federated Stochastic Gradient Descent, and Homomorphic Encryption, are dissected to reveal the underlying mechanisms that make Federated Learning a viable solution in various domains.

Nevertheless, the dissertation does not shy away from acknowledging the limitations and challenges of implementing Federated Learning. Communication overhead, device heterogeneity, model selection, data imbalance, regulatory compliance, and security risks are thoroughly examined, including specific threats like model poisoning

attacks, data leakage attacks, and differential privacy violations.

In conclusion, this dissertation provides a comprehensive overview of Federated Learning as a potent solution to the confluence of Deep Learning and Big Data challenges, while preserving data privacy and confidentiality. By illuminating the advantages, showcasing real-world applications, and addressing the limitations, this research equips organizations and practitioners with the knowledge required to navigate the evolving landscape of privacy-aware machine learning in an era of proliferating data.

Résumé

Ces dernières années, les domaines de l'apprentissage profond (Deep Learning) et du Big Data ont connu une croissance remarquable, entraînant des changements transformateurs dans divers domaines. L'apprentissage profond, composante essentielle de l'intelligence artificielle, a révolutionné notre capacité à extraire des informations précieuses à partir de données complexes. Le Big Data, caractérisé par les 4V (Volume, Vitesse, Variété et Véracité), présente à la fois des opportunités et des défis dans l'exploitation de vastes quantités d'informations. Cependant, avec ces avancées survient un besoin croissant de répondre à la question cruciale de la protection des données et de la confidentialité.

Cette thèse plonge dans l'intersection de l'apprentissage profond, du Big Data et des défis liés à la gestion de l'information confidentielle. Elle explore le paradigme classique de l'apprentissage machine centralisé et les risques associés à l'exposition des données, aux violations de la vie privée et aux limitations en termes de scalabilité. Alors que les organisations luttent contre ces préoccupations, l'attention se tourne vers une solution prometteuse : l'apprentissage fédéré.

Le cœur de cette thèse tourne autour de l'apprentissage fédéré, une approche décentralisée, préservant la vie privée et collaborative pour l'entraînement des modèles d'apprentissage machine. Elle offre un examen complet de la manière dont fonctionne l'apprentissage fédéré, englobant l'initialisation, la formation locale, les mises à jour des modèles, l'agrégation et le processus itératif. Les avantages de l'apprentissage fédéré, notamment la réduction du transfert de données, l'amélioration de la vie privée, la localisation des données et le support pour l'informatique périphérique, sont largement expliqués.

S'appuyant sur des exemples concrets et des cas d'utilisation du monde réel couvrant la santé, la finance, l'Internet des objets (IoT), les véhicules autonomes, la fabrication, l'agriculture, les réseaux énergétiques, les télécommunications, et bien d'autres, cette recherche démontre les applications pratiques de l'apprentissage fédéré. Les aspects techniques, tels que la moyenne fédérée, la descente de gradient stochas-

tique fédérée et le chiffrement homomorphe, sont disséqués pour révéler les mécanismes sous-jacents qui font de l'apprentissage fédéré une solution viable dans divers domaines.

Cependant, la thèse n'évite pas de reconnaître les limites et les défis de la mise en œuvre de l'apprentissage fédéré. La surcharge de communication, l'hétérogénéité des appareils, la sélection des modèles, le déséquilibre des données, la conformité réglementaire et les risques de sécurité sont examinés en détail, y compris des menaces spécifiques telles que les attaques d'empoisonnement de modèles, les attaques de fuite de données et les violations de la vie privée différentielle.

En conclusion, cette thèse offre un aperçu complet de l'apprentissage fédéré en tant que solution puissante à la convergence des défis de l'apprentissage profond et du Big Data, tout en préservant la protection des données et la confidentialité. En mettant en lumière les avantages, en présentant des applications réelles et en abordant les limites, cette recherche équipe les organisations et les praticiens avec les connaissances nécessaires pour naviguer dans le paysage en évolution de l'apprentissage automatique respectueux de la vie privée à l'ère de la prolifération des données.

مُلخص

في السنوات الأفي السنوات الأخيرة، شهدت مجالات التعلم العميق والبيانات الضخمة نموًا ملحوظًا، مما أدى إلى تغييرات تحولية في مجموعة متنوعة من المجالات. التعلم العميق، الجزء الأساسي من الذكاء الاصطناعي، ثور في قدرتنا على استخراج رؤى قيمة من البيانات المعقدة. البيانات الضخمة، التي تتميز بمفهوم الـ4V (الحجم، السرعة، التنوع، وصدق البيانات)، تقدم فرصًا وتحديات في استغلال كميات هائلة من المعلومات. ومع ذلك، مع هذه التطورات تأتي حاجة متزايدة دائمًا لمعالجة القضية الحرجة لحماية البيانات والسرية.

جاذبيةالتعلم التفاعلي تكمن في قدرته على تنسيق االستفادة الفعالة من البياناتتتعمق هذه الرسالة في تقاطع التعلم العميق والبيانات الضخمة والتحديات المتعلقة بمعالجة المعلومات السرية. إنها تستكشف نموذج التعلم الآلي المركزي التقليدي والمخاطر المرتبطة بتعريض البيانات وانتهاك الخصوصية والقيود على القدرة على التوسع. مع مواجهة المؤسسات لهذه المخاوف، ينتقل الضوء إلى حلاً واعدًا - وهو التعلم التفاعلي (Federated Learning).

يتمحور جوهر هذه الرسالة حول التعلم التفاعلي (Federated Learning)، وهو نهج لتدريب نماذج التعلم الآلي غير المركزي والمحافظ على الخصوصية والتعاوني. إنها توفر استعراضًا شاملاً لكيفية عمل التعلم التفاعلي، بما في ذلك البدء، والتدريب المحلي، وتحديث النموذج، والتجميع، والعملية التكرارية. يتم شرح بالتفصيل مزايا التعلم التفاعلي، بما في ذلك تقليل نقل البيانات، وتعزيز الخصوصية، وتوجيه البيانات، ودعم الحوسبة على الحواف.

خلال الاعتماد على أمثلة من العالم الحقيقي وحالات الاستخدام التي تشمل الرعاية الصحية، والأمور المالية، وإنترنت الأشياء، والمركبات المستقلة، والتصنيع، والزراعة، وشبكات الطاقة، والاتصالات، والمزيد، تظهر هذه البحث التطبيقات العملية للتعلم التفاعلي. يتم تفكيك الجوانب التقنية مثل المتوسط التفاعلي، والجراد الذي يتراوح على النموذج التفاعلي، والتشفير الهوموورفي للكشف عن الآليات الأساسية التي تجعل التعلم التفاعلي حلاً قابلاً للتنفيذ في مجموعة متنوعة من المجالات.

ومع ذمع ذلك، لا تتجنب الرسالة التعرف على الحدود والتحديات التي تواجه تنفيذ التعلم التفاعلي. يتم فحص

العبء التواصلي، وتنوع الأجهزة، واختيار النموذج، وعدم توازن البيانات، وامتثال التنظيم، ومخاطر الأمان بتفصيل، بما في ذلك التهديدات المحددة مثل هجمات تلويث النماذج، وهجمات تسرب البيانات، وانتهاكات الخصوصية التفاضلية.

باألساس،تقف هذه الرسالة كمعلم بارز في رحلة تحقيق إلمكانيات الحقيقية للتعلفي الختام، تقدم هذه الرسالة لمحة شاملة عن التعلم التفاعلي كحلاً قوياً لتقاطع تحديات التعلم العميق والبيانات الضخمة، مع الحفاظ على حماية البيانات والسرية. من خلال تسليط الضوء على المزايا، وعرض الحالات العملية الحقيقية، ومعالجة الحدود، تزود هذه البحث المؤسسات والممارسين بالمعرفة اللازمة للتنقل في المناظر المتطورة للتعلم الآلي الذي يحترم الخصوصية في عصر انتشار البيانات.