Projet d'apprentissage profond
« Differentially Private Releasing via Deep Generative Model »

Alexandre Huat
Master Science des Données
INSA Rouen Normandie
alexandre.huat@insa-rouen.fr

18 février 2018

Référence X. Zhang, S. Ji et T. Wang, « Differentially Private Releasing via Deep Generative Model », *ArXiv e-prints*, jan. 2018. arXiv : 1801.01594 [cs.CR].

Supposons un groupe de clients et un prestateur de services informatiques collectant et analysant leurs données (*e.g.* classification d'images). Au cours des tâches d'analyses, le prestateur est amené à traiter des données sensibles. Afin de respecter la vie privée de ses clients, il doit trouver un moyen de traiter efficacement ces données tout en conservant leur confidentialité. C'est à cette problématique que répondent Zhang *et al.* par l'architecture dp-GAN de l'article résumé ici.

La Figure 1 illustre le rôle de dp-GAN, qui est de générer des données synthétiques mais sémantiquement riches, *i.e.* suffisamment représentatives des clients, sans violation de leur vie privée. En introduction, les auteurs rappellent les défis à relever en apprentissage sous contrainte de confidentialité et présentent les apports de dp-GAN. En plus de sa cauthoryearcité à générer une infinité de données, dp-GAN garantit l'anonymisation des données réelles par respect du principe de « confidentialité différentielle » [1]. Pour ce faire, dp-GAN applique au réseau adverse génératif de Wisserstein (WGAN) amélioré des mécanismes d'anonymisation à l'état de l'art, alors optimisés pour gagner en stabilité et en scalabilité.

En Section 2, Zhang *et al.* font un rappel théorique sur les GAN et justifient leur utilisation du WGAN amélioré par une plus grande stabilité et un plus court temps d'apprentissage que le GAN originel. Ils font également la définition formelle de la confidentialité différentielle et citent des propriétés associées dont bénéficient dp-GAN.

La Section 3 présente dp-GAN dans sa version basique et fournit son algorithme d'apprentissage. Pour assurer sa confidentialité, à chaque mise-à-jour, l'algorithme bruite le gradient du discriminateur (bruitage gaussien et seuillage), à partir duquel un pirate pourrait autrement reconstruire les données privées ; cette technique est communément utilisée dans la littérature. Une preuve théorique du niveau de confidentialité différentielle atteint par l'algorithme est également apportée.

Néanmoins, cette version de dp-GAN souffre de trois inconvénients : elle génère des données de faible qualité ; elle converge moins rapidement que le GAN non-confidentiel, voire diverge ; elle est rigide et n'exploite aucune ressource bonus, *e.g.* des données publiques. Pour palier ces défauts, la version avancée de dp-GAN implémente : un regroupement des paramètres du réseau pour un réglage fin et spécifique de leurs bruits respectifs ; un seuillage adaptatif du gradient, qui évolue au cours des itérations ; une initialisation des paramètres du réseau par pré-apprentissage sur les données publiques disponibles. Ces améliorations boostent la vitesse de convergence et la confidentialité de dp-GAN. La Section 4 détaille l'algorithme de cette version avancée.

S'en suit un rapport d'expériences sur trois bases célèbres et *open source*, étendues à quatre :

1. Synthétiquement, la confidentialité différentielle mesure la cauthoryearcité d'un tiers à déduire des données privées des résultats d'un algorithme ; *cf.* exemple à https://fr.wikipedia.org/wiki/Confidentialité_différentielle#Formalisation.

MNIST [2], CelebA [3] et LSUN [4], LSUN étant divisée en deux bases, l'une labellisée (LSUN-L) et l'autre non-labellisée (LSUN-U). Les expériences sont réalisées avec TensorFlow mais le code n'est pas partagé par ses auteurs. Les paramètres testés sont cependant renseignése, entre autres, le ratio données publiques sur données privées est fixé à 2 contre 98 pour chaque. La Section 5 propose ainsi une évaluation qualitative et quantitative des performances du système. De mon point de vue, les images générées par dp-GAN, quelle que soit la base, sont assez vraisemblables ; MNIST en particulier est très bien simulée. Dans leur deux premières expériences, Zhang *et al.* comparent quantitativement la qualité des données générées par dp-GAN aux données réelles et à celles générées par le GAN classique non-confidentiel : ses performances sont légèrement moins bonnes pour les données labellisées (plus faible score d'Inception) et non-labellisées (plus grand score de Jensen-Schannon). Dans leur troisième expérience, ils comparent les performances atteintes en classification sur LSUN-L après apprentissage sur les données réelles seules, puis sur les données réelles jointes aux données synthétisées par un GAN non-confidentiel et par dp-GAN. Il en ressort que l'apprentissage avec les données synthétiques de dp-GAN permet une diminution systématique des taux d'erreurs (jusqu'à -3.3 %, contre -8.7 % pour le GAN non-confidentiel). Quatrièmement, en terme de score d'Inception et de Jensen-Schannon, une ultime expérimentation valide l'efficacité des stratégies optimisations de dp-GAN avancé.

Enfin, les auteurs consacrent une section aux travaux similaires de la littérature. Puis, ils concluent en rappelant l'intérêt et les améliorations apportées par dp-GAN et ouvrent la discussion sur la limite que l'architecture n'a été testée que sur des images ; une évaluation sur d'autres types de données (*e.g.* texte) étant bienvenue.

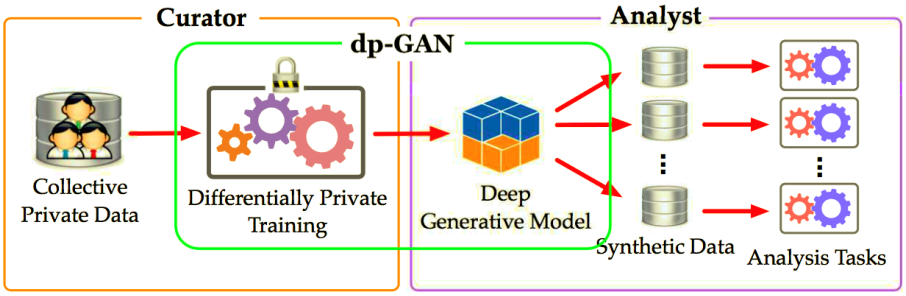

Figure 1. La place de dp-GAN dans la chaîne de traitement des données confidentielles. Le « *curator* » est l'entité qui anonymise les données pour l'*analyst*.

Implémentation Étant donnée ma formation en apprentissage statistique et profond, le véritable défi de ce projet est l'implémentation de dp-GAN basique. En effet, dp-GAN avancé consiste en l'enrichissement de dp-GAN basique par (*cf.* Algorithme 4) : des opérations sur des données publiques identiques à celles réalisées sur les données privées (apprentissage et descente de gradient) et un clustering hiérarchique ascendant des paramètres du réseau. Ces opérations ne présentent pas de plus-value pédagogique. En revanche, dp-GAN basique nécessite la compréhension de notions nouvelles (*cf.* Algorithme 1) : le fonctionnement détaillé de WGAN et ce qu'est un « *privacy accountant* ». Sachant le temps allouable au projet, il apparaît raisonnable de se concentrer sur l'implémentation de dp-GAN basique, testée sur MNIST, CelebA et/ou LSUN.

2. http://yann.lecun.com/exdb/mnist/
3. http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html
4. http://lsun.cs.princeton.edu/2015.html