Résumé de l'article « Differentially Private Releasing via Deep Generative Model »

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

17 février 2018

Référence : Zhang, Xinyang, Shouling Ji et Ting Wang (2018). « Differentially Private Releasing via Deep Generative Model ». Version 1. In : ArXiv e-prints. arXiv : 1801.01594 [cs.CR].

Supposons un groupe de clients et un prestateur de services informatiques collectant et analysant leurs données (*e.g.* image, texte, audio). 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, Ji et Wang (2018), via la méta-architecture dp-GAN [1] de l'article résumé ici.

Le rôle de dp-GAN est de générer des données synthétiques mais sémantiquement riches, *i.e.* suffisamment représentatives des clients, qui pourront être utilisées sans violation de leur vie privée (*cf.* Figure 1). 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 capacité à générer une infinité de données, dp-GAN garantit l'anonymisation des données réelles par respect du principe de « confidentialité différentielle » [2]. Pour ce faire, dp-GAN combine le réseau adverse génératif de Wisserstein amélioré (IWGAN) et 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 d'un IWGAN par une plus grande stabilité et un plus court temps d'apprentissage que le GAN

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

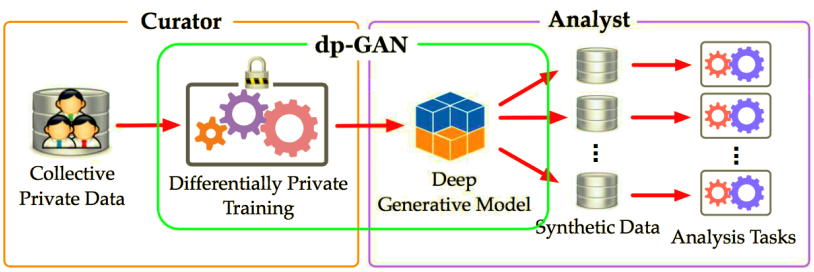

Figure 1. La place de dp-GAN dans la chaîne de traitement des données confidentielles. Le « *curator* » est l'entité qui anonymise les données privées pour l'*analyst*.

originel. Ils rappellent également la définition formelle de la confidentialité différentielle et citent des propriétés associées dont bénéficient dp-GAN.

La Section 3 de l'article 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), qui pourrait autrement être utilisé par un pirate pour 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 l'IWGAN 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 groupage des paramètres pour un réglage fin et spécifique de leurs bruits respectifs ; un seuillage du gradient qui évolue au cours des itérations ; une initialisation des paramètres du réseau par pré-apprentissage sur les données publiques à disposition. 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 : MNIST [3], CelebA [4] et LSUN [5], 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 renseignés. 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, quelque soit la base, sont assez vraisemblables ; MNIST en particulier est très bien representée mais il subsiste un léger bruit de pixels isolés sur chaque image. L'évaluation quantitative, quant à elle, repose sur deux métriques statistiques. D'une part, le score *Inception*, pour les données labellisées, défini par :

$$\mathrm{Inc}(G) = \exp(\mathbb{E}_{x\sim G(z)}[\mathrm{KL}(\mathbb{P}(y \mid x) \,\|\, \mathbb{P}(y))]) \tag{1}$$

où G est un générateur. D'autre part, le score de Jensen-Schannon, pour les données non-labellisées, défini par :

$$\mathrm{JS}(G) = \frac{1}{2}\mathrm{KL}(\mathbb{P}(y \mid x) \,\|\, \mathcal{B}_p) + \frac{1}{2}\mathrm{KL}(\mathcal{B}_p \,\|\, \mathbb{P}(y \mid x)) \tag{2}$$

où $\mathcal{B}_p$ est une loi de Bernouilli de paramètre $p = \frac{1}{2}$.

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