

CentraleSupélec
Projet long du cursus Supélec
Encadré par : J. Tomasik et A. Rimmel

Dessine-moi un mouton

Generative Adversarial Network

François Bouvier d'Yvoire
Matthieu Delmas
Romain Poirot
Paul Witz

Étude des réseaux de neurones en perceptrons
avec application au concept des Generative Adversarial Network

Années 2017-2018

Dessine-moi un mouton

Generative Adversarial Network

Résumé

Résumé

Mots-clefs

Mots-clefs

Table des matières

1	Introduction aux réseaux de neurones et premières applications	1
1.1	Outils utilisés pour le projet	1
1.2	Réseaux de neurones et fonctionnement	1
1.2.1	Le neurone	2
1.2.2	Réseau de neurones et perceptron	2
1.2.3	Apprentissage par rétro-propagation	3
1.3	Conception logicielle des réseaux de neurones	4
1.3.1	Structure du code	5
1.4	Application au problème du XOR	5
1.5	Application à la base de données MNIST	7
2	Generative Adversarial Networks	10
2.1	Principes	10
2.2	Apprentissage	11
2.3	Paramètres des GANS	12
2.4	Structure et utilisation du code	14
2.5	Premiers résultats pour des GANs simples	14
2.5.1	Méthodologie initiale	14
2.6	Mode Collapse et Bruit en entrée	15
2.7	Résultats sans collapse	15
3	Améliorations classiques des Réseaux de neurones appliqués à GAN	16
3.1	Algorithmes de descente de gradient à pas adaptatif	16
3.1.1	Momentum	17
3.1.2	AdaGrad	17

3.1.3	RMSProp	17
3.1.4	Adadelta	18
3.1.5	Adam	18
3.2	Réseaux à convolution : DCGAN	19
4	Améliorations spécifiques au GAN	20
5	Axes de Recherches : WGAN	21
5.1	Problématique de la descente de gradient simultanée	21
5.2	L'approche Wasserstein GAN	22
5.3	Mise en œuvre	24
5.4	Réflexion sur l'approche	24

Chapitre 1

Introduction aux réseaux de neurones et premières applications

Le première partie de ce projet a pour but de comprendre le fonctionnement des réseaux de neurones et leurs applications à la classification. On mettra également en place une structure informatique en python pour les utiliser.

1.1 Outils utilisés pour le projet

Ce projet possède une dimension de conception logiciel. Il s'agit de programmer sans utiliser de bibliothèques existantes des réseaux de neurones efficaces et performants adaptés aux problèmes que nous souhaitons résoudre.

Étant donné l'évolution prévue de notre code, perceptron simple pour XOR ou MNIST, puis mise en place du GAN, et enfin toute sortes d'améliorations utiles, nous devons être particulièrement vigilant sur la souplesse de notre code. La programmation en équipe, sur une longue durée et avec de telles contraintes nécessitent la mise en place d'outils et certains choix techniques.

1.2 Réseaux de neurones et fonctionnement

Les réseaux de neurones font partis des piliers de l'intelligence artificielle. Leur fonctionnement est basé sur une interprétation sommaire du cerveau humain. Des neurones seules reçoivent des signaux, les traitent et renvoient un signal de sortie. Les neurones sont alors agrégés en

réseaux avec des entrées du réseaux et des sorties. On modélise la plasticité du cerveaux par des paramètres variables qui changent au cours de l'apprentissage, ce dernier se faisant en comparant des sorties attendues aux sorties obtenues.

1.2.1 Le neurone

L'unité de base du réseau est le neurone, on peut l'imaginer comme une fonction mathématique. Lui sont attribuées n entrées, chacune affectée d'un poids w_i et une fonction mathématique de \mathbb{R} dans \mathbb{R} . Le rôle du neurone sera de renvoyer le résultat de la fonction, appliquée à la somme pondérée par le poids des entrées. On pourrait ajouter un biais comme paramètre de notre neurone afin d'ajuster notre résultat (choisir quand une fonction seuil renvoie 1 par exemple).

Un exemple simple du réseau de neurones est la séparation d'un plan en deux.

Imaginons un neurone à deux entrées e_1 et e_2 , chacune attribuée d'un poids w_1 et w_2 . On affecte au neurone un biais b et une fonction d'activation seuil (Heavyside par exemple).

Notre neurone renverra : 1 si $(e_1 * w_1 + e_2 * w_2) - b > 0$ et 0 sinon.

Si les e_1 et e_2 représentent les abscisses et ordonnées d'un point du plan, on reconnaît dans l'argument de la fonction d'activation l'équation d'une droite affine. Notre neurone pourra donc distinguer les points du plan selon le côté de la droite où ils se trouvent.

On peut déjà voir qu'une modification des poids entraînera une différente délimitation du plan.

On peut donc imaginer faire « apprendre » au réseau quels points délimiter en modifiant ses poids. Nous reviendrons sur ce concept par la suite.

Cependant, les applications d'un neurone seul sont vite limitées. C'est pourquoi on va s'intéresser à en connecter plusieurs entre eux.

1.2.2 Réseau de neurones et perceptron

On a déjà vu que le neurone se prêtait bien à une séparation binaire des données. On va voir que l'organisation de neurones en réseau permet de meilleures classifications.

L'organisation du réseau se fera au moyen de couches de neurones. Dans les structures les plus basiques, la sortie d'une couche est utilisée comme l'entrée de la couche suivante. On peut imaginer des réseaux plus complexes où la sortie n'est pas réutilisée dans la couche suivante mais plusieurs couches plus loin ou dans des couches antérieures. On appelle la dernière couche, celle qui donne le résultat du réseau de neurones, la couche de sortie ; et la première couche où l'on

donne les entrées est appelée couche d'entrée. Les autres couches sont appelées des couches cachées. Cette appellation vient du fait qu'a priori, nous n'avons aucun moyen de voir ou de corriger les comportements des neurones cachés. En effet, avec la seule donnée de la sortie, l'influence des poids des couches cachées sur celle-ci n'est pas évidente.

Le perceptron est un modèle de réseau de neurones auquel on va s'intéresser particulièrement. Il s'agit d'un réseau linéaire où chaque couche est entièrement connectée à la suivante, c'est-à-dire que chaque neurone d'une couche prend en entrée toutes les sorties de la couche précédente. On ne trouve aucune boucle dans le graphe d'un perceptron, c'est donc une propagation vers l'avant. L'utilité d'avoir plusieurs couches se comprend facilement. Si on reprend notre exemple du problème de classification des points, on peut imaginer, par exemple, quatre neurones qui enverront leur sortie sur un neurone à quatre entrées. Chacun des neurones réalisera la séparation du plan en deux selon le principe déjà évoqué précédemment. Le neurone de la couche de sortie pourra réaliser facilement le rôle d'un ET logique. On vient de sélectionner un carré dans le plan. En étendant le raisonnement, on voit qu'un réseau à deux couches permet de sélectionner n'importe quelle zone convexe de l'espace des entrées (ici du plan). De même, un réseau à trois couches pourra sélectionner n'importe quelle zone concave de l'espace des entrées.

1.2.3 Apprentissage par rétro-propagation

Il existe plusieurs types d'apprentissages du réseau. Les deux grandes catégories sont l'apprentissage supervisé et l'apprentissage non supervisé. Dans le cadre du perceptron, nous utilisons seulement un apprentissage supervisé, la rétro-propagation des erreurs. Un apprentissage supervisé nécessite une base d'apprentissage à enseigner au réseau. Elle est composée d'associations entre entrées et sorties voulues. Le réseau déduira de cette base les autres cas qu'on ne lui aura pas appris.

La rétro-propagation consiste à calculer l'influence de chaque paramètre sur la sortie et à les mettre à jour en fonction de cette influence.

Les paramètres que l'on fait évoluer sont les poids et les biais. La formule de mise à jour est la suivante :

$$W(t+1) = W(t) + \eta \frac{\partial E}{\partial W}$$

avec η le pas de convergence, $\frac{\partial E}{\partial W}$ la matrice de terme général $\frac{\partial E}{\partial W_{i,j}}$

Pour pouvoir mettre à jour les poids, il faut donc calculer les $\frac{\partial E}{\partial W_{i,j}}$.

A la couche k l'influence des poids est donnée par :

$$\frac{\partial E^p}{\partial W_k} = \frac{\partial F}{\partial W}(W_k, X_{k-1}) \frac{\partial E^p}{\partial X_k}$$

Avec $\frac{\partial F}{\partial W}(W_k, X_{k-1})$ la matrice jacobienne de F par rapport à la variable W_k

Pour pouvoir calculer l'influence des poids de toutes les couches, il faut donc calculer $\frac{\partial E^p}{\partial W_k}$

On peut calculer par récurrence cette valeur pour toutes les couches.

$$\frac{\partial E^p}{\partial X_{k-1}} = \frac{\partial F}{\partial X}(W_k, X_{k-1}) \frac{\partial E^p}{\partial X_k}$$

Avec $\frac{\partial F}{\partial X}(W_k, X_{k-1})$ la matrice jacobienne de F par rapport à la variable X_k . De plus, dans un perceptron on peut noter la sortie de la couche k :

$$Y_k = W_k X_k$$

$$X_k = F(Y_k)$$

On obtient donc ces 3 équations :

$$\begin{aligned} \frac{\partial E^p}{\partial y_k^i} &= f'(x_k^i) \frac{\partial E^p}{\partial x_k^i} \\ \frac{\partial E^p}{\partial w_{k-1}^{i,j}} &= x_{k-1}^j \frac{\partial E^p}{\partial y_k^i} \\ \frac{\partial E^p}{\partial x_{k-1}^m} &= \sum_i (w_k^{im} \frac{\partial E^p}{\partial x_k^i}) \end{aligned}$$

En forme matricielle, ces équations donnent :

$$\begin{aligned} \frac{\partial E^p}{\partial Y_k} &= \text{Diag}(f'(x_k^i)) \frac{\partial E^p}{\partial x_k^i} \\ \frac{\partial E^p}{\partial W_k} &= X_{k-1}^T \frac{\partial E^p}{\partial Y_k} \\ \frac{\partial E^p}{\partial X_{k-1}} &= W_k^T \frac{\partial E^p}{\partial X_k} \end{aligned}$$

1.3 Conception logicielle des réseaux de neurones

L'un des objectifs du projet est la conception d'une librairie permettant l'implémentation de réseaux de neurones. Notre démarche est la suivante, nous cherchons à mettre en place la structure la plus simple possible mais également la plus souple possible. Ainsi nous ne cherchons pas l'exhaustivité de notre librairie, mais nous pouvons facilement la compléter dès lors que nous avons besoin de fonctionnalité supplémentaire.

Notre code est structuré autour de 3 types de classes, les classes permettant la création et le fonctionnement d'un ou plusieurs réseaux de neurones (ce sont les classes qui font l'intelligence du programme, noté *brain*), les classes apportant des outils de compréhension et de travail sur les réseaux (affichage de résultats, chargement et sauvegarde de paramètres, etc) et les classes permettant de lancer une expérience (classes *main*, instanciant les objets et les expériences).

Le projet est découpé en 2 répertoire Github, le premier correspondant au code le plus simple, fonctionnel sur le problème du XOR, le second correspondant au développement suivant. Ces derniers développement correspondent à la généralisation à tout types de problèmes d'apprentissage de perception simple, puis à la mise en place du GAN et toutes les évolutions que nous avons mis en place.

Vous pouvez retrouver les codes sur <https://github.com/Supelec-GAN/Salamandre-XOR.git> et sur <https://github.com/Supelec-GAN/Salamandre-Code.git>.

1.3.1 Structure du code

Afin de pouvoir implémenter facilement les calculs matricielles obtenus plus tôt, nous définissons notre plus bas niveau d'intelligence par une classe *NeuronLayer* représentant une couche de neurone.

Une couche de neurones est défini par sa matrice de poids *weights*, son vecteur de biais *biais* ainsi que sa fonction d'activation *activation_function*, nécessairement commune à tout les neurones dans cette structure.

Présentation des méthodes

— *compute* : Propagation d'une entrée au sein de cette couche

—

1.4 Application au problème du XOR

Lorsque l'on souhaite travailler sur des algorithmes d'apprentissage par ordinateur, il est recommandé de les essayer sur des problèmes connus afin d'en vérifier les performances.

Le problème du XOR est l'un des plus classiques car il apporte de nombreuses difficultés.

L'objectif du XOR est de séparer le plan complexe en quatre cadrants, $(x > 0, y > 0)$, $(x > 0, y < 0)$, $(x < 0, y > 0)$ et $(x < 0, y < 0)$. Pour l'expérimentation, on restreint le plan à $[-1; 1]^2$. Les sorties attendues par le réseau de neurones sont alors 1 pour les points tel que $x * y > 0$ et -1 pour les points tels que $x * y < 0$.

Le premier intérêt de ce problème est qu'il est non linéaire. Cela se traduit par le fait qu'une droite séparant le plan en 2 ne répond pas du tout au problème.

C'est en se basant sur la résolution du XOR que nous avons construit notre structure de réseau et vérifié la cohérence de notre code. La littérature propose comme réseau le plus simple pour ce problème une couche cachée de 2 neurones, avec 2 entrées (x et y) et 1 sortie dans $[-1, 1]$. Nous avons étudié également quelques autres formes de réseaux pour comparer les résultats.

Notion de résultats La notion de résultats nécessite d'être correctement définie afin de pouvoir être interprétée correctement, en particulier pour la comparaison à d'autres résultats obtenus par nous-même ou par d'autres personnes.

La structure de perceptron sous cette forme classe les objets que l'on donne en entrée. Généralement, le résultat est défini par rapport à un pourcentage de succès dans cette classification. Pour l'obtenir, on commence par définir une erreur relative, c'est-à-dire une distance entre la sortie cible et la sortie obtenue. Un seuil est alors appliqué afin de définir une sortie booléenne de la classification de l'entrée.

Dans le cas du XOR on met en place un seuil de 0.5, c'est à dire que, si l'erreur est inférieure à 50%, le réseau a raison. Cela peut s'interpréter comme suit : le réseau donne un résultat qui indique sa confiance dans la sortie. 1 ou 0 si il est certain que la sortie doit être 1 ou 0, 0.5 si il ne peut départager l'un ou l'autre, le seuil consiste à dire que sa réponse est celle en qui il a le plus confiance. On cherche également à évaluer la vitesse d'apprentissage. Ainsi, on calcule le pourcentage de succès du réseau à intervalles réguliers au cours de l'apprentissage. Les réseaux étant soumis à une forte composante aléatoire (l'ordre d'apprentissage, ainsi que l'initialisation des poids), on effectue des apprentissages dans les mêmes conditions plusieurs fois afin d'obtenir des courbes moyennes, et des intervalles de confiances justifiant nos résultats.

Réseau en $2 \rightarrow 2 \rightarrow 1$ Les résultats obtenus au début sur ce réseau extrêmement simple semblaient tout à fait aléatoires et nous ont permis de détecter des erreurs de traduction des équations de rétro-propagation en code Python. Nous avons finalement pu obtenir des résultats

satisfaisants, comme le montre la figure ???. Cependant, ce résultat n'était pas obtenu dans l'intégralité des apprentissages, nous fournissant des résultats très différents, comme sur la figure ???. La littérature, et en particulier les rapports des années précédentes [4] et [2], nous ont montré que le XOR n'était effectivement pas juste dans 100% des cas.

Nous avons donc soumis le réseau à de nombreux apprentissages, en faisant varier les paramètres ainsi que la forme du réseau. Voici les résultats les plus intéressants :

Conclusion sur le XOR Instabilité du réseau $2 \rightarrow 2 \rightarrow 1$ et comparaison avec le $2 \rightarrow 4 \rightarrow 1$ et le $2 \rightarrow 2 \rightarrow 2 \rightarrow 1$

Pas d'apprentissage très petit par rapport à la littérature

Influence des fonctions d'activation

1.5 Application à la base de données MNIST

Description du problème

Pour le problème du MNIST qui consiste à apprendre à reconnaître des chiffres manuscrits, les données étaient les suivantes :

- 60000 images pour l'apprentissage, avec leurs étiquettes
- 10000 images de test

Toutes les images ont une dimension de 28×28 pixels en noir et blanc. Ces sets d'images sont récupérables sur le site <http://yann.lecun.com/exdb/mnist/> sous le format IDX. L'extraction de ce format vers une liste python est faite grâce au module python-mnist.

Paramètres généraux utilisés :

Les fonctions d'activation utilisées pour toutes les expériences ici sont des sigmoïdes de paramètre μ : $\sigma_\mu(x) = \frac{1}{1+e^{-\mu x}}$

On pourra étudier l'influence de μ sur la vitesse de convergence. Puisque les fonctions d'activation utilisées sont des sigmoïdes dont la sortie est dans $[0, 1]$, les valeurs d'entrées situées entre 0 et 255 sont normalisées entre 0 et 1.

L'erreur utilisée sur la couche de sortie est l'erreur quadratique.

Les poids sont initialisés avec une répartition gaussienne centrée réduite. Les biais sont initialisés


à 0.

De bons résultats ont été obtenus avec le réseau suivant, conformément à la littérature :

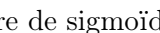
- Eta 0.2
- Sigmoides 0.1
- Réseau à une couche cachée de 300 neurones et 10 neurones de sortie (784-300-10)
- Apprentissage stochastique

Avec ce type de réseau, on obtient rapidement des taux de succès proche de 95% après 10 passes de l'ensemble du set d'apprentissages, et un écart-type final de?????. Nous allons maintenant voir l'influence des différents paramètres.

Variation d'êta :

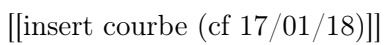
Sur le réseau 300-10 précédent, une augmentation du êta de 0.2 à 10 ne semble qu'améliorer la vitesse de convergence :  L'écart-type n'augmente pas et on obtient une meilleure précision à la fin.

Choix du paramètre de la sigmoïde :

Ici, l'influence du choix de la sigmoïde est observée. Les tests ont été effectués avec $\mu = 0.1$ et $\mu = 0.5$ comme paramètre de sigmoïdes.  On remarque qu'en tout point, choisir 0.1 en paramètre à la place de 0.5 est mieux : vitesse de convergence, précision finale, écart-type. Ce résultat empire avec un êta plus élevé, au point de ne plus réussir à apprendre. On restera donc sur une sigmoïde de paramètre 0.1 pour la suite des expériences.

Différents réseaux :

Intuitivement, un réseau avec plus de couches cachées devrait obtenir une meilleure précision, mais devrait avoir un temps d'apprentissage plus long. Ces résultats se confirment avec des expériences sur le réseau à une couche cachée (300 neurones) précédent, et un réseau sans couche cachée. Ce dernier converge très vite aussi bien vis-à-vis du nombre d'apprentissage nécessaire que du temps de calcul. Cependant, il est difficile de dépasser les 90% de succès. Alors que sur le réseau avec couche cachée, on arrive à obtenir moins de 5% d'erreur. En revanche, les temps de calculs sont plus élevés. Le réseau avec deux couches cachées, 1000 puis 300, a aussi été testé. Les résultats ici sont satisfaisants, cependant l'amélioration des résultats n'est pas très

importante, alors que les temps de calculs augmentent fortement.  Temps de calcul approximatifs pour une passe du set d'apprentissage, et un test tous les 1000 apprentissages :

- 5-6 minutes pour le 784-1000-300-10
- 4 minutes pour le 784-300-10

Ces temps peuvent être amélioré avec l'introduction du batch learning qui permet de calculer les résultats des tests plus rapidement.

Chapitre 2

Generative Adversarial Networks

La première partie de cette étude nous a permis de maîtriser l'utilisation de réseaux en perceptron et de structurer une architecture logicielle efficace et souple pour l'étude des GAN.

Après avoir obtenu des résultats satisfaisants dans la classification de motif sur la base MNIST, nous étudions la génération de données à l'aide de réseaux de neurones en nous basant sur le concept de GAN, introduit par I. Goodfellow en 2016 [?].

Notre objectif dans cette partie est d'appréhender le concept de GAN et de l'appliquer sur notre programme afin d'étudier les différents paramètres.

2.1 Principes

Le GAN s'inscrit dans les problèmes de générations de données par ordinateurs. Ses modèles cherchent à produire de données nouvelles respectant un certain nombre de contraintes. Les applications possible sont très nombreuses, tant au niveau scientifiques qu'industrielles, avec par exemple la modélisation de nouvelles protéines, le dessin de circuit intégrés, etc. Le but de nos GAN sera de générer des images que l'on ne pourra distinguer de « vraies » images, prises avec un appareil photo.

Le principe général est le suivant :

un GAN est constitué de deux réseaux de neurones, le Générateur (G) et le Discriminateur (D). Le Générateur a pour but de créer les images et le Discriminateur de déterminer si les images qu'on lui donne sont de « vraies » images ou ont été créées par le Générateur. Ces deux réseaux

sont mis en compétition : le Générateur a pour but de tromper le Discriminateur tandis que le Discriminateur doit détecter les « fausses » images.

L'apprentissage du Discriminateur se fait à la fois sur des images générées par le Générateur et de « vraies » images, issues d'une banque d'images afin de continuellement améliorer sa capacité de discernement.

L'apprentissage du Générateur dépend de la réponse du Discriminateur : lorsqu'il génère une image, on la donne au Discriminateur pour voir si le Générateur a réussi à le tromper. Le discriminateur sert donc de fonction d'erreur au Générateur.

De façon plus formelle, on travaille avec 3 distributions : p_x , la distribution idéale des vraies images, p_{data} , la distribution de l'échantillon des vraies images et p_{model} la distribution réalisée par les images issues du Générateur. Le but de l'apprentissage est de rapprocher p_{model} de p_x . Comme p_x nous est inconnu, on va plutôt s'approcher de p_{data} .

On dispose de deux fonctions de coûts $J_D(\theta_G, \theta_D)$ et $J_G(\theta_G, \theta_D)$, représentant respectivement les fonctions de coûts du Discriminateur et du Générateur. On note θ_G et θ_D les paramètres des réseaux. Les fonctions de coûts dépendent bien des paramètres des deux réseaux car le Discriminateur apprend à discerner les vraies images des fausses, dépendantes du générateur, et le générateur apprend via le résultat du Discriminateur.

C'est un problème d'optimisation simultanée. Il peut également être décrit comme un problème de jeux à informations complètes. G a accès aux données de D, mais ne peut influencer que sur θ_D et D a accès aux données de G, mais ne peut influencer que sur θ_G . Cette vision permet de déduire un algorithme où chaque joueur va faire un mouvement de manière optimale, afin de tendre vers un équilibre de Nash.

2.2 Apprentissage

L'apprentissage consiste à appliquer cette méthode de jeu à l'apprentissage des réseaux de neurones. Nous fournissons au Discriminateur des images x (de la BDD MNIST par exemple) et lui demandons de nous renvoyer un réel entre 0 et 1, qui représente son degré de confiance sur le fait que l'image fournie ait été tirée d'une banque de données authentiques ou du Générateur. Les réponses attendues sont respectivement ($D(x) = 1$) et ($D(x) = 0$) ce qui nous permet de calculer des erreurs pour la descente de gradient du Discriminateur.

Le Générateur, quant à lui, génère une image à partir d'un vecteur de bruit z . Cette image est

ensuite jugée par le Discriminateur : $D(G(z)) = 1$ si le Générateur a dupé le Discriminateur et 0 sinon. L'objectif du Générateur est d'être le plus proche possible de la première situation, l'erreur pour la descente du gradient du Générateur en est déduite.

L'apprentissage complet se fait en alternant les 2 phases successivement, chaque réseau jouant tour à tour. On parle de réseau concurrent car le Discriminateur cherche à obtenir $D(G(z)) = 0$ pour tout z et le Générateur $D(G(z)) = 1$.

Problèmes liés à la convergence L'apprentissage des GANs n'est pas simple à maîtriser car ils ne fonctionnent pas comme un seul réseau qui apprend avec une algorithme de descente de gradients. Il s'agit en réalité d'une descente de gradient simultanée (J_G et J_D) qui n'est pas un cas particulier, mais une généralisation du problème classique d'optimisation. La résolution mathématique de ce problème n'est pas trivial, et les méthodes ne s'adaptent pas facilement aux réseaux de neurones. La méthode proposée ci-dessous est donc une heuristique que l'on peut fortement adapter. Elle a fait ses preuves dans de nombreux cas, malgré son manque d'appui mathématique. Cependant d'autres façon de voir les choses permettent d'obtenir des résultats toujours corrects mais avec une plus grande rigueur. Les questionnements mathématiques seront abordés dans les axes de recherches.

2.3 Paramètres des GANS

Voici une description des paramètres principaux sur lequel on peut jouer pour l'implémentation d'un GAN. Ils sont nombreux car la description précédente est en réalité peu restrictive.

Fonctions de coût : Le premier critère intéressant des GANs est la fonction de coût (Loss Function dans la littérature) utilisé pour l'apprentissage. Il y a en réalité 2 fonctions de coût, celle pour l'apprentissage du Discriminateur et celle pour le Générateur.

La théorie des jeux nous donne, d'après [5], des fonctions qui se basent sur le maximum de vraisemblance (en particulier la Log-likelihood). Cependant l'expérience montre des limites à ces fonctions, et d'autres suggestions sont faites, par cette article ou par de nombreuses autres publications depuis, pour compenser une par. Il existe en effet de nombreuses fonctions intéressantes ayant les propriétés nécessaire : Monotonie, $\lim_{x \rightarrow 0} f(x) = 0$ et $\lim_{x \rightarrow 1} f(x) = +\infty$, le reste laissant la liberté pour les courbures

Attention : J'ai les idées en têtes mais pas encore les bons trucs mathématiques qui vont avec.

Le choix d'une fonction de coûts détermine le gradient initial que l'on propage en fonction du résultat du Discriminateur. On peut alors choisir une stratégie d'apprentissage pour le couple Discriminateur/Générateur, en insistant par exemple sur les erreurs grossières du générateur, ou alors en insistant sur ces succès.

Cela s'illustre bien avec le schéma ci-dessous, représentant différentes fonctions de coûts et les stratégies correspondantes.

Inserer ici une reprise des courbes du papier NIPS 2016 de Goodfellow, commentées

Ratios d'apprentissage : Dans l'aspect jeu de l'apprentissage, on considère que chacun des deux réseaux (Discriminateur et générateur) joue à son tour, en appliquant sa propre stratégie visant à minimiser sa fonction de coûts. Il est possible de changer les règles du jeu et d'autoriser les réseaux à jouer plusieurs coup d'affilé. Cela permet par exemple de déséquilibrer ou rééquilibrer un rapport de force, ou d'augmenter les convergences partielles du systèmes.

En effet la littérature insiste sur l'équilibre nécessaire entre les 2 structures pour obtenir le meilleur apprentissage possible. Un Discriminateur trop fort (c'est à dire $D(x) = 1$ et $D(G(z)) = 0$ de façon quasi certaines), empêche le générateur de trouver une direction d'apprentissage pertinente. L'idée est que même si le générateur s'approche d'une image réaliste, le Discriminateur lui dira qu'il se trompe. Par ailleurs, un Discriminateur trop faible ne pourrait pas non plus donner d'informations pertinentes au Générateur.

Le deuxième aspect consiste, sans déséquilibre cette fois, à augmenter le nombre de coups d'affilées des deux réseaux. Comme pour un réseau, il effectue une descente de gradient classique, avec l'autre réseau fixé, cette augmentation consiste à aller plus loin vers une convergence intermédiaire. Par exemple, pour un état D_f fixe du discriminateur, le générateur cherche à converger vers un état $G_c(D_f)$. Dans un cas idéal, une descente de gradient fera effectivement converger G vers $G_c(D_f)$. Ainsi, plus le Générateur va jouer de coup avec un Discriminateur fixe, plus il se rapproche de cette valeur.

Il n'est pas du tout évident de connaître l'influence de ses paramètres juste avec du bon sens paysan.

Paramètres classiques des réseaux de neurones : L'ensemble des paramètres des réseaux peuvent être choisis, c'est à dire la forme, les types de couches (Fully Connected, Convolution,

etc), le pas d'apprentissage (avec optimisation type RMSProp ou non), la forme de l'entrée ou tout autres optimisations existante.

En effet, la théorie du GAN n'est que très peu restrictive sur ce sujet, et libre à nous de tester les résultats avec les plus diverses paramètres.

2.4 Structure et utilisation du code

description des modifications importantes pour le GAN (or optimisation du code source)

2.5 Premiers résultats pour des GANs simples

Afin d'appréhender correctement le fonctionnement des GANs, il est nécessaire de tester de nombreux paramètres. Les résultats, fructueux ou non, permettent d'évaluer l'intérêt des paramètres et de comprendre le sens "physique" de leur influence.

Ces essais se feront sur la base MNIST, c'est-à-dire que notre objectif sera d'obtenir par le générateur des chiffres manuscrits qu'il aura dessiné par lui-même. Plus précisément on attend du générateur après apprentissage que chaque entrée génère un chiffre entre 0 et 9 qu'un être humain ne peut distinguer d'un chiffre manuscrit écrit par un humain (donc de la base MNIST). De plus, chaque entrée

2.5.1 Méthodologie initiale

L'un des objectifs de ce projet est de décortiquer au mieux la méthode GAN. Pour cela nous utilisons les réseaux de neurones les plus simples, ils seront complexifiés par la suite. La plupart des articles sur le sujet présentent des réseaux utilisant des structures avancées (couche convolutive, optimiseur de descente, etc.) il n'est donc pas possible de se référer aux paramètres de ses articles pour obtenir immédiatement des résultats et s'assurer que notre programme tournent correctement.

Pour le choix du réseau Discriminateur, le choix se porte sur des structures ayant eu de bonnes performance pour le MNIST, c'est-à-dire soit un très bon taux final, soit une vitesse de convergence élevée. En effet le Discriminateur n'est qu'un simple classificateur sur la base MNIST. Le choix du générateur est plus complexe car il faut que la structure soit assez puissante pour dessiner des chiffres. Cela semble bien plus difficile que simplement les reconnaître, car ils peuvent être

reconnus à l'aide de features très particulières, tandis que la génération exige une information complète. La première idée pour dimensionner le réseau est d'avoir une entrée suffisante pour générer la diversité de sortie souhaitée, mais pas nécessairement plus pour ne pas avoir un réseau trop lourd en calcul.

Pour les autres paramètres, aucune idée précise de leurs ordre de grandeur nécessaire, en particulier le pas d'apprentissage (très différents pour MNIST et XOR par exemple).

Nous avons donc balayer les paramètres possibles afin d'obtenir le maximum d'information avec des GANs simples.

2.6 Mode Collapse et Bruit en entrée

Le mode collapse aura été présenté à la section d'avant, tentative d'explication de pourquoi on en sort avec le bruit

2.7 Résultats sans collapse

même chose que la section premiers résultats, mais avec le bruit en entrée

Chapitre 3

Améliorations classiques des Réseaux de neurones appliqués à GAN

Les réseaux de neurones que nous utilisons pour le GAN sont de simples perceptrons. De nombreuses méthodes pour améliorer les résultats et/ou la convergence ont été proposés pour ces types de réseaux. Nous avons étudié en particulier les algorithmes de descente de gradient avec pas adaptatif et les réseaux de neurones à convolution.

3.1 Algorithmes de descente de gradient à pas adaptatif

En utilisant la descente de gradient classique, nous avons constaté que, dans le générateur, presque seule la couche de sortie travaillait. Le GAN ne générait alors pas des images d'assez bonne qualité ni assez diverses. Nous nous sommes donc intéressés à d'autres algorithmes de descente, dans l'espoir qu'ils soient plus efficaces et atteignent plus "en profondeur" les réseaux.

Les algorithmes de descente auxquels nous nous sommes intéressés sont notamment des algorithmes à pas adaptatif. En effet, dans ces algorithmes, le pas change au fur et à mesure de l'apprentissage. Il peut être grand au début, pour aller dans la bonne direction, et petit à la fin pour plus de précision.

3.1.1 Momentum

Dans une descente de gradient classique, la formule de mise à jour des poids est la suivante.

$$W_{k+1} = W_k - \eta * \frac{\partial J}{\partial W}$$

On peut également le noter :

$$\Delta W_k = -\eta * \frac{\partial J}{\partial W}$$

Une première méthode qui est compatible avec tous les algorithmes suivants est de rajouter une inertie au gradient, ou momentum. Le but est de limiter les oscillations "inutiles" qui peuvent arriver lors d'une descente de gradient. On a alors :

$$\Delta W_k = \mu * W_k - \eta * \frac{\partial J}{\partial W}$$

3.1.2 AdaGrad

AdaGrad (qui signifie Adaptive Gradient) est un algorithme où le pas change en fonction de l'erreur. On calcule la somme des carrés des gradients :

$$g_{k+1} = g_k + \left(\frac{\partial J}{\partial W}\right)^2$$

La formule de mise à jour des poids est alors :

$$\Delta W_k = -\frac{\partial J}{\partial W} * \frac{\eta}{\sqrt{g_{k+1}} + \epsilon}$$

ϵ est une valeur arbitrairement faible pour éviter une division par zéro et pour initialiser l'algorithme. L'inconvénient majeur de AdaGrad est que la quantité g_k ne peut qu'augmenter dans le temps, ce qui implique que le pas devient de plus en plus faible. Si l'apprentissage dure trop longtemps, les poids ne bougeront presque plus à cause du faible pas.

3.1.3 RMSProp

RMSProp est sensiblement identique à Adagrad mais avec une amélioration : au lieu de considérer la somme des carrés des gradients, on considère une pondération de cette somme. Cela permet de donner plus d'importance aux derniers gradients. On calcule donc :

$$g_{k+1} = \gamma * g_k + (1 - \gamma) * \left(\frac{\partial J}{\partial W}\right)^2$$

La formule de mise à jour des poids est donc identique à celle d'Adagrad :

$$\Delta W_k = -\frac{\partial J}{\partial W} * \frac{\eta}{\sqrt{g_{k+1}} + \epsilon}$$

Dans le calcul de g_k , il y a un terme quadratique. On appelle donc g_k **moment d'ordre 2**. Il existe également un moment d'ordre 1 qui se calcule par :

$$g_{k+1} = \gamma * g_k + (1 - \gamma) * \frac{\partial J}{\partial W}$$

Cela donne comme équation de mise à jour des poids :

$$\Delta W_k = -\frac{\partial J}{\partial W} * \frac{\eta}{\sqrt{g_{k+1}} + \epsilon}$$

3.1.4 Adadelta

Cet algorithme est similaire à RMSProp et utilise également une somme mobile pour calculer le moment d'ordre 2 du gradient, g_k . Cependant, au lieu d'avoir un η fixe, on introduit x_k , le moment d'ordre 2 de ΔW_k .

$$\begin{aligned} g_{k+1} &= \gamma * g_k + (1 - \gamma) * \left(\frac{\partial J}{\partial W}\right)^2 \\ x_{k+1} &= \gamma * x_k + (1 - \gamma) * (\Delta W_k)^2 \end{aligned}$$

On obtient donc :

$$\Delta W_k = -\frac{\partial J}{\partial W} * \frac{\sqrt{x_k} + \epsilon}{\sqrt{g_{k+1}} + \epsilon}$$

3.1.5 Adam

Adam (pour Adaptive Moment Estimation) adapte le pas en fonction des moments d'ordre 1 et 2 du gradient. Notons m_k le moment d'ordre 1 et v_k le moment d'ordre 2. On les calcule par :

$$\begin{aligned} m_{k+1} &= \beta_1 * m_k + (1 - \beta_1) * g_k \\ v_{k+1} &= \beta_2 * v_k + (1 - \beta_2) * g_k^2 \end{aligned}$$

Quand m_k et v_k sont initialisés à 0, ils sont biaisés vers 0. Pour pallier à cela, on considère \widehat{m}_k et \widehat{v}_k :

$$\begin{aligned} \widehat{m}_k &= \frac{m_k}{1 - \beta_1^k} \\ \widehat{v}_k &= \frac{v_k}{1 - \beta_2^k} \end{aligned}$$

On met à jour les poids avec :

$$\Delta W_k = -\frac{\eta}{\sqrt{\widehat{v}_k} + \epsilon} * \widehat{m}_k$$

3.2 Réseaux à convolution : DCGAN

Chapitre 4

Améliorations spécifiques au GAN

MiniBatch, etc

Chapitre 5

Axes de Recherches : WGAN

Introduction : Avec les différents résultats obtenus par nos premiers GAN, nous avons pu tirer, entre autres, deux conclusions importantes. Le GAN manque cruellement de stabilité, (par exemple un petit changement de paramètre l'empêche de converger correctement), et de métriques pertinentes, c'est à dire que les scores des générateurs et des discriminateurs n'ont pas d'interprétations en termes de progrès de la qualité d'image perçue.

Les chercheurs se sont beaucoup attardés depuis 2016 sur la première question, en comparant par exemple les différents optimiseurs possible [1], le deuxième point est moins souvent abordés. L'article de 2017 Wasserstein GAN [9] propose une méthode qui, en s'éloignant légèrement de la philosophie original du papier de Goodfellow [5], tente d'apporter une réponse à ces deux questions, avec en particulier une métrique pertinente.

5.1 Problématique de la descente de gradient simultanée

L'article de Goodfellow semble démontrer la convergence du système GAN, cependant la mise en œuvre montre que cette convergence n'est pas aussi évidente à obtenir. En effet, il semblerait que la stratégie de descente de gradient de l'algorithme de GAN ne permettent pas d'assurer cette convergence. Le blog inFERENCe [7] décrit une partie du problème en se basant sur l'article The Numerics of GANs [8].

Ces articles montrent que la descente de gradient simultanée n'est pas simplement une double descente de gradient, mais une descente de gradient dans un champ vectoriel. l'algorithme de

GAN effectue l'optimisation suivante :

$$x_{t+1} \leftarrow x_t + hv(x_t) \text{ avec } v(x) = \begin{pmatrix} \frac{\partial}{\partial \theta} f(\theta, \phi) \\ \frac{\partial}{\partial \phi} g(\theta, \phi) \end{pmatrix}.$$

f et g étant respectivement les fonctions de coût du Discriminateur et du Générateur. Cependant on constate 2 problèmes, d'une part l'algorithme basé sur la théorie des jeux, qui consiste à faire "jouer" tour à tour le Discriminateur et le Générateur pour optimiser ses paramètres ne consiste qu'en une approximation de la simultanéité de la descente, d'autre part il n'y a aucune preuve que le champ vectoriel x dans lequel l'on se déplace possède des propriétés conservatives. En particulier rien ne garantit que le rotationnel soit nul, ce qui implique que la descente de gradient ne soit pas garantie d'aller vers un minimum, même local !

On est donc à la recherche d'une autre approche qui contournerait ce problème.

5.2 L'approche Wasserstein GAN

Le papier Wasserstein GAN [3], propose une autre approche que celle de la théorie des jeux.

Il s'agit de calculer une divergence entre distribution afin de se servir de cette métrique pour faire l'apprentissage de l'une sur l'autre. Au premier abord cette approche ne semble pas si différente de l'approche de Goodfellow, mais elle en est sensiblement différente. Dans l'approche précédente, l'on tente de minimiser la divergence entre deux distribution par un algorithme utilisant 2 réseaux de neurones, cependant, nous n'avons jamais accès à cette divergence (que l'on utilise des fonctions d'erreur pour approcher la KL-divergence ou une autre), et comme nous l'avons montré précédemment la convergence n'est pas réellement assurée.

L'idée du papier Wasserstein GAN est de calculer explicitement une divergence entre deux ensemble. Cela n'est pas une question simple, comme nous avons pu le voir au chapitre 1. En effet nous n'avons généralement pas accès à la distribution $p_{\text{réel}}$ mais uniquement à un échantillon tiré de cette distribution.

Cependant il apparaît possible de calculer une divergence entre deux ensemble à l'aide des réseaux de neurones. Cela est possible en tout cas avec la divergence de Wasserstein, comme le montre ce papier, nous allons revenir sur les étapes de raisonnements.

L'objectif est de rapprocher 2 distributions en utilisant la métrique de Wasserstein à l'ordre 1.

Celle-ci s'écrit :

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in (\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y) \sim \gamma} [\|x - y\|]$$

On peut la nommer également distance Earth-Mover, c'est à dire distance du déplacement de terre. En effet cette métrique calcul l'effort à faire pour passer d'une distribution à l'autre. Par exemple si l'on a deux terrain contenant des tas de terre, la hauteur de terre en un point représente la densité de probabilité à cette endroit, le volume complet de terre étant le même (cela représente l'intégrale sur le terrain), alors la distance EM entre les deux terrains est l'effort minimal que l'on peut faire pour déplacer la terre de l'un des terrains pour le faire ressembler à l'autre. Il y a une infinité de façon de déplacer la terre, avec des efforts différents, la distance de Wasserstein, C'est le coût en utilisant le plan de transport optimal. Cela se traduit également en terme de probabilité jointes, pour une approche plus mathématique.

L'idée d'utiliser cette divergence de Wasserstein provient des propriétés mathématique qui devrait la rendre plus pertinente pour l'apprentissage des GANs, le détails se trouve dans se papier, mais cela se résume à dire que cette divergence donne plus d'information que la KL-Divergence et ses dérivées (Jensen-Shannon, etc) en étant entre autre bien défini lorsque les supports de distribution sont disjoints et en étant moins souvent constant, avec donc des gradients non nuls et plus adapté à l'apprentissage.

On ne peut toujours pas se servir de cette divergence, mais un théorème (la dualité Kantorovich-Rubinstein [10]) permet d'obtenir une nouvelle forme de cette divergence :

$$W(\mathbb{P}_r, \mathbb{P}_g) = \sup_{\|f\|_L < 1} \mathbb{E}_{x \sim \mathbb{P}_r} [f(x)] - \mathbb{E}_{x \sim \mathbb{P}_g} [f(x)]$$

Vous pouvez en voir une preuve simplifié sur le blog de Vincent Hermann [6] On se retrouve alors à calculer un sup sur un ensemble de fonction (les fonctions 1-Lipschitziennes), et cela est pratique car c'est justement ce que permet de faire un réseau de neurones : Simuler des fonctions que l'on optimise par rapport à un paramètre ! Il ne reste qu'à s'assurer que les réseaux de neurones peuvent garantir le caractère 1-Lipschitzien.

Une méthode pour s'assurer de cette propriété est de restreindre les poids dans un intervalle $[-c, c]$ (on parle de weight-clipping). Cependant on obtient un caractère K-Lipschitzien, avec un K inconnu. (En terme de preuve, il suffit de voir que des matrice avec ses propriétés sont toutes K-Lipschitziennes avec un même K.) cela nous assure que l'on peut calculer non pas $W(\mathbb{P}_r, \mathbb{P}_g)$ mais $K * W(\mathbb{P}_r, \mathbb{P}_g)$, mais le K étant fixe tout au long de l'apprentissage cela reste une métrique pertinente.

Nous avons donc la possibilité avec un réseau de neurone de calculer la métrique de Wasserstein et nous allons pouvoir nous en servir pour l'apprentissage d'un générateur

5.3 Mise en œuvre

5.4 Réflexion sur l'approche

Bibliographie

- [1] Comparaison des optimiser pour le gan.
- [2] Maxime Amossé, Julien Hemery, Hugo Hervieux, Sylvain Pascou, Arpad Rimmel, and Joanna Tomasik. PinaPL Réseaux de neurones & LSTM. Technical report, 2017.
- [3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv :1701.07875*, 2017.
- [4] Vincent Auriau, Laurent Beaughon, Marc Belicard, Yaqine Hechaichi, Thaïs Rahoul, Pierre Vigier, Joanna Tomasik, and Arpad Rimmel. Apprentissage Automatique de séquences. Technical report, 2017.
- [5] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.
<http://www.deeplearningbook.org>.
- [6] Vincent Hermann. Wasserstein gan and the kantorovich-rubinstein duality. <https://vincentherrmann.github.io/blog/wasserstein/>.
- [7] Ferenc Huszár. Gans are being fixed in more than a way. <http://www.inference.vc/gans-are-being-fixed-in-more-than-one-way/>.
- [8] Sebastian Nowozin Lars Mescheder and Andreas Geiger. The numerics of gan.
- [9] Soutmith Chintala Martin Arjovsky and Léon Bottou. Wasserstein gan. 2017.
- [10] Cédric Villani. Optimal transport : Old and new. 2009.